THE UNIVERSITY OF CHICAGO

WEAK FACTORS AND SUPERVISED PRINCIPAL COMPONENTS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE UNIVERSITY OF CHICAGO
BOOTH SCHOOL OF BUSINESS
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
DAKE ZHANG

CHICAGO, ILLINOIS
JUNE 2024

To everyone who has guided, encouraged, and supported me throughout this journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I extend my deepest gratitude to my advisor, Prof. Dacheng Xiu, who has guided me since my master's studies. Dacheng has not only been an exceptional academic advisor but also a friend. His work passion for research have profoundly influenced me and lead me to the worlds of econometrics and finance research.

I am also immensely grateful to another co-author Prof. Stefano Giglio for his knowledge and guidance on our research projects. My appreciation extends to the rest of my committee: Prof. Christian Hansen and Prof. Jeffrey Russell, whose insights and support have been invaluable throughout my PhD career.

My time at Booth has also been enriched by the exceptional support from our PhD office, for which I am thankful. Furthermore, I have been fortunate to collaborate with and learn alongside fellow doctoral students Qing Yan, Zhan Lin, Boxiang Lyu, Boxin Zhao, Jizhou Liu, Zhouyu Shen and others. The friendships we formed have been a great source of joy throughout my PhD and I am deeply thankful.

Finally, my deepest and most personal thanks go to my family. To my parents, Zhixiong Zhang and Jin Du, who have always encouraged my curiosity and supported my academic endeavors with unconditional love. To my fiancée, Yaxin Huang, whose support and companionship have been my anchor through my PhD journey; meeting you in Chicago is the best thing happens to me.

ABSTRACT

In macroeconomic forecasting, principal component analysis (PCA) has been the most prevalent approach to the recovery of factors, which summarize information in a large set of macro predictors. Nevertheless, the theoretical justification of this approach often relies on a convenient and critical assumption that factors are pervasive. This thesis, however, delves into the terrain of 'weak factors'—elements that are not pervasively influential but nonetheless critical for precise predictions.

To incorporate information from weaker factors, in Chapter 1, 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. Our approach is theoretically supported within an asymptotic framework where sample size and cross-sectional dimension may increase at potentially different rates.

In Chapter 2, we transition the discussion to empirical asset pricing, where weak factors and the selection of test assets are identified as interconnected challenges. Since weak factors are those to which test assets have limited exposure, an appropriate selection of test assets can improve the strength of factors. Building on this insight, we design the SPCA methodology for risk premia estimation and factor model diagnosis. The theoretical efficacy of this approach is validated through its asymptotic properties.

Chapter 3 showcases SPCA's empirical applications. The first application highlights the role of weak factors in predicting inflation, industrial production growth, and changes in unemployment. The second application employs SPCA to estimate the risk premia of a variety of observable factors, and to diagnose observable factor models. All chapters are adopted from my joint research work with Stefano Giglio and Dacheng Xiu in Giglio et al. [2023] and Giglio et al. [2021].

CHAPTER 1

PREDICTION WHEN FACTORS ARE WEAK

1.1 Introduction

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

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

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

In this paper we study a setting in which factors are sufficiently weak that PCA fails to recover them. We propose a new approach to dimension reduction for forecasting, based on *supervised PCA* (SPCA). The key idea of supervised PCA is to select a subset of predictors that are correlated with the prediction target before applying PCA. The concept of supervised PCA originated from a cancer diagnosis technique applied to DNA microarray

data by Bair and Tibshirani [2004], and was later formalized by Bair et al. [2006] in a prediction framework, in which some predictors are not correlated with the latent factors that drive the outcome of interest. Bai and Ng [2008] generalize this selection procedure (i.e., a form of hard-thresholding) to what they call the use of targeted predictors (that include soft-thresholding as well), and find it helpful in a macroeconomic forecasting environment.

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

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

and therefore missing it does not affect the consistency of the prediction.

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

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

Our paper relates to several strands of the literature on forecasting and on dimension reduction. Within the context of forecasting using latent factors, it focuses on static approximate factor models. Dynamic factor models are developed in Forni et al. [2000], Forni and Lippi [2001], Forni et al. [2004], and Forni et al. [2009], in which the lagged values of the unobserved factors may also affect the observed predictors. It is possible to extend our approach to the dynamic factor setting, which is beyond the scope of this paper. Chao

and Swanson [2022] study estimation and forecasting within a weak-factor-augmented VAR framework. They also use a pre-selection step since factors only have influence on a subset of predictors. A unique contribution of theirs is a self-normalized score statistics for selection in place of correlation screening as in supervised PCA, which ensures consistent selection of marginally correlated predictors with vanishing Type I and II errors. Similar to Bair et al. [2006], they assume all factors to have the same order of strength and all important predictors to be marginally correlated with the target, which our iterative procedure is designed to avoid.

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

Besides supervised PCA, an alternative route taken by an adjacent literature to resolving the inconsistency of PCA is sparse PCA, which imposes sparsity on population eigenvectors, see, e.g., Jolliffe et al. [2003], Zou et al. [2006], d'Aspremont et al. [2007], Johnstone and Lu [2009], and Amini and Wainwright [2009]. Uematsu and Yamagata [2022a] adopt a variant of the sparse PCA algorithm proposed in Uematsu et al. [2019] to estimate a sparsity-induced weak factor model. Bailey et al. [2020] and Freyaldenhoven [2022] adopt a similar framework for estimating factor strength and number of factors. Because sparsity is rotation dependent, such weak factor models require rotation-specific identification assumptions, whereas

standard factor models do not. The weak factor models we consider, for instance, avoid such a sparsity assumption, which makes our approach distinct from the sparse PCA.

Our approach also shares the spirit with Bai and Ng [2008] and Huang et al. [2022]. The former suggests a hard or soft thresholding procedure to select "targeted" predictors to which PCA is then applied, without providing theoretical justification. The latter suggests scaling each predictor with its predictive slope on the prediction target before applying the PCA. Our procedure and its asymptotic justification are more involved because the eigenvalues of the factor loadings in our setting can grow at distinct and slower rates.

The rest of this chapter is organized as follows. In Section 1.2 we introduce the model, provide examples to illustrate the impact of weak factors on prediction, and develop our supervised PCA procedure. In Section 1.3, we present our approach in general settings and provide asymptotic theory for our procedure. Section 1.4 provides Monte Carlo simulations demonstrating the finite-sample performance. Section 1.5 concludes. Section 1.6 provides mathematical proofs of the main theorems and propositions.

1.2 Methodology

1.2.1 Notation

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

We use $a \vee b$ to denote the max of a and b, and $a \wedge b$ as their min for any scalars a and b. We also use the notation $a \lesssim b$ to denote $a \leq Kb$ for some constant $K > 0$ and $a \lesssim_{\mathrm{P}} b$ to

denote $a = O_{\mathrm{P}}(b)$. If $a \lesssim b$ and $b \lesssim a$, we write $a \asymp b$ for short. Similarly, we use $a \asymp_{\mathrm{P}} b$ if $a \lesssim_{\mathrm{P}} b$ and $b \lesssim_{\mathrm{P}} a$.

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

1.2.2 Model Setup

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

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

$$x_t = \beta f_t + \beta_w w_t + u_t, \tag{1.1}$$

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

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

$$y_{t+h} = \alpha f_t + \alpha_w w_t + z_{t+h}, \tag{1.2}$$

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

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

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

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

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

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

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

where $\Sigma_f \in \mathbb{R}^{K\times K}$, $\Sigma_w \in \mathbb{R}^{M\times M}$ are positive-definite matrices with

$$\lambda_K\left(\Sigma_f\right) \gtrsim 1, \quad \lambda_M\left(\Sigma_w\right) \gtrsim 1, \quad \lambda_1\left(\Sigma_f\right) \lesssim 1 \text{ and } \lambda_1\left(\Sigma_w\right) \lesssim 1.$$

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

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

condition.

Next, we assume

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

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

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

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

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

Next, we need the following moment conditions on U.

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

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 $I \subset \langle N \rangle$,

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

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], provided that u_t is i.i.d. both in time and in the cross-section. While it is tempting to impose a stronger inequality that bounds $\sup_{I \subset \langle N \rangle} \left\|U_{[I]}\right\|$ uniformly over all index sets of a given size $|I|$, the rate $|I|^{1/2} + T^{1/2}$ we desire may not hold. In fact, assuming $|I|$ is small, Cai et al. [2021] establish a uniform bound that differs from our non-uniform rate only by a log factor. When $|I|$ is large, the result on uniform bounds no longer exists to the best of our knowledge. We thereby avoid making any assumption on uniform bound over all index sets.

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

Assumption 4. *For any non-random subset $I \subset \langle N \rangle$, the factor loading $\beta_{[I]}$, and the*

idiosyncratic error $U_{[I]}$ satisfy the following conditions:

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

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

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

Assumptions 2 and 3 are the key identification conditions of the weak factor model we consider. It is helpful to compare these conditions with those spelled out by Chamberlain and Rothschild [1983]. We do not require that u_t is stationary, but for the sake of comparison here, we assume that the covariance matrix of u_t exists, denoted by Σ_u and that $\beta_w = 0$. By model setup (1.1), we have $\Sigma := \mathrm{Cov}(x_t) = \beta\Sigma_f\beta' + \Sigma_u$. Chamberlain and Rothschild [1983] show that the model is identified if $\|\Sigma_u\| \lesssim 1$ and $\lambda_K \to \infty$, which guarantees the separation of the common and idiosyncratic components in the population model. To implement this strategy, Bai [2003] provides an alternative set of conditions (Assumption C therein) on the time-series and cross-sectional dependence of the idiosyncratic components that ensure the consistency of PCA, but in the case of pervasive factors, that is $\lambda_K(\beta'\beta) \gtrsim N$.

In fact, PCA can separate the factor and idiosyncratic components from the sample covariance matrix under much weaker conditions. To see this, note that from (1.1) and $\beta_w = 0$, we have $XX' = \beta FF'\beta' + UU' + \beta FU' + UF'\beta'$. Using random matrix theory

from Bai and Silverstein [2009], $\lambda_1(UU') \lesssim_{\mathrm{P}} T + N$, if u_t is i.i.d. with $\|\Sigma_u\| \lesssim 1$. Since $T\lambda_K(\beta'\beta) \asymp_{\mathrm{P}} \lambda_K(\beta FF'\beta')$ and because of the weak dependence between U and F as in Assumption 4, the eigenvalues corresponding to the factor component $\beta FF'\beta'$ dominate the three remainder terms that are related to the idiosyncratic component U asymptotically, if $(T+N)/(T\lambda_K(\beta'\beta)) \to 0$, enabling the factor components to be identified from XX'. Wang and Fan [2017] and Bai and Ng [2021] study the setting $N/(T\lambda_K(\beta'\beta)) \to 0$, in which case PCA remains consistent despite the fact that factor exposures are not pervasive. Wang and Fan [2017] also study the borderline case $N \asymp T\lambda_K(\beta'\beta)$, and document a bias term in the estimated eigenvalues and eigenvectors associated with factors.

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

1.2.3 Prediction via Supervised Principal Components

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

We use a simple one factor example to illustrate the procedure, before explaining its

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

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

$$\widehat{I} = \left\{ i \middle| T^{-1} |\underline{X}_{[i]} \overline{Y}'| \geq c \right\}, \tag{1.3}$$

where c is some threshold. Therefore, we keep predictors that covary sufficiently strongly (positively or negatively) with the target. This step involves a single tuning parameter, c, that effectively determines how many predictors we use to extract the factor. The fact that $\widehat{I}$ incorporates information from the target reflects the distinctive nature of a supervised procedure. Given the existence of I_0 by Assumption 2, there exists a choice of c such that predictors within the set $\widehat{I}$ have a strong factor structure. The rest of the procedure is a straightforward application of the principal component regression for prediction. Specifically, we apply PCA to extract factors $\{\widehat{f}_t\}_{t=1}^{T-h}$ from $\underline{X}_{[\widehat{I}]}$, which can be written as $\widehat{f}_t = \widehat{\zeta}' x_t$ for some loading matrix $\widehat{\zeta}$, then obtain $\widehat{\alpha}$ by regressing $\{y_t\}_{t=1+h}^{T}$ onto $\{\widehat{f}_t\}_{t=1}^{T-h}$ based on the predictive model (1.2). The resulting predictor for y_{T+h} is therefore given by: $\widehat{y}_{T+h} = \widehat{\alpha}\widehat{f}_T = \widehat{\alpha}\widehat{\zeta}' x_T$.

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

In contrast, we assume the existence of a set I_0 within which predictors have a strong factor structure, yet we do not make any assumptions on the correlation between the target and predictors outside this set I_0, nor on the strength of their factor structure. As a result,

I_0 under our Assumption 2 needs not be unique, and we will show that the validity of the prediction procedure does not rely on consistent recovery of any pre-determined set I_0. More importantly, since marginal screening is based on marginal covariances between $\overline{Y}$ and $\underline{X}$, in a multi-factor model the condition that marginal screening can recover a subset within which all factors are pervasive (even if such a subset is uniquely defined as in Bair et al. [2006]) is rather strong. On the one hand, marginal screening can be misguided by the correlation induced by a strong factor to the extent that weak factors after screening remain unidentifiable. On the other hand, predictors eliminated by marginal screening can be instrumental or even essential for prediction. We illustrate these points using examples of two-factor models below.

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

$$x_t = \left[\begin{array}{c|c} \beta_{11} & \beta_{12} \\ \hline \beta_{21} & 0 \end{array}\right] f_t + u_t, \quad y_{t+h} = \left[\begin{array}{cc} 1 & 1 \end{array}\right] f_t, \tag{1.4}$$

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

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

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

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

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

$$x_t = \begin{bmatrix} \beta_{11} & \beta_{11} \\ \hline 0 & \beta_{22} \end{bmatrix} f_t + u_t, \qquad y_{t+h} = \begin{bmatrix} 1 & 0 \end{bmatrix} f_t, \tag{1.5}$$

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

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

After the screening step the second group of predictors would be eliminated, because

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

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

1.2.4 Iterative Screening and Projection

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

It is straightforward to verify that this iterative screening and projection approach successfully addresses the issues with the aforementioned examples. Consider first Example

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

Formally, we present our algorithm for the general model given by (1.1) and (1.2):

Algorithm 1 (Prediction via SPCA)**.**

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

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

a. Select an appropriate subset $\widehat{I}_k \subset \langle N \rangle$ *via marginal screening.*

b. Estimate the k*th 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 using* $\widehat{\beta}_{(k-1)}$ *(defined in c.).*

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

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

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

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

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

We discuss the details of the algorithm below.

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

$$\widehat{I}_k = \left\{ i \middle| T^{-1} \left\| (X_{(k)})_{[i]} Y'_{(k)} \right\|_{\text{MAX}} \geq \widehat{c}_{qN}^{(k)} \right\},$$
$$\text{where } \widehat{c}_{qN}^{(k)} \text{ is the } (1-q)th\text{-quantile of } \left\{ T^{-1} \left\| (X_{(k)})_{[i]} Y'_{(k)} \right\|_{\text{MAX}} \right\}_{i=1,\ldots,N}. \quad (1.6)$$

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

Correspondingly, the algorithm terminates as soon as

$$\widehat{c}_{qN}^{(k+1)} < c, \quad \text{for some threshold } c. \quad (1.7)$$

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

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

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

The last step of the algorithm needs more explanations. Step S1. provides a set of factor estimates, $\widehat{\underline{F}}$, on the basis of $\overline{Y}$ and $\underline{X}$. Moreover, a time series regression of $\overline{Y}$ on $\widehat{\underline{F}}$ and $\underline{W}$ yields an estimator of α_w (coefficient defined in (1.2)). That is, $\widehat{\alpha}_w = \overline{Y}\mathbb{M}_{\widehat{\underline{F}}'}\underline{W}'\left(\underline{W}\mathbb{M}_{\widehat{\underline{F}}'}\underline{W}'\right)^{-1} = \overline{Y}\underline{W}'(\underline{W}\underline{W}')^{-1}$, since $\mathbb{M}_{\widehat{\underline{F}}'}\underline{W}' = \underline{W}'$ by construction, which explains the formula for $\widehat{\alpha}_w$ in Step S2.. Finally, with $\widehat{\alpha}$, $\widehat{\alpha}_w$, and $\widehat{f}_T$, it is sufficient to construct the predicted value of y_{T+h} by combining $\widehat{\alpha}\widehat{f}_T$ with $\widehat{\alpha}_w w_T$, which yields the final prediction formula for $\widehat{y}_{T+h}$, a projection on observables, x_T and w_T.

1.3 Asymptotic Theory

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

1.3.1 Consistency in Prediction

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

Recall that in Algorithm 1, we denote the selected subsets in the SPCA procedure as $\widehat{I}_k$, $k = 1, 2, \ldots$. We now construct their population counterparts iteratively, for any given choice

of c and q. This step is critical to characterize the exact factor space recovered by SPCA. For simplicity in notation and without loss of generality, we consider the case $\Sigma_f = \mathbb{I}_K$ here, because in the general case, we can simply replace β and α by $\beta^* = \beta\Sigma_f^{1/2}$ and $\alpha^* = \alpha\Sigma_f^{1/2}$ in the following construction.

In detail, we start with $a_i^{(1)} := \left\|\beta_{[i]}\alpha'\right\|_{\text{MAX}}$ and define $I_1 := \{i|a_i^{(1)} \geq c_{qN}^{(1)}\}$, where $c_{qN}^{(1)}$ is the $\lfloor qN \rfloor$th largest value in $\left\{a_i^{(1)}\right\}_{i=1,\ldots,N}$. Then, we denote the largest singular value of $\beta_{(1)} := \beta_{[I_1]}$ by $\lambda_{(1)}^{1/2}$ and the corresponding left and right singular vectors by $\varsigma_{(1)}$ and $b_{(1)}$. For $k > 1$, we obtain $a_i^{(k)} := \left\|\beta_{[i]}\prod_{j<k}\mathbb{M}_{b_{(j)}}\alpha'\right\|_{\text{MAX}}$, $I_k := \{i|a_i^{(k)} \geq c_{qN}^{(k)}\}$, 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 $b_{(j)}$'s are well defined and identifiable, we need that the top two singular values of $\beta_{(k)}$ are distinct at each stage k. We also need distinct values of $c_{qN}^{(k)}$ to ensure that I_k's are identifiable. More precisely, we say that two sequences of variables a_N and b_N are asymptotically distinct if there exists a constant $\delta > 0$ such that $|a_N - b_N| \geq \delta|b_N|$ for sufficiently large N. In light of the above discussion, we make the following assumption:

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

This assumption is rather mild as it only rules out corner cases, despite the fact that this is not very explicit. Excluding such corner cases is common in the literature on high dimensional PCA, see, e.g., Assumption 2.1 of Wang and Fan [2017]. Assumption 5 is closely tied to our choice of the number of predictors qN and the parameter c in the stopping rule. In particular, the current algorithm adopts a strategy where the same number of predictors is

selected at each step, representing one version of SPCA. An alternative approach may involve selecting predictors based on a predetermined threshold for their covariances and stopping the selection process when $|I_k|$ becomes smaller than another threshold. By allowing for the flexibility of using varying numbers of predictors at each step, this alternative approach can be particularly useful in addressing certain corner cases ruled out by the current version of Assumption 5.[5] Similar asymptotic results, akin to those presented in Theorem 1 through 3 below, can be derived with more intricate conditions regarding the rate of convergence, etc. However, the current version of SPCA, with its more concise theorems and superior performance in simulation, is the primary focus of our discussion in the main text. We now are ready to present the consistency of the estimated factors by SPCA:

Theorem 1. *Suppose that x_t follows* (1.1) *and y_t satisfies* (1.2)*, and that Assumptions 1-5 hold. If $\log(NT)(N_0^{-1} + T^{-1}) \to 0$, then for any tuning parameters c and q that satisfy*

$$c \to 0, \quad c^{-1}(\log NT)^{1/2}(q^{-1/2}N^{-1/2} + T^{-1/2}) \to 0, \quad qN/N_0 \to 0, \tag{1.8}$$

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

$$\left\|\underline{\widehat{F}}_{(k)}\right\|^{-1} \left\|\underline{\widehat{F}}_{(k)} - \underline{\widehat{F}}_{(k)}\mathbb{P}_{\underline{F}'}\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1}. \tag{1.9}$$

We make a few observations regarding this result. First, the assumptions in Theorem 1 do not guarantee a consistent estimate of the number of factors, K, because the SPCA procedure cannot guarantee to recover factors that are uninformative about y. At the same time, the factors recovered by SPCA are not necessarily useful for prediction, because it

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.

is possible that some strong factors with no predictive power are also recovered by SPCA. Ultimately, the factor space recoverable is determined by β, α, Σ_f, c, q, and N. For this reason, we have consistency of factor estimates up to the first $\tilde{K}$ factors. Moreover, $\widehat{K}$ is a consistent estimator of $\tilde{K}$, which we prove satisfies $\tilde{K} \leq K$. That is, SPCA omits $K - \tilde{K}$ factors. Also, the inequality (1.9) 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, $\mathbb{P}_{\underline{F}'}$. (1.9) shows that this angle vanishes asymptotically.

Second, with respect to the tuning parameters, the condition (1.8) implies that $c \to 0$, $c\sqrt{T} \to \infty$, and $c\sqrt{qN} \to \infty$. On the one hand, the threshold c needs be sufficiently small so that the iteration procedure continues until selected predictors have asymptotically vanishing predictive power; on the other hand, c needs be large enough that dominates error in the covariance estimates from the screening step. The estimation error consists of the usual error in the construction of the sample covariances between $X_{(1)}$ and $Y_{(1)}$, which introduces an error of order $T^{-1/2}$, as well as the construction of residuals in the projection step, $X_{(k)}$ and $Y_{(k)}$, for $k > 1$, as soon as multiple factors are involved (i.e., $\tilde{K} > 1$). As we show next, the factor estimation error is of order $(qN)^{-1/2} + T^{-1}$, which pollutes the residuals and hence affects screening. Taking these two points into consideration, the choice of c needs dominate $T^{-1/2} + (qN)^{-1/2}$. In terms of q, it appears that the maximal number of selected predictors, $\lfloor qN \rfloor$, allowed for should be of the same order as N_0. Nevertheless, since N_0 given by Assumption 2 is not precisely defined, in the sense that the assumption holds if N_0 is scaled by any non-zero constant, we require $qN/N_0 \to 0$ to ensure that the scaling constant of N_0 does not matter for the choice of q and that the selected $\lfloor qN \rfloor$ predictors are within the subset of N_0 predictors that guarantee a strong factor structure.

Third, the estimation error of factors are bounded from the above by $q^{-1/2}N^{-1/2} + T^{-1}$. Recall that in the strong factor case, the factor space can be recovered at the rate of $N^{-1/2} +$

T^{-1}, see, e.g., Bai [2003]. In our result, qN plays the same role as N in the strong factor case. Nevertheless, our Assumption 2 does not require all factors to have the same strength. It is possible that some factors could be recovered with a higher convergence rate, should we select a different number of predictors for each factor based on its strength. In fact, an alternative choice of $\widehat{I}_k$ based on (1.3) allows different numbers of predictors to be selected at each stage, since the threshold itself is a fixed level. While this approach may achieve a faster rate for relatively stronger factors, the prediction error rate is ultimately determined by the estimation error of the weakest factor. Yet, we find that the approach based on (1.6) offers more stable prediction out of sample, whereas prediction based on (1.3) can be sensitive to the tuning parameters. Given that our ultimate goal is about prediction rather than factor recovery, we prefer a more stable procedure and thereby focus our analysis on the former approach.

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

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

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

$$\widehat{\gamma}\beta = \widehat{\alpha}H, \quad \text{so that by Theorem 2} \quad \|\widehat{\alpha}H - \alpha\| = o_{\mathrm{P}}(1). \tag{1.10}$$

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

$$\left\|\underline{\widehat{F}}_{(k)}\right\|^{-1} \left\|\underline{\widehat{F}}_{(k)} - h_k \underline{F}\right\| \lesssim_{\mathrm{P}} q^{-1/2} N^{-1/2} + T^{-1}, \tag{1.11}$$

where h_k is the kth row of H. Therefore, $\widehat{\alpha}\underline{\widehat{F}} \overset{\text{by}(1.11)}{\approx} \widehat{\alpha} H \underline{F} \overset{\text{by}(1.10)}{\approx} \alpha \underline{F}$, which, together with $\widehat{\alpha}_w - \alpha_w = o_{\mathrm{P}}(1)$, leads to the consistency of prediction.

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

Moreover, this result does not rely on normally distributed error nor on the assumption that all factors share the same strength with respect to all predictors. The assumption on the relative size of N and T is also quite flexible, in contrast with existing results in the literature in which N cannot grow faster than a certain polynomial rate of T, e.g., Bai and Ng [2021], Huang et al. [2022].

1.3.2 Recovery of All Factors

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

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

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

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

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

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

not select more factors than needed asymptotically, because the iteration is terminated as soon as all covariances vanish. With a consistent estimator of the number of factors, we can recover the factor space as well as conduct inference on the prediction targets.

The inference theory on strong factor models also relies on a consistent estimator of the count of (strong) factors, e.g., Bai and Ng [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 Section 1.3.5, we show that our version of the PCA regression is more robust than the procedure of Stock and Watson [2002a] with respect to the error due to overestimating the number of factors. We also provide simulation evidence on the finite sample performance of our estimator of the number of factors.

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

Theorem 3. *Under the same assumptions as Theorem 2, if we further have* $\lambda_{\min}(\alpha'\alpha) \gtrsim 1$*, then for any tuning parameters* c *and* q *in* (1.6) *and* (1.7) *satisfying*

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

we have

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

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

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

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

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

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

1.3.3 Inference on the Prediction Target

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

To do so, we impose the following assumption:

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

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

where Ψ is a $K \times N$ matrix whose kth row is equal to $b_{(k)}'\beta_{[I_k]}'(\mathbb{I}_N)_{[I_k]}$ and $b_{(k)}$ is the first right singular vector of $\beta_{(k)} = \beta_{[I_k]}\prod_{j<k}\mathbb{M}_{b_{(j)}}$ as defined in Section 1.3.1.

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

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

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

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

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

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

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

Π_{ij} *is specified by Assumption 6,* $\Sigma_{f,w} = \text{diag}(\Sigma_f, \Sigma_w)$*,* $\Lambda = \text{diag}(\lambda_{(1)}, \ldots, \lambda_{(K)})$*, and* B *is a* $K \times K$ *matrix whose* k*th 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 1.3.1.*

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

Algorithm 1 (Step S2.) makes predictions by exploiting the projection of y_{T+h} onto x_T and w_T, with loadings given by γ and $\alpha_w - \gamma\beta_w$. This is convenient and easily extendable out of sample, as both x_T and w_T are directly observable, unlike latent factors. Section 1.3.5 investigates potential issues with plain PCA and PLS, as well as an alternative algorithm based on Stock and Watson [2002a], which does not involve the projection parameter γ.

1.3.4 Estimation of Φ_1 *and* Φ_2

Recall that from the outputs of Algorithm 1, we have defined $\underline{\widehat{F}}$, $\widehat{\beta}$, and $\widehat{\alpha}$. As a result, we can also estimate $\widehat{Z} = Y - \widehat{\alpha}\underline{\widehat{F}} - \widehat{\alpha}_w\underline{W}$ and $\underline{\widehat{U}} = \underline{X} - \widehat{\beta}\underline{\widehat{F}} - \widehat{\beta}_w\underline{W}$. 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 $\widehat{\Sigma}_f = T_h^{-1}\underline{\widehat{F}}\underline{\widehat{F}}'$ and $\widehat{\Sigma}_w = T_h^{-1}\underline{W}\underline{W}'$. 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 consistent estimators 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 $\Sigma_u = \text{Cov}(u_t)$ following Fan et al. [2013]. In detail, $\widehat{\Sigma}_u$ can be constructed by

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

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 $\tau_{ij} = C\left(\frac{1}{\sqrt{qN}} + \sqrt{\frac{\log N}{T}}\right)\sqrt{\widehat{\theta}_{ij}}$, where $C > 0$ is a sufficiently large constant and

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

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

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

Theorem 5. *Under the assumptions of Theorem 4, if 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* $\alpha(T) = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_T^\infty} |\mathrm{P}(A)\mathrm{P}(B) - \mathrm{P}(AB)|$ *and* $\mathcal{F}_{-\infty}^0$, $\mathcal{F}_T^\infty$ *are* σ*-algebras generated by* $\{u_t\}_{-\infty \leq t \leq 0}$, $\{u_t\}_{T \leq t \leq \infty}$.

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

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

1.3.5 Alternative Procedures

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

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

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

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

1.3.5.1 Principal Component Regression

Formally, we present the algorithm below:

Algorithm 2 (PCA Regression)**.**

Inputs: $\overline{Y}$, $\underline{X}$, $\underline{W}$, x_T, *and* w_T.

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

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

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

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

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

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

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

Proposition 1 first shows that even if the number of factors is known to be 1, the factor estimated by PCA is in general inconsistent, because the eigenvector η_{PCA} deviates from that of $T^{-1}\underline{F}'\underline{F}$, as the latter is polluted by A. In the special case where error is homoskedastic

and has no serial correlation, i.e., $A_1'A_1 = \mathbb{I}_{T-h}$, the estimated factor becomes consistent, in that $\delta A_1'A_1$ in M does not change the eigenvectors of $T^{-1}\underline{F}'\underline{F}$. This result echoes a similar result in Section 4 of Bai [2003], who established the consistency of factors with homoskedasticity and serially independent error even when T is fixed. That said, while factors can be estimated consistently in this special case, the prediction of y_{T+h} based on Algorithm 2 is not consistent.

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

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, 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.

1.3.5.2 Partial Least Squares

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

The PLS regression algorithm is formulated below:

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

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

a. *Obtain the weight vector $\widehat{\varsigma}_{(k)}$ from the largest left singular vector of $X_{(k)}Y'_{(k)}$.*

b. *Estimate the kth factor as $\underline{\widehat{F}}_{(k)} = \widehat{\varsigma}'_{(k)}X_{(k)}$.*

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

e. *Remove $\underline{\widehat{F}}_{(k)}$ to obtain residuals for the next step: $X_{(k+1)} = X_{(k)} - \widehat{\beta}_{(k)}\underline{\widehat{F}}_{(k)}$ and $Y_{(k+1)} = Y_{(k)} - \widehat{\alpha}_{(k)}\underline{\widehat{F}}_{(k)}$.*

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

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

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

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

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

Proposition 3. *In Example 4, suppose that $N/(N_0T) \to \delta \geq 0$ and $\|\beta\| \to \infty$, then the estimated factor $\underline{\widehat{F}}_{PLS}$ satisfies*

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

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

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

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

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

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

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

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

where $u_t s$ *follow i.i.d.* $\mathcal{N}(0, \mathbb{I}_N)$ *and* $z_t s$ *follow i.i.d.* $\mathcal{N}(0,1)$.

Proposition 5. *In Example 5, if we use* $\widehat{K} = 1$*, then we have*

$$\widehat{y}_{T+h}^{PLS} \gtrsim_{\mathrm{P}} N^{3/2}T^{1/2}/(N^2+T^2) \text{ while } \widehat{y}_{T+h}^{PCA} \lesssim_{\mathrm{P}} 1/(N^{1/2}+T^{1/2}).$$

Specifically, in the case of $N \asymp T$*,*

$$\widehat{y}_{T+h}^{PLS} \gtrsim_{\mathrm{P}} 1 \text{ and } \widehat{y}_{T+h}^{PCA} \lesssim_{\mathrm{P}} N^{-1/2}.$$

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 also avoid overfitting.

1.3.5.3 PCA Regression of Stock and Watson [2002a]

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

Formally, we present their algorithm below:

Algorithm 4 (SW-PCA)**.**

Inputs: $\overline{Y}$*,* X*, and* W*.*

S1. Apply SVD on X*, and obtain the estimated factors* $\widehat{F}_{SW} = \widehat{\varsigma}_*' X \mathbb{M}_{W'}$*, where* $\widehat{\varsigma}_* \in \mathbb{R}^{N\times K}$ *are the first* K *left singular vectors of* X*.*

S2. Estimate the coefficients by time-series regression: $\widehat{\alpha} = \overline{Y}\mathbb{M}_{\underline{W}'}\widehat{\underline{F}}_{SW}'\left(\widehat{\underline{F}}_{SW}\mathbb{M}_{\underline{W}'}\widehat{\underline{F}}_{SW}'\right)^{-1}$[6] *and* $\widehat{\alpha}_w = \overline{Y}M_{\widehat{\underline{F}}_{SW}'}\underline{W}'\left(\underline{W}\mathbb{M}_{\widehat{\underline{F}}_{SW}'}\underline{W}'\right)^{-1}$*.*

S3. Obtain the prediction $\widehat{y}_{T+h}^{SW} = \widehat{\alpha}\widehat{f}_T + \widehat{\alpha}_w w_T$*, where* $\widehat{f}_T$ *is the last column of* $\widehat{F}_{SW}$ *and* $\widehat{\alpha}_w = \overline{Y}\underline{W}'(\underline{W}\underline{W}')^{-1}$*.*

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

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

The advantage of SW-PCA is that the consistency of factors is sufficient for the consistency of the prediction, unlike PCA as shown by Proposition 2. In other words, even though this is not true in general, $\widehat{y}_{T+h}^{SW}$ 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 6. *Suppose* x_t *and* y_{t+h} *follow a "0-factor" model:*

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

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

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

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

1.3.6 Tuning Parameter Selection

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

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

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

1.4 Simulations

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

Specifically, we consider a 3-factor DGP as given by equation (1.1) with two strong factors f_{1t}, f_{2t} and one potentially weak factor f_{3t}. For strong factors f_{1t} and f_{2t}, we generate exposure to them independently from $\mathcal{N}(0,1)$. To simulate a weak factor f_{3t}, we generate exposure to it from a Gaussian mixture distribution, drawing values with probability a from $\mathcal{N}(0,1)$ and $1-a$ from $\mathcal{N}(0,0.1^2)$. The parameter a determines the strength of the third factor and it ranges from $\{0.5, 0.1, 0.05\}$ in the simulations.

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

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

Table 1.1 compares the finite sample performance of SPCA, PCA, and PLS in the first scenario. In both panels, the sample size is $T = 60, 120$, and around $aN = 100$ predictors have exposure to the factor f_{3t}. We simulate $N = 200$ ($a = 0.5$) predictors in the upper panel, so that f_{3t} is exposed to half of them and is thereby strong, and set $N = 2,000$

($a = 0.05$) in the lower panel, where f_{3t} becomes much weaker due to the large number of predictors that do not load on it.

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

The simulation results in Table 1.1 square well with our theoretical predictions. In the strong factor case (upper panel), PCA and SPCA perform similarly. They achieve minimum prediction error when K is set at the true value 3 in that the first two factors do not predict y. This suggests that tuning q does not worsen the performance of SPCA. PLS can also achieve desirable performance but typically with K smaller than 3. Interestingly, its performance deteriorates rapidly as K increases and surpasses the true value. The reason, as we explain in Proposition 5, is that PLS is more likely to overfit as it uses information about y to directly construct predictors. In contrast, PCA based approaches are more robust to noisy factors used in prediction.

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

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

Furthermore, Table 1.2 reports the performance of SPCA, PCA, and PLS for each entry of y in the multi-target scenario. In this case we only report results with parameters tuned.

Table 1.1: Finite Sample Comparison of Predictors (Univariate y)

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

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

As discussed previously, we expect the recovery of all factors using SPCA, because to each factor, at least one entry of y_t has exposure. We first report the distance between $\widehat{F}$ and the true factors F, defined by $d(\widehat{F}, F) = \left\| \mathbb{P}_{\widehat{F}'} - \mathbb{P}_{F'} \right\|$. We also report the MSE_is for $\widehat{y}_{i,T+1}$, $i = 1, 2, 3$, where MSE_3 is based on $y_{3,T+1}$, which depends on the potentially weak factor f_{3T} by construction. Again, we vary the value of a and N, while maintaining $aN = 100$, so that the number of predictors with exposure to the third factor is fixed throughout.

The findings here are again consistent with our theory. In particular, as a varies from 0.5 to 0.05, the third factor becomes increasingly difficult to detect. Both PCA and PLS report a substantially larger distance $d(\widehat{F}, F)$ than SPCA. In the mean-time, the distortion in the factor space translates to larger prediction errors for the third target y_3, in that it loads on the weak factor f_3 besides its own lag. Throughout this experiment, SPCA maintains almost the same level of performance as a varies, demonstrating its robustness to weak factors.

Last but not least, we report the histograms of the standardized prediction errors using the CLT of Theorem 4 in Figure 1.1. The setting is identical to that of Table 1.2 with $a = 0.05$ and $T = 120$. The histograms match well with the standard normal density for SPCA, and hence verifies the central limit result we derive. As to PCA, there is visible

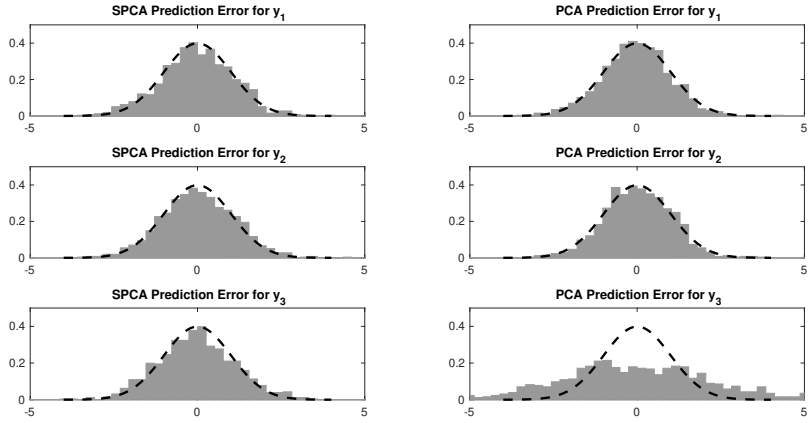
Table 1.2: Finite Sample Comparison of Predictors (Multivariate y)

	SPCA				PCA				PLS			
a	$d(\widehat{F}, F)$	MSE_1	MSE_2	MSE_3	$d(\widehat{F}, F)$	MSE_1	MSE_2	MSE_3	$d(\widehat{F}, F)$	MSE_1	MSE_2	MSE_3
						$T = 60$						
0.5	0.40	0.14	0.16	0.20	0.40	0.14	0.15	0.21	0.41	0.14	0.15	0.19
0.1	0.44	0.13	0.14	0.25	0.55	0.12	0.12	0.55	0.54	0.12	0.13	0.38
0.05	0.45	0.14	0.13	0.27	0.66	0.12	0.11	0.72	0.59	0.12	0.12	0.53
						$T = 120$						
0.5	0.30	0.07	0.08	0.10	0.30	0.07	0.08	0.10	0.31	0.08	0.08	0.10
0.1	0.31	0.07	0.07	0.12	0.36	0.06	0.06	0.22	0.39	0.07	0.06	0.17
0.05	0.31	0.07	0.07	0.11	0.39	0.06	0.06	0.29	0.44	0.06	0.06	0.22

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

distortion to normality for y_3, due to the presence of the weak factor f_3.

Figure 1.1: Histograms of the Standardized Prediction Errors



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

1.5 Conclusions

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

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

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

In practice, there is no guarantee for that to be the case. Whether a latent factor is strong or weak (and *how* strong) depends on how exposed the various predictors are to it – and each empirical applications could feature a different mix of strong and weak latent factors. Therefore, we propose a new SPCA approach that iterates a selection step, a factor extraction step, and a projection step. As we demonstrate in the paper, this procedure can consistently handle a whole range of latent factor strength. Our empirical analysis in section 3.2 shows that indeed this procedure fares well in an application with a large number of potentially noisy macroeconomic predictors.

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

Second, our theory highlights an interesting tradeoff that emerges when working with

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

1.6 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 (1.1) and α in (1.2) accordingly.

1.6.1 Proof of Theorem 1

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

$$\widehat{\varsigma}_{(k)} = T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}\widetilde{X}_{(k)}\widehat{\xi}_{(k)}, \quad \widehat{\xi}_{(k)} = T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}\widetilde{X}'_{(k)}\widehat{\varsigma}_{(k)}. \tag{1.12}$$

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

$$\widehat{\beta}_{(k)} = T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}X_{(k)}\widehat{\xi}_{(k)} \quad \text{and} \quad \widehat{\alpha}_{(k)} = T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}Y_{(k)}\widehat{\xi}_{(k)}. \tag{1.13}$$

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

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

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

$$\widetilde{X}_{(k)} = \left(X_{(k)}\right)_{[\widehat{I}_k]} = X_{[\widehat{I}_k]}\prod_{i=1}^{k-1}\mathbb{M}_{\widehat{F}'_{(i)}} = X_{[\widehat{I}_k]}\left(\mathbb{I}_{T_h} - \sum_{i=1}^{k-1}\widehat{\xi}_{(i)}\widehat{\xi}'_{(i)}\right), \tag{1.14}$$

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

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

Using (1.12), if $\widetilde{X}_{(i)} = \widetilde{D}_{(i)}X$ for any $i < k$ we can write (1.14) as

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

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

$$\widehat{F}_{(k)} = \widehat{\varsigma}'_{(k)}\widetilde{X}_{(k)} = \widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)}X, \tag{1.15}$$

for all k, and by definition, we have $\widehat{\zeta}_{(k)} = (\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)})'$. Moreover, using (1.13) the estimated coefficient $\widehat{\gamma}$ can be written as

$$\widehat{\gamma} = \sum_{k=1}^{\widehat{K}} \widehat{\alpha}_{(k)}\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)} = \sum_{k=1}^{\widehat{K}} T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2} Y \widehat{\xi}_{(k)}\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)}. \tag{1.16}$$

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

$$\widetilde{X}_{(k)} = \widetilde{\beta}_{(k)}F + \widetilde{U}_{(k)}. \tag{1.17}$$

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

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

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

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

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

Similar representations to (1.17) can be constructed for $Y_{(k)} := Y\prod_{i=1}^{k-1}\mathbb{M}_{\widehat{F}'_{(i)}}$ for each

k. Specifically, we have

$$Y_{(k)} = Y\left(\mathbb{I}_{T_h} - \sum_{i=1}^{k-1} \widehat{\xi}_{(i)}\widehat{\xi}_{(i)}'\right) = \widetilde{\alpha}_{(k)}F + \widetilde{Z}_{(k)}, \tag{1.18}$$

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

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

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

To prove Theorem 1, using (1.17), the estimated factors can be written as

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

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

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

□

1.6.2 Proof of Theorem 2

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

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

Therefore, using (1.18), we have

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

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

$$\left\|T_h^{-1}X_{[i]}Y'_{(\tilde{K}+1)}\right\|_{\text{F}}^2 \lesssim \left\|T_h^{-1}X_{[i]}Y'_{(\tilde{K}+1)}\right\|_{\text{MAX}}^2 \lesssim_{\text{P}} 1, \tag{1.19}$$

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

$$\begin{aligned}\sum_{i\in I_0}\left\|T_h^{-1}X_{[i]}Y'_{(\tilde{K}+1)}\right\|_{\text{F}}^2 &= \sum_{i\in I_0\cap S}\left\|T_h^{-1}X_{[i]}Y'_{(\tilde{K}+1)}\right\|_{\text{F}}^2 + \sum_{i\in I_0\cap S^c}\left\|T_h^{-1}X_{[i]}Y'_{(\tilde{K}+1)}\right\|_{\text{F}}^2 \\ &\lesssim_{\text{P}} |I_0\cap S| + |I_0\cap S^c|c^2 \leq qN + c^2N_0 = o(N_0),\end{aligned} \tag{1.20}$$

where we use (1.19), $|S| \leq qN-1$, $c \to 0$, and $qN/N_0 \to 0$. Consequently, (1.20) leads to $\left\|Y_{(\tilde{K}+1)}X'_{[I_0]}\right\| = o_{\text{P}}(TN_0^{1/2})$. Moreover, using (1.18) and that $X = \beta F + U$, we can decompose

$$Y_{(\tilde{K}+1)}X'_{[I_0]} = \widetilde{\alpha}_{(\tilde{K}+1)}FF'\beta'_{[I_0]} + \widetilde{\alpha}_{(\tilde{K}+1)}FU'_{[I_0]} + \widetilde{Z}_{(\tilde{K}+1)}F'\beta'_{[I_0]} + \widetilde{Z}_{(\tilde{K}+1)}U'_{[I_0]}. \tag{1.21}$$

Using (1.20), (1.21), Lemma 9(i)(ii), and the fact that $\left\|\beta_{[I_0]}\right\| \lesssim N_0^{1/2}$, we have

$$\left\|\widetilde{\alpha}_{(\tilde{K}+1)}\left(FF'\beta'_{[I_0]} + FU'_{[I_0]}\right)\right\| = o_{\text{P}}\left(N_0^{1/2}T\right). \tag{1.22}$$

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

$$|\sigma_K(FF'\beta'_{[I_0]} + FU'_{[I_0]}) - \sigma_K(T_h\beta_{[I_0]})| \leq \left\|FU'_{[I_0]}\right\| + \left\|T_h^{-1}FF' - \mathbb{I}_K\right\| \left\|T_h\beta_{[I_0]}\right\|$$
$$\lesssim_{\mathrm{P}} N_0^{1/2}T^{1/2}. \tag{1.23}$$

Since Assumption 2 implies that $\sigma_K(\beta_{[I_0]}) \asymp N_0^{1/2}$, we have $\sigma_K(FF'\beta'_{[I_0]}+FU'_{[I_0]}) \asymp N_0^{1/2}T$. Using this result, (1.22) and the inequality $\left\|\widetilde{\alpha}_{(\tilde{K}+1)}\left(FF'\beta'_{[I_0]} + FU'_{[I_0]}\right)\right\| \geq \sigma_K(FF'\beta_{[I_0]} + FU'_{[I_0]})\left\|\widetilde{\alpha}_{(\tilde{K}+1)}\right\|$, we have $\left\|\widetilde{\alpha}_{(\tilde{K}+1)}\right\| \xrightarrow{\mathrm{P}} 0$. That is, by definition of $\tilde{\alpha}_{(\tilde{K}+1)}$ in (1.18),

$$\left\|\alpha - \sum_{i=1}^{\tilde{K}} Y\widehat{\xi}_{(i)}\frac{\widehat{\varsigma}'_{(i)}\widetilde{\beta}_{(i)}}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\right\| = o_{\mathrm{P}}(1). \tag{1.24}$$

Next, (1.16) and $\widetilde{\beta}_{(k)} = \widetilde{D}_{(k)}\beta$ imply that

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

Therefore, (1.24) is equivalent to $\|\widehat{\gamma}\beta - \alpha\| = o_{\mathrm{P}}(1)$.

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

$$\widehat{y}_{T+h} - \mathrm{E}_T(y_{T+h}) = \widehat{\gamma}x_T + (\widehat{\alpha}_w - \widehat{\gamma}\widehat{\beta}_w)w_T - \alpha f_T - \alpha_w w_T \tag{1.25}$$
$$=(\widehat{\gamma}\beta - \alpha)\left(f_T - FW'(WW')^{-1}w_T\right) + \widehat{\gamma}(u_T - UW'(WW')^{-1}w_T) + ZW'(WW')^{-1}w_T.$$

Using (1.16) and $\|Y\| \leq \|\alpha F\| + \|Z\| \lesssim_{\mathrm{P}} T^{1/2}$ by Assumption 1, we have

$$\|\widehat{\gamma} u_T\| \leq \sum_{k \leq \tilde{K}} T_h^{-1/2} \widehat{\lambda}_{(k)}^{-1/2} \|Y\| \left\| \widehat{\xi}_{(k)} \right\| \left\| \widehat{\varsigma}_{(k)}' \widetilde{D}_{(k)} u_T \right\| \lesssim_{\mathrm{P}} \sum_{k \leq \tilde{K}} \widehat{\lambda}_{(k)}^{-1/2} \left\| \widehat{\varsigma}_{(k)}' \widetilde{D}_{(k)} u_T \right\|, \quad (1.26)$$

and

$$\begin{aligned} T_h^{-1} \left\| \widehat{\gamma} U W' \right\| &\leq \sum_{k \leq \tilde{K}} T_h^{-3/2} \widehat{\lambda}_{(k)}^{-1/2} \|Y\| \left\| \widehat{\xi}_{(k)} \right\| \left\| \widehat{\varsigma}_{(k)}' \widetilde{D}_{(k)} U W' \right\| \\ &\lesssim_{\mathrm{P}} \sum_{k \leq \tilde{K}} T_h^{-1} \widehat{\lambda}_{(k)}^{-1/2} \left\| \widehat{\varsigma}_{(k)}' \widetilde{U}_{(k)} W' \right\|. \end{aligned} \quad (1.27)$$

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

$$T_h^{-1} \widehat{\lambda}_{(k)}^{-1/2} \left\| \widehat{\varsigma}_{(k)} \widetilde{U}_{(k)} W' \right\| \lesssim_{\mathrm{P}} q^{-1} N^{-1} + T^{-1}, \widehat{\lambda}_{(k)}^{-1/2} \left\| \widehat{\varsigma}_{(k)}' \widetilde{D}_{(k)} u_T \right\| \lesssim_{\mathrm{P}} q^{-1/2} N^{-1/2} + T^{-1/2}. \quad (1.28)$$

Therefore, $\|\widehat{\gamma} u_T\| = o_{\mathrm{P}}(1)$. Furthermore, with $\left\|(WW')^{-1}\right\| \lesssim_{\mathrm{P}} T^{-1}$ from Assumption 1, we have $\left\|\widehat{\gamma} U W' (WW')^{-1}\right\| = o_{\mathrm{P}}(1)$. Together with $\left\|FW'\right\| \lesssim_{\mathrm{P}} T^{1/2}$, $\left\|ZW'\right\| \lesssim_{\mathrm{P}} T^{1/2}$ from Assumption 1 and $\|\widehat{\gamma}\beta - \alpha\| = o_{\mathrm{P}}(1)$, we show that each term of (1.25) vanishes, and hence $\widehat{y}_{T+h} - \mathrm{E}_T[y_{T+h}] \xrightarrow{\mathrm{P}} 0$. □

1.6.3 Proof of Theorem 3

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

Again, following the same argument above (1.25), we only need analyze the case without w_t. As $\widehat{F}_{(k)} = T_h^{1/2} \widehat{\lambda}_{(k)}^{1/2} \widehat{\xi}_{(k)}'$, Theorem 1 implies $\left\| \widehat{\xi}_{(k)}' \mathbb{M}_{F'} \right\| \lesssim_{\mathrm{P}} q^{-1/2} N^{-1/2} + T^{-1}$ for

$k \leq K$. Let v denote $F'(FF')^{-1/2}$, we have

$$\left\|\widehat{\xi} - \mathbb{P}_{F'}\widehat{\xi}\right\| = \left\|\widehat{\xi} - vv'\widehat{\xi}\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1}, \tag{1.29}$$

where $\widehat{\xi}$ is a $T \times K$ matrix with each column equal to $\widehat{\xi}_{(k)}$. (1.29) implies that

$$\left\|\widehat{\xi}'vv'\widehat{\xi} - \mathbb{I}_K\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1}.$$

By Weyl's inequality, $|\sigma_i(\widehat{\xi}'v) - 1| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1}$, for $1 \leq i \leq K$, and thus

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

Then, using this, (1.29), and the fact that $\|v\| = 1$ and $\left\|\widehat{\xi}\right\| = 1$, we have

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

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

$$\widehat{\gamma}\beta = \sum_{k=1}^{\tilde{K}} T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}Y\widehat{\xi}_{(k)}\widetilde{\varsigma}_{(k)}'\widetilde{\beta}_{(k)}. \tag{1.30}$$

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

$$b_{k1} = T^{-1/2}F\widehat{\xi}_{(k)}, \quad b_{k2} = \widehat{\lambda}_{(k)}^{-1/2}\widetilde{\beta}_{(k)}'\widehat{\varsigma}_{(k)}. \tag{1.31}$$

By Lemma 6,

$$\left\|T_h^{-1/2} Z\widehat{\xi}_{(k)} - T_h^{-1} ZF' b_{k2}\right\| \lesssim_{\mathrm{P}} T^{-1} + q^{-1}N^{-1}. \tag{1.32}$$

As we impose that $\widehat{K} = \tilde{K} = K$, combining (1.30), (1.31) and (1.32), with $\|B_1\| \lesssim_{\mathrm{P}} 1$, $\|B_2\| \lesssim_{\mathrm{P}} 1$ from Lemma 10, we have

$$\left\|\widehat{\gamma}\beta - \alpha B_1 B_2' - T_h^{-1} ZF' B_2 B_2'\right\| \lesssim_{\mathrm{P}} T^{-1} + q^{-1}N^{-1}. \tag{1.33}$$

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

1.6.4 Proof of Theorem 4

Proof. As in the proof of Theorem 2, we have $\left\|FW'(WW')^{-1}\right\| \lesssim_{\mathrm{P}} T^{-1/2}$ from Assumption 1 and $\left\|\widehat{\gamma}UW'(WW')^{-1}\right\| \lesssim_{\mathrm{P}} T^{-1} + q^{-1}N^{-1}$ as shown in (1.27) and (1.28). 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 (1.25) that:

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

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

$$\begin{aligned}
&\left\|ZW'(WW')^{-1} w_T - T_h^{-1} ZW'\Sigma_w^{-1} w_T\right\| \leq T_h^{-1} \left\|ZW'\right\| \left\|(T_h^{-1}WW')^{-1} - \Sigma_w^{-1}\right\| \|w_T\| \\
\lesssim_{\mathrm{P}} & T_h^{-1/2} \left\|T_h^{-1}\Sigma_w^{-1/2} WW' \Sigma_w^{-1/2} - \mathbb{I}_D\right\| = T_h^{-1/2} \max_{i \leq D} |\lambda_i\left(T_h^{-1}\Sigma_w^{-1/2} WW' \Sigma_w^{-1/2}\right)^{-1} - 1| \\
\lesssim_{\mathrm{P}} & T^{-1}.
\end{aligned} \tag{1.34}$$

For $\widehat{\gamma} u_T$, by (1.16), we have $\widehat{\gamma} u_T = \sum_{k=1}^K \widehat{\alpha}_{(k)} \widehat{\varsigma}'_{(k)} \widetilde{D}_{(k)} u_T$ and thus

$$\left\| \widehat{\gamma} u_T - \sum_{k=1}^K \lambda_{(k)}^{-1/2} \alpha b_{(k)} \varsigma'_{(k)} D_{(k)} u_T \right\| \leq \sum_{k=1}^K \left\| \widehat{\alpha}_{(k)} \widehat{\varsigma}'_{(k)} \widetilde{D}_{(k)} u_T - \lambda_{(k)}^{-1/2} \alpha b_{(k)} \varsigma'_{(k)} D_{(k)} u_T \right\|. \tag{1.35}$$

Lemma 8(vi) gives

$$q^{-1/2} N^{-1/2} |\widehat{\varsigma}'_{(k)} \widetilde{D}_{(k)} u_T - \varsigma'_{(k)} D_{(k)} u_T| \lesssim_{\mathrm{P}} T^{-1} + q^{-1} N^{-1}. \tag{1.36}$$

In addition, (1.13) and Lemma 1 give $\widehat{\lambda}_{(k)}^{1/2} \widehat{\alpha}_{(k)} = T_h^{-1/2} Y \widehat{\xi}_{(k)} = \alpha b_{k1} + T_h^{-1/2} Z \widehat{\xi}_{(k)}$. With (1.32), $\left\| Z F' \right\| \lesssim_{\mathrm{P}} T^{1/2}$ and $\| b_{k2} \| \lesssim_{\mathrm{P}} 1$ from Lemma 10(i), this equation leads to

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

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

$$\begin{aligned} \left\| \widehat{\alpha}_{(k)} - \widehat{\lambda}_{(k)}^{-1/2} \alpha b_{(k)} \right\| &\leq \left\| \widehat{\alpha}_{(k)} - \widehat{\lambda}_{(k)}^{-1/2} \alpha b_{k2} \right\| + \left\| \widehat{\lambda}_{(k)}^{-1/2} \alpha (b_{(k)} - b_{k2}) \right\| \\ &\lesssim_{\mathrm{P}} T^{-1/2} q^{-1/2} N^{-1/2} + q^{-1} N^{-1}. \end{aligned} \tag{1.37}$$

Also, with Lemma 3(iii), we have

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

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

$$\left\| \widehat{\alpha}_{(k)} - \lambda_{(k)}^{-1/2} \alpha b_{(k)} \right\| \leq T^{-1/2} q^{-1/2} N^{-1/2} + q^{-1} N^{-1}. \tag{1.38}$$

For each term in the summation of (1.35), we have

$$\begin{aligned}&\left\|\widehat{\alpha}_{(k)}\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)}u_T - \lambda_{(k)}^{-1/2}\alpha b_{(k)}\varsigma'_{(k)}D_{(k)}u_T\right\| \\ \leq &\left\|\widehat{\alpha}_{(k)}(\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)}u_T - \varsigma'_{(k)}D_{(k)}u_T)\right\| + \left\|(\widehat{\alpha}_{(k)} - \lambda_{(k)}^{-1/2}\alpha b_{(k)})\varsigma'_{(k)}D_{(k)}u_T\right\|. \end{aligned} \tag{1.39}$$

Note that (1.37) also implies $\left\|\widehat{\alpha}_{(k)}\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2}$ as $\widehat{\lambda}_{(k)} \asymp qN$, and that (1.36) implies the first term in (1.39) is $O_{\mathrm{P}}(T^{-1}+q^{-1}N^{-1})$. Furthermore, $|\varsigma'_{(k)}D_{(k)}u_T| \lesssim_{\mathrm{P}} 1$ from Lemma 5(iv) and (1.38) show that the second term in (1.39) is also $O_{\mathrm{P}}(T^{-1}+q^{-1}N^{-1})$. Given this, (1.35) becomes

$$\left\|\widehat{\gamma}u_T - \sum_{k=1}^{K}\lambda_{(k)}^{-1/2}\alpha b_{(k)}\varsigma'_{(k)}D_{(k)}u_T\right\| \lesssim_{\mathrm{P}} T^{-1}+q^{-1}N^{-1}. \tag{1.40}$$

To sum up, we have established that

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

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

1.6.5 *Proof of Theorem 5*

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

$$\max_{i \le N} T_h^{-1} \sum_t |u_{it} - \widehat{u}_{it}|^2 = O_{\text{P}}\left(\frac{1}{qN} + \frac{\log N}{T}\right).$$

These two estimates have been shown by Lemma 11(iii)(iv). If $m_{q,N}\left(\frac{1}{\sqrt{qN}} + \sqrt{\frac{\log N}{T}}\right)^{1-q} = o(1)$, then $\left\|\widehat{\Sigma}_u - \Sigma_u\right\| = o_{\text{P}}(1)$. With $\left\|\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)} - \varsigma'_{(k)}D_{(k)}\right\| \lesssim_{\text{P}} T^{-1/2} + q^{-1/2}N^{-1/2}$ from Lemma 8(iv) and $\widehat{\gamma} = \sum_{k \le K} \widehat{\alpha}_{(k)}\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)}$, rewrite the proof of (1.40), we have

$$\left\|\widehat{\gamma} - \sum_{k \le K} \lambda_{(k)}^{-1/2} \alpha b_{(k)} \varsigma'_{(k)} D_{(k)}\right\| \lesssim_{\text{P}} T^{-1/2} q^{-1/2} N^{-1/2} + q^{-1} N^{-1}. \tag{1.41}$$

The difference between the rate of (1.40) and this equation arises from the difference between Lemma 8(iv) and (vi). 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 (1.41) is equivalent to $\left\|\widehat{\gamma} - \alpha B \Lambda^{-1}\Omega'\Psi\right\|$. In addition, under the assumption $\text{Cov}(u_t) = \Sigma_u$, Π_{33} equals to $(qN)^{-1}\Psi\Sigma_u\Psi'$. Let $\tilde{\gamma}$ denote $\alpha B\Lambda^{-1}\Omega'\Psi$, then we have

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

Consequently, we have

$$\left\|\widehat{\Phi}_2 - \Phi_2\right\| \le qN\left\|\widehat{\gamma}(\widehat{\Sigma}_u - \Sigma_u)\widehat{\gamma}'\right\| + \left\|(\widehat{\gamma} - \tilde{\gamma})\Sigma_u\widehat{\gamma}'\right\| + \left\|\tilde{\gamma}\Sigma_u(\widehat{\gamma} - \tilde{\gamma})'\right\|. \tag{1.42}$$

Using the definition of $D_{(k)}$, $\left\|\beta_{[I_k]}\right\| \lesssim (qN)^{1/2}$, and $\lambda_{(k)} \asymp qN$, we have $\left\|D_{(k)}\right\| \lesssim 1$ and thus $\|\widehat{\gamma}\| \lesssim q^{-1/2}N^{-1/2}$. Using $\left\|\widehat{\Sigma}_u - \Sigma_u\right\| = o_{\text{P}}(1)$, (1.41), $\|\Sigma_u\| \lesssim 1$ from the assumption and $\|\widehat{\gamma}\| \lesssim q^{-1/2}N^{-1/2}$, all three terms in (1.42) are $o_{\text{P}}(1)$. □

1.6.6 Proofs from Section 1.3.5

1.6.6.1 Proof of Propositions 1 and 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:

$$X = \beta F + U = \beta F + \epsilon A_1 = \begin{pmatrix} \sqrt{\lambda}F + \epsilon_1 A_1 \\ \epsilon_2 A_1 \end{pmatrix}, \tag{1.43}$$

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

$$\widehat{\varsigma}_1 = \frac{(\sqrt{\lambda}F + \epsilon_1 A_1)\widehat{\xi}}{\sqrt{T\widehat{\lambda}}}, \quad \widehat{\varsigma}_2 = \frac{\epsilon_2 A_1 \widehat{\xi}}{\sqrt{T\widehat{\lambda}}}. \tag{1.44}$$

Since the entries of F are i.i.d. $\mathcal{N}(0,1)$, we have large deviation inequality $|T_h^{-1}FF' - 1| \lesssim_{\mathrm{P}} T^{-1/2}$. 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 $\left\|N^{-1}U'U - A_1'A_1\right\| \leq \|A_1\|^2 \left\|N^{-1}\epsilon'\epsilon - \mathbb{I}_{T_h}\right\| \lesssim_{\mathrm{P}} \sqrt{T/N}$. Next, by direct calculation using the previous inequalities we obtain

$$\left\| \frac{F'\epsilon_1 A_1 + A_1'\epsilon_1' F}{T_h\sqrt{\lambda}} + \frac{U'U - NA_1'A_1}{T_h\lambda} \right\| \lesssim_{\mathrm{P}} \frac{1}{\sqrt{\lambda}} + \frac{\sqrt{NT}}{T\lambda} \lesssim_{\mathrm{P}} \frac{1}{\sqrt{\lambda}}.$$

Together with (1.43), we have

$$\left\| \frac{X'X}{T_h\lambda} - \frac{F'F}{T_h} - \frac{NA_1'A_1}{T_h\lambda} \right\| \lesssim_{\mathrm{P}} \frac{1}{\sqrt{\lambda}}. \tag{1.45}$$

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 (1.45), 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

$$\lambda_i = \begin{cases} T_h^{-1}FF' + \delta & i = 1; \\ \delta & i \geq 2. \end{cases} \tag{1.46}$$

and the first eigenvector is $F'/\left\|F\right\|$. Since the largest eigenvalue of $X'X/(T_h\lambda)$ is $\widehat{\lambda}/\lambda$ with its corresponding eigenvector $\widehat{\xi}$, (1.45) and Weyl's theorem yield that

$$\frac{\widehat{\lambda}}{\lambda} = \frac{FF'}{T_h} + \frac{N}{T_h\lambda} + O_{\mathrm{P}}\left(\frac{1}{\sqrt{\lambda}}\right) = 1 + \delta + o_{\mathrm{P}}(1), \tag{1.47}$$

and the sin-theta theorem implies that

$$\left\| \mathbb{P}_{F'} - \mathbb{P}_{\widehat{\xi}} \right\| = \left\| F'(FF')^{-1}F - \widehat{\xi}\widehat{\xi}' \right\| = o_{\mathrm{P}}(1). \tag{1.48}$$

Furthermore, (1.48) 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

$$\frac{F\widehat{\xi}}{\sqrt{T_h}} - 1 = o_{\mathrm{P}}(1). \tag{1.49}$$

Therefore, we have

$$\widehat{y}_{t+h} = \widehat{\alpha}\widehat{\varsigma}' x_T = \frac{Y\widehat{\xi}\widehat{\varsigma}' x_T}{\sqrt{T_h\widehat{\lambda}}} = \alpha\frac{F\widehat{\xi}\widehat{\varsigma}' x_T}{\sqrt{T_h\widehat{\lambda}}} = \alpha\frac{\widehat{\varsigma}'\beta f_T + \widehat{\varsigma}' u_T}{\sqrt{\widehat{\lambda}}}(1 + o_{\mathrm{P}}(1)). \tag{1.50}$$

Using (1.44), 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 (1.47), (1.49), $\|A_1\| \leq 1$, and $\|\epsilon_1\| \lesssim_{\mathrm{P}} \sqrt{T}$, it follows that

$$\frac{\widehat{\varsigma}'\beta}{\sqrt{\widehat{\lambda}}} \xrightarrow{\mathrm{P}} \frac{1}{1+\delta}. \tag{1.51}$$

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 (1.50) and (1.51), we have $\widehat{y}_{T+h} \xrightarrow{\mathrm{P}} \frac{\alpha f_T}{1+\delta} = (1+\delta)^{-1}\mathrm{E}_T(y_{T+h})$. □

1.6.6.2 Proof of Propositions 3 and 4

Proof. In the case $d = K = 1$ and $z_t = 0$, the PLS estimate of the factor is $\widehat{F} = FX'X$. With (1.45) and $T_h^{-1}FF' - 1 = o_{\mathrm{P}}(T^{-1/2})$, we have

$$\left\| T_h^{-1}\lambda^{-1}\widehat{F} - F\left(\mathbb{I}_{T_h} + \delta A_1' A_1\right)\right\| = o_{\mathrm{P}}\left(T^{1/2}\right). \tag{1.52}$$

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 (1.52), 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 3. In the specical case $A_1'A_1 = \mathbb{I}_{T_h}$, as in Section 1.3.5.2, we can write

$$\widehat{y}_{T+h} = \left\|YX'X\right\|^{-2} YX'XY'YX'x_T = \alpha\left\|FX'X\right\|^{-2} FX'XF'FX'x_T. \tag{1.53}$$

We now analyze $\left\|FX'X\right\|$, $FX'XF'$, and $FX'x_T$, respectively. Recall that from (1.45), 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

$$\frac{1}{T_h^{3/2}\lambda}\left\|FX'X\right\| = \frac{1}{\sqrt{T_h}}\left\|F\left(\frac{F'F}{T_h} + \delta\mathbb{I}_{T_h}\right)\right\| + o_{\mathrm{P}}(1) = 1 + \delta + o_{\mathrm{P}}(1). \tag{1.54}$$

For the same reason, by direct calculation we have

$$\frac{1}{T_h^2\lambda}FX'XF' = \frac{1}{T_h}F\left(\frac{F'F}{T_h} + \delta\mathbb{I}_{T_h}\right)F' + o_{\mathrm{P}}(1) \xrightarrow{p} 1 + \delta. \tag{1.55}$$

Next, write X in the form of (1.43) as in the proof of Proposition 1. Then, using $\|\epsilon_1\| \lesssim_{\mathrm{P}} \sqrt{T}$, we have

$$\frac{1}{T_h\lambda}FX'\beta = \frac{FF'}{T_h} + \frac{FA_1'\epsilon_1'}{T\sqrt{\lambda}} \xrightarrow{\mathrm{P}} 1. \tag{1.56}$$

In addition, as u_T is independent of f_t and x_t for $t < T$, and (1.55), we have

$$\frac{1}{T_h\lambda}\left\|FX'u_T\right\| \lesssim_{\mathrm{P}} \frac{1}{T_h\lambda}\left\|FX'\right\| \xrightarrow{\mathrm{P}} 0. \tag{1.57}$$

In light of (1.54), (1.55), (1.56), (1.57) and (1.53), we have concluded the proof. □

1.6.6.3 Proof of Proposition 5

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

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

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

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

As z_i and u_i are generated from independent standard normal distribution, we have

$$\left\|ZU'\right\| \asymp_{\mathrm{P}} T^{1/2}N^{1/2} \text{ and } \|U\| \asymp_{\mathrm{P}} N^{1/2} + T^{1/2}.$$

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 $\left\|\widehat{\varsigma}'u_T\right\| \lesssim_{\mathrm{P}} 1$. Along with the fact that $\|U\| \gtrsim_{\mathrm{P}} N^{1/2} + T^{1/2}$, we have $\widehat{y}_{T+h}^{PCA} \lesssim_{\mathrm{P}} 1/(N^{1/2}+T^{1/2})$. □

1.6.6.4 Proof of Proposition 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 $\left\|N^{-1}U'U - \mathrm{diag}(1, \ldots, 1, 1+\epsilon)\right\| \lesssim_{\mathrm{P}} \sqrt{T/N}$. 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 $\text{Var}(\widehat{f}_{T+h}^{SW}) \to \infty$. 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 $\left\|\overline{Z}\widehat{\xi}\right\| \lesssim_{\text{P}} 1$ and $\left\|\widehat{\varsigma}' u_T\right\| \lesssim_{\text{P}} 1$. Along with the fact that $\|\underline{U}\| \gtrsim_{\text{P}} N^{1/2} + T^{1/2}$, we have $\widehat{y}_{T+h}^{PCA} \xrightarrow{\text{P}} 0$. □

1.6.7 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 3, we assume that $\widehat{K} = \tilde{K}$ and $\widehat{I}_k = I_k$ for $k \leq \tilde{K}$, which hold with probability approaching one as we will show in Lemma 3.

Lemma 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 (1.14) 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

$$\left\|\widetilde{X}_{(k)}(\widehat{\xi}_{(k)} - c_0\widehat{\xi}_{(j)})\right\| = \left\|\widetilde{X}_{(k)}\widehat{\xi}_{(k)} - c_0\widetilde{X}_{(k)}\widehat{\xi}_{(j)}\right\| = \left\|\widetilde{X}_{(k)}\widehat{\xi}_{(k)}\right\|, \tag{1.58}$$

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,

$$\left\|\widehat{\xi}_{(k)}\right\|^2 = \left\|\widehat{\xi}_{(k)} - c_0\widehat{\xi}_{(j)}\right\|^2 + \left\|c_0\widehat{\xi}_{(j)}\right\|^2 > \left\|\widehat{\xi}_{(k)} - c_0\widehat{\xi}_{(j)}\right\|^2. \tag{1.59}$$

Apparently, if $\left\|\widetilde{X}_{(k)}\right\| = 0$, the SPCA procedure will terminate so we have $\left\|\widetilde{X}_{(k)}\right\| > 0$ for

$k \le \widehat{K}$. Together with (1.58) and (1.59), we have

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

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

Lemma 2. *Under assumptions of Theorem 1, $b_{(k)}$, $\beta_{(k)}$ and $\tilde{K}$ in Section 1.3.1 satisfy*

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

(ii) $\lambda_{(k)}^{1/2} = \left\|\beta_{(k)}\right\| \asymp q^{1/2}N^{1/2}$.

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

(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 1, we have $b'_{(j)}b_{(k)} = \delta_{jk}$ for $j, k \le \tilde{K}$.

(ii) The selection rule at kth step implies that

$$|I_k|^{-1}\sum_{i\in I_k}\left\|\beta_{[i]}\prod_{j<k}\mathbb{M}_{b_{(j)}}\alpha'\right\|_{\text{MAX}}^2 \ge N_0^{-1}\sum_{i\in I_0}\left\|\beta_{[i]}\prod_{j<k}\mathbb{M}_{b_{(j)}}\alpha'\right\|_{\text{MAX}}^2. \tag{1.60}$$

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

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

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

$$\left\|A_{[I]}\right\|^2 \asymp \sum_{i\in I}\left\|A_{[i]}\right\|_{\text{MAX}}^2. \tag{1.61}$$

Using this result, (1.60) 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

$$\begin{aligned}&\frac{\left\|\beta_{(k)}\right\|}{\sqrt{|I_k|}}\left\|\prod_{j<k}\mathbb{M}_{b_{(j)}}\alpha'\right\| \geq \frac{1}{\sqrt{|I_k|}}\left\|\beta_{[I_k]}\prod_{j<k}\mathbb{M}_{b_{(j)}}\alpha'\right\| \gtrsim \frac{1}{\sqrt{N_0}}\left\|\beta_{[I_0]}\prod_{j<k}\mathbb{M}_{b_{(j)}}\alpha'\right\| \\ \geq&\frac{\sigma_K\left(\beta_{[I_0]}\right)}{\sqrt{N_0}}\left\|\prod_{j<k}\mathbb{M}_{b_{(j)}}\alpha'\right\|, \end{aligned} \tag{1.62}$$

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 $\sigma_K(\beta_{[I_0]}) \gtrsim \sqrt{N_0}$ from Assumption 2, (1.62) leads to $\left\|\beta_{(k)}\right\| \gtrsim |I_k|^{1/2}$. In addition, $\|\beta\|_{\text{MAX}} \lesssim 1$ from Assumption 2 leads to $\left\|\beta_{(k)}\right\| \lesssim |I_k|^{1/2}$. Therefore, we have $\left\|\beta_{(k)}\right\| \asymp |I_k|^{1/2} \asymp q^{1/2}N^{1/2}$.

(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, $\tilde{K} \leq K$.

(iv) Recall that $\tilde{K}$ is defined in Section 1.3.1. 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

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

where we use (1.61) and the assumptions $c \to 0$, $qN/N_0 \to 0$, and a similar argument for the proof of (1.20). With $\sigma_K(\beta_{[I_0]}) \gtrsim \sqrt{N_0}$ from Assumption 2, we have

$$\left\|\alpha\prod_{j\leq\tilde{K}}\mathbb{M}_{b_{(j)}}\right\| \leq \sigma_K(\beta_{[I_0]})^{-1}\left\|\beta_{[I_0]}\prod_{j\leq\tilde{K}}\mathbb{M}_{b_{(j)}}\alpha'\right\| = o(1). \tag{1.63}$$

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

$$\sigma_K(\alpha) \leq \sigma_1 \left(\alpha \prod_{j \leq \tilde{K}} \mathbb{M}_{b_{(j)}} \right) + \sigma_K \left(\alpha \sum_{j \leq \tilde{K}} b_{(j)} b'_{(j)} \right). \tag{1.64}$$

Since

$$\text{Rank} \left(\alpha \sum_{j \leq \tilde{K}} b_{(j)} b'_{(j)} \right) \leq \tilde{K} \leq K-1, \tag{1.65}$$

we have $\sigma_K \left(\alpha \sum_{j \leq \tilde{K}} b_{(j)} b'_{(j)} \right) = 0$. Therefore, by (1.64) and (1.63), we further have $\sigma_K(\alpha) \lesssim \sigma_1 \left(\alpha \prod_{j \leq \tilde{K}} \mathbb{M}_{b_{(j)}} \right) \to 0$. This contradicts with the assumption that $\lambda_K(\alpha'\alpha) \gtrsim 1$. Therefore, we have established that $\tilde{K} \geq K$. Together with (iii), we have $\tilde{K} = K$. □

Lemma 3. *Under assumptions of Theorem 1, for $k \leq \tilde{K}$, 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) $|\widehat{\lambda}_{(k)}^{1/2} / \lambda_{(k)}^{1/2} - 1| \lesssim_{\mathrm{P}} q^{-1/2} N^{-1/2} + T^{-1/2}$, *and* $\widehat{\lambda}_{(k)} \asymp_{\mathrm{P}} \lambda_{(k)} \asymp qN$.

