DISCUSSION PAPER SERIES

DP16307

Test Assets and Weak Factors

Stefano Giglio, Dacheng Xiu and Dake Zhang

FINANCIAL ECONOMICS

ISSN 0265-8003

Test Assets and Weak Factors

Stefano Giglio, Dacheng Xiu and Dake Zhang

Discussion Paper DP16307
Published 28 June 2021
Submitted 26 June 2021

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programmes:

- Financial Economics

Any opinions expressed here are those of the author(s) and not those of the Centre for Economic Policy Research. Research disseminated by CEPR may include views on policy, but the Centre itself takes no institutional policy positions.

The Centre for Economic Policy Research was established in 1983 as an educational charity, to promote independent analysis and public discussion of open economies and the relations among them. It is pluralist and non-partisan, bringing economic research to bear on the analysis of medium- and long-run policy questions.

These Discussion Papers often represent preliminary or incomplete work, circulated to encourage discussion and comment. Citation and use of such a paper should take account of its provisional character.

Test Assets and Weak Factors

Abstract

Estimation and testing of factor models in asset pricing requires choosing a set of test assets. The choice of test assets determines how well different factor risk premia can be identified: if only few assets are exposed to a factor, that factor is weak, which makes standard estimation and inference incorrect. In other words, the strength of a factor is not an inherent property of the factor: it is a property of the cross-section used in the analysis. We propose a novel way to select assets from a universe of test assets and estimate the risk premium of a factor of interest, as well as the entire stochastic discount factor, that explicitly accounts for weak factors and test assets with highly correlated risk exposures. We refer to our methodology as supervised principal component analysis (SPCA), because it iterates an asset selection step and a principal-component estimation step. We provide the asymptotic properties of our estimator, and compare its limiting behavior with that of alternative estimators proposed in the recent literature, which rely on PCA, Ridge, Lasso, and Partial Least Squares (PLS). We find that the SPCA is superior in the presence of weak factors, both in theory and in finite samples. We illustrate the use of SPCA by applying it to estimate the risk premia of several tradable and nontradable factors, to evaluate asset managers' performance, and to de-noise asset pricing factors.

JEL Classification: N/A

Keywords: N/A

Stefano Giglio - stefano.giglio@yale.edu
Yale University and CEPR

Dacheng Xiu - dacheng.xiu@chicagobooth.edu
University of Chicago

Dake Zhang - dkzhang@chicagobooth.edu
University of Chicago

Acknowledgements
We benefited tremendously from discussions with seminar and conference participants at the University of California, Los Angeles, University of North Carolina at Chapel Hill, University of Miami, Temple University, University of Michigan, London School of Economics, Tilburg University, INSEAD Business School, Yale School of Management, Chicago Booth School of Business, Rotterdam, Rutgers University, Cornell University, Duke University, University of Texas at Dallas, ESSEC Business School, Chinese University of Hong Kong, University of Connecticut, Federal Reserve Board, ITAM Business School, Vienna Graduate School Finance, Paris Dauphine University, Goethe University, and 13th Annual SoFiE Conference.

Test Assets and Weak Factors[*]

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

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

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

This Version: June 25, 2021

Abstract

Estimation and testing of factor models in asset pricing requires choosing a set of test assets. The choice of test assets determines how well different factor risk premia can be identified: if only few assets are exposed to a factor, that factor is weak, which makes standard estimation and inference incorrect. In other words, the strength of a factor is not an inherent property of the factor: it is a property of the cross-section used in the analysis. We propose a novel way to select assets from a universe of test assets and estimate the risk premium of a factor of interest, as well as the entire stochastic discount factor, that explicitly accounts for weak factors and test assets with highly correlated risk exposures. We refer to our methodology as supervised principal component analysis (SPCA), because it iterates an asset selection step and a principal-component estimation step. We provide the asymptotic properties of our estimator, and compare its limiting behavior with that of alternative estimators proposed in the recent literature, which rely on PCA, Ridge, Lasso, and Partial Least Squares (PLS). We find that the SPCA is superior in the presence of weak factors, both in theory and in finite samples. We illustrate the use of SPCA by applying it to estimate the risk premia of several tradable and nontradable factors, to evaluate asset managers' performance, and to de-noise asset pricing factors.

Key words: Supervised PCA, SPCA, PCA, risk premium, factor models, APT, Ridge, Lasso, stochastic discount factor

[*] We benefited tremendously from discussions with seminar and conference participants at the University of California, Los Angeles, University of North Carolina at Chapel Hill, University of Miami, Temple University, University of Michigan, London School of Economics, Tilburg University, INSEAD Business School, Yale School of Management, Chicago Booth School of Business, Rotterdam, Rutgers University, Cornell University, Duke University, University of Texas at Dallas, ESSEC Business School, Chinese University of Hong Kong, University of Connecticut, Federal Reserve Board, ITAM Business School, Vienna Graduate School Finance, Paris Dauphine University, Goethe University, and 13th Annual SoFiE Conference.

[†] Address: 165 Whitney Avenue, New Haven, CT 06520, USA. E-mail address: `stefano.giglio@yale.edu`.

[‡] Address: 5807 S Woodlawn Avenue, Chicago, IL 60637, USA. E-mail address: `dacheng.xiu@chicagobooth.edu`.

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

1 Introduction

Inference on factor risk premia is a central element of empirical work in asset pricing. An essential role in this exercise is played by the set of test assets used in the estimation, yet little work has been dedicated to investigating rigorously and systematically how they should be chosen. In this paper, we show that there is an important connection between the selection of test assets and the long-standing problem of weak factors in asset pricing – factors to which the test assets have little or no exposure, resulting in a well-known failure in risk premia inference.

Central to understanding this connection is an alternative perspective on the issue of weak factors. We argue the strength or weakness of a factor 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. By exploiting this insight, we propose a new methodology for risk premia estimation, *supervised principal component analysis* (SPCA), which tackles the issue of weak factors via supervised test asset selection.

As discussed in the literature (e.g., Jagannathan and Wang (1998) and Giglio and Xiu (2021)), estimating and testing the risk premia of some factors 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. Importantly, the choice of test assets determines the strength not only of the factor of interest (e.g., liquidity), but also of *all* the other factors that drive the stochastic discount factor. We design the SPCA procedure using an iterative algorithm that uses the factor of interest to guide the selection of test assets. At the same time, the algorithm uses PCA to recover the relevant latent factors iteratively, thus controlling for potentially omitted factors. The integration of supervised selection and PCA yields a general methodology that is robust to the omission of factors, even when these omitted factors are weak.

In a nutshell, the procedure estimates the risk premium of a factor g_t 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 these portfolios (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). 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 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 p times, where p can be either a prior estimate of the number of factors in the data or can be regarded as a tuning parameter to be determined by some validation step. This procedure recovers from the data p 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 p latent factors allows us to estimate the risk premium for g_t by linking it to the risk premia for these latent factors, yielding a consistent estimator of the risk premium of g_t.

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 standard approaches to choose test assets 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: the test assets do not contain sufficient information to identify their risk premia. Using a large cross-section of test assets (the second approach) may appear, on the surface, to address this issue: these assets contain returns that are exposed to a large number of underlying factors. However, and importantly, if only a *few* of those many assets are exposed to some factor, whereas most others are not, that factor will, again, be weak in this large cross-section, disrupting inference on the risk premium. Finally, the third approach – building targeted portfolios of assets sorted by the exposure to the factor of interest – is affected by the omitted factor problem, since it considers univariate exposures only (exposures with the factor of interest may also capture correlated exposures to other risks in the economy); 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, that the Lasso approach is consistent for the estimation of the stochastic discount factor (SDF), and hence 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 with the standard two-pass regressions, so that we can easily compare SPCA with other estimators. The simulations confirm the robustness of SPCA to both omitted factors and weak factors, as well as measurement error, which SPCA also tackles.

Finally, we illustrate the use of SPCA for estimating risk premia of a variety of tradable and nontradable factors proposed in the asset pricing literature. 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 the factor risk premia, and study the out-of-sample performance of SPCA. 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). In addition to estimating risk premia, we explore additional applications of SPCA, including the performance evaluation of money managers and the removal of measurement error (de-noising) of factors.

This paper builds on a large literature on risk premia and factor model estimation and their limits in the presence of weak and omitted factors. 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 if betas are relatively small.[1] This issue is quite relevant in practice because many test assets are not very sensitive to macroeconomic shocks. Moreover, the same rank-deficiency problem 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.[2]

[1] Also related is Pesaran and Smith (2019), who 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.

[2] 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

Several recent papers have proposed different methodologies to deal with weak factors. Lettau and Pelger (2020) are among the first to study the issue of weak latent factors in a related problem, that of estimation of the SDF. They propose an estimator of the SDF in the presence of weak factors, which generalizes PCA with a penalty term that accounts for expected returns; they refer to the estimator as risk premium PCA, or rpPCA. Their objective is different from ours, but the SDF estimated using this procedure can still be used to estimate risk premia, since risk premia are covariances with the SDF. 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.[3] Anatolyev and Mikusheva (2021) propose an complementary approach to dealing with weak factors, based on sample-splitting and instrumental variables. This 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.

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). It also relates to a growing strand of econometrics literature on weak factor models, like Bai and Ng (2008) and Huang et al. (2021). Our SPCA approach shares the spirit of these approaches, but is more involved because we do not assume all factors are of the same strength, which thereby requires multiple selection steps. Also, our focus is on risk premia estimation instead of forecasting, for which we also provide inference. Also related are papers that propose estimators of factor count and strength, like Freyaldenhoven (2019) and Bailey 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. 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, the econometrician may not, making the factor weak for the set of test assets available to the econometrician.

[3] 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 indeed, rpPCA does not recover consistently the SDF, but 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 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.

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 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 inference on parameters (i.e., risk premia), which involves an additional step and more intricate analysis for the asymptotic theory.

The paper is organized as follows. Section 2 first sets up the notation and model (Sections 2.1 and 2.2), then discusses the inconsistency of existing estimators in the presence of weak factors (Section 2.3), provides our methodology (Sections 2.4 and 2.5) and finally the inference theory (Section 2.6). Section 3 provides simulation evidence, followed by an empirical study in Section 4. The appendix provides technical details.

2 Methodology

2.1 Notation

Throughout the paper, we use (A, B) to denote the concatenation (by columns) of two matrices A and B. e_i is a vector with 1 in the ith entry and 0 elsewhere, whose dimension depends on the context. ι_k denotes a k-dimensional vector with all entries being 1, and $\mathbb{I}_d$ denotes the $d \times d$ identity matrix. 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. We denote $\mathbb{P}_A = A(A^\intercal A)^{-1}A^\intercal$ and $\mathbb{M}_A = \mathbb{I}_d - \mathbb{P}_A$, for some $d \times T$ matrix A. 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 $a \lesssim_p b$ to denote $a = O_p(b)$. If $a \lesssim b$ and $b \lesssim a$, we write $a \asymp b$ for short. Similarly, we use $a \asymp_p b$ if $a \lesssim_p b$ and $b \lesssim_p a$.

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 $\|A\|_1$, $\|A\|_\infty$, $\|A\|$, and $\|A\|_{\mathrm{F}}$ to denote the $\mathbb{L}_1$ norm, the $\mathbb{L}_\infty$ norm, the operator norm (or $\mathbb{L}_2$ norm), and the Frobenius norm of a matrix $A = (a_{ij})$, that is, $\max_j \sum_i |a_{ij}|$, $\max_i \sum_j |a_{ij}|$, $\sqrt{\lambda_{\max}(A^\intercal A)}$, and $\sqrt{\mathrm{Tr}(A^\intercal A)}$, respectively. We also use $\|A\|_{\mathrm{MAX}} = \max_{i,j}|a_{ij}|$ to denote the $\mathbb{L}_\infty$ norm of A on the vector space. When a is a vector, we use $\|a\|_0$ to denote $\sum_i 1_{\{a_i \neq 0\}}$. We also denote $Supp(a) = \{i : a_i \neq 0\}$. Finally, we use $[N]$ to denote the set of integers: $\{1, 2, \ldots, N\}$. For an index set $I \subset [N]$, we use $|I|$ to denote its cardinality. We use $A_{[I]}$ to denote a submatrix of A whose rows are indexed in I.

2.2 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:

$$r_t = \beta\gamma + \beta v_t + u_t, \quad \mathrm{E}(v_t) = \mathrm{E}(u_t) = 0 \text{ and } \mathrm{Cov}(v_t, u_t) = 0, \tag{1}$$

where β is an $N \times p$ matrix of factor exposures, v_t is a $p \times 1$ vector of factor innovations, and u_t is an $N \times 1$ vector of idiosyncratic errors.[4] The v_t vector is unobservable, even though it may include factor innovations of observable factors, f_t, i.e., $v_t = f_t - \mu_f$, since μ_f is an unknown parameter.