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

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

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

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

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

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

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

$$\begin{aligned}A_1 :&= \left\{ \left\|T_h^{-1}X_{[i]}Y'\right\|_{\text{MAX}} > (c_{qN}^{(1)} + c_{qN+1}^{(1)})/2 \text{ for all } i \in I_1\right\}, \\ A_2 :&= \left\{ \left\|T_h^{-1}X_{[i]}Y'\right\|_{\text{MAX}} < (c_{qN}^{(1)} + c_{qN+1}^{(1)})/2 \text{ for all } i \in I_1^c\right\}, \\ A_3 :&= \left\{ \left\|T_h^{-1}X_{[i]}Y' - \beta_{[i]}\alpha'\right\|_{\text{MAX}} \geq (c_{qN}^{(1)} - c_{qN+1}^{(1)})/2 \text{ for some } i \in [N]\right\}. \qquad (1.66)\end{aligned}$$

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 $\left\|\beta_{[i]}\alpha'\right\|_{\text{MAX}} \geq c_{qN}^{(1)}$ for all $i \in I_1$ and $\left\|\beta_{[i]}\alpha'\right\|_{\text{MAX}} \leq c_{qN+1}^{(1)}$ for all $i \in I_1^c$. Therefore, if A_1^c occurs, we have $\left\|T_h^{-1}X_{[i]}Y' - \beta_{[i]}\alpha'\right\|_{\text{MAX}} \geq (c_{qN}^{(1)} - c_{qN+1}^{(1)})/2$, 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

$$\text{P}(\widehat{I}_1 = I_1) \geq \text{P}(A_1 \cap A_2) = 1 - \text{P}(A_1^c \cup A_2^c) \geq 1 - \text{P}(A_3). \qquad (1.67)$$

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

(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 (1.14) 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 13, we have $\sigma_j(\beta_{(1)}F)/\sigma_j(\beta_{(1)}) = T_h^{1/2} + O_{\mathrm{P}}(1)$, which leads to

$$|\left\|\beta_{(1)}F\right\| - T_h^{1/2}\lambda_{(1)}^{1/2}| = |\sigma_1(\beta_{(1)}F) - T_h^{1/2}\sigma_1(\beta_{(1)})| \lesssim_{\mathrm{P}} q^{1/2}N^{1/2}, \tag{1.68}$$

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

$$|\left\|\widetilde{X}_{(1)}\right\| - \left\|\beta_{(1)}F\right\|| \leq \left\|\widetilde{X}_{(1)} - \beta_{(1)}F\right\| \lesssim_{\mathrm{P}} q^{1/2}N^{1/2} + T^{1/2}. \tag{1.69}$$

Using (1.68), (1.69) and $\lambda_{(1)}^{1/2} \asymp q^{1/2}N^{1/2}$, we have

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

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

(iv) Let $\tilde{\xi}_{(1)} \in \mathbb{R}^{T_h \times 1}$ denote the first right singular vector of $\beta_{(1)}F$. Lemma 13 yields

$$\left\|\mathbb{P}_{\tilde{\xi}_{(1)}} - T_h^{-1}F'\mathbb{P}_{b_{(1)}}F\right\| \lesssim_{\mathrm{P}} T^{-1/2} \tag{1.70}$$

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

$$\sigma_1(\beta_{(1)}F) - \sigma_2(\beta_{(1)}F) = T_h^{1/2}(\sigma_1(\beta_{(1)}) - \sigma_2(\beta_{(1)})) + O_{\mathrm{P}}(\sigma_1(\beta_{(1)})) \asymp_p T^{1/2}\sigma_1(\beta_{(1)}), \tag{1.71}$$

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), (1.71), Lemma 2 and Wedin [1972]'s sin-theta theorem for singular vectors, we have

$$\left\|\mathbb{P}_{\widehat{F}'_{(1)}} - \mathbb{P}_{\tilde{\xi}_{(1)}}\right\| \lesssim_{\mathrm{P}} \frac{q^{1/2}N^{1/2} + T^{1/2}}{\sigma_1(\beta_{(1)}F) - \sigma_2(\beta_{(1)}F)} \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1/2}. \tag{1.72}$$

In light of (1.70) and (1.72), we have 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

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

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

$$\begin{aligned}
&\left\|\beta\prod_{j=1}^{k-1}\mathbb{M}_{b_{(j)}}\alpha' - T_h^{-1}(\beta F + U)\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{F}'_{(j)}}(\alpha F + Z)'\right\|_{\mathrm{MAX}} \\
\leq &\left\|\beta\prod_{j=1}^{k-1}\mathbb{M}_{b_{(j)}}\alpha' - T_h^{-1}\beta F\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{F}'_{(j)}}F'\alpha'\right\|_{\mathrm{MAX}} + T_h^{-1}\left\|\beta F\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{F}'_{(j)}}Z'\right\|_{\mathrm{MAX}} \\
&+ T_h^{-1}\left\|U\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{F}'_{(j)}}F'\alpha'\right\|_{\mathrm{MAX}} + T_h^{-1}\left\|U\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{F}'_{(j)}}Z'\right\|_{\mathrm{MAX}}.
\end{aligned} \tag{1.73}$$

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

$$\left\| \sum_{j=1}^{k-1} \mathbb{P}_{\widehat{F}'_{(j)}} - T_h^{-1} F' \sum_{j=1}^{k-1} \mathbb{P}_{b_{(j)}} F \right\| = \left\| \sum_{j=1}^{k-1} \left(\mathbb{P}_{\widehat{F}'_{(j)}} - T_h^{-1} F' \mathbb{P}_{b_{(j)}} F \right) \right\| \lesssim_{\mathrm{P}} q^{-1/2} N^{-1/2} + T^{-1/2}. \tag{1.74}$$

Using Lemma 1 and Lemma 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, (1.74), and $\left\| T_h^{-1} F F' - \mathbb{I}_K \right\| \lesssim_{\mathrm{P}} T^{-1/2}$, we have

$$T_h^{-1/2} \left\| F \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(j)}} - \prod_{j=1}^{k-1} \mathbb{M}_{b_{(j)}} F \right\| = T_h^{-1/2} \left\| F \sum_{j=1}^{k-1} \mathbb{P}_{\widehat{F}'_{(j)}} - \sum_{j=1}^{k-1} \mathbb{P}_{b_{(j)}} F \right\| \lesssim_{\mathrm{P}} q^{-1/2} N^{-1/2} + T^{-1/2}. \tag{1.75}$$

Similarly, right multiplying F' to the term inside the $\|\cdot\|$ of (1.75), we have

$$\left\| T_h^{-1} F \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{F}'_{(j)}} F' - \prod_{j=1}^{k-1} \mathbb{M}_{b_{(j)}} \right\| \lesssim_{\mathrm{P}} q^{-1/2} N^{-1/2} + T^{-1/2}. \tag{1.76}$$

Next, we analyze the four terms in (1.73) one by one. For the first term, using (1.76) and Assumption 2, we have

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

For the second term, using (1.75), Assumption 2, we have

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

For the third term, using (1.75), we have

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

For the forth term, using (1.74), we have

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

Hence, we have

$$\left\|T_h^{-1}X\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{F}'_{(j)}}Y'-\beta\prod_{j=1}^{k-1}\mathbb{M}_{b_{(j)}}\alpha'\right\|_{\text{MAX}}\lesssim_{\text{P}}(\log NT)^{1/2}\left(q^{-1/2}N^{-1/2}+T^{-1/2}\right).\tag{1.77}$$

As in the case of $k=1$, with the assumption that $c^{-1}(\log NT)^{1/2}\left(q^{-1/2}N^{-1/2}+T^{-1/2}\right)\to 0$, and Assumption 5, we can reuse the arguments for (1.66) and (1.67) in the case of $k=1$

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

(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

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

Hence, using Assumption 2 and (1.75), we have

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

(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 (1.77), we have

$$\left\| T_h^{-1} X \prod_{j=1}^{\tilde{K}} \mathbb{M}_{\widehat{F}'_{(j)}} Y' - \beta \prod_{j=1}^{\tilde{K}} \mathbb{M}_{b_{(j)}} \alpha' \right\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} (\log NT)^{1/2} \left(q^{-1/2} N^{-1/2} + T^{-1/2} \right). \tag{1.78}$$

The assumption $c_{qN}^{(\tilde{K}+1)} \leq (1+\delta)^{-1} c$ in Assumption 5 implies that $c - c_{qN}^{(\tilde{K}+1)} \asymp c$. Together with $c^{-1} (\log NT)^{1/2} \left(q^{-1/2} N^{-1/2} + T^{-1/2} \right) \to 0$, we can reuse the arguments for (1.66)

and (1.67) with events

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

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

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 (1.77) and (1.78). □

Lemma 4. *Under assumptions of Theorem 1, for $k \leq \tilde{K}$, we have*

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

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

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

Proof. (i) Using Lemma 1, we have

$$T_h^{1/2} \widehat{\lambda}_{(k)}^{1/2} \widehat{\varsigma}_{(k)} = X_{[I_k]} \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{\xi}_{(j)}} \widehat{\xi}_{(k)} = X_{[I_k]} \widehat{\xi}_{(k)} = \beta_{[I_k]} F \widehat{\xi}_{(k)} + U_{[I_k]} \widehat{\xi}_{(k)}. \tag{1.79}$$

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

$$\begin{aligned} T_h^{1/2} \widehat{\lambda}_{(k)}^{1/2} \left\| \widehat{\varsigma}'_{(k)} U_{[I_k]} \right\| &\leq \left\| \widehat{\xi}'_{(k)} F' \beta'_{[I_k]} U_{[I_k]} \right\| + \left\| \widehat{\xi}'_{(k)} U'_{[I_k]} U_{[I_k]} \right\| \\ &\leq \|F\| \left\| \beta'_{[I_k]} U_{[I_k]} \right\| + \left\| U'_{[I_k]} U_{[I_k]} \right\| \lesssim_{\mathrm{P}} q^{1/2} N^{1/2} T + qN. \end{aligned} \tag{1.80}$$

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

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

$$T^{1/2}\widehat{\lambda}_{(k)}^{1/2}\left\|\widehat{\varsigma}'_{(k)}U_{[I_k]}F'\right\| \leq \left\|\widehat{\xi}'_{(k)}F'\beta'_{[I_k]}U_{[I_k]}F'\right\| + \left\|\widehat{\xi}'_{(k)}U'_{[I_k]}U_{[I_k]}F'\right\|$$
$$\leq \|F\|\left\|\beta'_{[I_k]}U_{[I_k]}F'\right\| + \left\|U_{[I_k]}\right\|\left\|U_{[I_k]}F'\right\| \lesssim_{\mathrm{P}} q^{1/2}N^{1/2}T + qNT^{1/2}. \tag{1.81}$$

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

(iii) By Assumption 1, Assumption 3 and Assumption 4(iii), we have

$$T_h^{1/2}\widehat{\lambda}_{(k)}^{1/2}|\widehat{\varsigma}'_{(k)}(u_T)_{[I_k]}| \leq |\widehat{\xi}'_{(k)}F'\beta'_{[I_k]}(u_T)_{[I_k]}| + |\widehat{\xi}'_{(k)}U'_{[I_k]}(u_T)_{[I_k]}|$$
$$\leq \|F\|\left\|\beta'_{[I_k]}(u_T)_{[I_k]}\right\| + \left\|U_{[I_k]}\right\|\left\|(u_T)_{[I_k]}\right\| \lesssim_{\mathrm{P}} q^{1/2}N^{1/2}T^{1/2} + qN.$$

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

Lemma 5. *Under assumptions of Theorem 1, for $k, l \leq \tilde{K}$, we have*

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

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

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

(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 (1.16)), we have

$$\widetilde{U}_{(k)} = U_{[I_k]} - \sum_{i=1}^{k-1}\frac{X_{[I_k]}\widehat{\xi}_{(i)}}{\sqrt{T_h}}\frac{\widehat{\varsigma}'_{(i)}\widetilde{U}_{(i)}}{\sqrt{\widehat{\lambda}_{(i)}}}. \tag{1.82}$$

Then, a direct multiplication of $\widehat{\varsigma}'_{(k)}/\sqrt{T_h\widehat{\lambda}_{(k)}}$ from the left side of (1.82) 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 $\left\|X_{[I_k]}\right\| \leq \left\|\beta_{[I_k]}\right\|\|F\| + \left\|U_{[I_k]}\right\| \lesssim_{\mathrm{P}} q^{1/2}N^{1/2}T^{1/2}$, $\widehat{\lambda}_{(k)} \asymp_{\mathrm{P}} qN$ and Lemma 4(i) we have

$$\begin{aligned}\left\|\frac{\widehat{\varsigma}'_{(k)}\widetilde{U}_{(k)}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}\right\| &\leq \left\|\frac{\widehat{\varsigma}'_{(k)}U_{[I_k]}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}\right\| + \sum_{i=1}^{k-1}\left\|\frac{X_{[I_k]}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}\right\|\left\|\frac{\widehat{\varsigma}'_{(i)}\widetilde{U}_{(i)}}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\right\| \\ &\lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1} + \sum_{i=1}^{k-1}\left\|\frac{\widehat{\varsigma}'_{(i)}\widetilde{U}_{(i)}}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\right\|. \end{aligned} \tag{1.83}$$

If $\left\|T_h^{-1/2}\widehat{\lambda}_{(i)}^{-1/2}\widehat{\varsigma}'_{(i)}\widetilde{U}_{(i)}\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1}$ holds for $i \leq k-1$, then (1.83) 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 4(i). Therefore, we have (i) holds for $k \leq \tilde{K}$ by induction.

Using (1.82) again, with Assumption 3, we have

$$\begin{aligned}\left\|\frac{\widetilde{U}_{(k)}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}\right\| &\leq \left\|\frac{U_{[I_k]}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}\right\| + \sum_{i=1}^{k-1}\left\|\frac{X_{[I_k]}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}\right\|\left\|\frac{\widetilde{U}_{(i)}}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\right\| \\ &\lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1/2} + \sum_{i=1}^{k-1}\left\|\frac{\widetilde{U}_{(i)}}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\right\|. \end{aligned} \tag{1.84}$$

When $k = 1$, Assumption 3 implies $\left\|T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}\widetilde{U}_{(k)}\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1/2}$. Then, using the same induction argument with (1.84), we have this inequality holds for $k \leq \tilde{K}$.

(ii) Similarly, by simple multiplication of F' from the right side of (1.82), 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

$$\begin{aligned}\left\|\frac{\widehat{\varsigma}'_{(k)}\widetilde{U}_{(k)}F'}{T_h\sqrt{\widehat{\lambda}_{(k)}}}\right\| &\leq \left\|\frac{\widehat{\varsigma}'_{(k)}U_{[I_k]}F'}{T_h\sqrt{\widehat{\lambda}_{(k)}}}\right\| + \sum_{i=1}^{k-1}\left\|\frac{X_{[I_k]}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}\right\|\left\|\frac{\widehat{\varsigma}'_{(i)}\widetilde{U}_{(i)}F'}{T_h\sqrt{\widehat{\lambda}_{(i)}}}\right\| \\ &\lesssim_{\mathrm{P}} q^{-1}N^{-1} + T^{-1} + \sum_{i=1}^{k-1}\left\|\frac{\widehat{\varsigma}'_{(i)}\widetilde{U}_{(i)}F'}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\right\|. \end{aligned} \tag{1.85}$$

When $k=1$, $\left\|T_h^{-1}\widehat{\lambda}_{(k)}^{-1/2}\widehat{\varsigma}'_{(k)}\widetilde{U}_{(k)}F'\right\| \lesssim_{\mathrm{P}} q^{-1}N^{-1}+T^{-1}$ is a result of Lemma 4(ii). Then, a direct induction argument using (1.85) leads to this inequality for $k \leq \tilde{K}$.

Replacing F by Z in the above proof, and using Lemma 4(ii), we have:

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

(iii) Recall that $\widetilde{X}_{(k)} = \widetilde{\beta}_{(k)}F + \widetilde{U}_{(k)}$ as defined in (1.14), we have

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

Along with (1.12), we have

$$|\frac{\widehat{\xi}'_{(l)}U'_{[I_k]}\widehat{\varsigma}_{(k)}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}| \leq \left\|\frac{\widehat{\varsigma}'_{(l)}\widetilde{\beta}_{(l)}}{\sqrt{\widehat{\lambda}_{(l)}}}\right\|\left\|\frac{FU'_{[I_k]}\widehat{\varsigma}_{(k)}}{T_h\sqrt{\widehat{\lambda}_{(k)}}}\right\| + \left\|\frac{U'_{[I_k]}\widehat{\varsigma}_{(k)}}{\sqrt{T_h\widehat{\lambda}_{(k)}}}\right\|\left\|\frac{\widetilde{U}'_{(l)}\widehat{\varsigma}_{(l)}}{\sqrt{T_h\widehat{\lambda}_{(l)}}}\right\|. \tag{1.86}$$

Using $\left\|\widehat{\lambda}_{(k)}^{-1/2}\widehat{\varsigma}'_{(k)}\widetilde{\beta}_{(k)}\right\| \lesssim_{\mathrm{P}} 1$ from Lemma 10, results of (i)(ii) and Lemma 4(i) completes the

proof. Replacing $U_{[I_k]}$ by $\widetilde{U}_{(k)}$ above and using the inequality that

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

and (1.12), we obtain the second equation in (iii).

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

$$|\widehat{\varsigma}'_{(k)}\widetilde{D}_{(k)}u_T| \leq |\widehat{\varsigma}_{(k)}(u_T)_{[I_k]}| + \sum_{i=1}^{k-1}\left\|\frac{X_{[I_k]}}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\right\| |\widehat{\varsigma}'_{(i)}\widetilde{D}_{(i)}u_T| \lesssim_{\mathrm{P}} 1 + T^{-1/2}q^{1/2}N^{1/2}.$$

For the second inequality, from Lemma 2, we have $b'_{(i)}b_{(j)} = 0$ when $i \neq j$. Thus, by definition, we have $\varsigma_{(k)} = \lambda_{(k)}^{-1/2}\beta_{(k)}b_{(k)} = \lambda_{(k)}^{-1/2}\beta_{[I_k]}b_{(k)}$. Using $\left\|\beta'_{[I_k]}(u_T)_{[I_k]}\right\| \lesssim_{\mathrm{P}} q^{1/2}N^{1/2}$ from Assumption 4(iii), $\lambda_{(k)} \asymp qN$ from Lemma 2 and the definition of $D_{(k)}$, we have

$$\begin{aligned}
|\varsigma'_{(k)}D_{(k)}u_T| \leq{}& |\varsigma'_{(k)}(u_T)_{[I_k]}| + \sum_{i<k}\lambda_{(i)}^{-1/2}\left\|\beta_{[I_k]}\right\|\left\|b_{(i)}\right\| |\varsigma'_{(i)}D_{(i)}u_T| \\
\leq{}& \lambda_{(k)}^{-1/2}\left\|b_{(k)}\right\|\left\|\beta'_{[I_k]}(u_T)_{[I_k]}\right\| + \sum_{i<k}\lambda_{(i)}^{-1/2}\left\|\beta_{[I_k]}\right\| |\varsigma'_{(i)}D_{(i)}u_T| \lesssim_{\mathrm{P}} 1 + \sum_{i<k}|\varsigma'_{(i)}D_{(i)}u_T|.
\end{aligned}$$

As $|\varsigma'_{(1)}D_{(1)}u_T| \leq \lambda_{(1)}^{-1/2}\left\|b_{(1)}\right\|\left\|\beta_{[I_1]}(u_T)_{[I_1]}\right\| \lesssim_{\mathrm{P}} 1$, $|\varsigma'_{(k)}D_{(k)}u_T| \lesssim_{\mathrm{P}} 1$ holds by induction.

□

Lemma 6. *Under assumptions of Theorem 1, for $k \leq \tilde{K}$, we have*

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

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

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

$$\widehat{\xi}_{(k)} - T_h^{-1/2}F'b_{k2} = \frac{\widetilde{U}'_{(k)}\widehat{\varsigma}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}. \tag{1.87}$$

Then, Lemma 5(i) leads to (i) directly. (ii) Similarly, Lemma 5(ii) yields (ii). □

Lemma 7. *Under assumptions of Theorem 1, for $k, j \leq \tilde{K}$, we have*

(i) $\left\|\beta'_{[I_k]} U_{[I_k]} \widehat{\xi}_{(j)}\right\| \lesssim_{\mathrm{P}} q^{1/2} N^{1/2} + T^{1/2}$.

(ii) $\left\|\beta'_{[I_k]} \widetilde{U}_{(k)} \widehat{\xi}_{(j)}\right\| \lesssim_{\mathrm{P}} q^{1/2} N^{1/2} + T^{1/2}$.

(iii) $\left\|(u_T)'_{[I_k]} U_{[I_k]} \widehat{\xi}_{(j)}\right\| \lesssim_{\mathrm{P}} T^{-1/2} qN + T^{1/2}$.

(iv) $\left\|(u_T)'_{[I_k]} \widetilde{U}_{(k)} \widehat{\xi}_{(j)}\right\| \lesssim_{\mathrm{P}} T^{-1/2} qN + T^{1/2}$.

Proof. (i) With Lemma 6 and Assumption 4, we have $\left\|\beta'_{[I_k]} U_{[I_k]} \widehat{\xi}_{(j)}\right\| \leq T_h^{-1/2}$ $\left\|\beta'_{[I_k]} U_{[I_k]} F' b_{j2}\right\| + \left\|\beta'_{[I_k]} U_{[I_k]}\right\| \left\|T_h^{-1/2} F' b_{j2} - \widehat{\xi}_{(j)}\right\| \lesssim_{\mathrm{P}} q^{1/2} N^{1/2} + T^{1/2}$. (ii) Assumptions 1, 2 and 3 imply that $\left\|X_{[I_k]}\right\| \leq \left\|\beta_{[I_k]}\right\| \|F\| + \left\|U_{[I_k]}\right\| \lesssim_{\mathrm{P}} q^{1/2} N^{1/2} T^{1/2}$. Together with (i) and Lemma 5(iii), we have

$$\begin{aligned}\left\|\beta'_{[I_k]} \widetilde{U}_{(k)} \widehat{\xi}_{(j)}\right\| \leq & \left\|\beta'_{[I_k]} U_{[I_k]} \widehat{\xi}_{(j)}\right\| + \sum_{i=1}^{k-1} \left\|\beta_{[I_k]}\right\| \left\|\frac{X_{[I_k]} \widehat{\xi}_{(i)}}{\sqrt{T_h \widehat{\lambda}_{(i)}}}\right\| \left\|\widehat{\varsigma}'_{(i)} \widetilde{U}_{(i)} \widehat{\xi}_{(j)}\right\| \\ \lesssim_{\mathrm{P}} & q^{1/2} N^{1/2} + T^{1/2}.\end{aligned}$$

(iii) With Lemma 6 and Assumption 4, we have

$$\begin{aligned}\left\|(u_T)'_{[I_k]} U_{[I_k]} \widehat{\xi}_{(j)}\right\| \leq & T_h^{-1/2} \left\|(u_T)'_{[I_k]} U_{[I_k]} F' b_{j2}\right\| + \left\|(u_T)'_{[I_k]} U_{[I_k]}\right\| \left\|T_h^{-1/2} F' b_{j2} - \widehat{\xi}_{(j)}\right\| \\ \lesssim_{\mathrm{P}} & T^{-1/2} qN + T^{1/2}.\end{aligned}$$

(iv) Similar to (ii), with Lemma 5, $\left\|X_{[I_k]}\right\| \lesssim_{\mathrm{P}} q^{1/2} N^{1/2} T^{1/2}$, and (iii), we have $\left\|(u_T)'_{[I_k]} \widetilde{U}_{(k)} \widehat{\xi}_{(j)}\right\| \lesssim_{\mathrm{P}} T^{-1/2} qN + T^{1/2}$. □

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

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

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

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

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

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

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

Proof. We prove (i) - (vi) by induction. Consider the $k = 1$ case. The definitions of $\widehat{\varsigma}_{(k)}$ in (1.12) and $\varsigma_{(k)}$ in Section 1.3.1 lead to

$$\widehat{\varsigma}_{(k)} - \varsigma_{(k)} = T_h^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}(\widetilde{D}_{(k)}\beta F\widehat{\xi}_{(k)} + \widetilde{D}_{(k)}U\widehat{\xi}_{(k)}) - \lambda_{(k)}^{-1/2}D_{(k)}\beta b_{(k)}, \tag{1.88}$$

when $k = 1$, as $\widetilde{D}_{(1)} = D_{(1)} = (\mathbb{I}_N)_{[I_1]}$, (1.88) becomes

$$\widehat{\varsigma}_{(1)} - \varsigma_{(1)} = (T_h^{-1/2}\widehat{\lambda}_{(1)}^{-1/2}\beta_{(1)}F\widehat{\xi}_{(1)} - \lambda_{(1)}^{-1/2}\beta_{(1)}b_{(1)}) + T_h^{-1/2}\widehat{\lambda}_{(1)}^{-1/2}U_{[I_1]}\widehat{\xi}_{(1)}. \tag{1.89}$$

As Lemma 3(iii) and (iv) imply that

$$\left\|T_h^{-1/2}\widehat{\lambda}_{(1)}^{-1/2}F\widehat{\xi}_{(1)} - \lambda_{(1)}^{-1/2}b_{(1)}\right\| \lesssim_{\mathrm{P}} q^{-1}N^{-1} + T^{-1/2}q^{-1/2}N^{-1/2}$$

and $\left\|\beta_{(1)}\right\| \lesssim q^{1/2}N^{1/2}$, to prove (i) it is sufficient to show that $T_h^{-1/2}q^{-1/2}N^{-1/2}\left\|U_{[I_1]}\right\| \lesssim_{\mathrm{P}} T^{-1/2} + q^{-1/2}N^{-1/2}$, which is given by Assumption 3.

(ii) As $\left\|\beta_{(1)}\right\| \lesssim q^{1/2}N^{1/2}$, (ii) is implied by (i).

(iii) Left-multiplying (1.89) by $(u_T)'_{[I_1]}$, as $\left\|(u_T)'_{[I_1]}\beta_{[I_1]}\right\| \lesssim_{\mathrm{P}} q^{1/2}N^{1/2}$, to prove (iii) when $k = 1$, it is sufficient to show that

$$T_h^{-1/2}q^{-1/2}N^{-1/2}\left\|(u_T)'_{[I_1]}U_{[I_1]}\widehat{\xi}_{(1)}\right\| \lesssim_{\mathrm{P}} T^{-1}q^{1/2}N^{1/2} + q^{-1/2}N^{-1/2},$$

which is implied by Lemma 7.

(iv)-(vi) are equivalent to (i)-(iii) when $k = 1$ as $\widetilde{D}_{(1)} = D_{(1)} = (\mathbb{I}_N)_{[I_1]}$.

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

(i) Similar to the $k = 1$ case, using (1.88) and Lemma 3(iii)(iv), it is sufficient to show that $T_h^{-1/2} q^{-1/2} N^{-1/2} \left\| \widetilde{U}_{(k)} \right\| \lesssim_{\mathrm{P}} T^{-1/2} + q^{-1/2} N^{-1/2}$ and $q^{-1/2} N^{-1/2} \left\| (\widetilde{D}_{(k)} - D_{(k)})\beta \right\| \lesssim_{\mathrm{P}} q^{-1/2} N^{-1/2} + T^{-1/2}$. The first inequality is the same as the $k = 1$ case, which is implied by Lemma 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 (v) holds for $i < k$, it is sufficient to show that

$$\left\| \frac{\beta_{[I_k]} b_{(i)}}{\sqrt{\lambda_{(i)}}} - \frac{X_{[I_k]} \widehat{\xi}_{(i)}}{\sqrt{T_h \widehat{\lambda}_{(i)}}} \right\| \lesssim_{\mathrm{P}} T^{-1/2} + q^{-1/2} N^{-1/2}. \tag{1.90}$$

Plugging $X_{[I_k]} = \beta_{[I_k]} F + U_{[I_k]}$ into (1.90) and using Lemma 3(iii)(iv) again, we only need to show that $T_h^{-1/2} q^{-1/2} N^{-1/2} \left\| U_{[I_k]} \widehat{\xi}_{(i)} \right\| \lesssim_{\mathrm{P}} q^{-1/2} N^{-1/2} + T^{-1/2}$, which holds by Assumption 3 and $\left\| \widehat{\xi}_{(i)} \right\| = 1$.

(ii) It is implied by (i) as $\left\| \beta_{[I_k]} \right\| \lesssim q^{1/2} N^{1/2}$.

(iii) By (1.88), we have

$$\begin{aligned}(u_T)'_{[I_k]} \widehat{\varsigma}_{(k)} - (u_T)'_{[I_k]} \varsigma_{(k)} =& (u_T)'_{[I_k]} (T_h^{-1/2} \widehat{\lambda}_{(k)}^{-1/2} \widetilde{D}_{(k)} \beta F \widehat{\xi}_{(k)} - \lambda_{(k)}^{-1/2} D_{(k)} \beta_{(k)} b_{(k)}) \\ &+ T_h^{-1/2} \widehat{\lambda}_{(k)}^{-1/2} (u_T)'_{[I_k]} \widetilde{U}_{(k)} \widehat{\xi}_{(k)}.\end{aligned}$$

As in the $k = 1$ case, for the second term, we have $T_h^{-1/2} q^{-1/2} N^{-1/2} \left\| (u_T)'_{[I_k]} \widetilde{U}_{(k)} \widehat{\xi}_{(k)} \right\| \lesssim_{\mathrm{P}} T^{-1} q^{1/2} N^{1/2} + q^{-1/2} N^{-1/2}$, as is given by Lemma 7(iv). For the first term, similar to the proof of (i), using Lemma 3(iii)(iv), it is sufficient to show that

$q^{-1/2}N^{-1/2}\left\|(u_T)'_{[I_k]}(\widetilde{D}_{(k)}-D_{(k)})\beta\right\| \lesssim_{\mathrm{P}} T^{-1}q^{1/2}N^{1/2}+q^{-1/2}N^{-1/2}$. Write

$$\begin{aligned}&(u_T)'_{[I_k]}(\widetilde{D}_{(k)}-D_{(k)})\beta\\ =&\sum_{i<k}\left(\frac{(u_T)'_{[I_k]}\beta_{[I_k]}b_{(i)}}{\sqrt{\lambda_{(i)}}}\varsigma'_{(i)}D_{(i)}\beta-\frac{(u_T)'_{[I_k]}(\beta_{[I_k]}F+U_{[I_k]})\widehat{\xi}_{(i)}}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\widehat{\varsigma}'_{(i)}\widetilde{D}_{(i)}\beta\right).\end{aligned}$$

As (v) holds for $i<k$, we only need to show that

$$\left\|\frac{(u_T)'_{[I_k]}\beta_{[I_k]}b_{(i)}}{\sqrt{\lambda_{(i)}}}-\frac{(u_T)'_{[I_k]}(\beta_{[I_k]}F+U_{[I_k]})\widehat{\xi}_{(i)}}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\right\|\lesssim_{\mathrm{P}} T^{-1}q^{1/2}N^{1/2}+q^{-1/2}N^{-1/2}. \quad (1.91)$$

Using Lemma 3(iii)(iv) again, it is sufficient to show

$$T_h^{-1/2}q^{-1/2}N^{-1/2}\left\|(u_T)'_{[I_k]}U_{[I_k]}\widehat{\xi}_{(i)}\right\|\lesssim_{\mathrm{P}} T^{-1}q^{1/2}N^{1/2}+q^{-1/2}N^{-1/2},$$

which holds by Lemma 7(iii).

(iv) By simple algebra, we have

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

Using the fact that (i) holds for k, (iii) holds for $i<k$ and (1.90), the proof is completed.

(v) Similarly, we have

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

Using the fact that (i) holds for k, (v) holds for $i < k$ and (1.90), the proof is completed.

(vi) Replace $q^{-1/2}N^{-1/2}\beta$ by u_T in the proof of (v), we obtain (vi). □

Lemma 9. *Under assumptions of Theorem 1, for $k \leq \tilde{K}+1$, we have*

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

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

Proof. (i) From the definition (1.18) 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 5(ii), we have

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

(ii) With (1.18) 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 5(i) and the assumptions on q, lead to

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

□

Lemma 10. *Under assumptions of Theorem 1, B_1, B_2 defined by* (1.31) *satisfy*

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

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

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

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

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

Proof. (i) Using the definition (1.31) 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 (1.31) of B_2, we have

$$\|b_{k2}\| = \left\|\frac{\widetilde{\beta}_{(k)}'\widehat{\varsigma}_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}}\right\| \leq q^{-1/2}N^{-1/2}\left\|\widetilde{\beta}_{(k)}\right\|. \tag{1.92}$$

Note that

$$\left\|\widetilde{\beta}_{(k)}\right\| \leq \left\|\beta_{[I_k]}\right\| + \sum_{i=1}^{k-1}\left\|\frac{X_{[I_k]}\widehat{\xi}_{(i)}}{\sqrt{T_h\widehat{\lambda}_{(i)}}}\right\|\left\|\widehat{\varsigma}_{(i)}'\widetilde{\beta}_{(i)}\right\| \lesssim_{\mathrm{P}} q^{1/2}N^{1/2} + \sum_{i=1}^{k-1}\left\|\widetilde{\beta}_{(i)}\right\| \tag{1.93}$$

and $\left\|\widetilde{\beta}_{(1)}\right\| = \left\|\widetilde{\beta}_{[I_1]}\right\| \lesssim q^{1/2}N^{1/2}$, we have $\left\|\widetilde{\beta}_{(k)}\right\| \lesssim q^{1/2}N^{1/2}$ by induction. Together with (1.92), we have $\|b_{k2}\| \lesssim_{\mathrm{P}} 1$ and thus $\|B_2\| \lesssim_{\mathrm{P}} 1$.

(ii) By (1.12) and Lemma 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 5(iii), we have $|b'_{l1}b_{k2} - \delta_{lk}| \lesssim_{\mathrm{P}} q^{-1}N^{-1} + T^{-1}$, and thus $\left\|B'_1B_2 - \mathbb{I}_{\widetilde{K}}\right\| \lesssim_{\mathrm{P}} q^{-1}N^{-1} + T^{-1}$.

(iii) Using (1.12) 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

$$b_{k1} = \frac{FF'}{T_h}b_{k2} + \frac{F\widetilde{U}'_{(k)}\widehat{\varsigma}_{(k)}}{T_h\sqrt{\widehat{\lambda}_{(k)}}}. \tag{1.94}$$

With $\|B_2\| \lesssim_{\mathrm{P}} 1$, Assumption 1 and Lemma 5(ii), (1.94) leads to

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

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

(iv) When $\tilde{K} = K$, B'_2B_2 is a $K \times K$ matrix. By (i), (ii) and (iii), we have

$$\left\|B'_2B_2 - \mathbb{I}_K\right\| \leq \left\|B'_1B_2 - \mathbb{I}_K\right\| + \|B_1 - B_2\|\,\|B_2\| \lesssim_{\mathrm{P}} T^{-1/2} + q^{-1}N^{-1}.$$

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

$$\left\|B_2B'_2 - \mathbb{I}_K\right\| = \max_{1\leq i\leq K}|\lambda_i(B'_2B_2) - 1| = \left\|B'_2B_2 - \mathbb{I}_K\right\| \lesssim_{\mathrm{P}} T^{-1/2} + q^{-1}N^{-1}.$$

(v) With respect to $B_1B'_2$, we have

$$\sigma_K(B_2)\left\|B_2B'_1 - \mathbb{I}_K\right\| \leq \left\|(B_2B'_1 - \mathbb{I}_K)B_2\right\| = \left\|B_2(B'_1B_2 - \mathbb{I}_K)\right\| \leq \sigma_1(B_2)\left\|B'_1B_2 - \mathbb{I}_K\right\|. \tag{1.95}$$

Since (iv) implies that $\sigma_1(B_2)/\sigma_K(B_2) \lesssim_{\mathrm{P}} 1$ when $\tilde{K} = K$, (ii) and (1.95) yield

$$\left\|B_1 B_2' - \mathbb{I}_K\right\| = \left\|B_2 B_1' - \mathbb{I}_K\right\| \leq \frac{\sigma_1(B_2)}{\sigma_K(B_2)} \left\|B_1' B_2 - \mathbb{I}_K\right\| \lesssim_{\mathrm{P}} T^{-1} + q^{-1}N^{-1}. \tag{1.96}$$

□

Lemma 11. *Under Assumptions 1-5, we have*

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

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

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

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

Proof. (i) Recall that by (1.87), (1.12), and (1.14), 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

$$\begin{aligned}&\left\|T_h^{1/2}\widehat{\xi}_{(k)} - b_{k2}'F\right\|_{\mathrm{MAX}} \\ &\lesssim_{\mathrm{P}} q^{-1}N^{-1}T^{-1/2}\left(\left\|\widehat{\xi}_{(k)}'F'\beta_{[I_k]}'\widetilde{U}_{(k)}\right\|_{\mathrm{MAX}} + \left\|\widehat{\xi}_{(k)}'U_{[I_k]}'\widetilde{U}_{(k)}\right\|_{\mathrm{MAX}}\right).\end{aligned} \tag{1.97}$$

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

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

Now suppose that this property holds for $i < k$, then for the first term in (1.97), we have

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

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

$$\left\|\widehat{\varsigma}'_{(i)}\widetilde{U}_{(i)}\right\|_{\text{MAX}} \lesssim_{\text{P}} (\log T)^{1/2} + q^{1/2}N^{1/2}T^{-1/2} + q^{-1/2}N^{-1/2}.$$

With $\left\|\beta_{[I_k]}\right\| \lesssim q^{1/2}N^{1/2}$ and $\left\|X_{[I_k]}\right\| \leq \left\|\beta_{[I_k]}\right\| \|F\| + \left\|U_{[I_k]}\right\| \lesssim_{\text{P}} q^{1/2}N^{1/2}T^{1/2}$ and Assumption 4(ii), we have the first term in (1.97) satisfies

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

For the second term in (1.97), we have

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

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

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

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

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

(iii) By triangle inequality, we have

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

By Assumptions 1, 2 and Lemma 10, the first term satisfies

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

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

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

which, together with (i)(ii), conclude 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 $T^{-1/2}\left\|\widehat{\Lambda}^{-1/2}\widehat{F} - B_2'F\right\|$

$\lesssim_{\mathrm{P}} T^{-1} + q^{-1/2}N^{-1/2}$ from Lemma 6, $\|\beta\|_{\mathrm{MAX}} \lesssim 1$, Lemma 10 and (ii), we have

$$
\begin{aligned}
&\max_i T_h^{-1/2} \left\| \beta_{[i]} \left(B_1 B_2' - \mathbb{I}_K \right) F \right\| \\
&\lesssim T_h^{-1/2} \|\beta\|_{\mathrm{MAX}} \left\| B_1 B_2' - \mathbb{I}_K \right\| \|F\| \lesssim_{\mathrm{P}} q^{-1}N^{-1} + T^{-1}, \\
&\max_i T_h^{-1/2} \left\| \beta_{[i]} B_1 \right\| \left\| \widehat{\Lambda}^{-1/2} \widehat{F} - B_2' F \right\| \\
&\lesssim T_h^{-1/2} \|\beta\|_{\mathrm{MAX}} \left\| \widehat{\Lambda}^{-1/2} \widehat{F} - B_2' F \right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1}, \\
&\max_i T_h^{-1/2} \left\| \beta_{[i]} B_1 - \widehat{\beta}_{[i]} \widehat{\Lambda}^{1/2} \right\| \left\| \widehat{\Lambda}^{-1/2} \widehat{F} \right\| \\
&\lesssim T_h^{-1/2} \left\| \beta B_1 - \widehat{\beta} \widehat{\Lambda}^{1/2} \right\|_{\mathrm{MAX}} \|F\| \lesssim_{\mathrm{P}} T^{-1/2} (\log N)^{1/2}.
\end{aligned}
$$

Consequently, we have the desired bound. □

Lemma 12. *Under Assumptions 1-4, for any $I \subset [N]$, we have the following results:*

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

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

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

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

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

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

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

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

(ix) $\left\| (u_T)_{[I]}' \underline{U}_{[I]} \mathbb{M}_{W'} F' \right\| \lesssim_{\mathrm{P}} |I| + |I|^{1/2} T^{1/2}$, $\left\| (u_T)_{[I]}' \underline{U}_{[I]} \mathbb{M}_{W'} Z' \right\| \lesssim_{\mathrm{P}} |I| + |I|^{1/2} T^{1/2}$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(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,

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

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

(ix) By Assumptions 1 and 4, we have

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

Replacing F by Z, we have the second inequality. □

Lemma 13. *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 $j \leq K$, $\sigma_j(\beta F)^2 = \lambda_j(\beta FF'\beta') = \lambda_j(\beta'\beta FF')$, 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 $T_h^{-1}\beta FF'\beta'$. As $\left\|\beta\beta' - T_h^{-1}\beta FF'\beta'\right\| \leq \|\beta\|^2 \left\|T_h^{-1}FF' - \mathbb{I}_K\right\| \lesssim_{\mathrm{P}} \sigma_1(\beta)^2 T^{-1/2}$ and $\sigma_1(\beta) - \sigma_2(\beta) \asymp \sigma_1(\beta)$, by sin-theta theorem

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

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,

$$\left\|\mathbb{P}_{\tilde{\xi}} - \frac{\sigma_1(\beta)^2}{\|\beta F\|^2} F'\mathbb{P}_b F\right\| = \left\|\tilde{\xi}\tilde{\xi}' - \frac{F'\beta'\varsigma\varsigma'\beta F}{\|\beta F\|^2}\right\| = \left\|\frac{F'\beta'\tilde{\varsigma}\tilde{\varsigma}'\beta F}{\|\beta F\|^2} - \frac{F'\beta'\varsigma\varsigma'\beta F}{\|\beta F\|^2}\right\| \lesssim_{\mathrm{P}} T^{-1/2}. \tag{1.98}$$

By Weyl's inequality, $T_h^{-1}\|\beta F\|^2 = \lambda_1(T^{-1}\beta FF'\beta') = \sigma_1(\beta)^2 + O_{\mathrm{P}}(\sigma_1(\beta)^2 T^{-1/2})$. In light of (1.98), we have $\left\|\mathbb{P}_{\tilde{\xi}} - T_h^{-1}F'\mathbb{P}_b F\right\| \lesssim_{\mathrm{P}} T^{-1/2}$. □

CHAPTER 2

TEST ASSETS AND WEAK FACTORS

2.1 Introduction

Estimation and inference on factor models are central elements of empirical work in asset pricing. Typically, a researcher starts with a given factor, for example an aggregate liquidity factor, motivated by economic theory. The objective of the researcher is to estimate and test its risk premium. To proceed, the researcher needs to decide which test assets to use in the estimation. While the literature has made a variety of choices for test assets, little work has been dedicated to investigating rigorously and systematically how they should be chosen.

Another issue the researcher has to face is the potential presence of weak factors. Broadly speaking, the factor of interest to the researcher is one of many factors potentially driving returns. Some may be weak: these are factors to which the available test assets have little or no exposure. This makes it difficult to learn about them using the available assets. Their presence also contaminates inference about the entire model: the literature shows the presence of a weak factor biases the estimation of the risk premia of all factors, including the one of interest to the researcher (whether that factor itself is strong or weak) as well as the inference about the pricing ability of the model. To make things worse, a weak factor could be latent, so that we may not even know it exists in the first place.

In this paper, we document a deep connection between the selection of test assets and the long-standing problem of weak factors in asset pricing. Exploiting this connection, we propose a novel methodology, supervised principal component analysis (SPCA), that serves a dual purpose: first, it provides a well-founded basis for the selection of test assets, and second, it leverages the selection to mitigate the bias in risk premium estimation for the factor of interest to the researcher, irrespective of its strength and the strength of (known or unknown) factors in the panel of test asset returns.

The connection we emphasize between weak factors and test assets is that the strength or weakness of a factor (whether it is observable or latent), should not be viewed as a property of the factor itself, as typical in the asset pricing literature; rather, it should be viewed as a property of the set of test assets used in the estimation. As an example, a liquidity factor may be weak in a cross-section of portfolios sorted by, say, size and value, but may be strong in a cross-section of assets sorted by characteristics that capture well exposure to liquidity.

This perspective provides clear guidance on how to choose test assets: select them in a way that yields a consistent estimate of the risk premium of the factor chosen by the researcher, and that is robust to the presence of observable or latent weak factors among those driving returns. This criterion is statistical in nature, and offers an agnostic selection and estimation technique that complements alternative selection strategies found in the literature, where researchers often use strong economic priors or ad-hoc methodologies to determine which test assets to include and which to exclude.

Estimating and testing the risk premium of a factor of interest requires properly controlling for all the other factors relevant to investors (whether they are observed or latent), in order to avoid an omitted variable bias (see for example Giglio and Xiu [2021]).[1] Giglio and Xiu [2021] propose to do so by first estimating a latent factor model for the stochastic discount factor (SDF) using principal component analysis (PCA), and then using it to estimate the factor of interest's risk premium. This approach eliminates the need for explicit specification of all the control factors, but relies on the assumption that all the latent factors driving the SDF are pervasive (i.e., strong). Our SPCA procedure also utilizes PCA for extracting latent factors while remaining agnostic about the identities of the control factors. However, it exploits correlations with the factor of interest as a guiding criterion for selecting

1. This is only necessary when the factor of interest is not itself a tradable portfolio (i.e., it is a non-tradable factor, like a macroeconomic risk). If the factor of interest is itself a portfolio (also referred to as tradable factor), like in the case of the CAPM, the computation of the risk premium just requires computing the average excess return of the portfolio. In practice, most economic models have predictions about the risk premia of non-tradable factors.

a subset of test assets, before applying PCA. This results in a versatile methodology that remains robust even in scenarios where certain factors are omitted, including cases where these omitted factors are weak.

Given a factor g_t specified ex-ante by the researcher, the procedure estimates its risk premium as follows. We start from a large universe of potential test assets. In a first step of the procedure (selection step), we compute the univariate correlation of each asset's return with g_t. We select a relatively small portion of assets, only keeping those with sufficiently high correlation (in absolute value): these are assets that are particularly informative about the factor of interest g_t. We then compute the first principal component of the returns of these assets (PCA step), which will be our first estimated latent factor. Next, we remove via linear projection from both g_t and all the returns of the test assets the part explained by this first latent factor (projection step). We then go back to the selection step, computing the univariate correlation of the *residuals* of the factor and the *residuals* of the assets from the projection step. Again, we select from the universe of test assets a subset for which this correlation is especially high, and compute the first principal component of these residuals. This will be our second estimated latent factor. We then further remove (from g_t and the test assets) the part explained by this second estimated factor as well, and iterate again on the residuals. We repeat this procedure $\widehat{p}$ times, where $\widehat{p}$ is a tuning parameter which can be determined by some validation step. In the most desirable scenario, $\widehat{p}$ serves as a desirable estimate of the actual number of factors, p, in the data. This procedure recovers from the data latent factors that are informative about the factor of interest g_t. Importantly, the fact that at each iteration only test assets that are sufficiently correlated with the factor g_t are selected ensures that not only strong, but also weak factors (relative to the entire cross-section) are captured by the procedure – contrary to standard PCA that uses *all* assets at all steps to extract latent factors. Finally, a time-series regression of g_t on the $\widehat{p}$ latent factors yields a consistent estimator of the risk premium of g_t, by linking it to the risk premia of

these latent factors. The latent factors themselves can be thought of as the part of the SDF that is related to g_t and determines its risk premium.

While the supervision of g_t aids in the recovery of factors, including weak ones, this procedure may not retrieve *all* the factors driving the cross-section of returns (i.e., the entire SDF). It specifically ensures the recovery of factors correlated with g_t, while uncorrelated factors, particularly if they are weak, may remain unrecoverable (so it may be true that $\widehat{p} < p$). Fortunately, but crucially, the omission of these factors by SPCA does not affect the consistency of the risk premium estimation for g_t, since such factors do not contribute to the pricing of g_t. That said, complete recovery of all factors remains feasible, contingent on including multiple variables in the target g_t and ensuring that each latent factor has at least one variable in g_t with a non-vanishing exposure to it.

Beyond risk premia estimation, SPCA can also be used to diagnose omitted factors in a model based on a set of observable factors in g_t. Supervised by g_t, SPCA recovers all the latent factors that drive the SDF and correlate with g_t. We prove that SPCA consistently recovers the true SDF if and only if g_t is spanned by all factors that drive the SDF. We apply this result to diagnose whether g_t misses any factors. This diagnosis on g_t can be executed as a simple comparison between the maximal Sharpe ratio achieved by g_t and that achieved by the factors recovered by SPCA. When the latter is larger than the former, it indicates that g_t misses some factor, and that the researcher should seek a better model. On the other hand, if the latter is smaller, it implies that g_t contains factors to which the given cross-section of test assets have insufficient exposures. In such a scenario, a richer set of test assets is needed.

The choice of test assets in the literature has mainly followed one of three approaches. The vast majority of the literature has adopted a "standard" set of portfolios sorted by a few characteristics, such as size and value, following the seminal work by Fama and French [1993]. A second approach, taken more recently, e.g., Kozak et al. [2020], has been to expand this

cross-section to include portfolios sorted by a much larger set of characteristics discovered in the last decades, on the order of hundreds of portfolios. Finally, a third approach, see, e.g., Ang et al. [2006], has been more "targeted" around the specific factor of interest: sorting assets into portfolios by their estimated exposure to the factor, and then estimating risk premia using only these sorted portfolios, that is, using a small cross-section expected to be particularly informative about that factor.

It is useful to contrast the asset selection procedure of SPCA with the three approaches summarized above. Using a standard, small cross-section (like the size- and value-sorted portfolios) to estimate risk premia has the problem that except for size and value, which are strong factors in this cross-section, many other factors are weak: these test assets do not contain sufficient information to identify their risk premia. The second approach may appear, on the surface, to address this issue: a large cross-section of test assets are likely exposed to many potential factors. However, if only a *few* of those many assets are exposed to some factor, whereas most others are not, that factor will, again, be weak. Finally, the third approach – building targeted portfolios of assets sorted by the exposure to the factor of interest to the researcher – is affected by the omitted factor problem, since it considers univariate exposures only; in general, it will fail in a multi-factor context.

In the paper, we derive the asymptotic properties of SPCA, in a setting that allows for weak factors and test assets with highly correlated risk exposures. The latter scenario potentially involves the same (asymptotically) rank-deficiency issue as weak factors. We also analyze in this setting alternative estimators that have been proposed in the recent literature, which rely on PCA, Ridge, Lasso, and Partial Least Squares (PLS). We show that the PCA (and some other variations of it), Ridge, and PLS are inconsistent in the presence of weak factors, and that the Lasso approach is consistent for the estimation of the SDF, as well as risk premia estimation, but is not as efficient as SPCA in general. Additionally, we perform an extensive set of simulations to study the performance of SPCA in different

scenarios. These simulations isolate issues in conventional two-pass regressions, facilitating a clear comparison of SPCA with other estimators. Our findings confirm SPCA's robustness to omitted factors and weak factors, as well as measurement error, which SPCA also tackles.

As expected, a trade-off exists between robustness and efficiency. In scenarios where all factors are strong, the PCA-based approach by Giglio and Xiu [2021] is consistent and likely to outperform SPCA in terms of efficiency. The potential efficiency loss associated with SPCA arises from its selective use of test assets when all of them are in fact informative, or the possibility that it may not recover all factors driving returns. However, the PCA-based estimator is biased in the presence of weak factors, a major concern in empirical applications. We therefore advocate for using SPCA to estimate risk premia due to its robustness when weak factors are potentially present: where consistency is compromised, prioritizing efficiency becomes irrelevant.

The problem of weak factors in latent factor models is closely connected to that of weak factors in observable factor models, which has been widely examined in the literature. The seminal contribution of Kan and Zhang [1999] shows that the inference on risk premia estimates from Fama-MacBeth regressions becomes invalid when a "useless" factor – a factor to which test assets have zero exposures – is included in the model. Kleibergen [2009] further points out the failure of the standard inference even for strong factors, if betas are relatively small.[2] In our paper, we show that the same logic applies in the context of latent factor models: if some (latent) factors are weak in a cross-section, the PCA estimator will not be able to disentangle them from idiosyncratic error, leading to biases in the estimated factors and their risk premia.

The issue of weak factors is particularly important in empirical work in asset pricing, because most economically-motivated factors (e.g., most macroeconomic factors) do seem to

2. Related literature also include Gospodinov et al. [2013] and Gospodinov et al. [2014]. On the other hand, Pesaran and Smith [2019] investigate the impact of factor strength and pricing error on risk premium estimation. They point out that the conventional two-pass risk premium estimator converges at a lower rate as the factors become weaker.

be weak in practice. Moreover, a statistical problem analogous to weak factors arises when betas are collinear, that is, some factors are redundant in terms of explaining the variation of expected returns. This is again a relevant issue in practice due to the existence of hundreds of factors discovered in the literature, see, e.g., Harvey et al. [2016], many of which are close cousins and do not add any explanatory power (Feng et al. [2020]). The weak factor problem appears to be caused by having seemingly more factors than necessary, which is why some suggest eliminating such factors (Bryzgalova [2015]) or shrinking their risk premia estimates (Bryzgalova et al. [2019]), so as to improve the estimates for strong factors. We instead argue that the weak factor problem is fundamentally an issue of test asset selection. Since weaker factors may still be priced, our solution is to accommodate them using an adapted procedure with carefully selected test assets.[3]

Several recent papers have proposed different methodologies to deal with weak factors. Lettau and Pelger [2020] propose an estimator of the SDF in the presence of weak factors, rpPCA, which generalizes PCA with a penalty term that accounts for expected returns. Whereas this estimator features desirable properties as explored by Lettau and Pelger [2020], we show that it is inconsistent for estimating risk premia in the weak-factor setting we consider.[4] Anatolyev and Mikusheva [2021] propose an complementary four-split approach to dealing with weak factors, based on sample-splitting and instrumental variables. This

3. It is worth noting that whereas some theories assume that only strong factors can be priced, this is not true in general for two reasons. First, many theoretical models – e.g., the consumption-CAPM – are silent on what assets are traded in equilibrium, and if markets are incomplete, it may very well be that some priced factors may not be reflected in many of the assets that are traded. Second, even if investors may have access to many assets exposed to a particular factor, the econometrician may not, making the factor weak for the set of test assets available to the econometrician.

4. Lettau and Pelger [2020] focus their analysis on the case where factors are extremely weak – so much so that they are not statistically distinguishable from idiosyncratic noise. In that case, no estimator can be consistent for either risk premia or the SDF. They show that rpPCA does not recover consistently the SDF, but in simulations it correlates with the SDF more than the SDF estimator obtained from standard PCA. Rather than focusing on this extreme case of weak factors, our theory covers a range of factor weaknesses, which includes the cases from strong to very weak, and which still permits consistent estimation of factors and risk premia. Formally, we study the case where the minimum eigenvalues of the factor component in the covariance matrix of returns diverges whereas the largest eigenvalue due to the idiosyncratic errors is bounded.

alternative procedure works well to address the weak factor bias, though it does not deal with omitted priced factors or with measurement error in the factors.[5]

Our paper also relates to a literature that has explored different methods to form portfolios to test asset pricing models, like Ahn et al. [2009] or Bryzgalova et al. [2020]. These methods are useful in helping to build or expand the starting cross-section for SPCA. In this paper, we use the simpler approach of working with an existing large cross-section of portfolios sorted by firm characteristics, as in Chen and Zimmermann [2020] and Hou et al. [2020].

The concept of supervised-PCA originated from a cancer diagnosis technique applied to DNA microarray data by Bair and Tibshirani [2004], and was later formalized by Bair et al. [2006] in a prediction framework, in which some predictors are not correlated with the latent factors that drive the outcome of interest. Bair et al. [2006] suggest a screening step using marginal correlations between predictors and the outcome variable to select the subset of useful predictors, before applying the standard PCA to this subset.[6] They prove the consistency of this so-called SPCA procedure, but relying on a restrictive identification assumption that any important predictor must also have a substantial marginal correlation with the outcome. We provide several examples of multivariate factor models in which this assumption fails. While the screening step of our SPCA procedure shares the spirit with theirs (in the sense that their outcome variable is our factor of interest, and their predictors

5. Our paper also relates to a growing strand of econometrics literature on weak factor models. Bai and Ng [2021] show that PCA can recover moderately weak factors at the cost of efficiency. Bai and Ng [2008] and Huang et al. [2022] propose supervised learning methods in the context of factor-based forecasting. Fan et al. [2021] also exploit information from observed proxies to improve the estimation of factor models, and Wan et al. [2023] consider moderately weak factors as in Bai and Ng [2021] in this context. Fan and Liao [2022] propose to extract factors by diversifying away idiosyncratic noise directly. Uematsu and Yamagata [2022a] adopt a variant of the sparse PCA algorithm proposed in Uematsu et al. [2019] to estimate a sparsity-induced weak factor model. Uematsu and Yamagata [2022b] provide inference results in that sparse model. Freyaldenhoven [2022] and Bailey et al. [2020] adopt a similar framework for estimating factor count and strength.

6. The screening approach has also been adopted in the contexts, such as classification and regression, see Fan and Fan [2008], Fan and Lv [2008].

are our test assets), our projection step and the subsequent iteration procedure are new, and are introduced precisely to eliminate the strong identification assumption used in the existing statistics literature. Also, our focus is not on prediction per se, but instead on parameter inference.

2.2 Methodology

To rigorously address the challenge of weak factors, our approach begins with the specification of a general Data Generating Process (DGP). It is crucial to underscore that within this population model, the concept of weak factors holds no relevance. In population, researchers aiming to identify the risk premium of a factor like g_t would ideally utilize all available assets for this purpose.

However, the real-world (finite-sample) scenario diverges from this idealized population model. We encounter practical constraints, such as a large number of assets (large N), relatively short time spans (small T), and a significant proportion of assets that are only weakly correlated with the target variable g_t. We characterize this finite sample context using asymptotic concepts, formally defining the notion of weak factors. This particular asymptotic perspective is useful as it enables us to investigate the issues of weak factors arising in finite samples with existing estimators, and understand the properties of our proposed solution.

2.2.1 Model Setup

We study a standard linear factor model setup. Suppose that an $N \times 1$ vector of test asset excess returns, r_t, follows:

$$r_t = \beta\gamma + \beta v_t + u_t, \quad \mathrm{E}(v_t) = \mathrm{E}(u_t) = 0 \text{ and } \mathrm{Cov}(v_t, u_t) = 0, \tag{2.1}$$

where β is an $N \times p$ matrix of factor exposures, v_t is a $p \times 1$ vector of innovations of p factors f_t (i.e., $v_t = f_t - \mu_f$, where $\mu_f = \mathrm{E}(f_t)$), and u_t is an $N \times 1$ vector of idiosyncratic errors.

We assume that the vector of factor innovations v_t is not fully observable. Specifically, we allow the asset pricing factors f_t to be either latent or observable. In the former case, innovations v_t are naturally also latent. Even in the latter case, when a factor f_t is observable, its innovation v_t is not directly observable because μ_f is an unknown parameter.[7]

Also, note that we model risk exposures (β) as constant: we implicitly assume that the test assets are portfolios sorted so that their factor exposures are modelled as constant, as in Giglio and Xiu [2021]. Alternatively, one could work directly with individual stocks (which generally have time-varying risk exposure), combining our procedure with the methodologies of Gagliardini et al. [2016], Kelly et al. [2019], or Kim et al. [2020] to account for the time-variation in betas.

We situate our discussion within the framework of two standard asset pricing exercises: the estimation of risk premia and the recovery of the SDF. Given our model, an SDF can be defined in terms of factors v_t as

$$m_t = 1 - \gamma^\intercal \Sigma_v^{-1} v_t, \tag{2.2}$$

where Σ_v is the covariance matrix of factor innovations, see, e.g., Giglio and Xiu [2021]. It also makes sense to consider the SDF represented in terms of the set of tradable test asset returns:

$$\widetilde{m}_t = 1 - b^\intercal (r_t - \mathrm{E}(r_t)), \tag{2.3}$$

where b is an $N \times 1$ vector of SDF loadings which satisfies $\mathrm{E}(r_t) = \Sigma b$, where Σ is the

7. In Appendix section 2.5.3.2 we discuss the case in which factors are observable, and in Appendix section 2.5.3.3 we discuss the case in which the zero-beta rate needs to be estimated.

covariance matrix of r_t, see, e.g., Kozak et al. [2020]. The relationship between the two SDFs depends on the degree of completeness of markets. As will be shown later, these two forms of the SDF are asymptotically equivalent in the asymptotic scheme we consider, with the number of assets N going to infinity, so that there is no ambiguity with respect to which estimand we consider.

In addition to the SDF, we are also interested in estimating the risk premium of one or more observable factors, summarized in a $d \times 1$ vector, g_t. It is important to emphasize that g_t is a proxy for some risks, constructed or otherwise chosen by the researcher ex-ante, not necessarily tradable, and typically motivated from economic theory or narratives. Following Giglio and Xiu [2021], we do not impose that g_t is part of or identical to v_t; instead, g_t and v_t are assumed (potentially) correlated:

$$g_t = \xi + \eta v_t + z_t, \tag{2.4}$$

where $\xi = \mathrm{E}(g_t)$, η is a $d \times p$ matrix, and z_t is measurement error orthogonal to v_t.[8] This model clearly nests the classic linear asset pricing model with observable factors only, in which case we can set $\eta = \mathbb{I}_p$ and $z_t = 0$. To price g_t, we can simply use the SDF given by (2.4), as g_t's risk premium is given by $\gamma_g = -\mathrm{Cov}(g_t, m_t) = \eta\gamma$.

To characterize the strength of a factor, we need to set up an asymptotic environment in which weak factors may arise. First, let us introduce some useful notation. We use the notation $a \lesssim b$ to denote $a \leq Kb$ for some constant $K > 0$, and if $a \lesssim b$ and $b \lesssim a$, we write $a \asymp b$ for short. We use similar notation $\lesssim_{\mathrm{P}}$ and $\asymp_{\mathrm{P}}$ for bounded in probability. Also, for any matrix A, we use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote its minimum and maximum eigenvalues, and $\lambda_i(A)$ is the i-th largest eigenvalue.

The environment in which we study weak factors is quite general, and is characterized by

8. When g_t is nontradable, measurement error could arise as the econometrician is implementing an empirical counterpart of some theory-predicted factor; when g_t is tradable, it captures the non-diversified errors in the portfolio.

three assumptions. First, we assume that both N and T go to ∞ (at arbitrary rates, unless we specify otherwise), whereas the number of factors p is fixed. Letting N go to infinity in addition to T is rather natural in the asset pricing context, as motivated in theory by Ross' APT (Ross [1976]) and given the proliferation of "anomalies" generated by the empirical literature in the past decades. Second, we assume that the $p \times p$ factor covariance matrix of the factor innovations, Σ_v, is asymptotically non-singular: $1 \lesssim \lambda_{\min}(\Sigma_v) \leq \lambda_{\max}(\Sigma_v) \lesssim 1$. This assumption is rather weak, as it only rules out factors whose risks are (asymptotically) negligible or exploding. Finally, we also maintain the assumption that $\|\Sigma_u\| \lesssim 1$, where $\|\|$ indicates the spectral norm of a matrix, so that there exists no factor structure in the residuals u_t. This assumption is widely adopted in the so-called approximate factor models proposed by Chamberlain and Rothschild [1983].[9]

We are now ready to characterize the strength of factors, as an exclusive function of test assets' *exposures* to the factors, as opposed to a property of the factors themselves. We formalize here the idea that, for instance, a momentum factor could be a strong factor when the test assets are momentum-sorted portfolios, but this same factor may be weak when the test assets are portfolios sorted by size or value: the latter portfolios may diversify away the exposures to the momentum factor, and therefore may be uninformative about momentum risk.

In the econometrics literature on factor models (for example, Bai and Ng [2002]), the setup described in (2.1) is typically complemented by the assumption that $\lambda_i(\beta^\intercal\beta) \asymp N$ for $i = 1, 2, \ldots, p$: all eigenvalues of the matrix $\beta^\intercal\beta$ grow at rate N, so that *all* factors are *pervasive*. Informally, even as the number of test assets N is large, there is a sufficiently large number of assets that are well exposed to each of the risk factors (their β with respect to all factors are non-vanishing for a large number of assets). Under this assumption, as we

9. We only need u_t to be stationary (so that Σ_u is well defined) when we discuss the SDF in Section 2.2.3. For risk premia estimation, we instead impose a weaker condition, namely, Assumption 10, which plays a similar role as $\|\Sigma_u\| \lesssim 1$.

will review later, standard PCA works well to recover the latent factors v_t.

This is the point of departure of our paper: we study situations in which this pervasiveness assumption fails, with respect to some or even all factors. Formally, we define the presence of *weak factors* as the case in which some of those eigenvalues, $\lambda_i(\beta^\intercal\beta)$, grow at a slower rate than N (which will be made more precise later). Intuitively, in this case, while the number of test assets N is large, many test assets may have small or zero exposures to some or all of the factors, making those factors weak. The lack of test asset exposures to a factor makes it more difficult for standard PCA to recover this factor; and in more extreme cases, PCA completely fails to recover either the factors or their loadings. In our setting, the strength/weakness of a factor is actually not a binary distinction. Rather, we allow for a continuum of strength/weakness of factors, determined by how large the exposures to the risk factors are (formally, by the asymptotic behavior of the eigenvalues $\lambda_i(\beta^\intercal\beta)$).

How relevant do we expect these weak factors to be in practice? Consider Figure 3.6, the scree plot of the eigenvalues of returns from our empirical analysis, which uses a large cross-section of 950 assets. The figure shows that the first one or two eigenvalues are clearly much larger than the others. But the absence of clear gaps among the remaining eigenvalues suggests that several factors beyond the first two may be weak. Despite the large cross-section, their eigenvalues remain relatively small and difficult to distinguish from idiosyncratic error.

Our model naturally allows g_t to be weak, since the true factors in v_t are potentially weak and the observable factors in g_t inherit this weakness through their loading on v_t, η. However, as N and T increase, the risk premium associated with g_t, $\eta\gamma$, may not necessarily converge to zero. This is because neither the risk exposure of g_t to v_t, represented by η, nor the risk premiums of v_t, denoted as γ, necessarily diminish asymptotically. In simpler terms, weak factors in this model can still have non-zero risk premia as the sample size and the cross-sectional dimension grow.

2.2.2 Estimating Risk Premia when Factors are Weak

We begin our analysis with risk premia estimation.

2.2.2.1 The Benchmark PCA-based Estimator

Giglio and Xiu [2021] study this problem in a similar setup as in this paper, except that all factors in v_t are assumed to be strong. They propose a three-pass procedure to estimate g_t's risk premium $\eta\gamma$: 1) apply PCA to the sample covariance matrix of returns to obtain estimates of the latent factors, $\widehat{v}_t$; 2) use Fama-MacBeth regressions to recover the risk premia of $\widehat{v}_t$, $\widehat{\gamma}$; 3) use time series regressions of g_t on $\widehat{v}_t$ to estimate $\widehat{\eta}$. The product of the estimates at steps 2 and 3 yields $\widehat{\eta}\widehat{\gamma}$, the estimate of risk premia. We summarize this procedure in the following algorithm:

Algorithm 5 (PCA-based Estimator of Risk Premia)**.** *The estimator proceeds as follows: Inputs: $\bar{R}$ and $\bar{G}$, the matrices of demeaned returns and demeaned g_t, respectively.*[10]

S1. Apply singular-value decomposition (SVD) on $\bar{R}$, and write the first p right singular vectors as ξ. The estimated factors are given by $\widehat{V} = \sqrt{T}\xi^{\intercal}$.

S2. Estimate the risk premia of $\widehat{V}$ by $\widehat{\gamma} = (\widehat{\beta}^{\intercal}\widehat{\beta})^{-1}\widehat{\beta}^{\intercal}\bar{r}$, where $\widehat{\beta} = \bar{R}\widehat{V}^{\intercal}(\widehat{V}\widehat{V}^{\intercal})^{-1}$ and $\bar{r}$ is the vector of average excess returns.

S3. Estimate the factor loading of g_t on v_t by $\widehat{\eta} = \bar{G}\widehat{V}^{\intercal}(\widehat{V}\widehat{V}^{\intercal})^{-1}$.

Outputs: $\widehat{V}, \widehat{\eta}$, $\widehat{\gamma}$, and $\widehat{\gamma}_g^{PCA} = \widehat{\eta}\widehat{\gamma}$.

As discussed in Giglio and Xiu [2021], one interpretation of this estimator is that it builds a mimicking portfolio for the factor g_t, by projecting it onto the first p principal components of the space of returns. A mimicking portfolio would be ideally built directly using *all*

10. For any time series of vectors $\{a_t\}_{t=1}^T$, we denote $\bar{a} = \frac{1}{T}\sum_{t=1}^T a_t$. In addition, we write $\bar{a}_t = a_t - \bar{a}$. We use the capital letter A to denote the matrix $(a_1, a_2, \cdots, a_T)$, and write $\bar{A} = A - \bar{a}\iota_T^{\intercal}$ correspondingly.

possible assets. But when N is large, this can be inefficient or even infeasible (if $N > T$). The three-step estimator effectively regularizes the mimicking portfolio problem by using only p portfolios appropriately constructed as basis assets, i.e., the principal components of the returns. Giglio and Xiu [2021] establish the consistency of this estimator and derive its asymptotic inference, in the case that all latent factors are strong. This procedure also recovers the SDF, because it consistently estimates all latent factors, $\widehat{v}_t$ (columns of $\widehat{V}$), that drive the SDF, along with their SDF loadings as in (2.2), $\widehat{\Sigma}_v^{-1}\widehat{\gamma}$.

This estimator is appealing for its simplicity, efficiency, and, importantly, robustness to missing factors (since the identity of any factors beyond g_t does not need to be specified). Unfortunately, it fails precisely when some latent factors are weak, which we will show next.

To understand this, it is sufficient to consider a one-factor model with $p = d = 1$ and $\Sigma_v = 1$, in which case the covariance matrix of returns satisfies: $\Sigma = \beta\beta^\intercal + \Sigma_u$. This matrix has a noisy low rank structure in that $\beta\beta^\intercal$ has rank 1, whereas Σ_u is a full-rank covariance matrix. To make the exposition simple, we also assume that g_t has no measurement error, i.e., $z_t = 0$ and $g_t = \eta v_t$.

As discussed above, the problem of weak factors stems from the fact that many assets may not have sufficiently strong exposure to the factor of interest, which hinders the construction of its mimicking portfolio, and in turn, the estimation of its risk premium. This intuition applies also when the weak factor is latent (v_t). In this case, the manifestation of the weak factor problem is that PCA will fail to recover this factor.

Estimation of the latent factors v_t via PCA involves recovering the matrix of risk exposures β from the covariance matrix of realized returns, $\widehat{\Sigma}$. A successful recovery of β via PCA of realized returns therefore requires a favorable signal-to-noise ratio. If the "signal", as measured by $\|\beta\|$, dominates the "noise", which arises from the idiosyncratic component Σ_u and the estimation error in the sample covariance matrix $\widehat{\Sigma} - \Sigma$, the first sample eigenvector of $\widehat{\Sigma}$ would (approximately) span the same space spanned by the true β. Thus using $\widehat{\beta}$, ef-

fectively the eigenvector of $\widehat{\Sigma}$, in the cross-sectional regression step (step 2 of the estimator) would yield a consistent estimator of the risk premium of the estimated latent factor, which in turn leads to a consistent estimator of the risk premium of g_t. Otherwise, if the signal $\|\beta\|$ is so weak that the estimation error in $\widehat{\beta}$ dominates, there would be a non-vanishing angle between the space spanned by $\widehat{\beta}$ and that by β. But estimating risk premia requires comparing the average returns of assets with different betas (e.g., computing the slope in a cross-sectional regression); "measurement" error in the betas thereby induces a bias in the risk premium estimate.

Proposition 7 below shows that the PCA-based estimator is consistent only if $N/(\|\beta\|^2 T) \to 0$. This condition formalizes our notion of factor weakness. In a one-factor model, the factor is weak if this condition fails. We generalize this definition for the case of multiple factors later.

Proposition 7. *Suppose that test asset returns follow a single-factor model in the form of* (2.1) *with* $p = 1$*, that* g_t *satisfies* (2.4) *with* $d = 1$*, that* u_t *and* v_t *are i.i.d. normally distributed and mutually independent, and that* $z_t = 0$*. In addition, suppose that* β *satisfies* $N/(\|\beta\|^2 T) \to B \geq 0$ *and* $\|\beta\| \to \infty$*. Then we have* $\widehat{\gamma}_g^{PCA} \xrightarrow{\text{P}} (1+B)^{-1}\eta\gamma$*.*

In the presence of strong factors, $\|\beta\| \asymp \sqrt{N}$, which leads to $B = 0$ as $T \to \infty$, so there is no bias. In general, the consistency depends on the relative magnitude of N, T, and $\|\beta\|$. When N and T are of the same order, $\|\beta\| \to \infty$ is sufficient for the consistency of risk premia estimation. This makes sense in that the eigenvalue of returns corresponding to this factor is proportional to $\|\beta\|^2$, whereas the eigenvalues for the idiosyncratic errors are bounded, so that $\|\beta\| \to \infty$ guarantees the separation between factors and errors and hence the identification of factors.

This example also shows that the risk premium estimator could be biased even if we have consistent estimator of the factors. In fact, the estimated factors in $\widehat{V}$ are consistent under

the assumptions of Proposition 7 in the sense that $|\text{Corr}(\widehat{V}, V)| \xrightarrow{\text{P}} 1$.[11] However, estimating a large-dimensional vector β given $\widehat{V}$ remains a challenging problem, which requires, additionally, $B = 0$, for consistency.

Section 2.5.1 of the appendix studies how several other estimators perform in a weak-factor setting, including PLS, Ridge regression, and rpPCA. The analysis there reveals that these estimators exhibit failures that mirror that of PCA, despite PLS leveraging information from g_t for supervision and rpPCA being specifically designed for weak factors. None of these estimators, therefore, can address the bias originating from the presence of weak factors.

2.2.2.2 Our Solution: Supervised-PCA and Test Asset Selection

The results in the previous section shed light on the detrimental influence of weak factors on the PCA-based estimator (as well as other existing approaches). As we mention in the introduction, an important difference with the prior literature is that we do not view the weakness of a factor as a property of the factor itself; rather, we see it as a property of the universe of test assets that are used in the estimation. This leads us to find a potential solution in modifying the set of test assets. The solution we propose is to *screen* test assets and only keep those that have nontrivial exposure to the factor of interest g_t. Then, if the factor is strong *within this smaller set of test assets*, it is possible to apply PCA (or other procedures discussed in the appendix) to recover its risk premium. The key idea behind the screening approach is to remove those uninformative assets, focusing the estimation on a set of assets whose exposures are large and dominate the estimation error in β.

To proceed with this idea, we formalize the problem by imposing an assumption that there exists a subset $I_0 \subset \langle N \rangle$,[12] within which test assets feature a strong factor structure. In other words, there exists a subset of assets that are sufficiently *informative* about latent

11. We can further establish that a sufficient condition for consistent recovery of factors is $N/(\|\beta\|^4 T) \to 0$, which clearly holds in the setup of Proposition 7.

12. We use $\langle N \rangle$ to denote the set of integers: $\{1, 2, \ldots, N\}$.

factors driving test asset returns. To be clear, we do not make any assumption about the remaining test assets in the complement set of I_0 — they may or may not be informative. Such a set is thus not uniquely defined. In this regard, this assumption is relatively mild.

To see how this assumption helps, note that in the population model of Proposition 7, the expected excess return of g_t's mimicking portfolio *built only with test assets in* I_0 is

$$\mathrm{Cov}(g_t, r_{t,[I_0]})\mathrm{Cov}(r_{t,[I_0]})^{-1}\mathrm{E}(r_{t,[I_0]}) = \eta\Sigma_v\beta_{[I_0]}^{\intercal}(\beta_{[I_0]}\Sigma_v\beta_{[I_0]}^{\intercal} + \Sigma_{u,[I_0]})^{-1}\beta_{[I_0]}\gamma,$$

where $r_{t,[I_0]}$ denotes the vector of returns of test assets in I_0, and $\beta_{[I_0]}$ is their corresponding beta.[13] It can be shown that (see the proof of a more general setting in Proposition 13 of the appendix)

$$\mathrm{Cov}(g_t, r_{t,[I_0]})\mathrm{Cov}(r_{t,[I_0]})^{-1}\mathrm{E}(r_{t,[I_0]}) = \eta\gamma + O(\left\|\beta_{[I_0]}\right\|^{-2}). \tag{2.5}$$

Since test assets in I_0 feature a strong factor structure, $\left\|\beta_{[I_0]}\right\|^2 \asymp |I_0| =: N_0$,[14] the approximation error is thereby $O(N_0^{-1})$. This result establishes the fact that in population using a smaller number of sufficiently informative assets leads to an asymptotically vanishing error in approximating the risk premium. Moreover, it holds that $N_0/(\left\|\beta_{[I_0]}\right\|^2 T) = O(T^{-1})$, i.e., factors are pervasive within this subset. Therefore, as long as we locate a subset that satisfies the properties of I_0, we can estimate g_t's risk premium consistently with PCA by only using test assets within this subset.

In practice, it is the researcher who decides which test assets to employ in an empirical study. Assuming that a strong factor structure exists at least within a subset of test assets seems practical and plausible. That said, this assumption does rule out the case in which exposures to a factor are uniformly small for all test assets. In this scenario, there is no

13. We use $A_{[I]}$ to denote a submatrix of A whose rows are indexed in I.

14. For an index set $I \subset \langle N \rangle$, we use $|I|$ to denote its cardinality.

guarantee that SPCA can recover this factor, a limitation shared with other estimators.

Unfortunately, we do not know ex-ante such a set, i.e., which assets are informative about the latent factor v_t. Rather than using all assets, the idea of SPCA revolves around selecting the most informative assets based on their covariances with g_t. In the DGP of Proposition 7, the group of assets exhibiting high covariances with g_t comprises those with large magnitudes of βs. Therefore screening via correlation selects a subset of assets satisfying the desirable properties of I_0.

Our proposed screening strategy echoes some of the practice in the empirical asset pricing literature. Very often, test assets are formulated using the exact characteristics-sorted portfolios that the factor of interest is generated from. For instance, Fama and French [1993] use size and value double-sorted portfolios as test assets when estimating a factor model that includes size and value as factors. In other cases, for nontradable factors, portfolios are sorted based on individual stock betas with respect to the factor of interest.

These choices seldom are justified formally, and are often only valid in very special cases. For example, building portfolios by sorting stocks on beta with respect to g_t may inadvertently incorporate compensation for other correlated risks, introducing a bias when omitted factors exist in the asset pricing model that is used to calculate the betas, not to mention the issue of propagation of errors that arise in the estimation of the beta. Similarly, using Fama-French portfolios as test assets assumes implicitly that they span the investment universe. This assumption contradicts the recent asset pricing literature, from which numerous factors or anomalies emerge. While our methodology formalizes the insight behind these traditional procedures, the fundamental motivation behind our approach is precisely to circumvent the adoption of arbitrary priors when selecting assets.

We next formally present our SPCA procedure in the simple one factor setting as discussed in the previous proposition, which helps illustrate the intuition behind our proposal and facilitates the comparison with existing estimators (the next section is devoted to the

general case).

Algorithm 6 (SPCA-based Estimator of Risk Premia for a Single Factor Model ($p = 1$))**.** *The procedure is as follows:*

Inputs: $\bar{R}$ *and* $\bar{G}$*, a* $1 \times T$ *vector.*[15]

S1. Select a subset $\widehat{I} \subset \langle N \rangle$*:* $\widehat{I} = \left\{ i \middle| T^{-1} |\bar{R}_{[i]} \bar{G}^{\intercal}| \geq c_q \right\}$*, where* c_q *is the* $(1-q)$*-quantile of* $\left\{ T^{-1} |\bar{R}_{[i]} \bar{G}^{\intercal}| \right\}_{i \in \langle N \rangle}$*.*

S2. Repeat S1. – S3. of Algorithm 5 with selected return matrix $\bar{R}_{[\widehat{I}]}$ *and* $\bar{G}$*, and* $p = 1$*.*

Outputs: $\widehat{\gamma}_g^{SPCA} := \widehat{\eta}\widehat{\gamma}$*,* $\widehat{V}, \widehat{\eta}$*, and* $\widehat{\gamma}$*.*

SPCA (Algorithm 6) adds the screening step, S1, to the PCA-based risk-premium estimation method of Giglio and Xiu [2021] (Algorithm 5). In this step, out of the N assets available, only a subset $\widehat{I}$ is selected, and the three steps of Algorithm 5 are applied to this subset only.

The selection is operated by computing the absolute value of the covariance between each of the N assets and the factor g_t: ($T^{-1} |\bar{R}_{[i]} \bar{G}^{\intercal}|$ for each asset i). Only those assets for which the magnitude of this covariance is large enough are selected: specifically, the top $q\%$ of them. Therefore, SPCA involves a tuning parameter, q, which plays a crucial role in determining how many assets we use to extract the factor. Note that the fact that $\widehat{I}$ incorporates information from the target, g_t, reflects the distinctive nature of a supervised procedure (from which the name *supervised*-PCA).

We next prove that SPCA is consistent in the presence of weak factors.

Proposition 8. *Suppose that* $\log N/T \to 0$ *and test asset returns follow a single-factor model in the form of* (2.1) *and that* g_t *satisfies* (2.4)*, with* u_t*,* v_t*, and* z_t *i.i.d. normally distributed and independent from each other. The loading matrix* β *satisfies* $\|\beta\|_{\text{MAX}} \lesssim 1$

15. We discuss the case of a multivariate $(d \times T)$ $\bar{G}$ in Section 2.2.2.4.

and there exists a subset $I_0 \subset \langle N \rangle$ *such that* $\left\|\beta_{[I_0]}\right\| \asymp \sqrt{N_0}$ *where* $N_0 = |I_0| \to \infty$. *Then, for any choice of* q *in Algorithm 6 such that* $qN/N_0 \to 0$[16] *and* $qN \to \infty$, *and that* $|\beta|_{\{qN+1\}} \leq (1+\delta)^{-1}|\beta|_{\{qN\}}$[17] *for some* $\delta > 0$, *where* $|\beta|_{\{k\}}$ *denotes the* k*th largest value in* $\left\{|\beta_{[i]}|\right\}_{i\in\langle N\rangle}$, *we have* $\widehat{\gamma}_g^{SPCA} \xrightarrow{\mathrm{P}} \eta\gamma$.

To gain a better understanding of the intuition, let us delve into some key steps of the proof, which is detailed in the appendix. Given a specific choice of the tuning parameter q, we can identify the population counterpart of $\widehat{I}$, denoted as I. This set I consists of the qN largest entries of β in terms of their magnitudes, as specified before Assumption 13 in the appendix.[18] The proof of Proposition 8 establishes the consistency of the selected set $\widehat{I}$ (which contains the top qN test assets with the largest sample covariances with g_t) with respect to I in the following sense: $\mathrm{P}(\widehat{I} = I) \to 1$.

This result is valid for two reasons. Firstly, the estimation error for the (population) covariance with g_t for any test asset is of order $T^{-1/2}$. By applying the large deviation bound in high-dimensional statistics, we can establish that the estimation error for covariances between g_t and all test assets is uniformly bounded by $(\log N)^{1/2}T^{-1/2}$. Consequently, to ensure consistent estimation of all covariances, it is necessary that $\log N/T \to 0$.

Secondly, the condition that there exists I_0 such that $\left\|\beta_{[I_0]}\right\|^2 \asymp N_0$ and $qN/N_0 \to 0$ guarantee the existence of at least qN test assets with non-zero population covariances with g_t. Thus, according to the definition of I, the smallest population covariance with g_t among all test assets in I must be non-zero. This suggests that $\left\|\beta_{[I]}\right\|^2 \asymp |I| = qN$. Furthermore, since we assume a non-vanishing gap between the (qN)th and $(qN+1)$th

16. It may be tempting to use $qN/N_0 \xrightarrow{\mathrm{P}} \text{const} < 1$. However, this is not viable because N_0 and I_0 are not precisely defined in the assumption $\left\|\beta_{[I_0]}\right\| \asymp \sqrt{N_0}$. That is, if we replace N_0 by $N_0/2$, the previous assumption still holds but qN/N_0 might be greater than 1.

17. This technical condition on $|\beta|_{\{qN+1\}}$ simply states that the test assets should have (asymptotically) distinct risk exposure. It is a rather mild assumption that simplifies the proof.

18. It is crucial to distinguish between I and I_0. I is uniquely defined for each q that satisfies the conditions of I_0, whereas I_0 is a general mathematical abstraction not uniquely defined.

population covariances, it thereby follows that the set of test assets with largest population covariances must coincide with those having the largest sample covariances, because the vanishing estimation error is dominated by this non-vanishing gap in the asymptotic context.

Given that the identified set I can function as I_0 (since $\left\|\beta_{[I]}\right\|^2 \asymp |I|$), and as demonstrated in equation (2.5), we can directly approximate the risk premium of g_t using its mimicking portfolio built on this subset I of test assets. The consistency of risk premium estimate thereby follows from the consistency of $\widehat{I}$ in the recovery of I.

Proposition 7 and Propositions 10 - 12 in the appendix show that in the single factor case, the consistency of PCA, Ridge, PLS, and rpPCA requires $B = 0$. Suppose $\|\beta\|^2 = N^v$, for some $v > 0$, then $B = 0$ is equivalent to $N^{1-v}/T \to 0$. The consistency of SPCA, as shown by Proposition 8, nonetheless, only requires $(\log N)/T \to 0$.[19]

2.2.2.3 SPCA in the General Case: Selection and Projection

Propositions 7 - 8 focus on an unrealistic single-factor model since they are meant to illustrate the failure of PCA due to the presence of a weak factor as well as the intuition behind our procedure. In general, the DGP of returns is likely driven by more than one factor; in addition, these factors will generally have different strength in any specific cross-section of test assets. Note also that g_t could have more than one dimension in the general setup (2.4). In this section, we show how to generalize SPCA to the case where multiple factors of distinct strength are present.

To begin with, in the same spirit of Proposition 7, we can show that a general necessary

19. Another idea that shares the spirit of SPCA is the scaled-PCA proposed by Huang et al. [2022], which uses regression coefficients of $\bar{G}$ on $\bar{R}$ to weight $\bar{R}$ before feeding it into the PCA procedure. An advantage of the scaled PCA approach is that it does not involve any tuning parameter. Nonetheless, the scaled PCA still assigns weights of $1/\sqrt{T}$ magnitude to assets that have zero-correlations with the target variable, whereas our approach assigns zero weights to such assets. As a result, our procedure only requires $\log N$ to be small relative to T, whereas both the scaled PCA and PCA require N to grow no faster than a certain polynomial rate relative to T.

condition for the consistency of PCA in a multi-factor model is that

$$N/(\lambda_{\min}(\beta^{\intercal}\beta)T) \to 0. \tag{2.6}$$

If this holds, it means that even the weakest one among all p factors in (2.1) is sufficiently strong that it can be recovered by PCA. Then, the three-pass estimator of Giglio and Xiu [2021] would properly recover risk premium for any factor g_t. We thereby define weak factors as those for which test asset exposures fail condition (2.6). This is a compact formal description of the non-ideal finite-sample environment encountered in practice.[20]

Just like in the single-factor case, in the multi-factor case condition (2.6) can fail if one of the factors is not pervasive. But in the multi-factor case, it can also happen that all factors are individually strong, and condition (2.6) still fails because the factors' exposures are highly correlated. Consider, for example, a two-factor model where the beta matrix has the following form:

$$\beta = \left[\begin{array}{c|c} \beta_{11} & \beta_{12} \\ \hline \\ \beta_{21} & \beta_{22} \\ \\ \end{array}\right], \tag{2.7}$$

where β_{11} and β_{12} are $N_0 \times 1$ vectors, β_{21} and β_{22} are $(N - N_0) \times 1$ vectors, and N_0 is small relative to N. Suppose that $\beta_{21} = \beta_{22}$. In this setup, we can identify two groups of test assets. The first one is a small group of N_0 test assets, with exposures β_{11} to the first factor and β_{12} to the second factor. The second is a large group of $(N - N_0)$ assets, that have the *same* exposure to both factors (since $\beta_{21} = \beta_{22}$). In this case, we can show that

20. Note that r_t is related to g_t through v_t. The loading of g_t on v_t is a low dimensional parameter η specific to each g_t, whereas the loading of r_t on v_t is a high-dimensional vector β independent of g_t. The advantage of formulating the condition in terms of $\lambda_{\min}(\beta^{\intercal}\beta)$ without η guarantees the applicability of our conclusion across all factors of interest.

condition (2.6) can fail: even if each factor is strong individually, there is a "rank deficiency" issue in the betas. The reason is that most of the asset (group 2) do not contain information that can separate the risk premia of the two factors, because they are equally exposed to the two. This loss of information turns out to have exactly the same effect on estimation and inference as the weak factor issue.[21] We need a procedure that consistently estimates risk premia in this case as well.

It is also important to note that in the general case with multiple factors of potentially different strength, a simple extension of Algorithm 6, operating an initial screening (S1) and then extracting *multiple* factors via PCA (S2) would *not* actually work to recover all factors. To see this, take (2.7) again as an example. Suppose now that $\beta_{21} \neq \beta_{22}$, but $\beta_{22} = 0$: that is, most of the assets have zero exposure to the second factor. Consequently, the first factor is strong, while the second factor is weak.[22] Now suppose that $\eta = (1, 1)$, implying that the observed factor g is correlated with both factors and, by extension, with all the test assets. In this scenario, the determination of which assets to exclude via screening hinges on the betas of these test assets. Should a majority of the selected assets pertain to the second group, the subsequent application of PCA in step S2 would only recover the first factor. This would occur if condition (2.6) fails for the selected assets. On the other hand, if many of the selected assets belong to the first group, PCA applied to them has the potential to recover both factors. In this scenario, the first principal component may capture a linear combination of both the strong and weak factors. This example demonstrates that even though screening assets ensures that the *first* principal component after screening recovers one factor (which could be the strong factor, the weak factor or their mixture on the basis of the original cross-section), there is no guarantee that this procedure can solve the weak

21. Formally, we can show that $\lambda_{\min}(\beta^\intercal\beta) \leq \|\beta_{11} - \beta_{12}\|^2 /2 \lesssim N_0$. As a result, $N/(\lambda_{\min}(\beta^\intercal\beta)T) \gtrsim N/(N_0T)$, which does not necessarily converge to 0 if N_0 and T are small, so that the condition (2.6) could fail.

22. It is easy to show that in this case $\lambda_{\min}(\beta^\intercal\beta) \leq \|\beta_{12}\|^2 \lesssim N_0$.

factor issue in one shot.

Next we provide another example, that shows that in some situations screening can sometimes eliminate *too many* assets, making a strong factor model become weak or even rank-deficient. Suppose that β has the following form:

$$\beta = \left[\begin{array}{c|c} \beta_{11} & \beta_{11} \\ \hline 0 & \beta_{22} \end{array}\right], \tag{2.8}$$

where β_{11} and β_{22} are $N/2 \times 1$ non-zero vectors satisfying $\|\beta_{11}\| \asymp \|\beta_{22}\| \asymp \sqrt{N}$. Clearly, β is full-rank and both factors are strong. Therefore, a standard PCA procedure should work smoothly. Suppose in addition that $\eta = (1, 0)$ (i.e., $g_t = v_{1t}$) and that v_{1t} and v_{2t} are uncorrelated. Then it implies that g_t is uncorrelated with the second half of test assets in r_t, so only those test assets within the first half would remain, should screening be applied with g_t before extracting the principal components. In this example, however, the *remaining* test assets have perfectly correlated exposures to both factors, so that effectively only *one* factor, $v_{1t} + v_{2t}$, is left. This example shows once again that the one-step supervised procedure (screening once and then applying PCA) may fail at extracting all factors in a multi-factor setting.[23]

To address the aforementioned issues, we propose a multi-step version of SPCA, that iteratively conducts selection and projection. Step S1 of Algorithm 6 described above – valid when there is only one factor – can help identify one strong factor from a selected subset of test assets. In a nutshell, the multi-step SPCA, described below in Algorithm 7 iteratively applies Algorithm 6 to extract a new factor, with a projection step designed to

23. This one-step procedure was originally called Supervised PCA, as proposed by Bair et al. [2006] in the context of prediction. We propose below an iterative version that can cope with a general multi-factor model. We still use the term Supervised PCA for this iterative procedure.

ensure that each new factor is orthogonal to the estimated factors in the previous steps, similar to the factors extracted by the standard PCA.

Formally, the algorithm is given by:

Algorithm 7 (Selection and Projection)**.** *The iterative SPCA procedure for risk premia estimation is as follows:*

Inputs: $\bar{R}_{(1)} := \bar{R}$, $\bar{r}_{(1)} := \bar{r}$, *and* $\bar{G}_{(1)} := \bar{G}$, *a* $d \times T$ *vector.*

S1. For $k = 1, 2, \ldots$ *iterate the following steps using* $\bar{R}_{(k)}$, $\bar{r}_{(k)}$, *and* $\bar{G}_{(k)}$*:*

a. Select an appropriate subset $\widehat{I}_k \subset \langle N \rangle$.

b. Repeat S1. – S3. of Algorithm 5 with selected return matrix $\left(\bar{R}_{(k)}\right)_{[\widehat{I}_k]}$ *and* $\bar{G}_{(k)}$ *to extract only the first principle component. Denote the estimates as* $\widehat{V}_{(k)}$, $\widehat{\eta}_{(k)}$, $\widehat{\gamma}_{(k)}$.

c. Estimate the exposure of $\bar{R}_{(k)}$ *to* $\widehat{V}_{(k)}$ *by* $\widehat{\beta}_{(k)} = T^{-1}\bar{R}_{(k)}\widehat{V}_{(k)}^{\intercal}$.

d. Obtain $\bar{R}_{(k+1)} = \bar{R}_{(k)} - \widehat{\beta}_{(k)}\widehat{V}_{(k)}$, $\bar{r}_{(k+1)} = \bar{r}_{(k)} - \widehat{\beta}_{(k)}\widehat{\gamma}_{(k)}$, *and* $\bar{G}_{(k+1)} = \bar{G}_{(k)} - \widehat{\eta}_{(k)}\widehat{V}_{(k)}$.

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

S2. Estimate risk premia by $\widehat{\gamma}_g^{SPCA} = \sum_{k=1}^{\widehat{p}} \widehat{\eta}_{(k)}\widehat{\gamma}_{(k)}$.

Outputs: $\widehat{\gamma}_g^{SPCA}$, $\widehat{\eta} = (\widehat{\eta}_{(1)}^{\intercal}, \cdots, \widehat{\eta}_{(\widehat{p})}^{\intercal})^{\intercal}$, $\widehat{\gamma} = (\widehat{\gamma}_{(1)}, \cdots, \widehat{\gamma}_{(\widehat{p})})^{\intercal}$, $\widehat{V} = (\widehat{V}_{(1)}^{\intercal}, \cdots, \widehat{V}_{(\widehat{p})}^{\intercal})^{\intercal}$ *and* $\widehat{\beta} = (\widehat{\beta}_{(1)}, \cdots, \widehat{\beta}_{(\widehat{p})})$.

Each iteration k of the procedure recovers one latent factor $\widehat{V}_{(k)}$, estimates its risk premium $\widehat{\gamma}_{(k)}$, and the exposure of g_t to that factor, $\widehat{\eta}_{(k)}$. In step S1, there is first asset selection (S1.a). Next, the three-step estimator of risk premia of Giglio and Xiu [2021] is applied using the selected assets (S1.b) to recover the kth factor $\widehat{V}_{(k)}$ in addition to $\widehat{\gamma}_{(k)}$ and $\widehat{\eta}_{(k)}$, which are specific to that factor. Then, in S1.c, we project the returns of all assets (not just those

selected) on the estimated factor $\widehat{V}_{(k)}$, and in step S1.d we compute the residuals of this projection for returns and the factor g_t itself. Therefore, at the end of step S1, we have completely eliminated the effect of the kth factor on returns and the target factor g_t. We then repeat S1 again, this time using the residuals of returns and g_t, looking for the next factor. Iteration continues for $\widehat{p}$ steps. At the end, step S2 combines the $\widehat{\gamma}_{(k)}$ and the $\widehat{\eta}_{(k)}$ obtained at each step into an estimator $\widehat{\gamma}_g^{SPCA}$ for the risk premia of g_t.

Algorithm 7 requires an appropriate choice of $\widehat{I}_k$ and a stopping rule. One choice for $\widehat{I}_k$ is:[24]

$$\widehat{I}_k = \left\{ i \middle| T^{-1} \left\| (\bar{R}_{(k)})_{[i]} \bar{G}_{(k)}^{\intercal} \right\|_{\text{MAX}} \geq c_q^{(k)} \right\},$$
$$\text{where } c_q^{(k)} \text{ is the } (1-q)th\text{-quantile of } \left\{ T^{-1} \left\| (\bar{R}_{(k)})_{[i]} \bar{G}_{(k)}^{\intercal} \right\|_{\text{MAX}} \right\}_{i \in \langle N \rangle}. \tag{2.9}$$

Correspondingly, we set the stopping criterion as:

$$c_q^{(k)} < c, \quad \text{for some threshold } c. \tag{2.10}$$

In other words, we select test assets that have predictive power for at least one variable in g_t and stop when most test assets are uncorrelated with all variables in g_t. With a good choice of tuning parameters, q and c, the iteration stops as soon as most projected residuals of returns appear uncorrelated with the projected residuals of g_t, which implies that all factors that are correlated with g_t are successfully recovered.

It is helpful to revisit the aforementioned examples and understand how the new procedure fixes issues with the one-step SPCA. Recall that in example (2.7), $\beta_{22} = 0$ and

24. Using covariance for screening allows us to replace all $\bar{G}_{(k)}$ in the definition of $\widehat{I}_k$ and Algorithm 7 by $\bar{G}$, that is, only the projections of $\bar{R}_{(k)}$ and $\bar{r}_{(k)}$ are needed, because this replacement would not affect the covariance between $\bar{G}_{(k)}$ and $\bar{R}_{(k)}$, and in turn, the test assets after screening and the estimates of $\widehat{\eta}_{(k)}$. We use this fact in the proofs, which simplifies the notation. We can also use correlation instead of covariance in constructing $\widehat{I}_k$. While this does not affect the asymptotic analysis, we find correlation screening performs slightly better in finite samples.

$g_t = v_{1t} + v_{2t}$. As discussed previously, screening will select a subset of q assets that are spread across both groups of assets since they are all correlated with g_t. Consequently, applying PCA to them will identify a factor that is in general spanned by v_{1t} and v_{2t}. Even if this first step only recovers the strong factor v_{1t}, once we project r_t and g_t onto this factor following Algorithm 6, both residuals should only depend on v_{2t}. Subsequently, applying screening again to these residuals will leave us with only the test assets within the first group of assets, to which applying PCA can recover v_{2t}. In cases where a linear combination of v_{1t} and v_{2t} are recovered in the first step, after projection the residuals feature a strong factor (again a linear combination of v_{1t} and v_{2t} but orthogonal to the first linear combination), since the second group of $N - N_0$ assets have exposure to it. Therefore, a subsequent screening and PCA suffice to recover this factor.

Similarly in example (2.8), the second half of the assets will be eliminated in the first step when using $g_t = v_{1t}$ to screen test assets. The returns for the remaining (first half) assets load on $v_{1t} + v_{2t}$ with a common loading matrix β_{11}. Applying PCA to these assets thereby finds $(v_{1t} + v_{2t})/\sqrt{2}$ as the first factor (up to a sign, assuming v_{1t} and v_{2t} share the same variance). Following Algorithm 6, we then obtain residuals from projections of r_t and g_t onto this factor. It is easy to see that the residuals of the second half of r_t and the residuals of g_t both load on a single strong factor $(v_{1t} - v_{2t})/\sqrt{2}$ yet the first half of the residuals are purely idiosyncratic. Applying screening plus PCA will successfully recover this factor, and hence the span of the factor space.

To formally establish the consistency of this estimator, we introduce an assumption akin to the single factor case. Specifically, we require that a subset of assets, indexed by I_0, satisfies that all factors are strong within this subset. In other words, $\lambda_{\min}(\beta_{[I_0]}^{\intercal}\beta_{[I_0]}) \asymp N_0$, where $N_0 = |I_0| \to \infty$. Because the number of factors, p, is finite, such a subset I_0 always exists as long as for each factor we can locate a sufficiently large subset, respectively, within

which this factor can be extracted consistently.[25] Proposition 13 of the appendix establishes that test assets in such a subset suffice to serve as basis assets, building on which a mimicking portfolio can approximate the risk premia of any observable factor. With this identification assumption, along with moment conditions given in the appendix, the following theorem establishes the consistency of the SPCA estimator:

Theorem 6. *Suppose that test asset returns in r_t follow* (2.1)*, the factor proxies in g_t satisfy* (2.4)*, and that Assumptions 7-14 hold. If* $\log(NT)(N_0^{-1} + T^{-1}) \to 0$ *then for any tuning parameters c and q that satisfy*

$$c \to 0, \quad c^{-1}(\log NT)^{1/2}(q^{-1/2}N^{-1/2} + T^{-1/2}) \to 0, \quad qN/N_0 \to 0, \tag{2.11}$$

we have $\widehat{\gamma}_g^{SPCA} \xrightarrow{\text{P}} \eta\gamma$.

The screening step in Algorithm 7 ensures that the selected test assets or their residuals must encompass one strong factor, as they have high correlations with g_t. As the SPCA procedure unfolds, each iteration selects a distinct subset of test assets. By amalgamating all such subsets, we obtain a subset of assets within which all factors are potentially strong, given that the number of factors is finite. However, this procedure may not recover all factors that drive returns. The number of factors that SPCA can recover depends on the interplay between η and β as well as the tuning parameters in a complex manner.[26] Some of the factors that SPCA omits might even be strong! Intuitively, only factors correlated with g_t are guaranteed to be recovered. This is the trade-off that arises for using g_t as a supervisory signal.[27] Nonetheless, missing any factors in the SDF that are uncorrelated with g_t does

25. This assumption is weak in that it does not imply all factors should have identical strength with respect to the entire cross-section of assets in r_t. In addition, different groups of assets could be exposed to different factors.

26. We explicitly characterize this number, denoted by $\tilde{p}$, given in the appendix following Assumption 13.

27. In the context of forecasting, Giglio et al. [2023] provide convergence rate of the estimated factor space, spanned by the factors that are correlated with the variables used for supervision in a similar SPCA procedure.

not affect the consistency of the estimate of the risk premia of g_t. This holds true because such factors do not help price g_t. Of course, this result will need to be strengthened if the objective is to recover the entire SDF, a problem we tackle in Section 2.2.3.

The consistency result in Theorem 6 does not rely on Gaussian error assumptions nor on an assumption that all factors have the same strength with respect to all test assets. The assumption on the relative size of N and T is also quite flexible, in contrast with existing results on factor models in the literature, where N cannot grow at a rate exceeding a certain polynomial function of T.

2.2.2.4 Asymptotic Inference on Risk Premia

In this section we develop the asymptotic distribution of the risk premia estimator from Algorithm 7. Naturally, deriving asymptotic inference requires stronger assumptions than those required for consistency discussed above. To consistently estimate the risk premia of g_t, one only needs recover factors that are correlated with g_t. Nonetheless, if SPCA misses factors that are in the SDF but are not correlated with g_t, consistency is maintained, but inference is undermined, because the omitted factors may contribute a higher-order error that invalidates the central limit result.

More specifically, the conditions in Theorem 6 do not guarantee that $\widehat{\gamma}_g^{SPCA}$ converges to $\eta\gamma$ at the desirable rate $T^{-1/2}$. The major obstacle lies in the recovery of factors not strongly correlated with g_t, which we can explain with the previous single-factor example.

Recall that we use the sample correlation/covariance between r_t and g_t to screen test assets. Condition (2.11) necessitates two key considerations: First, it requires that $c \to 0$, allowing the iteration procedure to continue until the selected r_t exhibit asymptotically diminishing correlations with g_t. Simultaneously, it demands that $c\sqrt{T} \to \infty$ and $c\sqrt{qN} \to \infty$. In other words, c must be sufficiently large to supersede the estimation error in covariance

estimates during the screening step, which is of order $T^{-1/2}$,[28] and to dominate error in the construction of residuals in the projection step when multiple steps are involved, an error of order $T^{-1/2}+(qN)^{-1/2}$. However, for any given threshold, say, $c = T^{-1/4}$, if it happens that $\eta \asymp T^{-1/3} < T^{-1/4}$, then screening based on g_ts' correlation with r_t will likely not select any assets, which in turn leads to the termination of Algorithm 7 and no discovery of factors. Our procedure thereby gives a risk premium estimate of 0, which is certainly consistent, but the estimation error is of an order $T^{-1/3} > T^{-1/2}$, so that the usual central limit theorem (CLT) fails. In general, this problem arises due to the possibility of not identifying all factors in the DGP. Once all factors are recovered, the CLT holds regardless of the magnitude of η. To make correct inference, we thus need a stronger assumption that eliminates scenarios like this.

It appears that if $\eta \in \mathbb{R}^{d\times p}$ meets the condition $\lambda_{\min}(\eta^\intercal\eta) \gtrsim 1$, we can rule out the possibility of missing factors. This condition necessitates that each latent factor maintains a correlation with at least one of the observable variables within g_t. Consequently, this implies that d must be greater than or equal to p, meaning we require g_t to possess at least the same number of variables as the true number of factors. Meanwhile, our algorithm will not select more factors than needed, as we stop the iteration as soon as $c_q^{(k)}$ is sufficiently small (below c), at which points no common factors are left in the residuals of g_t and r_t. We thus obtain the consistency result on the number of factors, which in turn leads to the CLT result on risk premia. Formally, we have

Theorem 7. *Under the same assumptions as Theorem 6, if we further have $T^{-1/2}N_0 \to \infty$, Assumption 15 and $\lambda_{\min}(\eta^\intercal\eta) \gtrsim 1$, then for any tuning parameters c and q in* (2.9) *and* (2.10) *satisfying*

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

28. Even if g_t is uncorrelated with the test assets, their sample covariances can be as large as $T^{-1/2}$.

we have that $\widehat{p}$ *defined in Algorithm 7 satisfies* $\mathrm{P}(\widehat{p}=p)\to 1$*, and that the estimator constructed via Algorithm 7 satisfies* $\widehat{\gamma}_g^{SPCA}-\eta\gamma=O_{\mathrm{P}}(T^{-1/2})+O_{\mathrm{P}}(q^{-1}N^{-1})$*. Furthermore, we obtain a CLT:*

$$\sqrt{T}\left(\widehat{\gamma}_g^{SPCA}-\eta\gamma\right)\xrightarrow{d}\mathcal{N}\left(0,\Phi\right),$$

where Φ *is given by*

$$\Phi=\left(\gamma^{\intercal}\Sigma_v^{-1}\otimes\mathbb{I}_d\right)\Pi_{11}\left(\Sigma_v^{-1}\gamma\otimes\mathbb{I}_d\right)+\left(\gamma^{\intercal}\Sigma_v^{-1}\otimes\mathbb{I}_d\right)\Pi_{12}\eta^{\intercal}+\eta\Pi_{12}^{\intercal}\left(\Sigma_v^{-1}\gamma\otimes\mathbb{I}_d\right)+\eta\Pi_{22}\eta^{\intercal},$$

and Π_{11}*,* Π_{12}*, and* Π_{22} *are* $dp\times dp$*,* $dp\times p$*, and* $p\times p$ *matrices, respectively, defined as:*

$$\begin{aligned}
\Pi_{11}&=\lim_{T\to\infty}\frac{1}{T}\mathrm{E}\left(\mathrm{vec}(ZV^{\intercal})\mathrm{vec}(ZV^{\intercal})^{\intercal}\right),\\
\Pi_{12}&=\lim_{T\to\infty}\frac{1}{T}\mathrm{E}\left(\mathrm{vec}(ZV^{\intercal})\iota_T^{\intercal}V^{\intercal}\right),\\
\Pi_{22}&=\lim_{T\to\infty}\frac{1}{T}\mathrm{E}\left(V\iota_T\iota_T^{\intercal}V^{\intercal}\right).
\end{aligned}$$

In regard to our theoretical findings, several key points merit attention. Firstly, Theorem 7 hinges on the existence of a tuning parameter, q, which must satisfy two conditions: $q^{-1}N^{-1}T^{1/2}\to 0$ and $qN/N_0\to 0$. A necessary condition for the existence of such a q is thus $T^{1/2}/N_0\to 0$.

Secondly, the estimation error of $\widehat{\gamma}_g^{SPCA}-\eta\gamma$ consists of two components. A portion of this error stems from the error accumulation at each step of the iteration in Algorithm 7. This accumulated error is compounded in each step k at most by a factor of $\sqrt{|\widehat{I}_k|/\widehat{\lambda}_{(k)}}$, where $\widehat{\lambda}_{(k)}=\left\|\left(\bar{R}_{(k)}\right)_{[\widehat{I}_k]}\right\|^2/T$. Importantly, the assumption that there exists a subset within which factors are pervasive ensures that $\widehat{\lambda}_{(k)}\asymp_{\mathrm{P}}qN=|\widehat{I}_k|$, implying that the accumulated error is only magnified by a constant factor with each iteration of SPCA. Ultimately, our proof establishes that this iterative process results in an overall estimation error in risk premia

estimates, that is of the order $O_{\mathrm{P}}(T^{-1/2} + q^{-1}N^{-1})$. The condition $q^{-1}N^{-1}T^{1/2} \to 0$ thus guarantees that the $O_{\mathrm{P}}(q^{-1}N^{-1})$ term does not influence the asymptotic distribution. The derivation of the error rate for an iterative procedure is non-trivial, constituting our primary contribution to the econometric literature on factor models.

Thirdly, the estimation error of the factor loading has no impact on the asymptotic variance of risk premia, as the expression of Φ demonstrates. This stands in contrast to the classical Fama-MacBeth regression setting, where Shanken's adjustment term (Shanken [1992]) is crucial. This difference is due to the fact that when dealing with a large cross-sectional dimension ($N \to \infty$), this adjustment term vanishes asymptotically.[29] To make inference feasible, we implement the same Newey-West-type estimator for Φ as in Section 4.5 of Giglio and Xiu [2021], since each component of Φ can be estimated from the outputs of the SPCA algorithm. These estimates are consistent up to some rotation matrices which will cancel each other and yield a consistent estimate of Φ.

Fourthly, Theorem 7 suggests that, with probability approaching one, we can expect a perfect recovery of the number of factors p. Yet, in any finite sample, perfect recovery remains challenging. Notably, the assumptions made here are considerably less stringent compared to the prevalent factor assumptions found in the literature, see, e.g., Bai [2003] and Bai and Ng [2002]. In these previous studies, inference theory for factor models also relies on the perfect recovery of the count of (strong) factors. We explore the finite sample behavior of SPCA through simulations in Section 2.3.

Lastly, in the special case when the returns of test assets are *exclusively* driven by strong factors, SPCA is asymptotically equivalent to PCA, contingent upon the appropriate selection of the tuning parameters c and q. Otherwise, SPCA is less efficient – either due to an excessively small choice of q to the extent that the $O_{\mathrm{P}}(q^{-1}N^{-1})$ term plays a dominant role in the estimation error in finite sample (note that PCA corresponds to the case of $q = 1$), or

29. For a more detailed discussion on this point, please refer to equation 45 of Giglio et al. [2022], and the discussion that follows it.

to the fact that some factors (specifically those uncorrelated with g_t) may not be recovered by SPCA. The former loss of efficiency can be mitigated through careful tuning parameter selection; the latter typically hinges on the unknown values of β and η, which can be resolved with a multivariate target satisfying $\lambda_{\min}(\eta^\intercal \eta) \gtrsim 1$.

2.2.2.5 Tuning Parameter Selection

While the enhanced robustness to weak factors is an advantage, it comes at the expense of introducing an additional tuning parameter. To employ the SPCA estimator, we need make choices regarding two tuning parameters: q and c. The parameter q governs the subset size employed in PCA construction, while c determines the stopping rule and consequently the number of factors, p. In contrast, PCA (and other estimators like PLS) essentially require the selection of only p. We have established in Theorem 7 that we can consistently recover p, as long as certain conditions are met by q and c.

In theory, the textbook approach to choosing a tuning parameter for parameter estimation revolves around the analytical minimization of the root-mean-squared error (RMSE) of the estimator.[30] This approach effectively balances the trade-off between bias and variance inherent in the estimation. Regrettably, this method necessitates intricate finite sample analytical calculations of the RMSE, often relying on strong assumptions regarding the DGP. In our context, assumptions of normal distribution for returns and certain distributional properties and sparsity conditions for betas are likely necessary. Complicating matters further, our iterative SPCA procedure compounds the difficulty of this analysis, rendering it practically infeasible. Additionally, this RMSE-based criterion primarily hinges on statistical considerations, lacking economic relevance.

In lieu of this, we instead opt for the utilization of the R^2 of the hedging portfolio for g_t

30. Note that in the realm of machine learning, the prevailing approach involves leaning on the prediction RMSE derived from a validation sample, where the actual values of the prediction target are available. This stands in contrast to the estimation problem, where the true values are never known.

built by SPCA as a criterion, which is both simpler to apply and justified from an economic perspective. Recall that any estimator of risk premia for a nontradable factor explicitly or implicitly builds a hedging portfolio, or a factor-mimicking portfolio, for g_t, and computes the risk premium as the average excess return of that portfolio. The empirical R^2 obtained by different estimators then has an economic meaning: it reveals the hedging efficacy of the factor mimicking portfolios constructed (explicitly or implicitly) by any risk premia estimators.[31]

Beyond the economic motivation, the R^2 is a useful criterion from a statistical perspective, because attaining an optimal R^2 in a validation sample stands as a sufficient condition for valid selection of tuning parameters, which in turn guarantees consistency of risk premia estimates, see Proposition 14 in the appendix for a rigorous statement.

Furthermore, in practice we can consider directly tuning the parameter p instead of c, as it offers greater interpretability, restricts itself to integer values, and is well-informed by the scree plot, providing insights into reasonable ranges for p. Regarding the parameter q, opting for larger values makes SPCA's performance resemble that of PCA, thus reducing its robustness against weak factors. Conversely, smaller values of q raise the risk of overfitting, resulting in a high in-sample R^2 but a low out-of-sample one. We suggest tuning $\lfloor qN \rfloor$ instead of q, because the former can only take integer values, and that multiple choices of the latter may lead to the same integer values of the former.

In our applications, we select tuning parameters based on cross-validation (CV) in a training sample, that proceeds as follows. We split the sample into three folds. We then use each of the three folds, in turn, for validation while the other two are used for training. We select the optimal tuning parameters according to the average time series R^2 in the validation folds.

31. To be clear, while comparing R^2s provides an insightful depiction of the empirical performance of the hedging portfolios, this cannot be interpreted as proof of the superiority of one estimator over another (which is instead established based on the theoretical properties, like consistency and efficiency, discussed in the previous sections).

2.2.3 Recovery of the Stochastic Discount Factor

The estimation of risk premia for observable factors g_t, studied in Section 2.2.2, is a natural application of the supervised PCA approach, since g_t can be used to supervise the latent factor extraction. In this section we explore another application in which observable factors help extract latent factors: a *diagnostic procedure* for observable factor models.

The asset pricing literature has proposed a variety of models composed of a small number of tradable factors g_t: the CAPM, the Fama-French 3 or 5 factor models, etc. These models are typically evaluated by computing the alphas of a universe of test assets, and testing whether these alphas are different from zero. This is clearly a valid test for a model, but it gives only limited insights about the reason why the model is (as is often the case) rejected statistically. Specifically, it does not clarify if the model's failure is due to the presence of true alphas or the omission of priced factors. Our SPCA procedure helps shed light on this by recovering strong and weak latent factors that drive the cross-section of returns, and evaluating whether those factors are indeed spanned by the observable factor model g_t. This helps ascertain whether the model is lacking certain factors.

A last point relates to the universe of test assets. The asset pricing literature (e.g. Lewellen et al. [2010]) has emphasized that using a large cross-section of test assets is important for evaluating asset pricing model, as it can improve the power of the tests. There is, however, a downside in expanding the set of test assets: the possibility that many of the added assets may have little exposure to some factors, introducing a weak factor problem. The ability of SPCA to handle weak factors also frees the researcher from worrying about adding assets to the universe, not only in risk premia estimation, but also in performing diagnostic tests like the one we explore in this section.

2.2.3.1 Consistency of the SDF Estimator

We first prove that, under certain conditions, SPCA does consistently recover the SDF even in the presence of weak factors. Using the outputs of Algorithm 7, we can estimate the SDF as:

$$\widehat{m}_t^{SPCA} = 1 - \widehat{\gamma}^\intercal \widehat{v}_t, \quad \text{where } \widehat{v}_1, \cdots, \widehat{v}_T \text{ are the columns of } \widehat{V}. \tag{2.12}$$

In the appendix, we prove the following theorem, which not only shows the consistency of the recovery of the SDF, but also derives the rate at which the recovery occurs.

Theorem 8. *Suppose the same assumptions as in Theorem 7 hold. In addition, we have Assumption 16. Then the estimator* (2.12) *satisfies*

$$\frac{1}{T}\sum_{t=1}^{T} |\widehat{m}_t^{SPCA} - m_t|^2 \lesssim_{\mathrm{P}} \frac{1}{T} + \frac{\log N_0}{N_0}. \tag{2.13}$$

The theorem shows that consistent estimation of the entire SDF time-series is possible in terms of average ℓ_2-distance, but under specific conditions. Firstly, for every weak latent factor in v_t, there must be a sufficiently large subset of assets with exposure to that factor. This condition, reflected in the requirement of a large N_0, is also necessary for the consistent estimation of risk premia.

In addition, for each latent factor in v_t, there must be at least one observable factor in g_t that is correlated with that latent factor. This second assumption is not only needed for asymptotic inference on risk premia but also for SDF recovery here. In cases where g_t does not correlate with a latent factor, that latent factor can potentially be missed by SPCA, thereby hindering SDF recovery.

2.2.3.2 Comparison with Alternative Procedures of SDF Estimation

There are a number of alternative approaches for SDF estimation with latent factors proposed in the literature, e.g., the selection/shrinkage approach by Kozak et al. [2020] and the risk premia PCA by Lettau and Pelger [2020]. In what follows, we provide a theoretical comparison of Lasso- and Ridge-based estimators in our general framework where factors can potentially be weak. The ridge estimator shares the spirit of PCA-based estimators as shown by Giglio and Xiu [2021] and propositions in previous sections. Examining the asymptotic behavior of these two approaches provides useful insights that may guide their applications in practice. Developing the asymptotic guarantee of these estimators is yet another contribution we make to the existing literature on SDF recovery.

Kozak et al. [2020] consider an SDF in the form of (2.3), whereas we represent it as in (2.2). Prior to the asymptotic analysis of their estimators, we first establish the asymptotic equivalence of these two definitions in our large-N setting:

Proposition 9. *Suppose that test asset returns in r_t follow* (2.1)*, and Assumption 16 holds. Then as $N \to \infty$, we have*

$$\frac{1}{T}\sum_{t=1}^{T}|m_t - \widetilde{m}_t|^2 \lesssim_{\mathrm{P}} \frac{1}{\lambda_{\min}(\beta^{\intercal}\beta)}.$$

Effectively, Proposition 9 proves that there is no ambiguity with respect to the definition of the estimand, since the two estimands are asymptotically equivalent as long as $\lambda_{\min}(\beta^{\intercal}\beta) \to \infty$. Given that this exact assumption is necessary for Theorem 8, and that $\lambda_{\min}(\beta^{\intercal}\beta) \gtrsim N_0$, we can replace m_t in the left-hand side of (2.13) by $\widetilde{m}_t$.

Kozak et al. [2020] suggest estimating the SDF by solving an optimization problem:

$$\widehat{b} = \arg\min_{b}\left\{(\bar{r} - \widehat{\Sigma}b)^{\intercal}\widehat{\Sigma}^{-1}(\bar{r} - \widehat{\Sigma}b) + p_{\mu}(b)\right\}, \tag{2.14}$$

with which the estimated SDF is given by

$$\widehat{m}_t = 1 - \widehat{b}^{\intercal}(r_t - \bar{r}). \tag{2.15}$$

In the above, $\widehat{\Sigma}$ is the sample covariance matrix of r_t and $p_\mu(b)$ is a penalty term through which economic priors are imposed. Depending on the penalty function, we will denote the resulting estimator of m by $\widehat{m}_t^{Ridge}$ or $\widehat{m}_t^{Lasso}$.

The objective function in (2.14) appears to require the inverse of $\widehat{\Sigma}$, which is not well-defined when $N > T$. Instead, we suggest optimizing an equivalent but different form of (2.14):

$$\widehat{b} = \arg\min_b \left\{ b^{\intercal}\widehat{\Sigma}b - 2b^{\intercal}\bar{r} + b^{\intercal}\widehat{\Sigma}b + p_\mu(b) \right\}, \tag{2.16}$$

which avoids the calculation of $\widehat{\Sigma}^{-1}$.

The following result sheds light on the asymptotic properties of this estimator in the cases of $p_\mu(b) = \mu \|b\|_1$ and $p_\mu(b) = \mu \|b\|^2$, respectively.[32]

Theorem 9. *We investigate two distinct scenarios.*

(a) Suppose that r_t is driven by p latent factors as in (2.1). With $p_\mu(b) = \mu \|b\|^2$, if $(N+T)/(\lambda_p T) \to 0$ and Assumptions 10-13, 16-18 hold, we have

$$\frac{1}{T}\sum_{t=1}^{T} |\widehat{m}_t^{Ridge} - m_t|^2 \lesssim_{\mathrm{P}} \frac{1}{T} + \frac{N+T}{\lambda_p T},$$

where λ_p is the p-th largest eigenvalue of $\beta\Sigma_v\beta^{\intercal}$. Since $\lambda_p \asymp \lambda_{\min}(\beta^{\intercal}\beta)$, we can replace m_t in the above equation by $\widetilde{m}_t$.

(b) Suppose that the true SDF satisfies $\mathrm{E}(\widetilde{m}_t^2) \lesssim 1$. With $p_\mu(b) = \mu \|b\|_1$, if Assumptions

32. We use $\|\|_0$, $\|\|_1$, and $\|\|$ to denote the ℓ_0-, ℓ_1-, and ℓ_2-norms of a vector, respectively.

16, 17 hold, we have

$$\frac{1}{T}\sum_{t=1}^{T}|\widehat{m}_t^{Lasso}-\widetilde{m}_t|^2 \lesssim_{\mathrm{P}} \|b\|_1\sqrt{\frac{\log N}{T}}. \tag{2.17}$$

If, in addition, it holds that $\lambda_{\min}(\Sigma) \gtrsim 1$*, and* $\|b\|_0^2 \log N/T \to 0$*, then we have a stronger result*

$$\frac{1}{T}\sum_{t=1}^{T}|\widehat{m}_t^{Lasso}-\widetilde{m}_t|^2 \lesssim_{\mathrm{P}} \|b\|_0\frac{\log N}{T}. \tag{2.18}$$

Interestingly, both the Ridge and Lasso approaches deliver consistent estimates of the SDF, albeit under distinct sets of assumptions.

In the case of Ridge, its convergence rate hinges significantly on the strength of the weakest factor. If condition (2.6) is not met, the SDF consistency is compromised. The failure of this condition is a clear symptom of weak factors, precisely the scenario for which our SPCA estimator is designed.

In contrast, the Lasso approach replaces the explicit factor model assumption on r_t with a sparsity assumption on the vector b. This sparsity assumption dictates that the SDF should be represented as a sparse linear combination of the test assets but imposes no explicit assumptions on the DGP of these test assets. This implies that the Lasso estimator remains consistent regardless of the strength of the factors but converges at a rather slow rate, as indicated in (2.17), which is $\|b\|_1\sqrt{\log N/T}$. Consequently, it is not as efficient as our SPCA estimator, which leverages the factor structure to achieve faster convergence. Nevertheless, under a much stronger sparsity assumption where $\|b\|_0^2 \log N/T \to 0$, the Lasso estimator can attain a comparable convergence rate to that of the SPCA. This more stringent notion of sparsity essentially asserts that the set of true factors must be part of the test assets. In contrast, our SPCA estimator allows for the presence of idiosyncratic components in any of

the test assets, enhancing its practicality in real-world applications.

We can adapt any SDF estimator to obtain an estimator of risk premia, because of the relationship $-\text{Cov}(m_t, g_t) = \eta\gamma$. In light of this, we have a Lasso-based risk premia estimator:[33]

$$\widehat{\gamma}_g^{Lasso} = -\frac{1}{T}\sum_{t=1}^{T} \widehat{m}_t^{Lasso} \times (g_t - \bar{g}).$$

Furthermore, the consistency of the SDF estimator translates to the consistency of the resulting risk premia estimator.[34] Deriving a valid inference procedure is possible for Lasso-based risk premia estimator, if we employ an additional de-biasing step, see, Feng et al. [2020], which is beyond the scope of the current paper.

2.2.3.3 Diagnosis of SDF Models using Sharpe Ratios

We now discuss the diagnosis of SDF models that consist of tradable factors exclusively. Recall that the projection of the SDF on the space of returns achieves the highest possible Sharpe ratio. Given that the factors recovered by SPCA are themselves portfolios, as long as SPCA recovers the entire SDF these factors should achieve the maximal Sharpe ratio. We can then diagnose a model g_t by comparing its Sharpe ratio with that achieved by the estimated SDF supervised by g_t. If g_t contains all the factors that drive the SDF, then the maximal Sharpe ratio achieved by factors in g_t should be on par with the Sharpe ratio of the SDF. Otherwise, if g_t achieves a lower Sharpe ratio, it is a sign that g_t is missing some factors; if g_t's Sharpe ratio is higher than that achieved by SPCA, it indicates that g_t has alpha relative to the entire cross-section of test asset returns.

33. The SDF-induced Ridge estimator is numerically equivalent to (2.21), so we do not introduce it again.

34. By Assumption 17(1), Cauchy-Schwartz and triangle inequalities, we have

$$\left\|\widehat{\gamma}_g^{Lasso} - \gamma_g\right\|_{\text{MAX}} \lesssim_{\text{P}} \sqrt{\frac{1}{T}\sum_{t=1}^{T} |\widehat{m}_t^{Lasso} - \widetilde{m}_t|^2} + \sqrt{\frac{\log N}{T}}.$$

For this purpose, it is more convenient to rewrite our SPCA estimator of the SDF given by equation (2.12) in the form of portfolio returns as in (2.15), so that we can directly evaluate its Sharpe ratio. In other words, we need an SPCA based estimate of b in the definition of SDF given by equation (2.3). Formally, we provide the following algorithm:[35]

Algorithm 8. *The SPCA based procedure for estimating SDF loadings is as follows:*
Inputs: $\bar{R}_{(1)} := \bar{R}$, $\bar{r}_{(1)} := \bar{r}$, *and* $\bar{G}_{(1)} := \bar{G}$, *a* $d \times T$ *vector.*

S1. For $k = 1, 2, \ldots$ *iterate the following steps using* $\bar{R}_{(k)}$, $\bar{r}_{(k)}$, *and* $\bar{G}_{(k)}$ *and construct an* $N \times p$ *matrix* B*:*

a. Run S1.a of Algorithm 7 to obtain $\widehat{I}_k$

b. Run S1. - S3. of Algorithm 5 with selected return matrix $\left(\bar{R}_{(k)}\right)_{[\widehat{I}_k]}$ *and* $\bar{G}_{(k)}$*. Construct the* k*th column of* B *as:* $B_{[\widehat{I}_k],k} = \varsigma_{(k)}$ *and* $B_{[\widehat{I}_k^c],k} = 0$*, where* $\varsigma_{(k)}$ *is the left singular vector of* $\left(\bar{R}_{(k)}\right)_{[\widehat{I}_k]}$*. Also, obtain* $\widehat{V}_{(k)}$ *and* $\widehat{\eta}_{(k)}$*.*

c. Run S1.c of Algorithm 7 to obtain $\widehat{\beta}_{(k)}$*.*

d. Run S1.d of Algorithm 7 to obtain $\bar{R}_{(k+1)}$ *and* $\bar{G}_{(k+1)}$*.*

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

S2. Estimate the SDF loading b *as:*

$$\widehat{b}^{SPCA} = TB\left(B^{\intercal}\bar{R}\bar{R}^{\intercal}B\right)^{-1}B^{\intercal}\bar{r}. \tag{2.19}$$

35. The effectiveness of this procedure stems from the fact that the SPCA estimates of $\widehat{V}$ can be written as a rotation of $B^{\intercal}\bar{R}$. Given that b is invariant to rotations of factors, we can exploit this invariance property to construct a convenient estimator $\widehat{b}$. To elaborate, if we use $B^{\intercal}\bar{R}$ as the factors, denoted by, $\widetilde{V}$, with their risk premia and covariance denoted as $\widetilde{\gamma}$ and $\widetilde{\Sigma}$ respectively, we can express the SDF as $m_t = 1 - \widehat{\gamma}^{\intercal}(\widehat{\Sigma}_v)^{-1}\widehat{v}_t = 1 - \widetilde{\gamma}^{\intercal}(\widetilde{\Sigma}_v)^{-1}\widetilde{v}_t = 1 - \widetilde{\gamma}^{\intercal}(\widetilde{\Sigma}_v)^{-1}B^{\intercal}(r_t - \bar{r})$. Consequently, we can deduce that:

$$\widehat{b} = B(\widetilde{\Sigma}_v)^{-1}\widetilde{\gamma} = B(\frac{1}{T}B^{\intercal}\bar{R}\bar{R}^{\intercal}B)^{-1}B^{\intercal}\bar{r}.$$

Outputs: $\widehat{b}^{SPCA}$

Similarly, we can construct estimates of b using PCA and PLS.[36] With $\widehat{b}$ it is convenient to build SDFs (optimal portfolios) and evaluate their Sharpe ratio.

Theorem 10. *Under the same assumptions as Theorem 6, if Assumption 16 holds, the Sharpe ratio of the optimal portfolio constructed by* $\widehat{b}^{SPCA}$ *in* (2.19) *satisfies*

$$\sqrt{\gamma^{\intercal}\Sigma_v^{-1}\gamma} \geq \lim_{N,T\to\infty} \frac{\widehat{b}^{SPCA\intercal}\mathrm{E}(r_t)}{\sqrt{\widehat{b}^{SPCA\intercal}\Sigma\widehat{b}^{SPCA}}} \geq \sqrt{\gamma^{\intercal}\eta^{\intercal}(\eta\Sigma_v\eta^{\intercal})^{\dagger}\eta\gamma}, \tag{2.20}$$

where † *denotes the Moore–Penrose inverse of a matrix.*

In the inequality (2.20), the upper bound corresponds to the optimal Sharpe ratio of the SDF, while the middle term represents the optimal Sharpe ratio achieved by the SPCA estimator. Meanwhile, the lower bound corresponds to the optimal Sharpe ratio achieved by $\eta(v_t + \gamma)$. This lower bound also matches the bound attained by g_t, except for any undiversified idiosyncratic errors that may persist in g_t. These errors would further reduce the Sharpe ratio, but for the sake of our discussion exclusively on observable factor models in the literature, we follow the convention and assume that g_t comprises well-diversified portfolios, so we can ignore this aspect in this section. A sufficient condition for the upper and lower bounds to be equal is that $\lambda_{\min}(\eta^{\intercal}\eta) \gtrsim 1$. In this case, the SPCA-based SDF estimator also achieves the optimal Sharpe ratio. This result is not surprising, especially considering the consistency result outlined in Theorem 8.

Theorem 10 serves as the basis for diagnosing SDF models. We do not observe the left side of the equation (the true maximal Sharpe ratio), but can estimate and compare the middle term (Sharpe ratio obtained by the SPCA-recovered SDF) and the right term (Sharpe ratio of g_t). If we find in the data that the Sharpe ratio from SPCA is higher, then we learn that

36. For PCA, the kth column of B can be chosen as the left singular vectors of $\bar{R}$, then (2.19) yields the standard PCA based SDF loadings. For PLS, B is a similar weight matrix given by the iterative procedure. We compare these SDF estimators in simulations.

g_t must be missing a factor. If we instead find that the Sharpe ratio from g_t is higher, it means that there are factors in g_t that are insufficiently represented in r_t (for example, if none of the assets in r_t has exposure to those factors): this points to an insufficiently rich set of test assets r_t.[37]

2.3 Simulations

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

2.3.1 Results on Risk Premia

We implement a number of risk premia estimators for comparison, some of which are robust to omitted or weak factors, including PCA and its related estimators (Ridge, PLS, and rpPCA), Lasso, as well as the four-split estimator by Anatolyev and Mikusheva [2021].[38] Both the standard two-pass and four-split methods directly use g_t as if they were the true factors in their regressions. The PCA, rpPCA, Ridge, and Lasso effectively construct the SDF first without knowledge of g_t, then estimate the risk premia of g_t factor by factor, using the covariance between each factor and the resulting SDF. PLS and SPCA use all variables in g_t to supervise the estimation procedure.

To implement the SPCA estimator, we select the tuning parameters p and $\lfloor qN \rfloor$ by CV using the procedure detailed in Section 2.2.2.5. To ensure a conservative basis for comparison, all methods, except for SPCA, use optimal (albeit infeasible) tuning parameters. Specifically, for PCA, PLS and rpPCA, we make use of the true number of factors, $p = 4$, even though

37. Of course, it can also be that the two Sharpe ratios are the same. In that case, g_t and the latent-factor model recovered by SPCA are equivalent in terms of their pricing ability.

38. The four-split estimator, which does not rely on dimension reduction, selection, or shrinkage techniques, is valid in the presence of weak observable factors and strong omitted factors that are *not* priced. However, it does not have asymptotic guarantees against omitted and priced strong/weak factors, or measurement error in the observed factors.

it is difficult to obtain a consistent estimator of p in the regime of weak factors. The tuning parameter μ of the Ridge estimator is determined via maximum likelihood estimation, with perfect knowledge of $\Sigma = \text{Cov}(r_t)$ and $\text{E}(r)$. The second tuning parameter of rpPCA is selected by maximizing the theoretical Sharpe ratio of the estimated SDF, using, again, perfect knowledge of Σ and $\text{E}(r)$. Due to limited sample size, estimating the sample mean and sample covariances in a separate validation sample is rather challenging, which would further deteriorate their performance.

To demonstrate and compare the performance of different estimators, we consider various DGPs of returns and/or the observed variables in g_t. We start with the benchmark case (a), in which all factors are strong and observed. Specifically, we consider a 4-factor DGP as given by equation (2.1), where the first three factors are calibrated to match the three Fama-French factors (RmRf, SMB, HML) as in Giglio and Xiu [2021], and the last one is a potentially weak factor, denoted by V. We calibrate the parameters such that the monthly Sharpe ratio for the optimal portfolio out of these factors is about 0.256. The process generating u_t is modeled as a vector autoregressive process: $u_t = 0.8u_{t-1} + \epsilon_t$, where ϵ_t is drawn from a Gaussian distribution with a diagonal covariance matrix.[39] The standard deviation of u_t is calibrated at 0.04. For comparison, the standard deviations of the four factors are calibrated at 0.04, 0.03, 0.03, and 0.02. The loadings of RmRf are generated independently from $\mathcal{N}(1, 1)$ and the loadings of SMB and HML are generated independently from $\mathcal{N}(0, 1)$. We generate the exposure to the fourth factor V, $\beta_{i,V}$, independently from a Gaussian mixture distribution, with probability a from $\mathcal{N}(0, 1)$ and $1 - a$ from $\mathcal{N}(0, 0.1^2)$. Our calibration suggests that $a = 0.5$ ensures the factor V is sufficiently strong with respect to the cross-section of assets in simulations. g_t includes exactly these four factors in the DGP (RmRF, SMB, HML, and V), and we set $\eta = \mathbb{I}_4$, and measurement error is absent.

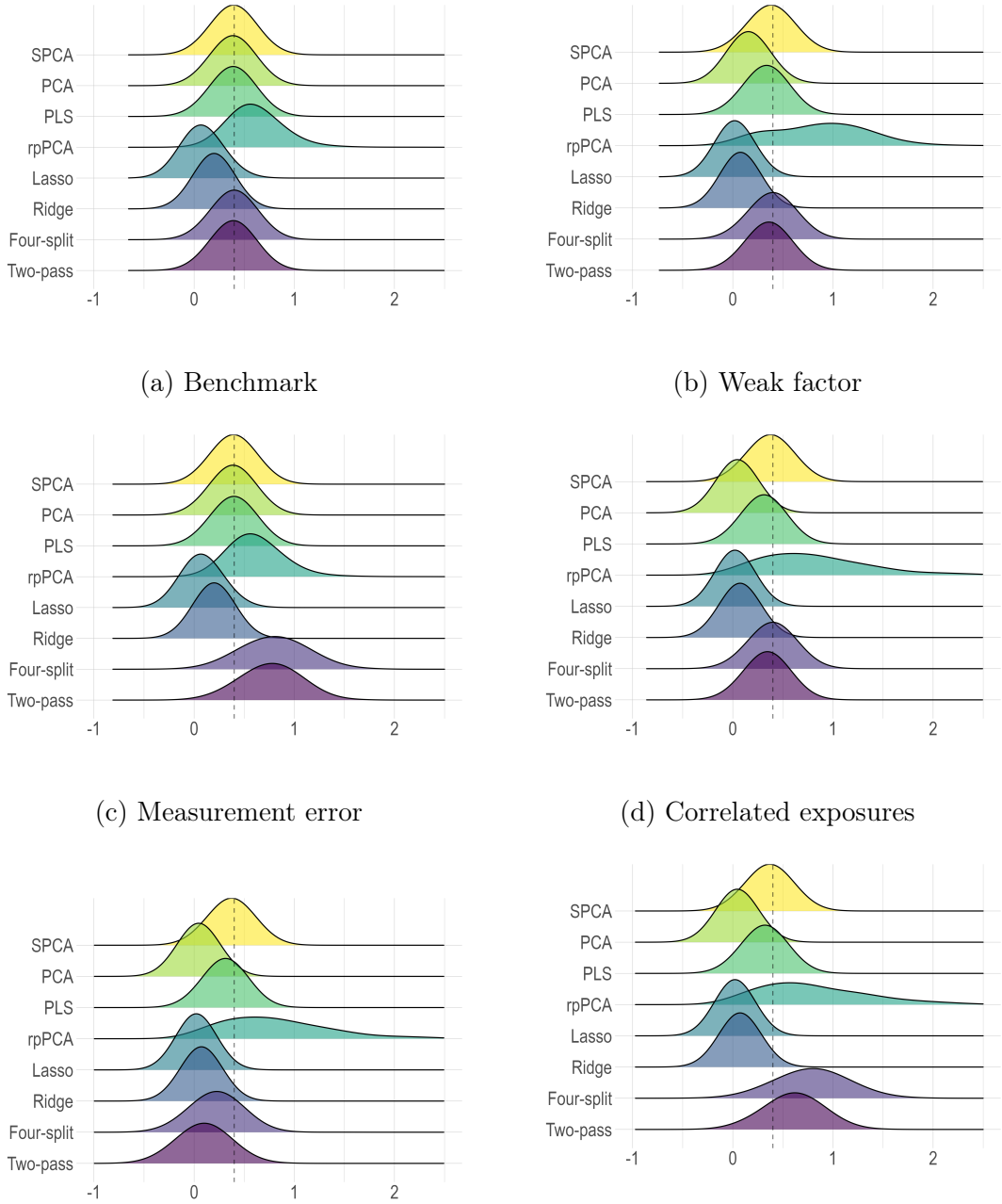
In scenario b), we choose $a = 0.1$ so that V is weak in that for almost all test assets

39. Although it is conceivable to employ a more complex covariance matrix for u_t, calibrating such a model can be a challenging endeavor. We thereby simulate u_ts that are cross-sectionally uncorrelated for simplicity.

their factor loadings to V are tiny: only 10% of the assets have nontrivial exposure to this factor. In scenario c), the DGP is the same as that of the benchmark case, except that we add Gaussian measurement error, z_t, to each of the factors in g_t. In scenario d), we simulate β for V according to $\beta_{i,V} = -\beta_{i,HML} + e_i$ instead, where e_is are generated independently from the same mixture Gaussian distribution as above with $a = 0.1$. This nearly results in a rank deficiency in the factor loading matrix due to their correlated exposures. The variable g_t contains all four factors with no measurement error. In scenario e), we consider the same DGP of returns as in scenario d), but in g_t we omit the HML factor. Finally, in scenario f), we further add measurement error to scenario d).

For each of these six scenarios (including the benchmark), we plot in Figure 2.1 the histograms of the estimated risk premium of V (one entry in g_t) for all estimators. If an estimator is consistent, then the histogram is expected to be centered around the true risk premium of V, whose value is represented by a vertical dashed line. This is indeed the case for SPCA in *all* scenarios. It is also the case for almost all estimators in the benchmark scenario, a), when factors are strong (except for Lasso and Ridge, which have a large shrinkage bias). This suggests that the latter two estimators are not suitable for *inference* on risk premia. Furthermore, in scenario b), when weak factors are present, only SPCA and four-split are consistent. The same is true for scenario d) in which a similar rank-deficiency issue arises. In scenario c) the four-split estimator becomes inconsistent due to measurement error, and it is also ill-behaved in scenario e) because the omitted variable, HML, is priced. The PCA and PLS estimators are consistent in scenario c) but also fail in e), because they are robust to measurement error but not to omitted weak factors. The standard two-pass estimator is only consistent in the benchmark scenario. Overall, the simulation evidence is in agreement with our theoretical predictions.

Next, we focus on the last scenario f), which includes the case of weak factors as well as measurement error. For this case, we report in Table 2.1 the bias and the RMSE (root-



(a) Benchmark

(b) Weak factor

(c) Measurement error

(d) Correlated exposures

(e) Weak + omitted factor

(f) Correlated exposures + measurement error

Figure 2.1: Histogram of Risk Premium Estimates of V

Note: The figure provides histograms of the risk premium estimates in six scenarios for eight estimators we compare, including SPCA, PCA, PLS, rpPCA, Lasso, Ridge, four-split, and the standard two-pass estimator. We simulate the models with $N = 1,000$ and $T = 240$. The number of Monte Carlo repetitions is 1,000. Values reported are percentages.

mean-square error) of all estimators for various sample size T. The four rows in each panel provide the results of risk premia estimation for RmRf, SMB, HML, and the weak factor V, respectively. We find that our SPCA approach has smaller biases for the weak factors, whereas the remaining estimators have larger biases and RMSEs, which agrees with our theoretical analysis and Figure 2.1. Notably, PLS ranks the second. All estimators perform better in terms of RMSE as T increases.

In the appendix, we also report a scenario similar to c) except that the last factor is a pure noise. In other words, the DGP is driven by the first three factors, but econometricians, lacking knowledge of the true model, include these three factors alongside this pure noise variable in their attempt to estimate risk premia. This scenario closely resembles the one extensively discussed by Kan and Zhang [1999] and Kleibergen [2009]. For the sake of comparison, PLS and SPCA incorporate this pure noise variable along with the aforementioned three factors into g_t. The histograms corresponding to the risk premium estimates associated with the noise factor suggest that SPCA, PCA, PLS, rpPCA, Lasso, and Ridge remain consistent and cluster around zero. The consistency stems from the fact that none of these methods involve a cross-sectional regression on the estimated beta for the noise factor. In contrast, the four-split and two-pass methods seem to exhibit considerable variances.

We then investigate the finite sample performance of the inference result developed in Theorem 7. Figure 2.2 plots histograms of the standardized risk premia estimators using the estimated asymptotic standard errors for SPCA and PCA, respectively, using the DGP in scenario f) as an example. The histograms of PCA deviate from the standard Gaussian distribution for the two highly correlated factors, V and HML. In contrast, the histograms corresponding to SPCA closely align with the standard Gaussian distribution, showcasing significantly reduced bias for these two factors. A portion of this small bias stems from the population-level approximation as demonstrated in (2.5) (see also Proposition 13). This phenomenon thereby likely persists irrespective of the value of T. Finally, we also investigate

			SPCA		PCA		rpPCA		PLS	
T	Param	True	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
	RmRf	53.7	0.2	39.2	0.4	38.9	1.8	66.4	0.2	39.1
120	SMB	21.7	-0.0	29.0	0.6	28.4	1.7	65.1	0.4	28.7
	HML	25.4	-6.7	29.3	-38.0	43.9	114.6	205.8	-15.7	30.6
	V	40.0	-6.6	20.9	-37.0	38.9	109.9	195.8	-15.7	22.6
	RmRf	53.7	0.7	29.7	0.6	29.6	1.3	36.4	0.7	29.7
240	SMB	21.7	0.2	20.1	0.6	19.5	1.2	27.8	0.4	19.8
	HML	25.4	-3.3	19.7	-36.3	39.3	64.1	111.9	-8.0	20.1
	V	40.0	-3.4	14.6	-35.5	36.5	63.0	109.0	-8.2	15.4
	RmRf	53.7	-0.1	20.2	0.0	20.2	0.2	20.7	0.0	20.2
480	SMB	21.7	-0.3	14.2	-0.2	14.0	-0.2	14.7	-0.2	14.1
	HML	25.4	-2.6	14.6	-13.4	18.6	22.3	34.6	-4.1	14.5
	V	40.0	-3.1	10.3	-13.7	16.1	20.7	32.7	-4.7	10.6
			Lasso		Ridge		Four-split		Two-pass	
T	Param	True	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
	RmRf	53.7	-27.6	37.0	-8.1	32.4	12.4	52.0	11.5	48.1
120	SMB	21.7	-12.6	16.5	-5.1	16.9	4.9	47.2	5.4	41.8
	HML	25.4	-30.6	31.6	-33.4	36.2	12.9	50.5	-6.1	40.1
	V	40.0	-38.3	38.6	-36.0	36.8	32.3	58.6	9.1	32.6
	RmRf	53.7	-31.6	37.4	-4.2	25.8	13.4	40.1	12.4	37.9
240	SMB	21.7	-14.0	16.3	-3.0	13.9	6.1	33.3	5.9	29.5
	HML	25.4	-29.9	30.7	-31.5	33.7	16.2	37.3	2.5	27.4
	V	40.0	-37.6	37.9	-32.7	33.4	38.8	51.2	20.7	32.1
	RmRf	53.7	-18.5	24.7	-1.7	19.1	12.6	29.5	11.9	27.3
480	SMB	21.7	-9.0	11.9	-1.5	12.0	4.3	24.0	4.7	20.9
	HML	25.4	-32.8	33.5	-29.1	30.9	16.6	29.4	8.3	22.1
	V	40.0	-36.8	37.1	-29.5	30.1	38.6	45.6	28.0	33.5

Table 2.1: Simulation Results for Risk Premia Estimators

Note: In this table, we report the bias (Column "Bias") and the root-mean-square error (Column "RMSE") of the risk premia estimates using SPCA, PCA, rpPCA, Lasso, PLS, Ridge, four-split, and the standard two-pass regression approaches, respectively. The true data-generating process, given by scenario f), has four factors, driven by RmRf, SMB, HML, and V, whereas we estimate the risk premia for noisy versions of these four factors. Their true risk premia are provided in Column "True." We fix $N = 1,000$ while varying $T = 120$, 240, and 480 in this experiment. All values reported are in basis points.

the statistical power of SPCA in strong and weak cases, respectively, and draw a comparative analysis with PCA. We report these results in the appendix.

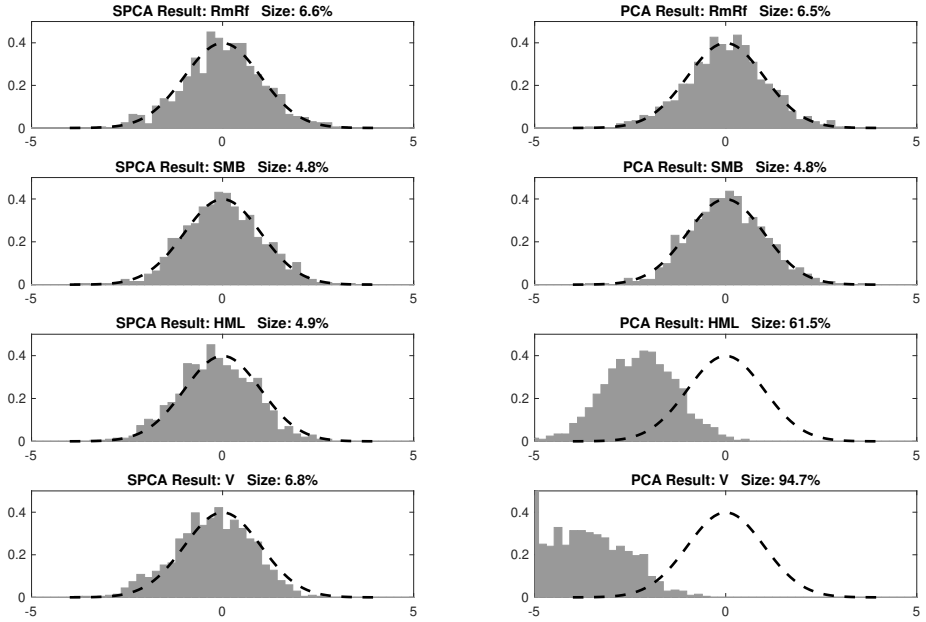


Figure 2.2: Histogram of the Standardized Estimates in Simulations

Note: The left panels provide the histograms of the standardized SPCA estimates as in Algorithm 7 with asymptotic standard errors given by Theorem 7, whereas the right panels provide those of the standardized PCA-based risk premia estimates as in Algorithm 5. We simulate the model in scenario f) with $N = 1,000$ and $T = 240$. The number of Monte Carlo repetitions is 1,000. These standardized statistics serve as the basis for testing the null hypotheses that the risk premia are equal to their true values. The sizes of these t-tests at 5% level are reported in the figure subtitles, allowing us to assess the tail behavior of our asymptotic approximations.

2.3.2 Results on SDF recovery

Next, we study the finite sample behavior of the SDF estimators. We compare the performance of SPCA, PCA, rpPCA, Lasso and Ridge estimators in scenario f). We report in Table 2.2 the MSE of the SDF estimators where the true SDF is defined by equation (2.3). We also include the tuned number of factors determined through our SPCA approach. Additionally, we report in Table 2.3 the out-of-sample Sharpe ratios of different methods,

	SPCA		PCA	rpPCA	PLS	Lasso	Ridge
T	$\widehat{p}$	MSE	MSE	MSE	MSE	MSE	MSE
120	4.186	0.044	0.074	9.200	0.050	0.056	0.054
	(0.389)	(0.030)	(0.026)	(11.332)	(0.026)	(0.010)	(0.013)
240	4.011	0.021	0.058	1.901	0.025	0.055	0.045
	(0.104)	(0.014)	(0.013)	(3.313)	(0.013)	(0.009)	(0.010)
480	4.004	0.010	0.018	0.087	0.012	0.050	0.036
	(0.063)	(0.007)	(0.007)	(0.083)	(0.007)	(0.007)	(0.008)

Table 2.2: Simulation Results for SDF estimators

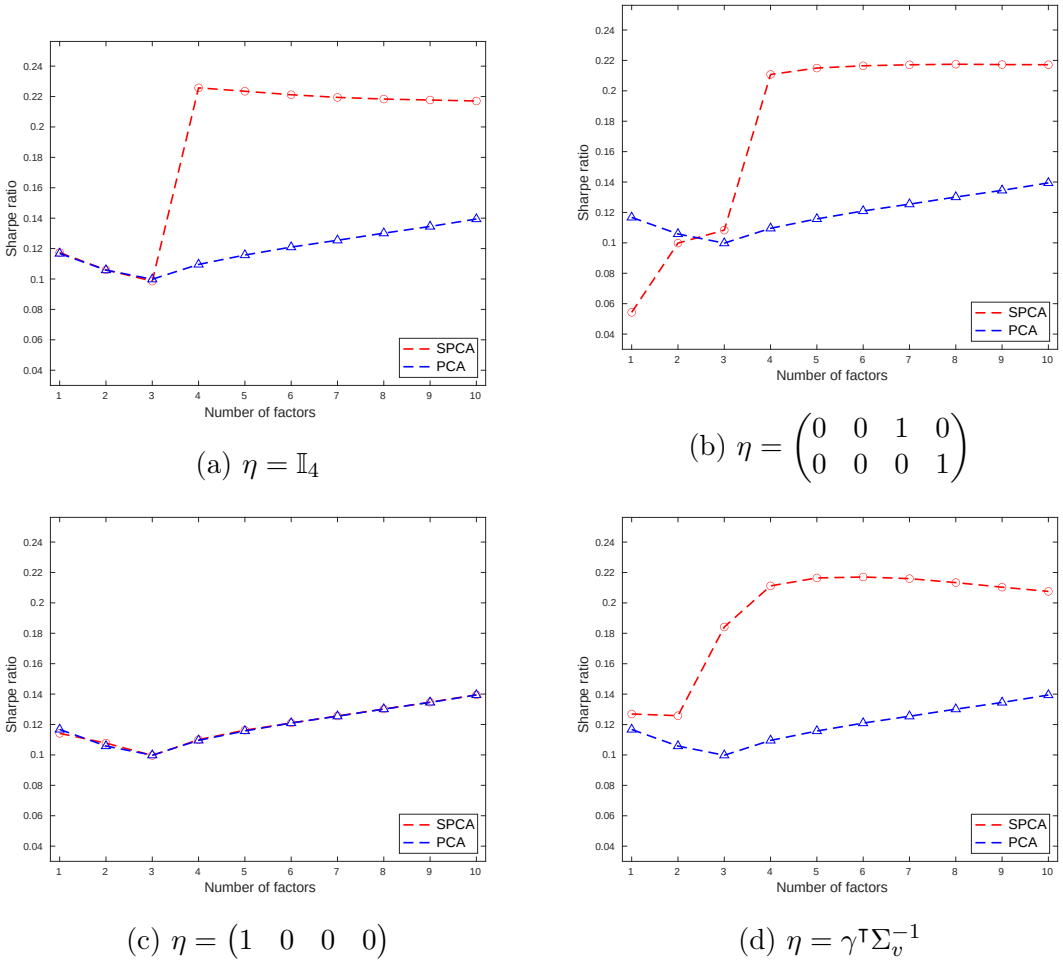
Note: In this table, we report the mean-squared errors (Column "MSE") defined by $\frac{1}{T}\sum_{t=1}^{T}|\widehat{m}_t - \widetilde{m}_t|^2$ for various SDF estimates using SPCA, PCA, rpPCA, PLS, Lasso, and Ridge approaches, respectively. The reported MSEs are the sample average over 1,000 Monte Carlo repetitions and their standard deviations are reported in the brackets. We also report the mean and standard deviation of the estimated number of factors $\widehat{p}$ using the SPCA approach. The true data-generating process, given by scenario f), has four factors, driven by RmRf, SMB, HML, and a weak factor V, whereas we estimate the SDF using a vector of factor proxies, g_t, that includes noisy versions of the four factors. We compare three scenarios with $T = 120$, 240, and 480, where $N = 1,000$ is fixed.