In order to study the statistical properties of risk premia estimators in the presence of weak factors, we first define our asymptotic scheme. We will assume that both N and T go to ∞, whereas p is fixed. The $p \times p$ factor covariance matrix Σ_v is asymptotically non-singular in the sense that $1 \lesssim \lambda_{\min}(\Sigma_v) \lesssim \lambda_{\max}(\Sigma_v) \lesssim 1$. This assumption is rather weak as it only rules out factors whose risks are (asymptotically) negligible or exploding. We also maintain the assumption that $\|\Sigma_u\| \lesssim 1$, so that there exists no factor structure in the residuals u_t. This condition is useful for identification purposes, and ensures that all factors must be distinguishable from the idiosyncratic errors, regardless of their strength, which we turn to next.

In this setting, a factor's strength is entirely determined by test assets' exposures to it, since all factors have non-negligible or non-exploding risks. In light of this, the strength of a factor is context specific — the selection of test assets dictates its strength. For instance, a momentum factor could be a strong factor for momentum-sorted portfolios, but this factor may be weak with portfolios sorted by size or value as test assets, because the latter portfolios may diversify the exposure to the momentum factor.

In the econometrics literature on factor models, the most prevalent assumption adopted by, e.g., Bai and Ng (2002), is that all factors are strong or pervasive, that is, $\lambda_i(\beta^\intercal\beta) \asymp N$ for $i = 1, 2, \ldots, p$, which dominates the strength of the idiosyncratic component, as measured by $\|\Sigma_u\|$. Our focus is on the regime of *weak factors*, which covers a wide range of factor strength. In particular, the norm of columns of β is allowed to diverge at different and slower rates, which will be made more precise later.

The fact that weak factors are relevant in practice can be illustrated from a scree plot of eigenvalues of returns (for example, see Figure 3 based on the large cross-section we use in our empirical analysis). Factors with a spectrum of strength, as indicated by various magnitudes of eigenvalues, are clearly present. Except for the first one or two eigenvalues, there is not a clear-cut gap between the next few eigenvalues (that would correspond to weaker factors) and the remaining eigenvalues that correspond to idiosyncratic components.

[4] Our model is set up for portfolios as test assets. To generalize this model for individual stocks, more structures should be imposed to address time-varying risk exposures, see, e.g., Gagliardini et al. (2016), Kelly et al. (2019), and Kim et al. (2020).

We develop our discussion of weak factors in the context of two standard asset pricing exercises: the estimation of risk premia and the recovery of the stochastic discount factor (SDF). In this model, an SDF can be defined in terms of asset pricing factors v_t as

$$m_t = 1 - \gamma^\intercal \Sigma_v^{-1} v_t, \tag{2}$$

where Σ_v is the covariance matrix of factor innovations. It also makes sense to consider the SDF represented in terms of the tradable test asset returns:

$$\widetilde{m}_t = 1 - b^\intercal (r_t - \mathrm{E}(r_t)), \tag{3}$$

where b is an $N \times 1$ vector of SDF loadings which satisfies $\mathrm{E}(r_t) = \Sigma b$, where Σ is the covariance matrix of r_t. 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 risk premia of some observable factors, summarized in a $d \times 1$ vector, g_t. Following Giglio and Xiu (2021), we do not impose that g_t is part of or is identical to v_t; instead, we assume g_t and v_t are (potentially) correlated:

$$g_t = \xi + \eta v_t + z_t, \tag{4}$$

where $\xi = \mathrm{E}(g_t)$, η is a $d \times p$ matrix, and z_t is measurement error orthogonal to v_t.[5] The risk premia of the factors g_t are $\eta\gamma$, our parameter of interest in this paper. 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 $z_t = 0$.

Since the true factors in v_t are potentially weak, the observable factors in g_t may therefore also be weak because the exposure of r_t to g_t is partially determined by that to v_t. The risk exposure of g_t (to v_t), η, and risk premia, γ, are not necessarily diminishing (asymptotically). Specifically, $\eta\gamma$ could be a fixed parameter that does not vary with sample size.

2.3 Inconsistency of Existing Estimators

While the literature has proposed many different estimators of the SDF and risk premia, their properties in the weak factor setting have not been studied. In what follows, we revisit a number of existing procedures for estimating risk premia, and show that they are inconsistent in the presence of weak factors using a simple model with a single weak factor.

[5] 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.

2.3.1 PCA

Giglio and Xiu (2021) suggest a three-pass procedure to estimate $\eta\gamma$: 1) apply PCA to the sample covariance matrix of returns to obtain estimates of the latent factors, $\widehat{v}_t$;[6] 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 1 (PCA-based Estimator of Risk Premia)**.** *The estimator proceeds as follows:*
Inputs: $\bar{R}$ and $\bar{G}$.

S1. Apply SVD on $\bar{R}$, and write the first p right singular vector 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 $\widehat{\beta} = \bar{R}\widehat{V}^\intercal(\widehat{V}\widehat{V}^\intercal)^{-1}$.

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 $\widehat{\gamma}_g^{PCA} = \widehat{\eta}\widehat{\gamma}$.

Giglio and Xiu (2021) establish the consistency of this estimator and derive its asymptotic inference, in the case that all latent factors are pervasive, whereas g_t can be either strong or weak (depending on the magnitude of η). This risk-premia estimator is appealing for its simplicity, efficiency, and robustness to missing factors. Unfortunately, it fails when some latent factors are weak, which we will show next.

To explain the intuition behind the failure of PCA, 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 it simple, we also assume that the factor of interest g_t has no measurement error, i.e., $z_t = 0$ and $g_t = \eta v_t$.

A successful recovery of β via PCA of realized returns requires a favorable signal-to-noise ratio. If the "signal" as measured by $\|\beta\|$, dominates "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 $\widehat{\beta}$, effectively the eigenvector of $\widehat{\Sigma}$, in the cross-sectional regression 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 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, which eventually results in an inconsistent estimate of the risk premium $\eta\gamma$. Proposition 1 below shows that the PCA-based risk premium estimator is consistent only if $N/(\|\beta\|^2 T) \to 0$.

[6] Equivalently, one can directly apply the singular value decomposition (SVD) on $\bar{R}$.

Proposition 1. *Suppose that test asset returns follow a single-factor model in the form of* (1) *with* $p = 1$*,* g_t *satisfies* (4) *with* $d = 1$*, and* 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* $\widehat{\gamma}_g^{PCA} \xrightarrow{p} (1 + B)^{-1}\eta\gamma$*.*

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 are T are of the same order, $\|\beta\| \to \infty$ is sufficient for the consistency of risk pemia 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 1 in the sense that $|\text{Corr}(\widehat{V}, V)| \xrightarrow{p} 1$.[7] However, estimating a large-dimensional vector β given $\widehat{V}$ remains a challenging problem, which requires this additional condition, $B = 0$, to achieve consistency.

2.3.2 PLS

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 now 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 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. We formulate a general PLS-based algorithm for a $d \times 1$ vector of g_t below:

Algorithm 2 (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 $w$ from the largest left singular vector of $\bar{R}_{(k)}\bar{G}^\intercal$.
```

[7] 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 1.

b. *Estimate the kth factor as* $\widehat{V}_{(k)} = \sqrt{T}w^\intercal \bar{R}_{(k)} / \left\| w^\intercal \bar{R}_{(k)} \right\|$. *Here,* $\widehat{V}_{(k)}$ *is normalized to have norm* $\sqrt{T}$.

c. *Estimate the risk premium of* $\widehat{V}_{(k)}$ *by* $\widehat{\gamma}_{(k)} = \sqrt{T}w^\intercal \bar{r}_{(k)} / \left\| w^\intercal \bar{R}_{(k)} \right\|$.

d. *Estimate the kth 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 2. *Suppose that test asset returns follow a single-factor model in the form of* (1) *with* $p = 1$, g_t *satisfies* (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* $\widehat{\gamma}_g^{PLS} \xrightarrow{p} (1+B)^{-1}\eta\gamma$.

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.

2.3.3 Ridge

Next, we consider an alternative ridge regression approach to the construction of mimicking portfolios, and the resulting risk premia estimator can be written as:

$$\widehat{\gamma}_g^{Ridge} = \bar{G}\bar{R}^\intercal \left(\bar{R}\bar{R}^\intercal + \mu \mathbb{I}_N \right)^{-1} \bar{r}, \tag{5}$$

where $\mu > 0$ is some tuning parameter. In the case of pervasive factors, Giglio and Xiu (2021) show that the ridge estimator yields consistent estimate of $\eta\gamma$. However, the ridge estimator also fails in the presence of weak factors:

Proposition 3. *Suppose that test asset returns follow a single-factor model in the form of* (1) *with* $p = 1$, g_t *satisfies* (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* $\widehat{\gamma}_g^{Ridge} \xrightarrow{p} (1 + B + D)^{-1}\eta\gamma$.

Even though the ridge-based risk premia estimator seemingly accounts for the impact of all eigenvectors as factors instead of only the first p of them, the resulting estimator remains inadequate for consistency. Intuitively, the tuning parameter μ in the ridge procedure serves as a threshold that impedes 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.

2.3.4 Risk Premium PCA

Finally, we consider an estimator of $\eta\gamma$ based on the risk premium PCA (rpPCA) estimator proposed by Lettau and Pelger (2020) in the context of SDF estimation.

Algorithm 3 (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* ς. *The estimated factors are given by* $\widehat{V} = \varsigma^\intercal \bar{R}$.
- *S2. Estimate the risk premia of* $\widehat{V}$ *by* $\widehat{\gamma} = \varsigma^\intercal \bar{r}$.
- *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: $\widehat{\gamma}_g^{rpPCA} = \widehat{\eta}\widehat{\gamma}$.

The 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 improves the Sharpe ratio of the estimated SDF.[8] While this estimator was originally proposed for estimating the SDF, it can be used to estimate risk premia as well (since risk premia are just covariances with the SDF). We discuss here this risk premium estimator, in a setting where a single factor can be weak yet its strength is of a distinct order relative to idiosyncratic components asymptotically. This setting is more informative for comparing different approaches, because in this setting a consistent estimation procedure exists.

[8] They 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 so weak 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 rpPCA, which, despite being more correlated with the SDF than PCA, is also inconsistent.

Proposition 4. *Suppose that test asset returns follow a single-factor model in the form of* (1) *with* $p = 1$, g_t *satisfies* (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*

$$\widehat{\gamma}_g^{rpPCA} \xrightarrow{p} w(1+B)^{-1}\eta\gamma + (1-w)\eta(\gamma + \gamma^{-1}B),$$

where

$$w = \frac{2+2B}{1+2B+\sqrt{(1-a)^2+4(1+\mu)\gamma}+a}, \qquad a = (1+\mu)(\gamma^2+B) - B.$$

Proposition 4 suggests 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 $B > 0$, we may design a different asymptotic setting, in which the tuning parameter $\mu \to \infty$, under which the rpPCA estimator converges to $\eta(\gamma + \gamma^{-1}B)$. If we further assume $\gamma \to \infty$ (while keeping $\eta\gamma$ constant), this estimator can be consistent as long as $\eta\gamma^{-1}B \xrightarrow{p} 0$. This suggests that rpPCA can be robust to weak factors if the information about β from the expected return dominates the information from return covariances (in which case factors have a diverging Sharpe ratio.)

An alternative approach to Algorithm 3, based on rpPCA, is to adapt Algorithm 1 by replacing its step S1 by S1 of Algorithm 3. It turns out that this approach yields the same asymptotic behavior as the PCA estimator of Algorithm 3, which is characterized by Proposition 1.[9] Because its performance is essentially identical with that of PCA, we omit the discussion of this version of rpPCA from the rest of the paper.

2.4 Our Solution: Test Asset Selection

The results in the previous section shed light on the limitation of dimension reduction or shrinkage estimators, when factors are not pervasive.[10] One potential solution is to screen test assets and only keep those that have nontrivial exposure to the factor of interest. Then, if the factor is strong *within this smaller set of test assets*, it is possible to apply PCA or any of the above procedures to recover its risk premium, as long as there remains a sufficient number of test assets.

This strategy echoes some of the practice in the empirical asset pricing literature. Very often,

[9] As shown by Giglio and Xiu (2021), using either left or right singular vectors of $\bar{R}$ as factors yields asymptotically equivalent PCA-based estimators of risk premia. This is, however, not true for rpPCA, because its estimated "eigenvectors" do not correspond to any singular vectors of $\bar{R}$. This is the reason why using a rpPCA adapted Algorithm 1 would lead to a different asymptotic result (equivalent to Proposition 1), as opposed to Proposition 4 based on Algorithm 3.

[10] These results should not be regarded as evidence against the use of above estimators in all scenarios. Rather, we only establish that for data generating processes in the regime of weak factors we define, none of these estimators are consistent. It is however possible that, for some alternative sequences of data generating processes, or for purposes other than risk premia estimation, these estimators may perform well.

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 of test assets indeed help address the weak factor problem, though, as discussed in the introduction, they do not address the other issue that is relevant in practical applications – omitted factors. Our methodology formalizes the insight behind these traditional procedures and combines it with the use of PCA to address the omitted factor bias.

We start with a simple one factor setting as discussed in the previous propositions, which helps illustrate the intuition behind our proposal and facilitates the comparison with existing estimators (the next section is devoted to the general case). To ensure sufficient test assets after screening, we assume that there exists a subset $I_0 \subset [N]$ such that $\left\|\beta_{[I_0]}\right\| \asymp \sqrt{N_0}$, where $N_0 = |I_0| \to \infty$. Consequently, as long as we locate this subset of assets, within which there exists a strong factor structure, we can recover risk pemia consistently. 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.

We next formally present our SPCA procedure for test assets selection and risk premia estimation.

Algorithm 4 (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.[11]

- *S1. Select a subset $\widehat{I} \subset [N]$: $\widehat{I} = \left\{i \middle| T^{-1}|\bar{R}_{[i]}\bar{G}^\intercal| \geq c_q\right\}$, where c_q is the $(1-q)$-quantile of $\left\{T^{-1}|\bar{R}_{[i]}\bar{G}^\intercal|\right\}_{i\in[N]}$.*
- *S2. Repeat S1. – S3. of Algorithm 1 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 $\widehat{\gamma}$.

We establish the consistency of the SPCA estimator in the following proposition:

Proposition 5. *Suppose that $\log N/T \to 0$ and test asset returns follow a single-factor model in the form of (1) and that g_t satisfies (4), with u_t, v_t, and z_t i.i.d. normally distributed and independent from each other. The loading matrix β satisfies $\|\beta\|_{\mathrm{MAX}} \lesssim 1$ and there exists a subset $I_0 \subset [N]$ 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 4 such that $qN/N_0 \to 0$ and $qN \to \infty$, and that $|\beta|_{\{qN+1\}} \leq (1+\delta)^{-1}|\beta|_{\{qN\}}$ for some $\delta > 0$, where $|\beta|_{\{k\}}$ denotes the kth largest value in $\left\{|\beta_{[i]}|\right\}_{i\in[N]}$, we have $\widehat{\gamma}_g^{SPCA} \xrightarrow{p} \eta\gamma$.*

Algorithm 4 involves a single tuning parameter q that determines how many assets we use to extract the factor. We select the first qN assets sorted by their covariances with the target variable

[11] We discuss the case of a multivariate ($d \times T$) $\bar{G}$ in Section 2.6.

$\bar{G}$. The fact that $\widehat{I}$ incorporates information from the target reflects the distinctive nature of a supervised procedure. The technical condition on $|\beta|_{\{qN+1\}}$ simply states that these test assets should have (asymptotically) distinct risk exposure, which is a rather mild assumption used in the proof.

Propositions 2 - 4 show that in the single factor case, the consistency of PCA, 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 5, nonetheless, only requires $\log N/T \to 0$.[12] That said, the condition $\|\beta\|^2 \gtrsim N_0 \to \infty$ rules out the case that the factor strength is of the same magnitude as that of idiosyncratic errors.[13]

2.5 The General Case: Selection and Projection

Propositions 1 - 5 focus on a perhaps unrealistic single-factor model since they are meant to illustrate the intuition behind our procedure as well as the failure of existing approaches due to the presence of a weak factor. In general, the DGP of returns is likely driven by more than one factors, some of which may be weak. In the same spirit of Proposition 1, we can show that a more general necessary condition for the consistency of PCA in a multi-factor model is that

$$N/(\lambda_{\min}(\beta^\intercal\beta)T) \to 0. \tag{6}$$

Intuitively, this condition requires that the weakest one among all p factors in (1) is sufficiently strong that it can be recovered by PCA. Once again, we consider below more challenging regimes in which the condition (6) fails.

In a multi-factor model, even if all factors are strong by themselves, a related problem arises when some of the factors' exposures are highly correlated. Consider, for example, a two-factor model where the beta matrix has the following form:

$$\beta = \left[\begin{array}{c|c} \beta_{11} & \beta_{12} \\ \hline \\ \beta_{21} & \beta_{22} \\ \\ \end{array}\right], \tag{7}$$

[12] Another idea that shares this spirit is the scaled-PCA proposed by Huang et al. (2021), 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.