T	SPCA	PCA	rpPCA	PLS	Lasso	Ridge	Theoretical Value
120	0.193	0.084	0.134	0.164	0.113	0.109	0.256
	(0.049)	(0.046)	(0.035)	(0.051)	(0.024)	(0.046)	
240	0.226	0.110	0.192	0.214	0.122	0.137	0.256
	(0.026)	(0.036)	(0.033)	(0.031)	(0.019)	(0.032)	
480	0.241	0.227	0.242	0.238	0.127	0.162	0.256
	(0.012)	(0.019)	(0.008)	(0.015)	(0.021)	(0.019)	

Table 2.3: Simulation Results for Out-of-Sample Sharpe Ratios of Optimal Portfolios

Note: In this table, we report the mean and standard deviation of the out-of-sample Sharpe ratios for various optimal portfolios constructed by SPCA, PCA, rpPCA, PLS, Lasso, and Ridge approaches, respectively. The true data-generating process, given by scenario f), has four factors, driven by RmRf, SMB, HML, and a weak factor V, whereas we estimate the SDF using a vector of factor proxies, g_t, that includes noisy versions of the four factors. The reported Sharpe ratios are the sample average over 1,000 Monte Carlo repetitions and their standard errors are reported in the brackets. Column "'Theoretical Value" provides the benchmark Sharpe ratio calculated by $b^\intercal \mathrm{E}(r)/\sqrt{b'\Sigma b}$ using true parameter values. We compare three scenarios with $T = 120$, 240, and 480, where $N = 1,000$ is fixed.

given by $\widehat{b}^{\intercal}\mathrm{E}(r)/\sqrt{\widehat{b}^{\intercal}\Sigma\widehat{b}}$, where $\mathrm{E}(r)$ and Σ are the true mean and covariance of all test assets and $\widehat{b}$ is the estimated SDF loading using each method. Overall, we find that SPCA outperforms all other methods. PLS ranks second, while rpPCA performs the worst. rpPCA is only competitive in terms of the out-of-sample Sharpe ratio. For risk premia estimation, the disadvantage of rpPCA relative to other methods may not only stem from its inherent bias but also from its tuning parameters being primarily oriented towards maximizing the out-of-sample Sharpe ratio. Last but not least, the tuning parameter $\widehat{p}$ is found to be in close proximity to the truth value of 4.



(a) $\eta = \mathbb{I}_4$

(b) $\eta = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$

(c) $\eta = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix}$

(d) $\eta = \gamma^{\intercal}\Sigma_v^{-1}$

Figure 2.3: Out-of-sample Sharpe Ratio Patterns with Different Models of g_t

Note: Each panel reports the out-of-sample Sharpe ratios for PCA (blue) and SPCA (red) as a function of number of factors, p, for a specific model of $g_t = \eta v_t + z_t$.

Finally, we investigate the pattern of out-of-sample Sharpe ratios for various models g_t in Figure 2.3. The setting resembles scenario (f), except that we consider different models g_t to examine the role of g_t in supervising the procedure. We report Sharpe ratios as a function of number of factors $\widehat{p}$ used in the PCA and SPCA procedure. For SPCA, we select $\lfloor qN \rfloor$ via CV using the time series R^2 for each given $\widehat{p}$. The sample size T is fixed at 240. The theoretical value of the optimal Sharpe ratio is 0.256, as shown in Table 2.3, though in finite sample the maximum Sharpe ratio achieved by SPCA is around 0.226.

We consider four cases of $g_t = \eta v_t + z_t$ here. In case (a), we set $\eta = \mathbb{I}_4$, so all factors are included in g_t to supervise the procedure. In case (b), only the factor V and HML are included in g_t. In case (c), we fix $\eta = (1, 0, 0, 0)$, that is, g_t only includes the (strong) market factor. Finally, in case (d), we let $\eta = \gamma^\intercal \Sigma_v^{-1}$, so that g_t is a noisy measure of the SDF. In light of Theorem 10, SPCA should achieve the maximal out-of-sample Sharpe ratio in cases (a) and (d), provided appropriate tuning parameters. Figure 2.3 confirms this result. In case (a), SPCA reaches its highest Sharpe ratio out-of-sample precisely at $\widehat{p} = 4$, and the Sharpe ratio declines slightly as $\widehat{p}$ increases beyond 4, since these additional factors only add noise. Case (d) exhibits a similar pattern. In contrast, the PCA approach cannot achieve the maximal Sharpe ratio, even as $\widehat{p}$ increases to 10, because PCA cannot recover the weak factor, which contributes to the SDF. In case (b), SPCA is supervised by two factors with highly correlated loadings, so it can recover the part of the SDF spanned by the weak factors. With a large enough p, we force the procedure of SPCA to continue, then it will also extract the strong factors and achieve the maximal Sharpe ratio. In case (c), however, SPCA and PCA provide similar results — neither achieves the optimum — because g_t only includes the market factor, which does not help SPCA recover the missing weak factor.

2.4 Conclusions

The choice of test assets plays a fundamental role in empirical asset pricing tests. The recent explosion of anomaly discoveries and related characteristics in the empirical literature has provided researchers with a large universe of potential test assets to choose from. On the one hand, the availability of so many different characteristics gives us hope that the returns of these portfolios can help us uncover and identify the pricing of various dimensions of risk, including those that are not well captured by standard cross-sections. On the other hand, the large dimensionality goes hand in hand with the weak factor issue: a factor may well be captured by *some* assets within the large cross-section, but if most assets do not have exposure to that factor, it will be weak and inference will be incorrect.

Traditional methodologies take the cross-section of assets as given. In this paper, we present a new methodology, SPCA, that instead actively selects assets in order to estimate risk premia of factors of interest, whether they are strong or weak, and at the same time addresses the issue of potentially omitted factors, again regardless of whether they are strong or weak. In addition, SPCA can exploit its ability to recover weak latent factors to help diagnose omitted factors in observable-factor models. The paper confirms the good performance of SPCA for both of these tasks in a variety of simulations, and illustrates the application of the methodology in various empirical contexts in Section 3.3.

While the road to a full understanding of risk and risk premia in financial markets is still long, we believe that systematically tackling weak factors in empirical asset pricing is an important step forward, that opens the door to the study of factors that, while important to investors, may be not pervasive in either the standard cross-sections or the recently developed large universes of test assets.

Two pressing issues on the debates related to the factor zoo are the economic interpretability and the overwhelming amount of degrees of freedom in empirical asset pricing research. The central issue we address in this paper is to evaluate factors motivated by economic the-

ories. Our proposal eliminates two critical degrees of freedom altogether from this exercise: the choice of control factors when estimating risk premia of economically motivated factors, and the choice of test assets used for estimation and testing. Our study thereby contributes to a promising agenda developing a fusion of asset pricing theory and machine learning. It does so by using the factor structure as a main theoretical foundation, and applying to it tools and results from machine learning, in order to exploit these statistical advances while maintaining economic interpretability.

2.5 Appendix

2.5.1 Alternative Estimators and Their Asymptotic Behavior

While the literature has proposed several different estimators of the SDF and risk premia, their properties in a general weak factor setting like ours have not been investigated. In this section we revisit a number of estimation procedures, and show that they are inconsistent in the presence of weak factors, using a simple model with a single weak factor.

We focus our discussion of alternative estimators on those that can be used when factors are latent. In this setting, the researcher does not need to know the identities of all true factors, which yields a risk premium estimator that is robust to potentially omitted factors.

PLS

As reviewed in the main text, Giglio and Xiu [2021] show that the PCA-based estimation procedure effectively constructs a mimicking portfolio for g_t via a principal component regression (PCR) on r_t, which amounts to a projection of g_t onto the first few PCs of the sample covariance matrix of r_t. This is an unsupervised approach, in that the PCs are obtained without any information from g_t. Therefore, PCA might be misled by large idiosyncratic errors in r_t when the signal is not sufficiently strong. In contrast with PCA, partial least

squares (PLS) is a supervised procedure, which has been shown to work better than PCA in other settings, see, e.g., Kelly and Pruitt [2013]. In the same spirit, we propose a PLS-based approach for risk premia estimation, exploiting variation of returns that is relevant to the target factor of interest.

The key difference between the two approaches is that PCA seeks linear combinations of r_t that maximize variation, ignoring information from the target g_t, whereas PLS seeks linear combinations that have the largest covariance with g_t. The PLS-based risk premia estimator effectively uses PLS instead of PCA in the first step of Algorithm 5 described in the main text.

We formulate a general PLS-based algorithm for a $d \times 1$ vector of g_t below:

Algorithm 9 (PLS-based Estimator of Risk Premia)**.** *The estimator proceeds as follows:*
Inputs: $\bar{R}_{(1)} := \bar{R}$, $\bar{r}_{(1)} := \bar{r}$ *and* $\bar{G}$, *a* $d \times T$ *matrix.*

S1. For $k = 1, 2, \cdots, p$, *repeat the following steps using* $\bar{R}_{(k)}$, $\bar{r}_{(k)}$ *and* $\bar{G}$.

   a. *Obtain the weight vector* $\widehat{w}$ *from the largest left singular vector of* $\bar{R}_{(k)}\bar{G}^{\intercal}$.

   b. *Estimate the* k*th factor as* $\widehat{V}_{(k)} = \sqrt{T}\widehat{w}^{\intercal}\bar{R}_{(k)} / \left\|\widehat{w}^{\intercal}\bar{R}_{(k)}\right\|$.

   c. *Estimate the risk premium of* $\widehat{V}_{(k)}$ *by* $\widehat{\gamma}_{(k)} = \sqrt{T}\widehat{w}^{\intercal}\bar{r}_{(k)} / \left\|\widehat{w}^{\intercal}\bar{R}_{(k)}\right\|$.

   d. *Estimate the* k*th factor loading of* r_t *by* $\widehat{\beta}_{(k)} = T^{-1}\bar{R}_{(k)}\widehat{V}_{(k)}^{\intercal}$.

   e. *Remove* $\widehat{V}_{(k)}$ *to obtain residuals for the next step:* $\bar{R}_{(k+1)} = \bar{R}_{(k)} - \widehat{\beta}_{(k)}\widehat{V}_{(k)}$ *and* $\bar{r}_{(k+1)} = \bar{r}_{(k)} - \widehat{\beta}_{(k)}\widehat{\gamma}_{(k)}$.

S2. Estimate the factor loading of g_t *on* v_t *by* $\widehat{\eta} = T^{-1}\bar{G}\widehat{V}^{\intercal}$, *where* $\widehat{V} = (\widehat{V}_{(1)}^{\intercal}, \cdots, \widehat{V}_{(p)}^{\intercal})^{\intercal}$, *and denote their risk premia estimated above as* $\widehat{\gamma} = (\widehat{\gamma}_{(1)}, \cdots, \widehat{\gamma}_{(p)})^{\intercal}$.

Output: $\widehat{\gamma}_g^{PLS} = \widehat{\eta}\widehat{\gamma}$.

The PLS estimator has a closed-form formula if $\bar{G}$ is a $1 \times T$ vector and a single-factor is extracted ($p = 1$):

$$\widehat{\gamma}_g^{PLS} = \left\|\bar{G}\bar{R}^\intercal \bar{R}\right\|^{-2} \bar{G}\bar{R}^\intercal \bar{R}\bar{G}^\intercal \bar{G}\bar{R}^\intercal \bar{r}.$$

While the PLS procedure seems appealing, the next proposition shows that this approach is asymptotically equivalent to the PCA-based procedure, hence it fails in exactly the same weak factor setting as PCA.

Proposition 10. *Suppose that test asset returns follow a single-factor model in the form of* (2.1) *with* $p = 1$, g_t *satisfies* (2.4) *with* $d = 1$, u_t *and* v_t *i.i.d. normally distributed and independent from each other, and* $z_t = 0$. *In addition, suppose that* β *satisfies* $N/(\|\beta\|^2 T) \to B \geq 0$ *and* $\|\beta\| \to \infty$. *Then we have* $\widehat{\gamma}_g^{PLS} \xrightarrow{\text{P}} (1+B)^{-1}\eta\gamma$.

Intuitively, the covariance information embedded in the objective function of PLS is dominated by its variance component, hence PLS yields the same asymptotic behavior as PCA with respect to estimating β, and therefore risk premia.

Ridge

Next, we consider an alternative, ridge-regression-based approach to the construction of mimicking portfolios, which instead directly regularizes the projection of g_t on the vector of returns. The Ridge-based estimator can be written as:

$$\widehat{\gamma}_g^{Ridge} = \bar{G}\bar{R}^\intercal \left(\bar{R}\bar{R}^\intercal + \mu \mathbb{I}_N\right)^{-1} \bar{r}, \tag{2.21}$$

where $\mu > 0$ is some tuning parameter. In the case of pervasive factors, Giglio and Xiu [2021] show that the ridge estimator yields a consistent estimate of $\eta\gamma$. However, we show that the ridge estimator also fails in the presence of weak factors:

Proposition 11. *Suppose that test asset returns follow a single-factor model in the form of* (2.1) *with* $p = 1$, g_t *satisfies* (2.4) *with* $d = 1$, u_t *and* v_t *i.i.d. normally distributed and independent from each other, and* $z_t = 0$. *In addition, suppose that* β *satisfies* $N/(\|\beta\|^2 T) \to B \geq 0$ *and* $\|\beta\| \to \infty$, *and the tuning parameter* μ *satisfies* $\mu/(\|\beta\|^2 T) \to D$ *for some constant* $D \geq 0$ *such that* $B + D > 0$. *Then we have* $\widehat{\gamma}_g^{Ridge} \xrightarrow{\text{P}} (1 + B + D)^{-1}\eta\gamma$.

Intuitively, the Ridge-based estimator fails because the tuning parameter μ in the ridge procedure serves as a threshold that suppresses the influence of eigenvectors corresponding to small eigenvalues, just like in PCA and PLS (which explains the appearance of B in the limit). The presence of μ also leads to a shrinkage bias to the first few eigenvectors (i.e., factors), which is why an extra term D appears in the limit as well.

Risk Premium PCA

Finally, we consider an estimator based on the approach of Lettau and Pelger [2020]. This approach was designed to estimate a latent-factor SDF, but can also be used to estimate the risk premium of a factor g_t, by replacing the PCA step of Algorithm 5 with the risk premia PCA procedure of Lettau and Pelger [2020]:

Algorithm 10 (rpPCA-based Estimator of Risk Premia)**.** *The estimator proceeds as follows: Inputs:* $\bar{R}$ *and* $\bar{G}$.

S1. Apply PCA on $T^{-1}RR^\intercal + \mu\bar{r}\bar{r}^\intercal$, *where* μ *is a tuning parameter, and write the first* p *eigenvectors as* $\widehat{\varsigma}$. *The estimated factors are given by* $\widehat{V} = \widehat{\varsigma}^\intercal \bar{R}$.

S2. Estimate the risk premia of $\widehat{V}$ *by* $\widehat{\gamma} = \widehat{\varsigma}^\intercal \bar{r}$.

S3. Estimate the factor loading of g_t *on* v_t *by* $\widehat{\eta} = \bar{G}\widehat{V}^\intercal(\widehat{V}\widehat{V}^\intercal)^{-1}$.

Outputs: $\widehat{\gamma}_g^{rpPCA} = \widehat{\eta}\widehat{\gamma}$.

Standard PCA is applied to the covariance matrix of returns, that is $T^{-1}RR^{\intercal} - \bar{r}\bar{r}^{\intercal}$. Lettau and Pelger [2020] show that assigning a larger weight $\mu > -1$ to the term related to average returns (the second term) improves the Sharpe ratio of the estimated SDF. Lettau and Pelger [2020] derive asymptotic properties of rpPCA in a setting where all factors are weak and N and T increase to infinity at the same rate. The setting they analyze is one where all factors are extremely weak, so that they cannot be recovered – specifically, the strength of weak factors remains indistinguishable from that of idiosyncratic errors as N and T increase. Under this assumption, consistent estimation of the SDF is impossible, including by rpPCA, which, despite being more correlated with the SDF than PCA, is also inconsistent. In contrast, we preclude this extreme case from our discussion because no estimators under consideration could achieve consistency and a harmless modeling choice would be to treat these extremely weak factors as noise: their risk premia cannot be distinguished from alpha. The weak-factor setting we investigate permits consistency, and allows for asymptotic comparison of different estimators. The following proposition shows that, like PCA, rpPCA is also inconsistent for estimating risk premia.

Proposition 12. *Suppose that test asset returns follow a single-factor model in the form of* (2.1) *with* $p = 1$, g_t *satisfies* (2.4) *with* $d = 1$, u_t *and* v_t *i.i.d. normally distributed and independent from each other, and* $z_t = 0$. *In addition, suppose that* β *satisfies* $N/(\|\beta\|^2 T) \to B \geq 0$ *and* $\|\beta\| \to \infty$, *that the factor has a non-zero risk premia, i.e.,* $\gamma \neq 0$. *Then for some tuning parameter* $\mu > -1$, *we have*

$$\widehat{\gamma}_g^{rpPCA} \xrightarrow{\text{P}} w(1+B)^{-1}\eta\gamma + (1-w)\eta(\gamma + \gamma^{-1}B),$$

where w *is a constant that depends on* B, μ, γ, *explicitly given by equation* (2.119) *in the proof.*

Proposition 12 shows that this rpPCA estimator is inconsistent in the presence of a weak

factor, with a more involved bias term compared to the above estimators. Like PCA and PLS, this estimator is consistent when all factors are strong ($B = 0$). When $B > 0$, the estimator is inconsistent.

Different asymptotic settings can affect the asymptotic behavior of rpPCA. For example, if one assumes that the tuning parameter $\mu \to \infty$, the rpPCA estimator converges to $\eta(\gamma + \gamma^{-1}B)$ (so it still displays a bias). If we further assume $\gamma \to \infty$ (while keeping $\eta\gamma$ constant), this estimator can be consistent as long as $\eta\gamma^{-1}B \xrightarrow{\text{P}} 0$. This suggests that rpPCA can be robust to weak factors if the information about β from the expected return $\beta\gamma$ dominates the information from return covariances (when $\gamma \to \infty$). But this is only the case if factors have diverging Sharpe ratios, i.e., $\Sigma_v^{-1/2}\gamma \to \infty$.

2.5.2 Model Assumptions

To derive the asymptotic properties of the SPCA and alternative estimators, we need the following high-level assumptions, which can be easily verified by standard and more primitive assumptions. We start with assumptions that characterize the data generating process (DGP) of returns and factor proxies.

Assumption 7. *The factor innovation V satisfies:*

$$\|\bar{v}\| \lesssim_{\text{P}} T^{-1/2}, \quad \left\|T^{-1}VV^{\intercal} - \Sigma_v\right\| \lesssim_{\text{P}} T^{-1/2}, \quad \|V\|_{\text{MAX}} \lesssim_{\text{P}} (\log T)^{1/2},$$

where $\Sigma_v \in \mathbb{R}^{p\times p}$ is a positive-definite matrix with $\lambda_p(\Sigma_v) \gtrsim 1$ and $\lambda_1(\Sigma_v) \lesssim 1$.

Assumption 8. *The residual innovation Z satisfies:*

$$\|\bar{z}\| \lesssim_{\text{P}} T^{-1/2}, \quad \left\|T^{-1}ZZ^{\intercal} - \Sigma_z\right\| \lesssim_{\text{P}} T^{-1/2}, \quad \|Z\|_{\text{MAX}} \lesssim_{\text{P}} (\log T)^{1/2},$$

where $\Sigma_z \in \mathbb{R}^{d\times d}$ *is a positive-definite matrix with* $\lambda_d(\Sigma_z) \gtrsim 1$ *and* $\lambda_1(\Sigma_z) \lesssim 1$. *In addition,*

$$\|ZV^\intercal\| \lesssim_{\mathrm{P}} T^{1/2}.$$

Assumptions 7 and 8 impose rather weak conditions on the time series behavior of the factors and measurement error. Since v_t and z_t have a finite cross-sectional dimension, both assumptions hold if these processes are stationary, strong mixing, and satisfy some moment conditions.

Assumption 9. *The factor loading matrix* β *satisfies*

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

for some index set I_0, *where* $N_0 = |I_0|$.

Assumption 9 implies that there exists a subset of test assets, within which all latent factors are strong. It does not imply all factors should have identical strength with respect to the entire cross-section of assets in r_t.

Next, we need the following moment conditions.

Assumption 10. *The idiosyncratic component* U *satisfies:*

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

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

$$\left\|U_{[I]}\right\| \lesssim_{\mathrm{P}} |I|^{1/2} + T^{1/2}, \quad \left\|\bar{u}_{[I]}\right\| \lesssim_{\mathrm{P}} |I|^{1/2} T^{-1/2}.$$

Assumption 10 imposes restrictions on the time-series dependence and heteroskedasticity of u_t. We do not necessarily need stationarity on u_t. That said, the first two inequalities

can be shown by some large deviation theorem, see, e.g., Fan et al. [2011]; the last two inequalities can be shown by random matrix theory, see Bai and Silverstein [2009], if u_t is i.i.d. both in time and in the cross-section.

Assumption 11. *For any non-random subset $I \subset \langle N \rangle$, the factor loading $\beta_{[I]}$ and the idiosyncratic error $U_{[I]}$ satisfy the following conditions:*

$$(i) \quad \left\| (\beta_{[I]}^{\intercal}\beta_{[I]})^{-1/2}\beta_{[I]}^{\intercal}U_{[I]} \right\| \lesssim_{\mathrm{P}} T^{1/2}.$$
$$(ii) \quad \left\| (\beta_{[I]}^{\intercal}\beta_{[I]})^{-1/2}\beta_{[I]}^{\intercal}U_{[I]}\iota_T \right\| \lesssim_{\mathrm{P}} T^{1/2}.$$

If $\beta_{[I]}^{\intercal}\beta_{[I]}$ is singular, we need replace the matrix inverse above by the Moore-Penrose inverse.

Assumption 12. *The following conditions hold for U, V, β, and any non-random subset $I \subset \langle N \rangle$:*

$$(i) \quad \left\| U_{[I]}V^{\intercal} \right\| \lesssim_{\mathrm{P}} |I|^{1/2}T^{1/2}, \quad \left\| U_{[I]}V^{\intercal} \right\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} (\log N)^{1/2}T^{1/2}.$$
$$(ii) \quad \left\| (\beta_{[I]}^{\intercal}\beta_{[I]})^{-1/2}\beta_{[I]}^{\intercal}U_{[I]}V^{\intercal} \right\| \lesssim_{\mathrm{P}} T^{1/2}.$$

Assumption 13. *The following conditions hold for U, Z, β, and any non-random subset $I \subset \langle N \rangle$:*

$$(i) \quad \left\| U_{[I]}Z^{\intercal} \right\| \lesssim_{\mathrm{P}} |I|^{1/2}T^{1/2}, \quad \left\| U_{[I]}Z^{\intercal} \right\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} (\log N)^{1/2}T^{1/2}.$$
$$(ii) \quad \left\| (\beta_{[I]}^{\intercal}\beta_{[I]})^{-1/2}\beta_{[I]}^{\intercal}U_{[I]}Z^{\intercal} \right\| \lesssim_{\mathrm{P}} T^{1/2}.$$

Assumptions 11 - 13 resemble Assumptions A.7, A.9, and A.10 of Giglio and Xiu [2021], except that here we impose their stronger versions which hold for any non-random subset $I \subset \langle N \rangle$. Of course, these two sets of assumptions are equivalent if u_t is identically distributed along the cross-sectional dimension.

In the main text, we denote the selected subsets in the SPCA procedure as $\widehat{I}_k$, $k = 1, 2, \ldots$. We now define their population counterparts. Because SPCA is an iterative procedure, we need these quantities to characterize the limiting behavior of the procedure.

Without loss of generality, we consider the case $\Sigma_v = \mathbb{I}_p$ here. In the general case, we can simply replace β and η by $\beta' = \beta\Sigma_v^{1/2}$ and $\eta' = \eta\Sigma_v^{1/2}$ in the following definitions. In detail, we start with $a_i^{(1)} := \left\|\beta_{[i]}\eta^\intercal\right\|_{\text{MAX}}$ and define $I_1 := \{a_i^{(1)} \geq c_{qN}^{(1)}\}$, where $c_{qN}^{(1)}$ is the (qN)th largest value in $\left\{a_i^{(1)}\right\}_{i=1,\ldots,N}$. Then, we denote the largest right singular vector of $\beta_{(1)} := \beta_{[I_1]}$ by b_1. For $k > 1$, we obtain $a_i^{(k)} := \left\|\beta_{[i]}\prod_{j<k}\mathbb{M}_{b_j}\eta^\intercal\right\|_{\text{MAX}}$, $I_k := \{a_i^{(k)} \geq c_{qN}^{(k)}\}$ and b_k is the largest right singular vector of $\beta_{(k)} := \beta_{[I_k]}\prod_{j<k}\mathbb{M}_{b_j}$. This procedure is stopped at step $\tilde{p}$ (for some $\tilde{p}$ that is not necessarily equal to p) if $c_{qN}^{(\tilde{p}+1)} < c$. In a nutshell, I_k's are what we will select if we do SPCA directly on $\beta \in \mathbb{R}^{N\times p}$ and $\eta \in \mathbb{R}^{d\times p}$, while $\widehat{I}_k$'s are obtained by SPCA on $\bar{R} \in \mathbb{R}^{N\times T}$ and $\bar{G} \in \mathbb{R}^{d\times T}$. We need the following assumption to guarantee the selection consistency, that is, $\mathrm{P}(\widehat{I}_k = I_k) \to 1$ for any $1 \leq k \leq \tilde{p}$.

Assumption 14. *We assume that $\beta_{(k)}$, $a_i^{(k)}$ and c in the above procedure satisfy:*

(i) $\sigma_1(\beta_{(k)})$ and $\sigma_2(\beta_{(k)})$ are distinct in the sense that there exists a constant $\delta > 0$ such that

$$\sigma_2(\beta_{(k)}) \leq (1+\delta)^{-1}\sigma_1(\beta_{(k)}).$$

(ii) $c_{qN}^{(k)}$ and $c_{qN+1}^{(k)}$ are distinct in the sense that there exists a constant $\delta > 0$ such that

$$c_{qN+1}^{(k)} \leq (1+\delta)^{-1}c_{qN}^{(k)},$$

where $c_{qN}^{(k)}$ and $c_{qN+1}^{(k)}$ are the (qN)th and $(qN+1)$th largest value in $\left\{a_i^{(k)}\right\}_{i=1,\ldots,N}$, respectively.

(iii) $c_{qN}^{(\tilde{p}+1)}$ and c are distinct in the sense that there exists a constant $\delta > 0$ such that

$$c_{qN}^{(\tilde{p}+1)} \leq (1+\delta)^{-1}c.$$

Assumption 14 requires that these singular values are distinguishable, so that their (relative) differences will not vanish asymptotically. This assumption is rather mild, despite not being very explicit.

Assumption 15. *As $T \to \infty$, the following joint central limit theorem holds:*

$$T^{1/2} \begin{pmatrix} T^{-1}\mathrm{vec}(ZV^\intercal) \\ \bar{v} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{12}^\intercal & \Pi_{22} \end{pmatrix} \right),$$

where Π_{11}, Π_{12}, Π_{22} are $dp \times dp$, $dp \times p$, and $p \times p$ matrices, respectively, defined as:

$$\begin{aligned} \Pi_{11} &= \lim_{T\to\infty} \frac{1}{T} \mathrm{E}\left(\mathrm{vec}(ZV^\intercal)\mathrm{vec}(ZV^\intercal)^\intercal\right), \\ \Pi_{12} &= \lim_{T\to\infty} \frac{1}{T} \mathrm{E}\left(\mathrm{vec}(ZV^\intercal)\iota_T^\intercal V^\intercal\right), \\ \Pi_{22} &= \lim_{T\to\infty} \frac{1}{T} \mathrm{E}\left(V\iota_T\iota_T^\intercal V^\intercal\right). \end{aligned}$$

Assumption 15 characterizes the joint asymptotic distribution of $ZV^\intercal$ and $V\iota_T$. Since the dimensions of these random processes are finite, this CLT is a standard result of a central limit theory for mixing processes.

Blow we introduce assumptions needed for the SDF estimation. Assumption 16 ensures that the SDF concept is well defined. Assumption 17 again can be shown by some large deviation result and certain central limit theorem.

Assumption 16. *Suppose that v_t and u_t are stationary time series independent of β, and that $\Sigma_v = \mathrm{Cov}(v_t)$ and $\Sigma_u = \mathrm{Cov}(u_t)$ satisfy $\lambda_{\min}(\Sigma_v) \gtrsim 1$ and $\lambda_{\max}(\Sigma_u) \lesssim 1$. Consequently, $\Sigma = \mathrm{Cov}(r_t) = \beta\Sigma_v\beta^\intercal + \Sigma_u$.*

Assumption 17. *The time series r_t and the SDF defined by $m_t = 1 - b^\intercal(r_t - \mathrm{E}(r))$ with*

$b = \Sigma^{-1}\mathrm{E}(r_t)$ *satisfy:*

$$(1) \quad \left\| T^{-1}\sum_{t=1}^{T}(r_t - \bar{r}_t)(m_t - \bar{m}_t) - \mathrm{Cov}(r_t, m_t) \right\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} (\log N)^{1/2} T^{-1/2}.$$

$$(2) \quad \left\| T^{-1}\sum_{t=1}^{T}(r_t - \bar{r}_t)(r_t - \bar{r}_t)^{\intercal} - \mathrm{Cov}(r_t) \right\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} (\log N)^{1/2} T^{-1/2}.$$

$$(3) \quad \left| T^{-1}\sum_{t=1}^{T} m_t - \mathrm{E}(m_t) \right| \lesssim_{\mathrm{P}} T^{-1/2}.$$

$$(4) \quad \left\| T^{-1}\sum_{t=1}^{T} r_t - \mathrm{E}(r_t) \right\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} (\log N)^{1/2} T^{-1/2}.$$

Finally, we need the following assumption for establishing the convergence of the ridge-based SDF estimator. It ensures that all eigenvalues of $\beta\Sigma_v\beta^{\intercal}$ are well separated. This assumption shares the spirit with Assumption 14. A similar assumption has been adopted by, e.g., Wang and Fan [2017].

Assumption 18. *The eigenvalues of $\beta\Sigma_v\beta^{\intercal}$ are separated in the sense that*

$$(\lambda_j - \lambda_{j+1})/\lambda_j \geq \delta$$

for some constant $\delta > 0$, where $\lambda_j := \lambda_j(\beta\Sigma_v\beta^{\intercal})$ is the jth eigenvalue of $\beta\Sigma_v\beta^{\intercal}$.

2.5.3 Additional Theoretical Results

In this section, we present additional theoretical results.

2.5.3.1 Mimicking Portfolio Built From I_0

Proposition 13 establishes that test assets in a subset I_0 are adequate to serve as basis assets, building on which a mimicking portfolio can approximate the risk premium of any observable

factor in g_t.

Proposition 13. *Suppose that r_t and g_t follow (2.1) and (2.4), respectively, and that Assumption 16 holds. Then for any subset $I_0 \subset \langle N \rangle$, we have*

$$\mathrm{Cov}(g_t, r_{t,[I_0]})\mathrm{Cov}(r_{t,[I_0]})^{-1}\mathrm{E}(r_{t,[I_0]}) = \eta\gamma + O\left(1/\lambda_{\min}(\beta_{[I_0]}^{\intercal}\beta_{[I_0]})\right).$$

Next, we provide a result that sheds light on the effectiveness of out-of-sample R^2 as a criterion for tuning parameter selection. In the main text, we partition the complete dataset into two segments, one for training and the other for evaluation (testing). Within the training sample, we employ cross-validation to determine the optimal tuning parameters. A more detailed procedure is outlined in Appendix 2.5.5. For the sake of simplicity, our theoretical analysis is based on a "validation" procedure instead of "cross-validation." In this context, the phrase "out of sample" specifically refers to the validation sample, used to select the tuning parameters.

For each combination of tuning parameter values $\breve{q}$ and $\breve{p}$, the application of SPCA to the in-sample data produces factor estimates, with each estimate representing a portfolio. Consequently, this process gives rise to a mimicking portfolio for g_t, characterized by weights denoted as $w(\breve{p}, \breve{q}) \in \mathbb{R}^{d\times N}$. The expected return of this portfolio is thus estimated as $w(\breve{p}, \breve{q})\bar{r}$. We also write the matrix forms of de-meaned r_t and g_t ouf of sample as $\bar{R}_{\mathrm{oos}} \in \mathbb{R}^{N\times T_{\mathrm{oos}}}$ and $\bar{G}_{\mathrm{oos}} \in \mathbb{R}^{d\times T_{\mathrm{oos}}}$, where T_{oos} represents the sample size of out of sample data. The time series R^2 of the ith factor's hedging portfolio out of sample is thus given by

$$R_i^2(\breve{p}, \breve{q}) = 1 - \frac{\left\|(\bar{G}_{\mathrm{oos}})_{[i]} - (w(\breve{p}, \breve{q}))_{[i]}\bar{R}_{\mathrm{oos}}\right\|^2}{\left\|(\bar{G}_{\mathrm{oos}})_{[i]}\right\|^2}.$$

To derive theoretical results for parameter tuning using these R^2 values, we require additional assumptions about the underlying DGP out of sample. In essence, the following

assumption asserts that the DGP remains unchanged between the in-sample and out-of-sample contexts.

Assumption 19. *Assumptions 7, 8, and 10 hold when V, U, Z, and T are replaced by V_{oos}, U_{oos}, Z_{oos}, and T_{oos}, respectively.*

Moreover, we need the following assumption regarding the relationship between in-sample estimates and out-of-sample DGP.

Assumption 20. *U_{oos}, V_{oos}, Z_{oos}, and $\breve{w} = w(\breve{p}, \breve{q}) \in \mathbb{R}^{d\times N}$ constructed by in-sample data satisfy:*

$$\left\|\breve{w}_{[i]}U_{\text{oos}}\right\| \asymp_{\text{P}} \left\|\breve{w}_{[i]}\right\| T_{\text{oos}}^{1/2}, \quad \left\|\breve{w}_{[i]}U_{\text{oos}}A^{\intercal}\right\| \lesssim_{\text{P}} \left\|\breve{w}_{[i]}\right\| T_{\text{oos}}^{1/2},$$

for $A = V_{\text{oos}}, Z_{\text{oos}}, \iota_T^{\intercal}$ and $i \leq d$.

Assumption 20 shares the same spirit of Assumptions 11 - 13. However, the key distinction lies in the direct imposition of constraints on the relationship between $\breve{w}$ and out-of-sample data. Given that $\breve{w}$ is constructed solely from in-sample data, these conditions can be interpreted as restrictions on the dependence between in-sample and out-of-sample data.

It is important to note that the first equality effectively imposes both an upper bound and a lower bound on $\left\|\breve{w}_{[i]}U_{\text{oos}}\right\|$. In the special case that $\text{Cov}(u_{\text{oos},t}) = \Sigma_u$ and in-sample data is independent of the out-of-sample data, each element in $\breve{w}_{[i]}U_{\text{oos}}$ has variance $\breve{w}_{[i]}\Sigma_u\breve{w}_{[i]}^{\intercal}$, which is bounded within $\left\|\breve{w}_{[i]}\right\|^2 \lambda_{\min}(\Sigma_u)$ and $\left\|\breve{w}_{[i]}\right\|^2 \lambda_{\max}(\Sigma_u)$. Therefore, the first equality becomes a standard concentration result under the restriction $1 \lesssim \lambda_{\min}(\Sigma_u) \leq \lambda_{\max}(\Sigma_u) \lesssim 1$.

The next proposition shows that selecting tuning parameters using out of sample R^2 leads to consistent estimates of risk premia, both in sample and out of sample:

Proposition 14. *Suppose that the out-of-sample DGP also follows* (2.1) *and* (2.4)*, and satisfies Assumptions 19-20. In addition, let* p^* *and* q^* *be defined by*

$$(p^*, q^*) = \text{argmax}_{\breve{p} \leq p_{\max}, \breve{q} \in \mathcal{Q}} \sum_{i=1}^{d} R_i^2(\breve{p}, \breve{q}),$$

where $p_{\max}$ *is some finite upper bound on the number of factors, and* $\mathcal{Q} = \{N^{-\alpha_j} | j = 1, \cdots, n_q\}$*,* $0 \leq \alpha_1 \leq \alpha_2 \cdots < \alpha_{n_q} < 1$ *is a finite grid of tuning parameter values. If* $p_{\max} \geq p$*,* $N^{1-\alpha_{n_q}}/N_0 \to 0$ *and* $\log T/N^{1-\alpha_{n_q}} \to 0$*, under assumptions of Theorem 6, as* $T_{\text{oos}} \to \infty$*, we have* $w^* = w(p^*, q^*)$ *satisfies* $\|w^* \bar{r}_{\text{oos}} - \eta\gamma\| \xrightarrow{\text{P}} 0$*. In addition, if* $q^* N \log N = O(T)$*, we have* $\left\| \widehat{\gamma}_g^{SPCA} - \eta\gamma \right\| = \|w^* \bar{r} - \eta\gamma\| \xrightarrow{\text{P}} 0$*.*

2.5.3.2 The Case of Observable Factors

The theoretical setup in this paper does not assume any knowledge of the identities of the factors v_t in (2.1). If v_t corresponds to innovations of known and observable factors, denoted by f_t, say, the Fama-French five factors, our procedure can be greatly simplified. It is meaningful to study this case, because it is common in the empirical literature, albeit having perfect knowledge of the factor model is a rather strong assumption.

Suppose first that factors in f_t are tradable. If the factor of interest g_t is one of the factors in f_t (therefore also tradable), then we can estimate the risk premium of g_t by simply taking its time-series average. If g_t is either spanned by f_t or not tradable, then a simple time series regression of g_t onto the factors f_t can recover its loading, η, which along with the risk premia estimates of f_t by their time-series averages yields the risk premium estimate of g_t. These scenarios are simple, and do not require cross-sectional regressions.

If some of the observed factors in f_t are not tradable, say, GDP growth, then a cross-sectional regression is necessary, which effectively constructs the mimicking portfolios for the non-tradable factors. In this setting, a weak factor problem potentially arises as documented

in the literature, see, e.g., Kan and Zhang [1999], Kleibergen [2009]. To tackle this issue, one could adopt a simplified version of Algorithm 7, to supervise the construction of mimicking portfolios for each of the observed non-tradable factors (in this case GDP growth), while using residuals from the projection of test asset returns onto tradable factors as new test assets.

2.5.3.3 The Case of Unknown Zero-beta Rate

In the theoretical setup, we focus on the case where the zero-beta rate is known. When it is not known, we need to modify our SPCA procedure slightly. Suppose that the DGP of returns follows

$$r_t = \gamma_0 \iota + \beta\gamma + \beta v_t + u_t, \tag{2.22}$$

where γ_0 is the zero-beta rate, and ι is a vector of 1s.

To proceed, we multiply $\mathbb{M}_\iota = \mathbb{I}_N - N^{-1}\iota\iota^\intercal$, from the left on both sides of equation (2.22). This results in a similar form of (2.1):

$$\tilde{r}_t = \tilde{\beta}\gamma + \tilde{\beta} v_t + \tilde{u}_t,$$

where $\tilde{a}_t = \mathbb{M}_\iota a_t$, for $a = r, \beta$, and u. Subsequently, we can readily apply Algorithm 7 to the transformed returns, $\tilde{r}_t$. To better grasp the reasoning behind this adjustment, let us consider a one-factor scenario. Choosing assets with strong absolute correlations with $\tilde{r}_t$ amounts to selecting assets characterized by large magnitudes of $\tilde{\beta}$. This choice, in turn, leads to the selection of assets that exhibit high cross-sectional dispersion in their β values.

2.5.4 Additional Simulation Results

In this section, we present a scenario akin to situation c) in the simulation setting of the main text, with the key distinction being that the final factor in this scenario is purely random noise. To elaborate, the DGP of r_t is driven by the first three factors. However, econometricians, who lack knowledge of the true model, include these three factors along with a random noise variable in their attempt to estimate risk premia. This scenario resembles a setting extensively discussed by Kan and Zhang [1999] and Kleibergen [2009].

For the sake of comparison, both PLS and SPCA incorporate this random noise variable alongside the aforementioned three factors, considering them collectively as g_t. The histograms provided in Figure 2.4 depicting the estimated risk premiums associated with this noise factor reveal that SPCA, PCA, PLS, rpPCA, Lasso, and Ridge methods produce estimates around zero – the true value. The consistency arises because none of these methods entail a cross-sectional regression on the estimated beta of the noise factor. In contrast, the four-split and two-pass methods seem to display substantial variances in this context.

Finally, we investigate the statistical power of SPCA in strong and weak cases, respectively, and draw a comparative analysis with PCA. We adopt the setting in scenario c) of the main text, as the case of strong factors. To simulate a weak factor scenario, we simply replace $a = 0.5$ in c) by $a = 0.1$. We consider a null hypothesis that the risk premium of V is zero, whereas the true risk premium of V ranges from -0.01 to 0.01. In Figure 2.5, we present the rejection rates for both SPCA and PCA. The left panel demonstrates that when all factors are strong, SPCA and PCA yield almost identical results. However, the right panel indicates that SPCA exhibits greater power than PCA across most ranges of risk premium values. The rejection rate for SPCA is around 5% when the null hypothesis is true, and it escalates to 100% as the actual risk premium value diverges from zero.

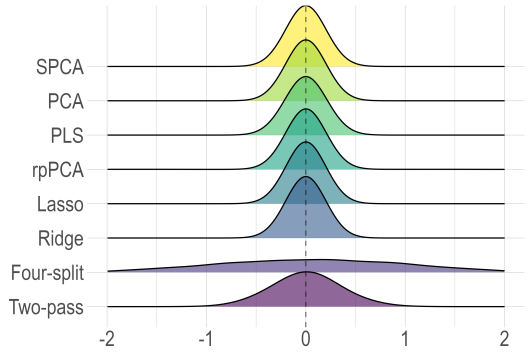
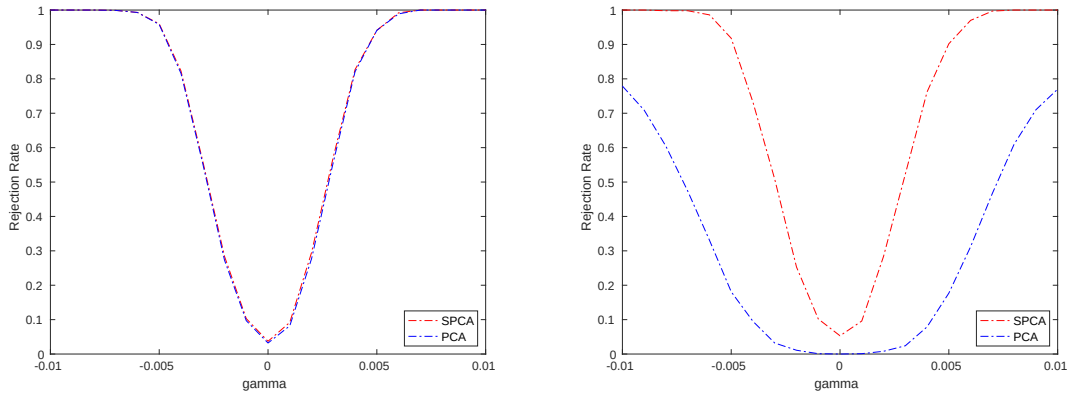


Figure 2.4: Histogram of Risk Premium Estimates of the noise factor

Note: The figure provides histograms of the risk premium estimates of the noise variable for eight estimators we compare, including SPCA, PCA, PLS, rpPCA, Lasso, Ridge, four-split, and the standard two-pass estimator. We simulate a model of returns driven by three strong factors, whereas g_t includes a pure noise variable, in addition to these three factors. All estimators attempt to estimate risk premia for the three factors and the noise variable altogether. We set $N = 1,000$ and $T = 240$. The number of Monte Carlo repetitions is 1,000.



(a) $a = 0.5$ (b) $a = 0.1$

Figure 2.5: Rejection Rate

Note: We simulate the model in scenario c) of the main text, using two different values for the parameter a: 0.5 for the strong factor case, and 0.1 for the weak factor case. We fix $N = 1,000$ and $T = 240$. The null hypothesis we test is that the risk premium of V is equal to zero, whereas its actual value varies between -0.01 to 0.01. We conduct a total of 1,000 Monte Carlo repetitions.

2.5.5 Implementation Details

In this section we detail the steps to compute the out-of-sample R^2 used in tuning parameter selection in the empirical analysis.

1. Inputs

   (a) Training sample data: returns for the N assets and target factor g_t, for the first half of the sample

   (b) Evaluation sample data: returns for the N assets and target factor g_t, for the second half of the sample

2. For each value of the number of factors p, execute the following steps:

   (a) Run 100 times the following cross-validation steps:

      i. Divide the training sample data into three folds (subsamples), chosen randomly without replacement

      ii. Choose the first of the three folds as validation (and the other two folds as training)

      iii. For each value of $\lfloor qN \rfloor$ between 100 and the maximum number of assets in the universe N in increments of 50:

         A. Estimate SPCA in the two training folds using p and $\lfloor qN \rfloor$

         B. Compute the R^2 of the mimicking portfolio in the validation fold

      iv. Repeat steps ii and iii using folds 2 and then 3 as validation samples (with the remaining two folds as training in each case)

      v. Find the tuning parameter $\lfloor qN \rfloor$ (and therefore the corresponding q) that maximizes the average R^2 in the validation samples across the three folds

   (b) Choose $\lfloor qN \rfloor$ as the median of the 100 tuning choices obtained across the cross-validation runs in (a)

(c) Estimate SPCA in the training sample using the tuning parameters p and the choice of $\lfloor qN \rfloor$ from (b).

(d) Compute the out-of-sample R^2 achieved by the SPCA mimicking portfolio estimated in (c) in the evaluation sample.

(e) Repeat (a)-(d) for every value of p.

2.6 Mathematical Proofs

2.6.1 Proofs from Section 2.2.2

2.6.1.1 Proof of Proposition 7

Proof. Note that for any orthogonal matrix $\Gamma \in \mathbb{R}^{N\times N}$, the estimators based on PCA, PLS and Ridge on $R' = \Gamma R$ are the same as those based on R. Thus, without loss of generality, we can assume $\beta = (\lambda^{1/2}, 0, \cdots, 0)^\intercal$, where $\lambda = \|\beta\|^2$. The same simplifying assumption is adopted in the proofs of Propositions 7, 10, and 11. Also, since $z_t = 0$, $\bar{G} = \eta\bar{V}$.

We start with $\widehat{\gamma}_g^{PCA}$. We write $\bar{R}$ in the following form:

$$\bar{R} = \beta\bar{V} + \bar{U} = \begin{pmatrix} \sqrt{\lambda}\bar{V} + \bar{U}_1 \\ \bar{U}_2 \end{pmatrix}, \tag{2.23}$$

where $\bar{U}_1$ is the first row of $\bar{U}$ and $\bar{U}_2$ contains the remaining rows. Correspondingly, we write the largest left singular vector of $\bar{R}$ as $\widehat{\varsigma} = (\widehat{\varsigma}_1, \widehat{\varsigma}_2^\intercal)^\intercal$, 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 $\widehat{\varsigma}$. Recall that in Algorithm 5, we denote $\widehat{\xi}$ and $\widehat{\varsigma}$ as the largest right and left singular vectors of $\bar{R}$ with the singular value $\sqrt{T\widehat{\lambda}}$, so that by simple algebra we have

$$\widehat{\varsigma}_1 = \frac{(\sqrt{\lambda}\bar{V} + \bar{U}_1)\widehat{\xi}}{\sqrt{T\widehat{\lambda}}}, \quad \widehat{\varsigma}_2 = \frac{\bar{U}_2\widehat{\xi}}{\sqrt{T\widehat{\lambda}}}. \tag{2.24}$$

Since the entries of U and V are i.i.d $\mathcal{N}(0,1)$, we have

$$|T^{-1}\bar{V}\bar{V}^\intercal - 1| = |T^{-1}V(\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)V^\intercal - 1| \leq |T^{-1}VV^\intercal - 1| + |\bar{v}|^2 \lesssim_{\mathrm{P}} T^{-1/2},$$

where we use large deviation results $|T^{-1}VV^\intercal - 1| \lesssim_{\mathrm{P}} T^{-1/2}$ and $|\bar{v}| \lesssim_{\mathrm{P}} T^{-1/2}$ in the last equation. This equation also implies that $\left\|\bar{V}\right\| - \sqrt{T} \lesssim_{\mathrm{P}} 1$.

Similarly, we can get $|T^{-1}\bar{U}_1\bar{U}_1^\intercal - 1| \lesssim_{\mathrm{P}} T^{-1/2}$ and $\left\|\bar{U}_1\right\| - \sqrt{T} \lesssim_{\mathrm{P}} 1$.

In addition, by Lemma A.1 in Wang and Fan [2017], we have $\left\|N^{-1}U^\intercal U - \mathbb{I}_T\right\| \lesssim_{\mathrm{P}} \sqrt{T/N}$, which leads to

$$\begin{aligned}\left\|N^{-1}\bar{U}^\intercal\bar{U} - (\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)\right\| &= \left\|(\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)(N^{-1}U^\intercal U - \mathbb{I}_T)(\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)\right\| \\ &\lesssim_{\mathrm{P}} \sqrt{T/N}.\end{aligned}$$

Next, by direct calculation using the above inequalities we obtain

$$\left\|\frac{\bar{V}^\intercal\bar{U}_1 + \bar{U}_1^\intercal\bar{V}}{T\sqrt{\lambda}} + \frac{\bar{U}^\intercal\bar{U} - N(\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)}{T\lambda}\right\| \lesssim_{\mathrm{P}} \frac{1}{\sqrt{\lambda}} + \frac{\sqrt{NT}}{T\lambda} \lesssim_{\mathrm{P}} \frac{1}{\sqrt{\lambda}}.$$

Together with (2.23), we have

$$\left\|\frac{\bar{R}^\intercal\bar{R}}{T\lambda} - \frac{\bar{V}^\intercal\bar{V}}{T} - \frac{N(\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)}{T\lambda}\right\| \lesssim_{\mathrm{P}} \frac{1}{\sqrt{\lambda}}. \tag{2.25}$$

Because of this result, to study the eigenstructure of $\bar{R}^\intercal\bar{R}/(T\lambda)$, we need analyze the eigenstructure of

$$M := \frac{\bar{V}^\intercal\bar{V}}{T} + \frac{N(\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)}{T\lambda} = \frac{\bar{V}^\intercal\bar{V}}{T} + \tilde{B}(\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal),$$

where $\tilde{B} = N/(T\lambda)$ and the assumption of the proposition implies that $\tilde{B} \to B$ for a constant B.

Note that $\bar{V}\iota_T = 0$, the eigenvalues of M can be explicitly given by:

$$\lambda_i = \begin{cases} T^{-1}\bar{V}\bar{V}^\intercal + \tilde{B} & i = 1; \\ \tilde{B} & 2 \leq i \leq T-1; , \\ 0 & i = T. \end{cases} \tag{2.26}$$

and the first eigenvector is $\bar{V}^\intercal / \left\|\bar{V}^\intercal\right\|$. Since the largest eigenvalue of $\bar{R}^\intercal\bar{R}/(T\lambda)$ is $\widehat{\lambda}/\lambda$ with its corresponding eigenvector $\widehat{\xi}$, Weyl's theorem yields that

$$\frac{\widehat{\lambda}}{\lambda} = \frac{\bar{V}\bar{V}^\intercal}{T} + \tilde{B} + O_{\mathrm{P}}\left(\frac{1}{\sqrt{\lambda}}\right) = 1 + \tilde{B} + O_{\mathrm{P}}\left(\frac{1}{\sqrt{\lambda}} + \frac{1}{\sqrt{T}}\right), \tag{2.27}$$

and the sin-theta theorem in Davis and Kahan [1970] implies that

$$\left\|\mathbb{P}_{\bar{V}^\intercal} - \mathbb{P}_{\widehat{\xi}}\right\| = \left\|\bar{V}^\intercal(\bar{V}\bar{V}^\intercal)^{-1}\bar{V} - \widehat{\xi}\widehat{\xi}^\intercal\right\| \lesssim_{\mathrm{P}} \frac{1}{\sqrt{\lambda}}, \tag{2.28}$$

which implies that $(\bar{V}\bar{V})^{-1}(\bar{V}\widehat{\xi})^2 = \widehat{\xi}^\intercal\bar{V}^\intercal(\bar{V}\bar{V})^{-1}\bar{V}\widehat{\xi} = 1 + O_{\mathrm{P}}(\lambda^{-1/2} + T^{-1/2})$. Together with $|T^{-1}\bar{V}\bar{V}^\intercal - 1| \lesssim T^{-1/2}$, we have

$$\frac{|\bar{V}\widehat{\xi}|}{\sqrt{T}} = 1 + O_{\mathrm{P}}\left(\frac{1}{\sqrt{\lambda}} + \frac{1}{\sqrt{T}}\right). \tag{2.29}$$

It is easy to observe that the sign of $\widehat{\xi}$ plays no role in the estimator $\widehat{\gamma}_g^{PCA}$, we can choose $\widehat{\xi}$ such that

$$\frac{\bar{V}\widehat{\xi}}{\sqrt{T}} = 1 + O_{\mathrm{P}}\left(\frac{1}{\sqrt{\lambda}} + \frac{1}{\sqrt{T}}\right). \tag{2.30}$$

Recall that the risk premium estimator is $\widehat{\gamma}_g^{PCA} = \widehat{\eta}\widehat{\gamma}$, where

$$\widehat{\eta} = \frac{\bar{G}\widehat{\xi}}{\sqrt{T}} \quad \text{and } \widehat{\gamma} = \frac{\widehat{\varsigma}^{\intercal}\bar{r}}{\sqrt{\widehat{\lambda}}}. \tag{2.31}$$

Using $\bar{G} = \eta\bar{V}$ and (2.30), we have

$$\widehat{\eta} = \eta + O_{\mathrm{P}}\left(\frac{1}{\sqrt{\lambda}} + \frac{1}{\sqrt{T}}\right). \tag{2.32}$$

Write

$$\widehat{\gamma} = \frac{\widehat{\varsigma}^{\intercal}\bar{r}}{\sqrt{\widehat{\lambda}}} = \frac{\widehat{\varsigma}^{\intercal}\beta(\gamma+\bar{v})}{\sqrt{\widehat{\lambda}}} + \frac{\widehat{\varsigma}^{\intercal}\bar{u}}{\sqrt{\widehat{\lambda}}} = \frac{\sqrt{\lambda}\widehat{\varsigma}_1}{\sqrt{\widehat{\lambda}}}(\gamma+\bar{v}) + \frac{\widehat{\varsigma}^{\intercal}\bar{u}}{\sqrt{\widehat{\lambda}}}, \tag{2.33}$$

where we use $\beta = (\sqrt{\lambda}, 0, \ldots, 0)^{\intercal}$ in the last step. Now we analyze the two terms on the right hand side of (2.33) one by one. For the first term, using (2.24), we have

$$\frac{\sqrt{\lambda}\widehat{\varsigma}_1}{\sqrt{\widehat{\lambda}}} = \frac{\lambda}{\widehat{\lambda}}\frac{(\bar{V} + \lambda^{-1/2}\bar{U}_1)\widehat{\xi}}{\sqrt{T}} = \frac{\lambda}{\widehat{\lambda}}\left(\frac{\bar{V}\widehat{\xi}}{\sqrt{T}} + \frac{\bar{U}_1\widehat{\xi}}{\sqrt{T\lambda}}\right).$$

Using (2.27) and (2.30) and $\left\|\bar{U}_1\right\| \lesssim_{\mathrm{P}} \sqrt{T}$, it follows that

$$\frac{\sqrt{\lambda}\widehat{\varsigma}_1}{\sqrt{\widehat{\lambda}}} = \frac{1}{1+\tilde{B}} + O_{\mathrm{P}}\left(\frac{1}{\sqrt{\lambda}} + \frac{1}{\sqrt{T}}\right). \tag{2.34}$$

For the second term in (2.33), using (2.24) again, we can write

$$\frac{\widehat{\varsigma}^{\intercal}\bar{u}}{\sqrt{\widehat{\lambda}}} = \frac{\widehat{\varsigma}_1 U_1 \iota_T}{T\sqrt{\widehat{\lambda}}} + \frac{\widehat{\varsigma}_2^{\intercal} U_2 \iota_T}{T\sqrt{\widehat{\lambda}}} = \frac{\widehat{\varsigma}_1 U_1 \iota_T}{T\sqrt{\widehat{\lambda}}} + \frac{\widehat{\xi}^{\intercal}(\mathbb{I}_T - T^{-1}\iota_T\iota_T^{\intercal})U_2^{\intercal}U_2\iota_T}{T^{3/2}\widehat{\lambda}}. \tag{2.35}$$

The condition that entries of U are independent $\mathcal{N}(0,1)$ implies that $\|U_1\iota_T\| \lesssim_{\mathrm{P}} \sqrt{T}$, with $\widehat{\lambda}/\lambda \xrightarrow{\mathrm{P}} 1 + B$ as shown in (2.27), the first term in (2.35) is of order $O_{\mathrm{P}}(T^{-1/2}\lambda^{-1/2})$. For

the second term in (2.35), using $\left\|(N-1)^{-1}U_2^\intercal U_2 - \mathbb{I}_T\right\| \lesssim_{\mathrm{P}} \sqrt{T/N}$, we have

$$\begin{aligned}
&|\frac{\widehat{\xi}^\intercal(\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)U_2^\intercal U_2\iota_T}{T^{3/2}\widehat{\lambda}}| \\
\leq&|\frac{(N-1)\widehat{\xi}^\intercal(\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)\iota_T}{T^{3/2}\widehat{\lambda}}| + \frac{N-1}{T\widehat{\lambda}}\left\|(N-1)^{-1}U_2^\intercal U_2 - \mathbb{I}_T\right\| \\
=&\frac{N-1}{T\widehat{\lambda}}\left\|(N-1)^{-1}U_2^\intercal U_2 - \mathbb{I}_T\right\| \lesssim_{\mathrm{P}} \frac{1}{\sqrt{\lambda}},
\end{aligned}$$

which leads to $|\widehat{\lambda}^{-1/2}\widehat{\varsigma}^\intercal\bar{u}| \lesssim_{\mathrm{P}} \lambda^{-1/2}$. Plugging this and (2.34) into (2.33), we obtain

$$\widehat{\gamma} = \frac{\widehat{\varsigma}^\intercal\bar{r}}{\sqrt{\widehat{\lambda}}} = \frac{\gamma}{1+\tilde{B}} + O_{\mathrm{P}}\left(\frac{1}{\sqrt{\lambda}} + \frac{1}{\sqrt{T}}\right), \tag{2.36}$$

and thus $\widehat{\gamma}_g^{PCA} \xrightarrow{\mathrm{P}} (1+B)^{-1}\eta\gamma$ by (2.32), (2.36) and $\tilde{B} \to B$. □

2.6.1.2 Proof of Proposition 8

Proof. Consider the set $I = \{|\beta_{[i]}| \geq \beta_{\{qN\}}\}$, where $|\beta|_{\{qN\}}$ is the (qN)th largest value in $\left\{|\beta_{[i]}|\right\}_{i\in\langle N\rangle}$. Since

$$T^{-1}\bar{R}\bar{G}^\intercal - \beta\eta^\intercal = \beta\left(T^{-1}\bar{V}\bar{V}^\intercal - 1\right)\eta^\intercal + T^{-1}\bar{U}\bar{V}^\intercal\eta^\intercal + T^{-1}\beta\bar{V}\bar{Z}^\intercal + T^{-1}\bar{U}\bar{Z}^\intercal,$$

we have

$$\begin{aligned}
\left\|T^{-1}\bar{R}\bar{G}^\intercal - \beta\eta^\intercal\right\|_{\mathrm{MAX}} \lesssim& \|\beta\|_{\mathrm{MAX}}\left|T^{-1}\bar{V}\bar{V}^\intercal - 1\right|\|\eta\| + T^{-1}\left\|\bar{U}\bar{V}^\intercal\right\|_{\mathrm{MAX}}\|\eta\| \\
&+ T^{-1}\|\beta\|_{\mathrm{MAX}}\left\|\bar{V}\bar{Z}^\intercal\right\| + T^{-1}\left\|\bar{U}\bar{Z}^\intercal\right\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} (\log N)^{1/2}T^{-1/2}.
\end{aligned}$$

In other words, the difference between $T^{-1}\bar{R}\bar{G}^\intercal$ and $\beta\eta^\intercal$ for all test assets is bounded by $O_{\mathrm{P}}\left((\log N)^{1/2}T^{-1/2}\right)$, which is $o_{\mathrm{P}}(1)$ under our assumption.

On the other hand, with the assumption that $\|\beta\|_{\mathrm{MAX}} \lesssim 1$ and the definition of $|\beta|_{\{qN\}}$,

we have $\left\|\beta_{[I_0]}\right\|^2 \lesssim qN + (N_0 - qN)|\beta|^2_{\{qN\}}$. Together with the assumption that $qN/N_0 \to 0$ and $\left\|\beta_{[I_0]}\right\| \asymp \sqrt{N_0}$, it leads to $|\beta|^2_{\{qN\}} \gtrsim \left\|\beta_{I_0}\right\|^2 / N_0 \asymp 1$. Then, with the assumption that $|\beta|_{\{qN+1\}} \leq (1+\delta)^{-1}|\beta|_{\{qN\}}$, we have that the difference between $|\beta|_{\{qN+1\}}$ and $|\beta|_{\{qN\}}$ should be at the same rate as $|\beta|_{\{qN\}} \gtrsim 1$, which is larger than the difference between $T^{-1}\bar{R}\bar{G}^\intercal$ and $\beta\eta^\intercal$. Therefore, with probability approaching one, we have $\widehat{I} = I$. In what follows, we only need consider the case of $\widehat{I} = I$.

Since $qN/N_0 \to 0$, by the definition of I, we have $\left\|\beta_{[I]}\right\| / \sqrt{|I|} \geq \left\|\beta_{[I_0]}\right\| / \sqrt{|I_0|}$. Together with the assumption that $\left\|\beta_{[I_0]}\right\| \asymp \sqrt{N_0}$, $\left\|\beta_{[I_0]}\right\| \to \infty$ and $|I| = qN \to \infty$, we have $|I|/(T\left\|\beta_{[I]}\right\|^2) \to 0$ and $\left\|\beta_{[I]}\right\| \to \infty$. Now compared to the case with PCA, the expansion on $\widehat{\gamma}_g^{SPCA}$ resembles that of (2.33), except for an extra term that depends on $\bar{Z}$ and the restriction of $\bar{r}$ on I:

$$\widehat{\gamma}_g^{SPCA} = \frac{\eta\bar{V}\widehat{\xi}}{\sqrt{T}}\frac{\widehat{\varsigma}^\intercal\bar{r}_{[I]}}{\sqrt{\widehat{\lambda}}} + \frac{\bar{Z}\widehat{\xi}}{\sqrt{T}}\frac{\widehat{\varsigma}^\intercal\bar{r}_{[I]}}{\sqrt{\widehat{\lambda}}}. \tag{2.37}$$

In restriction to the smaller set I, the first term matches exactly the case of $|I|/(T\left\|\beta_{[I]}\right\|^2) \to 0 = B$ in Proposition 7, which yields

$$\frac{\eta\bar{V}\widehat{\xi}}{\sqrt{T}}\frac{\widehat{\varsigma}^\intercal\bar{r}_{[I]}}{\sqrt{\widehat{\lambda}}} = \eta\gamma + o_{\mathrm{P}}(1).$$

We now analyze the second term in (2.37). As shown in (2.36), we have

$$\left\|\frac{\widehat{\varsigma}^\intercal\bar{r}_{[I]}}{\sqrt{\widehat{\lambda}}}\right\| \lesssim_{\mathrm{P}} 1,$$

so to prove that SPCA is consistent in this case, it is sufficient to show that $T^{-1/2}\left\|\bar{Z}\widehat{\xi}\right\| \xrightarrow{\mathrm{P}} 0$, where $\widehat{\xi}$ is the largest right singular vector of $\bar{R}_{[I]}$. Similar to the proof of (2.28) in Proposition 7, we can show that the difference between projection matrices, $\mathbb{P}_{\widehat{\xi}}$ and $\mathbb{P}_{\bar{V}^\intercal}$ is

small by sin-theta theorem. That is to say, we have $\left\|\widehat{\xi}\widehat{\xi}^{\intercal} - \bar{V}^{\intercal}(\bar{V}\bar{V}^{\intercal})^{-1}\bar{V}\right\| \xrightarrow{\mathrm{P}} 0$. Then, with the fact that

$$\left\|\bar{Z}\bar{V}^{\intercal}(\bar{V}\bar{V}^{\intercal})^{-1}\bar{V}\right\| \leq \left\|\bar{Z}\bar{V}^{\intercal}\right\| \left\|(\bar{V}\bar{V}^{\intercal})^{-1}\right\| \left\|\bar{V}\right\| \lesssim_{\mathrm{P}} T^{1/2} \times T^{-1} \times T^{1/2} \lesssim_{\mathrm{P}} 1,$$

we have $T^{-1/2}\left\|\bar{Z}\widehat{\xi}\widehat{\xi}^{\intercal}\right\| \xrightarrow{\mathrm{P}} 0$. Consequently,

$$T^{-1/2}\left\|\bar{Z}\widehat{\xi}\right\| = T^{-1/2}\left\|\bar{Z}\widehat{\xi}\widehat{\xi}^{\intercal}\widehat{\xi}\right\| \leq T^{-1/2}\left\|\bar{Z}\widehat{\xi}\widehat{\xi}^{\intercal}\right\| \left\|\widehat{\xi}\right\| \xrightarrow{\mathrm{P}} 0.$$

Hence, z_t does not affect the consistency of the SPCA estimator. This completes the proof.

□

2.6.1.3 Proof of Theorem 6

Proof. It is sufficient to consider the case $\Sigma_v = \mathbb{I}_p$. Otherwise, we can do transformation $V' = \Sigma_v^{-\frac{1}{2}}V$, $\beta'_{[I]} = \beta_{[I]}\Sigma_v^{\frac{1}{2}}$, $\eta' = \eta\Sigma_v^{\frac{1}{2}}$ and $\gamma' = \Sigma_v^{-\frac{1}{2}}\gamma$. All the Assumptions 7-14 still hold for the new V', $\beta'_{[I]}$. Therefore, we only need analyze the case of $\Sigma_v = \mathbb{I}_p$.

For notation simplicity, throughout the proofs of Theorems 6-8, we use $\widetilde{R}_{(k)} := \left(\bar{R}_{(k)}\right)_{[\widehat{I}_k]}$ to denote the matrix on which we perform SVD in each step of Algorithm 7. Similarly, we use $\widetilde{r}_{(k)} := \left(\bar{r}_{(k)}\right)_{[\widehat{I}_k]}$. The first left and right singular vectors of $\widetilde{R}_{(k)}$ are denoted by $\widehat{\varsigma}_{(k)}$ and $\widehat{\xi}_{(k)}$, while the largest singular value of $\widetilde{R}_{(k)}$ is denoted by $\sqrt{T\widehat{\lambda}_{(k)}}$. As a result, $\widehat{\lambda}_{(k)} = T^{-1}\left\|\widetilde{R}_{(k)}\right\|^2$ and

$$\widehat{\varsigma}_{(k)} = \frac{\widetilde{R}_{(k)}\widehat{\xi}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}, \quad \widehat{\xi}_{(k)} = \frac{\widetilde{R}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}. \tag{2.38}$$

Using the above notation, our estimated factor at k-th step is $\widehat{V}_{(k)} = \sqrt{T}\widehat{\xi}_{(k)}^{\intercal} \in \mathbb{R}^{1\times T}$, the risk premium of this factor is given by $\widehat{\gamma}_{(k)} = \widehat{\lambda}_{(k)}^{-1/2}\widehat{\varsigma}_{(k)}^{\intercal}\widetilde{r}_{(k)}$, the loading matrix of R on

this factor is $\widehat{\beta}_{(k)} = T^{-1/2}\bar{R}\widehat{\xi}_{(k)}$, and the loading of G on this factor is $\widehat{\eta}_{(k)} = T^{-1/2}\bar{G}\widehat{\xi}_{(k)}$. By footnote 24, we can use $\bar{G}$ instead of $\bar{G}_{(k)}$ in Algorithm 7 and throughout the proof. We denote $\widehat{\eta} = (\widehat{\eta}_{(1)}, \ldots, \widehat{\eta}_{(\tilde{p})})$ and $\widehat{\gamma} = (\widehat{\gamma}_{(1)}, \ldots, \widehat{\gamma}_{(\tilde{p})})^{\intercal}$, so the risk premium estimator is $\widehat{\gamma}_g^{SPCA} = \widehat{\eta}\widehat{\gamma}$.

To see more clear the relationship between $\widetilde{R}_{(k)}$ and $\bar{R}$, we define matrix $D_{(k)} \in \mathbb{R}^{|\widehat{I}_k| \times N}$ iteratively:

$$D_{(k)} = \mathbb{I}_{[\widehat{I}_k]} - \sum_{i=1}^{k-1} \bar{R}_{[\widehat{I}_k]}\widehat{\xi}_{(i)} \frac{\widehat{\varsigma}'_{(i)} D_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}$$

with $D_{(1)} = \mathbb{I}_{[\widehat{I}_1]}$. We can show by induction that $\widetilde{R}_{(k)} = D_{(k)}\bar{R}$. In fact, by Lemma 15, we have $\widehat{\xi}_{(i)}^{\intercal}\widehat{\xi}_{(j)} = 0$. With $\widehat{V}_{(i)} = \sqrt{T}\widehat{\xi}_{(i)}^{\intercal}$ and the definition of $\widetilde{R}_{(k)}$, we have

$$\widetilde{R}_{(k)} := \left(\bar{R}_{(k)}\right)_{[\widehat{I}_k]} = \bar{R}_{[\widehat{I}_k]} \prod_{i=1}^{k-1} \mathbb{M}_{\widehat{V}_{(i)}^{\intercal}} = \bar{R}_{[\widehat{I}_k]} \left(\mathbb{I}_T - \sum_{i=1}^{k-1} \widehat{\xi}_{(i)}\widehat{\xi}_{(i)}^{\intercal} \right), \tag{2.39}$$

for $k > 1$ and when $k = 1$, $\widetilde{R}_{(1)} = \bar{R}_{[\widehat{I}_1]} = \beta_{[\widehat{I}_1]}\bar{V} + \bar{U}_{[\widehat{I}_1]}$. Using (2.38), if $\widetilde{R}_{(i)} = D_{(i)}\bar{R}$ for $i < k$, we can write (2.39) as

$$\widetilde{R}_{(k)} = \bar{R}_{[\widehat{I}_k]} \left(\mathbb{I}_T - \sum_{i=1}^{k-1} \widehat{\xi}_{(i)}\widehat{\xi}_{(i)}^{\intercal} \right) = \bar{R}_{[\widehat{I}_k]} - \sum_{i=1}^{k-1} \bar{R}_{[\widehat{I}_k]}\widehat{\xi}_{(i)} \frac{\widehat{\varsigma}_{(i)}^{\intercal}\widetilde{R}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} = D_{(k)}\bar{R} \tag{2.40}$$

As $\widetilde{R}_{(1)} = \bar{R}_{[\widehat{I}_1]} = D_{(1)}\bar{R}$ holds immediately by the definition, we have $\widetilde{R}_{(k)} = D_{(k)}\bar{R}$ by induction. If we further define $\widetilde{\beta}_{(k)} = D_{(k)}\beta$ and $\widetilde{U}_{(k)} = D_{(k)}\bar{U}$, then $\widetilde{R}_{(k)}$ can be written in the form $\widetilde{R}_{(k)} = \widetilde{\beta}_{(k)}\bar{V} + \widetilde{U}_{(k)}$. Similarly, we can write

$$\widetilde{r}_{(k)} = \widetilde{\beta}_{(k)}(\gamma + \bar{v}) + \widetilde{u}_{(k)}, \tag{2.41}$$

where $\widetilde{r}_{(k)} = D_{(k)}\bar{r}$ and $\widetilde{u}_{(k)} = D_{(k)}\bar{u}$.

We also create similar representations for $\widetilde{G}_{(k)} := \bar{G}\prod_{i=1}^{k-1}\mathbb{M}_{\widehat{V}_{(i)}^{\intercal}}$. Specifically, we have

$$\begin{aligned}\widetilde{G}_{(k)} :=& \bar{G}\left(\mathbb{I}_T - \sum_{i=1}^{k-1}\widehat{\xi}_{(i)}\widehat{\xi}_{(i)}^{\intercal}\right)\\ =& \bar{G} - \sum_{i=1}^{k-1}\bar{G}\widehat{\xi}_{(i)}\frac{\widehat{\varsigma}_{(i)}^{\intercal}\widetilde{R}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} = \eta\bar{V} + \bar{Z} - \sum_{i=1}^{k-1}\bar{G}\widehat{\xi}_{(i)}\frac{\widehat{\varsigma}_{(i)}^{\intercal}\widetilde{\beta}_{(i)}\bar{V}}{\sqrt{T\widehat{\lambda}_{(i)}}} - \sum_{i=1}^{k-1}\bar{G}\widehat{\xi}_{(i)}\frac{\widehat{\varsigma}_{(i)}^{\intercal}\widetilde{U}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}\\ =& \left(\eta - \sum_{i=1}^{k-1}\bar{G}\widehat{\xi}_{(i)}\frac{\widehat{\varsigma}_{(i)}^{\intercal}\widetilde{\beta}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}\right)\bar{V} + \left(\bar{Z} - \sum_{i=1}^{k-1}\bar{G}\widehat{\xi}_{(i)}\frac{\widehat{\varsigma}_{(i)}^{\intercal}\widetilde{U}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}\right).\end{aligned}$$

In light of this equation, if we define

$$\widetilde{\eta}_{(k)} := \eta - \sum_{i=1}^{k-1}\bar{G}\widehat{\xi}_{(i)}\frac{\widehat{\varsigma}_{(i)}^{\intercal}\widetilde{\beta}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}, \quad \text{and} \quad \widetilde{Z}_{(k)} := \bar{Z} - \sum_{i=1}^{k-1}\bar{G}\widehat{\xi}_{(i)}\frac{\widehat{\varsigma}_{(i)}^{\intercal}\widetilde{U}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}, \tag{2.42}$$

$\widetilde{G}_{(k)}$ can be written as $\widetilde{G}_{(k)} = \widetilde{\eta}_{(k)}\bar{V} + \widetilde{Z}_{(k)}$.

To sum up, we have defined $\widetilde{R}_{(k)}, \widetilde{r}_{(k)}, \widetilde{\beta}_{(k)}, \widetilde{U}_{(k)}, \widetilde{u}_{(k)}, \widetilde{\eta}_{(k)}$ and $\widetilde{Z}_{(k)}$ at the kth step of the algorithm. Note that $\widetilde{\beta}_{(k)} \in \mathbb{R}^{|I_k|\times p}$ and $\widetilde{\eta}_{(k)} \in \mathbb{R}^{d\times p}$ can be viewed as the loading of $\widetilde{R}_{(k)}$ and $\widetilde{G}_{(k)}$ on $\bar{V}$, but they are not the same as the estimators defined in Algorithm 7, $\widehat{\beta}_{(k)} \in \mathbb{R}^{N\times 1}$ and $\widehat{\eta}_{(k)} \in \mathbb{R}^{d\times 1}$, which are the estimated loadings of R and G on the kth factor.

By Lemma 17, we have $\mathrm{P}(\widehat{I}_k = I_k) \to 1$ for $k \leq \tilde{p}$ and $\mathrm{P}(\widehat{p} = \tilde{p}) \to 1$. Thus, we can impose that $\widehat{I}_k = I_k$ for any k and $\widehat{p} = \tilde{p}$ in what follows. In addition, Lemma 16(ii) and Lemma 17(iii) imply that $\widehat{\lambda}_{(k)} \asymp qN$ and that $|I_k| = qN$. Therefore, the assumptions of Lemmas 19-22 hold.

Since our algorithm stops at $\tilde{p}$, it implies that at most $qN - 1$ test assets satisfy

$$T^{-1}\left\|\widetilde{G}_{(\tilde{p}+1)}\bar{R}_{[i]}^{\intercal}\right\|_{\mathrm{MAX}} = T^{-1}\left\|\left(\bar{R}_{(\tilde{p}+1)}\right)_{[i]}\bar{G}^{\intercal}\right\|_{\mathrm{MAX}} \geq c.$$

Let $\mathcal{S}$ denote the set of these assets. For asset $i \in \mathcal{S}$, we have

$$\left\|T^{-1}\widetilde{G}_{(\tilde{p}+1)}\bar{R}_{[i]}^{\intercal}\right\|_{\mathrm{F}}^{2} \lesssim \left\|T^{-1}\widetilde{G}_{(\tilde{p}+1)}\bar{R}_{[i]}^{\intercal}\right\|_{\mathrm{MAX}}^{2} \lesssim 1.$$

Consider the test assets in I_0, we have

$$\begin{aligned}\sum_{i\in I_0}\left\|T^{-1}\widetilde{G}_{(\tilde{p}+1)}\bar{R}_{[i]}^{\intercal}\right\|_{F}^{2} &= \sum_{i\in I_0\cap\mathcal{S}}\left\|T^{-1}\widetilde{G}_{(\tilde{p}+1)}\bar{R}_{[i]}^{\intercal}\right\|_{F}^{2} + \sum_{i\in I_0\cap\mathcal{S}^c}\left\|T^{-1}\widetilde{G}_{(\tilde{p}+1)}\bar{R}_{[i]}^{\intercal}\right\|_{F}^{2} \\ &\lesssim_{\mathrm{P}} qN + c^2 N_0 = o\left(N_0\right), \end{aligned} \tag{2.43}$$

where we use the the assumptions $c \to 0$ and $qN/N_0 \to 0$ in the last equation. Consequently, (2.43) leads to $\left\|T^{-1}\widetilde{G}_{(\tilde{p}+1)}\bar{R}_{[I_0]}^{\intercal}\right\| = o_{\mathrm{P}}(N_0^{1/2})$. Write

$$\widetilde{G}_{(\tilde{p}+1)}\bar{R}_{[I_0]}^{\intercal} = \widetilde{\eta}_{(\tilde{p}+1)}\bar{V}\bar{V}^{\intercal}\beta_{[I_0]} + \widetilde{\eta}_{(\tilde{p}+1)}\bar{V}\bar{U}_{[I_0]}^{\intercal} + \bar{Z}_{(\tilde{p}+1)}\bar{V}^{\intercal}\beta_{[I_0]} + \bar{Z}_{(\tilde{p}+1)}\bar{U}_{[I_0]}^{\intercal}. \tag{2.44}$$

Using (2.43), (2.44) and Lemma 21(i)(ii), we have

$$\left\|\widetilde{\eta}_{(\tilde{p}+1)}\left(\bar{V}\bar{V}^{\intercal}\beta_{[I_0]} + \bar{V}\bar{U}_{[I_0]}^{\intercal}\right)\right\| = o_{\mathrm{P}}\left(N_0^{1/2}T\right). \tag{2.45}$$

Also, using Assumption 12, Lemma 14(i) and Weyl's theorem, we have

$$\begin{aligned}|\sigma_p(\bar{V}\bar{V}^{\intercal}\beta_{[I_0]} + \bar{V}\bar{U}_{[I_0]}^{\intercal}) - \sigma_p(T\beta_{[I_0]})| &\le \left\|\bar{V}\bar{U}_{[I_0]}^{\intercal}\right\| + \left\|T^{-1}\bar{V}\bar{V}^{\intercal} - \mathbb{I}_p\right\|\left\|T\beta_{[I_0]}\right\| \\ &\lesssim_{\mathrm{P}} N_0^{1/2}T^{1/2}.\end{aligned}$$

Since Assumption 9 implies that $\sigma_p(\beta_{[I_0]}) \asymp N_0^{1/2}$, we have $\sigma_p(\bar{V}\bar{V}^{\intercal}\beta_{[I_0]}+\bar{V}\bar{U}_{[I_0]}^{\intercal}) \asymp_{\mathrm{P}} N_0^{1/2}T$. Using this result, (2.45) and the inequality $\left\|\widetilde{\eta}_{(\tilde{p}+1)}\left(\bar{V}\bar{V}^{\intercal}\beta_{[I_0]} + \bar{V}\bar{U}_{[I_0]}^{\intercal}\right)\right\| \ge \sigma_p(\bar{V}\bar{V}^{\intercal}\beta_{[I_0]} +$

$\bar{V}\bar{U}_{[I_0]}^{\intercal}) \left\| \widetilde{\eta}_{(\tilde{p}+1)} \right\|$, we have $\left\| \widetilde{\eta}_{(\tilde{p}+1)} \right\| \xrightarrow{\mathrm{P}} 0$. That is, by definition of $\tilde{\eta}_{(\tilde{p}+1)}$ in (2.42),

$$\left\| \eta - \sum_{i=1}^{\tilde{p}} \bar{G}\widehat{\xi}_{(i)} \frac{\widehat{\varsigma}_{(i)}^{\intercal} \widetilde{\beta}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \right\| = o_{\mathrm{P}}(1). \tag{2.46}$$

Multiplying (2.46) by γ from the right-hand side, we have

$$\left\| \eta\gamma - \sum_{i=1}^{\tilde{p}} \bar{G}\widehat{\xi}_{(i)} \frac{\widehat{\varsigma}_{(i)}^{\intercal} \widetilde{\beta}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \gamma \right\| = o_{\mathrm{P}}(1). \tag{2.47}$$

Recall that our final estimator of γ_g is

$$\widehat{\gamma}_g^{SPCA} = \sum_{i=1}^{\tilde{p}} \widehat{\eta}_{(i)} \widehat{\gamma}_{(i)} = \sum_{i=1}^{\tilde{p}} \bar{G}\widehat{\xi}_{(i)} \frac{\widehat{\varsigma}_{(i)}^{\intercal} \widetilde{r}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} = \sum_{i=1}^{\tilde{p}} \bar{G}\widehat{\xi}_{(i)} \frac{\widehat{\varsigma}_{(i)}^{\intercal} \widetilde{\beta}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} (\gamma + \bar{v}) + \sum_{i=1}^{\tilde{p}} \bar{G}\widehat{\xi}_{(i)} \frac{\widehat{\varsigma}_{(i)}^{\intercal} \widetilde{u}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}. \tag{2.48}$$

Combining (2.47) and (2.48), we have

$$\|\eta\gamma - \widehat{\eta}\widehat{\gamma}\| \leq \sum_{i=1}^{\tilde{p}} \left\| \bar{G}\widehat{\xi}_{(i)} \frac{\widehat{\varsigma}_{(i)}^{\intercal} \widetilde{\beta}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \bar{v} \right\| + \sum_{i=1}^{\tilde{p}} \left\| \bar{G}\widehat{\xi}_{(i)} \frac{\widehat{\varsigma}_{(i)}^{\intercal} \widetilde{u}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \right\| + o_{\mathrm{P}}(1). \tag{2.49}$$

Using $\left\| \bar{G} \right\| \lesssim_{\mathrm{P}} T^{1/2}$, Lemma 20(ii), Lemma 22(i) and the assumptions that $qN \to \infty$, we have

$$\left\| \bar{G}\widehat{\xi}_{(i)} \frac{\widehat{\varsigma}_{(i)}^{\intercal} \widetilde{\beta}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \bar{v} \right\| \leq \left\| \bar{G}\widehat{\xi}_{(i)} \right\| \left\| \frac{\widehat{\varsigma}_{(i)}^{\intercal} \widetilde{\beta}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \right\| \|\bar{v}\| = o_{\mathrm{P}}(1),$$

and

$$\left\| \bar{G}\widehat{\xi}_{(i)} \frac{\widehat{\varsigma}_{(i)}^{\intercal} \widetilde{u}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \right\| \leq \left\| \bar{G}\widehat{\xi}_{(i)} \right\| \left\| \frac{\widehat{\varsigma}_{(i)}^{\intercal} \widetilde{u}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \right\| = o_{\mathrm{P}}(1).$$

Plugging them into (2.49) completes the proof.

2.6.1.4 Proof of Theorem 7

To derive the asymptotic distribution, we need a more intricate analysis. As in the proof of Theorem 6, we impose that $\widehat{p} = \tilde{p}$ and $\widehat{I}_k = I_k$, since Lemma 17 shows that both events occur with probability approaching 1.

Recall that in Algorithm 7 the SPCA estimator is written as

$$\widehat{\gamma}_g^{SPCA} = \widehat{\eta}\widehat{\gamma} = \sum_{k=1}^{\widehat{p}} \widehat{\eta}_{(k)}\widehat{\gamma}_{(k)},$$

where $\widehat{p}$ is the number of factors selected and, with the notation defined in the proof of Theorem 6,

$$\widehat{\eta}_{(k)} = \frac{\bar{G}\widehat{\xi}_{(k)}}{\sqrt{T}} = \frac{\eta\bar{V}\widehat{\xi}_{(k)}}{\sqrt{T}} + \frac{\bar{Z}\widehat{\xi}_{(k)}}{\sqrt{T}}, \qquad \widehat{\gamma}_{(k)} = \frac{\widehat{\varsigma}_{(k)}^{\intercal}\widetilde{r}_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}} = \frac{\widehat{\varsigma}_{(k)}^{\intercal}\widetilde{\beta}_{(k)}(\gamma + \bar{v})}{\sqrt{\widehat{\lambda}_{(k)}}} + \frac{\widehat{\varsigma}_{(k)}^{\intercal}\widetilde{u}_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}}. \quad (2.50)$$

Denote $H_1 = (h_{11}, \ldots, h_{\widehat{p}1})$, $H_2 = (h_{12}, \ldots, h_{\widehat{p}2})$, where

$$h_{k1} = T^{-1/2}\bar{V}\widehat{\xi}_{(k)}, \quad h_{k2} = \widehat{\lambda}_{(k)}^{-1/2}\widetilde{\beta}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}. \quad (2.51)$$

Therefore, we can write (2.50) as

$$\widehat{\eta}_{(k)} - \eta h_{k1} = \frac{\bar{Z}\widehat{\xi}_{(k)}}{\sqrt{T}}, \quad \widehat{\gamma}_{(k)} - h_{k2}^{\intercal}(\gamma + \bar{v}) = \frac{\widehat{\varsigma}_{(k)}^{\intercal}\widetilde{u}_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}}. \quad (2.52)$$

Since $\widehat{\xi}_{(k)}$ and $\widehat{\varsigma}_{(k)}$ are the largest singular vectors of $\widetilde{R}_{(k)}$ with the singular value $\sqrt{T\widehat{\lambda}_{(k)}}$,

we have

$$\widehat{\varsigma}_{(k)} = \frac{\widetilde{R}_{(k)}\widehat{\xi}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}, \quad \widehat{\xi}_{(k)} = \frac{\widetilde{R}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}. \tag{2.53}$$

From (2.53), we have

$$\frac{\bar{Z}\widehat{\xi}_{(k)}}{\sqrt{T}} = \frac{\bar{Z}}{\sqrt{T}}\frac{\widetilde{R}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}} = \frac{\bar{Z}\bar{V}^{\intercal}}{T}\frac{\widetilde{\beta}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}} + \frac{\bar{Z}\widetilde{U}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}} = \frac{\bar{Z}\bar{V}^{\intercal}}{T}h_{k2} + \frac{\bar{Z}\widetilde{U}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}}.$$

Using Lemma 20(ii), we have

$$\left\| \frac{\bar{Z}\widetilde{U}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}} \right\| \lesssim_{\mathrm{P}} \frac{1}{T} + \frac{1}{qN}, \quad \left\| \frac{\widehat{\varsigma}_{(k)}^{\intercal}\widetilde{u}_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}} \right\| \lesssim_{\mathrm{P}} \frac{1}{T} + \frac{1}{qN}.$$

Then, along with (2.52) and Lemma 14(vi), the above equations lead to

$$\left\| \widehat{\eta} - \eta H_1 - \frac{ZV^{\intercal}}{T}H_2 \right\| \lesssim_{\mathrm{P}} \frac{1}{T} + \frac{1}{qN}, \tag{2.54}$$

and

$$\left\| \widehat{\gamma} - H_2^{\intercal}\gamma - H_2^{\intercal}\bar{v} \right\| \lesssim_{\mathrm{P}} \frac{1}{T} + \frac{1}{qN}. \tag{2.55}$$

Combining (2.54) and (2.55), with $\|H_1\| \lesssim_{\mathrm{P}} 1$, $\|H_2\| \lesssim_{\mathrm{P}} 1$ from Lemma 22 and Assumptions 7, 8, we have

$$\left\| \widehat{\eta}\widehat{\gamma} - \eta H_1 H_2^{\intercal}(\gamma + \bar{v}) - \frac{ZV^{\intercal}}{T}H_2 H_2^{\intercal}\gamma \right\| \lesssim_{\mathrm{P}} \frac{1}{T} + \frac{1}{qN}. \tag{2.56}$$

As shown in Lemma 16(iv), under the assumption that $\lambda_p(\eta^{\intercal}\eta) \gtrsim 1$, we have $\tilde{p} = p$. Together with $\mathrm{P}(\widehat{p} = \tilde{p}) \to 1$, we can impose that $\widehat{p} = p$ for derivations below. To analyze

$H_1 H_2^\intercal$ and $H_2 H_2^\intercal$ in (2.56), using Lemma 22 and the assumptions on q, we have

$$\left\|H_2^\intercal H_2 - \mathbb{I}_p\right\| \leq \left\|H_1^\intercal H_2 - \mathbb{I}_p\right\| + \|H_1 - H_2\| \, \|H_2\| \lesssim_{\mathrm{P}} T^{-1/2}. \tag{2.57}$$

Then, for the term $H_2 H_2^\intercal$, we have

$$\left\|H_2 H_2^\intercal - \mathbb{I}_p\right\| = \max_{1\leq i\leq p} |\lambda_i(H_2 H_2^\intercal) - 1| = \max_{1\leq i\leq p} |\lambda_i(H_2^\intercal H_2) - 1| = \left\|H_2^\intercal H_2 - \mathbb{I}_p\right\| \lesssim_{\mathrm{P}} T^{-1/2} \tag{2.58}$$

since H_2 is a $p \times p$ matrix.

For the term $H_1 H_2^\intercal$, by Lemma 22, we have

$$\left\|H_1^\intercal H_2 - \mathbb{I}_p\right\| \lesssim_{\mathrm{P}} \frac{1}{T} + \frac{1}{qN}. \tag{2.59}$$

In addition, we have

$$\sigma_p(H_2) \left\|H_2 H_1^\intercal - \mathbb{I}_p\right\| \leq \left\|(H_2 H_1^\intercal - \mathbb{I}_p) H_2\right\| = \left\|H_2 (H_1^\intercal H_2 - \mathbb{I}_p)\right\| \leq \|H_2\| \left\|H_1^\intercal H_2 - \mathbb{I}_p\right\|. \tag{2.60}$$

Since (2.57) implies that $\sigma_1(H_2)/\sigma_p(H_2) = \lambda_1(H_2 H_2^\intercal)^{1/2}/\lambda_p(H_2 H_2^\intercal)^{1/2} \lesssim_{\mathrm{P}} 1$, (2.59) and (2.60) give

$$\left\|H_1 H_2^\intercal - \mathbb{I}_p\right\| = \left\|H_2 H_1^\intercal - \mathbb{I}_p\right\| \leq \frac{\sigma_1(H_2)}{\sigma_p(H_2)} \left\|H_1^\intercal H_2 - \mathbb{I}_p\right\| \lesssim_{\mathrm{P}} \frac{1}{T} + \frac{1}{qN}. \tag{2.61}$$

Combining (2.56), (2.58), (2.61) and the assumption $q^{-1} N^{-1} T^{1/2} \to 0$, we obtain

$$\left\|\widehat{\widehat{\eta}}\widehat{\gamma} - \eta(\gamma + \bar{v}) - T^{-1} Z V^\intercal \gamma\right\| \lesssim_{\mathrm{P}} \frac{1}{T} + \frac{1}{qN}. \tag{2.62}$$

(2.62) implies that $\|\widehat{\widehat{\eta}}\widehat{\gamma} - \eta\gamma\| \lesssim_{\mathrm{P}} T^{-1/2} + (qN)^{-1}$. In addition, with the assumption

$q^{-1}N^{-1}T^{1/2} \to 0$, (2.62) becomes $\left\|\widehat{\widehat{\eta}\widehat{\gamma}} - \eta(\gamma + \bar{v}) - T^{-1}ZV^{\intercal}\gamma\right\| = o_{\mathrm{P}}(T^{-1/2})$. Using Delta method and Assumption 15, it is straightforward to obtain: $\sqrt{T}\left(\widehat{\widehat{\eta}\widehat{\gamma}} - \eta\gamma\right) \xrightarrow{d} \mathcal{N}\left(0, \Phi\right)$, where Φ is as defined in Theorem 7. □

2.6.2 Proofs from Section 2.2.3

2.6.2.1 Proof of Theorem 8

Proof. As shown in the proof of Theorem 7, we have $\mathrm{P}(\widehat{p} = p) \to 1$ and $\mathbb{P}(\widehat{I}_k = I_k) \to 1$ for $k \leq p$. Thus, we impose $\widehat{p} = \tilde{p} = p$ and $\widehat{I}_k = I_k$ below. Using the same notation as in the proof of Theorem 7 and (2.55), we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}|m_t - \widehat{m}_t|^2 = \frac{1}{T}\left\|\widehat{V}^{\intercal}\widehat{\gamma} - V^{\intercal}\gamma\right\|^2 &= \frac{1}{T}\left\|\sqrt{T}\widehat{\xi}(H_2^{\intercal}\gamma + O_{\mathrm{P}}(T^{-1/2})) - V^{\intercal}\gamma\right\|^2 \\ &= \frac{1}{T}\left\|\sqrt{T}\widehat{\xi}H_2^{\intercal}\gamma - \bar{V}^{\intercal}\gamma\right\|^2 + O_{\mathrm{P}}\left(T^{-1}\right), \qquad (2.63)\end{aligned}$$

where $\widehat{\xi} = (\widehat{\xi}_{(1)}, \ldots, \widehat{\xi}_{(p)})$.

Using (2.53), we can write

$$\sqrt{T}\widehat{\xi}_{(k)}h_{k2}^{\intercal} = \frac{\widetilde{R}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}}h_{k2}^{\intercal} = \frac{\bar{V}^{\intercal}\widetilde{\beta}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}}h_{k2}^{\intercal} + \frac{\widetilde{U}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}}h_{k2}^{\intercal}. \qquad (2.64)$$

Using Lemma 20(i), Lemma 22(i) and $\widehat{\lambda}_{(k)} \asymp_{\mathrm{P}} |I_k|$, $|I_k| = qN$, we can derive from (2.64) that

$$\sqrt{T}\widehat{\xi}_{(k)}h_{k2}^{\intercal} = \bar{V}^{\intercal}h_{k2}h_{k2}^{\intercal} + O_{\mathrm{P}}\left(q^{-1/2}N^{-1/2}T^{1/2} + T^{-1/2}\right).$$

That is,

$$\sqrt{T}\widehat{\xi}H_2^{\intercal} = \bar{V}^{\intercal}H_2H_2^{\intercal} + O_{\mathrm{P}}\left(q^{-1/2}N^{-1/2}T^{1/2} + T^{-1/2}\right). \tag{2.65}$$

Therefore, using (2.65), (2.58) and the assumptions on q, we have

$$\begin{aligned} T^{-1/2}\left\|\sqrt{T}\widehat{\xi}H_2^{\intercal}\gamma - \bar{V}^{\intercal}\gamma\right\| &\lesssim_{\mathrm{P}} T^{-1/2}\left\|\bar{V}^{\intercal}H_2H_2^{\intercal} - \bar{V}^{\intercal}\right\|\|\gamma\| + q^{-1/2}N^{-1/2} + T^{-1} \\ &\lesssim_{\mathrm{P}} T^{-1/2}\left\|\bar{V}\right\|\left\|H_2H_2^{\intercal} - \mathbb{I}_p\right\| + q^{-1/2}N^{-1/2} + T^{-1} \\ &\lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1/2}. \end{aligned}$$

Therefore, it follows from (2.63) that

$$\frac{1}{T}\sum_{t=1}^{T}|m_t - \widehat{m}_t|^2 = \frac{1}{T}\left\|\widehat{V}^{\intercal}\widehat{\gamma} - V^{\intercal}\gamma\right\|^2 \lesssim_{\mathrm{P}} \frac{1}{T} + \frac{1}{qN}.$$

In light of the assumptions on q, we can choose q such that $qN \gtrsim N_0/\log N_0$, which leads to

$$\frac{1}{T}\sum_{t=1}^{T}|m_t - \widehat{m}_t|^2 \lesssim_{\mathrm{P}} \frac{1}{T} + \frac{\log N_0}{N_0}.$$

□

2.6.2.2 Proof of Proposition 9

Proof. Write $\widetilde{\beta} = \Sigma_u^{-1/2}\beta\Sigma_v^{1/2}$, then by definition $\widetilde{m}_t$ can be written as

$$\widetilde{m}_t = 1 - \gamma^{\intercal}\beta^{\intercal}\Sigma_r^{-1}(\beta v_t + u_t) = 1 - \gamma^{\intercal}\Sigma_v^{-1/2}\tilde{\beta}^{\intercal}\left(\widetilde{\beta}\widetilde{\beta}^{\intercal} + \mathbb{I}_N\right)^{-1}(\widetilde{\beta}\Sigma_v^{-1/2}v_t + \Sigma_u^{-1/2}u_t), \tag{2.66}$$

or in matrix form

$$\widetilde{M} = 1 - \gamma^\intercal \beta^\intercal \Sigma_r^{-1}(\beta V + U) = 1 - \gamma^\intercal \Sigma_v^{-1/2} \tilde{\beta}^\intercal \left(\widetilde{\beta}\widetilde{\beta}^\intercal + \mathbb{I}_N\right)^{-1} (\widetilde{\beta}\Sigma_v^{-1/2} V + \Sigma_u^{-1/2} U), \tag{2.67}$$

where $\widetilde{M} = (\widetilde{m}_1, \ldots, \widetilde{m}_T)$, $V = (v_1, \ldots, v_T)$ and $U = (u_1, \ldots, u_t)$. Suppose that the SVD of $\widetilde{\beta}$ can be written as $\widetilde{\beta} = B\Lambda^{1/2}\Gamma$, where $B \in \mathbb{R}^{N\times p}$ and $\Gamma \in \mathbb{R}^{p\times p}$ are matrices of left and right singular vectors, $\Lambda^{1/2} = \mathrm{diag}(\widetilde{\lambda}_1^{1/2}, \cdots, \widetilde{\lambda}_p^{1/2})$ is a diagonal matrix and $\widetilde{\lambda}_i$ is the ith eigenvalue of $\widetilde{\beta}^\intercal\widetilde{\beta}$. Write $B = (b_1, \cdots, b_p)$, then $b_i^\intercal b_j = 0$ for $i \neq j$. Using the SVD of $\widetilde{\beta}$, we have

$$\widetilde{\beta}^\intercal \left(\widetilde{\beta}\widetilde{\beta}^\intercal + \mathbb{I}_N\right)^{-1} = \Gamma^\intercal \Lambda^{1/2}(\Lambda + \mathbb{I}_p)^{-1} B^\intercal.$$

Hence, we have

$$\begin{aligned} \left\| \widetilde{\beta}^\intercal \left(\widetilde{\beta}\widetilde{\beta}^\intercal + \mathbb{I}_N\right)^{-1} \widetilde{\beta} - \mathbb{I}_p \right\| &= \left\| \Gamma^\intercal \Lambda^{1/2}(\Lambda + \mathbb{I}_p)^{-1} \Lambda^{1/2}\Gamma - \mathbb{I}_p \right\| \\ &= \left\| \Lambda^{1/2}(\Lambda + \mathbb{I}_p)^{-1}\Lambda^{1/2} - \mathbb{I}_p \right\| \lesssim_{\mathrm{P}} \widetilde{\lambda}_p^{-1}, \end{aligned} \tag{2.68}$$

and

$$\left\| \widetilde{\beta}^\intercal \left(\widetilde{\beta}\widetilde{\beta}^\intercal + \mathbb{I}_N\right)^{-1} \Sigma_u^{-1/2} U \right\| = \left\| \Gamma^\intercal \Lambda^{1/2}(\Lambda + \mathbb{I}_p)^{-1} B^\intercal \Sigma_u^{-1/2} U \right\| \lesssim_{\mathrm{P}} \widetilde{\lambda}_p^{-1/2} \left\| B^\intercal \Sigma_u^{-1/2} U \right\|. \tag{2.69}$$

Since $\mathrm{Cov}(B^\intercal \Sigma_u^{-1/2} u_t) = \mathbb{I}_p$, we have $\mathrm{E}\left(\left\| B^\intercal \Sigma_u^{-1/2} U \right\|_{\mathrm{F}}^2 \right) = pT$, which leads to

$$\left\| B^\intercal \Sigma_u^{-1/2} U \right\| \leq \left\| B^\intercal \Sigma_u^{-1/2} U \right\|_{\mathrm{F}} \lesssim_{\mathrm{P}} T^{1/2}. \tag{2.70}$$

For the same reason, we have $\left\|\Sigma_v^{-1/2}V\right\| \lesssim_{\mathrm{P}} T^{1/2}$. Then, with Assumption 16, (2.67), (2.68), (2.69), and (2.70), we have

$$
\begin{aligned}
&\sqrt{\sum_{t=1}^{T}|m_t-\widetilde{m}_t|^2} \\
&\leq \left\|\gamma^\intercal\Sigma_v^{-1/2}\left(\widetilde{\beta}^\intercal\left(\widetilde{\beta}\widetilde{\beta}^\intercal+\mathbb{I}_N\right)^{-1}\widetilde{\beta}-\mathbb{I}_p\right)\Sigma_v^{-1/2}V\right\| + \left\|\gamma^\intercal\Sigma_v^{-1}\widetilde{\beta}^\intercal\left(\widetilde{\beta}\widetilde{\beta}^\intercal+\mathbb{I}_N\right)^{-1}\Sigma_u^{-1/2}U\right\| \\
&\lesssim \left\|\widetilde{\beta}^\intercal\left(\widetilde{\beta}\widetilde{\beta}^\intercal+\mathbb{I}_N\right)^{-1}\widetilde{\beta}-\mathbb{I}_p\right\|\left\|\Sigma_v^{-1/2}V\right\| + \left\|\widetilde{\beta}^\intercal\left(\widetilde{\beta}\widetilde{\beta}^\intercal+\mathbb{I}_N\right)^{-1}\Sigma_u^{-1/2}U\right\| \\
&\lesssim_{\mathrm{P}} T^{1/2}\widetilde{\lambda}_p^{-1/2},
\end{aligned}
$$

which in turn leads to

$$
\frac{1}{T}\sum_{t=1}^{T}|m_t-\widetilde{m}_t|^2 \lesssim_{\mathrm{P}} \widetilde{\lambda}_p^{-1},
$$

where

$$
\widetilde{\lambda}_p = \lambda_p\left(\Sigma_v^{1/2}\beta^\intercal\Sigma_u^{-1}\beta\Sigma_v^{1/2}\right) \geq \lambda_p(\beta\Sigma_v\beta^\intercal)\lambda_{\min}(\Sigma_u^{-1}) \asymp_{\mathrm{P}} \lambda_p(\beta^\intercal\beta)\lambda_{\max}^{-1}(\Sigma_u) \gtrsim \lambda_p(\beta^\intercal\beta),
$$

which concludes the proof. □

2.6.2.3 Proof of Theorem 9(a)

Proof. For Ridge SDF estimator $\widehat{m}_t$, we have

$$
\frac{1}{T}\sum_{t=1}^{T}|m_t-\widehat{m}_t|^2 = \frac{1}{T}\left\|\bar{R}^\intercal(\widehat{\Sigma}+\mu\mathbb{I}_N)^{-1}\bar{r}-V^\intercal\gamma\right\|^2. \tag{2.71}
$$

Recall that in the proof of Proposition 11, we have a condensed form of SVD on $\bar{R}$:

$$\bar{R} = \sqrt{T}\widehat{\varsigma}\widehat{\Lambda}^{1/2}\widehat{\xi}^{\intercal} + \sqrt{T}\widehat{\varsigma}_*\widehat{\Lambda}_*^{1/2}\widehat{\xi}_*^{\intercal},$$

where $\widehat{\Lambda}^{1/2}$ is the diagonal matrix of the first p singular values of $T^{-1/2}\bar{R}$ and $\widehat{\varsigma}$, $\widehat{\xi}$ are the corresponding left and right singular vectors. $\widehat{\varsigma}_* \in \mathbb{R}^{N\times K}$, $\widehat{\xi}_* \in \mathbb{R}^{T\times K}$ are the singular vectors corresponding to the remaining K nonzero singular values in $\widehat{\Lambda}_*^{1/2} \in \mathbb{R}^{K\times K}$, where $K = \min\{N, T-1\} - p$. Using this representation, (2.71) becomes

$$\begin{aligned}
\sqrt{\sum_{t=1}^{T}|m_t - \widehat{m}_t|^2} &= \left\|(\bar{V}^{\intercal}\beta^{\intercal} + \bar{U}^{\intercal})\widehat{\varsigma}(\widehat{\Lambda} + \mu I)^{-1}\widehat{\varsigma}^{\intercal}\bar{r} - V^{\intercal}\gamma + (\bar{V}^{\intercal}\beta^{\intercal} + \bar{U}^{\intercal})\widehat{\varsigma}_*(\widehat{\Lambda}_* + \mu I)^{-1}\widehat{\varsigma}_*^{\intercal}\bar{r}\right\| \\
&\le \left\|\bar{V}^{\intercal}\beta^{\intercal}\widehat{\varsigma}(\widehat{\Lambda} + \mu I)^{-1}\widehat{\varsigma}^{\intercal}\beta\gamma - \bar{V}^{\intercal}\gamma\right\| + \left\|\bar{V}^{\intercal}\beta^{\intercal}\widehat{\varsigma}(\widehat{\Lambda} + \mu I)^{-1}\widehat{\varsigma}^{\intercal}(\beta\bar{v} + \bar{u})\right\| \\
&\quad + \left\|\bar{U}^{\intercal}\widehat{\varsigma}(\widehat{\Lambda} + \mu I)^{-1}\widehat{\varsigma}^{\intercal}\bar{r}\right\| + \left\|\bar{V}^{\intercal}\beta^{\intercal}\widehat{\varsigma}_*(\widehat{\Lambda}_* + \mu I)^{-1}\widehat{\varsigma}_*^{\intercal}\bar{r}\right\| \\
&\quad + \left\|\bar{U}^{\intercal}\widehat{\varsigma}_*(\widehat{\Lambda}_* + \mu I)^{-1}\widehat{\varsigma}_*^{\intercal}\bar{r}\right\| + \left\|V^{\intercal}\gamma - \bar{V}^{\intercal}\gamma\right\| \qquad (2.72)
\end{aligned}$$

We analyze these terms one-by-one. Firstly, we consider the properties of $\widehat{\varsigma}$ and $\widehat{\xi}$. Let $\widehat{\varsigma}_k$ and $\widehat{\xi}_k$ denote the kth columns of $\widehat{\varsigma}$ and $\widehat{\xi}$, respectively. Note that $\widehat{\varsigma}_k$ and $\widehat{\xi}_k$ can be regarded as the $\widehat{\varsigma}_{(k)}$ and $\widehat{\xi}_{(k)}$ in our SPCA procedure with $I_k = \langle N\rangle$, where $\widehat{\varsigma}_k$ and $\widehat{\xi}_k$ are the singular vectors at the kth stage. This means we can reuse some of the proofs without repeating essentially the same arguments therein.

Similar to (2.51), we define

$$\tilde{h}_{k1} = T^{-1/2}\bar{V}\widehat{\xi}_k, \quad \tilde{h}_{k2} = \widehat{\lambda}_k^{-1/2}\beta^{\intercal}\widehat{\varsigma}_k, \qquad (2.73)$$

and $\tilde{H}_1 = (\tilde{h}_{11}, \ldots, \tilde{h}_{p1})$, $\tilde{H}_2 = (h_{12}, \ldots, \tilde{h}_{p2})$. Using Lemma 27, we can obtain

$$\left\|\tilde{H}_1\tilde{H}_2^\intercal - \mathbb{I}_p\right\| \lesssim_{\mathrm{P}} T^{-1} + \lambda_p^{-1}(T^{-1}N + 1), \quad \left\|\tilde{H}_1 - \tilde{H}_2\right\| \lesssim_{\mathrm{P}} T^{-1/2} + \lambda_p^{-1}(T^{-1}N + 1). \tag{2.74}$$

Using (2.74) and Lemma 27(i), we have $\left\|\tilde{H}_2\tilde{H}_2^\intercal - \mathbb{I}_p\right\| \le \left\|\tilde{H}_1\tilde{H}_2^\intercal - \mathbb{I}_p\right\| + \left\|\tilde{H}_1 - \tilde{H}_2\right\| \left\|\tilde{H}_2\right\| \lesssim_{\mathrm{P}}$ $T^{-1/2} + \lambda_p^{-1}(T^{-1}N + 1)$, which, by (2.73), is equivalent to

$$\left\|\beta^\intercal\widehat{\varsigma}\widehat{\Lambda}^{-1}\widehat{\varsigma}^\intercal\beta - \mathbb{I}_p\right\| \lesssim_{\mathrm{P}} \frac{1}{\sqrt{T}} + \frac{N+T}{T\lambda_p}. \tag{2.75}$$

Consequently, with Lemma 24 and $\left\|\beta^\intercal\widehat{\varsigma}\widehat{\Lambda}^{-1/2}\right\| = \left\|\tilde{H}_2\right\| \lesssim_{\mathrm{P}} 1$, we have

$$\begin{aligned}
&\left\|\beta^\intercal\widehat{\varsigma}\left(\widehat{\Lambda} + \mu I\right)^{-1}\widehat{\varsigma}^\intercal\beta - \mathbb{I}_p\right\| \\
&\le \left\|\beta^\intercal\widehat{\varsigma}\widehat{\Lambda}^{-1/2}\left(\widehat{\Lambda}^{1/2}\left(\widehat{\Lambda} + \mu I\right)^{-1}\widehat{\Lambda}^{1/2} - \mathbb{I}_p\right)\widehat{\Lambda}^{-1/2}\widehat{\varsigma}^\intercal\beta\right\| + \left\|\beta^\intercal\widehat{\varsigma}\widehat{\Lambda}^{-1}\widehat{\varsigma}^\intercal\beta - \mathbb{I}_p\right\| \\
&\le \left\|\beta^\intercal\widehat{\varsigma}\widehat{\Lambda}^{-1/2}\right\|^2 \left\|\widehat{\Lambda}^{1/2}\left(\widehat{\Lambda} + \mu I\right)^{-1}\widehat{\Lambda}^{1/2} - \mathbb{I}_p\right\| + \left\|\beta^\intercal\widehat{\varsigma}\widehat{\Lambda}^{-1}\widehat{\varsigma}^\intercal\beta - \mathbb{I}_p\right\| \\
&\lesssim_{\mathrm{P}} \frac{1}{\sqrt{T}} + \frac{N+T}{T\lambda_p} + \frac{\mu}{\lambda_p},
\end{aligned} \tag{2.76}$$

where we use $\left\|\widehat{\Lambda}^{1/2}\left(\widehat{\Lambda} + \mu I\right)^{-1}\widehat{\Lambda}^{1/2} - \mathbb{I}_p\right\| = \max_{j\le p}(\widehat{\lambda}_j + \mu)^{-1}\mu \lesssim_{\mathrm{P}} \lambda_p^{-1}\mu$ in the last step.

With $\left\|\bar{V}\right\| \lesssim_{\mathrm{P}} T^{1/2}$ from Lemma 14, it implies from (2.76) that the first term in (2.72) can be bounded:

$$\left\|\bar{V}^\intercal\beta^\intercal\widehat{\varsigma}(\widehat{\Lambda} + \mu I)^{-1}\widehat{\varsigma}^\intercal\beta\gamma - \bar{V}^\intercal\gamma\right\| \lesssim_{\mathrm{P}} 1 + \frac{N+T}{\sqrt{T}\lambda_p} + \frac{\mu\sqrt{T}}{\lambda_p}.$$

For the second term in (2.72), using Lemma 24, we have

$$\left\|\bar{V}^{\intercal}\beta^{\intercal}\widehat{\varsigma}(\widehat{\Lambda}+\mu I)^{-1}\widehat{\varsigma}^{\intercal}(\beta\bar{v}+\bar{u})\right\| \leq \left\|\bar{V}\right\| \left\|\beta^{\intercal}\widehat{\varsigma}\widehat{\Lambda}^{-1/2}\right\| \left\|\widehat{\Lambda}^{1/2}(\widehat{\Lambda}+\mu I)^{-1}\right\| \left\|\beta\bar{v}+\bar{u}\right\| \lesssim_{\mathrm{P}} \sqrt{\frac{N}{\lambda_p}}. \tag{2.77}$$

Next, recall that $\widehat{\varsigma}_*$ and $\widehat{\xi}_*$ are singular vectors of $\bar{R}$, we have

$$\bar{V}^{\intercal}\beta^{\intercal}\widehat{\varsigma}_* + \bar{U}^{\intercal}\widehat{\varsigma}_* = \bar{R}^{\intercal}\widehat{\varsigma}_* = \sqrt{T}\widehat{\xi}_*\widehat{\Lambda}_*^{1/2}. \tag{2.78}$$

By Weyl's theorem and Assumption 10, we have

$$|\sigma_j(T^{-1/2}\bar{R}) - \sigma_j(T^{-1/2}\beta\bar{V})| \leq T^{-1/2}\left\|\bar{R}-\beta\bar{V}\right\| = T^{-1/2}\left\|\bar{U}\right\| \lesssim_{\mathrm{P}} \sqrt{\frac{N}{T}}+1, \tag{2.79}$$

for $j \leq \min\{N,T\}$. Since $\mathrm{Rank}(T^{-1/2}\beta\bar{V}) \leq p$, we have $\sigma_j(T^{-1/2}\beta\bar{V}) = 0$ for $j > p$ and thus

$$\left\|\widehat{\Lambda}_*^{1/2}\right\| = \sigma_{p+1}(T^{-1/2}\bar{R}) \lesssim_{\mathrm{P}} \sqrt{\frac{N}{T}}+1. \tag{2.80}$$

Left multiplying (2.78) by $\bar{V}$, we obtain

$$\bar{V}\bar{V}^{\intercal}\beta^{\intercal}\widehat{\varsigma}_* = \sqrt{T}\bar{V}\widehat{\xi}_*\widehat{\Lambda}_*^{1/2} - \bar{V}\bar{U}^{\intercal}\widehat{\varsigma}_*. \tag{2.81}$$

Together with (2.80) and Assumption 12, we have

$$\left\|\beta^{\intercal}\widehat{\varsigma}_*\right\| \leq \left\|\left(\bar{V}\bar{V}^{\intercal}\right)^{-1}\right\| \left(\sqrt{T}\left\|\bar{V}\right\| \left\|\widehat{\Lambda}_*^{1/2}\right\| + \left\|\bar{V}\bar{U}^{\intercal}\right\|\right) \lesssim_{\mathrm{P}} \sqrt{\frac{N}{T}}+1, \tag{2.82}$$

and consequently,

$$\left\|\widehat{\varsigma}_*^\intercal \bar{r}\right\| \leq \left\|\widehat{\varsigma}_*^\intercal \beta\right\| \|\gamma + \bar{v}\| + \left\|\widehat{\varsigma}_*^\intercal \bar{u}\right\| \lesssim_{\mathrm{P}} \sqrt{\frac{N}{T}} + 1. \tag{2.83}$$

Using (2.82), (2.83), Lemma 26(iv) and $\left\|\bar{U}\right\| \lesssim_{\mathrm{P}} N^{1/2} + T^{1/2}$, we have

$$\left\|\beta^\intercal \widehat{\varsigma}_* (\widehat{\Lambda}_* + \mu I)^{-1} \widehat{\varsigma}_*^\intercal \bar{r}\right\| \leq \|\beta^\intercal \widehat{\varsigma}_*\| \left\|(\widehat{\Lambda}_* + \mu I)^{-1}\right\| \left\|\widehat{\varsigma}_*^\intercal \bar{r}\right\| \lesssim_{\mathrm{P}} \frac{N+T}{\mu T}, \tag{2.84}$$

and

$$\left\|\bar{U}^\intercal \widehat{\varsigma}_* (\widehat{\Lambda}_* + \mu I)^{-1} \widehat{\varsigma}_*^\intercal \bar{r}\right\| \leq \left\|\bar{U}\right\| \left\|(\widehat{\Lambda}_* + \mu I)^{-1}\right\| \left\|\widehat{\varsigma}_*^\intercal \bar{r}\right\| \lesssim_{\mathrm{P}} \frac{N+T}{\mu\sqrt{T}}. \tag{2.85}$$

Using Lemma 26(iii), we have

$$\left\|\widehat{\Lambda}^{-1/2} \widehat{\varsigma}^\intercal \bar{r}\right\| \lesssim_{\mathrm{P}} \left\|\widehat{\Lambda}^{-1/2} \widehat{\varsigma}^\intercal \beta\right\| + \left\|\widehat{\Lambda}^{-1/2} \widehat{\varsigma}^\intercal \bar{u}\right\| \lesssim_{\mathrm{P}} 1 + \frac{N+T}{T\lambda_p} \lesssim_{\mathrm{P}} 1,$$

where we use $\left\|\widehat{\Lambda}^{-1/2} \widehat{\varsigma}^\intercal \beta\right\| = \left\|\tilde{H}_2\right\| \lesssim_{\mathrm{P}} 1$. Then, with Lemma 26(iv), we have

$$\left\|\bar{U}^\intercal \widehat{\varsigma} (\widehat{\Lambda} + \mu I)^{-1} \widehat{\varsigma}^\intercal \bar{r}\right\| \leq \left\|\bar{U}^\intercal \widehat{\varsigma}\right\| \left\|(\widehat{\Lambda} + \mu I)^{-1} \widehat{\Lambda}^{1/2}\right\| \left\|\widehat{\Lambda}^{-1/2} \widehat{\varsigma}^\intercal \bar{r}\right\| \lesssim_{\mathrm{P}} \sqrt{\frac{T}{\lambda_p}} + \frac{N+T}{\sqrt{T}\lambda_p}. \tag{2.86}$$

Plugging (2.76), (2.77), (2.84), (2.85) and (2.86) into (2.72) and using $\left\|\bar{V} - V\right\| \lesssim_{\mathrm{P}} 1$, we obtain

$$\frac{1}{T}\sum_{t=1}^{T} |m_t - \widehat{m}_t|^2 \lesssim_{\mathrm{P}} \frac{\mu^2}{\lambda_p^2} + \frac{1}{T} + \frac{N+T}{T\lambda_p} + \frac{N^2+T^2}{\mu^2 T^2}.$$

With $\mu^2 \asymp T^{-1}\lambda_p(N+T)$, we achieve the best rate from the above bound:

$$\frac{1}{T}\sum_{t=1}^{T}|m_t - \widehat{m}_t|^2 \lesssim_{\mathrm{P}} \frac{1}{T} + \frac{N+T}{T\lambda_p}.$$

□

2.6.2.4 Proof of Theorem 9(b)

Proof. i. (Slow rate) Note that (2.14) is equivalent to a constrained optimization problem:

$$\widehat{b} = \arg\min_{b} \left\| \widehat{\Sigma}^{-1/2}\bar{r} - \widehat{\Sigma}^{1/2}b \right\|^2, \quad \text{subject to } \|b\|_1 \le \mu,$$

for some tuning parameter μ. This implies that the vector of the true SDF loadings, b, satisfies that

$$\left\| \widehat{\Sigma}^{-1/2}\bar{r} - \widehat{\Sigma}^{1/2}\widehat{b} \right\|^2 \le \left\| \widehat{\Sigma}^{-1/2}\bar{r} - \widehat{\Sigma}^{1/2}b \right\|^2 \quad \text{and} \quad \left\| \widehat{b} \right\|_1 \le \mu, \text{ for some } \mu \ge s.$$

Equivalently, expanding the left- and right-hand sides of the above we have

$$\widehat{b}^{\intercal}\widehat{\Sigma}\widehat{b} - b^{\intercal}\widehat{\Sigma}b \le 2(\widehat{b}-b)^{\intercal}\bar{r},$$

which leads to

$$(\widehat{b}-b)^{\intercal}\widehat{\Sigma}(\widehat{b}-b) \le 2(\widehat{b}-b)^{\intercal}(\bar{r}-\widehat{\Sigma}b) \le 2\left\|\widehat{b}-b\right\|_1 \left\|\bar{r}-\widehat{\Sigma}b\right\|_\infty.$$

With a tuning parameter $\mu \asymp s$, we have

$$(\widehat{b}-b)^{\intercal}\widehat{\Sigma}(\widehat{b}-b) \lesssim s\left\|\bar{r}-\widehat{\Sigma}b\right\|_\infty. \tag{2.87}$$

With Lemma 28, we have

$$\left\|\widehat{\Sigma}^{1/2}(\widehat{b}-b)\right\|^2 \lesssim_{\mathrm{P}} s\sqrt{\frac{\log N}{T}}. \tag{2.88}$$

Therefore, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\widehat{m}_t-\widetilde{m}_t\|^2 =&\frac{1}{T}\sum_{t=1}^{T}\left\|\widehat{b}^{\intercal}(r_t-\bar{r})-b^{\intercal}(r_t-\mathrm{E}(r_t))\right\|^2\\ \leq&\frac{2}{T}\sum_{t=1}^{T}\left\|(\widehat{b}-b)^{\intercal}(r_t-\bar{r})\right\|^2+\frac{2}{T}\sum_{t=1}^{T}\|b^{\intercal}(\bar{r}-\mathrm{E}(r_t))\|^2\\ \leq&2\left\|\widehat{\Sigma}^{1/2}(\widehat{b}-b)\right\|^2+2\|b\|_1^2\|\bar{r}-\mathrm{E}(r_t)\|_\infty^2 \lesssim_{\mathrm{P}} s\sqrt{\frac{\log N}{T}}+s^2\frac{\log N}{T}.\end{aligned}$$

Since $s\asymp\mu\gtrsim\|b\|_1$, plugging in the optimal rate choice $s\asymp\|b\|_1$, we complete the proof.

ii. (Fast rate) Since $\widehat{b}$ is the optimal solution of the minimization problem, it implies that

$$b^{\intercal}\widehat{\Sigma}b-2b^{\intercal}\bar{r}+b^{\intercal}\widehat{\Sigma}b+\mu\|b\|_1\geq\widehat{b}^{\intercal}\widehat{\Sigma}\widehat{b}-2\widehat{b}^{\intercal}\bar{r}+\widehat{b}^{\intercal}\widehat{\Sigma}\widehat{b}+\mu\|\widehat{b}\|_1. \tag{2.89}$$

Rewrite (2.89) as

$$(\widehat{b}-b)^{\intercal}\widehat{\Sigma}(\widehat{b}-b)\leq 2(\widehat{b}-b)^{\intercal}(\bar{r}-\widehat{\Sigma}b)+\mu(\|b\|_1-\|\widehat{b}\|_1). \tag{2.90}$$

If $\mu\geq 4\left\|\bar{r}-\widehat{\Sigma}b\right\|_\infty$, (2.90) becomes

$$\begin{aligned}\left\|\widehat{\Sigma}^{1/2}(\widehat{b}-b)\right\|^2\leq&2\left\|\widehat{b}-b\right\|_1\left\|\bar{r}-\widehat{\Sigma}b\right\|_\infty+\mu(\|b\|_1-\|\widehat{b}\|_1)\\ \leq&\frac{1}{2}\mu\left\|\widehat{b}-b\right\|_1+\mu(\|b\|_1-\|\widehat{b}\|_1).\end{aligned} \tag{2.91}$$

Let J denote the support of $\widehat{b}$, then (2.91) can be written as

$$\begin{aligned}\left\|\widehat{\Sigma}^{1/2}(\widehat{b}-b)\right\|^2 \leq & \frac{1}{2}\mu\left\|\widehat{b}_J - b_J\right\|_1 + \frac{1}{2}\mu\left\|\widehat{b}_{J^c}\right\|_1 + \mu\left\|\widehat{b}_J - b_J\right\|_1 - \mu\left\|\widehat{b}_{J^c}\right\|_1 \\ = & \frac{3}{2}\mu\left\|\widehat{b}_J - b_J\right\|_1 - \frac{1}{2}\mu\left\|\widehat{b}_{J^c}\right\|_1. \end{aligned} \tag{2.92}$$

Define $b^* = \widehat{b} - b$, then (2.92) implies that $3\left\|b^*_J\right\|_1 \geq \left\|b^*_{J^c}\right\|_1$, and we have

$$\frac{b^{*\intercal}(\Sigma - \widehat{\Sigma})b^*}{\|b^*\|^2} \leq \left\|\Sigma - \widehat{\Sigma}\right\|_{MAX}\frac{\|b^*\|_1^2}{\|b^*\|^2} \lesssim_{\mathrm{P}} \sqrt{\frac{\log N}{T}}\left(\frac{4\left\|b^*_J\right\|_1}{\left\|b^*_J\right\|}\right)^2 \lesssim_{\mathrm{P}} |J|\sqrt{\frac{\log N}{T}}.$$

Consequently, with the assumption $|J|\sqrt{\log N/T} \to 0$ and $\lambda_{\min}(\Sigma) \gtrsim 1$, we have

$$\frac{b^{*\intercal}\widehat{\Sigma}b^*}{\|b^*\|^2} = \frac{b^{*\intercal}\Sigma b^*}{\|b^*\|^2} + \frac{b^{*\intercal}(\Sigma - \widehat{\Sigma})b^*}{\|b^*\|^2} \gtrsim_{\mathrm{P}} 1.$$

Therefore, we have

$$\left\|\widehat{\Sigma}^{1/2}(\widehat{b}-b)\right\|^2 = b^{*\intercal}\widehat{\Sigma}b^* \gtrsim_{\mathrm{P}} \|b^*\|^2 \geq \left\|b^*_J\right\|^2 \geq |J|^{-1}\left\|b^*_J\right\|_1^2 = |J|^{-1}\left\|\widehat{b}_J - b_J\right\|_1^2. \tag{2.93}$$

Plugging (2.93) into (2.92), we have

$$\left\|\widehat{\Sigma}^{1/2}(\widehat{b}-b)\right\|^2 \leq \frac{3}{2}\mu\left\|\widehat{b}_J - b_J\right\|_1 \lesssim_{\mathrm{P}} \mu|J|^{1/2}\left\|\widehat{\Sigma}^{1/2}(\widehat{b}-b)\right\|.$$

Thus,

$$\left\|\widehat{\Sigma}^{1/2}(\widehat{b}-b)\right\|^2 \lesssim_{\mathrm{P}} \mu^2|J|. \tag{2.94}$$

Choosing $\mu = 4\left\|\bar{r} - \widehat{\Sigma}b\right\|_\infty$ and by Lemma 28, we obtain

$$\left\|\widehat{\Sigma}^{1/2}(\widehat{b}-b)\right\|^2 \lesssim_{\mathrm{P}} |J|\frac{\log N}{T}. \tag{2.95}$$

Similar to the slow rate case, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\widehat{m}_t-\widetilde{m}_t\|^2=&\frac{1}{T}\sum_{t=1}^{T}\left\|\widehat{b}^\intercal(r_t-\bar{r})-b^\intercal(r_t-\mathrm{E}(r_t))\right\|^2\\ \le&\frac{2}{T}\sum_{t=1}^{T}\left\|(\widehat{b}-b)^\intercal(r_t-\bar{r})\right\|^2+\frac{2}{T}\sum_{t=1}^{T}\|b^\intercal(\bar{r}-\mathrm{E}(r_t))\|^2\\ \le&2\left\|\widehat{\Sigma}^{1/2}(\widehat{b}-b)\right\|^2+2\|b^\intercal(\bar{r}-\mathrm{E}(r_t))\|^2\lesssim_{\mathrm{P}}\|b\|_0\frac{\log N}{T}.\end{aligned}$$

□

2.6.2.5 Proof of Theorem 10

Proof. To simplify the notation, we assume $\Sigma_v=\mathbb{I}_p$ without loss of generality and define a function $\mathrm{sr}(\cdot)$:

$$\mathrm{sr}(x_t)=\underset{0\neq b\in\mathbb{R}^N}{\arg\max}\frac{\mathrm{E}(b^\intercal x_t)}{\sqrt{\mathrm{Var}(b^\intercal x_t)}}.$$

In other words, $\mathrm{sr}(\mathrm{x_t})$ is the optimal Sharpe ratio we can get from x_t. It is well known that $\mathrm{sr}(x_t)=\sqrt{\mathrm{E}(x_t)^\intercal\mathrm{Cov}(x_t)^{-1}\mathrm{E}(x_t)}$ and the optimal value can be achieved by $b=\mathrm{Cov}(x_t)^{-1}\mathrm{E}(x_t)$, where $\mathrm{Cov}(x_t)^{-1}$ can be replaced by the Moore–Penrose inverse if it is singular.

Recall that our estimated factors are $\widehat{F}_{(k)}=\widehat{\varsigma}_{(k)}^\intercal\widetilde{R}_{(k)}/\sqrt{\widehat{\lambda}_{(k)}}=\widehat{\varsigma}_{(k)}^\intercal D_{(k)}\bar{R}/\sqrt{\widehat{\lambda}_{(k)}}$. Define $B=(b_1,\ldots,b_{\tilde{p}})^\intercal\in\mathbb{R}^{\tilde{p}\times N}$, where $b_i=D_{(k)}^\intercal\widehat{\varsigma}_{(k)}/\sqrt{\widehat{\lambda}_{(k)}}$. From Lemma 20(i), we have $\|B\|=O_{\mathrm{P}}(1)$. With the notation of B, the estimated factors can be written as $\widehat{F}_{(k)}=B\bar{R}$ and the SDF loading we get from SPCA is $\widehat{\gamma}^\intercal B$. Therefore, the out-of-sample Sharpe ratio for SPCA estimator is $\mathrm{sr}(\widehat{\gamma}^\intercal Br_t)$. To prove that $\lim_{N,T\to\infty}\mathrm{sr}(\widehat{\gamma}^\intercal Br_t)\geq\sqrt{\gamma^\intercal\mathbb{P}_{\eta^\intercal}\gamma}$, we first show that $\lim_{N,T\to\infty}\mathrm{sr}(Br_t)\geq\mathrm{sr}(\eta(v_t+\gamma))=\sqrt{\gamma^\intercal\mathbb{P}_{\eta^\intercal}\gamma}$.

The definition of $\mathrm{sr}(Br_t)$ implies that

$$\mathrm{sr}(Br_t) \geq \frac{\mathrm{E}(b^\intercal \widehat{\eta} B r_t)}{\sqrt{\mathrm{Var}(b^\intercal \widehat{\eta} B r_t)}},$$

for any $b \in \mathbb{R}^d$. Therefore, $\mathrm{sr}(Br_t) \geq \mathrm{sr}(\widehat{\eta} B r_t)$. Note that (2.46) can be written as $\|\widehat{\eta} B \beta - \eta\| = o_{\mathrm{P}}(1)$ and

$$\widehat{\eta} B r_t - \eta(v_t + \gamma) = (\widehat{\eta} B \beta - \eta)(v_t + \gamma) + \widehat{\eta} B u_t,$$

with $\|B\| = o_{\mathrm{P}}(1)$, we have $\|\widehat{\eta} B r_t - \eta(v_t + \gamma)\| = o_{\mathrm{P}}(1)$. Consequently, for any $b \in \mathbb{R}^d$, $\mathrm{E}(b^\intercal \widehat{\eta} B r_t) \to \mathrm{E}(b^\intercal \eta(v_t + \gamma))$ and $\mathrm{Cov}(b^\intercal \widehat{\eta} B r_t) \to \mathrm{Cov}(b^\intercal \eta(v_t + \gamma))$. Therefore, $\mathrm{sr}(\widehat{\eta} B r_t) \xrightarrow{\mathrm{P}} \mathrm{sr}(\eta(v_t + \gamma))$.

$\mathrm{sr}(\eta(v_t + \gamma))$ can be calculated by $\mathrm{E}(\eta(v_t + \gamma)) = \eta\gamma$ and $\mathrm{Cov}(\eta(v_t + \gamma)) = \eta\eta^\intercal$ we have

$$\begin{aligned} \mathrm{sr}(\eta(v_t + \gamma)) &= \sqrt{\mathrm{E}(\eta(v_t + \gamma))^\intercal \mathrm{Cov}(\eta(v_t + \gamma))^{-1} \mathrm{E}(\eta(v_t + \gamma))} \\ &= \sqrt{\gamma^\intercal \eta^\intercal (\eta\eta^\intercal)^{-1} \eta\gamma} = \sqrt{\gamma^\intercal \mathbb{P}_{\eta^\intercal} \gamma} \end{aligned}$$

Again, $(\eta\eta^\intercal)^{-1}$ here will be the Moore-Perose inverse if it is singular. To sum up, we have $\lim_{N,T\to\infty} \mathrm{sr}(Br_t) \geq \mathrm{sr}(\eta(v_t + \gamma)) = \sqrt{\gamma^\intercal \mathbb{P}_{\eta^\intercal} \gamma}$. Then, we will show that the optimal Sharpe ratio from Br_t can be achieved approximately by the portfolio $\widehat{\gamma}^\intercal B r_t$.

Note that $B\beta = H_2^\intercal$ from the definition of H_2 in (2.51), we have $\mathrm{E}(Br_t) = B\beta\gamma = H_2^\intercal \gamma$. With (2.55), it leads to $\left\|\widehat{\gamma} - H_2^\intercal \gamma\right\| = o_{\mathrm{P}}(1)$. For the covariance matrix, write

$$\left\|\mathrm{Cov}(Br_t) - \mathbb{I}_{\tilde{p}}\right\| \leq \left\|B\beta\beta^\intercal B^\intercal - \mathbb{I}_{\tilde{p}}\right\| + \|B\Sigma_u B^\intercal\| \leq \left\|H_2^\intercal H_2 - \mathbb{I}_{\tilde{p}}\right\| + \|B\|^2 \|\Sigma_u\| = o_{\mathrm{P}}(1).$$

where we use Lemma 27(ii),(iii), $\|B\| = o_{\mathrm{P}}(1)$ and the assumption $\|\Sigma_u\| \lesssim 1$ in the last

equation. Consequently, $\widehat{\gamma} \xrightarrow{\mathrm{P}} \mathrm{Cov}(Br_t)^{-1}\mathrm{E}(Br_t)$. With this equation, we have

$$\mathrm{sr}(\widehat{\gamma}^\intercal Br_t) = \frac{\mathrm{E}(\widehat{\gamma}^\intercal Br_t)}{\sqrt{\mathrm{Cov}(\widehat{\gamma}^\intercal Br_t)}} \xrightarrow{\mathrm{P}} \sqrt{\mathrm{E}(Br_t)^\intercal \mathrm{Cov}(Br_t)^{-1}\mathrm{E}(Br_t)} = \mathrm{sr}(Br_t).$$

The proof of the lower bound $\lim_{N,T\to\infty} \mathrm{sr}(\widehat{\gamma}^\intercal Br_t) \geq \sqrt{\gamma^\intercal \mathbb{P}_{\eta^\intercal}\gamma}$ is completed. For the upper bound, for the same reason as $\mathrm{sr}(Br_t) \geq \mathrm{sr}(\widehat{\eta}Br_t)$, it is straightforward to obtain $\mathrm{sr}(\widehat{\gamma}^\intercal Br_t) \leq \mathrm{sr}(r_t) \leq \sqrt{\gamma^\intercal\gamma}$. In the general case that $\Sigma_v \neq \mathbb{I}_p$, replace η, γ by $\eta\Sigma_v^{1/2}$, $\Sigma_v^{-1/2}\gamma$ to obtain the results. □

2.6.3 Proofs from Section 2.5.1

2.6.3.1 Proof of Proposition 10

Proof. Recall that in Section 2.5.1, we have

$$\widehat{\gamma}_g^{PLS} = \left\|\bar{G}\bar{R}^\intercal\bar{R}\right\|^{-2}\bar{G}\bar{R}^\intercal\bar{R}\bar{G}^\intercal\bar{G}\bar{R}^\intercal\bar{r}. \tag{2.96}$$

We analyze $\left\|\bar{G}\bar{R}^\intercal\bar{R}\right\|$, $\bar{G}\bar{R}^\intercal\bar{R}\bar{G}^\intercal$ and $\bar{G}\bar{R}^\intercal\bar{r}$ separately. Recall that from (2.25), we have

$$\left\|\frac{\bar{R}^\intercal\bar{R}}{T\lambda} - \frac{\bar{V}^\intercal\bar{V}}{T} - \tilde{B}(\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)\right\| \lesssim_{\mathrm{P}} \frac{1}{\sqrt{\lambda}},$$

where $\tilde{B} = N/(T\lambda)$ satisfies $\tilde{B} \to B$. Together with $\bar{G} = \eta\bar{V}$ and $\left\|\bar{G}\right\| \lesssim_{\mathrm{P}} \sqrt{T}$, we have

$$\begin{aligned}\frac{1}{T\lambda\sqrt{T}}\left\|\bar{G}\bar{R}^\intercal\bar{R}\right\| &= \frac{1}{\sqrt{T}}\left\|\bar{G}\left(\frac{\bar{V}^\intercal\bar{V}}{T} + \tilde{B}(\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)\right)\right\| + O_{\mathrm{P}}\left(\frac{1}{\sqrt{\lambda}}\right) \\ &= \frac{\eta}{\sqrt{T}}\left\|\frac{\bar{V}\bar{V}^\intercal\bar{V}}{T} + \tilde{B}\bar{V}\right\| + O_{\mathrm{P}}\left(\frac{1}{\sqrt{\lambda}}\right) \xrightarrow{\mathrm{P}} \eta(1+B),\end{aligned} \tag{2.97}$$

where we use $|T^{-1}\bar{V}\bar{V}^\intercal - 1| \lesssim_{\mathrm{P}} T^{-1/2}$ and $\left\|\bar{V}\right\| - \sqrt{T} \lesssim_{\mathrm{P}} 1$ in the last equation. For the same reason, by direct calculation we have

$$\begin{aligned}\frac{1}{T^2\lambda}\bar{G}\bar{R}^\intercal\bar{R}\bar{G}^\intercal &= \frac{1}{T}\bar{G}\left(\frac{\bar{V}^\intercal\bar{V}}{T} + \tilde{B}(\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)\right)\bar{G}^\intercal + O_{\mathrm{P}}\left(\frac{1}{\sqrt{\lambda}}\right) \\ &= \eta^2\frac{\bar{V}\bar{V}^\intercal\bar{V}\bar{V}^\intercal}{T^2} + \eta^2\tilde{B}\frac{\bar{V}\bar{V}^\intercal}{T} + O_{\mathrm{P}}\left(\frac{1}{\sqrt{\lambda}}\right) \xrightarrow{\mathrm{P}} \eta^2(1+B). \end{aligned} \tag{2.98}$$

Next, we write

$$\frac{1}{T\lambda}\bar{G}\bar{R}^\intercal\bar{r} = \frac{1}{T\lambda}\bar{G}\bar{R}^\intercal\beta(\gamma+\bar{v}) + \frac{1}{T\lambda}\bar{G}\bar{R}^\intercal\bar{u}. \tag{2.99}$$

We analyze these two terms in (2.99) separately. For the first term, we can write $\bar{R}$ in the form of (2.23) as in the proof of Proposition 7. Then, using $\left\|\bar{U}_1\right\| \lesssim_{\mathrm{P}} \sqrt{T}$ we have

$$\frac{1}{T\lambda}\bar{G}\bar{R}^\intercal\beta = \eta\frac{\bar{V}\bar{V}^\intercal}{T} + \eta\frac{\bar{V}\bar{U}_1^\intercal}{T\sqrt{\lambda}} = \eta\frac{\bar{V}\bar{V}^\intercal}{T} + O_{\mathrm{P}}\left(\frac{1}{\sqrt{\lambda}}\right). \tag{2.100}$$

For the second term in (2.99), we have

$$\begin{aligned}\frac{1}{T\lambda}\bar{G}\bar{R}^\intercal\bar{u} &= \eta\frac{1}{T^2\sqrt{\lambda}}\bar{V}\bar{V}^\intercal\bar{U}_1\iota_T + \eta\frac{1}{T^2\lambda}\bar{V}\bar{U}^\intercal U\iota_T = \eta\frac{1}{\sqrt{\lambda}}\frac{\bar{V}\bar{V}^\intercal}{T}\frac{\bar{U}_1\iota_T}{T} + \eta\frac{1}{T^2\lambda}\bar{V}U^\intercal U\iota_T \\ &= O_{\mathrm{P}}\left(\frac{1}{\sqrt{T\lambda}}\right) + \eta\frac{N}{T^2\lambda}\bar{V}\left(N^{-1}U^\intercal U - \mathbb{I}_T\right)\iota_T + \eta\frac{N}{T^2\lambda}\bar{V}\iota_T \\ &= O_{\mathrm{P}}\left(\frac{1}{\sqrt{T\lambda}}\right) + O_{\mathrm{P}}\left(\frac{1}{\sqrt{\lambda}}\right), \end{aligned} \tag{2.101}$$

where we use $\left\|N^{-1}U^\intercal U - \mathbb{I}_T\right\| \lesssim_{\mathrm{P}} \sqrt{T/N}$ and $\bar{V}\iota_T = 0$ in the last equation. Plugging (2.100) and (2.101) into (2.99), we have

$$\frac{1}{T\lambda}\bar{G}\bar{R}^\intercal\bar{r} = \eta\frac{\bar{V}\bar{V}^\intercal}{T}(\gamma+\bar{v}) + O_{\mathrm{P}}\left(\frac{1}{\sqrt{\lambda}}\right) \xrightarrow{\mathrm{P}} \eta\gamma. \tag{2.102}$$

Plug (2.97), (2.98), (2.102) into (2.96), we have

$$\widehat{\gamma}_g^{PLS} \xrightarrow{\text{P}} \frac{1}{\eta^2(1+B)^2}\eta^2(1+B)\eta\gamma = \frac{1}{1+B}\eta\gamma.$$

□

2.6.3.2 Proof of Proposition 11

Proof. Since $\text{Rank}(\bar{R}) \leq \min\{N, T-1\}$, and the assumptions of the proposition imply that $N/T \to \infty$, we thereby have a condensed SVD of $\bar{R}$ as

$$\bar{R} = \sqrt{T}(\widehat{\varsigma}, \widehat{\varsigma}_*)\widehat{\Lambda}^{1/2}(\widehat{\xi}, \widehat{\xi}_*)^\intercal = \sqrt{T}\widehat{\varsigma}\widehat{\lambda}^{1/2}\widehat{\xi}^\intercal + \sqrt{T}\widehat{\varsigma}_*\widehat{\Lambda}_*^{1/2}\widehat{\xi}_*^\intercal,$$

where $\widehat{\Lambda}^{1/2}$ is the diagonal matrix of $T-1$ singular values, $\widehat{\varsigma}$, $\widehat{\xi}$ are the left and right singular vectors corresponding to the largest singular value of $T^{-1/2}\bar{R}$, which is denoted by $\widehat{\lambda}^{1/2}$. In addition, $\widehat{\varsigma}_* \in \mathbb{R}^{N\times(T-2)}$ and $\widehat{\xi}_* \in \mathbb{R}^{T\times(T-2)}$ are the singular vectors corresponding to the rest $T-2$ nonzero singular values, $\widehat{\Lambda}_*^{1/2} \in \mathbb{R}^{(T-2)\times(T-2)}$. By direct calculation, we have

$$\begin{aligned}\sqrt{T}\bar{R}^\intercal\left(\bar{R}\bar{R}^\intercal + \mu I\right)^{-1} &= (\widehat{\xi}, \widehat{\xi}_*)\widehat{\Lambda}^{1/2}(\widehat{\Lambda} + T^{-1}\mu I)^{-1}(\widehat{\varsigma}, \widehat{\varsigma}_*)^\intercal \\ &= \frac{\widehat{\lambda}^{1/2}}{\widehat{\lambda} + T^{-1}\mu}\widehat{\xi}\widehat{\varsigma}^\intercal + \widehat{\xi}_*\widehat{\Lambda}_*^{1/2}\left(\widehat{\Lambda}_* + T^{-1}\mu I\right)^{-1}\widehat{\varsigma}_*^\intercal,\end{aligned}$$

and thus, with $\bar{G} = \eta\bar{V}$, the Ridge estimator can be written as

$$\begin{aligned}\widehat{\gamma}_g^{Ridge} = \bar{G}\bar{R}^\intercal\left(\bar{R}\bar{R}^\intercal + \mu I\right)^{-1}\bar{r} &= \frac{\widehat{\lambda}}{\widehat{\lambda} + T^{-1}\mu}\frac{\eta\bar{V}\widehat{\xi}}{\sqrt{T}}\frac{\widehat{\varsigma}^\intercal\bar{r}}{\sqrt{\widehat{\lambda}}} + \frac{\eta\bar{V}\widehat{\xi}_*}{\sqrt{T}}\widehat{\Lambda}_*^{1/2}\left(\widehat{\Lambda}_* + T^{-1}\mu\right)^{-1}\widehat{\varsigma}_*^\intercal\bar{r} \\ &= \frac{\widehat{\lambda}}{\widehat{\lambda} + T^{-1}\mu}\widehat{\gamma}_g^{PCA} + \frac{\eta\bar{V}\widehat{\xi}_*}{\sqrt{T}}\widehat{\Lambda}_*^{1/2}\left(\widehat{\Lambda}_* + T^{-1}\mu\right)^{-1}\widehat{\varsigma}_*^\intercal\bar{r}.\end{aligned} \tag{2.103}$$

Using (2.27) and the fact that $T^{-1}\lambda^{-1}\mu \to D$ and Proposition 7, we can show that the first term in (2.103) converges to $(1+B+D)^{-1}\eta\gamma$. With respect to the second term, as shown in (2.25), we have

$$\left\| \frac{\bar{R}^\intercal \bar{R}}{T\lambda} - \frac{\bar{V}^\intercal \bar{V}}{T} - \frac{N(\mathbb{I}_T - T^{-1}\iota_T \iota_T^\intercal)}{T\lambda} \right\| \lesssim_{\mathrm{P}} \frac{1}{\sqrt{\lambda}},$$

and the eigenvalues of

$$M = \frac{\bar{V}^\intercal \bar{V}}{T} + \frac{N(\mathbb{I}_T - T^{-1}\iota_T \iota_T^\intercal)}{T\lambda}$$

are given by (2.26), it then follows from Weyl's theorem that $\lambda_i(T^{-1}\lambda^{-1}\bar{R}^\intercal\bar{R}) = \tilde{B} + O_{\mathrm{P}}(\lambda^{-1/2})$ for $2 \le i \le T-1$. Note that $\widehat{\Lambda}_*^{1/2}\left(\widehat{\Lambda}_* + T^{-1}\mu\right)^{-1}$ is a $(T-2)\times(T-2)$ diagonal matrix and the ith element on the diagonal is

$$\frac{\lambda_{i+1}(T^{-1}\bar{R}^\intercal\bar{R})^{1/2}}{\lambda_{i+1}(T^{-1}\bar{R}^\intercal\bar{R}) + T^{-1}\mu} = \frac{1}{\sqrt{\lambda}} \frac{\lambda_{i+1}(T^{-1}\lambda^{-1}\bar{R}^\intercal\bar{R})^{1/2}}{\lambda_{i+1}(T^{-1}\lambda^{-1}\bar{R}^\intercal\bar{R}) + T^{-1}\lambda^{-1}\mu}.$$

Together with $T^{-1}\lambda^{-1}\mu \to D$, we have

$$\left\| \widehat{\Lambda}_*^{1/2}\left(\widehat{\Lambda}_* + T^{-1}\mu\right)^{-1} \right\| = \max_{1\le i\le T-2} \frac{\lambda_{i+1}(T^{-1}\bar{R}^\intercal\bar{R})^{1/2}}{\lambda_{i+1}(T^{-1}\bar{R}^\intercal\bar{R}) + T^{-1}\mu} \lesssim_{\mathrm{P}} \frac{1}{\sqrt{\lambda}}. \tag{2.104}$$

Also, with $\|\bar{u}\| \lesssim_{\mathrm{P}} \sqrt{N/T}$, we have

$$\left\|\widehat{\varsigma}_*^\intercal \bar{r}\right\| \le \left\|\widehat{\varsigma}_*^\intercal \beta(\gamma+\bar{v})\right\| + \left\|\widehat{\varsigma}_*^\intercal \bar{u}\right\| \le \|\beta(\gamma+\bar{v})\| + \|\bar{u}\| \lesssim_{\mathrm{P}} \sqrt{\lambda} + \sqrt{N/T} \lesssim_{\mathrm{P}} \sqrt{\lambda} \tag{2.105}$$

and

$$\left\|\frac{\bar{V}\widehat{\xi}_*}{\sqrt{T}}\right\|^2 = \left\|\frac{\bar{V}(\widehat{\xi},\widehat{\xi}_*)}{\sqrt{T}}\right\|^2 - \left\|\frac{\bar{V}\widehat{\xi}}{\sqrt{T}}\right\|^2 \leq \left\|\frac{\bar{V}}{\sqrt{T}}\right\|^2 - \left\|\frac{\bar{V}\widehat{\xi}}{\sqrt{T}}\right\|^2 = 1 + O_{\mathrm{P}}\left(\frac{1}{\sqrt{T}}\right) - \left\|\frac{\bar{V}\widehat{\xi}}{\sqrt{T}}\right\|^2 \lesssim_{\mathrm{P}} \frac{1}{\sqrt{\lambda}} + \frac{1}{\sqrt{T}}, \tag{2.106}$$

where we use (2.30) in the last inequality. Consequently, using (2.104), (2.105) and (2.106), we have

$$|\frac{\eta\bar{V}\widehat{\xi}_*}{\sqrt{T}}\widehat{\Lambda}_*^{1/2}\left(\widehat{\Lambda}_* + T^{-1}\mu\right)^{-1}\widehat{\varsigma}_*^{\intercal}\bar{r}| \leq \left\|\frac{\eta\bar{V}\widehat{\xi}_*}{\sqrt{T}}\right\| \left\|\widehat{\Lambda}_*^{1/2}\left(\widehat{\Lambda}_* + T^{-1}\mu\right)^{-1}\right\| \|\widehat{\varsigma}_*^{\intercal}\bar{r}\| \lesssim T^{-1/4} + \lambda^{-1/4}.$$

By comparing this with the limit of the first term in (2.103), we obtain

$$\widehat{\gamma}_g^{Ridge} \xrightarrow{\mathrm{P}} \frac{1}{1+B+D}\eta\gamma.$$

□

2.6.3.3 Proof of Proposition 12

Proof. By direct calculation, we can write

$$RR^{\intercal} + T\mu\bar{r}\bar{r}^{\intercal} = R\left(\mathbb{I}_T + \frac{\mu}{T}\iota_T\iota_T^{\intercal}\right)R^{\intercal} = R\left(\mathbb{I}_T + \frac{\tilde{\mu}}{T}\iota_T\iota_T^{\intercal}\right)^2 R^{\intercal}, \tag{2.107}$$

where $\tilde{\mu} = \sqrt{\mu+1} - 1$. Hence, the eigenvectors of $RR^{\intercal} + T\mu\bar{r}\bar{r}^{\intercal}$ are equivalent to the left singular vectors of $R\left(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^{\intercal}\right)$. Let $\widehat{\varsigma}$ and $\widehat{\xi}$ denote the largest left and right singular vectors of $R\left(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^{\intercal}\right)$. Note that $\widehat{\xi}$ can be viewed as the largest eigenvector of

$$(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^{\intercal})R^{\intercal}R(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^{\intercal}),$$

we analyze the eigenspace of this matrix first. Similar to (2.25) in the PCA case, we have the following approximation of $R^\intercal R$

$$\left\| \frac{R^\intercal R}{T\lambda} - \frac{\bar{V}^\intercal \bar{V}}{T} - \gamma \frac{\iota_T \bar{V} + \bar{V}^\intercal \iota_T^\intercal}{T} - \gamma^2 \frac{\iota_T \iota_T^\intercal}{T} - \frac{N}{T\lambda} \mathbb{I}_T \right\| \lesssim_{\mathrm{P}} \frac{1}{\sqrt{T}} + \frac{1}{\sqrt{\lambda}}, \tag{2.108}$$

by $|T^{-1}\bar{V}\bar{V}^\intercal - 1| \lesssim_{\mathrm{P}} T^{-1/2}$, $\left\|\bar{U}_1\right\| \lesssim_{\mathrm{P}} T^{1/2}$ and $\left\|N^{-1}\bar{U}^\intercal \bar{U} - (\mathbb{I}_T - T^{-1}\iota_T \iota_T^\intercal)\right\| \lesssim_{\mathrm{P}} \sqrt{T/N}$.