[13] Throughout this paper, an extremely weak factor is referred to as a factor whose strength is of the same order of magnitude as that of idiosyncratic errors. 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 effectively become alpha. The weak-factor setting we investigate permits consistency, and allows for asymptotic comparison of different estimators.

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 $\beta_{21} = \beta_{22}$. Then we can show that $\lambda_{\min}(\beta^\intercal\beta) \leq \|\beta_{11} - \beta_{12}\|^2 /2 \lesssim N_0$. As a result, $N/(\lambda_{\min}(\beta^\intercal\beta)T) \gtrsim N/(N_0 T)$, which does not necessarily converge to 0 if N_0 and T are small, so that the condition (6) could fail. In this example, while either factor could be strong, the same "rank deficiency" issue may arise, since these factors could have highly correlated exposures.

Another important consideration is that applying the screening approach only once would in general not work in a multi-factor model. Take (7) again as an example. Suppose that $\beta_{21} \neq \beta_{22} = 0$, then it is easy to show that $\lambda_{\min}(\beta^\intercal\beta) \leq \|\beta_{12}\|^2 \lesssim N_0$, thus in light of the above discussion, the weak factor problem could occur in this example. In this case, it is the second factor that is weak since most of test assets' exposure to it is zero. Now suppose that $\eta = (1, 1)$: the observed factor g is correlated with both factors and hence with all test assets. But in that case, the screening would not eliminate any test asset – and yet PCA with all test assets would not recover the weak factor, should $N/(N_0 T)$ not vanish. This example demonstrates that even though screening assets ensures that the *first* principal component after screening is strong, there is no guarantee that this procedure can solve the weak factor issue in one step if additional factors are weak.

It is worth pointing out that the two aforementioned cases are in fact equivalent, because we can rotate the beta matrix in the second case into the form of the first case. Thanks to the rotation invariance property illustrated in Giglio and Xiu (2021), both the risk premia and the SDF estimands remain unchanged after rotation, and hence the equivalence.

In the examples above, the problem was that the first screening step did not eliminate any assets, and therefore could not solve the weak factor issue. We provide next 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 β has the following form:

$$\beta = \left[\begin{array}{c|c} \beta_{11} & \beta_{11} \\ \hline 0 & \beta_{22} \end{array}\right], \tag{8}$$

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 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 only one factor, $v_{1t} + v_{2t}$, is left. This example shows that the one-step supervised procedure (screening plus PCA) proposed by Bair et al. (2006), may be counterproductive for factor extraction in a multi-factor setting.

To resolve the issue of weak factors and avoid these screening traps, 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, Step S1 of Algorithm 4 can help identify one strong factor from a selected subset of test assets. Once we have estimated this factor, we project the returns of *all* test assets r_t (not just those selected at the first step) and g_t 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 factors extracted by standard PCA. It is easy to check that this iterative screening and projection approach successfully addresses the problems of all three examples above. Formally, the algorithm is given by:

Algorithm 5 (Selection and Projection)**.** *The selection and projection based procedure for risk premium 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* $\widehat{I}_k \subset [N]$*.*
- *b. Repeat S1. – S3. of Algorithm 1 with selected return matrix* $\big(\bar{R}_{(k)}\big)_{[\widehat{I}_k]}$ *and* $\bar{G}_{(k)}$*. Denote the estimates as* $\widehat{\lambda}_{(k)}$*,* $\widehat{V}_{(k)}$*,* $\widehat{\eta}_{(k)}$*,* $\widehat{\gamma}_{(k)}$*.*
- *c. Estimate the exposure of* $\bar{R}_{(k)}$ *on* $\widehat{V}_{(k)}$ *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 the risk premium 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* $\widehat{\beta} = (\widehat{\beta}_{(1)}, \cdots, \widehat{\beta}_{(\widehat{p})})$*.*

In Algorithm 5, we recover one latent factor and obtain its risk premium at each stage of S1. Both the factor and its risk premium are estimated using a subset of rows in the stage-k return residual matrix $\bar{R}_{(k)}$, within which this factor is strong. We then project all observables onto this factor and proceed again with residuals. Because each row of $\bar{R}_{(k+1)}$ is orthogonal to $\widehat{V}_{(j)}$ for $j \leq k$ the factors we obtain are orthogonal with each other, as is the case with PCA.

Algorithm 5 yields a consistent estimator of γ_g as long as an appropriate choice of $\widehat{I}_k$ and a stopping rule are adopted. One possible choice for $\widehat{I}_k$ is:[14]

$$\widehat{I}_k = \left\{ i \middle| T^{-1} \left\| (\bar{R}_{(k)})_{[i]} \bar{G}_{(k)}^{\intercal} \right\|_{\mathrm{MAX}} \geq c_q^{(k)} \right\},$$
$$\text{where } c_q^{(k)} \text{ is the } (1-q)th\text{-quantile of } \left\{ T^{-1} \left\| (\bar{R}_{(k)})_{[i]} \bar{G}_{(k)}^{\intercal} \right\|_{\mathrm{MAX}} \right\}_{i \in [N]}. \tag{9}$$

Correspondingly, we set the stopping criterion as:

$$c_q^{(k)} < c, \quad \text{for some threshold } c. \tag{10}$$

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 of the rows of the 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.

To establish the consistency of this estimator, we need a subset of assets, indexed by I_0, such that within this subset all factors are strong, that is, $\lambda_{\min}(\beta_{[I_0]}^{\intercal}\beta_{[I_0]}) \asymp N_0$, 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 is strong.[15] With this identification assumption, along with moment conditions given in the appendix, the following theorem establishes the consistency of the SPCA estimator:

Theorem 1. *Suppose that test asset returns in r_t follow* (1)*, the factor proxies in g_t satisfy* (4)*, and that Assumptions A.1-A.8 hold. If* $\log(NT)(N_0^{-1} + T^{-1}) \to 0$ *then for any tuning parameters c and q that satisfy*

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

we have $\widehat{\gamma}_g^{SPCA} \xrightarrow{p} \eta\gamma$.

The consistency result in Theorem 1 does not require a full recovery of all factors that drive the SDF. In fact, only factors correlated with g_t will be recovered. Missing any uncorrelated factors in the SDF does not affect the consistency of the risk premium of g_t because such factors do not help price g_t.

Moreover, this result 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

[14] Using covariance for screening allows us to replace all $\bar{G}_{(k)}$ in the definition of $\widehat{I}_k$ and Algorithm 5 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 $\widehat{I}_k$. Despite this does not affect the asymptotic analysis, we find correlation screening performs slightly better in finite samples.

[15] This assumption is weak in that it does not imply all factors should have identifical strength with respect to the entire cross-section of assets in r_t. A detailed discussion on this point follows Assumption A.3 in the appendix.

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.

2.6 Asymptotic Inference on Risk Premia

In this section we develop the asymptotic distribution of the risk premium estimator from Algorithm 5. Not surprisingly, the conditions in Theorem 1 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, 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. Even if g_t is independent with respect to the test assets, their sample correlation can be as large as $T^{-1/2}$. Therefore, the threshold needs no smaller than $T^{-1/2}$. However, for any given threshold, say, $T^{-1/4}$, if it happens that $\eta \asymp T^{-1/3} < T^{-1/4}$, then it suggests that g_t is not too different from random noise, so that screening based on its correlation with r_t will likely not select any assets, which in turn leads to 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 $T^{-1/3}$, so that the usual central limit theorem (CLT) fails.

Generally speaking, this issue arises because of the potential failure to identify all factors in the DGP. Once all factors are identified, the central limit theorem holds regardless of the magnitude of η. So to make 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 though it does not affect consistency. It turns out that so long as $\eta \in \mathbb{R}^{d\times p}$ satisfies $\lambda_{\min}(\eta^\intercal\eta) \gtrsim 1$, we can rule out the possibility of missing factors. On the other hand, our algorithm will not select more factors than needed, if we stop the iteration as soon as $c_q^{(k)}$ is sufficiently small. Of course, in a finite sample, a perfect recovery of the factor space is a stretch, but the assumptions here are substantially weaker than the pervasive factor assumption adopted in the literature, e.g., Bai (2003). The inference theory on factor models also relies on a perfect recovery of the count of (strong) factors, e.g., Bai and Ng (2002). We provide below the consistency result on the number of factors and the CLT result on risk premium, and investigate the finite sample behavior of SPCA in Section 3.

Theorem 2. *Under the same assumptions as Theorem 1, if we further have $T^{-1/2}N_0 \to \infty$, Assumption A.9 and $\lambda_{\min}(\eta^\intercal\eta) \gtrsim 1$, then for any tuning parameters c and q in (9) and (10) satisfying*

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

we have that $\widehat{p}$ defined in Algorithm 5 satisfies: $\widehat{p} \xrightarrow{p} p$, and that the estimator constructed via Algorithm 5 satisfies

$$\sqrt{T}\left(\widehat{\gamma}_g^{SPCA} - \eta\gamma\right) \xrightarrow{d} \mathcal{N}\left(0, \Phi\right),$$

where Φ is given by

$$\Phi = \left(\gamma^\intercal \Sigma_v^{-1} \otimes \mathbb{I}_d\right) \Pi_{11} \left(\gamma^\intercal \Sigma_v^{-1} \otimes \mathbb{I}_d\right) + \left(\gamma^\intercal \Sigma_v^{-1} \otimes \mathbb{I}_d\right) \Pi_{12} \eta^\intercal + \eta \Pi_{12}^\intercal \left(\gamma^\intercal \Sigma_v^{-1} \otimes \mathbb{I}_d\right) + \eta \Pi_{22} \eta^\intercal,$$

and Π_{11}, Π_{12}, and Π_{22} are specified by Assumption A.9.

We can adopt 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.

The condition $\lambda_{\min}(\eta^\intercal \eta) \gtrsim 1$ implies $d \geq p$, that is, we need g_t to have at least equal number of variables as the true number of factors. Moreover, the condition also implies that for each factor in v_t, there is at least one variable in g_t with a non-vanishing exposure to it.

2.7 The Case of Observable Factors

The previous discussion does not assume any knowledge of the identities of the factors v_t in (1). If v_t corresponds to innovations of observable factors, denoted by f_t, which were known (by assumption), say, the Fama-French five factors, our procedure can be greatly simplified. It is meaningful to study this case, because it is most common in the empirical literature, albeit this is a (rather) strong assumption.