Then, with (2.108) and $N/(T\lambda) \to B$, we have

$$\left\| T^{-1}\lambda^{-1}(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T \iota_T^\intercal) R^\intercal R (\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T \iota_T^\intercal) - M^* \right\| = o_{\mathrm{P}}(1) \tag{2.109}$$

where the matrix M^* here is defined by

$$M^* := B\mathbb{I}_T + T^{-1}\bar{V}^\intercal \bar{V} + T^{-1}(1+\tilde{\mu})\gamma(\iota_T \bar{V} + \bar{V}^\intercal \iota_T^\intercal) + T^{-1}\Big((1+\tilde{\mu})^2\gamma^2 + \tilde{\mu}^2 B + 2\tilde{\mu}B\Big)\iota_T \iota_T^\intercal.$$

Recall that $\widehat{\xi}$ is the eigenvector of $T^{-1}\lambda^{-1}(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T \iota_T^\intercal) R^\intercal R (\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T \iota_T^\intercal)$, we can analyze the eigenspace of M^* first and then use sin-theta theorem to characterize $\widehat{\xi}$.

Firstly, the rank of $M^* - B\mathbb{I}_T$ is at most 2. Using the fact that $\bar{V}\iota_T = 0$, by direct calculation, we have the two nozero eigenvalues of $M^* - B\mathbb{I}_T$ are the solutions of the equation

$$(x - a_1)(x - a_3) - a_2^2 = 0, \tag{2.110}$$

where $a_1 = T^{-1}\left\|\bar{V}\right\|^2$, $a_2 = T^{-1/2}(1+\tilde{\mu})\gamma\left\|\bar{V}\right\|$ and $a_3 = (1+\tilde{\mu})^2\gamma^2 + \tilde{\mu}^2 B + 2\tilde{\mu}B$. Since the larger solution of (2.110) is

$$\frac{a_1 + a_3 + \sqrt{(a_1 - a_3)^2 + 4a_2^2}}{2} \geq a_1 > 0 \tag{2.111}$$

with probability 1, it is also the largest eigenvalue of $M^* - B\mathbb{I}_T$. In addition, the second largest eigenvalue of $M^* - B\mathbb{I}_T$ should be distinct with $\lambda_1(M^* - B\mathbb{I}_T)$. To see this, if the

second eigenvalue is the other solution of (2.110), we have $\lambda_1(M^* - B\mathbb{I}_T) - \lambda_2(M^* - B\mathbb{I}_T) = \sqrt{(a_1 - a_3)^2 + 4a_2^2} \geq \max\{2a_2, |a_1 - a_3|\} > 0$. If the second eigenvalue is 0 (in which case the second solution of the above equation must be negative), we have shown in (2.111) that $\lambda_1(M^* - B\mathbb{I}_T) - \lambda_2(M^* - B\mathbb{I}_T) = \lambda_1(M^* - B\mathbb{I}_T) \geq a_1 > 0$. In both cases, we have $\lambda_1(M^* - B\mathbb{I}_T) - \lambda_2(M^* - B\mathbb{I}_T) \geq \delta$ for some constant $\delta > 0$. Consequently,

$$\lambda_1(M^*) - \lambda_2(M^*) = \lambda_1(M^* - B\mathbb{I}_T) - \lambda_2(M^* - B\mathbb{I}_T) \geq \delta, \tag{2.112}$$

for some constant $\delta > 0$. Now we calculate the first eigenvector of M^*, which should also be the first eigenvector of $M^* - B\mathbb{I}_T$. We use $\tilde{\xi}$ to denote this eigenvector. Since we already know that the largest eigenvalue of $\lambda_1(M^* - B\mathbb{I}_T)$ is a solution of (2.110), which means that $\tilde{\xi}$ should be in the space spanned by $\bar{V}^\intercal$ and ι_T. Writing $\tilde{\xi} = K_1 \left\|\bar{V}\right\|^{-1} \bar{V}^\intercal + K_2 T^{-1/2} \iota_T$ and plugging the largest eigenvalue of $\lambda_1(M^* - B\mathbb{I}_T)$ of (2.111) into $\lambda_1(M - B\mathbb{I}_T)\tilde{\xi} = (M - B\mathbb{I}_T)\tilde{\xi}$, we directly get

$$\frac{K_2}{K_1} = \frac{\sqrt{(a_1 - a_3)^2 + 4a_2^2} + a_3 - a_1}{2a_2}, \tag{2.113}$$

which will pin down K_1 and K_2 because we also have $\left\|\tilde{\xi}\right\| = 1$.

Using $\left\|T^{-1}\lambda^{-1}(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^\intercal)R^\intercal R(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^\intercal) - M\right\| = o_{\mathrm{P}}(1)$, (2.112) and sin-theta theorem, we have

$$\left\|\mathbb{P}_{\widehat{\xi}} - \mathbb{P}_{\tilde{\xi}}\right\| \leq \frac{o_{\mathrm{P}}(1)}{\delta - o_{\mathrm{P}}(1)} = o_{\mathrm{P}}(1),$$

which implies that $|\tilde{\xi}^\intercal \widehat{\xi}| \xrightarrow{\mathrm{P}} 1$ and consequently,

$$\left\|\widehat{\xi} - K_1 \left\|\bar{V}\right\|^{-1} \bar{V}^\intercal - K_2 T^{-1/2} \iota_T\right\| = o_{\mathrm{P}}(1) \text{ or } \left\|\widehat{\xi} + K_1 \left\|\bar{V}\right\|^{-1} \bar{V}^\intercal + K_2 T^{-1/2} \iota_T\right\| = o_{\mathrm{P}}(1).$$

Since the sign of $\widehat{\xi}$ plays no role in the estimator $\widehat{\gamma}_g^{rpPCA}$, we can simply assume the former one.

Also, the relationship between singular vectors implies that

$$\widehat{F} = \widehat{\varsigma}^\intercal R = \left\| R(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^\intercal) \right\|^{-1} \widehat{\xi}^\intercal(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^\intercal)R^\intercal R. \tag{2.114}$$

With the approximation of $R^\intercal R$ in (2.108), $\bar{V}\iota_T = 0$, $T^{-1}\bar{V}\bar{V}^\intercal = 1 + O_{\mathrm{P}}(T^{-1/2})$ and $N/(T\lambda) \to B$, by direct calculation, we have

$$\left\| \left\|\bar{V}\right\|^{-1} \bar{V}(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^\intercal)R^\intercal R - \lambda T^{1/2}\Big((1+B)\bar{V} + \gamma\iota_T^\intercal\Big) \right\| = o_{\mathrm{P}}(\lambda T), \tag{2.115}$$

and

$$\left\| T^{-1/2}\iota_T^\intercal(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^\intercal)R^\intercal R - \lambda T^{1/2}(1+\tilde{\mu})\Big(\gamma\bar{V} + (\gamma^2+B)\iota_T^\intercal\Big) \right\| = o_{\mathrm{P}}(\lambda T). \tag{2.116}$$

Plugging (2.115), (2.116) and $\left\| \widehat{\xi} - K_1 \left\|\bar{V}\right\|^{-1} \bar{V}^\intercal + K_2 T^{-1/2}\iota_T \right\| = o_{\mathrm{P}}(1)$ into (2.114) we have

$$\left\| \left\| R(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^\intercal) \right\| \widehat{F} - \lambda T^{1/2}(L_1\bar{V} + L_2\iota_T^\intercal) \right\| = o_{\mathrm{P}}(\lambda T), \tag{2.117}$$

where

$$L_1 = K_1(1+B) + K_2(1+\tilde{\mu})\gamma, \qquad L_2 = K_1\gamma + K_2(1+\tilde{\mu})(\gamma^2+B). \tag{2.118}$$

It is easy to observe that scalar plays no role in the estimator $\widehat{\gamma}_g^{rpPCA}$, we can redefine

$$\widehat{F}^* = \lambda^{-1}T^{-1/2}L_1^{-1} \left\| R(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^\intercal) \right\| \widehat{F}$$

and use $\widehat{F}^*$ to create $\widehat{\gamma}_g^{rpPCA}$. Then, (2.117) becomes $\left\|\widehat{F}^* - \bar{V} - L_1^{-1}L_2\iota_T^{\intercal}\right\| = o_{\mathrm{P}}\left(T^{1/2}\right)$. Consequently,

$$\left\|\widehat{V} - \bar{V}\right\| = \left\|\widehat{F}^*(\mathbb{I}_T - T^{-1}\iota_T\iota_T^{\intercal}) - \bar{V}\right\| = o_{\mathrm{P}}\left(T^{1/2}\right), \quad \widehat{\gamma} = T^{-1}\widehat{F}^*\iota_T = L_1^{-1}L_2 + o_{\mathrm{P}}(1),$$

and

$$\widehat{\eta} = \bar{G}\widehat{V}^{\intercal}(\widehat{V}\widehat{V}^{\intercal})^{-1} = \eta\bar{V}\widehat{V}^{\intercal}(\widehat{V}\widehat{V}^{\intercal})^{-1} = \eta\Big(\bar{V}\bar{V}^{\intercal} + o_{\mathrm{P}}(T)\Big)\Big(\bar{V}\bar{V}^{\intercal} + o_{\mathrm{P}}(T)\Big)^{-1} = \eta + o_{\mathrm{P}}(1),$$

and the estimator $\widehat{\gamma}_g^{rpPCA} = \widehat{\eta}\widehat{\gamma} \xrightarrow{\mathrm{P}} \eta L_1^{-1}L_2$, where L_1 and L_2 are defined in (2.118).

In light of that $a_1 \xrightarrow{\mathrm{P}} 1$, $a_2 \xrightarrow{\mathrm{P}} (1+\tilde{\mu})\gamma$, $\tilde{\mu} = \sqrt{1+\mu} - 1$, $\widehat{\gamma}_g^{rpPCA} \xrightarrow{\mathrm{P}} \eta L_2/L_1$, (2.113) and the definitions of L_1 and L_2 in (2.118), we have

$$\widehat{\gamma}_g^{rpPCA} \xrightarrow{\mathrm{P}} w(1+B)^{-1}\eta\gamma + (1-w)\eta(\gamma + \gamma^{-1}B),$$

where

$$w = \frac{2+2B}{1+2B+\sqrt{(1-a)^2+4(1+\mu)\gamma}+a}, \qquad a = (1+\mu)(\gamma^2+B) - B. \tag{2.119}$$

□

2.6.4 Proofs from Section 2.5.3

2.6.4.1 Proof of Proposition 13

Proof. Using (2.1) and (2.4), we have

$$\mathrm{Cov}(g_t, r_{t,[I_0]})\mathrm{Cov}(r_{t,[I_0]})^{-1}\mathrm{E}(r_{t,[I_0]}) = \eta\Sigma_v\beta_{[I_0]}^{\intercal}(\beta_{[I_0]}\Sigma_v\beta_{[I_0]}^{\intercal} + \Sigma_{u,[I_0]})^{-1}\beta_{[I_0]}\gamma.$$

Therefore, it is sufficient to show that

$$\left\|\Sigma_v\beta_{[I_0]}^\intercal(\beta_{[I_0]}\Sigma_v\beta_{[I_0]}^\intercal+\Sigma_{u,[I_0]})^{-1}\beta_{[I_0]}-\mathbb{I}_{N_0}\right\|=O\left(1/\lambda_{\min}(\beta_{[I_0]}^\intercal\beta_{[I_0]})\right) \tag{2.120}$$

with $N_0=|I_0|$. Write $\widetilde{\beta}_{[I_0]}=\Sigma_{u,[I_0]}^{-1/2}\beta_{[I_0]}\Sigma_v^{1/2}$, then (2.120) becomes

$$\left\|\Sigma_v^{1/2}\widetilde{\beta}_{[I_0]}^\intercal(\widetilde{\beta}_{[I_0]}\widetilde{\beta}_{[I_0]}^\intercal+\mathbb{I}_{N_0})^{-1}\widetilde{\beta}_{[I_0]}\Sigma_v^{-1/2}-\mathbb{I}_{N_0}\right\|=O\left(1/\lambda_{\min}(\beta_{[I_0]}^\intercal\beta_{[I_0]})\right) \tag{2.121}$$

Suppose that the SVD of $\widetilde{\beta}_{[I_0]}$ can be written as $\widetilde{\beta}_{[I_0]}=B\Lambda^{1/2}\Gamma$, where $B\in\mathbb{R}^{N_0\times p}$ and $\Gamma\in\mathbb{R}^{p\times p}$ are matrices of left and right singular vectors, $\Lambda^{1/2}=\mathrm{diag}(\widetilde{\lambda}_1^{1/2},\cdots,\widetilde{\lambda}_p^{1/2})$ is a diagonal matrix and $\widetilde{\lambda}_i$ is the ith eigenvalue of $\widetilde{\beta}_{[I_0]}^\intercal\widetilde{\beta}_{[I_0]}$. Using the SVD of $\widetilde{\beta}_{[I_0]}$, we have

$$\widetilde{\beta}_{[I_0]}^\intercal\left(\widetilde{\beta}_{[I_0]}\widetilde{\beta}_{[I_0]}^\intercal+\mathbb{I}_N\right)^{-1}\widetilde{\beta}_{[I_0]}=\Gamma^\intercal\Lambda^{1/2}(\Lambda+\mathbb{I}_p)^{-1}\Lambda^{1/2}\Gamma.$$

Consequently, with $\lambda_{\max}(\Sigma_v)\lesssim 1$ and $\lambda_{\min}(\Sigma_v)\gtrsim 1$, the left hand side of (2.121) becomes

$$\begin{aligned}&\left\|\Sigma_v^{1/2}\widetilde{\beta}_{[I_0]}^\intercal(\widetilde{\beta}_{[I_0]}\widetilde{\beta}_{[I_0]}^\intercal+\mathbb{I}_{N_0})^{-1}\widetilde{\beta}_{[I_0]}\Sigma_v^{-1/2}-\mathbb{I}_{N_0}\right\|\\&=\left\|\Sigma_v^{1/2}\Gamma^\intercal\left(\Lambda^{1/2}(\Lambda+\mathbb{I}_p)^{-1}\Lambda^{1/2}-\mathbb{I}_p\right)\Gamma\Sigma_v^{-1/2}\right\|\\&\lesssim\left\|\Lambda^{1/2}(\Lambda+\mathbb{I}_p)^{-1}\Lambda^{1/2}-\mathbb{I}_p\right\|=\frac{1}{1+\widetilde{\lambda}_p}.\end{aligned}$$

Note that

$$\begin{aligned}\widetilde{\lambda}_p&=\lambda_p\left(\Sigma_v^{1/2}\beta_{[I_0]}^\intercal\Sigma_u^{-1}\beta_{[I_0]}\Sigma_v^{1/2}\right)\geq\lambda_p(\beta_{[I_0]}\Sigma_v\beta_{[I_0]}^\intercal)\lambda_{\min}(\Sigma_u^{-1})\asymp_{\mathrm{P}}\lambda_p(\beta_{[I_0]}^\intercal\beta_{[I_0]})\lambda_{\max}^{-1}(\Sigma_u)\\&\gtrsim\lambda_p(\beta_{[I_0]}^\intercal\beta_{[I_0]})=\lambda_{\min}(\beta_{[I_0]}^\intercal\beta_{[I_0]}),\end{aligned}$$

we have obtained (2.121) and this concludes the proof. □

2.6.4.2 Proof of Proposition 14

Proof. We consider the case $d = 1$ first. Recall that $(\breve{p}, \breve{q})$ denotes an arbitrary pair of tuning parameter values and that $\breve{w}$ is a short-hand notation for $w(\breve{p}, \breve{q})$. By definition, maximizing $R^2(\breve{p}, \breve{q})$ is equivalent to minimizing

$$\mathrm{MSE}(\breve{w}) = T_{\mathrm{oos}}^{-1} \left\| \bar{G}_{\mathrm{oos}} - \breve{w}\bar{R}_{\mathrm{oos}} \right\|^2 = \left\| (\breve{w}\beta - \eta)\bar{V}_{\mathrm{oos}} + \breve{w}\bar{U}_{\mathrm{oos}} - \bar{Z}_{\mathrm{oos}} \right\|^2. \tag{2.122}$$

As shown in the proof of Theorem 6, the estimated factor at the kth step is given by $\widehat{V}_{(k)} = \sqrt{T}\widehat{\xi}_{(k)}^{\intercal}$ and the loading of G on $\widehat{V}_{(k)}$ is $\widehat{\eta}_{(k)} = T^{-1/2}\bar{G}\widehat{\xi}_{(k)}$. Using (2.38) and (2.40), we have

$$\widehat{V}_{(k)} = \lambda_{(k)}^{-1/2}\widehat{\varsigma}_{(k)}^{\intercal}\tilde{R}_{(k)} = \lambda_{(k)}^{-1/2}\widehat{\varsigma}_{(k)}^{\intercal}D_{(k)}R_{(k)}. \tag{2.123}$$

Thus, the mimicking portfolio of g_t is given by $\sum_{i=1}^{\breve{p}} \widehat{\eta}_{(k)}\widehat{V}_{(k)} =: \breve{w}\bar{R}$, where

$$\breve{w} = \sum_{i=1}^{\breve{p}} \bar{G}\widehat{\xi}_{(i)} \frac{\widehat{\varsigma}_{(i)}^{\intercal}D_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}. \tag{2.124}$$

Next, we claim that there exists a pair of (c_0, q_0) that satisfy $q_0 \in \mathcal{Q}$ and (2.11) in Theorem 6. This holds because we can set $q_0 = N^{-\alpha_{nq}}$, and the existence of c_0 is guaranteed under the assumptions that $N^{1-\alpha_{nq}}/N_0 \to 0$ and $\log T/N^{1-\alpha_{nq}} \to 0$. Given c_0 and q_0, let p_0 denote the number of factors extracted. As shown in the proof of Theorem 6, we have $\mathrm{P}(p_0 = \tilde{p}) \to 1$ and $\tilde{p} \leq p$. Therefore, since $p_{\max} \geq p$, it implies that $(p_0, q_0) \in \langle p_{\max} \rangle \times \mathcal{Q}$ with probability approaching 1, and (p_0, q_0) corresponds to (c_0, q_0) mentioned above.

Denote $w_0 = w(p_0, q_0)$. Using (2.46) and the definition $\tilde{\beta}_{(k)} = D_{(k)}\beta$, we have

$$\|w_0\beta - \eta\| = o_{\mathrm{P}}(1). \tag{2.125}$$

Also, with $\left\|D_{(i)}\right\| \lesssim_{\mathrm{P}} 1$ from Lemma 20(i), we have

$$\|w_0\| \leq \sum_{i=1}^{p_0} T^{-1/2} \widehat{\lambda}_{(i)}^{-1/2} \left\|\bar{G}\right\| \left\|D_{(i)}\right\| = o_{\mathrm{P}}(1). \tag{2.126}$$

Next, we write (2.122) as

$$\begin{aligned} \mathrm{MSE}(w_0) = (w_0\beta - \eta) \left(\frac{\bar{V}_{\mathrm{oos}} \bar{V}_{\mathrm{oos}}^{\intercal}}{T_{\mathrm{oos}}}\right) (w_0\beta - \eta)^{\intercal} + \frac{w_0 \bar{U}_{\mathrm{oos}} (w_0 \bar{U}_{\mathrm{oos}})^{\intercal}}{T_{\mathrm{oos}}} + \frac{\bar{Z}_{\mathrm{oos}} \bar{Z}_{\mathrm{oos}}^{\intercal}}{T_{\mathrm{oos}}} \\ + 2\frac{w_0 \bar{U}_{\mathrm{oos}} \bar{V}_{\mathrm{oos}}^{\intercal}}{T_{\mathrm{oos}}} (w_0\beta - \eta)^{\intercal} - 2(w_0\beta - \eta)\frac{\bar{V}_{\mathrm{oos}} \bar{Z}_{\mathrm{oos}}^{\intercal}}{T_{\mathrm{oos}}} - 2\frac{w_0 \bar{U}_{\mathrm{oos}} \bar{Z}_{\mathrm{oos}}^{\intercal}}{T_{\mathrm{oos}}}, \end{aligned} \tag{2.127}$$

and analyze these terms one by one. Under Assumptions 19 and 20, rewriting the proof of Lemma 14 leads to

$$\left\|\frac{\bar{V}_{\mathrm{oos}} \bar{V}_{\mathrm{oos}}^{\intercal}}{T_{\mathrm{oos}}} - \Sigma_v\right\| \lesssim_{\mathrm{P}} T_{\mathrm{oos}}^{-1/2}, \quad \left\|\frac{\bar{Z}_{\mathrm{oos}} \bar{Z}_{\mathrm{oos}}^{\intercal}}{T_{\mathrm{oos}}} - \Sigma_z\right\| \lesssim_{\mathrm{P}} T_{\mathrm{oos}}^{-1/2}, \quad \left\|w_0 \bar{U}_{\mathrm{oos}}\right\| \lesssim_{\mathrm{P}} \|w_0\| \, T_{\mathrm{oos}}^{1/2}, \tag{2.128}$$

and

$$\left\|w_0 \bar{U}_{\mathrm{oos}} \bar{V}_{\mathrm{oos}}^{\intercal}\right\| \lesssim_{\mathrm{P}} \|w_0\| \, T_{\mathrm{oos}}^{1/2}, \quad \left\|\bar{V}_{\mathrm{oos}} \bar{Z}_{\mathrm{oos}}^{\intercal}\right\| \lesssim_{\mathrm{P}} T_{\mathrm{oos}}^{1/2}, \quad \left\|w_0 \bar{U}_{\mathrm{oos}} \bar{Z}_{\mathrm{oos}}^{\intercal}\right\| \lesssim_{\mathrm{P}} \|w_0\| \, T_{\mathrm{oos}}^{1/2}. \tag{2.129}$$

Therefore, together with (2.125) and (2.126), (2.127) implies that

$$\begin{aligned} \|\mathrm{MSE}(w_0) - \Sigma_z\| \lesssim_{\mathrm{P}} & \|w_0\beta - \eta\|^2 \lambda_{\min}(\Sigma_v) + \|w_0\|^2 + T_{\mathrm{oos}}^{-1/2} \\ & + \|w_0\beta - \eta\| \, \|w_0\| \, T_{\mathrm{oos}}^{-1/2} + \|w_0\beta - \eta\| \, T_{\mathrm{oos}}^{-1/2} + \|w_0\| \, T_{\mathrm{oos}}^{-1/2} \\ = & o_{\mathrm{P}}(1). \end{aligned} \tag{2.130}$$

In other words, there exists $(p_0, q_0) \in \langle p_{\max} \rangle \times \mathcal{Q}$ such that $\mathrm{MSE}(w(p_0, q_0)) \xrightarrow{\mathrm{P}} \Sigma_z$.

Given that (p^*, q^*) minimize $\mathrm{MSE}(w(\breve{p}, \breve{q}))$, we have $\mathrm{MSE}(w^*) \leq \mathrm{MSE}(w_0) = \Sigma_z + o_{\mathrm{P}}(1)$.

Let δ_1 and δ_2 denote $\|w^*\beta - \eta\|$ and $\|w^*\|$, we will show that $\delta_1 = o_{\mathrm{P}}(1)$ and $\delta_2 = o_{\mathrm{P}}(1)$. To see this, with (2.128) and the assumption $\lambda_{\min}(\Sigma_v) \gtrsim 1$, we have

$$(w^*\beta - \eta)\left(\frac{\bar{V}_{\mathrm{oos}}\bar{V}_{\mathrm{oos}}^{\intercal}}{T_{\mathrm{oos}}}\right)(w^*\beta - \eta)^{\intercal} \geq \delta_1^2 \lambda_{\min}\left(\frac{\bar{V}_{\mathrm{oos}}\bar{V}_{\mathrm{oos}}^{\intercal}}{T_{\mathrm{oos}}}\right) \gtrsim_{\mathrm{P}} \delta_1^2, \tag{2.131}$$

and

$$\frac{\left\|w^*\bar{U}\right\|}{\delta_2 T_{\mathrm{oos}}^{1/2}} = \frac{\left\|w^*\bar{U}\right\|}{\|w^*\|\, T_{\mathrm{oos}}^{1/2}} \geq \min_{\breve{p}\leq p_{\max}, \breve{q}\in\mathcal{Q}} \frac{\left\|w(\breve{p},\breve{q})\bar{U}\right\|}{\|w(\breve{p},\breve{q})\|\, T_{\mathrm{oos}}^{1/2}} \gtrsim_{\mathrm{P}} 1, \tag{2.132}$$

where we use the assumption that $p_{\max}$ and $|\mathcal{Q}| = n_q$ are finite and Assumption 20 to construct the uniform bound. Combining (2.131) and (2.132), we have

$$\begin{aligned}
&(w^*\beta - \eta)\left(\frac{\bar{V}_{\mathrm{oos}}\bar{V}_{\mathrm{oos}}^{\intercal}}{T_{\mathrm{oos}}}\right)(w^*\beta - \eta)^{\intercal} + \frac{w^*\bar{U}_{\mathrm{oos}}(w^*\bar{U}_{\mathrm{oos}})^{\intercal}}{T_{\mathrm{oos}}} \\
&\geq \max\left\{(w^*\beta - \eta)\left(\frac{\bar{V}_{\mathrm{oos}}\bar{V}_{\mathrm{oos}}^{\intercal}}{T_{\mathrm{oos}}}\right)(w^*\beta - \eta)^{\intercal}, \frac{w^*\bar{U}_{\mathrm{oos}}(w^*\bar{U}_{\mathrm{oos}})^{\intercal}}{T_{\mathrm{oos}}}\right\} \\
&\gtrsim_{\mathrm{P}} \max\{\delta_1^2, \delta_2^2\} \gtrsim (\delta_1 + \delta_2)^2
\end{aligned} \tag{2.133}$$

where the first inequality stems from the fact that the two quadratic forms on the left-hand-side are positive numbers. On the other hand, we have

$$\frac{\left\|w^*\bar{U}_{\mathrm{oos}}\bar{A}_{\mathrm{oos}}^{\intercal}\right\|}{\|w^*\|} \leq \max_{\breve{p}\leq p_{\max}, \breve{q}\in\mathcal{Q}} \frac{\left\|w(\breve{p},\breve{q})\bar{U}_{\mathrm{oos}}\bar{A}_{\mathrm{oos}}^{\intercal}\right\|}{\|w(\breve{p},\breve{q})\|} \lesssim_{\mathrm{P}} T_{\mathrm{oos}}^{1/2}, \tag{2.134}$$

for $A = V$ and Z. The decomposition of $\mathrm{MSE}(w^*)$ also has the form (2.127) by replacing

w_0 by w^*. With the decomposition, (2.129) and (2.134), we have

$$\begin{aligned}(\delta_1+\delta_2)^2 \lesssim_{\mathrm{P}} & (w^*\beta-\eta)\left(\frac{\bar{V}_{\mathrm{oos}}\bar{V}_{\mathrm{oos}}^{\intercal}}{T_{\mathrm{oos}}}\right)(w^*\beta-\eta)^{\intercal}+\frac{w^*\bar{U}_{\mathrm{oos}}(w^*\bar{U}_{\mathrm{oos}})^{\intercal}}{T_{\mathrm{oos}}} \\ \leq & \mathrm{MSE}(w^*)-\Sigma_z+O_{\mathrm{P}}(T_{\mathrm{oos}}^{-1/2}\delta_1)+O_{\mathrm{P}}(T_{\mathrm{oos}}^{-1/2}\delta_1\delta_2)+O_{\mathrm{P}}(T_{\mathrm{oos}}^{-1/2}\delta_2)+o_{\mathrm{P}}(1) \\ \leq & o_{\mathrm{P}}\big(1+(\delta_1+\delta_2)+\delta_1\delta_2\big)=o_{\mathrm{P}}\big(1+(\delta_1+\delta_2)^2\big),\end{aligned} \tag{2.135}$$

where we use $\delta_1+\delta_2 \leq (1+(\delta_1+\delta_2)^2)/2$ and $\delta_1\delta_2 \leq (\delta_1+\delta_2)^2/4$ in the last equation. This leads to $\delta_1 = o_{\mathrm{P}}(1)$ and $\delta_2 = o_{\mathrm{P}}(1)$ as δ_1 and δ_2 are non-negative. Plugging them into the second inequality of (2.135), we obtain $\mathrm{MSE}(w^*)-\Sigma_z \geq o_{\mathrm{P}}(1)$. With $\mathrm{MSE}(w^*) \leq \mathrm{MSE}(w_0)$ and (2.130), we have

$$\mathrm{MSE}(w^*)-\Sigma_z = o_{\mathrm{P}}(1). \tag{2.136}$$

In addition, in out-of-sample data, the expected return of g_t's mimicking portfolio based on w^* satisfies

$$\begin{aligned}\|w^*\bar{r}_{\mathrm{oos}}-\eta\gamma\| &= \|(w^*\beta-\eta)(\gamma+\bar{v}_{\mathrm{oos}})+\eta\bar{v}_{\mathrm{oos}}-w^*\bar{u}_{\mathrm{oos}}\| \\ &\lesssim_{\mathrm{P}} \|w^*\beta-\eta\|\,\|\gamma+\bar{v}_{\mathrm{oos}}\|+T_{\mathrm{oos}}^{-1/2}\,\|\eta\|+T_{\mathrm{oos}}^{-1}\,\|w^*U_{\mathrm{oos}}\|\,\big\|\iota_{T_{\mathrm{oos}}}\big\| \\ &\lesssim_{\mathrm{P}} \delta_1+T_{\mathrm{oos}}^{-1/2}+\delta_2 = o_{\mathrm{P}}(1),\end{aligned} \tag{2.137}$$

where we use $\|w^*\|^{-1}\,\|w^*U_{\mathrm{oos}}\| \leq \max_{\breve{p}\leq p_{\max},\breve{q}\in\mathcal{Q}}\|w(\breve{p},\breve{q})\|^{-1}\,\|w(\breve{p},\breve{q})U_{\mathrm{oos}}\| \lesssim_{\mathrm{P}} T_{\mathrm{oos}}^{1/2}$ and $\big\|\iota_{T_{\mathrm{oos}}}\big\| = \sqrt{T_{\mathrm{oos}}}$ in the second last inequality. This concludes the proof of the first part of the theorem.

For in-sample data, as we only use q^*N assets at each step, $\|w^*\|_0$ is no larger than

$p_{\max}q^*N$. Let S_{w^*} be the set $\{i|(w(p_i^*,q_i^*))_{[i]}\neq 0\}$, we have

$$\begin{aligned}\|w^*\bar{u}\| &= |\sum_{i\in S_{w^*}}(w(p_i^*,q_i^*))_{[i]}\bar{u}_i| \lesssim_{\mathrm{P}} \|w^*\|\left\|\bar{u}_{[S_{w^*}]}\right\| \leq \|w^*\|\sqrt{|S_{w^*}|}\,\|\bar{u}\|_{\mathrm{MAX}} \\ &= o_{\mathrm{P}}\left(\sqrt{\frac{p_{\max}q^*N\log N}{T}}\right).\end{aligned} \tag{2.138}$$

With the assumption that $q^*N\log N = O(T)$ and $p_{\max}$ is finite, we have $\|w^*\bar{u}\| = o_{\mathrm{P}}(1)$ and thus

$$\begin{aligned}\left\|\widehat{\gamma}_g - \eta\gamma\right\| = \|w^*\bar{r} - \eta\gamma\| &= \|(w^*\beta-\eta)(\gamma+\bar{v}) + \eta\bar{v} - w^*\bar{u}\| \\ &\lesssim_{\mathrm{P}} \|w^*\beta-\eta\|\,\|\gamma+\bar{v}\| + T^{-1/2}\|\eta\| + \|w^*\bar{u}\| = o_{\mathrm{P}}(1).\end{aligned} \tag{2.139}$$

Finally, we consider the general case $d \geq 1$. Note that in this case, $w(\breve{p},\breve{q}) \in \mathbb{R}^{d\times N}$. We use $(w(\breve{p},\breve{q}))_{[i]}$ to denote the ith row of $w(\breve{p},\breve{q})$. Suppose that R_i^2s are maximized separately for each i and denote

$$(p_i^*,q_i^*) = \underset{\breve{p}\leq p_{\max},\breve{q}\in\mathcal{Q}}{\arg\min}\ \mathrm{MSE}_i(w(\breve{p},\breve{q})),$$

where $\mathrm{MSE}_i(w) = T_{\mathrm{oos}}^{-1}\left\|(\bar{G}_{\mathrm{oos}})_{[i]} - w_{[i]}\bar{R}_{\mathrm{oos}}\right\|^2$. Replacing w_0 and w^* by $(w(p_0,q_0))_{[i]}$ and $(w(p_i^*,q_i^*))_{[i]}$ in the above proof, (2.130) and (2.136) become

$$(\Sigma_z)_{ii} + o_{\mathrm{P}}(1) = \mathrm{MSE}_i(w(p_i^*,q_i^*)) \leq \mathrm{MSE}_i(w(p_0,q_0)) = (\Sigma_z)_{ii} + o_{\mathrm{P}}(1). \tag{2.140}$$

Recall that $w^* = w(p^*,q^*)$ is obtained by maximizing the sum of R^2 and

$$(p^*,q^*) = \underset{\breve{p}\leq p_{\max},\breve{q}\in\mathcal{Q}}{\arg\max}\sum_{i=1}^{d}R_i^2(\breve{p},\breve{q}), \tag{2.141}$$

we then show that $\mathrm{MSE}_i(w^*) = (\Sigma_z)_{ii} + o_{\mathrm{P}}(1)$ also holds. To see this, using the definition

of (p_i^*, q_i^*) and (2.140), we have

$$\mathrm{MSE}_i(w^*) = \mathrm{MSE}_i(w(p^*, q^*)) \geq \mathrm{MSE}_i(w(p_i^*, q_i^*)) = (\Sigma_z)_{ii} + o_{\mathrm{P}}(1) = \mathrm{MSE}_i(w_0) + o_{\mathrm{P}}(1). \tag{2.142}$$

Using the definition of (p^*, q^*) in (2.141), we have

$$0 \leq \sum_{i=1}^{d} R_i^2(p^*, q^*) - \sum_{i=1}^{d} R_i^2(p_0, q_0) = \sum_{i=1}^{d} \frac{\mathrm{MSE}_i(w_0) - \mathrm{MSE}_i(w^*)}{T_{\mathrm{oos}}^{-1} \left\| \bar{G}_{[i]} \right\|^2}. \tag{2.143}$$

Together with $T_{\mathrm{oos}}^{-1} \left\| \bar{G}_{[i]} \right\|^2 \asymp_{\mathrm{P}} 1$ and (2.142), we have

$$\mathrm{MSE}_i(w_0) - \mathrm{MSE}_i(w^*) \geq \sum_{j \neq i} \frac{\left\| \bar{G}_{[i]} \right\|^2}{\left\| \bar{G}_{[j]} \right\|^2} \Big(\mathrm{MSE}_j(w^*) - \mathrm{MSE}_j(w_0) \Big) \geq o_{\mathrm{P}}(1). \tag{2.144}$$

Combining (2.142) and (2.144), we have

$$\mathrm{MSE}_i(w^*) = \mathrm{MSE}_i(w_0) + o_{\mathrm{P}}(1) = (\Sigma_z)_{ii} + o_{\mathrm{P}}(1). \tag{2.145}$$

With this equation, replacing w^*, Σ_z by $w_{[i]}^*$ and $(\Sigma_z)_{ii}$ in (2.135) leads to $\left\| w_{[i]}^* \eta - \beta \right\| = o_{\mathrm{P}}(1)$ and $\left\| w_{[i]}^* \right\| = o_{\mathrm{P}}(1)$. Consequently, we have

$$\left\| w_{[i]}^* \bar{r}_{\mathrm{oos}} - \eta_{[i]} \gamma \right\| = o_{\mathrm{P}}(1) \text{ and } \left\| w_{[i]}^* \bar{r} - \eta_{[i]} \gamma \right\| = o_{\mathrm{P}}(1),$$

which concludes the proof. □

2.6.5 Technical Lemmas and Their Proofs

Without loss of generality, we assume that $\Sigma_v = \mathbb{I}_p$ in the following lemmas. Also, except for Lemma 17, we assume that $\widehat{p} = \tilde{p}$ and $\widehat{I}_k = I_k$ for $k = 1, \ldots, \tilde{p}$, which hold with probability approaching one as we will show in Lemma 17.

Lemma 14. *Under Assumptions 7-13, for any $I \subset \langle N \rangle$, we have the following results:*

(i) $\left\|T^{-1}\bar{V}\bar{V}^\intercal - \Sigma_v\right\| \lesssim_{\mathrm{P}} T^{-1/2}, \quad \left\|T^{-1}\bar{Z}\bar{Z}^\intercal - \Sigma_z\right\| \lesssim_{\mathrm{P}} T^{-1/2}.$

(ii) $\left\|\left(\beta_{[I]}^\intercal \beta_{[I]}\right)^{-\frac{1}{2}} \beta_{[I]}^\intercal \bar{U}_{[I]}\right\| \lesssim_{\mathrm{P}} T^{1/2}.$

(iii) $\left\|\left(\beta_{[I]}^\intercal \beta_{[I]}\right)^{-\frac{1}{2}} \beta_{[I]}^\intercal \bar{U}_{[I]}\bar{V}^\intercal\right\| \lesssim_{\mathrm{P}} T^{1/2}, \quad \left\|\left(\beta_{[I]}^\intercal \beta_{[I]}\right)^{-\frac{1}{2}} \beta_{[I]}^\intercal \bar{U}_{[I]}\bar{Z}^\intercal\right\| \lesssim_{\mathrm{P}} T^{1/2}.$

(iv) $\left\|\bar{U}\right\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} (\log NT)^{1/2}$, $\left\|\bar{U}\bar{V}^\intercal\right\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} (\log N)^{1/2} T^{1/2}$,
$\left\|\bar{U}\bar{Z}^\intercal\right\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} (\log N)^{1/2} T^{1/2}.$

(v) $\left\|\bar{U}_{[I]}\right\| \lesssim_{\mathrm{P}} |I|^{1/2} + T^{1/2}, \quad \left\|\bar{U}_{[I]}\bar{V}^\intercal\right\| \lesssim_{\mathrm{P}} |I|^{1/2} T^{1/2}, \quad \left\|\bar{U}_{[I]}\bar{Z}^\intercal\right\| \lesssim_{\mathrm{P}} |I|^{1/2} T^{1/2}.$

(vi) $\left\|\bar{V}\right\| \lesssim_{\mathrm{P}} T^{1/2}, \quad \left\|\bar{Z}\right\| \lesssim_{\mathrm{P}} T^{1/2}, \quad \left\|\bar{V}\bar{Z}^\intercal\right\| \lesssim_{\mathrm{P}} T^{1/2}, \left\|\bar{V}\bar{Z}^\intercal - VZ^\intercal\right\| \lesssim_{\mathrm{P}} 1$

Proof. (i) Using Assumption 7, we have

$$\left\|\frac{\bar{V}\bar{V}^\intercal}{T} - \Sigma_v\right\| \leq \left\|\frac{VV^\intercal}{T} - \Sigma_v\right\| + \left\|\frac{V\iota_T\iota_T^\intercal V^\intercal}{T^2}\right\| = \left\|\frac{VV^\intercal}{T} - \mathbb{I}_p\right\| + \|\bar{v}\|^2 \lesssim_{\mathrm{P}} T^{-1/2}.$$

Replacing $\bar{V}$ by $\bar{Z}$, we also have the second inequality.

(ii) Using Assumption 11, we have

$$\left\|\left(\beta_{[I]}^\intercal \beta_{[I]}\right)^{-\frac{1}{2}} \beta_{[I]}^\intercal \bar{U}_{[I]}\right\| \leq \left\|\left(\beta_{[I]}^\intercal \beta_{[I]}\right)^{-\frac{1}{2}} \beta_{[I]}^\intercal U_{[I]}\right\| + T^{-1}\left\|\left(\beta_{[I]}^\intercal \beta_{[I]}\right)^{-\frac{1}{2}} \beta_{[I]}^\intercal U_{[I]}\iota_T\iota_T^\intercal\right\|$$
$$\lesssim_{\mathrm{P}} T^{1/2}.$$

(iii) By Assumptions 7, 11 and 12, we have

$$\begin{aligned}
&\left\| \left(\beta_{[I]}^{\intercal}\beta_{[I]}\right)^{-\frac{1}{2}} \beta_{[I]}^{\intercal}\bar{U}_{[I]}\bar{V}^{\intercal} \right\| \\
\leq &\left\| \left(\beta_{[I]}^{\intercal}\beta_{[I]}\right)^{-\frac{1}{2}} \beta_{[I]}^{\intercal}U_{[I]}V^{\intercal} \right\| + T^{-1}\left\| \left(\beta_{[I]}^{\intercal}\beta_{[I]}\right)^{-\frac{1}{2}} \beta_{[I]}^{\intercal}U_{[I]}\iota_T\iota_T^{\intercal}V \right\| \\
\leq &\left\| \left(\beta_{[I]}^{\intercal}\beta_{[I]}\right)^{-\frac{1}{2}} \beta_{[I]}^{\intercal}U_{[I]}V^{\intercal} \right\| + \left\| \left(\beta_{[I]}^{\intercal}\beta_{[I]}\right)^{-\frac{1}{2}} \beta_{[I]}^{\intercal}U_{[I]}\iota_T \right\| \|\bar{v}\| \lesssim_{\mathrm{P}} T^{1/2}.
\end{aligned}$$

Replacing $\bar{V}$ by $\bar{Z}$ in the above proof, with Assumptions 8, 11 and 13, we also have

$$\left\| \left(\beta_{[I]}^{\intercal}\beta_{[I]}\right)^{-\frac{1}{2}} \beta_{[I]}^{\intercal}\bar{U}_{[I]}\bar{Z}^{\intercal} \right\| \lesssim_{\mathrm{P}} T^{1/2}.$$

(iv) Using Assumption 10, we have

$$\begin{aligned}
\left\|\bar{U}\right\|_{\mathrm{MAX}} &\leq \|U\|_{\mathrm{MAX}} + T^{-1}\left\|U\iota_T\iota_T^{\intercal}\right\|_{\mathrm{MAX}} \leq \|U\|_{\mathrm{MAX}} + \|\bar{u}\|_{\mathrm{MAX}}\|\iota_T\| \\
&\lesssim_{\mathrm{P}} (\log N)^{1/2} + (\log T)^{1/2}.
\end{aligned}$$

Using Assumptions 7, 10, 12, we have

$$\begin{aligned}
\left\|\bar{U}\bar{V}^{\intercal}\right\|_{\mathrm{MAX}} &\leq \|UV^{\intercal}\|_{\mathrm{MAX}} + T^{-1}\left\|U\iota_T\iota_T^{\intercal}V^{\intercal}\right\|_{\mathrm{MAX}} \leq \|UV^{\intercal}\|_{\mathrm{MAX}} + T\|\bar{u}\|_{\mathrm{MAX}}\|\bar{v}\| \\
&\lesssim_{\mathrm{P}} (\log N)^{1/2}T^{1/2}.
\end{aligned}$$

Replacing $\bar{V}$ by $\bar{Z}$ in the above proof, with Assumptions 8, 10 and 13, we also have

$$\left\|\bar{U}\bar{Z}^{\intercal}\right\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} (\log N)^{1/2}T^{1/2}.$$

(v) Using Assumption 10 , we have

$$\left\|\bar{U}_{[I]}\right\| \leq \left\|U_{[I]}\right\| + T^{-1}\left\|U_{[I]}\iota_T\iota_T^{\intercal}\right\| \leq \left\|U_{[I]}\right\| + \left\|\bar{u}_{[I]}\right\| \|\iota_T\| \lesssim_{\mathrm{P}} |I|^{1/2} + T^{1/2}.$$

Using Assumptions 7, 10, 12, we have

$$\left\|\bar{U}_{[I]}\bar{V}^{\intercal}\right\| \leq \left\|U_{[I]}V^{\intercal}\right\| + T^{-1}\left\|U_{[I]}\iota_T\iota_T^{\intercal}V^{\intercal}\right\| \leq \left\|U_{[I]}V^{\intercal}\right\| + T\left\|\bar{u}_{[I]}\right\| \|\bar{v}\| \lesssim_{\mathrm{P}} |I|^{1/2}T^{1/2}.$$

Replacing $\bar{V}$ by $\bar{Z}$ in the above proof, with Assumptions 8, 10 and 13, we also have

$$\left\|\bar{U}_{[I]}\bar{Z}^{\intercal}\right\| \lesssim_{\mathrm{P}} |I|^{1/2}T^{1/2}.$$

(vi) Using Assumption 7, we have

$$\left\|\bar{V}\right\| \leq \|V\| + T^{-1}\left\|V\iota_T\iota_T^{\intercal}\right\| \leq \|V\| + \|\bar{v}\| \|\iota_T\| \lesssim_{\mathrm{P}} T^{1/2}.$$

Using Assumption 8, we have

$$\left\|\bar{Z}\right\| \leq \|Z\| + T^{-1}\left\|Z\iota_T\iota_T^{\intercal}\right\| \leq \|Z\| + \|\bar{z}\| \|\iota_T\| \lesssim_{\mathrm{P}} T^{1/2}.$$

Using Assumptions 7 and 8, we have

$$\left\|\bar{V}\bar{Z}^{\intercal}\right\| \leq \|VZ\| + T^{-1}\left\|V\iota_T\iota_T^{\intercal}Z\right\| \leq \|V\| + T\|\bar{v}\| \|\bar{z}\| \lesssim_{\mathrm{P}} T^{1/2},$$

and

$$\left\|\bar{V}\bar{Z}^{\intercal} - VZ^{\intercal}\right\| = \left\|T^{-1}V\iota_T\iota_T^{\intercal}Z\right\| = T\|\bar{v}\| \|\bar{z}\| \lesssim_{\mathrm{P}} 1.$$

□

Lemma 15. *The singular vectors $\widehat{\xi}_{(k)}$s we obtain from Algorithm 7 satisfy $\widehat{\xi}_{(j)}^{\mathsf{T}}\widehat{\xi}_{(k)} = \delta_{jk}$ for $j, k \leq \widehat{p}$.*

Proof. If $j = k$, this result holds from the definition of $\widehat{\xi}_{(k)}$. If $j < k$, recall that $\widetilde{R}_{(k)}$ is defined in (2.39) and $\widehat{\xi}_{(k)}$ is the first right singular vector of $\widetilde{R}_{(k)}$, we have

$$\widetilde{R}_{(k)} = \bar{R}_{[I_k]} \prod_{i<k} \left(\mathbb{I}_T - \widehat{\xi}_{(i)}\widehat{\xi}_{(i)}^{\mathsf{T}} \right) \quad \text{and} \quad \widehat{\xi}_{(k)} = \arg\max_{\alpha} \frac{\left\| \widetilde{R}_{(k)}\alpha \right\|}{\|\alpha\|}.$$

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

$$\left\| \widetilde{R}_{(k)}(\widehat{\xi}_{(k)} - c_0\widehat{\xi}_{(j)}) \right\| = \left\| \widetilde{R}_{(k)}\widehat{\xi}_{(k)} - c_0\widetilde{R}_{(k)}\widehat{\xi}_{(j)} \right\| = \left\| \widetilde{R}_{(k)}\widehat{\xi}_{(k)} \right\|, \tag{2.146}$$

since the definition of $\widetilde{R}_{(k)}$ implies that $\widetilde{R}_{(k)}\widehat{\xi}_{(j)} = 0$ for $j < k$.

On the other hand, since $\widehat{\xi}_{(k)}^{\mathsf{T}}\widehat{\xi}_{(j)} = c_0 \neq 0$, we have $(\widehat{\xi}_{(k)} - c_0\widehat{\xi}_{(j)})^{\mathsf{T}}\widehat{\xi}_{(j)} = 0$, and consequently,

$$\left\| \widehat{\xi}_{(k)} \right\|^2 = \left\| \widehat{\xi}_{(k)} - c_0\widehat{\xi}_{(j)} \right\|^2 + \left\| c_0\widehat{\xi}_{(j)} \right\|^2 > \left\| \widehat{\xi}_{(k)} - c_0\widehat{\xi}_{(j)} \right\|^2. \tag{2.147}$$

Apparently, if $\left\| \widetilde{R}_{(k)} \right\| = 0$, the process will stop so we have $\left\| \widetilde{R}_{(k)} \right\| > 0$ for $k \leq \widehat{p}$. Together with (2.146) and (2.147), we have

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

which contradicts with the definition of $\widehat{\xi}_{(k)}$. Therefore, $\widehat{\xi}_{(k)}^{\mathsf{T}}\widehat{\xi}_{(j)} = 0$ for $j < k$. This completes the proof. □

Lemma 16. *Under Assumption 9, if $c \to 0$, $qN/N_0 \to 0$ then b_k, $\beta_{(k)}$ and $\tilde{p}$ defined in Section 2.5.2 satisfy*

(i) $\langle b_j, b_k \rangle = \delta_{jk}$ *for* $j \leq k \leq \tilde{p}$.

(ii) $\left\| \beta_{(k)} \right\| \asymp q^{1/2} N^{1/2}$.

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

(iv) $\tilde{p} = p$, *if we further have* $\lambda_p(\eta^\intercal \eta) \gtrsim 1$.

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 arguments as in the proof of Lemma 15, we have $\langle b_j, b_k \rangle = \delta_{jk}$ for $j \leq k \leq \tilde{p}$.

(ii) The selection rule at kth step implies that

$$\frac{1}{|I_k|} \sum_{i \in I_k} \left\| \beta_{[i]} \prod_{j<k} \mathbb{M}_{b_j} \eta^\intercal \right\|_{\mathrm{MAX}}^2 \geq \frac{1}{N_0} \sum_{i \in I_0} \left\| \beta_{[i]} \prod_{j<k} \mathbb{M}_{b_j} \eta^\intercal \right\|_{\mathrm{MAX}}^2 . \tag{2.148}$$

For any matrix $A \in \mathbb{R}^{N \times d}$ and set $I \subset \langle N \rangle$, we have

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

and

$$\|A\|^2 \leq \|A\|_{\mathrm{F}}^2 \leq d \|A\|^2 ,$$

we thereby have

$$\|A\|^2 \asymp \sum_{i \in I} \left\| A_{[i]} \right\|_{\mathrm{MAX}}^2 . \tag{2.149}$$

Using this result, (2.148) becomes

$$\frac{1}{|I_k|} \left\| \beta_{[I_k]} \prod_{j<k} \mathbb{M}_{b_j} \eta^\intercal \right\|^2 \gtrsim \frac{1}{N_0} \left\| \beta_{[I_0]} \prod_{j<k} \mathbb{M}_{b_j} \eta^\intercal \right\|^2 .$$

Then, we have

$$\frac{1}{\sqrt{|I_k|}}\left\|\beta_{(k)}\right\|\left\|\prod_{j<k}\mathbb{M}_{b_j}\eta^{\intercal}\right\| \geq \frac{1}{\sqrt{|I_k|}}\left\|\beta_{[I_k]}\prod_{j<k}\mathbb{M}_{b_j}\eta^{\intercal}\right\| \\ \gtrsim \frac{1}{\sqrt{N_0}}\left\|\beta_{[I_0]}\prod_{j<k}\mathbb{M}_{b_j}\eta^{\intercal}\right\| \geq \frac{1}{\sqrt{N_0}}\sigma_p\left(\beta_{[I_0]}\right)\left\|\prod_{j<k}\mathbb{M}_{b_j}\eta^{\intercal}\right\|, \tag{2.150}$$

where we use $\beta_{[I_k]}\prod_{j<k}\mathbb{M}_{b_j}\eta^{\intercal} = \beta_{[I_k]}(\prod_{j<k}\mathbb{M}_{b_j})^2\eta^{\intercal} = \beta_{(k)}\prod_{j<k}\mathbb{M}_{b_j}\eta^{\intercal}$ in the first inequality. With $\sigma_p(\beta_{[I_0]}) \gtrsim \sqrt{N_0}$ from Assumption 9, (2.150) leads to $\left\|\beta_{(k)}\right\| \gtrsim |I_k|^{1/2}$. In addition, $\|\beta\|_{\text{MAX}} \lesssim 1$ from Assumption 9 leads to $\left\|\beta_{(k)}\right\| \lesssim |I_k|^{1/2}$. Therefore, we have $\left\|\beta_{(k)}\right\| \asymp |I_k|^{1/2} \asymp q^{1/2}N^{1/2}$.

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

(iv) Recall that $\tilde{p}$ is defined in Section 2.5.2. Since the procedure in its definition stops at $\tilde{p}+1$, we have at most $qN-1$ rows of β satisfying $\left\|\beta_{[i]}\prod_{j\leq\tilde{p}}\mathbb{M}_{b_j}\eta^{\intercal}\right\|_{\text{MAX}} \geq c$, which implies

$$\left\|\beta_{[I_0]}\prod_{j\leq\tilde{p}}\mathbb{M}_{b_j}\eta^{\intercal}\right\|^2 \lesssim qN + (N_0 - qN)c^2 = o(N_0),$$

where we use (2.149) and the assumptions $c \to 0$, $qN/N_0 \to 0$. With $\sigma_p(\beta_{[I_0]}) \gtrsim \sqrt{N_0}$ from Assumption 9, we have

$$\left\|\eta\prod_{j\leq\tilde{p}}\mathbb{M}_{b_j}\right\| \leq \sigma_p(\beta_{[I_0]})^{-1}\left\|\beta_{[I_0]}\prod_{j\leq\tilde{p}}\mathbb{M}_{b_j}\eta^{\intercal}\right\| = o(1). \tag{2.151}$$

If $\tilde{p} \leq p-1$, using (i), we have

$$\eta \prod_{j \leq \tilde{p}} \mathbb{M}_{b_j} = \eta - \eta \sum_{j \leq \tilde{p}} b_j b_j^\intercal,$$

which implies that

$$\sigma_p(\eta) \leq \sigma_1 \left(\eta \prod_{j \leq \tilde{p}} \mathbb{M}_{b_j} \right) + \sigma_p \left(\eta \sum_{j \leq \tilde{p}} b_j b_j^\intercal \right). \tag{2.152}$$

Since

$$\text{Rank} \left(\eta \sum_{j \leq \tilde{p}} b_j b_j^\intercal \right) \leq \tilde{p} \leq p-1, \tag{2.153}$$

we have $\sigma_p \left(\eta \sum_{j \leq \tilde{p}} b_j b_j^\intercal \right) \leq 0$ and thus (2.152) and (2.151) lead to

$$\sigma_p(\eta) \lesssim \sigma_1 \left(\eta \prod_{j \leq \tilde{p}} \mathbb{M}_{b_j} \right) \longrightarrow 0.$$

This contradicts with the assumption that $\lambda_p(\eta^\intercal \eta) \gtrsim 1$. Therefore, we have $\tilde{p} \geq p$. Together with the result in (iii), we have $\tilde{p} = p$. □

Lemma 17. *Suppose Assumptions 7-14 hold. If* $c^{-1} \log(NT)^{1/2} \left(q^{-1/2} N^{-1/2} + T^{-1/2} \right) \to 0$ *and* $c \to 0$, *then for* $k \leq \tilde{p}$ *and for* I_k, $\tilde{p}$ *and* $\beta_{(k)}$ *defined in Section 2.5.2, we have*

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

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

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

(iv) $\left\| \mathbb{P}_{\widehat{V}_{(k)}^\intercal} - T^{-1} \bar{V}^\intercal \mathbb{P}_{b_k} \bar{V} \right\| \lesssim_{\mathrm{P}} q^{-1/2} N^{-1/2} + T^{-1/2}$.

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

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 $T^{-1}\bar{R}\bar{G}^\intercal$ and I_1 based on $\beta\eta^\intercal$. With simple algebra, we have

$$T^{-1}\bar{R}\bar{G}^\intercal - \beta\eta^\intercal = \beta\left(T^{-1}\bar{V}\bar{V}^\intercal - \mathbb{I}_p\right)\eta^\intercal + T^{-1}\bar{U}\bar{V}^\intercal\eta^\intercal + T^{-1}\beta\bar{V}\bar{Z}^\intercal + T^{-1}\bar{U}\bar{Z}^\intercal.$$

With Assumptions 7, 8, 9, 12 13, we have

$$\begin{aligned}\left\|T^{-1}\bar{R}\bar{G}^\intercal - \beta\eta^\intercal\right\|_{\mathrm{MAX}} \lesssim \|\beta\|_{\mathrm{MAX}}\left\|T^{-1}\bar{V}\bar{V}^\intercal - \mathbb{I}_p\right\|\|\eta\| + T^{-1}\left\|\bar{U}\bar{V}^\intercal\right\|_{\mathrm{MAX}}\|\eta\| \\ + T^{-1}\|\beta\|_{\mathrm{MAX}}\left\|\bar{V}\bar{Z}^\intercal\right\| + T^{-1}\left\|\bar{U}\bar{Z}^\intercal\right\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} (\log N)^{1/2}T^{-1/2}.\end{aligned}$$

From Assumption 14, we have $c_{qN}^{(1)} - c_{qN+1}^{(1)} \gtrsim c_{qN}^{(1)}$ and the the definition of $\tilde{p}$ implies that $c_{qN}^{(k)} \geq c$ for $k \leq \tilde{p}$. Thus, we have $c_{qN}^{(1)} - c_{qN+1}^{(1)} \gtrsim c$. Define the events

$$\begin{aligned}A_1 &:= \left\{\left\|T^{-1}\bar{R}_{[i]}\bar{G}^\intercal\right\|_{\mathrm{MAX}} > (c_{qN}^{(1)} + c_{qN+1}^{(1)})/2 \text{ for all } i \in I_1\right\}, \\ A_2 &:= \left\{\left\|T^{-1}\bar{R}_{[i]}\bar{G}^\intercal\right\|_{\mathrm{MAX}} < (c_{qN}^{(1)} + c_{qN+1}^{(1)})/2 \text{ for all } i \in I_1^c\right\}, \\ A_3 &:= \left\{\left\|T^{-1}\bar{R}_{[i]}\bar{G}^\intercal - \beta_{[i]}\eta^\intercal\right\|_{\mathrm{MAX}} \geq (c_{qN}^{(1)} - c_{qN+1}^{(1)})/2 \text{ for some } i \in \langle N\rangle\right\}. \qquad (2.154)\end{aligned}$$

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 $\left\|\beta_{[i]}\eta^\intercal\right\|_{\mathrm{MAX}} \geq c_{qN}^{(1)}$ for all $i \in I_1$ and $\left\|\beta_{[i]}\eta^\intercal\right\|_{\mathrm{MAX}} \leq c_{qN+1}^{(1)}$ for all $i \in I_1^c$. Therefore, if A_1^c occurs, we have

$$\left\|T^{-1}\bar{R}_{[i]}\bar{G}^\intercal - \beta_{[i]}\eta^\intercal\right\|_{\mathrm{MAX}} \geq (c_{qN}^{(1)} - c_{qN+1}^{(1)})/2,$$

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

$$\mathrm{P}(\widehat{I}_1 = I_1) \geq \mathrm{P}(A_1 \cap A_2) = 1 - \mathrm{P}(A_1^c \cup A_2^c) \geq 1 - \mathrm{P}(A_3). \tag{2.155}$$

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 high probability, we impose $\widehat{I}_1 = I_1$ below. Then, we have $\widetilde{R}_{(1)} = \bar{R}_{[I_1]}$ and Assumption 18 gives $\left\|\widetilde{R}_{(1)} - \beta_{(1)}\bar{V}\right\| = \left\|\bar{U}_{[I_1]}\right\| \lesssim_{\mathrm{P}} q^{1/2}N^{1/2} + T^{1/2}$.

(iii) From Lemma 23, we have $\sigma_j(\beta_{(1)}\bar{V})/\sigma_j(\beta_1) = T^{1/2} + O_{\mathrm{P}}(1)$. The result in (ii) implies that

$$\left|\left\|\widetilde{R}_{(1)}\right\| - \left\|\beta_{(1)}\bar{V}\right\|\right| \leq \left\|\widetilde{R}_{(1)} - \beta_{(1)}\bar{V}\right\| \lesssim_{\mathrm{P}} q^{1/2}N^{1/2} + T^{1/2}.$$

Together with $\left\|\beta_{(1)}\right\| \asymp qN$ from Lemma 16, we have

$$\begin{aligned}\left|\frac{\widehat{\lambda}_{(1)}^{1/2}}{\left\|\beta_{(k)}\right\|} - 1\right| = \left|\frac{\left\|\widetilde{R}_{(1)}\right\|}{T^{1/2}\left\|\beta_{(1)}\right\|} - 1\right| &\leq \frac{\left|\left\|\widetilde{R}_{(1)}\right\| - \left\|\beta_{(1)}\bar{V}\right\|\right|}{T^{1/2}\left\|\beta_{(1)}\right\|} + \frac{\left|\left\|\beta_{(1)}\bar{V}\right\| - T^{1/2}\left\|\beta_{(1)}\right\|\right|}{T^{1/2}\left\|\beta_{(1)}\right\|} \\ &\lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1/2}.\end{aligned}$$

(iv) Let $\tilde{\xi}_{(1)} \in \mathbb{R}^{T\times 1}$ denote the first right singular vector of $\beta_{(1)}\bar{V}$. From Lemma 23, we have

$$\left\|\mathbb{P}_{\tilde{\xi}_{(1)}} - T^{-1}\bar{V}^{\intercal}\mathbb{P}_{b_k}\bar{V}\right\| \lesssim_{\mathrm{P}} T^{-1/2} \tag{2.156}$$

and $\sigma_j(\beta_{(1)}\bar{V})/\sigma_j(\beta_{(1)}) = T^{1/2} + O_{\mathrm{P}}(1)$ for $j \leq p$, which leads to

$$\sigma_1(\beta_{(1)}\bar{V}) - \sigma_2(\beta_{(1)}\bar{V}) = T^{1/2}(\sigma_1(\beta_{(1)}) - \sigma_2(\beta_{(1)})) + O_{\mathrm{P}}(\sigma_1(\beta_{(1)})) \asymp_{\mathrm{P}} T^{1/2}\sigma_1(\beta_{(1)}), \tag{2.157}$$

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

Using $\left\|\widetilde{R}_{(1)} - \beta_{(1)}\bar{V}\right\| \lesssim_{\mathrm{P}} q^{1/2}N^{1/2} + T^{1/2}$ as proved in (ii), (2.157), Lemma 16 and Wedin's sin-theta theorem for singular vectors in Wedin [1972], we have

$$\left\|\mathbb{P}_{\widehat{V}_{(k)}^{\intercal}} - \mathbb{P}_{\tilde{\xi}_{(1)}}\right\| \lesssim_{\mathrm{P}} \frac{q^{1/2}N^{1/2} + T^{1/2}}{\sigma_1(\beta_{(1)}\bar{V}) - \sigma_2(\beta_{(1)}\bar{V})} \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1/2}, \tag{2.158}$$

In light of (2.156) and (2.158), we have that (iv) holds for $k = 1$.

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 are tiny. At the kth step, the difference can be written as

$$\begin{aligned}
&\left\|\beta\prod_{j=1}^{k-1}\mathbb{M}_{b_j}\eta^{\intercal} - T^{-1}(\beta\bar{V}+\bar{U})\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}}(\eta\bar{V}+\bar{Z})^{\intercal}\right\|_{\mathrm{MAX}} \\
\leq &\left\|\beta\prod_{j=1}^{k-1}\mathbb{M}_{b_j}\eta^{\intercal} - T^{-1}\beta\bar{V}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}}\bar{V}^{\intercal}\eta^{\intercal}\right\|_{\mathrm{MAX}} + T^{-1}\left\|\beta\bar{V}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}}\bar{Z}^{\intercal}\right\|_{\mathrm{MAX}} \\
&+ T^{-1}\left\|\bar{U}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}}\bar{V}^{\intercal}\eta^{\intercal}\right\|_{\mathrm{MAX}} + T^{-1}\left\|\bar{U}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}}\bar{Z}^{\intercal}\right\|_{\mathrm{MAX}}
\end{aligned} \tag{2.159}$$

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

$$\left\| \sum_{j=1}^{k-1} \mathbb{P}_{\widehat{V}_{(j)}^{\intercal}} - T^{-1}\bar{V}^{\intercal} \sum_{j=1}^{k-1} \mathbb{P}_{b_j} \bar{V} \right\| = \left\| \sum_{j=1}^{k-1} \left(\mathbb{P}_{\widehat{V}_{(j)}^{\intercal}} - T^{-1}\bar{V}^{\intercal}\mathbb{P}_{b_j}\bar{V} \right) \right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1/2}. \tag{2.160}$$

Using Lemma 15 and Lemma 16(i), we have

$$\prod_{j=1}^{k-1} \mathbb{M}_{b_j} = \mathbb{I}_p - \sum_{j=1}^{k-1} \mathbb{P}_{b_j}, \quad \text{and} \quad \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{V}_{(j)}} = \mathbb{I}_T - \sum_{j=1}^{k-1} \mathbb{P}_{\widehat{V}_{(j)}}.$$

Using the above equations, (2.160), and $\left\| T^{-1}\bar{V}\bar{V}^{\intercal} - \mathbb{I}_p \right\| \lesssim_{\mathrm{P}} T^{-1/2}$, we have

$$\begin{aligned} T^{-1/2} \left\| \bar{V} \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{V}_{(j)}^{\intercal}} - \prod_{j=1}^{k-1} \mathbb{M}_{b_j} \bar{V} \right\| &= T^{-1/2} \left\| \bar{V} \sum_{j=1}^{k-1} \mathbb{P}_{\widehat{V}_{(j)}^{\intercal}} - \sum_{j=1}^{k-1} \mathbb{P}_{b_j} \bar{V} \right\| \\ &\lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1/2}. \end{aligned} \tag{2.161}$$

Similarly, right multiplying $\bar{V}^{\intercal}$ to the term inside the $\|\cdot\|$ of (2.161), we have

$$\left\| T^{-1}\bar{V} \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{V}_{(j)}^{\intercal}} \bar{V}^{\intercal} - \prod_{j=1}^{k-1} \mathbb{M}_{b_j} \right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1/2}. \tag{2.162}$$

Then, we analyze these four terms in (2.159) one by one. For the first term, using (2.162) and Assumption 9, we have

$$\begin{aligned} &\left\| \beta \prod_{j=1}^{k-1} \mathbb{M}_{b_j} \eta^{\intercal} - T^{-1}\beta\bar{V} \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{V}_{(j)}^{\intercal}} \bar{V}^{\intercal}\eta^{\intercal} \right\|_{\mathrm{MAX}} \\ \lesssim &\|\beta\|_{\mathrm{MAX}} \left\| \prod_{j=1}^{k-1} \mathbb{M}_{b_j} - T^{-1}\bar{V} \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{V}_{(j)}^{\intercal}} \bar{V}^{\intercal} \right\| \|\eta\| \\ \lesssim_{\mathrm{P}} &q^{-1/2}N^{-1/2} + T^{-1/2}. \end{aligned}$$

For the second term, using (2.161), Lemma 14 and Assumptions 9 and 8, we have

$$\begin{aligned} T^{-1}\left\|\beta\bar{V}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}}\bar{Z}^{\intercal}\right\|_{\mathrm{MAX}} \lesssim & T^{-1}\left\|\beta\right\|_{\mathrm{MAX}}\left\|\prod_{j=1}^{k-1}\mathbb{M}_{b_j}\right\|\left\|\bar{V}\bar{Z}^{\intercal}\right\| \\ & + T^{-1}\left\|\beta\right\|_{\mathrm{MAX}}\left\|\bar{V}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}}-\prod_{j=1}^{k-1}\mathbb{M}_{b_j}\bar{V}\right\|\left\|\bar{Z}\right\| \\ \lesssim_{\mathrm{P}} & q^{-1/2}N^{-1/2}+T^{-1/2}. \end{aligned}$$

For the third term, using (2.161) and Lemma 14, we have

$$\begin{aligned} T^{-1}\left\|\bar{U}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}}\bar{V}^{\intercal}\eta^{\intercal}\right\|_{\mathrm{MAX}} \lesssim & T^{-1}\left\|\bar{U}\bar{V}^{\intercal}\right\|_{\mathrm{MAX}}\left\|\prod_{j=1}^{k-1}\mathbb{M}_{b_j}\right\|\left\|\eta\right\| \\ & + T^{-1}\left\|\bar{U}\right\|_{\mathrm{MAX}}T^{1/2}\left\|\bar{V}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}}-\prod_{j=1}^{k-1}\mathbb{M}_{b_j}\bar{V}\right\|\left\|\eta\right\| \\ \lesssim_{\mathrm{P}} & (\log NT)^{1/2}\left(q^{-1/2}N^{-1/2}+T^{-1/2}\right). \end{aligned}$$

For the forth term, using (2.160) and Lemma 14, we have

$$\begin{aligned} T^{-1}\left\|\bar{U}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}}\bar{Z}^{\intercal}\right\|_{\mathrm{MAX}} \lesssim & T^{-1}\left\|\bar{U}\bar{Z}^{\intercal}\right\|_{\mathrm{MAX}}+T^{-2}\left\|\bar{U}\bar{V}^{\intercal}\right\|_{\mathrm{MAX}}\left\|\sum_{j=1}^{k-1}\mathbb{P}_{b_j}\right\|\left\|\bar{V}\bar{Z}^{\intercal}\right\| \\ & + T^{-1/2}\left\|\bar{U}\right\|_{\mathrm{MAX}}\left\|T^{-1}\bar{V}^{\intercal}\sum_{j=1}^{k-1}\mathbb{P}_{b_j}\bar{V}-\sum_{j=1}^{k-1}\mathbb{P}_{\widehat{V}_{(j)}^{\intercal}}\right\|\left\|\bar{Z}\right\| \\ \lesssim_{\mathrm{P}} & (\log NT)^{1/2}\left(q^{-1/2}N^{-1/2}+T^{-1/2}\right). \end{aligned}$$

Hence, we have

$$\left\| T^{-1}\bar{R}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}}\bar{G}^{\intercal} - \beta\prod_{j=1}^{k-1}\mathbb{M}_{b_j}\eta^{\intercal}\right\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} (\log NT)^{1/2}\left(q^{-1/2}N^{-1/2}+T^{-1/2}\right). \tag{2.163}$$

As in the case of $k=1$, from Assumption 14, we have $c_{qN}^{(k)} - c_{qN+1}^{(k)} \gtrsim c_{qN}^{(k)}$. In addition, since the stopping rule for the procedure in Section 2.5.2 is $c_{qN}^{(\tilde{p}+1)} < c$, we have $c_{qN}^{(k)} \geq c$ for $k \leq \tilde{p}$. With the assumption that

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

we can reuse the arguments for (2.154) and (2.155) in the case of $k=1$ and obtain $\mathrm{P}(\widehat{I}_k = I_k) \to 1$.

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

$$\begin{aligned}\widetilde{R}_{(k)} - \beta_{(k)}\bar{V} &= \bar{R}_{[I_k]}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}} - \beta_{(k)}\bar{V} \\ &= \bar{\beta}_{[I_k]}\left(\bar{V}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}} - \prod_{j=1}^{k-1}\mathbb{M}_{b_j}\bar{V}\right) + \bar{U}_{[I_k]}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}}.\end{aligned}$$

Hence, using Assumptions 9, Lemma 14, and (2.161), we have

$$\begin{aligned}\left\|\widetilde{R}_{(k)} - \beta_{(k)}\bar{V}\right\| &\leq \left\|\beta_{[I_k]}\right\|\left\|\bar{V}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}} - \prod_{j=1}^{k-1}\mathbb{M}_{b_j}\bar{V}\right\| + \left\|\bar{U}_{[I_k]}\right\|\left\|\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}}\right\| \\ &\lesssim_{\mathrm{P}} q^{1/2}N^{1/2} + T^{1/2}.\end{aligned}$$

(iii) The proof of (iii) is analogous to the case $k=1$. Rewrite the proof of the case $k=1$ by replacing $\widetilde{R}_{(1)}$ and $\beta_{(1)}$ by $\widetilde{R}_{(k)}$ and $\beta_{(k)}$. We have $|\widehat{\lambda}_{(k)}^{1/2}/\left\|\beta_{(k)}\right\| - 1| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1/2}$.

(iv) The proof of (iv) is analogous to the case $k = 1$. Let $\tilde{\xi}_{(k)}$ denote the first right singular vector of $\beta_{(k)}\bar{V}$, then we have $\left\|\mathbb{M}_{\tilde{\xi}_{(k)}} - T^{-1}\bar{V}^{\intercal}\mathbb{M}_{b_k}\bar{V}\right\| \lesssim_{\mathrm{P}} T^{-1/2}$ from Lemma 23. Since we have $\left\|\widetilde{R}_{(k)} - \beta_{(k)}\bar{V}\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1/2}$ from (ii), using the same proof as in the case $k = 1$, we have

$$\left\|\mathbb{M}_{\widehat{V}_{(k)}^{\intercal}} - \mathbb{M}_{\tilde{\xi}_{(k)}}\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1/2},$$

by Wedin's sin-theta theorem. Combining these two inequalities completes the proof.