Suppose factors in f_t are tradable. If g_t is part of them, 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 averages, give rise to the risk premium estimate of g. 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 their mimicking portfolios. 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 5, 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.

2.8 Asymptotic Inference on Alpha

As a by-product, we can also make inference on the pricing error, α_g, defined as $\mathrm{E}(g_t) - \gamma_g$, when g_t is tradable. In practice, this exercise is most relevant for inferring the "skill" of a fund manager; we explore this application in section 4.2.1. Using the SPCA estimator $\widehat{\gamma}_g^{SPCA}$, we can directly construct $\widehat{\alpha}_g = \bar{g} - \widehat{\gamma}_g^{SPCA}$. We now provide its corresponding CLT result.

Theorem 3. *Suppose the same assumptions as those in Theorem 2 hold. If we further have Assumption A.10, then the estimator $\widehat{\alpha}_g$ satisfies*

$$\sqrt{T}(\widehat{\alpha}_g - \alpha_g) \xrightarrow{d} \mathcal{N}(0, \widetilde{\Phi}),$$

where $\widetilde{\Phi}$ is given by

$$\widetilde{\Phi} = \left(\gamma^{\intercal}\Sigma_v^{-1} \otimes \mathbb{I}_d\right) \Pi_{11} \left(\Sigma_v^{-1}\gamma \otimes \mathbb{I}_d\right) - \left(\gamma^{\intercal}\Sigma_v^{-1} \otimes \mathbb{I}_d\right) \Pi_{13} - \Pi_{13}^{\intercal} \left(\Sigma_v^{-1}\gamma \otimes \mathbb{I}_d\right) + \Pi_{33}.$$

It is straightforward to construct a Newey-West-type estimator of the asymptotic variance $\widetilde{\Phi}$ via its sample analog.

2.9 Recovery of the Stochastic Discount Factor

The main focus of the previous sections is on risk premia, whose consistency does not require a consistent recovery of the SDF, since some of these factors driving SDF might be uncorrelated with the factors of interest, and will therefore not play any role in the consistency of risk premia. Nonetheless, we have pointed out that constructing valid asymptotic inference requires the recovery of all factors that drive the SDF. In this case, we can also reconstruct the SDF. More specifically, from the outputs of Algorithm 5, we can estimate the SDF by:

$$\widehat{m}_t^{SPCA} = 1 - \widehat{\gamma}^{\intercal}\widehat{v}_t, \qquad \text{where } \widehat{v}_1, \cdots, \widehat{v}_T \text{ are the columns of } \widehat{V}. \tag{11}$$

Theorem 4. *Suppose the same assumptions as in Theorem 2 hold. In addition, we have Assumption A.11. Then the estimator (11) satisfies*

$$\frac{1}{T}\sum_{t=1}^{T}|\widehat{m}_t^{SPCA} - m_t|^2 \lesssim_P \frac{1}{T} + \frac{\log N_0}{N_0}. \tag{12}$$

There are a number of alternative approaches for SDF estimation 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 same 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 will provide useful insights that may guide their applications in practice.

Kozak et al. (2020) consider an SDF in the form of (3), whereas we represent it as in (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 6. *Suppose that test asset returns in r_t follow (1), and Assumption A.11 holds. Then as $N \to \infty$, we have*

$$\frac{1}{T}\sum_{t=1}^{T}|m_t - \widetilde{m}_t|^2 \lesssim_p \frac{1}{\lambda_{\min}(\beta^\intercal \beta)}.$$

Effectively, Proposition 6 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 4, and that $\lambda_{\min}(\beta^\intercal\beta) \gtrsim N_0$, we can replace m_t in the left-hand side of (12) by $\widetilde{m}_t$.

Kozak et al. (2020) suggest estimating the SDF by solving an optimization problem:

$$\widehat{b} = \arg\min_b \left\{(\bar{r} - \widehat{\Sigma}b)^\intercal \widehat{\Sigma}^{-1}(\bar{r} - \widehat{\Sigma}b) + p_\mu(b)\right\}, \tag{13}$$

with which the estimated pricing kernel is given by

$$\widehat{m}_t = 1 - \widehat{b}^\intercal(r_t - \bar{r}). \tag{14}$$

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 (13) appears to require the inverse of the sample covariance matrix $\widehat{\Sigma}^{-1}$, which is not well-defined when $N > T$. Instead, we suggest optimizing an equivalent but different form of (13):

$$\widehat{b} = \arg\min_b \left\{b^\intercal\widehat{\Sigma}b - 2b^\intercal\bar{r} + b^\intercal\widehat{\Sigma}b + p_\mu(b)\right\}, \tag{15}$$

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.

Theorem 5. *We investigate two distinct scenarios.*

(a) Suppose that r_t is driven by p latent factors as in (1). With $p_\mu(b) = \mu\|b\|^2$, if $(N+T)/(\lambda_p T) \to 0$ and Assumptions A.4-A.7, A.11-A.13 hold, we have

$$\frac{1}{T}\sum_{t=1}^{T}|\widehat{m}_t^{Ridge} - m_t|^2 \lesssim_p \frac{1}{T} + \frac{N+T}{\lambda_p T},$$

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* $p_\mu(b) = \mu \left\|b\right\|_1$*, if Assumptions A.11, A.12 hold, we have*

$$\frac{1}{T}\sum_{t=1}^{T} |\widehat{m}_t^{Lasso} - \widetilde{m}_t|^2 \lesssim_P \left\|b\right\|_1 \sqrt{\frac{\log N}{T}}. \tag{16}$$

If, in addition, we assume that $\lambda_{\min}(\Sigma) \gtrsim 1$*, and* $\left\|b\right\|_0^2 \log N/T \to 0$*, then we have a stronger result*

$$\frac{1}{T}\sum_{t=1}^{T} |\widehat{m}_t^{Lasso} - \widetilde{m}_t|^2 \lesssim_P \left\|b\right\|_0 \frac{\log N}{T}. \tag{17}$$

Interestingly, both the Ridge and Lasso approaches deliver consistent estimates of the SDF, though under rather different sets of assumptions. First of all, the convergence rate of the Ridge approach depends critically on the strength of the weakest factor. If condition (6) fails, then the SDF is not consistent. The failure of this condition is precisely a symptom of weak factors which our SPCA estimator is designed for.

Second, with respect to the estimator using the Lasso penalty, the explicit factor model assumption on r_t is replaced by the sparsity assumption on b. The latter assumption requires that the SDF is spanned by a sparse linear combination of test assets, but place no explicit assumptions on the DGP of these test assets. This suggests that the Lasso estimator remains consistent regardless of the factor strength, but converges at a rather slow rate, $\left\|b\right\|_1 \sqrt{\log N/T}$ as shown in (16), so it is not as efficient as our SPCA estimator that exploits the factor structure. Nonetheless, under a much stronger sparsity assumption that $\left\|b\right\|_0^2 \log N/T \to 0$, the Lasso estimator can achieve a comparable rate to that of the SPCA. This stronger notion of sparsity effectively says that the set of true factors must be part of the test assets. In contrast, our SPCA estimator allows for idiosyncratic components in any of the test assets, which is a more acceptable assumption in practice.

Just like for the risk premia estimator based on rpPCA, we can adapt any SDF estimator to obtain an estimator of risk premia, because $-\mathrm{Cov}(m_t, g_t) = \eta\gamma$. Naturally we have a Lasso-based risk premia estimator:[16]

$$\widehat{\gamma}_g^{Lasso} = -\frac{1}{T}\sum_{t=1}^{T} \widehat{m}_t^{Lasso} \times (g_t - \bar{g}).$$

Furthermore, the consistency of the SDF estimator translates to the consistency of the resulting risk premia estimator.[17] Deriving a valid inference procedure is possible for Lasso, if we employ an

[16] The SDF-induced Ridge estimator is numerically equivalent to (5), so we do not mention it again.

[17] By Assumption A.12(1), Cauchy-Schwartz and triangle inequalities, we have

$$\left\|\widehat{\gamma}_g^{Lasso} - \gamma_g\right\|_{\mathrm{MAX}} \lesssim_P \sqrt{\frac{1}{T}\sum_{t=1}^{T} |\widehat{m}_t^{Lasso} - \widetilde{m}_t|^2} + \sqrt{\frac{\log N}{T}}.$$

additional de-biasing step, see, Feng et al. (2020), which is beyond the scope of the current paper.

As a side note, the SPCA estimator given by equation (11) can also be rewritten in the form of (14), so that it can yield an estimate of b in the definition of SDF given by equation (3). The reason is that $\widehat{v}_t$ is in fact a linear combination of r_t. Given that b is invariant to rotations of factors, we can use any rotation of $\widehat{v}_t$ to reconstruct an estimate of b. We can exploit this invariance property to construct a convenient estimator $\widehat{b}$. In fact, in S1.b of Algorithm 5, we can construct an $N \times p$ matrix B such that the kth column of B is defined as: $B_{[I_k],k} = \varsigma_{(k)}$ and $B_{[I_k^c],k} = 0$, where $\varsigma_{(k)}$ is the left singular vector of $\left(\bar{R}_{(k)}\right)_{[I_k]}$. It turns out the SPCA estimates of $\widehat{V}$ can be written as a rotation of $B^\intercal \bar{R}$, so to estimate $\widehat{b}$ we can use $B^\intercal \bar{R}$ as factors, denoted by, $\widetilde{V}$, whose risk premia and covariance are denoted by $\widetilde{\gamma}$ and $\widetilde{\Sigma}$. Indeed, since the SDF is $m_t = 1 - \widetilde{\gamma}^\intercal (\widehat{\Sigma}_v)^{-1} \widehat{v}_t = 1 - \widetilde{\gamma}^\intercal (\widetilde{\Sigma}_v)^{-1} \widetilde{v}_t = 1 - \widetilde{\gamma}^\intercal (\widetilde{\Sigma}_v)^{-1} B^\intercal (r_t - \bar{r})$, it follows that the SPCA-based estimate of b is given by

$$\widehat{b} = B(\widetilde{\Sigma}_v)^{-1}\widetilde{\gamma} = TB\left(B^\intercal \bar{R}\bar{R}^\intercal B^\intercal\right)^{-1} B^\intercal \bar{r}.$$

Similarly, we can construct estimates of b using PCA, PLS, and rpPCA. With $\widehat{b}$ it is convenient to build out-of-sample SDF (optimal portfolios).

3 Simulations

In this section, we study the finite sample performance of our SPCA procedure using Monte Carlo simulations. We also implement a number of alternative 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).[18] 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 qN (or equivalently q) by cross-validation using the time series R^2 of the hedging portfolio for g_t built by SPCA as the criterion.[19] Recall that any estimator of risk premia for a nontradable factor explicitly or implicitly builds a hedging portfolio exposed to g_t and not exposed to the other factors. We therefore use as a criterion for the choice of the tuning parameters the ability of this portfolio to hedge g_t in the validation sample. In order to produce a conservative comparison, except for SPCA, all the

[18] 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.

[19] In finite samples, we find it more effective and more convenient to tune p and q than q and c. This is because the former are direct input to SPCA, which only take values from integers, so that multiple choices of the latter lead to the same integer values of the former.

remaining methods use optimal (even if infeasible) tuning parameters. Specifically, for PCA, PLS and rpPCA, we make use of the true number of factors, $p = 4$, even though it is difficult to obtain a consistent estimator of p in the regime of weak factors. The tuning parameter μ of Ridge estimator is determined via maximum likelihood estimation, with perfect knowledge of Σ_r and $\mathrm{E}(r)$. The second tuning parameter of rpPCA is selected by maximizing the theoretical Sharpe ratio of the estimated SDF, using, again, perfect knowledge of Σ_r and $\mathrm{E}(r)$. 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 DGP 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 (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 0.25. The realizations of u_t are generated independently from a Gaussian distribution with mean 0 and standard deviation σ_u calibrated such that the time-series R^2 ranges from 50-90%. 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 $\mathcal{N}(0,0.1^2)$. Based on our calibration, we choose $a = 0.5$, so that 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), so that $\eta = \mathbb{I}_4$, and measurement error is absent.