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

(v) Recall that $\tilde{p}$ is determined by

$$\beta_{[i]}\prod_{j<k}\mathbb{M}_{b_j}\eta^{\intercal}$$

whereas $\widehat{p}$ is determined by $T^{-1}\bar{R}_{[i]}\prod_{j<k}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}}\bar{G}^{\intercal}$. Since (iv) holds for $j \leq \tilde{p}$ as shown above, using the same proof for (2.163), we have

$$\left\|T^{-1}\bar{R}\prod_{j=1}^{\tilde{p}}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}}\bar{G}^{\intercal} - \beta\prod_{j=1}^{\tilde{p}}\mathbb{M}_{b_j}\eta^{\intercal}\right\|_{\mathrm{MAX}} \lesssim_{\mathrm{P}} (\log NT)^{1/2}\left(q^{-1/2}N^{-1/2} + T^{-1/2}\right). \tag{2.164}$$

The assumption $c_{qN}^{(\tilde{p}+1)} \leq (1+\delta)^{-1}c$ in Assumption 14 implies that $c - c_{qN}^{(\tilde{p}+1)} \asymp c$. Together with

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

we can reuse the arguments for (2.154) and (2.155) with events

$$B_1 := \left\{ \left\| T^{-1}\bar{R}_{[i]} \prod_{j=1}^{\tilde{p}} \mathbb{M}_{\widehat{V}_{(j)}^{\intercal}} \bar{G}^{\intercal} \right\|_{\mathrm{MAX}} > (c + c_{qN}^{(\tilde{p}+1)})/2 \text{ for at most } qN-1 \text{ rows } i \in \langle N \rangle \right\},$$
$$B_2 := \left\{ \left\| T^{-1}\bar{R}_{[i]} \prod_{j=1}^{\tilde{p}} \mathbb{M}_{\widehat{V}_{(j)}^{\intercal}} \bar{G}^{\intercal} - \beta_{[i]} \prod_{j=1}^{\tilde{p}} \mathbb{M}_{b_j} \eta^{\intercal} \right\|_{\mathrm{MAX}} \geq (c - c_{qN}^{(\tilde{p}+1)})/2 \text{ for some } i \in \langle N \rangle \right\}, \tag{2.165}$$

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

Lemma 18. *Suppose that $\Gamma_{(k)} \in \mathbb{R}^{|I_k| \times |I_k|}$ is an orthogonal matrix with the first p rows equals to $\left(\beta_{[I_k]}^{\intercal}\beta_{[I_k]}\right)^{-\frac{1}{2}} \beta_{[I_k]}^{\intercal}$ and we define*

$$\begin{pmatrix} s_{(k)}^1 \\ s_{(k)}^2 \end{pmatrix} := \Gamma_{(k)}\widehat{\varsigma}_{(k)} \quad \text{and} \quad \begin{pmatrix} \widetilde{U}_{(k)}^1 \\ \widetilde{U}_{(k)}^2 \end{pmatrix} := \Gamma_{(k)}\bar{U}_{[I_k]},$$

where $s_{(k)}^1 \in \mathbb{R}^{p \times 1}$ and $\widetilde{U}_{(k)}^1 \in \mathbb{R}^{p \times T}$ are the first p rows of $\Gamma_{(k)}\widehat{\varsigma}_{(k)}$ and $\Gamma_{(k)}\bar{U}_{[I_k]}$, respectively. Then, under Assumptions 7-14, we have

(i) $\left\| s_{(k)}^2 \right\| \lesssim_{\mathrm{P}} T^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}(|I_k|^{1/2} + T^{1/2})$.

(ii) $\left\| \widetilde{U}_{(k)}^1 \right\| \lesssim_{\mathrm{P}} T^{1/2}$, $\left\| \widetilde{U}_{(k)}^1 \bar{V}^{\intercal} \right\| \lesssim_{\mathrm{P}} T^{1/2}$, $\left\| \widetilde{U}_{(k)}^1 \bar{Z}^{\intercal} \right\| \lesssim_{\mathrm{P}} T^{1/2}$.

Proof. (i) The assumption $\widehat{I}_k = I_k$ and the definition (2.39) of $\widetilde{R}_{(k)}$ together lead to

$$\widetilde{R}_{(k)} = \bar{R}_{[I_k]} \prod_{i<k} \left(\mathbb{I}_T - \widehat{\xi}_{(i)}\widehat{\xi}_{(i)}^{\intercal} \right).$$

Then, with (2.53) and Lemma 15, we have $\widehat{\varsigma}_{(k)} = \bar{R}_{[I_k]}\widehat{\xi}_{(k)}/\sqrt{T\widehat{\lambda}_{(k)}}$. From the construction

of $\Gamma_{(k)}$, we have

$$\Gamma_{(k)}\bar{R}_{(k)} = \begin{pmatrix} \left(\beta_{[I_k]}^{\intercal}\beta_{[I_k]}\right)^{\frac{1}{2}}\bar{V} + \widetilde{U}_{(k)}^{1} \\ \widetilde{U}_{(k)}^{2} \end{pmatrix},$$

which in turn gives

$$\begin{pmatrix} s_{(k)}^{1} \\ s_{(k)}^{2} \end{pmatrix} = \Gamma_{(k)}\widehat{\varsigma}_{(k)} = \frac{1}{\sqrt{T\widehat{\lambda}_{(k)}}} \begin{pmatrix} \left(\beta_{[I_k]}^{\intercal}\beta_{[I_k]}\right)^{\frac{1}{2}}\bar{V} + \widetilde{U}_{(k)}^{1} \\ \widetilde{U}_{(k)}^{2} \end{pmatrix} \widehat{\xi}_{(k)}.$$

With Lemma 14(v), we have

$$\left\|s_{(k)}^{2}\right\| = \left\|\frac{\widetilde{U}_{(k)}^{2}}{\sqrt{T\widehat{\lambda}_{(k)}}}\right\| \leq \left\|\frac{\bar{U}_{[I_k]}}{\sqrt{T\widehat{\lambda}_{(k)}}}\right\| \lesssim_{\mathrm{P}} T^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}(|I_k|^{1/2} + T^{1/2}).$$

(ii) With Lemma 14(ii)(iii) and the definition of $\Gamma_{(k)}$, these results follow immediately. □

Lemma 19. *Under Assumptions 7-14, if $\widehat{\lambda}_{(k)} \asymp_{\mathrm{P}} |I_k|$ and $|I_k| \asymp qN$ for $k \leq \tilde{p}$, then we have*

(i) $\left\|\frac{\bar{U}_{[I_k]}^{\intercal}\widehat{\varsigma}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}\right\| \lesssim_{\mathrm{P}} \frac{1}{\sqrt{qN}} + \frac{1}{T}.$

(ii) $\left\|\frac{\bar{V}\bar{U}_{[I_k]}^{\intercal}\widehat{\varsigma}_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}}\right\| \lesssim_{\mathrm{P}} \frac{1}{qN} + \frac{1}{T}$, $\left\|\frac{\bar{Z}\bar{U}_{[I_k]}^{\intercal}\widehat{\varsigma}_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}}\right\| \lesssim_{\mathrm{P}} \frac{1}{qN} + \frac{1}{T}$, $|\frac{\widehat{\varsigma}_{(k)}^{\intercal}\bar{u}_{[I_k]}}{\sqrt{\widehat{\lambda}_{(k)}}}| \lesssim_{\mathrm{P}} \frac{1}{qN} + \frac{1}{T}.$

Proof. (i) Using the equation $\widehat{\varsigma}_{(k)}^{\intercal}\bar{U}_{[I_k]} = (s_{(k)}^{1})^{\intercal}\widetilde{U}_{(k)}^{1} + (s_{(k)}^{2})^{\intercal}\widetilde{U}_{(k)}^{2}$ and Lemma 18, we have

$$\begin{aligned} \left\|\widehat{\varsigma}_{(k)}^{\intercal}\bar{U}_{[I_k]}\right\| &\leq \left\|s_{(k)}^{1}\right\|\left\|\widetilde{U}_{(k)}^{1}\right\| + \left\|s_{(k)}^{2}\right\|\left\|\widetilde{U}_{(k)}^{2}\right\| \leq \left\|s_{(k)}^{1}\right\|\left\|\widetilde{U}_{(k)}^{1}\right\| + \left\|s_{(k)}^{2}\right\|\left\|\bar{U}_{[I_k]}\right\| \\ &\lesssim_{\mathrm{P}} \sqrt{T} + \frac{|I_k| + T}{\sqrt{T\widehat{\lambda}_{(k)}}}, \end{aligned} \tag{2.166}$$

which leads to

$$\left\|\frac{\bar{U}_{[I_k]}^{\intercal}\widehat{\varsigma}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}\right\| \lesssim_{\mathrm{P}} \frac{1}{\sqrt{\widehat{\lambda}_{(k)}}} + \frac{|I_k|+T}{T\widehat{\lambda}_{(k)}} \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1}.$$

(ii) From Lemmas 14 and 18, we have

$$\begin{aligned}\left\|\bar{V}\bar{U}_{[I_k]}^{\intercal}\widehat{\varsigma}^{(k)}\right\| \leq \left\|\bar{V}\left(\widetilde{U}_{(k)}^{1}\right)^{\intercal}s_{(k)}^{1}\right\| + \left\|\bar{V}\left(\widetilde{U}_{(k)}^{2}\right)^{\intercal}s_{(k)}^{2}\right\| &\leq \left\|\bar{V}\left(\widetilde{U}_{(k)}^{1}\right)^{\intercal}\right\| + \left\|\bar{V}\bar{U}_{[I_k]}^{\intercal}\right\|\left\|s_{(k)}^{2}\right\| \\ &\lesssim_{\mathrm{P}} \sqrt{T} + \frac{|I_k|+T}{\sqrt{\widehat{\lambda}_{(k)}}},\end{aligned}$$

which leads to

$$\left\|\frac{\bar{V}\bar{U}_{[I_k]}^{\intercal}\widehat{\varsigma}_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}}\right\| \lesssim_{\mathrm{P}} \frac{1}{\sqrt{T\widehat{\lambda}_{(k)}}} + \frac{|I_k|+T}{T\widehat{\lambda}_{(k)}} \lesssim_{\mathrm{P}} q^{-1}N^{-1} + T^{-1}.$$

Replacing $\bar{V}$ by $\bar{Z}$ and $\iota_T^{\intercal}$ in the above proof and using Lemmas 14 and 18, we have similar results:

$$\left\|\frac{\bar{Z}\bar{U}_{[I_k]}^{\intercal}\widehat{\varsigma}_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}}\right\| \lesssim_{\mathrm{P}} q^{-1}N^{-1} + T^{-1}, \quad \text{and} \quad \left|\frac{\bar{u}_{[I_k]}^{\intercal}\widehat{\varsigma}_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}}\right| \lesssim_{\mathrm{P}} q^{-1}N^{-1} + T^{-1}. \tag{2.167}$$

□

Lemma 20. *Under Assumptions 7-14, if $\widehat{\lambda}_{(j)} \asymp_{\mathrm{P}} |I_j|$ and $|I_j| \asymp qN$ for $j \leq \tilde{p}$, then for $k, l \leq \tilde{p}$, we have*

(i) $\left\|\frac{\widetilde{U}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}\right\| \lesssim_{\mathrm{P}} \frac{1}{\sqrt{qN}} + \frac{1}{T}$, $\left\|\frac{\widetilde{U}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}\right\| \lesssim_{\mathrm{P}} \frac{1}{\sqrt{qN}} + \frac{1}{\sqrt{T}}$, $\left\|D_{(k)}\right\| \lesssim_{\mathrm{P}} 1$.

(ii) $\left\|\frac{\bar{V}\widetilde{U}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}}\right\| \lesssim_{\mathrm{P}} \frac{1}{qN} + \frac{1}{T}$, $\left\|\frac{\bar{Z}\widetilde{U}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}}\right\| \lesssim_{\mathrm{P}} \frac{1}{qN} + \frac{1}{T}$, $\left|\frac{\widehat{\varsigma}_{(k)}^{\intercal}\widetilde{u}_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}}\right| \lesssim_{\mathrm{P}} \frac{1}{qN} + \frac{1}{T}$.

(iii) $\left|\frac{\widehat{\xi}_{(l)}^{\intercal}\widetilde{U}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}\right| \lesssim_{\mathrm{P}} \frac{1}{qN} + \frac{1}{T}$.

Proof. (i) Recall that the definition of $U_{(k)}$ is

$$\widetilde{U}_{(k)} = \bar{U}_{[I_k]} - \sum_{i=1}^{k-1} \frac{\bar{R}_{[I_k]}\widehat{\xi}_{(i)}}{\sqrt{T}} \frac{\widehat{\varsigma}_{(i)}^{\intercal}\widetilde{U}_{(i)}}{\sqrt{\widehat{\lambda}_{(i)}}}. \tag{2.168}$$

Then, a direct multiplication of $\widehat{\varsigma}_{(k)}^{\intercal}/\sqrt{T\widehat{\lambda}_{(k)}}$ from the left side of (2.168) leads to

$$\frac{\widehat{\varsigma}_{(k)}^{\intercal}\widetilde{U}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}} = \frac{\widehat{\varsigma}_{(k)}^{\intercal}\bar{U}_{[I_k]}}{\sqrt{T\widehat{\lambda}_{(k)}}} - \sum_{i=1}^{k-1} \frac{\widehat{\varsigma}_{(k)}^{\intercal}\bar{R}_{[I_k]}\widehat{\xi}_{(i)}}{\sqrt{T\widehat{\lambda}_{(k)}}} \frac{\widehat{\varsigma}_{(i)}^{\intercal}\widetilde{U}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}.$$

Consequently, with Lemma 19(i) we have

$$\begin{aligned}
\left\| \frac{\widehat{\varsigma}_{(k)}^{\intercal}\widetilde{U}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}} \right\| &\leq \left\| \frac{\widehat{\varsigma}_{(k)}^{\intercal}\bar{U}_{[I_k]}}{\sqrt{T\widehat{\lambda}_{(k)}}} \right\| + \sum_{i=1}^{k-1} \left\| \frac{\bar{R}_{[I_k]}}{\sqrt{T\widehat{\lambda}_{(k)}}} \right\| \left\| \frac{\widehat{\varsigma}_{(i)}^{\intercal}\widetilde{U}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \right\| \\
&\lesssim_{\mathrm{P}} \frac{1}{\sqrt{\widehat{\lambda}_{(k)}}} + \frac{|I_k| + T}{T\widehat{\lambda}_{(k)}} + \sqrt{\frac{|I_k|}{\widehat{\lambda}_{(k)}}} \sum_{i=1}^{k-1} \left\| \frac{\widehat{\varsigma}_{(i)}^{\intercal}\widetilde{U}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \right\| \\
&\lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1} + \sum_{i=1}^{k-1} \left\| \frac{\widehat{\varsigma}_{(i)}^{\intercal}\widetilde{U}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \right\|.
\end{aligned} \tag{2.169}$$

If $\left\| T^{-1/2}\widehat{\lambda}_{(i)}^{-1/2}\widehat{\varsigma}_{(i)}^{\intercal}\widetilde{U}_{(i)} \right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1}$ holds for $i \leq k-1$, then (2.169) implies that this inequality also holds for k. In addition, when $k = 1$, $\widetilde{U}_{(1)} = \bar{U}_{[I_1]}$ and this equation is implied from Lemma 19(i). Therefore, we have $\left\| T^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}\widehat{\varsigma}_{(k)}^{\intercal}\widetilde{U}_{(k)} \right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1}$ for $k \leq \tilde{p}$ by induction.

Using (2.168) again, with Assumption 10, we have

$$\begin{aligned}\left\|\frac{\widetilde{U}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}\right\| &\leq \left\|\frac{\bar{U}_{[I_k]}}{\sqrt{T\widehat{\lambda}_{(k)}}}\right\| + \sum_{i=1}^{k-1}\left\|\frac{\bar{R}_{[I_k]}}{\sqrt{T\widehat{\lambda}_{(k)}}}\right\|\left\|\frac{\widetilde{U}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}\right\| \\ &\lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1/2} + \sum_{i=1}^{k-1}\left\|\frac{\widetilde{U}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}\right\|.\end{aligned} \tag{2.170}$$

When $k = 1$, Assumption 10 implies that $\left\|T^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}\widetilde{U}_{(k)}\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2} + T^{-1/2}$. Then, using the same induction argument with (2.170), we have this ineqaulity holds for $k \leq \tilde{p}$.

Recall that $D_{(k)}$ is defined by

$$D_{(k)} = \mathbb{I}_{[I_k]} - \sum_{i=1}^{k-1}\bar{R}_{[I_k]}\widehat{\xi}_{(i)}\frac{\widehat{\varsigma}_{(i)}'D_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}$$

and $D_{(1)} = \mathbb{I}_{[I_1]} \lesssim 1$, we have $\left\|D_{(k)}\right\| \lesssim_{\mathrm{P}} 1$ by induction as $\left\|T^{-1/2}\widehat{\lambda}_{(i)}^{-1/2}\bar{R}_{[I_k]}\right\| \lesssim_{\mathrm{P}} 1$.

(ii) Similarly, by simple multiplication of $\bar{V}^{\intercal}$ from the right side of (2.168), we have

$$\frac{\widehat{\varsigma}_{(k)}^{\intercal}\widetilde{U}_{(k)}\bar{V}^{\intercal}}{T\sqrt{\widehat{\lambda}_{(k)}}} = \frac{\widehat{\varsigma}_{(k)}^{\intercal}\bar{U}_{[I_k]}\bar{V}^{\intercal}}{T\sqrt{\widehat{\lambda}_{(k)}}} - \sum_{i=1}^{k-1}\frac{\widehat{\varsigma}_{(k)}^{\intercal}\bar{R}_{[I_k]}\widehat{\xi}_{(i)}}{\sqrt{T\widehat{\lambda}_{(k)}}}\frac{\widehat{\varsigma}_{(i)}^{\intercal}\widetilde{U}_{(i)}\bar{V}^{\intercal}}{T\sqrt{\widehat{\lambda}_{(i)}}}.$$

Consequently, we have

$$\begin{aligned}\left\|\frac{\widehat{\varsigma}_{(k)}^{\intercal}\widetilde{U}_{(k)}\bar{V}^{\intercal}}{T\sqrt{\widehat{\lambda}_{(k)}}}\right\| &\leq \left\|\frac{\widehat{\varsigma}_{(k)}^{\intercal}\bar{U}_{[I_k]}\bar{V}^{\intercal}}{T\sqrt{\widehat{\lambda}_{(k)}}}\right\| + \sum_{i=1}^{k-1}\left\|\frac{\bar{R}_{[I_k]}}{\sqrt{T\widehat{\lambda}_{(k)}}}\right\|\left\|\frac{\widehat{\varsigma}_{(i)}^{\intercal}\widetilde{U}_{(i)}\bar{V}^{\intercal}}{T\sqrt{\widehat{\lambda}_{(i)}}}\right\| \\ &\lesssim_{\mathrm{P}} q^{-1}N^{-1} + T^{-1} + \sqrt{\frac{|I_k|}{\widehat{\lambda}_{(k)}}}\sum_{i=1}^{k-1}\left\|\frac{\widehat{\varsigma}_{(i)}^{\intercal}\widetilde{U}_{(i)}\bar{V}^{\intercal}}{\sqrt{T\widehat{\lambda}_{(i)}}}\right\| \\ &\lesssim_{\mathrm{P}} q^{-1}N^{-1} + T^{-1} + \sum_{i=1}^{k-1}\left\|\frac{\widehat{\varsigma}_{(i)}^{\intercal}\widetilde{U}_{(i)}\bar{V}^{\intercal}}{\sqrt{T\widehat{\lambda}_{(i)}}}\right\|. \end{aligned} \tag{2.171}$$

When $k=1$, $\left\|T^{-1}\widehat{\lambda}_{(k)}^{-1/2}\widehat{\varsigma}_{(k)}^{\intercal}\widetilde{U}_{(k)}\bar{V}^{\intercal}\right\| \lesssim_{\mathrm{P}} q^{-1}N^{-1}+T^{-1}$ is a result of Lemma 19(ii). Then, a direct induction argument using (2.171) leads to this inequality for $k \leq \tilde{p}$.

Replacing $\bar{V}$ by $\bar{Z}$ and $\iota_T^{\intercal}$ in the above proof, and using Lemma 19(ii), we have the following results:

$$\left\|\frac{\bar{Z}\widetilde{U}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}}\right\| \lesssim_{\mathrm{P}} q^{-1}N^{-1}+T^{-1} \quad \text{and} \quad \left|\frac{\widetilde{u}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}}\right| \lesssim_{\mathrm{P}} q^{-1}N^{-1}+T^{-1}.$$

(iii) Recall that $\widetilde{R}_{(k)} = \widetilde{\beta}_{(k)}\bar{V} + \widetilde{U}_{(k)}$ as defined in (2.39), we have

$$\begin{aligned}|\widehat{\varsigma}_{(l)}^{\intercal}\widetilde{R}_{(l)}\widetilde{U}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}| &\leq |\widehat{\varsigma}_{(l)}^{\intercal}\widetilde{\beta}_{(l)}\bar{V}\widetilde{U}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}| + |\widehat{\varsigma}_{(l)}^{\intercal}\widetilde{U}_{(l)}\widetilde{U}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}| \\ &\leq \left\|\widehat{\varsigma}_{(l)}^{\intercal}\widetilde{\beta}_{(l)}\right\|\left\|\bar{V}\widetilde{U}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}\right\| + \left\|\widehat{\varsigma}_{(l)}^{\intercal}\widetilde{U}_{(l)}\right\|\left\|\widetilde{U}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}\right\|. \end{aligned}$$

Using (2.53), we have

$$\left|\frac{\widehat{\xi}_{(k)}^{\intercal}\widetilde{U}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}\right| = \left|\frac{\widehat{\varsigma}_{(l)}^{\intercal}\widetilde{R}_{(l)}\widetilde{U}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}\widehat{\lambda}_{(l)}}}\right| \leq \left\|\frac{\widehat{\varsigma}_{(l)}^{\intercal}\widetilde{\beta}_{(l)}}{\sqrt{\widehat{\lambda}_{(l)}}}\right\|\left\|\frac{\bar{V}\widetilde{U}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}}\right\| + \left\|\frac{\widetilde{U}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}\right\|\left\|\frac{\widetilde{U}_{(l)}^{\intercal}\widehat{\varsigma}_{(l)}}{\sqrt{T\widehat{\lambda}_{(l)}}}\right\|. \tag{2.172}$$

With Lemma 14 and (i), we have

$$T^{1/2}\left\|\widetilde{\beta}_{(k)}\right\| \lesssim_{\mathrm{P}} \sigma_p(\bar{V})\left\|\widetilde{\beta}_{(k)}\right\| \leq \left\|\widetilde{\beta}_{(k)}\bar{V}\right\| \leq \left\|\widetilde{U}_{(k)}\right\| + \left\|\widetilde{R}_{(k)}\right\| \leq \left\|\widetilde{U}_{(k)}\right\| + \left\|\bar{R}_{[I_k]}\right\| \lesssim_{\mathrm{P}} T^{1/2}q^{1/2}N^{1/2},$$

which leads to $\left\|\widehat{\lambda}_{(k)}^{-1/2}\widehat{\varsigma}_{(k)}^{\intercal}\widetilde{\beta}_{(k)}\right\| \lesssim_{\mathrm{P}} q^{-1/2}N^{-1/2}\left\|\widetilde{\beta}_{(k)}\right\| \lesssim_{\mathrm{P}} 1$. Using this inequality and results of (i) and (ii) in (2.172) completes the proof. □

Lemma 21. *Under Assumptions 7-14, if* $\widehat{\lambda}_{(j)} \asymp_{\mathrm{P}} |I_j|$ *and* $|I_j| \asymp qN$ *for* $j \leq \tilde{p}$*, then for* $k \leq \tilde{p}+1$*, we have*

(i) $\left\|\widetilde{Z}_{(k)}\bar{V}^{\intercal}\right\| \lesssim_{\mathrm{P}} T^{1/2} + Tq^{-1}N^{-1}$.

(ii) $\left\|\widetilde{Z}_{(k)}\bar{U}_{[I_0]}^{\intercal}\right\| \lesssim_{\mathrm{P}} N_0^{1/2}T^{1/2} + Tq^{-1/2}N^{-1/2}$.

Proof. (i) From the definition (2.42) of $\widetilde{Z}_{(k)}$, we have

$$\widetilde{Z}_{(k)}\bar{V}^{\intercal} = \bar{Z}\bar{V}^{\intercal} - \sum_{i=1}^{k-1}\bar{G}\widehat{\xi}_{(i)}\frac{\widehat{\varsigma}_{(i)}^{\intercal}\widetilde{U}_{(i)}\bar{V}^{\intercal}}{\sqrt{T\widehat{\lambda}_{(i)}}}.$$

Then, with Lemma 20(ii), we have

$$\left\|\widetilde{Z}_{(k)}\bar{V}^{\intercal}\right\| \leq \|\bar{Z}\bar{V}^{\intercal}\| + \sum_{i=1}^{k-1}\left\|\bar{G}\widehat{\xi}_{(i)}\right\|\left\|\frac{\widehat{\varsigma}_{(i)}^{\intercal}\widetilde{U}_{(i)}\bar{V}^{\intercal}}{\sqrt{T\widehat{\lambda}_{(i)}}}\right\| \lesssim_{\mathrm{P}} T^{1/2} + T\left(q^{-1}N^{-1} + T^{-1}\right) \lesssim_{\mathrm{P}} T^{1/2} + Tq^{-1}N^{-1}.$$

(ii) With (2.42) again, we have

$$\widetilde{Z}_{(k)}\bar{U}_{[I_0]}^{\intercal} = \bar{Z}\bar{U}_{[I_0]}^{\intercal} - \sum_{i=1}^{k-1}\bar{G}\widehat{\xi}_{(i)}\frac{\widehat{\varsigma}_{(i)}^{\intercal}\widetilde{U}_{(i)}\bar{U}_{[I_0]}^{\intercal}}{\sqrt{T\widehat{\lambda}_{(i)}}},$$

which, along with Lemma 20(i) and the assumptions on q, lead to

$$\begin{aligned}\left\|\widetilde{Z}_{(k)}\bar{U}_{[I_0]}^{\intercal}\right\| &\leq \left\|\bar{Z}\bar{U}_{[I_0]}^{\intercal}\right\| + \sum_{i=1}^{k-1}\left\|\bar{G}\widehat{\xi}_{(i)}\right\|\left\|\frac{\widehat{\varsigma}_{(i)}^{\intercal}\widetilde{U}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}\right\|\left\|\bar{U}_{[I_0]}\right\| \\ &\lesssim_{\mathrm{P}} N_0^{1/2}T^{1/2} + \left(q^{-1/2}N^{-1/2}+T^{-1}\right)\left(N_0^{1/2}T^{1/2}+T\right) \\ &\lesssim_{\mathrm{P}} N_0^{1/2}T^{1/2} + Tq^{-1/2}N^{-1/2}.\end{aligned}$$

□

Lemma 22. *Suppose that Assumptions 7-14 hold. If $\widehat{\lambda}_{(j)} \asymp_{\mathrm{P}} |I_j|$ and $|I_j| \asymp qN$ for $j \leq \tilde{p}$, then H_1, H_2 defined by* (2.51) *satisfy*

(i) $\|H_1\| \lesssim_{\mathrm{P}} 1$, $\|H_2\| \lesssim_{\mathrm{P}} 1$.

(ii) $\left\|H_1^{\intercal}H_2 - \mathbb{I}_{\tilde{p}}\right\| \lesssim_{\mathrm{P}} T^{-1} + q^{-1}N^{-1}$.

(iii) $\|H_1 - H_2\| \lesssim_{\mathrm{P}} T^{-1/2} + q^{-1}N^{-1}$.

Proof. (i) Using the definition (2.51) of H_1 and Lemma 14, we have

$$\|h_{k1}\| = \left\|\frac{\bar{V}\widehat{\xi}_{(k)}}{\sqrt{T}}\right\| \leq T^{-1/2}\left\|\bar{V}\right\| \lesssim_{\mathrm{P}} 1,$$

which leads to $\|H_1\| \lesssim_{\mathrm{P}} 1$.

Using the definition (2.51) of H_2, we have

$$\|h_{k2}\| = \left\|\frac{\widetilde{\beta}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}}\right\| \leq q^{-1/2}N^{-1/2}\left\|\widetilde{\beta}_{(k)}\right\|. \tag{2.173}$$

With Lemma 14 and Lemma 20(i), we have

$$T^{1/2}\left\|\widetilde{\beta}_{(k)}\right\| \lesssim_{\mathrm{P}} \sigma_p(\bar{V})\left\|\widetilde{\beta}_{(k)}\right\| \leq \left\|\widetilde{\beta}_{(k)}\bar{V}\right\| \leq \left\|\widetilde{U}_{(k)}\right\| + \left\|\widetilde{R}_{(k)}\right\| \leq \left\|\widetilde{U}_{(k)}\right\| + \left\|\bar{R}_{[I_k]}\right\| \lesssim_{\mathrm{P}} T^{1/2}q^{1/2}N^{1/2}. \tag{2.174}$$

Combining (2.173) and (2.174), we have $\|h_{k2}\| \lesssim_{\mathrm{P}} 1$ and thus $\|H_2\| \lesssim_{\mathrm{P}} 1$.

(ii) By (2.53) and Lemma 15, we have

$$\delta_{lk} = \widehat{\xi}_{(l)}^{\intercal}\widehat{\xi}_{(k)} = \frac{\widehat{\xi}_{(l)}^{\intercal}\bar{V}^{\intercal}\widetilde{\beta}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}} + \frac{\widehat{\xi}_{(l)}^{\intercal}\widetilde{U}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}} = h_{l1}^{\intercal}h_{k2} + \frac{\widehat{\xi}_{(l)}^{\intercal}\widetilde{U}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}.$$

By Lemma 20(iii), we have

$$|h_{l1}^{\intercal}h_{k2} - \delta_{lk}| \lesssim_{\mathrm{P}} q^{-1}N^{-1} + T^{-1},$$

and thus $\left\|H_1^{\intercal}H_2 - \mathbb{I}_{\tilde{p}}\right\| \lesssim_{\mathrm{P}} q^{-1}N^{-1} + T^{-1}$.

(iii) Using (2.53), we have

$$\bar{V}\widehat{\xi}_{(k)} = \frac{\bar{V}\bar{V}^{\intercal}\widetilde{\beta}_{(k)}^{\intercal}}{\sqrt{T\widehat{\lambda}_{(k)}}}\widehat{\varsigma}_{(k)} + \frac{\bar{V}\widetilde{U}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}.$$

With the definition of h_{k1} and h_{k2}, it becomes

$$h_{k1} = \frac{\bar{V}\bar{V}^{\intercal}}{T}h_{k2} + \frac{\bar{V}\widetilde{U}_{(k)}^{\intercal}\widehat{\varsigma}_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}}. \tag{2.175}$$

With $\|h_{k2}\| \lesssim_{\mathrm{P}} 1$, Lemma 14 and Lemma 20(ii), (2.175) leads to

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

This completes the proof. □

Lemma 23. *For any $N \times p$ matrix β, if $\left\|T^{-1}\bar{V}\bar{V}^\intercal - \mathbb{I}_p\right\| \lesssim_{\mathrm{P}} T^{-1/2}$, we have*

(i) $\sigma_j(\beta\bar{V})/\sigma_j(\beta) = T^{1/2} + O_{\mathrm{P}}(1)$ *for* $j \leq p$.

(ii) If $\sigma_1(\beta) - \sigma_2(\beta) \asymp \sigma_1(\beta)$, *then* $\left\|\mathbb{P}_{\tilde{\xi}} - T^{-1}\bar{V}^\intercal \mathbb{P}_b \bar{V}\right\| \lesssim_{\mathrm{P}} T^{-1/2}$, *where b is the first right singular vector of β and $\tilde{\xi}$ is the first right singular vector of $\beta\bar{V}$.*

Proof. (i) For $j \leq p$, $\sigma_j(\beta\bar{V})^2 = \lambda_j(\beta\bar{V}\bar{V}^\intercal\beta^\intercal) = \lambda_j(\beta^\intercal\beta\bar{V}\bar{V}^\intercal)$ which implies

$$\lambda_j(\beta^\intercal\beta)\lambda_p(\bar{V}\bar{V}^\intercal) \leq \sigma_j(\beta\bar{V})^2 \leq \lambda_j(\beta^\intercal\beta)\lambda_1(\bar{V}\bar{V}^\intercal).$$

With the assumption $\left\|T^{-1}\bar{V}\bar{V} - \mathbb{I}_p\right\| \lesssim_{\mathrm{P}} T^{-1/2}$, we have

$$T^{-1/2}\sigma_j(\beta\bar{V})/\sigma_j(\beta) = 1 + O_{\mathrm{P}}\left(T^{-1/2}\right)$$

by Weyl's theorem.

(ii) Let $\widehat{\varsigma}$ and $\tilde{\varsigma}$ be the first left singular vectors of β and $\beta\bar{V}$, respectively. Equivalently, $\widehat{\varsigma}$ and $\tilde{\varsigma}$ are the eigenvectors of $\beta\beta^\intercal$ and $T^{-1}\beta\bar{V}\bar{V}^\intercal\beta^\intercal$. Since $\left\|\beta\beta^\intercal - T^{-1}\beta\bar{V}\bar{V}^\intercal\beta^\intercal\right\| \leq \|\beta\|^2\left\|T^{-1}\bar{V}\bar{V}^\intercal - \mathbb{I}_p\right\| \lesssim_{\mathrm{P}} \sigma_1(\beta)^2 T^{-1/2}$ and $\sigma_1(\beta) - \sigma_2(\beta) \asymp \sigma_1(\beta)$, by sin-theta theorem we have

$$\|\widehat{\varsigma}\widehat{\varsigma}^\intercal - \tilde{\varsigma}\tilde{\varsigma}^\intercal\| \lesssim \frac{\left\|\beta\beta^\intercal - T^{-1}\beta\bar{V}\bar{V}^\intercal\beta^\intercal\right\|}{\sigma_1(\beta)^2 - \sigma_2(\beta)^2 - O(\left\|\beta\beta^\intercal - T^{-1}\beta\bar{V}\bar{V}^\intercal\beta^\intercal\right\|)} \lesssim_{\mathrm{P}} T^{-1/2}.$$

Using the relationship between left and right singular vectors, we have

$$b^\intercal = \frac{\widehat{\varsigma}^\intercal\beta}{\sigma_1(\beta)}, \quad \tilde{\xi}^\intercal = \frac{\tilde{\varsigma}^\intercal\beta\bar{V}}{\left\|\beta\bar{V}\right\|}.$$

Therefore,

$$\left\|\mathbb{P}_{\tilde{\xi}}-\frac{\sigma_1(\beta)^2}{\left\|\beta\bar{V}\right\|^2}\bar{V}^\intercal\mathbb{P}_b\bar{V}\right\|=\left\|\tilde{\xi}\tilde{\xi}^\intercal-\frac{\bar{V}^\intercal\beta^\intercal\widehat{\varsigma}\widehat{\varsigma}^\intercal\beta\bar{V}}{\left\|\beta\bar{V}\right\|^2}\right\|=\left\|\frac{\bar{V}^\intercal\beta^\intercal\tilde{\varsigma}\tilde{\varsigma}^\intercal\beta\bar{V}}{\left\|\beta\bar{V}\right\|^2}-\frac{\bar{V}^\intercal\beta^\intercal\widehat{\varsigma}\widehat{\varsigma}^\intercal\beta\bar{V}}{\left\|\beta\bar{V}\right\|^2}\right\|\lesssim_{\mathrm{P}}T^{-1/2}. \tag{2.176}$$

By Weyl's inequality, we have

$$T^{-1}\left\|\beta\bar{V}\right\|^2=\lambda_1(T^{-1}\beta\bar{V}\bar{V}^\intercal\beta^\intercal)=\lambda_1(\beta\beta^\intercal)+O_{\mathrm{P}}(\sigma_1(\beta)^2T^{-1/2})=\sigma_1(\beta)^2+O_{\mathrm{P}}(\sigma_1(\beta)^2T^{-1/2}).$$

Plugging this result into (2.176), we have $\left\|\mathbb{P}_{\tilde{\xi}}-T^{-1}\bar{V}^\intercal\mathbb{P}_b\bar{V}\right\|\lesssim_{\mathrm{P}}T^{-1/2}$.

□

Lemmas 24-26 below are concerned with the singular values and singular vectors of $T^{-1/2}\bar{R}$. We use $\widehat{\varsigma}_j$, $\widehat{\xi}_j$ and $\widehat{\lambda}_j^{1/2}$, $j\leq p$ to denote them throughout Lemmas 24-26.

Lemma 24. *Under the assumptions of Theorem 9(a), we have*

$$\frac{\widehat{\lambda}_j}{\lambda_j}-1\lesssim_{\mathrm{P}}\lambda_j^{-1/2}(T^{-1/2}N^{1/2}+1)+T^{-1/2},$$

where $\lambda_j=\lambda_j(\beta^\intercal\beta)$ *and* $\widehat{\lambda}_j=\lambda_j(T^{-1}\bar{R}\bar{R}^\intercal)$.

Proof. Since $\lambda_j\left(\beta\bar{V}\bar{V}^\intercal\beta^\intercal\right)=\lambda_j\left(\beta^\intercal\beta\bar{V}\bar{V}^\intercal\right)$, we have

$$\lambda_j\left(\beta^\intercal\beta\right)\lambda_p\left(\frac{\bar{V}\bar{V}^\intercal}{T}\right)\leq\frac{\lambda_j\left(\beta^\intercal\beta\bar{V}\bar{V}^\intercal\right)}{T}\leq\lambda_j\left(\beta^\intercal\beta\right)\lambda_1\left(\frac{\bar{V}\bar{V}^\intercal}{T}\right). \tag{2.177}$$

By Lemma 14(i) and Weyl's inequality, we have $\lambda_j\left(T^{-1}\bar{V}\bar{V}^\intercal\right)-1\lesssim_{\mathrm{P}}T^{-1/2}$ for $j\leq p$. Then, (2.177) becomes

$$\frac{\lambda_j\left(\beta\bar{V}\bar{V}^\intercal\beta^\intercal\right)}{T\lambda_j(\beta^\intercal\beta)}-1\lesssim_{\mathrm{P}}T^{-1/2},$$

which is equivalent to

$$\frac{\sigma_j\left(\beta\bar{V}\right)}{\sqrt{T}\sigma_j(\beta)} - 1 \lesssim_{\mathrm{P}} T^{-1/2}. \tag{2.178}$$

Using Weyl's inequality again, we have $|\sigma_j\left(\bar{R}\right) - \sigma_j\left(\beta\bar{V}\right)| \leq \left\|\bar{U}\right\| \lesssim_{\mathrm{P}} N^{1/2} + T^{1/2}$, which is equivalent to

$$\frac{\widehat{\lambda}_j^{1/2}}{\lambda_j^{1/2}} - \frac{\sigma_j(\beta\bar{V})}{\sqrt{T}\sigma_j(\beta)} \lesssim_{\mathrm{P}} \frac{1}{\sqrt{T}} + \frac{\sqrt{N}+\sqrt{T}}{\sqrt{T\lambda_j}}. \tag{2.179}$$

Combine (2.178) and (2.179), we complete the proof. □

Lemma 25. *Suppose that the SVD of β is given by:*

$$\beta = \Gamma^{\intercal} \begin{pmatrix} \Lambda^{\frac{1}{2}} \\ 0 \end{pmatrix} H, \tag{2.180}$$

where $\Gamma \in \mathbb{R}^{N\times N}$, $H \in \mathbb{R}^{p\times p}$ are orthogonal matrices, and Λ is a diagonal matrix of the eigenvalues of $\beta^{\intercal}\beta$. If we write $\Gamma\widehat{\varsigma}_j = (s_{j1}^{\intercal}, s_{j2}^{\intercal})^{\intercal}$, where $s_{j1} \in \mathbb{R}^p$, $s_{j2} \in \mathbb{R}^{N-p}$. Then under the assumptions of Theorem 9(a), we have

(i) $\left\|\left(\Lambda/\lambda_j\right)^{1/2}\left(s_{j1} - \langle s_{j1}, e_{j1}\rangle e_{j1}\right)\right\| \lesssim_{\mathrm{P}} \lambda_j^{-1/2}(T^{-1/2}N^{1/2}+1)$, *where e_{i1} is a $p\times 1$ unit vector with the ith entry being equal to 1.*

(ii) $\left\|s_{j1} - \langle s_{j1}, e_{j1}\rangle e_{j1}\right\| \lesssim_{\mathrm{P}} \lambda_j^{-1/2}(T^{-1/2}N^{1/2}+1)$.

(iii) $\left\|\left(\Lambda/\lambda_j\right)^{1/2} s_{j1}\right\| \lesssim_{\mathrm{P}} 1$.

(iv) $\left\|s_{j2}\right\| \lesssim_{\mathrm{P}} \lambda_j^{-1/2}(T^{-1/2}N^{1/2}+1)$.

Proof. With the orthogonal matrix Γ defined above, we can write

$$\widetilde{U} = \Gamma \bar{U} = \begin{pmatrix} \widetilde{U}_{1_{p\times T}} \\ \widetilde{U}_{2_{(N-p)\times T}} \end{pmatrix}, \tag{2.181}$$

so that

$$\Gamma \bar{R} = \begin{pmatrix} \Lambda^{\frac{1}{2}} \\ 0 \end{pmatrix} \bar{V} + \tilde{U} = \begin{pmatrix} \Lambda^{\frac{1}{2}} \bar{V} + \widetilde{U}_1 \\ \widetilde{U}_2 \end{pmatrix}.$$

The relationship between singular vectors $\widehat{\varsigma}_j$ and $\widehat{\xi}_j$ can be written as

$$\Gamma \widehat{\varsigma}_j = \frac{\left(\Gamma \bar{R}\right) \widehat{\xi}_j}{\sqrt{T\widehat{\lambda}_j}}, \quad \widehat{\xi}_j = \frac{\left(\Gamma \bar{R}\right)^{\intercal} \left(\Gamma \widehat{\varsigma}_j\right)}{\sqrt{T\widehat{\lambda}_j}}. \tag{2.182}$$

Specifically, we have

$$s_{j1} = \frac{\left(\Lambda^{\frac{1}{2}} \bar{V} + \widetilde{U}_1\right) \widehat{\xi}_j}{\sqrt{T\widehat{\lambda}_j}}, \quad s_{j2} = \frac{\widetilde{U}_2 \widehat{\xi}_j}{\sqrt{T\widehat{\lambda}_j}}, \quad \widehat{\xi}_j = \frac{\left(\Lambda^{\frac{1}{2}} \bar{V} + \widetilde{U}_1\right)^{\intercal} s_{j1} + \widetilde{U}_2^{\intercal} s_{j2}}{\sqrt{T\widehat{\lambda}_j}}. \tag{2.183}$$

From (2.183), we have

$$\left(\Lambda^{\frac{1}{2}} \bar{V} + \widetilde{U}_1\right) \left(\Lambda^{\frac{1}{2}} \bar{V} + \widetilde{U}_1\right)^{\intercal} s_{j1} + \left(\Lambda^{\frac{1}{2}} \bar{V} + \widetilde{U}_1\right) \widetilde{U}_2^{\intercal} s_{j2} = T\widehat{\lambda}_j s_{j1}. \tag{2.184}$$

We can rewrite (2.184) as

$$\begin{aligned} \left(\mathbb{I}_p - \frac{\Lambda}{\lambda_j}\right) s_{j1} =& \frac{1}{T\lambda_j} \left(\Lambda^{\frac{1}{2}} \bar{V} + \widetilde{U}_1\right) \widetilde{U}_2^{\intercal} s_{j2} + \frac{1}{\lambda_j} \Lambda^{\frac{1}{2}} \left(\frac{\bar{V}\bar{V}^{\intercal}}{T} - I\right) \Lambda^{\frac{1}{2}} s_{j1} + \frac{\Lambda^{\frac{1}{2}} \bar{V} \widetilde{U}_1^{\intercal}}{T\lambda_j} s_{j1} \\ &+ \frac{\widetilde{U}_1 \bar{V}^{\intercal} \Lambda^{\frac{1}{2}}}{T\lambda_j} s_{j1} + \frac{\widetilde{U}_1 \widetilde{U}_1^{\intercal}}{T\lambda_j} s_{j1} - \left(\frac{\widehat{\lambda}_j}{\lambda_j} - 1\right) s_{j1}. \end{aligned} \tag{2.185}$$

Define $L = \mathrm{diag}(l_1, \ldots, l_p)$, where l_i is equal to $\lambda_j/(\lambda_j - \lambda_i)$ if $i \neq j$ and 0 otherwise.

By left multiplying L to both sides of (2.185), we have

$$\begin{aligned} s_{j1} - \langle s_{j1}, e_{j1}\rangle e_{j1} =& \frac{1}{T\lambda_j} L\Lambda^{\frac{1}{2}} \bar{V} \frac{\widetilde{U}_2^\intercal \widetilde{U}_2}{\sqrt{T\widehat{\lambda}_j}} \widehat{\xi}_j + \frac{1}{T\lambda_j} L\widetilde{U}_1 \frac{\widetilde{U}_2^\intercal \widetilde{U}_2}{\sqrt{T\widehat{\lambda}_j}} \widehat{\xi}_j + \frac{1}{\lambda_j} L\Lambda^{\frac{1}{2}} \left(\frac{\bar{V}\bar{V}^\intercal}{T} - \mathbb{I}_p \right) \Lambda^{\frac{1}{2}} s_{j1} \\ &+ \frac{L\Lambda^{\frac{1}{2}} \bar{V} \widetilde{U}_1^\intercal}{T\lambda_j} s_{j1} + L \frac{\widetilde{U}_1 \bar{V}^\intercal \Lambda^{\frac{1}{2}}}{T\lambda_j} s_{j1} + L \frac{\widetilde{U}_1 \widetilde{U}_1^\intercal}{T\lambda_j} s_{j1} - \left(\frac{\widehat{\lambda}_j}{\lambda_j} - 1 \right) L s_{j1}. \end{aligned} \tag{2.186}$$

Now left multiplying $\left(\frac{\Lambda}{\lambda_j}\right)^{\frac{1}{2}}$ again, we have

$$\left(\frac{\Lambda}{\lambda_j} \right)^{\frac{1}{2}} \left(s_{j1} - \langle s_{j1}, e_{j1}\rangle e_{j1} \right) \tag{2.187}$$

$$\begin{aligned} =& \frac{1}{T\lambda_j^{3/2}} \Lambda^{\frac{1}{2}} L\Lambda^{\frac{1}{2}} \bar{V} \frac{\widetilde{U}_2^\intercal \widetilde{U}_2}{\sqrt{T\widehat{\lambda}_j}} \widehat{\xi}_j + \frac{1}{T\lambda_j^{3/2}} \Lambda^{\frac{1}{2}} L\widetilde{U}_1 \frac{\widetilde{U}_2^\intercal \widetilde{U}_2}{\sqrt{T\widehat{\lambda}_j}} \widehat{\xi}_j \\ &+ \frac{1}{\lambda_j} \Lambda^{\frac{1}{2}} L\Lambda^{\frac{1}{2}} \left(\frac{\bar{V}\bar{V}^\intercal}{T} - \mathbb{I}_p \right) \left(\frac{\Lambda}{\lambda_j} \right)^{\frac{1}{2}} s_{j1} + \Lambda^{\frac{1}{2}} L\Lambda^{\frac{1}{2}} \frac{\bar{V}\widetilde{U}_1^\intercal}{T\lambda_j^{3/2}} s_{j1} \\ &+ \Lambda^{\frac{1}{2}} L \frac{\widetilde{U}_1 \bar{V}^\intercal}{T\lambda_j} \left(\frac{\Lambda}{\lambda_j} \right)^{\frac{1}{2}} s_{j1} + \Lambda^{\frac{1}{2}} L \frac{\widetilde{U}_1 \widetilde{U}_1^\intercal}{T\lambda_j^{3/2}} s_{j1} - \left(\frac{\widehat{\lambda}_j}{\lambda_j} - 1 \right) \left(\frac{\Lambda}{\lambda_j} \right)^{\frac{1}{2}} L s_{j1} \\ =& K_1 + K_2 + K_3 + K_4 + K_5 + K_6 + K_7. \end{aligned} \tag{2.188}$$

Before we analyze these seven terms in (2.187), we first analyze $\|L\|$, $\left\|L\Lambda^{1/2}\right\|$ and $\|L\Lambda\|$. Since L and Λ are diagonal matrices, by Assumption 18 we can easily show that

$$\|L\| \lesssim 1, \quad \left\|L\Lambda^{1/2}\right\| \lesssim \lambda_j^{1/2}, \quad \|L\Lambda\| \lesssim \lambda_j. \tag{2.189}$$

In addition, Lemma 14(ii)(iii)(v) imply that

$$\begin{aligned}&\left\|\widetilde{U}_1\right\| = \left\|(\beta^\intercal\beta)^{-1/2}\beta^\intercal\bar{U}\right\| \lesssim_{\mathrm{P}} T^{1/2}, \quad \left\|\widetilde{U}_1\bar{V}^\intercal\right\| = \left\|(\beta^\intercal\beta)^{-1/2}\beta^\intercal\bar{U}\bar{V}^\intercal\right\| \lesssim_{\mathrm{P}} T^{1/2}, \\ &\left\|\widetilde{U}_2\right\| \leq \left\|\bar{U}\right\| \lesssim_{\mathrm{P}} N^{1/2} + T^{1/2}.\end{aligned} \tag{2.190}$$

Using Lemma 14(i)(vi), Lemma 24, (2.189) and (2.190), we analyze these seven terms in (2.187) one by one. For the first term, we have

$$\|K_1\| \leq T^{-3/2}\lambda_j^{-3/2}\widehat{\lambda}_j^{-1/2}\left\|L\Lambda\right\|\left\|\bar{V}\right\|\left\|\widetilde{U}_2^\intercal\widetilde{U}_2\right\|\left\|\widehat{\xi}_j\right\| \lesssim_{\mathrm{P}} \lambda_j^{-1}(T^{-1}N+1),$$

where we also use $\left\|\widetilde{U}_2^\intercal\widetilde{U}_2\right\| \leq \left\|\bar{U}^\intercal\bar{U}\right\| \lesssim_{\mathrm{P}} N+T$ in the last equation. For the second term, we have

$$\|K_2\| \leq T^{-3/2}\lambda_j^{-3/2}\widehat{\lambda}_j^{-1/2}\left\|\Lambda^{1/2}L\right\|\left\|\widetilde{U}_1\right\|\left\|\widetilde{U}_2^\intercal\widetilde{U}_2\right\|\left\|\widehat{\xi}_j\right\| \lesssim_{\mathrm{P}} \lambda_j^{-3/2}(T^{-1}N+1).$$

For the third term, we have

$$\|K_3\| \leq \lambda_j^{-1}\left\|L\Lambda\right\|\left\|T^{-1}\bar{V}\bar{V}^\intercal - \mathbb{I}_p\right\|\left\|(\Lambda/\lambda_j)^{1/2}s_{j1}\right\| \lesssim_{\mathrm{P}} T^{-1/2}\left\|(\Lambda/\lambda_j)^{1/2}s_{j1}\right\|.$$

For the forth term, we have

$$\|K_4\| \leq T^{-1}\lambda_j^{-3/2}\left\|L\Lambda\right\|\left\|\bar{V}\widetilde{U}_1^\intercal\right\| \lesssim_{\mathrm{P}} \lambda_j^{-1/2}T^{-1/2},$$

where we use $\left\|\bar{V}\widetilde{U}_1^\intercal\right\| \lesssim_{\mathrm{P}} T^{1/2}$ from Lemma 14. For the fifth term, we have

$$\|K_5\| \leq T^{-1}\lambda_j^{-1}\left\|L\Lambda^{1/2}\right\|\left\|\widetilde{U}_1\bar{V}^\intercal\right\|\left\|(\Lambda/\lambda_j)^{1/2}s_{j1}\right\| \lesssim_{\mathrm{P}} \lambda_j^{-1/2}T^{-1/2}\left\|(\Lambda/\lambda_j)^{1/2}s_{j1}\right\|.$$

For the sixth term, we have

$$\|K_6\| \leq T^{-1}\lambda_j^{-3/2} \left\|L\Lambda^{1/2}\right\| \left\|\widetilde{U}_1\widetilde{U}_1^\intercal\right\| \lesssim_{\mathrm{P}} \lambda_j^{-1},$$

where we use $\left\|\widetilde{U}_1\widetilde{U}_1^\intercal\right\| \lesssim_{\mathrm{P}} T$ as shown in Lemma 14. For the last term, we have

$$\|K_7\| \leq \lambda_j^{-2}|\widehat{\lambda}_j - \lambda_j| \left\|L\Lambda^{1/2}\right\| \lesssim_{\mathrm{P}} \lambda_j^{-1/2}(T^{-1/2}N^{1/2}+1) + T^{-1/2}.$$

To sum up, (2.187) gives

$$\begin{aligned} &\left\|(\Lambda/\lambda_j)^{1/2}\left(s_{j1} - \langle s_{j1}, e_{j1}\rangle e_{j1}\right)\right\| \\ \lesssim_{\mathrm{P}} &\lambda_j^{-1/2}(T^{-1/2}N^{1/2}+1) + T^{-1/2} + T^{-1/2}\left\|(\Lambda/\lambda_j)^{1/2}s_{j1}\right\|. \end{aligned} \tag{2.191}$$

Note that

$$\begin{aligned} \left\|(\Lambda/\lambda_j)^{1/2}s_{j1}\right\| &\leq \left\|(\Lambda/\lambda_j)^{1/2}\left(s_{j1} - \langle s_{j1}, e_{j1}\rangle e_{j1}\right)\right\| + \left\|(\Lambda/\lambda_j)^{1/2}\langle s_{j1}, e_{j1}\rangle e_{j1}\right\| \\ &\leq \left\|(\Lambda/\lambda_j)^{1/2}\left(s_{j1} - \langle s_{j1}, e_{j1}\rangle e_{j1}\right)\right\| + |\langle s_{j1}, e_{j1}\rangle|\sqrt{\lambda_j^{-1}e_{j1}^\intercal\Lambda e_{j1}} \\ &= \left\|(\Lambda/\lambda_j)^{1/2}\left(s_{j1} - \langle s_{j1}, e_{j1}\rangle e_{j1}\right)\right\| + O_{\mathrm{P}}(1). \end{aligned}$$

Plugging this into (2.191), we have

$$\left\|(\Lambda/\lambda_j)^{1/2}\left(s_{j1} - \langle s_{j1}, e_{j1}\rangle e_{j1}\right)\right\| \lesssim_{\mathrm{P}} \lambda_j^{-1/2}(T^{-1/2}N^{1/2}+1) + T^{-1/2}, \tag{2.192}$$

which in turn leads to $\left\|(\Lambda/\lambda_j)^{1/2}s_{j1}\right\| \lesssim_{\mathrm{P}} 1$ as by assumption $\lambda_j^{-1/2}(T^{-1/2}N^{1/2}+1) \to 0$. Similarly, we can analyze corresponding terms in (2.186), and obtain

$$\begin{aligned} \left\|s_{j1} - \langle s_{j1}, e_{j1}\rangle e_{j1}\right\| &\lesssim_{\mathrm{P}} T^{-1/2}\left\|(\Lambda/\lambda_j)^{1/2}s_{j1}\right\| + \lambda_j^{-1/2}(T^{-1/2}N^{1/2}+1) \\ &\lesssim_{\mathrm{P}} \lambda_j^{-1/2}(T^{-1/2}N^{1/2}+1) + T^{-1/2}. \end{aligned}$$

From (2.183), we have

$$\|s_{j2}\| \leq \left\| \frac{\widetilde{U}_2}{\sqrt{T\lambda_j}} \right\| \left\| \left(\frac{\lambda_j}{\widehat{\lambda}_j} \right)^{\frac{1}{2}} \right\| \left\| \widehat{\xi}_j \right\| \lesssim_{\mathrm{P}} \lambda_j^{-1/2} (T^{-1/2} N^{1/2} + 1). \tag{2.193}$$

This concludes the proof. □

Lemma 26. *Under the assumptions of Theorem 9(a), we have*

(i) $\left\| \frac{\widehat{\xi}_i^{\intercal} \bar{U}^{\intercal} \widehat{\varsigma}_j}{\sqrt{T\widehat{\lambda}_j}} \right\| \lesssim_{\mathrm{P}} \frac{1}{T} + \frac{N+T}{T\lambda_i} + \frac{N+T}{T\lambda_j}.$

(ii) $\left\| \frac{\bar{V}\bar{U}^{\intercal}\widehat{\varsigma}_i}{T\sqrt{\widehat{\lambda}_i}} \right\| \lesssim_{\mathrm{P}} \frac{1}{T} + \frac{N+T}{T\lambda_i}, \quad |\frac{\widehat{\varsigma}_i^{\intercal}\bar{u}}{\sqrt{\widehat{\lambda}_i}}| \lesssim_{\mathrm{P}} \frac{1}{T} + \frac{N+T}{T\lambda_i}.$

(iv) $\left\| \frac{\widehat{\varsigma}_i^{\intercal}\bar{U}}{\sqrt{T\widehat{\lambda}_i}} \right\| \lesssim_{\mathrm{P}} \frac{1}{\sqrt{\lambda_i}} + \frac{N+T}{T\lambda_i}.$

Proof. (i) From (2.182), we have

$$\frac{\widehat{\xi}_i^{\intercal} \bar{U}^{\intercal} \widehat{\varsigma}_j}{\sqrt{T\widehat{\lambda}_j}} = \frac{\widehat{\varsigma}_i^{\intercal} \bar{R} \bar{U}^{\intercal} \widehat{\varsigma}_j}{T\sqrt{\widehat{\lambda}_i \widehat{\lambda}_j}}.$$

Using the orthogonal matrix Γ and the notations in Lemma 24 and Lemma 25, we have

$$\begin{aligned}
\widehat{\varsigma}_i^{\intercal} \bar{R} \bar{U}^{\intercal} \widehat{\varsigma}_j = s_i^{\intercal} \left(\Gamma \beta \bar{V} + \widetilde{U} \right) \widetilde{U}^{\intercal} s_j =& s_{i1}^{\intercal} \left(\Lambda^{\frac{1}{2}} \bar{V} + \widetilde{U}_1 \right) \widetilde{U}_1^{\intercal} s_{j1} + s_{i2}^{\intercal} \widetilde{U}_2 \widetilde{U}_1^{\intercal} s_{j1} \\
&+ s_{i1}^{\intercal} \left(\Lambda^{\frac{1}{2}} \bar{V} + \widetilde{U}_1 \right) \widetilde{U}_2^{\intercal} s_{j2} + s_{i2}^{\intercal} \widetilde{U}_2 \widetilde{U}_2^{\intercal} s_{j2} \\
=& K_1 + K_2 + K_3 + K_4.
\end{aligned}$$

Recall that from Lemma 25, we have $\left\| (\Lambda/\lambda_j)^{1/2} s_{j1} \right\| \lesssim_{\mathrm{P}} 1$. Using this result and Lemma 14, we analyze these four terms one by one. For the first term, we have

$$\|K_1\| \leq \left\| s_{i1}^{\intercal} \Lambda^{\frac{1}{2}} \right\| \left\| \bar{V} \widetilde{U}_1^{\intercal} \right\| \|s_{j1}\| + \|s_{i1}\| \left\| \widetilde{U}_1 \widetilde{U}_1^{\intercal} \right\| \|s_{j1}\| \lesssim_{\mathrm{P}} \sqrt{\lambda_i T} + T.$$

For the second term, we have

$$\|K_2\| \le \|s_{i2}\| \left\|\widetilde{U}_2\right\| \left\|\widetilde{U}_1\right\| \lesssim_{\mathrm{P}} \sqrt{\frac{N+T}{T\lambda_i}} \left(\sqrt{N}+\sqrt{T}\right)\sqrt{T} \lesssim_{\mathrm{P}} \lambda_i^{-1/2}(N+T).$$

For the third term, we have

$$\begin{aligned}\|K_3\| &\le \left(\left\|s_{i1}^{\intercal}\Lambda^{\frac{1}{2}}\right\| \|\bar{V}\| + \left\|\widetilde{U}_1\right\|\right) \left\|\widetilde{U}_2\right\| \|s_{j2}\| \\ &\lesssim_{\mathrm{P}} \sqrt{\lambda_i T}\left(\sqrt{N}+\sqrt{T}\right)\sqrt{\frac{N+T}{T\lambda_j}} = \lambda_j^{-1/2}\lambda_i^{1/2}(N+T).\end{aligned}$$

For the last term, we have

$$\|K_4\| \le \left\|\widetilde{U}_2\widetilde{U}_2^{\intercal}\right\| \|s_{i2}\| \left\|s_{j2}\right\| \lesssim_{\mathrm{P}} \lambda_i^{-1/2}\lambda_j^{-1/2}T^{-1}(N+T)^2.$$

Using above equations and Lemma 24, we get

$$\left\|\frac{\widehat{\xi}_i^{\intercal}\bar{U}^{\intercal}\widehat{\varsigma}_j}{\sqrt{T\widehat{\lambda}_j}}\right\| = \left\|\frac{\widehat{\varsigma}_i^{\intercal}\bar{R}\bar{U}^{\intercal}\widehat{\varsigma}_j}{T\sqrt{\widehat{\lambda}_i\widehat{\lambda}_j}}\right\| \lesssim_{\mathrm{P}} \frac{1}{T} + \frac{N+T}{T\lambda_i} + \frac{N+T}{T\lambda_j}.$$

(ii) Using $\bar{U}^{\intercal}\widehat{\varsigma}_i = \widetilde{U}_1^{\intercal}s_{i1} + \widetilde{U}_2^{\intercal}s_{i2}$ and (2.190), we have

$$\left\|\bar{V}\bar{U}^{\intercal}\widehat{\varsigma}_i\right\| \le \left\|\bar{V}\widetilde{U}_1^{\intercal}s_{i1}\right\| + \left\|\bar{V}\widetilde{U}_2^{\intercal}s_{i2}\right\| \le \left\|\bar{V}\widetilde{U}_1^{\intercal}\right\| + \|\bar{V}\|\,\|\bar{U}\|\,\|s_{i2}\| \lesssim_{\mathrm{P}} \sqrt{T} + \frac{N+T}{\sqrt{\lambda_i}}.$$

Then, with Lemma 24, we have $\left\|T^{-1}\widehat{\lambda}_i^{-1/2}\bar{V}\bar{U}^{\intercal}\widehat{\varsigma}_i\right\| \lesssim_{\mathrm{P}} T^{-1} + \lambda_i^{-1}(T^{-1}N+1)$.

Replace $\bar{V}$ in the above proof by $\iota_T^{\intercal}$, we can get $\left\|\widehat{\lambda}_i^{-1/2}\bar{u}^{\intercal}\widehat{\varsigma}_i\right\| \lesssim_{\mathrm{P}} T^{-1} + \lambda_i^{-1}(T^{-1}N+1)$.

(iii) Using $\bar{U}^{\intercal}\widehat{\varsigma}_i = \widetilde{U}_1^{\intercal}s_{i1} + \widetilde{U}_2^{\intercal}s_{i2}$ and (2.190), we have

$$\left\|\widehat{\varsigma}_i^{\intercal}\bar{U}\right\| \le \left\|s_{i1}^{\intercal}\widetilde{U}_1\right\| + \left\|s_{i2}^{\intercal}\widetilde{U}_2\right\| \le \left\|\widetilde{U}_1\right\| + \|\bar{U}\| \lesssim_{\mathrm{P}} \sqrt{T} + \frac{N+T}{\sqrt{T\lambda_i}}.$$

Applying Lemma 24 again completes the proof. □

Lemma 27. *Under the assumptions of Theorem 9(a),* $\tilde{H}_1$, $\tilde{H}_2$ *defined by* (2.73) *satisfy*

(i) $\left\|\tilde{H}_1\right\| \lesssim_{\mathrm{P}} 1$, $\left\|\tilde{H}_2\right\| \lesssim_{\mathrm{P}} 1$.

(ii) $\left\|\tilde{H}_1^{\intercal}\tilde{H}_2 - \mathbb{I}_{\tilde{p}}\right\| \lesssim_{\mathrm{P}} T^{-1} + \lambda_p^{-1}(T^{-1}N + 1)$.

(iii) $\left\|\tilde{H}_1 - \tilde{H}_2\right\| \lesssim_{\mathrm{P}} T^{-1/2} + \lambda_p^{-1}(T^{-1}N + 1)$.

Proof. (i) Using the definition of $\tilde{H}_1$ in (2.73) and Lemma 14, we have

$$\left\|\tilde{h}_{k1}\right\| = \left\|\frac{\bar{V}\widehat{\xi}_k}{\sqrt{T}}\right\| \leq T^{-1/2}\left\|\bar{V}\right\| \lesssim_{\mathrm{P}} 1,$$

which leads to $\left\|\tilde{H}_1\right\| \lesssim_{\mathrm{P}} 1$.

Using $\Gamma\widehat{\varsigma}_k = (s_{k1}^{\intercal}, s_{k2}^{\intercal})^{\intercal}$, the SVD of β in (2.180), the definition of $\tilde{H}_2$ in (2.73), Lemma 24 and Lemma 25(iii), we have

$$\left\|\tilde{h}_{k2}\right\| = \left\|\frac{\beta^{\intercal}\widehat{\varsigma}_k}{\sqrt{\widehat{\lambda}_k}}\right\| = \left\|\frac{\Lambda^{1/2}s_{k1}}{\sqrt{\widehat{\lambda}_k}}\right\| \lesssim_{\mathrm{P}} 1, \tag{2.194}$$

which leads to $\left\|\tilde{H}_2\right\| \lesssim_{\mathrm{P}} 1$.

(ii) By (2.182) and Lemma 15, for $l, k \leq p$, we have

$$\delta_{lk} = \widehat{\xi}_l^{\intercal}\widehat{\xi}_k = \frac{\widehat{\xi}_l^{\intercal}\bar{V}^{\intercal}\beta^{\intercal}\widehat{\varsigma}_k}{\sqrt{T\widehat{\lambda}_k}} + \frac{\widehat{\xi}_l^{\intercal}\bar{U}^{\intercal}\widehat{\varsigma}_k}{\sqrt{T\widehat{\lambda}_k}} = \tilde{h}_{l1}^{\intercal}\tilde{h}_{k2} + \frac{\widehat{\xi}_l^{\intercal}\bar{U}^{\intercal}\widehat{\varsigma}_k}{\sqrt{T\widehat{\lambda}_k}}.$$

By Lemma 26(i), we have

$$|\tilde{h}_{l1}^{\intercal}\tilde{h}_{k2} - \delta_{lk}| \lesssim_{\mathrm{P}} \frac{1}{T} + \frac{N+T}{T\min\{\lambda_l, \lambda_k\}} \leq \frac{1}{T} + \frac{N+T}{T\lambda_p},$$

and thus $\left\|\tilde{H}_1^{\intercal}\tilde{H}_2 - \mathbb{I}_p\right\| \lesssim_{\mathrm{P}} T^{-1} + \lambda_p^{-1}(T^{-1}N + 1)$.

(iii) Using (2.182), we have

$$\bar{V}\widehat{\xi}_k = \frac{\bar{V}\bar{V}^\intercal\beta^\intercal}{\sqrt{T\widehat{\lambda}_k}}\widehat{\varsigma}_k + \frac{\bar{V}\bar{U}^\intercal\widehat{\varsigma}_k}{\sqrt{T\widehat{\lambda}_k}}.$$

With the definition of h_{k1} and h_{k2}, it becomes

$$\tilde{h}_{k1} = \frac{\bar{V}\bar{V}^\intercal}{T}\tilde{h}_{k2} + \frac{\bar{V}\bar{U}^\intercal\widehat{\varsigma}_k}{T\sqrt{\widehat{\lambda}_k}}. \tag{2.195}$$

With $\left\|\tilde{h}_{k2}\right\| \lesssim_{\mathrm{P}} 1$, Lemma 14 and Lemma 26(ii), (2.195) leads to

$$\left\|\tilde{h}_{k1} - \tilde{h}_{k2}\right\| \leq \left\|T^{-1}\bar{V}\bar{V}^\intercal - \mathbb{I}_p\right\| \left\|\tilde{h}_{k2}\right\| + \left\|\frac{\bar{V}\bar{U}^\intercal\widehat{\varsigma}_k}{T\sqrt{\widehat{\lambda}_k}}\right\| \lesssim_{\mathrm{P}} T^{-1/2} + \lambda_p^{-1}(T^{-1}N + 1),$$

which concludes the proof of (iii). □

Lemma 28. *Under Assumption 18, we have*

$$\left\|\bar{r} - \widehat{\Sigma}b\right\|_\infty \lesssim_{\mathrm{P}} \sqrt{\frac{\log N}{T}}, \qquad \|b^\intercal(\bar{r} - \mathrm{E}(r_t))\| \lesssim_{\mathrm{P}} \frac{1}{\sqrt{T}}.$$

Proof. For the first inequality, we have

$$\left\|\bar{r} - \widehat{\Sigma}b\right\|_\infty \leq \|\bar{r} - \mathrm{E}(r)\|_\infty + \left\|\Sigma b - \widehat{\Sigma}b\right\|_\infty \lesssim_{\mathrm{P}} \sqrt{\frac{\log N}{T}},$$

where we use large deviation inequalities in Assumption 17:

$$\|\bar{r} - \mathrm{E}(r_t)\|_\infty \lesssim_{\mathrm{P}} \sqrt{\frac{\log N}{T}} \text{ and } \left\|\Sigma b - \widehat{\Sigma}b\right\|_\infty = \left\|\frac{1}{T}\bar{R}\bar{R}^\intercal b - \mathrm{Cov}(r_t, r_t^\intercal b)\right\|_\infty \lesssim_{\mathrm{P}} \sqrt{\frac{\log N}{T}}.$$

The second inequality follows immediately from Assumption 17:

$$\|b^{\intercal}(\bar{r} - \mathrm{E}(r_t))\| = |\frac{1}{T}\sum_{t=1}^{T} m_t - \mathrm{E}(m_t)| \lesssim_{\mathrm{P}} \frac{1}{\sqrt{T}}.$$

□

CHAPTER 3

EMPIRICAL ANALYSIS WITH SUPERVISED PRINCIPAL COMPONENTS

3.1 Introduction

This chapter illustrates the use of SPCA with two empirical applications. In Section 3.2, we explore its application to macroeconomic forecasting. For this purpose, we combine the standard Fred-Md dataset of 127 macroeconomic and financial variables with the Blue Chip Financial Forecasts dataset, that contains hundreds of forecasts of various variables (like interest rates and inflation) from professional forecasters, thus obtaining a large dataset of predictors. We then apply different prediction and dimension reduction methods to forecast quarterly inflation, industrial production growth, and changes in unemployment. We compare the results using SPCA to those obtained using PCA (as in Stock and Watson [2002a]) and PLS (as in Kelly and Pruitt [2013]). We show that in a setting with a large number of (potentially noisy and/or redundant) predictors, SPCA performs well in forecasting macroeconomic quantities out of sample. We also investigate the selection that SPCA operates, and find that it isolates, for each target, a different group of useful predictors; it also focuses on a few financial forecasters, whose survey responses are selected particularly often. Finally, we illustrate the use of SPCA with multiple targets at the same time (macroeconomic variables forecasted at different horizons: 1, 2, 3, 6 and 12 months).

In Section 3.3, we illustrate the use of SPCA in estimating risk premia of a variety of tradable and nontradable factors proposed in the asset pricing literature, and for diagnosing observable factor models. We use the large cross-section of test portfolios produced by Chen and Zimmermann [2020] and Hou et al. [2020], covering more than 900 and 1600 portfolios, respectively, for the period 1976-2020. We apply SPCA to estimate factor risk premia, and evaluate its out-of-sample performance. Almost none of the non-tradable factors are priced,

except for the intermediary capital factor. We also explore the robustness of SPCA to the weakness of factors, by artificially changing the set of test assets used in the estimation: for example, we show that SPCA is able to recover the risk premium for momentum even when momentum assets are removed from the original set of test assets (and therefore the momentum factor is weak in the cross-section). Moreover, we illustrate empirically how SPCA can be used to diagnose whether observable factor models are missing important priced factors.

3.2 Macroeconomic Prediction

In this section we apply the SPCA methodology developed in Giglio et al. [2023] to a standard macroeconomic prediction exercise, using a large set of predictors to forecast inflation, industrial production, and unemployment.

3.2.1 Empirical Context

Predicting macroeconomic variables like output and inflation is a central exercise in empirical macroeconomics. The availability of large macroeconomic datasets that contain many potentially useful predictors has spurred the application of a variety of methods of dimension reduction to this objective. Some of these methods, like those based on principal component analysis (PCA), reduce the dimensionality of the predictors universe without using information in the target of the forecast (see Stock and Watson [2002b]). Others instead use information from the target to help the dimension reduction focus on the most valuable predictors; examples include partial least squares (PLS, Kelly and Pruitt [2015]), targeted PCA (Bai and Ng [2008]), and scaled PCA (Huang et al. [2022]). SPCA belongs to the latter group, as it employs an iterative screening step based on correlation with the target to eliminate useless or noisy predictors.

Because the selection step is designed to eliminate irrelevant predictors (as opposed to

downweight them as, for example, PLS does) we expect SPCA to perform best when faced with a large number of predictors that are potentially irrelevant, noisy, or redundant. In our empirical analysis, we therefore explore a context in which a large number of predictors are available to be used for forecasting. Specifically, we include in our set of predictors not only a standard panel of macroeconomic variables, but also a large dataset of individual forecasts of different macroeconomic quantities by professional forecasters. Macroeconomic forecasts have often been included in forecasting exercises, either by using the consensus forecast as an additional predictor (Faust and Wright [2013]) or in the context of optimal forecast combination (Genre et al. [2013]). In our context, we let SPCA decide if and which individual forecasts to use to complement the macroeconomic predictors – so the forecast combination will be decided automatically by SPCA.

3.2.2 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.[1]

1. 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)[2], CPI: All Items Less Medical Care (CUSR0000SA0L5), Personal Cons. Exp: Chain Index (PCEPI), Personal

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[3] and provides forecasts of interest rates and many other macroeconomic quantities[4] 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.

3.2.3 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

Cons. Exp: Durable Goods (DDURRG3M086SBEA), Personal Cons. Exp: Nondurable Goods (DNDGRG3M086SBEA), Personal Cons. Exp: Services (DSERRG3M086SBEA), Avg Hourly Earnings: Goods-Producing (CES0600000008), Avg 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).

3. For instance, Bank of America, Goldman Sachs & Co. and J.P. MorganChase.

4. 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).

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 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.[5]

Recall that the SPCA procedure presented in Giglio et al. [2023] 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

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

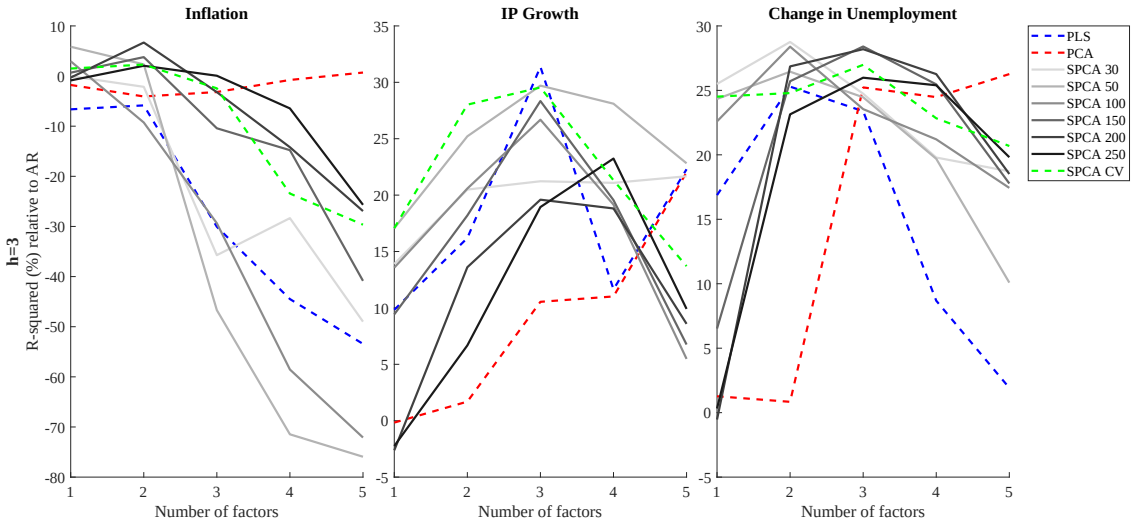
of $\lfloor qN \rfloor$ from 50 to 300.

3.2.4 Results

3.2.4.1 Forecasting Performance

We begin by focusing on prediction at the quarterly (3-month) horizon, which is a standard horizon studied in the literature. Figure 3.1 reports the out of sample R^2 of different forecasting methodologies relative to the AR benchmark, for inflation (left panel), industrial production growth (center panel), and change in unemployment (right panel). In this figure, the prediction exercise is performed by fixing the number of factors K. For PCA (red line) and PLS (blue line), there are no tuning parameters beyond K. For SPCA, we report separate results for each choice of the tuning parameter K (grey lines), as well as for the value for $\lfloor qN \rfloor$ chosen by CV (green line).