In scenario b), we choose $a = 0.05$ so that V is weak in that for almost all test assets their factor loadings to V are tiny. 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 $a = 0.05$. In this case, the loading matrices of V and HML are very similar, which (almost) leads to a rank deficient factor loading matrix due to highly 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 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.[20] 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 and omitted factors as well as measurement error. For this case, we report in Table 1 the bias and the RMSE (root-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 1.

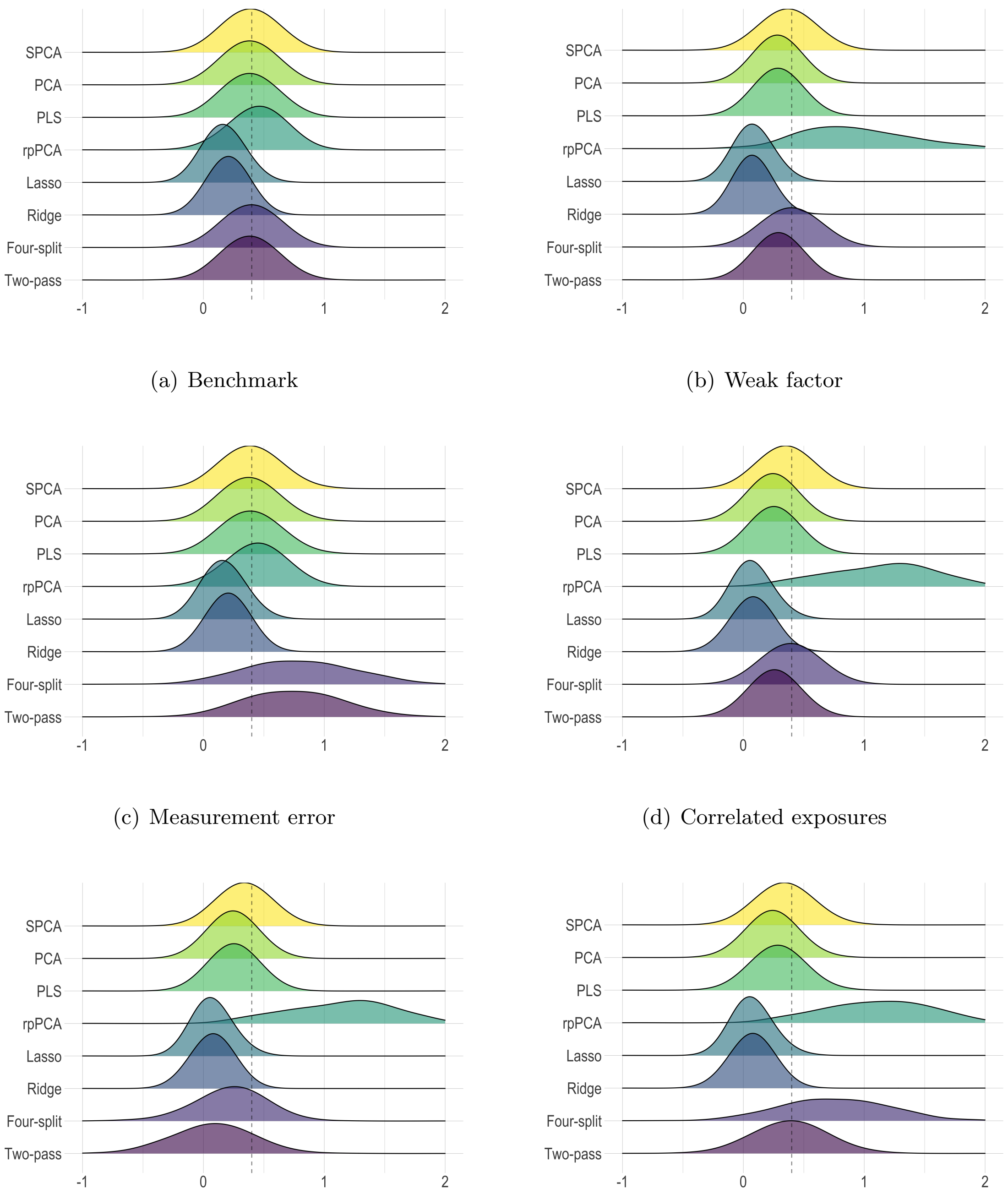
We then investigate the finite sample performance of the inference result developed in Theorem 2. Figure 2 plots histograms of the standardized risk premia estimators using the estimated asymptotic standard errors for SPCA and PCA, respectively, using the DGP in scenario f) as an example. The histograms of PCA deviate from the standard normal distribution for the two highly correlated factors, V and HML. In contrast, the histograms corresponding to the SPCA match the normal distribution well, which verifies our central limit results.

Finally, 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 the MSE of the SDF estimators where the true SDF is defined by equation (3). The estimated number of factors from our SPCA approach is also reported. We also report in Table 3 the out-of-sample Sharpe ratios of different methods, 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. We find that in terms of the RMSE, SPCA outperforms all other methods, and that rpPCA performs the worst. That said, rpPCA performs the best in terms of the out-of-sample Sharpe ratio, followed by the SPCA. Last but not least, SPCA produces a decent estimator of p when T is large.

4 Empirical Analysis

In this section we perform different empirical exercises to illustrate the use of SPCA. First, we apply it to estimate the risk premia of a number of tradable and nontradable factors proposed in the literature, and we evaluate its out-of-sample performance. Second, we evaluate the robustness of SPCA as we change the universe of test assets to make the factors stronger or weaker. Third, we

[20] The standard two-pass estimator appears to have a small bias in scenario f), but this happens to be true only for V since all sources of bias happen to balance out, as we show in Table 1.



(a) Benchmark

(b) Weak factor

(c) Measurement error

(d) Correlated exposures

(e) Weak + omitted factor

(f) Correlated exposures + measurement error

Figure 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 $N = 2,000$ and $T = 120$. The number of Monte Carlo repetitions is 1,000.

			SPCA		PCA		rpPCA		PLS	
T	Param	True	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
120	RmRf	53.7	0.9	39.2	1.0	38.9	0.8	42.4	1.0	39.1
	SMB	21.7	0.2	29.0	0.3	28.5	-0.1	32.7	0.4	28.7
	HML	25.4	-4.8	26.6	-15.3	28.1	76.4	93.5	-10.9	26.8
	V	40.0	-5.5	20.9	-15.7	23.1	74.5	89.7	-11.3	21.4
180	RmRf	53.7	0.8	33.9	0.8	33.8	0.8	34.9	0.9	33.8
	SMB	21.7	0.4	23.1	0.4	22.7	0.6	24.9	0.3	22.9
	HML	25.4	-3.7	21.5	-11.9	22.9	49.0	62.4	-7.4	21.7
	V	40.0	-3.6	17.0	-11.7	18.6	48.6	60.5	-7.1	17.1
240	RmRf	53.7	0.7	29.6	0.8	29.5	0.7	30.0	0.8	29.6
	SMB	21.7	0.5	20.2	0.5	20.0	0.3	21.4	0.5	20.1
	HML	25.4	-2.8	18.3	-9.4	19.3	35.5	45.7	-5.0	18.4
	V	40.0	-3.6	14.3	-10.1	16.0	33.9	42.7	-5.8	14.5
			Lasso		Ridge		Four-split		Two-pass	
T	Param	True	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
120	RmRf	53.7	-16.3	28.8	-2.9	35.2	16.0	53.8	14.7	51.2
	SMB	21.7	-8.1	15.3	-3.1	20.4	7.1	48.1	7.1	44.7
	HML	25.4	-28.8	31.2	-31.1	35.9	19.2	50.2	-12.1	39.4
	V	40.0	-32.8	34.5	-32.7	34.8	36.3	57.8	-1.3	29.4
180	RmRf	53.7	-11.7	26.9	-1.5	31.3	16.9	46.7	15.0	44.9
	SMB	21.7	-6.4	14.2	-2.1	17.8	6.4	37.7	6.9	35.9
	HML	25.4	-29.7	31.6	-28.3	32.6	20.1	42.5	-6.0	31.8
	V	40.0	-31.6	33.0	-28.8	30.9	39.1	53.2	7.6	26.4
240	RmRf	53.7	-6.7	24.7	-0.6	28.1	16.4	41.7	14.9	39.9
	SMB	21.7	-3.7	14.8	-1.1	16.9	7.3	33.6	7.3	31.9
	HML	25.4	-24.9	28.0	-25.3	29.3	21.4	38.4	-0.8	27.0
	V	40.0	-26.5	28.6	-26.2	28.1	38.8	49.1	12.4	25.6

Table 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 $N = 2,000$ while varying $T = 120, 180$, and 240 in this experiment. All values are in basis points.

	SPCA		PCA	rpPCA	PLS	Lasso	Ridge
T	$\widehat{p}$	MSE	MSE	MSE	MSE	MSE	MSE
120	4.080	0.036	0.037	0.387	0.040	0.044	0.050
	(0.339)	(0.026)	(0.025)	(0.505)	(0.025)	(0.012)	(0.018)
180	4.000	0.024	0.025	0.163	0.027	0.041	0.041
	(0.000)	(0.017)	(0.017)	(0.209)	(0.017)	(0.011)	(0.015)
240	4.000	0.018	0.019	0.085	0.020	0.035	0.035
	(0.000)	(0.013)	(0.013)	(0.088)	(0.013)	(0.011)	(0.013)

Table 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 errors 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 $T = 120$, 180, and 240, where $N = 2,000$ is fixed.

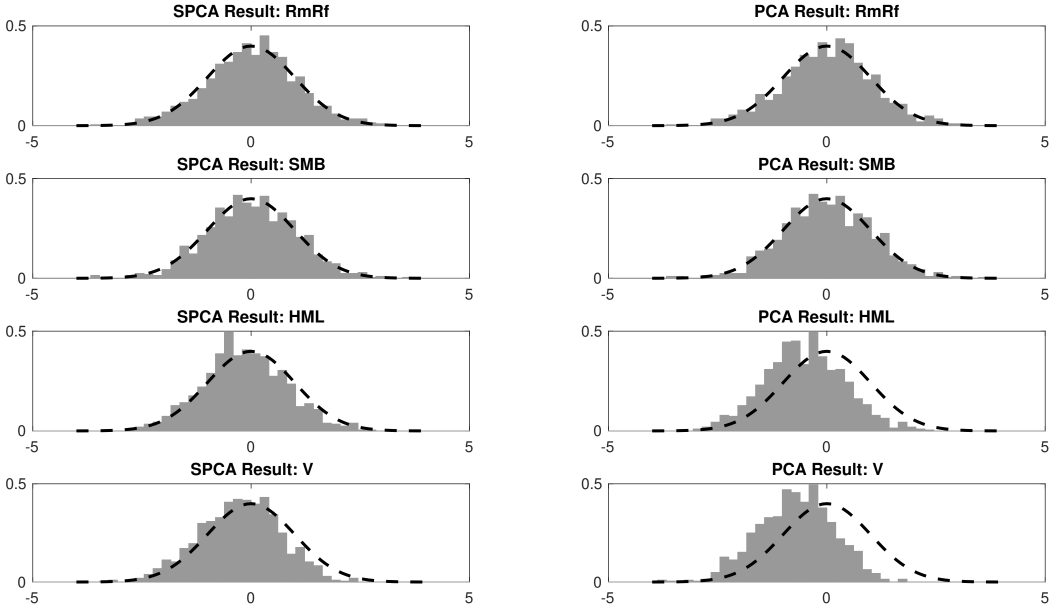


Figure 2: Histogram of the Standardized Estimates in Simulations

Note: The left panels provide the histograms of the standardized SPCA estimates as in Algorithm 5 with asymptotic standard errors given by Theorem 2, whereas the right panels provide those of the standardized PCA-based risk premia estimates as in Algorithm 1. We simulate the model in scenario f) with $N = 2,000$ and $T = 240$. The number of Monte Carlo repetitions is 1,000.

T	SPCA	PCA	rpPCA	PLS	Lasso	Ridge	Theoretical Value
120	0.186	0.159	0.214	0.155	0.133	0.126	0.245
	(0.042)	(0.045)	(0.025)	(0.045)	(0.035)	(0.044)	
180	0.204	0.186	0.224	0.183	0.144	0.148	0.245
	(0.032)	(0.037)	(0.017)	(0.037)	(0.035)	(0.041)	
240	0.214	0.202	0.229	0.201	0.160	0.164	0.245
	(0.025)	(0.029)	(0.014)	(0.029)	(0.033)	(0.035)	

Table 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 $T = 120$, 180, and 240, where $N = 2,000$ is fixed.

propose two additional examples of applications of the SPCA methodology: the estimation of fund alpha, and the de-noising of observable factors (similar in spirit to Daniel et al. (2020)).

4.1 Estimation of Risk Premia using SPCA

4.1.1 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.

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, intangibles, frictions. The two datasets of Hou et al. (2020) and Chen and Zimmermann (2020) 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 (e.g., see section 4.1.6).

We study the risk premium of both tradable and nontradable factors, 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.[21]

[21] 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).

4.1.2 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.

For ease of presentation, we choose to select only one tuning parameter, q, for each plausible choice of p in our analysis. This approach reduces the number of tuning parameters to only one, and also conveniently serves as a robustness check.

To determine reasonable candidates of p, we examine the factor structure of the panel of test asset returns. Figure 3 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, though weaker. 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.

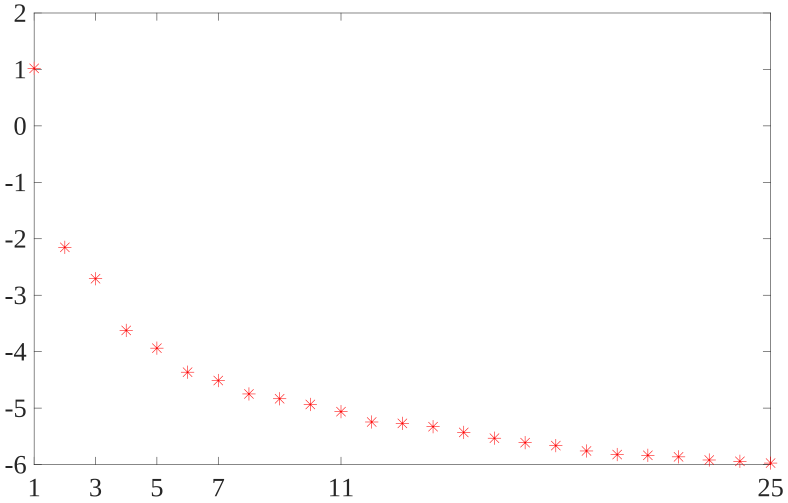


Figure 3: Logarithm of the First 25 Eigenvalues in the Chen-Zimmerman data

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.

To choose the tuning parameter q, we adopt the same 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. More concretely, recall that all estimators of risk premia (e.g., the standard Fama-MacBeth estimator, the PCA-based estimator of Giglio and Xiu (2021), and SPCA) recover the risk premium of a factor g_t by, implicitly or explicitly, building a tradable portfolio that isolates exposure to g_t. We therefore compute the weights of the hedging portfolio built by SPCA using the training data only, and calculate the mean-squared-error of hedging g_t over the validation period using that portfolio (effectively, the out-of-sample R^2). We apply this criterion to pick q using cross-validation (CV) within the training sample.

Our empirical analysis proceeds as follows. We first choose the number of factors p in the model, based on the scree plot. Then, working exclusively in the training sample, we run 3-fold cross-validation 100 times. In each cross-validation run, the tuning parameter q is chosen to maximize the R^2 of the hedging portfolio described above (requiring a minimum of 100 assets selected). Our choice for q is the median across the 100 cross-validation runs, and the risk-premium estimate is the one obtained under that choice of q; we also compute the weights of the hedging portfolio for the factor. All the analysis described so far uses only data from the training period. Next, we evaluate the out-of-sample performance of the estimator. Using the test asset returns from the evaluation period, together with the portfolio weights estimated previously, we compute the return of the hedging portfolio in the evaluation sample, and we calculate the fraction of the variance of g_t hedged by that portfolio, i.e., R^2. This calculation does not re-estimate any parameter or coefficient in the evaluation data, and is thus a fully out-of-sample R^2.

4.1.3 Results: Estimation of Risk Premia and Out-of-sample Evaluation

We report the main empirical results in Table 4 and Figures 4 and 5. Each row of Table 4 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 samples, 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 $d = 1$). As discussed above, in this case, assets are sorted by the maximum of the correlation with any of the factors in g_t for the purpose of the selection step. 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 to satisfy the more stringent assumptions for the CLT of Section 2.6 (because for the CLT to work, we need to assume that g_t has exposure to the entire SDF, while this is not required for the consistency of the estimates for an individual factor).

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 out-of-sample performance of the estimator, we can examine the heatmap in Figure 4, 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 SPCA, that is, the ability of the SPCA hedging portfolio to hedge g_t in the evaluation sample. The heatmap gives a complete description of the out-of-sample properties of SPCA as a function of the two parameters p and q. Panel (a) shows that for all combinations of p and q (that is, throughout the heatmap), 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 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 cross-validation (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 perform 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 SPCA estimator and of the tuning parameter selection procedure.

Consider now another tradable factor, CMA, in the 5th row of Table 4. Like for the market, the estimated risk premium for CMA is not statistically significantly different from the average 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 our latent factor model is able to capture the majority of the variation on CMA out of sample.[22]