Figure 3.1: OOS Performance of SPCA, PCA and PLS (for different number of factors)



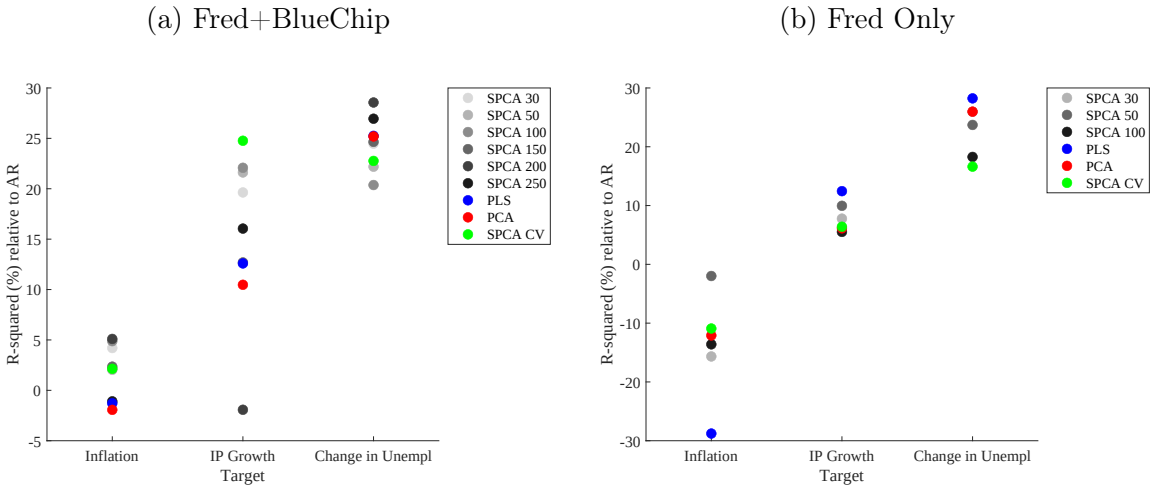
Notes: Each panel reports the out-of-sample R^2 relative to the AR model for a different target, aggregated over 3 months. The three panels predict inflation, industrial production growth and change in unemployment rate, respectively. The green dashed line shows the performance of SPCA with 3-fold cross 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 shows several interesting results. First, it is in general hard to predict inflation beyond what an AR model predicts (see also Faust and Wright [2013]): the out of sample R^2s are close to zero or even negative. Only SPCA, among all methods, produces positive R^2s, and it does so using a small number of factors. Predictability beyond the AR model is much higher for IP growth and change in unemployment. Second, the predictive performance of SPCA is generally higher than that of PCA and PLS for most choices of the number of factors. Third, the performance of PCA does depend on the tuning parameter, but in different ways for different targets. For inflation, for example, a lower value of $\lfloor qN \rfloor$ seems to predict better; for industrial production and unemployment, higher values work better. Finally, the performance of all these methods varies quite dramatically with the number of factors, with substantial declines for the methods that use target information (PLS and SPCA with a smaller $\lfloor qN \rfloor$) as the number of factors increases, because of their overfitting issue we explained earlier.

Given how important the number of factors is for the out-of-sample performance, in what follows we choose the number of factors via cross-validation for all three methods (so for SPCA both $\lfloor qN \rfloor$ and K are jointly selected via CV). The left panel of Figure 3.2 shows the results. Now all three targets (inflation, industrial production growth and change in unemployment rate) appear in the same panel. The panel confirms that SPCA generally performs well in predicting out of sample, doing better than the alternatives (in the case of unemployment, several choices of the tuning parameter $\lfloor qN \rfloor$ outperform PCA and PLS, but not the one chosen by cross-validation). Overall, SPCA tends to do comparatively well when choosing all parameters via cross-validation.

Given the way SPCA chooses the set of predictors, we would expect it to perform best in contexts where there are a large number of predictors, that overall contain valuable information, even if some predictors are redundant or noisy. The forecasting experiment we run here falls in this category: it contains both macroeconomic and financial data (which are likely

Figure 3.2: OOS Performance of SPCA, PCA and PLS (using CV to choose the number of factors)



Notes: The left panel of this figure repeats the analysis of Figure 3.1, but chooses the number of factors via CV. The right panel performs the same analysis as the left panel, but using only Fred data.

to contain important individual predictors), as well as a large number of individual forecasts that we would expect to be informative beyond macroeconomic quantities but where a large part of the observed variation is likely dominated by noise. To better gauge the importance of this additional data in the performance of SPCA, the right panel of Figure 3.2 shows the results of running the same analysis (using the same sample) but with only the Fred data. The figure shows that while the performance of SPCA remains broadly comparable with the other predictors, it deteriorates compared to PCA and PLS (PLS itself has very mixed performance, though, predicting well IP growth and unemployment, and failing to predict inflation). So, on the one hand, this figure shows that individual expert forecasts are useful for prediction of macroeconomic variables, confirming the results in Faust and Wright [2013]; on the other hand, it shows that SPCA does particularly well when working with this large and informative, yet noisy, universe of individual forecasts.

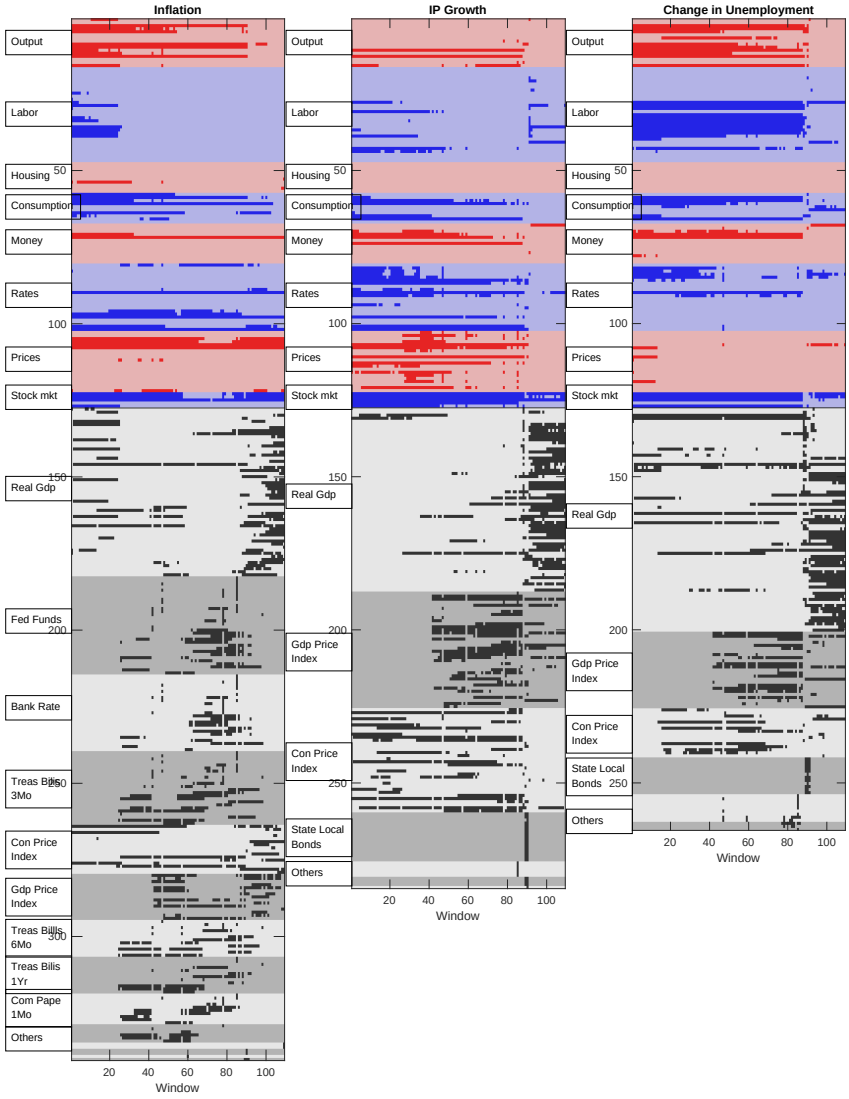
3.2.4.2 Predictors Selected by SPCA

Next, we study in detail how SPCA selects predictors. Figure 3.3 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 sample. GDP forecasts are used throughout the sample to predict changes in

Figure 3.3: Top 50 Predictors Selected by SPCA



Notes: Under the same settings as Figure 3.1, 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.

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.

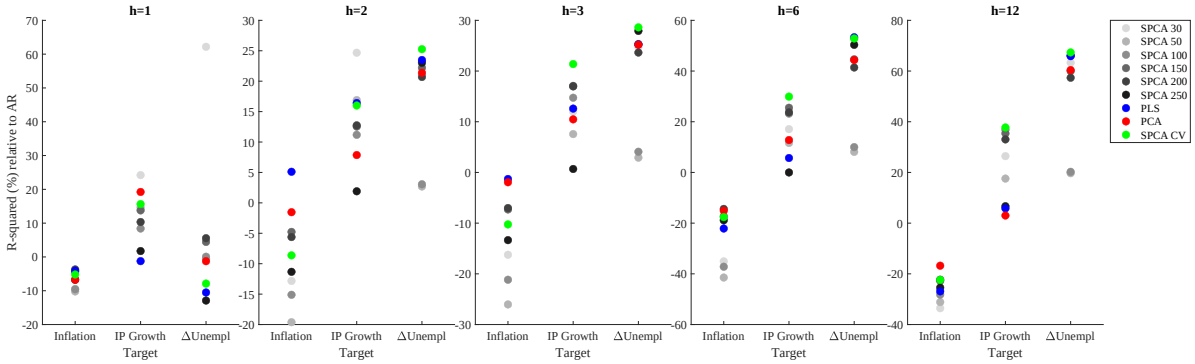
Finally, we note that not all Blue Chip forecasters are the same in terms of forecasting ability. Among the institutions whose forecasts are included in our analysis because they have a sufficiently long time series (each providing tens of forecasts, of different variables at different horizons), we 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.

3.2.4.3 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. [2023]) 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 3.4 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 3.2, 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 3.4: OOS Performances - Different Targeted Horizons



Notes: Similar to Figure 3.2, but showing the out of sample R^2s at different horizons, and using all the horizons concurrently to estimate the factors in SPCA.

3.2.4.4 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 3.2.4.3, for horizons of 1, 2, 3, 6 and 12 months. Figure 3.5 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. [2023], 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.

3.3 Risk Premia Estimation and Factor Model Diagnosis

In this section we apply SPCA to estimate the risk premia of a variety of observable factors, and to diagnose observable factor models.

3.3.1 Data

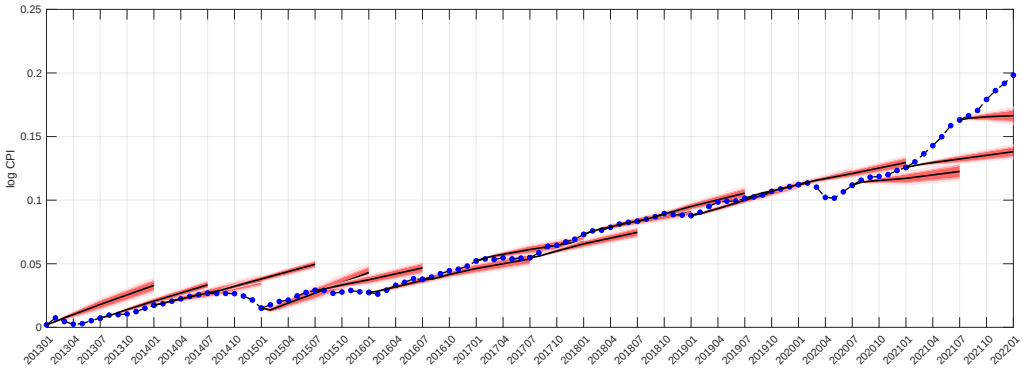
Our main dataset is the Chen and Zimmermann [2020] data, which includes a large number of equity portfolios sorted by characteristics. Specifically, we employ the April 2021 release of the data. For each characteristic considered, Chen and Zimmermann [2020] construct a variable number of portfolios (as many as are used in the original papers that introduced the anomaly in the literature: typically 2, 5, or 10). Not all test assets are available for the entire time period; for our analysis, we study the time period 1976m3 to 2020m12, for which 901 test portfolios are available without missing values. To these sorted portfolios, we add 49 industry portfolios from Ken French's website. All of our results are at the monthly frequency.[6]

We also consider an alternative dataset, proposed by Hou et al. [2020], that includes for the same period 1672 portfolios sorted by characteristics without missing values. Hou et al. [2020] classify their portfolios in six groups: momentum, value, investment, profitability,

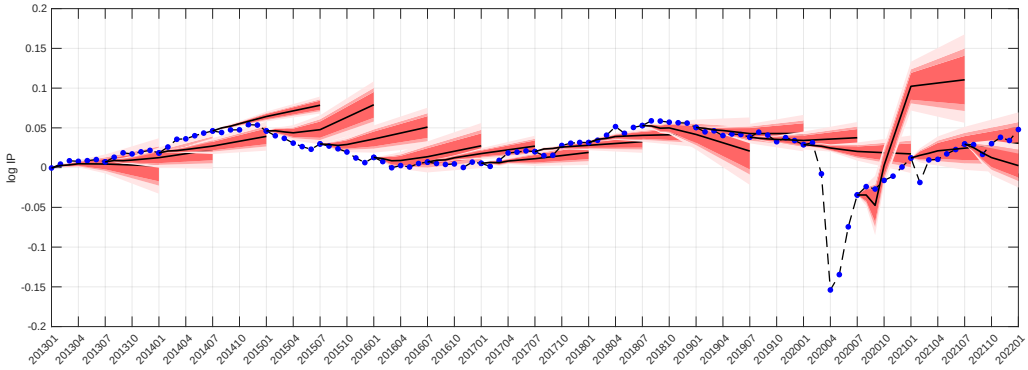
6. The theory is silent on what the correct frequency of the data to study is. Here we follow the literature and focus on monthly frequency; we leave for future work a more comprehensive study and comparison across frequencies.

Figure 3.5: Fan Charts

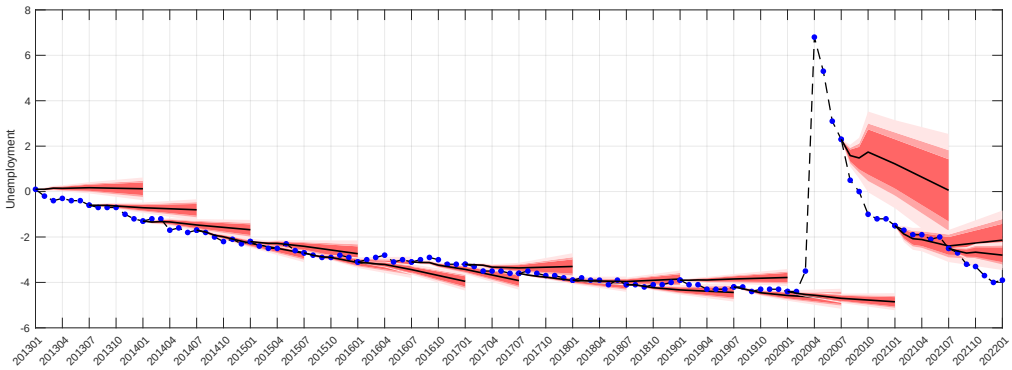
(a) Inflation



(b) IP Growth



(c) Change in Unemployment



Notes: Using the same estimates as Figure 3.4, 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.

intangibles, and frictions. These two datasets are similar and yield comparable results. Rather than producing two versions of each result using the two datasets, we choose Chen and Zimmermann [2020] to be our main dataset and report the robustness of the main results using the Hou et al. [2020] data (see Section 3.3.2.6). What both datasets have in common is that they capture a wide universe of anomaly equity portfolios discovered in the last four decades of asset pricing research.

We consider both tradable and nontradable factors in our analysis, focusing on the best-known ones from the literature. The tradable factors are: the market (in excess of the risk-free rate); size (SMB); value (HML); profitability (RMW); investment (CMA); momentum (MOM); betting-against-beta (BAB, from Frazzini and Pedersen [2014]); and quality-minus-junk (QMJ, from Asness et al. [2013]). The nontradable factors are: the liquidity factor from Pástor and Stambaugh [2003]; the intermediary capital factor from He et al. [2017]; AR(1) innovations in industrial production growth (IP); VAR(1) innovations in the first three principal components of 279 macro-finance variables from Ludvigson and Ng [2010]; AR(1) innovations in the three uncertainty indexes of Jurado et al. [2015], representing financial uncertainty, macroeconomic uncertainty, and real uncertainty; AR(1) innovations in the term spread, the credit spread, and the unemployment rate; AR(1) innovations in two sentiment indexes, one from Huang et al. [2015] and one from Baker and Wurgler [2006]; oil price growth AR(1) innovations; and consumption growth AR(1) innovations.[7]

7. The market factor, SMB, HML, RMW, CMA and MOM are from Ken French's website. BAB and QMJ are from AQR's website. The liquidity factor is from Lubos Pastor's website. The intermediary capital factor is from Asaf Manela's website. The macro principal components and the uncertainty indexes are from Sydney Ludvigson's website. Industrial production, the credit spread, unemployment rate, the term spread, and oil price are from Fred-MD. The Huang et al. [2015] sentiment index is from Huang's webpage. The Baker and Wurgler [2006] sentiment index is from Wurgler's website. The consumption factor was built from NIPA data using the methodology of Schorfheide et al. [2018].

3.3.2 Estimation of Risk Premia using SPCA

In this section we estimate the risk premia of a variety of tradable and nontradable factors. We begin by discussing some details of the implementation of the estimator.

3.3.2.1 Choice of Tuning Parameters and Implementation Details

To apply SPCA to the estimation of the risk premia and to evaluate its out-of-sample performance, we split the sample period into two equal-sized subsamples. The first half of the sample (training period) is used to choose the tuning parameters and produce the risk premium estimate. The second half of the sample (evaluation period) is used to evaluate the out-of-sample performance of the estimator and the choice of the tuning parameter.

For ease of presentation, we choose to select only one tuning parameter, q (or, equivalently, the number of assets selected $\lfloor qN \rfloor$), for each plausible choice of p (the number of factors) in our analysis. This approach reduces the number of tuning parameters to only one, and also conveniently serves as a robustness check with respect to the number of factors.

To determine reasonable candidates for p, we examine the factor structure of the panel of test asset returns. Figure 3.6 provides the scree plot of the log of the first 25 eigenvalues. There appear to be at least three strong factors. In addition, it appears that factors 4-11 might also be relevant, but *weak*. Motivated by the scree plot, in the empirical study below we highlight results for p equal to 3, 5, 7, and 11, therefore showing the robustness of our results to a wide range of model dimensions.

To choose the tuning parameter q, we adopt the same R^2 criterion as in simulations to evaluate the estimator's out-of-sample performance, namely, the hedging ability of the portfolio built by SPCA for g_t. Guided by this *statistical* justification, in our empirical work we choose q by 3-fold CV(100 runs) within the training sample, maximizing the hedging R^2 for g_t. Appendix 2.5.5 describes in detail the steps for the cross-validation. Once we have tuned q, we use it to compute the SPCA risk premium estimate for g_t.

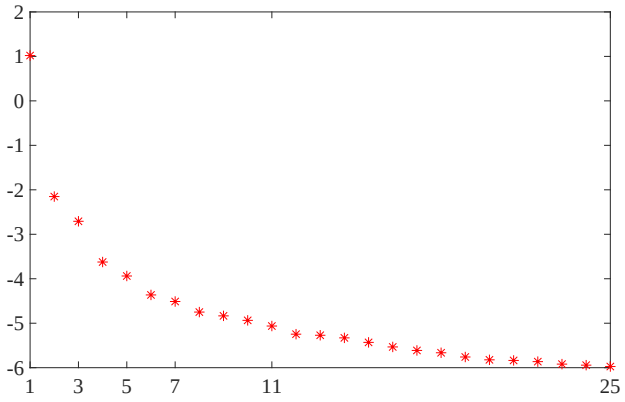


Figure 3.6: Logarithm of the First 25 Eigenvalues in the Chen-Zimmerman data

Note: The figure plots the logarithm of the first 25 eigenvalues of the data, obtained from Chen and Zimmermann [2020] plus 49 industry portfolios, covering the period 1976-2020.

3.3.2.2 Results: Estimation of Risk Premia and Out-of-sample Evaluation

We report the main empirical results in Table 3.1 and Figures 3.7 and 3.8. Each row of Table 3.1 corresponds to one factor; the first 8 are tradable, the rest are nontradable. For tradable factors, the first two columns show the average excess return of the factor, in the training sample and in the evaluation sample, respectively; these numbers correspond to model-free estimates of the risk premia of tradable factors, and can be directly compared with the model-based estimate obtained from SPCA.

The next columns of the table show the SPCA results in 4 groups of columns, corresponding to the number of latent factors $p = 3$, 5, 7, and 11, respectively. For each choice of p, we report the risk-premium estimate (obtained in the training sample, in bp per month), the number of assets selected by SPCA (determined by q), and the out-of-sample R^2 obtained in the evaluation period. These estimates are obtained factor by factor: that is, in each case, g_t contains one factor, and the asset selection is driven by that factor only. In the last two columns of the table, we repeat the exercise (with $p = 11$) but estimate all risk premia simultaneously: g_t contains all the factors and the selection of the assets is based on all of them simultaneously (so that $d \geq p$ as opposed to $d = 1$). In theory, both approaches are

consistent. In practice, estimating risk premia factor by factor has the advantage that the latent factors zoom in immediately on the assets relevant for each factor. On the other hand, the joint estimation is required for the CLT of Section 2.2.2.4.

Consider first the market portfolio (first row of the table), a strong factor in this dataset. The average return of the market in the training sample is 74bp per month, and 62bp in the evaluation period. The SPCA estimates of the market risk premium, for the four chosen values of p, are 68, 70, 72, and 74bp per month, respectively, all close to the average excess return. To obtain these estimates, SPCA estimates the latent factors picking, in each iteration, 100 assets out of the total of 950. Finally, the portfolio that SPCA builds to hedge the market achieves, not surprisingly, a very high out-of-sample R^2, above 0.98 for all p.

To better understand the performance of the estimator and the tuning parameter choice, we can examine the heatmap in Figure 3.7, panel (a), which focuses on the market factor. In the heatmap, the x axis reports the number of factors p; the y axis reports the number of test assets selected by SPCA (in turn determined by q); for each combination of p and q, the heatmap reports the out-of-sample R^2 of the hedging portfolio built by SPCA.

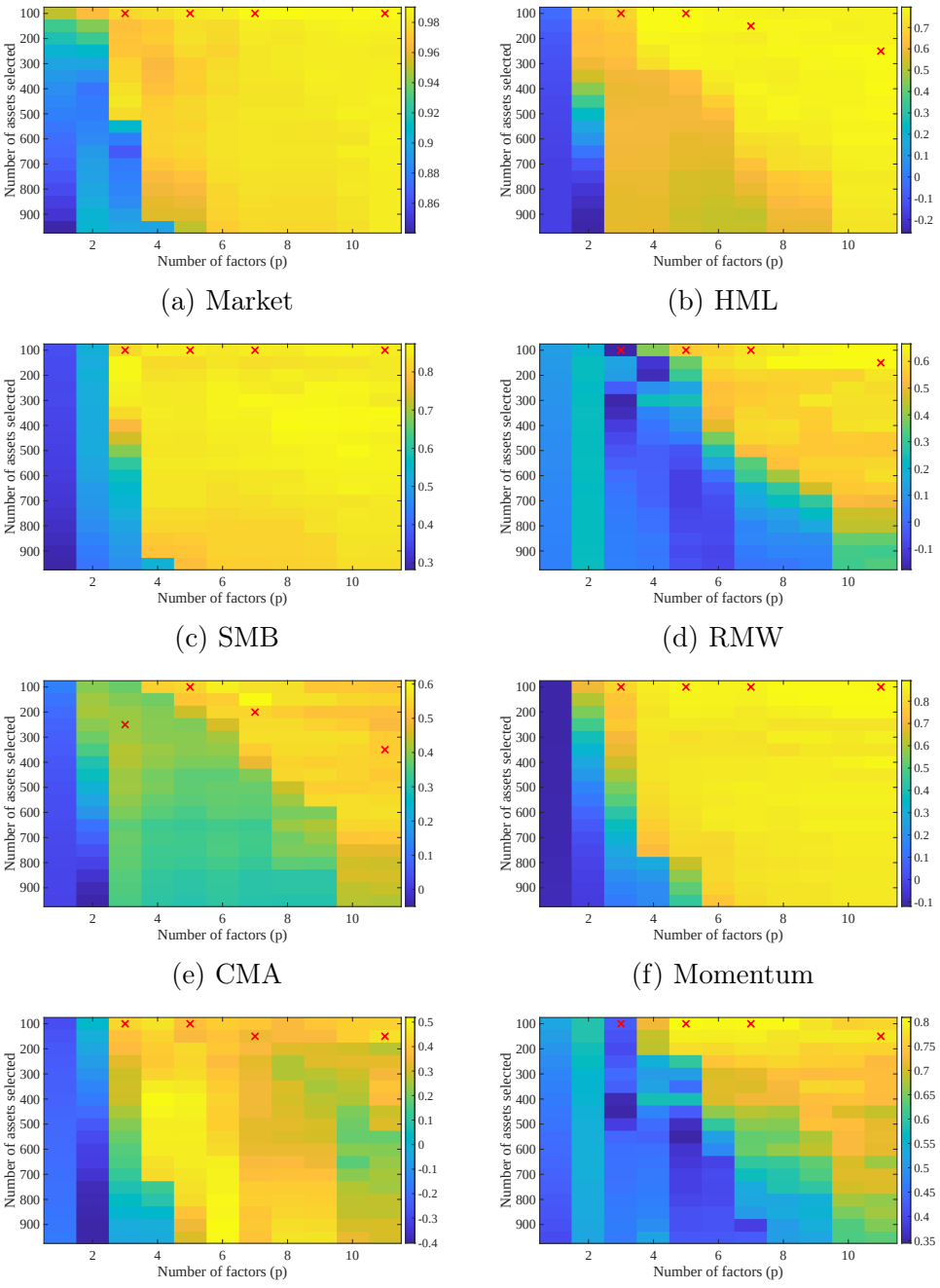
Panel (a) shows that for all combinations of p and q, out-of-sample R^2s are overall very high for the market portfolio, above 85%. However, there appears to be a subset of the parameter space where hedging performance is especially good: combinations with high p and low q. The red marks in the heatmap correspond to the values of q chosen by CV *in the training sample* (one for each value of p considered in the table: 3, 5, 7, 11). Ideally, the values of q chosen by CV in the training sample would yield a hedging portfolio that performs well out of sample: that is, the marks should lie in areas in the heatmap with high out-of-sample R^2s. This is indeed the case, as the figure shows, indicating good out-of-sample performance of the tuning parameter selection procedure.

Consider now another tradable factor, CMA, in the 5th row of Table 3.1. Like for the market, the estimated risk premium for CMA is not significantly different from the average

Table 3.1: Risk premia estimates

	Avg. ret.	Avg. ret.	3 Latent Factors			5 Latent Factors			7 Latent Factors			11 Latent Factors			Joint estim, 11 factors	
	(train.)	(eval.)	RP	# Assets	R^2	RP	# Assets	R^2	RP	# Assets	R^2	RP	# Assets	R^2	RP	Stderr
Market	74	62	68	100	0.98	70	100	0.98	72	100	0.99	74	100	0.99	73	26
HML	39	-7	50	100	0.70	37	100	0.79	39	150	0.78	44	250	0.79	54	18
SMB	12	25	15	100	0.82	5	100	0.85	10	100	0.85	10	100	0.85	7	18
RMW	37	28	-8	100	-0.18	40	100	0.56	33	100	0.61	27	150	0.66	23	9
CMA	26	19	36	250	0.41	40	100	0.55	27	200	0.55	31	350	0.53	34	11
Momentum	91	30	67	100	0.79	86	100	0.87	102	100	0.87	101	100	0.88	96	23
BAB	126	56	112	100	0.43	120	100	0.38	112	150	0.35	128	150	0.45	93	20
QMJ	41	39	-9	100	0.43	28	100	0.81	31	100	0.80	36	150	0.78	20	10
Liquidity			70	550	0.01	85	650	0.02	83	700	0.04	95	900	0.03	105	25
Intermed. Cap.			112	100	0.59	101	100	0.56	121	150	0.55	116	350	0.52	109	41
IP growth			-4	950	-0.01	-4	950	-0.02	-5	950	-0.03	-2	950	0.00	-2	3
LN 1			225	550	-0.28	202	650	-0.19	150	700	-0.11	54	950	-0.12	35	146
LN 2			-70	950	-0.05	-79	950	-0.12	-24	950	-0.16	-29	950	-0.17	-53	82
LN 3			96	400	0.03	86	650	0.06	16	700	0.06	-21	850	0.05	-92	78
Consumption			2	950	-0.01	3	950	0.00	3	950	-0.01	2	950	-0.01	2	2
Fin. Unc.			-61	350	-0.08	-48	750	0.00	-40	850	0.09	-41	950	0.10	-46	17
Real Unc.			-6	950	0.05	-7	950	0.04	-9	950	0.04	-11	950	0.06	-17	12
Macro Unc.			-7	950	0.08	-10	950	0.08	-10	950	0.08	-16	950	0.09	-19	10
Term			229	950	-0.11	81	950	-0.36	-57	950	-0.54	262	950	-0.59	384	372
Credit			41	950	-0.03	62	950	-0.03	41	950	-0.02	-43	950	-0.03	-32	77
Unempl.			65	950	0.00	109	950	-0.01	112	950	-0.01	110	950	0.00	45	108
Sentiment HJTZ			-24	950	0.01	-27	950	-0.03	-18	950	-0.06	-40	950	-0.07	-34	76
Sentiment BW			57	950	0.00	64	950	0.00	50	950	0.01	16	950	-0.02	44	71
Oil			-37	950	-0.05	-62	950	-0.02	-42	950	-0.03	-20	950	-0.02	-9	41

Note: In this table, we report the estimation results for tradable and nontradable factors using SPCA. The first two columns report the average excess returns for tradable factors, in the training sample (first half of the sample period) and in the evaluation sample (second half of the sample period). The remaining columns report, for different values of the number of factors p, the risk premia estimates (in basis points per month, computed in the training period), the number of assets selected by SPCA (governed by the parameter q), and the out-of-sample R^2 of the implied hedging portfolio. The last two columns report risk premia estimates and standard errors including all factors in g_t simultaneously, with $p = 11$. Sample is the Chen and Zimmermann [2020] test portfolios plus 49 industry portfolios, over the period 1976-2020.



(a) Market (b) HML

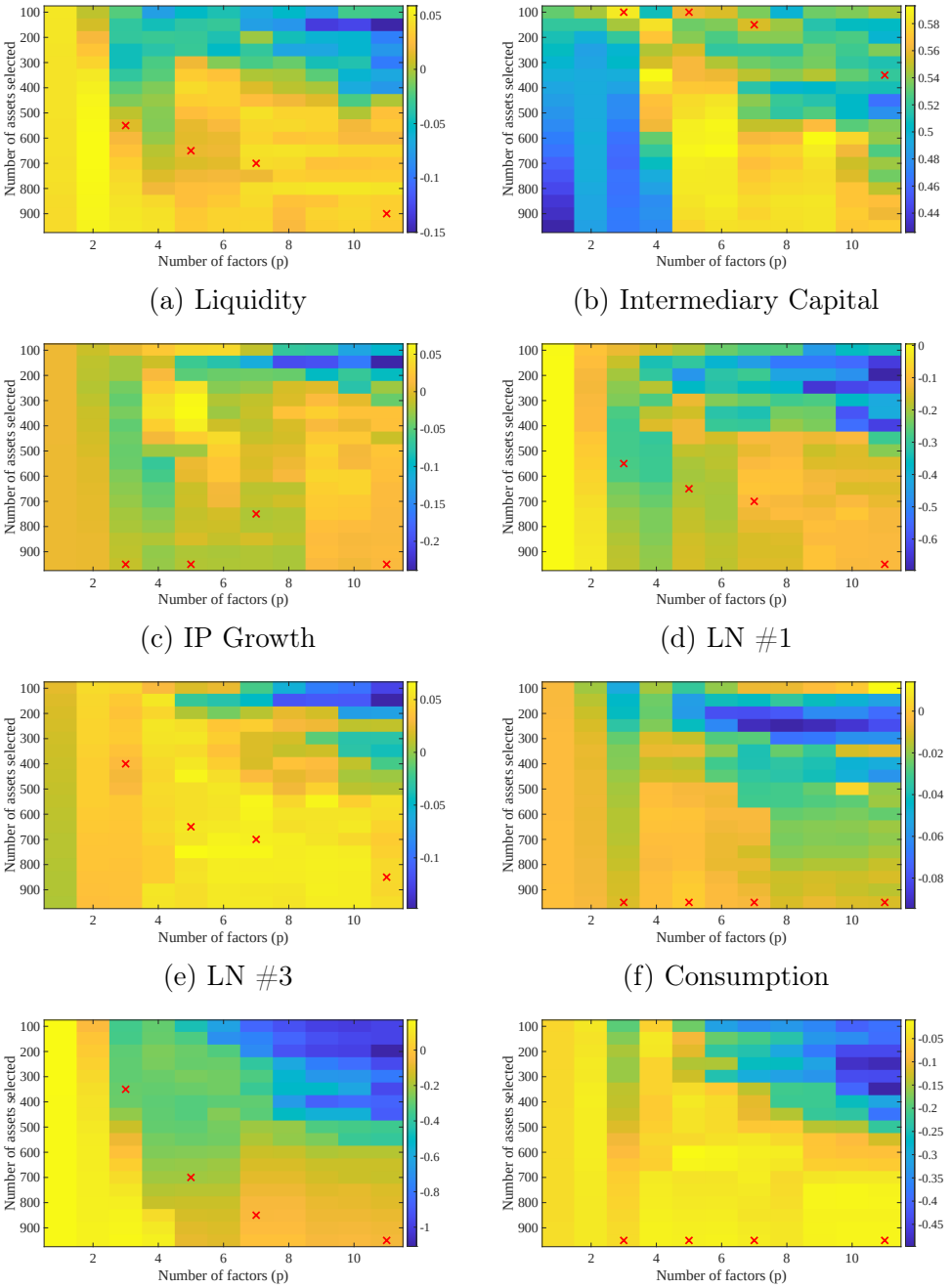
(c) SMB (d) RMW

(e) CMA (f) Momentum

(g) BAB (h) QMJ

Figure 3.7: Out-of-sample R^2 Heatmaps, Tradable Factors

Note: Each panel reports the out-of-sample R^2 heatmap for a different factor. X-axis reports p. Y-axis reports the number of assets selected, governed by q. The colors in the heatmap correspond to the out-of-sample R^2 of the SPCA-implied hedging portfolio for the factor g_t; this R^2 is computed entirely in the evaluation period. The red marks are the points chosen by CV within the training sample.



(a) Liquidity (b) Intermediary Capital

(c) IP Growth (d) LN #1

(e) LN #3 (f) Consumption

(g) Fin. Unc. (h) Oil

Figure 3.8: Out-of-sample R^2 Heatmaps, Nontradable Factors

Note: Same as Figure 3.7, but for a subset of nontradable factors.

excess return of the factor. The number of assets selected by SPCA ranges between 100 and 350, and the out-of-sample R^2 is above 50%, indicating that the hedge portfolio built by our latent factor model is able to capture the majority of the variation on CMA out of sample.[8]

The heatmap of the out-of-sample R^2 for the hedging portfolio of this factor is panel (e) of Figure 3.7. The figure shows that for the case of CMA, different combinations of p and q yield very different out-of-sample hedging performance, with R^2s ranging from above 50% to below 0. Ideally, if the tuning parameter were chosen properly, we would see the hedging portfolio also does well out of sample. The red marks in the figure show that this is indeed the case, especially for $p = 5$ and above.

These heatmaps also allow us to compare the results with the PCA-based estimator of Giglio and Xiu [2021]. This is because the last row of the heatmap corresponds to the case $q = 1$, that is, all assets are used to estimate the factors; so PCA corresponds to a particular choice for the tuning parameter. Looking across the various panels of Figure 3.7, it is clear that while for some factors (like the market) similar R^2 can be obtained by PCA and SPCA, for other factors (like CMA and RMW) the out-of-sample R^2s obtained by SPCA are substantially higher than those by PCA. This is not surprising given that the scree plot has shown the presence of several weak factors in the data.

One additional advantage of SPCA that is clearly visible in the heatmaps is that SPCA often manages to achieve the same (or better) R^2 than PCA, while estimating a much smaller number of factors. For example, consider the momentum factor in panel (f). The last row of the heatmap shows that extracting factors via PCA achieves an R^2 above 70% only once at least 6 factors are included; SPCA gets there even with 3 factors. The reason is intuitive:

8. Given that the universe of test assets includes portfolios sorted by the same characteristics used to construct the tradable factors like CMA, one may wonder why an out-of-sample R^2 of 100% is not always obtained for tradable factors. The reason is that SPCA attempts to build a hedging portfolio for the target g_t with factors that must also explain covariation among the universe of test assets. An advantage of our approach is that the hedging portfolio is able to avoid fitting the "measurement error" component in g_t, which, as discussed above, can be thought of as non-diversified idiosyncratic error for tradable factors, or more literally measurement error for nontradables.

SPCA focuses on the test assets most informative about g_t, and therefore can zoom in quickly on the most relevant latent factors.

For nontradable factors, we cannot compare the risk premium estimate from SPCA with the average excess return; beyond relying on the theory and simulations, we can look at the out-of-sample R^2 for suggestive evidence about the empirical performance of the estimator. Note that it is well known in the literature that it is difficult to hedge nontradable factors, like consumption or IP growth, in equity markets. We will however show that SPCA gives a hedging portfolio that successfully hedges at least a part of the variation in many nontradable factors.

Consider first the liquidity factor of Pástor and Stambaugh [2003], in row 9 of Table 3.1 and panel (a) of Figure 3.8. The out-of-sample R^2 achieved by SPCA is above 0 (up to 4%), and the estimated risk premium appears to be high (between 70 and 95bp per month). Panel (a) of Figure 3.8 shows how strongly this R^2 depends on p and q. Among all combinations of parameters, a large fraction actually delivers a negative out-of-sample R^2. This simply stresses how difficult it is to hedge this factor (like most macro factors) using equity markets, and indicates again the relatively good performance of SPCA as tuned in the training sample.

The remainder of the table and of the two figures shows the results for all the other factors (for reasons of space, the heatmaps only report a subset of the factors, while the table reports them all). A few interesting patterns emerge. First, for tradable factors, SPCA gives risk premia estimates that are always close to the model-free estimates obtained from average excess returns: the two are never statistically different at the 5% level (with the only exception of QMJ with $p = 3$). Second, confirming the previous literature, nontradable factors are much harder to hedge than tradable factors; in fact, for several factors – like the first two JLN macro factors – we do not get positive R^2 at all. For those factors, there is so little exposure in equity returns that SPCA cannot build a proper hedging portfolio. However, SPCA is able to hedge out of sample at least a part of the variation of many factors, like the

third LN factor, the three uncertainty measures, the liquidity factor and the intermediary capital factor (for which it achieves an R^2 above 50%). Third, the risk premia estimated by SPCA – for those factors where SPCA can actually hedge some of the variation – make economic sense: for example, the liquidity and intermediary factors command significantly positive risk premia, whereas the three uncertainty measures command negative risk premia.

3.3.2.3 Asset Selection

To better understand how SPCA estimates risk premia, we can study which assets are selected when extracting the latent factors. Table 3.2 shows, for four representative factors (two tradables, Momentum and RMW, and two nontradables, liquidity and intermediary capital), the top 10 test assets (by absolute value of correlation) selected at each step. The names of the portfolios follow Chen and Zimmermann [2020], with the numbers indicating the quintile or decile of the characteristic.

Consider Momentum in the first set of rows. To extract the first factor, SPCA selects the assets with the highest correlation with the momentum factor. The table indicates that the highest correlation, at 0.44, is with IntMom09, an intermediate momentum portfolio. The other assets with high correlation are all momentum-related, not surprisingly. In the next columns, the table shows the assets selected at the second iteration of SPCA, after orthogonalizing g_t and the test assets to the first factor. Interestingly, the correlations among these residuals are even higher, up to 0.79 for a different momentum sort (Mom12mOffSeason, momentum without the seasonal component). This suggests that the first factor captures some of the asset variation that is not exclusively specific to momentum (for example, part of the market factor), which the projection step of SPCA removes.

The remainder of the table shows which assets are selected at the different iterations for RMW, Liquidity, and Intermediary Capital. For RMW (a profitability factor), the assets selected are often based on accounting measures, like asset growth, accruals, leverage, and

Table 3.2: Assets Selected by SPCA

	Factor #1		Factor #2		Factor #3	
	Asset	| Corr |	Asset	| Corr |	Asset	| Corr |
Mom	IntMom09	0.44	Mom12mOffSeason02	0.79	Mom12m08	0.64
	IntMom10	0.4	Mom12mOffSeason03	0.76	BMdec05	0.63
	MomVol10	0.37	Size01	0.74	IntMom03	0.63
	MomVol09	0.36	ResidualMomentum01	0.73	SP05	0.62
	IntMom08	0.36	ResidualMomentum02	0.73	ShareIss5Y05	0.62
	Mom12m10	0.36	NumEarnIncrease01	0.72	BookLeverage02	0.62
	FirmAgeMom05	0.35	ShareIss5Y01	0.7	cfp05	0.61
	Mom12mOffSeason10	0.34	MomVol03	0.69	BMdec04	0.61
	Mom12mOffSeason09	0.33	CompEquIss01	0.68	ShareIss1Y05	0.6
	Mom12m09	0.33	Mom12m03	0.68	LRreversal04	0.6
RMW	Industry:Gold	0.27	OperProf05	0.54	OperProfRD01	0.53
	MomOffSeason10	0.27	OperProfRD09	0.53	RoE01	0.47
	AccrualsBM02	0.27	CBOperProf09	0.5	GP01	0.45
	DelEqu05	0.27	RoE05	0.49	CBOperProf02	0.45
	LRreversal05	0.27	CBOperProf10	0.49	DolVol01	0.44
	roaq01	0.26	Leverage02	0.49	OperProfRD02	0.44
	AssetGrowth10	0.26	OperProfRD08	0.49	CBOperProf01	0.43
	DolVol05	0.25	realestate03	0.49	OperProf01	0.41
	ChEQ05	0.25	GP05	0.49	RoE02	0.4
	Price05	0.25	GP04	0.48	VolMkt02	0.4
Liq.	InvGrowth06	0.47	InvGrowth06	0.28	InvGrowth06	0.3
	NetPayoutYield07	0.47	BetaFP09	0.26	DolVol01	0.27
	PayoutYield05	0.46	EntMult06	0.25	XFIN08	0.26
	PayoutYield07	0.46	NetPayoutYield07	0.24	MeanRankRevGrowth01	0.26
	BetaFP03	0.46	PayoutYield07	0.24	BetaFP03	0.25
	DelLTI02	0.46	PayoutYield05	0.24	ShortInterest01	0.25
	IntanBM03	0.46	cfp04	0.23	BetaFP09	0.24
	EntMult06	0.46	BetaFP10	0.23	EntMult06	0.24
	VolMkt04	0.46	XFIN08	0.23	PayoutYield07	0.24
	PayoutYield06	0.46	ShortInterest01	0.22	ChEQ04	0.23
Interm.	Industry:Banks	0.9	Industry:banks	0.76	Industry:banks	0.7
	Industry:Fin	0.84	Industry:Fin	0.56	Industry:Fin	0.47
	IntMom05	0.8	DelEqu02	0.46	DebtIssuance02	0.38
	EquityDuration04	0.8	grcapx3y02	0.44	NOA10	0.36
	IdioVolAHT05	0.8	OScore02	0.43	ChAssetTurnover04	0.35
	IdioVol3F05	0.79	GrLTNOA10	0.43	HerfAsset05	0.35
	MaxRet08	0.79	ChAssetTurnover04	0.43	ShareRepurchase01	0.35
	Illiquidity01	0.79	IntMom05	0.43	HerfBE05	0.35
	IdioRisk05	0.79	IdioVolAHT05	0.42	DelEqu05	0.32
	CBOperProf03	0.78	Tax01	0.42	Beta05	0.32

Note: For each factor (one per panel) the table shows the top-10 assets selected by SPCA in extracting the latent factors. Assets are sorted by absolute value of the correlation. For each factor from 1 to 3, the table reports the names of the portfolios selected, and the absolute value of the correlation with g_t. Naming convention for the portfolios follows Chen and Zimmermann [2020].

operating profits. For liquidity, portfolios sorted by payout yield and beta seem to play an important role in hedging the risk. Finally, for intermediary capital, the portfolios selected

by SPCA relate to idiosyncratic volatility, liquidity, as well as two industry portfolios (not surprisingly, banking and financials).

The selection of particularly informative assets is the central mechanism through which SPCA addresses the issue of weak factors. It is also responsible for the parsimony of SPCA to the number of factors used, since SPCA zooms in on the most informative assets.

3.3.2.4 Strength of the Factors

We next report the strength of the factors extracted by SPCA at each step. To make the results comparable across iterations of SPCA, and between SPCA and PCA, we compute the strength of a latent factor as the eigenvalue of the factor normalized by the number of assets used to extract it. Figure 3.9 reports, in each panel, the log normalized eigenvalues for the factors extracted from PCA (dashed line) and for the factors extracted by SPCA, grouped across panels for the various targets (since the factors extracted by SPCA are different for different targets g_t): panels (a) and (b) show the factors extracted when the targets are tradable factors, panels (c) use a subset of nontradables, and (d) the remaining nontradables. The figure shows eigenvalues corresponding to the first 5 factors.

As expected, the log eigenvalues for PCA decrease as lower-variance factors are extracted. This is also mostly (but not always) the case for SPCA, where however we see a large difference across factors. For some factors (like most nontradables, which, as discussed above, are mostly noise factors), SPCA chooses a large number of assets, so the results look very similar to PCA (e.g. see panel (d)). For factors where SPCA chooses a small number of assets (e.g., intermediary capital and many tradables) we see that the strength of the factor extracted is higher than with PCA. This effect is strongest for the first eigenvalue (the log scale hides it somewhat), but is there for subsequent factors as well. In general, it appears that SPCA indeed strengthens the factor extracted from the cross-section, compared

to PCA, and especially so when fewer assets are selected.[9]

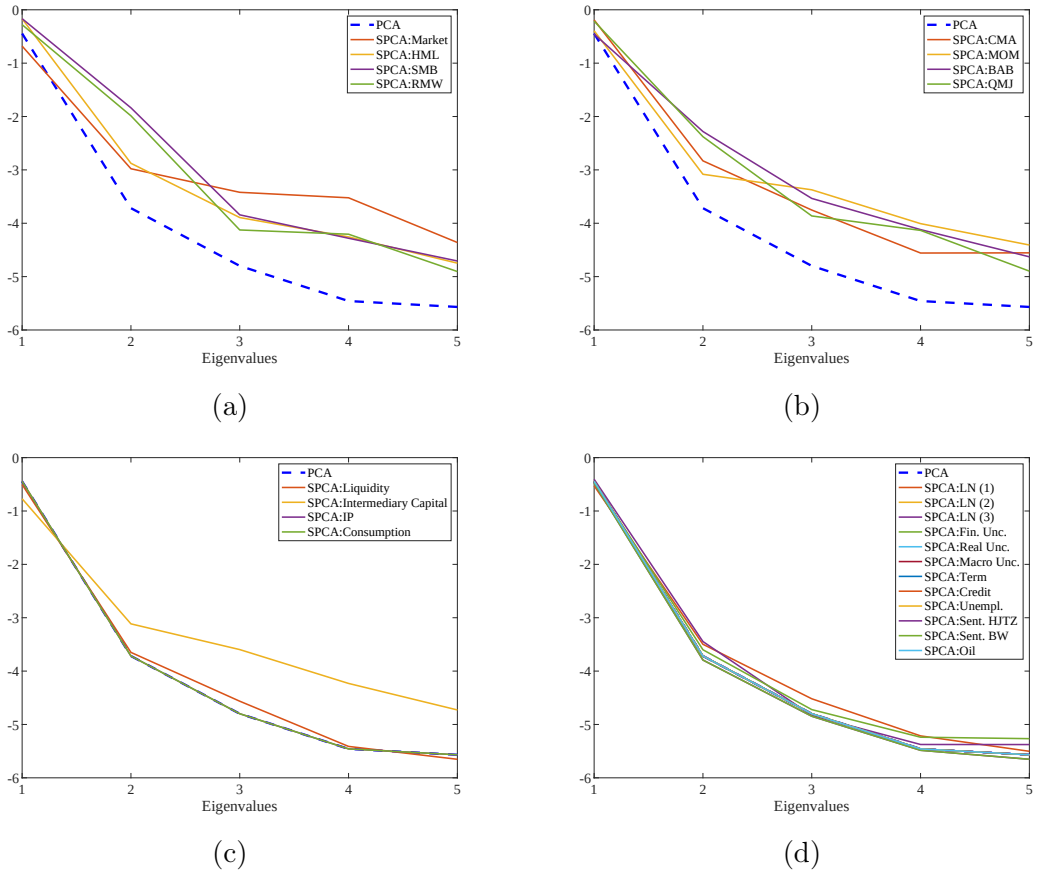


Figure 3.9: Strength of the Latent Factors

Note: Each panel of the figure shows the log eigenvalues extracted by PCA from the universe of all assets in the training sample, as well as the log eigenvalues extracted by SPCA at each iteration (for the first 5 factors), for the tuning parameter selected by CV. All eigenvalues are normalized by the number of assets used, which is a measure of strength of the factor that is directly comparable. Panels (a) and (b) study two groups of tradable factors, panel (c) a selection of the nontradables, and panel (d) the remaining nontradables.

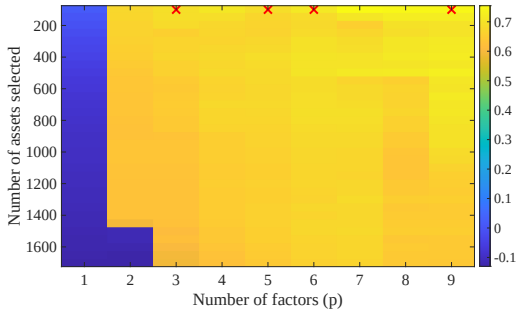
3.3.2.5 SPCA and the Universe of Test Assets

The fact that SPCA estimates the latent factors using the most informative assets also makes it particularly robust to the universe of test assets used in the estimation. We explore this

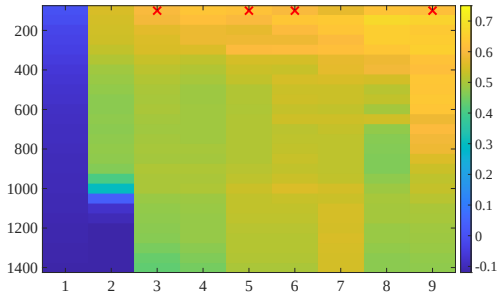
9. One caveat is that once the main factors are extracted, and mostly noise is left in the cross-section, noise itself could lead to higher normalized eigenvalues. This is why the criterion for tuning the parameter q of SPCA is the out-of-sample R^2 of the hedging portfolio, and not this measure of factor strength.

here in detail by considering three factors, value, momentum, and profitability, for which we can easily identify test assets informative about them. Specifically, we consider (for this section only) the dataset from Hou et al. [2020], which, as discussed in Section 3.3.1, collects test portfolios by characteristics in six groups, among which one is labeled "value vs. growth", one "momentum", and one "profitability." We can then ask: how does SPCA perform in estimating the value risk premium if we exclude the value and growth sorts from the universe? Similarly, how does it perform in estimating the momentum and profitability risk premia if momentum and profitability test assets, respectively, are removed? Once the sorted portfolios are removed, the corresponding factors naturally become weaker. However, we expect SPCA to still perform well, as long as sufficient exposure to the factor is present in the remaining test assets. On the contrary, we expect PCA's performance to deteriorate more sharply.

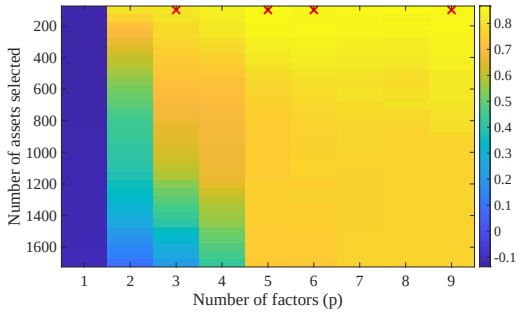
We again look at the performance of SPCA through the lens of the hedging portfolio R^2. Figure 3.10 reports the out-of-sample time-series R^2 heatmap for the three factors: value, momentum and profitability. On the left of each row we can see the R^2 obtained using all assets from the Hou et al. [2020] dataset; on the right we can see the results excluding the test assets corresponding to each factor. By looking at the last row of each heatmap, which corresponds to the PCA estimate with no selection, it is clear that the hedging performance of a portfolio built via PCA deteriorates significantly when the most informative assets are removed. Consider for example the case $p = 9$. For value, the PCA hedging portfolio's out-of-sample R^2 decreases from 64% to 47%, as value and growth assets are removed; SPCA's R^2 decreases by substantially less, from 74% to 62%. In the case of momentum, the R^2 decreases from 76% to 48% for PCA, but only from 86% to 77% for SPCA. Finally, for profitability, the R^2 decreases from 41% to 14% for PCA, but only from 71% to 60% for SPCA. In all cases, the SPCA portfolio hedging ability deteriorates little when the relative sorts are removed and the factor is made weaker, whereas the deterioration is much larger



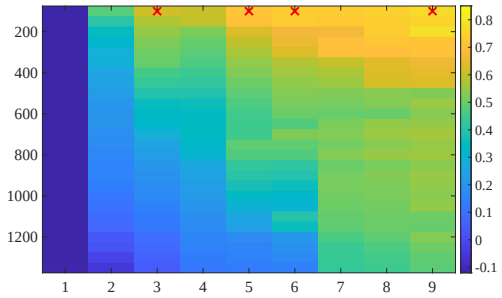
(a) Value



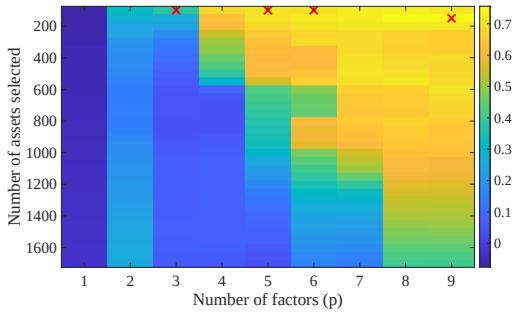
(b) Value w/o value vs. growth test assets



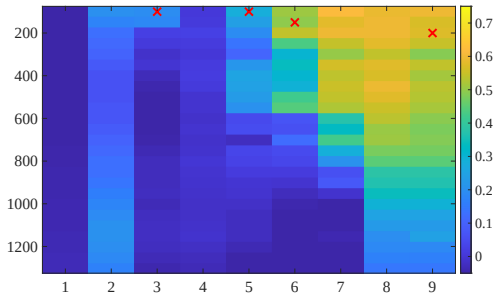
(c) Momentum



(d) Momentum w/o momentum test assets



(e) Profitability



(f) Profitability w/o profitability test assets

Figure 3.10: Varying the Universe of Test Assets

Note: For value, momentum and RMW (profitability), the figure shows the out-of-sample R^2 heatmaps when all the test assets from Hou et al. [2020] are used in the estimation (left), and when value portfolios, momentum portfolios, or profitability portfolios, respectively, are excluded (right).

for PCA.

To sum up, these empirical results mirror the simulations in Section 2.3, which show that SPCA performs well even when the factor of interest is weak in the universe of test assets considered.

3.3.2.6 Robustness

We conclude by reporting in Table 3.3 a version of Table 3.1 obtained using the Hou et al. [2020] dataset instead of the Chen and Zimmermann [2020] data. The results are qualitatively similar to the ones obtained using the Chen and Zimmermann [2020] data, and, with a few exceptions, not statistically different. This confirms that, broadly, the results do not depend on using one particular universe of test assets. That said, the results also suggest some differences between these two universes of test assets, which our analysis in the next section sheds some light on.

3.3.3 Diagnosing Factor Models via SPCA

In the previous section we apply SPCA to the estimation of risk premia. In this section, we illustrate the use of SPCA to diagnose missing factors in observable-factor models, applying the theory developed in Section 2.2.3. Recall that given an observable-factor model g_t, and a set of test assets r_t, we can use SPCA to recover the latent-factor SDF (using g_t to supervise the extraction of weak factors). If we find that the Sharpe ratio achieved by the latent factors recovered by SPCA is higher than that achieved by g_t, we can conclude that the factor model using g_t to span the SDF, is missing some factor. This is not just a test of whether g_t explains r_t, as it instead focuses on shedding light on *why* a model may be rejected in the data.

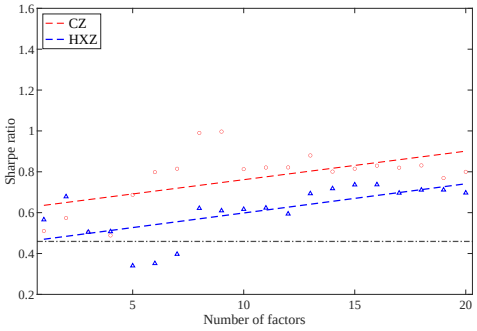
We consider five observable factor models: the CAPM, the Fama-French 3-factor model (FF3), the Fama-French 5-factor model (FF5), and finally two richer models: one with the FF5 factors plus momentum, and one with FF5 plus momentum, BAB, and QMJ. We diagnose these models using both the CZ and the HXZ datasets.

We divide the sample into two parts as in Section 2.2.2, and use the first half for training (and selection of the tuning parameter) and the second half for out-of-sample evaluation. Maximal Sharpe ratios achieved using the factors in g_t and using the factors from SPCA are calculated out of sample.

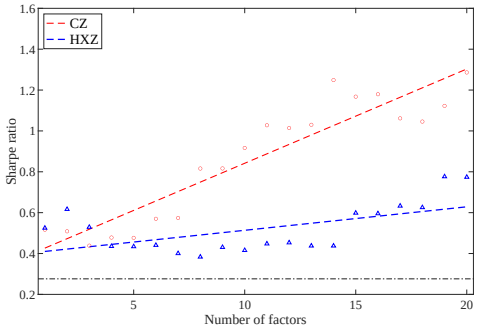
Table 3.3: Risk Premia Estimates, Hou et al. [2020] Data

	Avg. ret.	Avg. ret.	3 Latent Factors			5 Latent Factors			6 Latent Factors			9 Latent Factors			Joint estim, 9 factors	
	(train.)	(eval.)	RP	# Assets	R^2	RP	# Assets	R^2	RP	# Assets	R^2	RP	# Assets	R^2	RP	Stderr
Market	74	62	72	100	0.98	74	100	0.99	74	100	0.99	74	100	0.99	71	26
HML	39	-7	22	100	0.69	20	100	0.69	16	100	0.71	18	100	0.74	24	16
SMB	12	25	-12	100	0.74	-13	100	0.72	-16	100	0.74	-15	100	0.79	-9	18
RMW	37	28	12	100	0.38	26	100	0.71	25	100	0.71	36	150	0.76	18	9
CMA	26	19	8	100	0.70	11	100	0.65	12	100	0.66	4	300	0.65	8	11
Momentum	91	30	68	100	0.81	60	100	0.86	57	100	0.85	55	100	0.86	58	20
BAB	126	56	47	100	0.08	37	100	0.07	31	100	0.08	27	200	0.04	37	12
QMJ	41	39	-3	150	0.68	15	100	0.82	15	100	0.83	17	150	0.80	6	10
Liquidity			28	1700	0.05	35	1700	0.06	42	1700	0.06	35	1700	0.06	37	18
Intermed. Cap.			107	100	0.49	98	100	0.46	91	100	0.51	63	100	0.43	90	37
IP growth			-2	1700	0.01	-4	1700	-0.02	-3	1700	-0.01	-3	1700	-0.01	-1	2
LN 1			171	1200	-0.11	215	1650	-0.15	151	1700	-0.11	169	1700	-0.08	110	93
LN 2			-19	1700	-0.08	-17	1700	-0.08	-13	1700	-0.08	-4	1700	-0.12	-83	55
LN 3			16	1000	0.03	69	1550	0.04	26	1700	0.02	15	1700	0.03	11	62
Consumption			0	1700	0.00	0	1700	0.00	1	1700	0.00	0	1700	0.00	1	1
Fin. Unc.			-5	1600	0.18	-15	1700	0.16	-15	1700	0.16	-9	1700	0.14	-18	11
Real Unc.			-4	1700	0.02	-5	1700	0.02	-8	1700	0.02	-6	1700	-0.03	-11	6
Macro Unc.			-2	1700	0.05	-4	1700	0.05	-6	1700	0.05	-4	1700	0.04	-9	5
Term			-11	1700	-0.11	24	1700	-0.10	77	1700	-0.08	24	1700	-0.14	261	240
Credit			24	1700	-0.02	29	1700	-0.03	0	1700	-0.06	8	1700	-0.09	16	40
Unempl.			42	1700	0.00	116	1700	-0.01	112	1700	-0.01	101	1700	-0.02	89	61
Sentiment HJTZ			-44	1700	0.01	-39	1700	0.01	-22	1700	0.02	-20	1700	0.02	-39	44
Sentiment BW			-29	1700	0.03	-31	1700	0.02	-21	1700	0.02	-25	1700	-0.01	9	43
Oil			-8	1600	-0.03	-39	1500	0.00	-35	1600	0.00	-26	1550	-0.01	-47	28

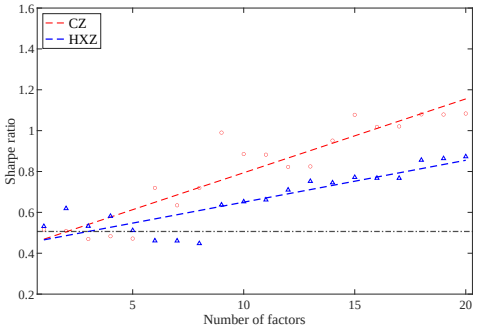
Note: Same as Figure 3.1, but using the characteristic-sorted portfolios from Hou et al. [2020] instead of those from Chen and Zimmermann [2020].



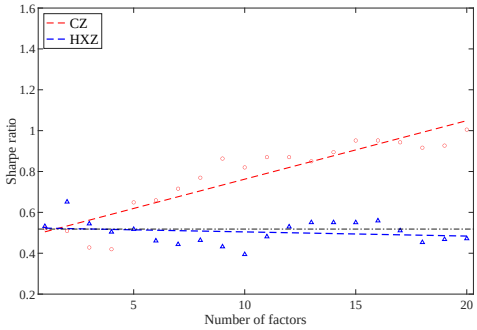
(a) Model 1 (CAPM)



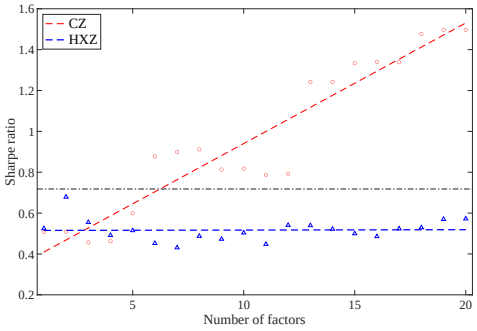
(b) Model 2 (FF3)



(c) Model 3 (FF5)



(d) Model 4 (FF5+Momentum)



(e) Model 5 (FF5+Momentum+BAB+QMJ)

Figure 3.11: Out-of-sample Sharpe Ratios of Different Factor Models

Note: Each panel reports the out-of-sample Sharpe ratio of an observable-factor model g_t (dashed line), together with the out-of-sample Sharpe ratio obtained from the factors recovered using SPCA, in the HXZ data (triangles) and in CZ (circles). The x axis corresponds to the number of factors used in SPCA (p).

Figure 3.11 reports the results. Each panel corresponds to a different model. The x axis in each figure corresponds to the number of factors extracted via SPCA. The y axis is the out-of-sample Sharpe ratio. The Sharpe ratio achieved by g_t is represented by a dashed solid

line, which naturally does not depend on the number of latent factors. In each graph, we overlay the SPCA results with the HXZ and CZ data, respectively, using different markers (blue triangles for HXZ and red circles for CZ). Not surprisingly, the out-of-sample Sharpe ratios are somewhat noisy; we also plot fitted lines using raw estimates to help visualize the trend.

Consider panel (a), in which g_t is just the market. The market in our out-of-sample period achieves a Sharpe ratio of 0.46 (dashed line). SPCA factors extracted using g_t achieve significantly higher Sharpe ratios, both in the HXZ and CZ data. The Sharpe ratio increases with the number of factors, indicating that the CAPM misses several sources of risk. Results for the FF3 and FF5 models (panels (b) and (c)) are similar: for both, once the number of factors is sufficiently large, SPCA produces a Sharpe ratio that is superior to either model. Once momentum is included (pane (d)), the model does perform as well as SPCA in the HXZ data. This suggests that relative to the universe of test assets in the HXZ dataset, this model (FF5+momentum) appears to be spanned by almost all sources of risk driving this dataset (but not so in the CZ dataset).

As more observable factors are added to these models (panel (e) that includes BAB and QMJ), we should expect the Sharpe ratio of the model to increase, as long as more latent factors adds risk factors and not noise. We indeed find that this is the case. Overall, this suggests that these richer models do a better job in capturing the fundamental sources of risk in these dataset, although some amount of misspecification remains visible in the CZ dataset.

The differences between the results using the HXZ and CZ datasets also emphasize the importance of the choice of test assets. Ideally, to have as powerful tests as possible, we would want to have a large and varied universe of test assets. The number of assets in a datasets is, however, not a perfect proxy for the richness of the universe in terms of risk exposures. In fact, as we have remarked in this paper, a universe with large N but low exposures to some

factors can introduce a weak factor problem. Here we see another case in which the size of the dataset does not necessarily translate into richer risk exposure: HXZ contains more assets than CZ; yet, the results in this section show that using the test assets, r_t, from CZ, SPCA diagnoses additional factors compared to the ones diagnosed using HXZ (this could reflect, for example, a different construction of the portfolios in the different datasets, or a different selection of characteristics).

Overall, these results illustrate that the ability of SPCA to recover weak latent factors can prove useful as a diagnostic tool for observable factor models, and once again highlight the importance of the choice of test assets in performing asset pricing tests.

REFERENCES

Dong-Hyun Ahn, Jennifer Conrad, and Robert F. Dittmar. Basis assets. *The Review of Financial Studies*, 22(12):5133–5174, 2009.

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

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

Stanislav Anatolyev and Anna Mikusheva. Factor models with many assets: strong factors, weak factors, and the two-pass procedure. *Journal of Econometrics, forthcoming*, 2021.

Andrew Ang, Robert Hodrick, Yuhang Xing, and Xiaoyan Zhang. The cross-section of volatility and expected returns. *Journal of Finance*, 61:259–299, 2006.

Clifford S. Asness, Andrea Frazzini, and Lasse Heje Pedersen. Quality Minus Junk. Technical report, AQR, 2013. URL `http://papers.ssrn.com/abstract=2312432`.

Jushan Bai. Inferential Theory for Factor Models of Large Dimensions. *Econometrica*, 71 (1):135–171, 2003. ISSN 0012-9682. doi:10.1111/1468-0262.00392.

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70:191–221, 2002.

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

Jushan Bai and Serena Ng. Approximate factor models with weaker loading. Technical report, Columbia University, 2021.

Zhidong Bai and Jack W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer, 2009.

Natalia Bailey, George Kapetanios, and M Hashem Pesaran. Measurement of factor strenght: Theory and practice. 2020.

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

Eric Bair, Trevor Hastie, Debashis Paul, and Robert Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137, 2006.

Malcolm Baker and Jeffrey Wurgler. Investor sentiment and the cross-section of stock returns. *The journal of Finance*, 61(4):1645–1680, 2006.

Svetlana Bryzgalova. Spurious Factors in Linear Asset Pricing Models. Technical report, Stanford University, 2015.

Svetlana Bryzgalova, Jiantao Huang, and Christian Julliard. Bayesian solutions for the factor zoo: We just ran two quadrillion models. *Available at SSRN 3481736*, 2019.

Svetlana Bryzgalova, Markus Pelger, and Jason Zhu. Forest through the trees: Building cross-sections of asset returns. Technical report, London School of Business and Stanford University, 2020.

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

Gary Chamberlain and Michael Rothschild. Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51:1281–1304, 1983.

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

Andrew Y Chen and Tom Zimmermann. Open source cross-sectional asset pricing. *Available at SSRN*, 2020.

Alexandre d'Aspremont, Laurent El Ghaoui, Michael I Jordan, and Gert R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49 (3):434–448, January 2007.

Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Eugene F. Fama and Kenneth R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56, 1993. ISSN 0304405X. doi:10.1016/0304-405X(93)90023-5.

Jianqing Fan and Yingying Fan. High dimensional classification using features annealed independence rules. *The Annals of Statistics*, 36:2605–2637, 2008.

Jianqing Fan and Yuan Liao. Learning latent factors from diversified projections and its applications to over-estimated and weak factors. *Journal of the American Statistical Association*, 117(538):909–924, 2022.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, B*, 70(5):849–911, 2008.

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

Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society, B*, 75: 603–680, 2013.

Jianqing Fan, Yuan Ke, and Yuan Liao. Augmented factor models with applications to validating market risk factors and forecasting bond risk premia. *Journal of Econometrics*, 222(1):269–294, 2021.

Jon Faust and Jonathan H Wright. Forecasting inflation. In *Handbook of economic forecasting*, volume 2, pages 2–56. Elsevier, 2013.

Guanhao Feng, Stefano Giglio, and Dacheng Xiu. Taming the factor zoo: A test of new factors. *Journal of Finance*, 75(3):1327–1370, 2020.

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

Mario Forni, Marc Hallin, Marco Lippi, and Lucrezia Reichlin. The generalized dynamic-factor model: Identification and estimation. *The Review of Economics and Statistics*, 82: 540–554, 2000.

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

Mario Forni, Domenico Giannone, Marco Lippi, and Lucrezia Reichlin. Opening the black box: Structural factor models with large cross sections. *Econometric Theory*, 25:1319–1347, 2009.

Andrea Frazzini and Lasse Heje Pedersen. Betting against beta. *Journal of Financial Economics*, 111(1):1–25, 2014. ISSN 0304405X. doi:10.1016/j.jfineco.2013.10.005. URL `http://dx.doi.org/10.1016/j.jfineco.2013.10.005`.

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

Patrick Gagliardini, Elisa Ossola, and Olivier Scaillet. Time-varying risk premium in large cross-sectional equity datasets. *Econometrica*, 84(3):985–1046, 2016.

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

Stefano Giglio and Dacheng Xiu. Asset pricing with omitted factors. *Journal of Political Economy*, 129(7):1947–1990, 2021.

Stefano Giglio, Dacheng Xiu, and Dake Zhang. Test assets and weak factors. Technical report, National Bureau of Economic Research, 2021.

Stefano Giglio, Bryan Kellly, and Dacheng Xiu. Factor models, machine learning, and asset pricing. *Annual Review of Financial Economics*, 14:337–368, 2022.

Stefano Giglio, Dacheng Xiu, and Dake Zhang. Prediction when factors are weak. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2023-47), 2023.

Nikolay Gospodinov, Raymond Kan, and Cesare Robotti. Chi-squared tests for evaluation and comparison of asset pricing models. *Journal of Econometrics*, 173(1):108–125, 2013. ISSN 03044076. doi:10.1016/j.jeconom.2012.11.002. URL `http://dx.doi.org/10.1016/j.jeconom.2012.11.002`.

Nikolay Gospodinov, Raymond Kan, and Cesare Robotti. Misspecification-Robust Inference in Linear Asset-Pricing Models with Irrelevant Risk Factors. *The Review of Financial Studies*, 27(7):2139–2170, 2014. ISSN 14657368. doi:10.1093/rfs/hht135.

Campbell R. Harvey, Yan Liu, and Heqing Zhu. ...and the Cross-Section of Expected Returns. *The Review of Financial Studies*, 29(1):5–68, 2016. ISSN 0893-9454. doi:10.1093/rfs/hhv059.

Zhiguo He, Bryan Kelly, and Asaf Manela. Intermediary asset pricing: New evidence from many asset classes. *Journal of Financial Economics*, 126(1):1–35, 2017.

Kewei Hou, Chen Xue, and Lu Zhang. Replicating anomalies. *Review of Financial Studies*, 33(5):2019–2133, 2020.

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

Dashan Huang, Fuwei Jiang, Jun Tu, and Guofu Zhou. Investor sentiment aligned: A powerful predictor of stock returns. *The Review of Financial Studies*, 28(3):791–837, 2015.

Dashan Huang, Fuwei Jiang, Kunpeng Li, Guoshi Tong, and Guofu Zhou. Scaled pca: A new approach to dimension reduction. *Management Science*, 68(3):1678–1695, 2022.

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

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

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

Kyle Jurado, Sydney C Ludvigson, and Serena Ng. Measuring uncertainty. *The American Economic Review*, 105(3):1177–1216, 2015.

Raymond Kan and Chu Zhang. Two-Pass Tests of Asset Pricing Models with Useless Factors. *The Journal of Finance*, 54(1):203–235, 1999.

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

Bryan Kelly and Seth Pruitt. The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics*, 186(2):294–316, 2015.

Bryan Kelly, Seth Pruitt, and Yinan Su. Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, 134(3):501–524, 2019.

Soohum Kim, Robert A. Korajczyk, and Andreas Neuhierl. Arbitrage portfolios. *Review of Financial Studies, Forthcoming*, 2020.

Frank Kleibergen. Tests of risk premia in linear factor models. *Journal of Econometrics*, 149(2):149–173, 2009. ISSN 03044076. doi:10.1016/j.jeconom.2009.01.013. URL `http://dx.doi.org/10.1016/j.jeconom.2009.01.013`.

Serhiy Kozak, Stefan Nagel, and Shrihari Santosh. Shrinking the cross-section. *Journal of Financial Economics*, 135(2):271–292, 2020.

Martin Lettau and Markus Pelger. Estimating latent asset-pricing factors. *Journal of Econometrics*, 218:1–31, 2020.

Jonathan Lewellen, Stefan Nagel, and Jay Shanken. A skeptical appraisal of asset pricing tests. *Journal of Financial Economics*, 96(2):175–194, 2010. ISSN 0304405X. doi:10.1016/j.jfineco.2009.09.001. URL `http://dx.doi.org/10.1016/j.jfineco.2009.09.001`.

Sydney C Ludvigson and Serena Ng. A factor analysis of bond risk premia. In Aman Ulah and David E. A. Giles, editors, *Handbook of empirical economics and finance*, volume 1, chapter 12, pages 313–372. Chapman and Hall, Boca Raton, FL, 2010.

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

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

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

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

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

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

Luboš Pástor and Robert F Stambaugh. Liquidity risk and expected stock returns. *Journal of Political Economy*, 111(3):642–685, 2003.

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

M Hashem Pesaran and Ron Smith. The role of factor strength and pricing errors for estimation and inference in asset pricing models. 2019.

Stephen A. Ross. The Arbitrage Theory of Capital Asset Pricing. *Journal of Economics Theory*, 13:341–360, 1976.

Frank Schorfheide, Dongho Song, and Amir Yaron. Identifying long-run risks: A bayesian mixed-frequency approach. *Econometrica*, 86(2):617–654, 2018.

Jay Shanken. On the Estimation of Beta Pricing Models. *The Review of Financial Studies*, 5(1):1–33, 1992. ISSN 1098-6596. doi:10.1017/CBO9781107415324.004.

James H Stock and Mark W Watson. Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association*, 97(460):1167–1179, 2002a. ISSN 0162-1459. doi:10.1198/016214502388618960. URL `http://www.jstor.org/stable/3085839`.

James H Stock and Mark W Watson. Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162, 2002b.

Yoshimasa Uematsu and Takashi Yamagata. Estimation of sparsity-induced weak factor models. *Journal of Business & Economic Statistics*, 41(1):213–227, 2022a.

Yoshimasa Uematsu and Takashi Yamagata. Inference in sparsity-induced weak factor models. *Journal of Business & Economic Statistics*, 41(1):126–139, 2022b.

Yoshimasa Uematsu, Yingying Fan, Kun Chen, Jinchi Lv, and Wei Lin. Sofar: Large-scale association network learning. *IEEE transactions on information theory*, 65(8):4924–4939, 2019.

Runzhe Wan, Yingying Li, Wenbin Lu, and Rui Song. Mining the factor zoo: Estimation of latent factor models with sufficient proxies. *Journal of Econometrics*, 2023.

Weichen Wang and Jianqing Fan. Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Ann. Statist.*, 45(3):1342–1374, 06 2017. doi:10.1214/16-AOS1487. URL `https://doi.org/10.1214/16-AOS1487`.

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

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15:265–286, 2006.