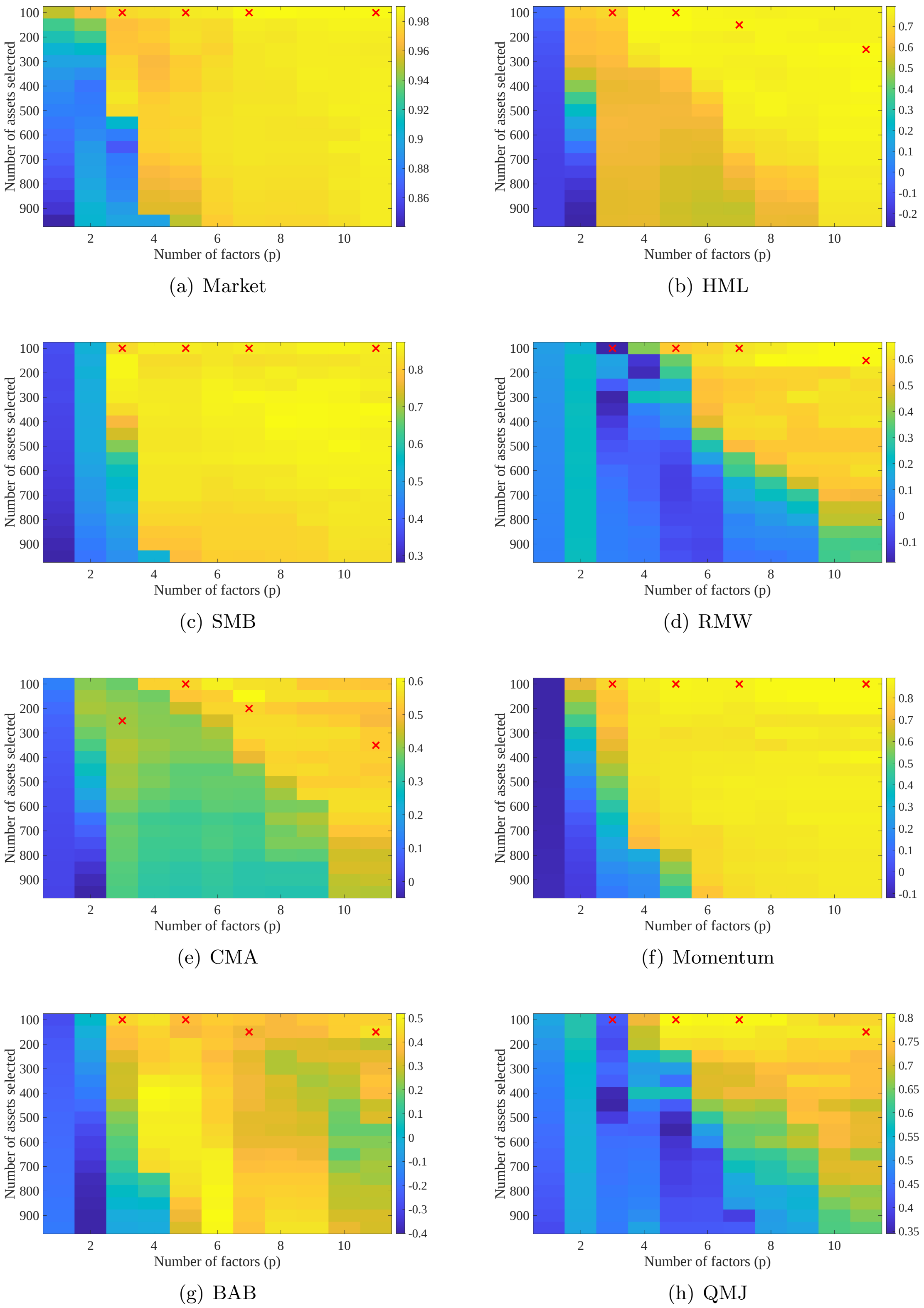
The heatmap of the out-of-sample R^2 for this factor is panel (e) of Figure 4. The figure shows that for the case of CMA, different combinations of p and q yield very different hedging performance out-of-sample, with R^2s ranging from above 50% to below 0. It is especially important then that our tuning parameter selection procedure yields good results out-of-sample. The red marks in the figure

[22] 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 is trying 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. We come back to this point in section 4.2.2.

Table 4: 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

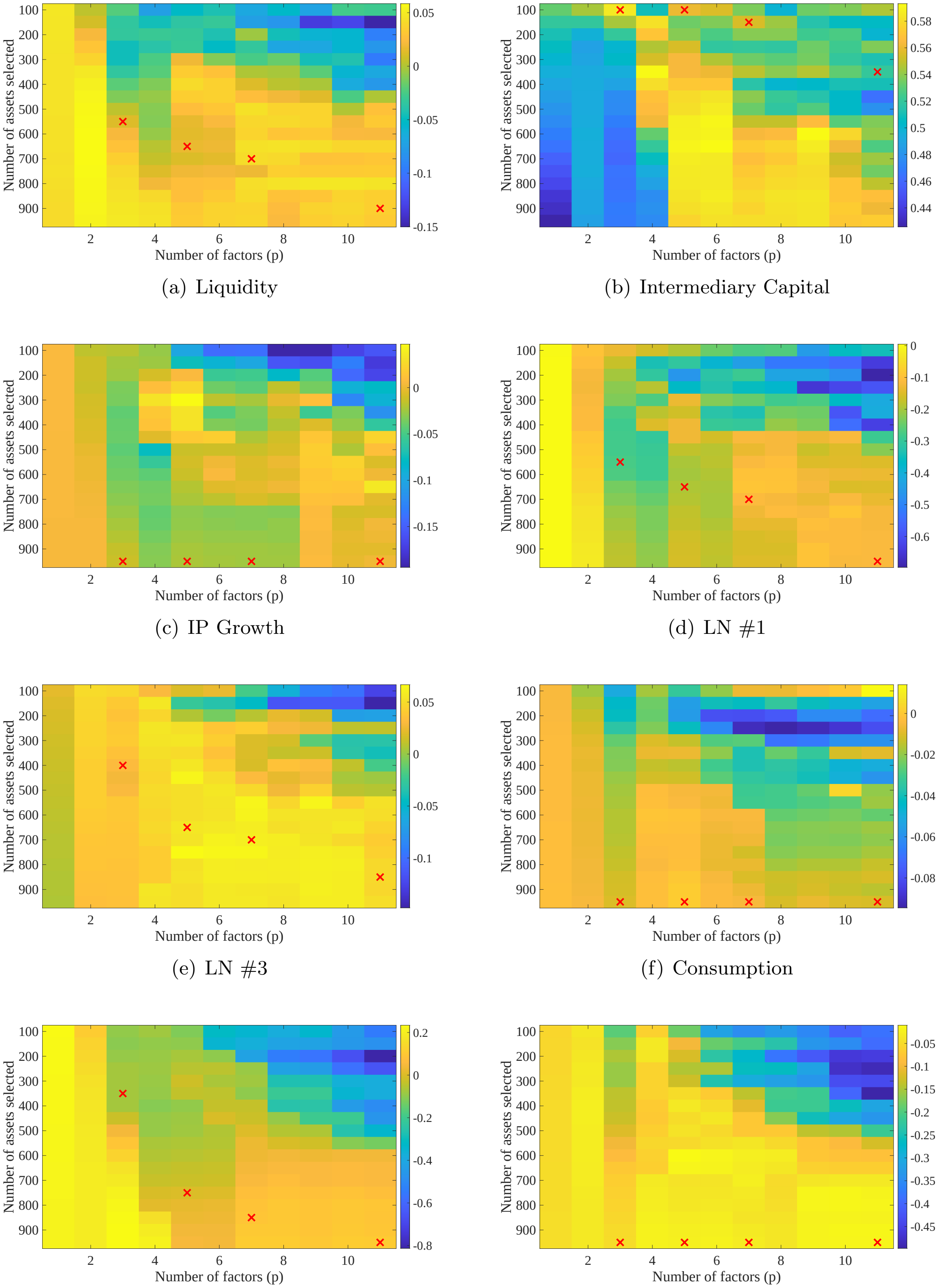
(f) Momentum

(g) BAB

(h) QMJ

Figure 4: 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 5: Out-of-sample R^2 Heatmaps, Nontradable Factors

Note: Same as figure 4, but for a subset of nontradable factors.

show that this is indeed the case, especially for $p = 5$ and above.

In addition to showing the performance of SPCA, 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; but in that case, SPCA coincides with PCA. Looking across the various panels of Figure 4, it is clear that while for some factors (like the market) similar R^2 can be obtained (for appropriate choices of p) 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 obtainable by PCA, for any choice of p (graphically: the area with the highest R^2s is concentrated in the upper part of the heatmap, where $q < 1$). This shows that the case of weak factors studied in this paper is relevant in empirical applications.

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: SPCA focuses on the test assets that are most informative about g_t, and therefore can zoom in quickly on the most relevant latent factors. This robustness to the number of factors is another advantage of SPCA that is relevant in practical applications.

For nontradable factors, we cannot compare the risk premium estimate from SPCA with the average excess return; the out-of-sample R^2 therefore plays an even more important role in evaluating the 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 4 and panel (a) of Figure 5. 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 5 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 emphasizes again the good performance of SPCA.

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 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.

4.1.4 Asset Selection

To better understand how SPCA estimates the risk premium, we can study which assets are selected when extracting the latent factors. Table 5 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 of the table. To extract the first latent factor, SPCA selects the assets with the highest correlation with the momentum factor. As the table shows, the highest correlation is achieved by IntMom09 (an intermediate momentum portfolio). The correlation is 0.44. 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 residuals of the factor and the portfolio are then *more* correlated than the original factors and portfolios after the influence of the first factor is eliminated. In any case, momentum portfolios appear again at the second iteration, and, in part, at the third iteration.

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 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 robustness of SPCA to the number of factors used, as we noted above: given that SPCA zooms in on the most informative assets, it can build a good hedging portfolio (and therefore a good estimate of the risk premium) even with a small number of factors.

Table 5: 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).

4.1.5 SPCA and the Universe of Test Assets

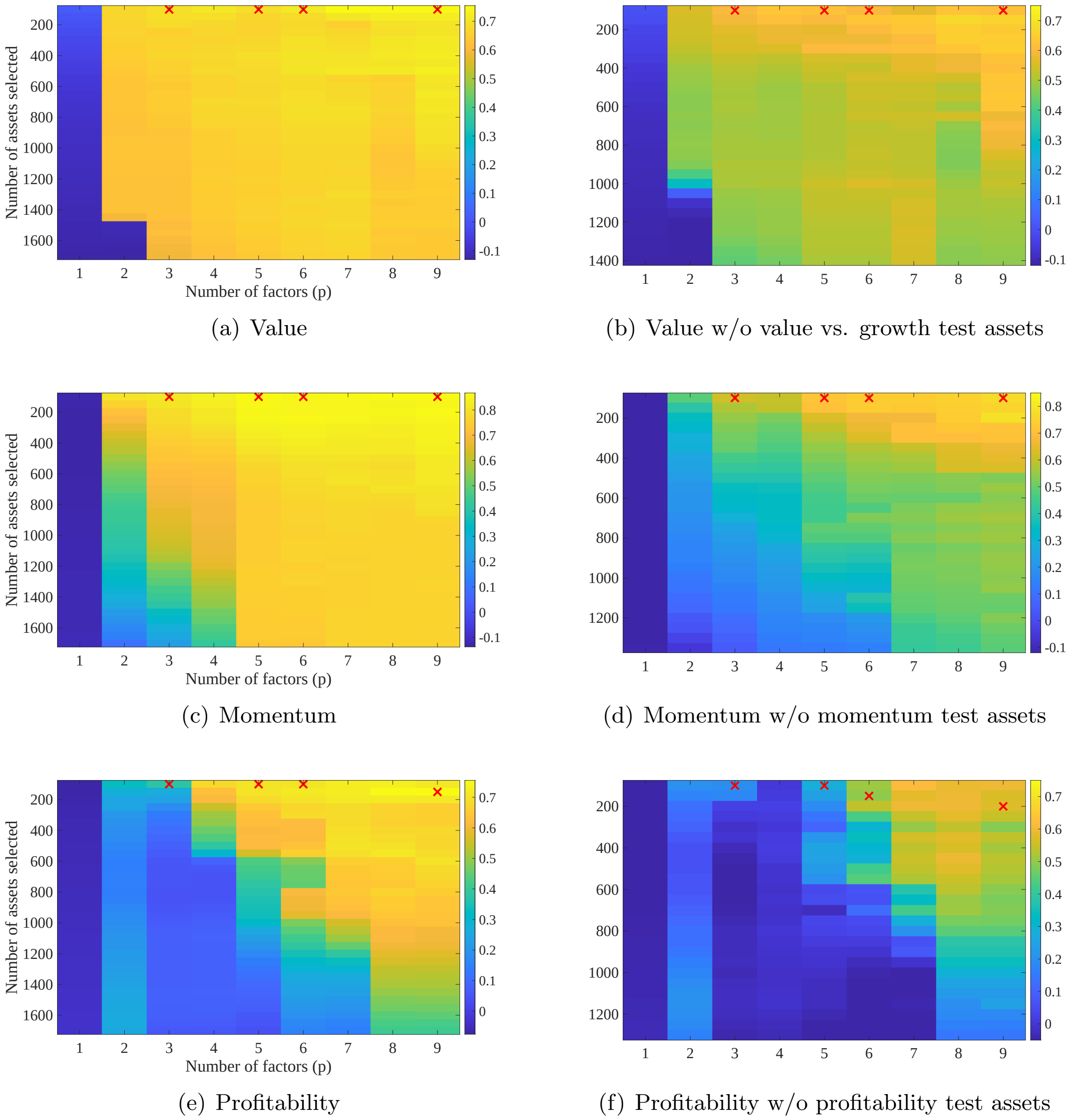
The fact that SPCA estimates the latent factors using the most informative assets also makes it particularly robust to the universe of test assets used in the estimation. We explore this here in detail by considering three factors, value, momentum, and profitability, for which we can easily identify test assets that are informative about them. Specifically, we consider (for this section only) the dataset from Hou et al. (2020), which, as discussed in section 4.1.1, 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 premium if momentum and profitability test assets, respectively, are removed? When the corresponding sorted portfolios are removed, the factors naturally become weaker. However, we should expect SPCA to still perform well, at least 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.

Figure 6 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 performance of PCA deteriorates significantly when the most informative assets are removed. Consider for example the case $p = 9$. For value, the PCA estimator'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 estimator deteriorates little when the relative sorts are removed and the factor is made weaker, whereas the deterioration in performance is much larger for PCA.

To conclude, these empirical results mirror the simulations in section 3, that show SPCA performing well even when the factor of interest is weak in the universe of test assets considered. This is important in practical applications: given a certain factor g_t, we do not know ex ante if many or just a few assets are exposed to that factor. SPCA builds a good hedging portfolio and provides a consistent estimate of the risk premium in either case.

4.1.6 Robustness

We conclude by reporting in Table 6 a version of Table 4 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



(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 6: 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).

test assets.

4.2 Other Applications

In this section we study two additional applications of SPCA: the estimation of the alpha of a fund using latent factors to capture risk exposures, and using it to de-noise factor returns.

Table 6: 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 4, but using the characteristic-sorted portfolios from Hou et al. (2020) instead of those from Chen and Zimmermann (2020).

4.2.1 Estimating Buffett's alpha using Latent Factors

When evaluating the performance of money managers, a crucial step is the choice of the benchmark against which the alpha of the fund is calculated. The benchmark model is often a standard tradable factor model like the Fama-French 3 or 5 factor model. How this benchmark should be selected is not entirely clear a priori; different benchmarks lead to different conclusions about the ability of managers to generate alpha. The asset pricing literature has proposed two approaches to address this issue that do not require arbitrarily choosing a benchmark: one approach is to use machine learning methods to select an optimal, parsimonious benchmark, chosen from a large set of candidate benchmarks in the "factor zoo" (Feng et al. (2020)). Another approach, proposed by Connor and Korajczyk (1986), uses latent factors to extract the relevant benchmarks, thus avoiding taking a stand on the identity of the factors (see also Giglio et al. (2020)).

SPCA offers a natural way to expand the second approach: by using a fund return as g_t, SPCA allows us to extract from the panel of returns all and only those factors that are informative about the fund's risk exposures (and therefore compute the alpha after accounting for all risk exposures, including those to weak factors). In this section, we illustrate this possibility by applying SPCA to understand the alpha of Berkshire Hathaway, similar to Frazzini et al. (2013).

One of the headline results in Frazzini et al. (2013) is that Berhshire Hathaway's returns display large alpha (13.4% annualized, statistically significant) when the fund's return is benchmarked to the market factor alone. However, the alpha becomes much smaller (5.7%) and statistically insignificant when the benchmark model also includes SMB, HML, Momentum, BAB, and QMJ. Note that all the factor exposures together capture 29% of the time-series variation of the fund return.

We apply SPCA using the same data as in our main analysis (Chen-Zimmerman data plus 49 industry portfolios). The first factor in g_t will be Buffett's return; we add to the vector of factors g_t all the other factors in our dataset, in order to get correct asymptotic standard errors. All the result here are in-sample, and we select the tuning parameter q by 3-fold cross-validation in the full sample.

The results are remarkably similar to those of Frazzini et al. (2013), with the main difference that we do not need to specify the identity of the factors. Specifically, when only one latent factor is extracted ($p = 1$), we obtain an alpha of 13%, with a t-stat of 4. The one-factor benchmark only explains 19% of the variation in the fund's return. As more latent factors are included by SPCA, the R^2 increases and the alpha decreases. Once $p = 6$, the results effectively coincide with the ones of Frazzini et al. (2013): the time-series R^2 reaches 30% and the alpha drops to 6.1%, statistically insignificant (t-stat of 1.82).

The assets selected by SPCA give us some insights on what are the main risk exposures that determine Berkshire Hathaway's risk premium. Specifically, among the portfolios most correlated with it, a large role is played by idiosyncratic volatility sorts, followed by a variety of portfolios related to profitability and leverage.

The results confirm the main insight of Frazzini et al. (2013): that Berkshire Hathaway's return can be attributed in large part to exposure to priced factors; we however obtain this result without having to take a stand on the entire benchmark model, and relying only on SPCA to extract the latent factors. We can however gain some insights on the drivers of the risk premium by studying the assets selected by SPCA to build the hedging portfolio.

4.2.2 De-noising Factor Models

In a recent paper, Daniel et al. (2020) argue that the way standard factors are constructed based on characteristics sorts could be suboptimal as the portfolios used as factors may be contaminated by exposure to unpriced factors. They propose a procedure to remove the unpriced risk from observed factors, and produce a version of the Fama-French 5-factor model that achieves a higher Sharpe ratio and has better pricing ability for a cross-section of specifically-sorted test portfolios.

As discussed in the previous sections, SPCA constructs a hedging portfolio for any (tradable or nontradable) factor g_t that is built to capture the fundamental factors in the panel of test assets. By eliminating "measurement error" from g_t, SPCA effectively helps de-noise the factor from idiosyncratic (and therefore plausibly unpriced) risk. Similar in spirit to Daniel et al. (2020), we can use SPCA to strip out measurement error and build a de-noised version of the 5 factors in the FF5 model.

A natural exercise is then to compare the pricing ability for the panel of test assets of the original Fama-French 5-factor model, the de-noised model of Daniel et al. (2020), and the Fama-French 5 factors de-noised via SPCA. Table 7 reports the average absolute alphas of these models (for SPCA, we consider different values of p to de-noise FF5 factors). The left set of columns reports the results using the Chen and Zimmermann (2020) data (CZ), the right column using the Hou et al. (2020) data (HXZ). For each set of results, we consider two versions of each model, restricting the zero-beta rate to be equal to the Tbill rate (left column), or with a free zero-beta rate (right column).

The table shows that for both the version with and without the zero-beta rate, SPCA produces an improvement over both the Fama-French 5-factor model and the Daniel et al. (2020) model, suggesting that removing the measurement error from the factor helps isolate the priced component of the factors.[23] The magnitude of the improvement of SPCA is as large as that by the Daniel et al. (2020) model.

5 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

[23] The assumptions on the zero-beta rate have a first-order effect only on the Daniel et al. (2020) results.

Table 7: Average absolute alphas across models

	CZ		HXZ	
	no zero-beta	w/ zero-beta	no zero-beta	w/ zero-beta
FF5	19	19	12	12
Daniel et al. (2020)	39	17	35	13
SPCA (5 factors)	20	20	10	10
SPCA (7 factors)	17	17	11	11
SPCA (11 factors)	17	17	11	11

Note: The table reports the average absolute alpha, in basis points per month, among the Chen and Zimmermann (2020) test assets (left) and the Hou et al. (2020) test assets (right), for different models: the Fama-French 5-factor model, the model of Daniel et al. (2020), and different versions of SPCA with different p. We consider two versions of each model, restricting the zero-beta rate to be equal to the Tbill rate (left), or with a free zero-beta rate (right).

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 to estimate risk premia 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.

The paper confirms the performance of SPCA in a variety of simulations, and explores different empirical applications of SPCA to risk premia estimation, fund performance evaluation, and factor de-noising. Overall, the simulations and empirical analysis highlight a few important features of SPCA, that are particularly relevant for empirical applications: its robustness to the number of factors used, to the universe of test assets employed in the estimation, and, most importantly, to the strength of the factors in the data.

While the road to a full understanding of risk and risk premia in financial markets is still long, we believe that addressing systematically the issue of 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 the standard cross-sections.

References

Ahn, D.-H., J. Conrad, and R. F. Dittmar (2009). Basis assets. *The Review of Financial Studies 22*(12), 5133–5174.

Anatolyev, S. and A. Mikusheva (2021). Factor models with many assets: strong factors, weak factors, and the two-pass procedure. *Journal of Econometrics, forthcoming*.

Ang, A., R. Hodrick, Y. Xing, and X. Zhang (2006). The cross-section of volatility and expected returns. *Journal of Finance 61*, 259–299.

Asness, C. S., A. Frazzini, and L. H. Pedersen (2013). Quality Minus Junk. Technical report, AQR.

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

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

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

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

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

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

Baker, M. and J. Wurgler (2006). Investor sentiment and the cross-section of stock returns. *The journal of Finance 61*(4), 1645–1680.

Bryzgalova, S. (2015). Spurious Factors in Linear Asset Pricing Models. Technical report, Stanford University.

Bryzgalova, S., J. Huang, and C. Julliard (2019). Bayesian solutions for the factor zoo: We just ran two quadrillion models. *Available at SSRN 3481736*.

Bryzgalova, S., M. Pelger, and J. Zhu (2020). Forest through the trees: Building cross-sections of asset returns. Technical report, London School of Business and Stanford University.

Chen, A. Y. and T. Zimmermann (2020). Open source cross-sectional asset pricing. *Available at SSRN*.

Connor, G. and R. A. Korajczyk (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics 15*(3), 373–394.

Daniel, K., L. Mota, S. Rottke, and T. Santos (2020). The cross-section of risk and returns. *The Review of Financial Studies 33*(5), 1927–1979.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds.

Journal of Financial Economics 33(1), 3–56.

Feng, G., S. Giglio, and D. Xiu (2020). Taming the factor zoo: A test of new factors. *Journal of Finance 75*(3), 1327–1370.

Frazzini, A., D. Kabiller, and L. H. Pedersen (2013). Buffett's alpha. Technical report, National Bureau of Economic Research.

Frazzini, A. and L. H. Pedersen (2014). Betting against beta. *Journal of Financial Economics 111*(1), 1–25.

Freyaldenhoven, S. (2019). A generalized factor model with local factors.

Gagliardini, P., E. Ossola, and O. Scaillet (2016). Time-varying risk premium in large cross-sectional equity datasets. *Econometrica 84*(3), 985–1046.

Giglio, S., Y. Liao, and D. Xiu (2020). Thousands of alpha tests. *Chicago Booth Research Paper* (18-09), 2018–16.

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

Harvey, C. R., Y. Liu, and H. Zhu (2016). ...and the Cross-Section of Expected Returns. *The Review of Financial Studies 29*(1), 5–68.

He, Z., B. Kelly, and A. Manela (2017). Intermediary asset pricing: New evidence from many asset classes. *Journal of Financial Economics 126*(1), 1–35.

Hou, K., C. Xue, and L. Zhang (2020). Replicating anomalies. *Review of Financial Studies 33*(5), 2019–2133.

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

Huang, D., F. Jiang, J. Tu, and G. Zhou (2015). Investor sentiment aligned: A powerful predictor of stock returns. *The Review of Financial Studies 28*(3), 791–837.

Jagannathan, R. and Z. Wang (1998). An asymptotic theory for estimating beta-pricing models using cross-sectional regression. *The Journal of Finance 53*(4), 1285–1309.

Jurado, K., S. C. Ludvigson, and S. Ng (2015). Measuring uncertainty. *The American Economic Review 105*(3), 1177–1216.

Kan, R. and C. Zhang (1999). Two-Pass Tests of Asset Pricing Models with Useless Factors. *The Journal of Finance 54*(1), 203–235.

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

Kelly, B., S. Pruitt, and Y. Su (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics 134*(3), 501–524.

Kim, S., R. A. Korajczyk, and A. Neuhierl (2020). Arbitrage portfolios. *Review of Financial Studies,* *Forthcoming*.

Kleibergen, F. (2009). Tests of risk premia in linear factor models. *Journal of Econometrics 149*(2), 149–173.

Kozak, S., S. Nagel, and S. Santosh (2020). Shrinking the cross-section. *Journal of Financial Economics 135*(2), 271–292.

Lettau, M. and M. Pelger (2020). Estimating latent asset-pricing factors. *Journal of Econometrics 218*, 1–31.

Ludvigson, S. C. and S. Ng (2010). A factor analysis of bond risk premia. In A. Ulah and D. E. A. Giles (Eds.), *Handbook of empirical economics and finance*, Volume 1, Chapter 12, pp. 313–372. Chapman and Hall, Boca Raton, FL.

Pástor, L. and R. F. Stambaugh (2003). Liquidity risk and expected stock returns. *Journal of Political Economy 111*(3), 642–685.

Pesaran, M. H. and R. Smith (2019). The role of factor strength and pricing errors for estimation and inference in asset pricing models.

Schorfheide, F., D. Song, and A. Yaron (2018). Identifying long-run risks: A bayesian mixed-frequency approach. *Econometrica 86*(2), 617–654.

Supplement to

Test Assets and Weak Factors

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

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

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

This Version: June 25, 2021

Abstract

This supplementary appendix provides model assumptions and mathematical proofs.

[*] Address: 165 Whitney Avenue, New Haven, CT 06520, USA. E-mail address: stefano.giglio@yale.edu.
[†] Address: 5807 S Woodlawn Avenue, Chicago, IL 60637, USA. E-mail address: dacheng.xiu@chicagobooth.edu.
[‡] Address: 5807 S Woodlawn Avenue, Chicago, IL 60637, USA. Email: dkzhang@chicagobooth.edu.

A 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 DGP of returns and factor proxies.

Assumption A.1. *The factor innovation V satisfies:*

$$\|\bar{v}\| \lesssim_p T^{-1/2}, \quad \left\|T^{-1}VV^\intercal - \Sigma_v\right\| \lesssim_p T^{-1/2}, \quad \|V\|_{\mathrm{MAX}} \lesssim_p \sqrt{\log T},$$

where $\Sigma_v \in \mathbb{R}^{p\times p}$ is a positive-definite matrix with $\lambda_p\left(\Sigma_v\right) \gtrsim 1$ and $\lambda_1\left(\Sigma_v\right) \lesssim 1$.

Assumption A.2. *The residual innovation Z satisfies:*

$$\|\bar{z}\| \lesssim_p T^{-1/2}, \quad \left\|T^{-1}ZZ^\intercal - \Sigma_z\right\| \lesssim_p T^{-1/2}, \quad \|Z\|_{\mathrm{MAX}} \lesssim_p \sqrt{\log T}.$$

where $\Sigma_z \in \mathbb{R}^{d\times d}$ is a positive-definite matrix with $\lambda_d\left(\Sigma_z\right) \gtrsim 1$ and $\lambda_1\left(\Sigma_z\right) \lesssim 1$. In addition,

$$\|ZV^\intercal\| \lesssim_p T^{1/2}.$$

Assumptions A.1 and A.2 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 A.3. *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 A.3 implies that there exists a subset of test assets, within which all latent factors are strong. Because the number of factors is finite, requiring *all* factors to be strong within a *common* index set I_0 is equivalent to requiring each factor to be strong in its own index set. One direction of the equivalence is trivial. To prove the other direction, suppose that for factor i, there exists an index set, I_i, in which this factor is strong, that is, $\lambda_1(\beta_{[I_i]}^\intercal \beta_{[I_i]}) \gtrsim |I_i|$. Then we can find $k^\star := \min_k |I_k|$, and build up I_0 from I_{k^*} (so that $|I_0| \geq |I_{k^\star}|$) by adding randomly selected $|I_{k^\star}|$ number of assets from each $I_j, j = 1, 2, \ldots, p, j \neq k^\star$. The resulting index set I_0 contains at most $p \times |I_{k^\star}|$ number of test assets, barring from repeated counts. We thereby construct a common index set such that all factors are strong within this set.

Next, we need the following moment conditions.

Assumption A.4. *The idiosyncratic component U satisfies:*

$$\|U\|_{\mathrm{MAX}} \lesssim_p (\log T)^{1/2} + (\log N)^{1/2}, \quad \|\bar{u}\|_{\mathrm{MAX}} \lesssim_p T^{-1/2}(\log N)^{1/2}.$$

In addition, for any non-random subset $I \subset [N]$,

$$\left\|U_{[I]}\right\| \lesssim_p |I|^{1/2} + T^{1/2}, \quad \left\|\bar{u}_{[I]}\right\| \lesssim_p |I|^{1/2}T^{-1/2}.$$

Assumption A.4 imposes restrictions on the time-series dependence and heteroskedasticity of u_t. The first two inequalities are results of some large deviation theorem, see, e.g., Fan et al. (2011). The last inequality 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 A.5. *For any non-random subset $I \subset [N]$, the factor loading $\beta_{[I]}$ and the idiosyncratic error $U_{[I]}$ satisfy the following conditions:*

$$\begin{aligned} &(i) \quad \left\|(\beta_{[I]}^{\intercal}\beta_{[I]})^{-1/2}\beta_{[I]}^{\intercal}U_{[I]}\right\| \lesssim_p T^{1/2}. \\ &(ii) \quad \left\|(\beta_{[I]}^{\intercal}\beta_{[I]})^{-1/2}\beta_{[I]}^{\intercal}U_{[I]}\iota_T\right\| \lesssim_p T^{1/2}. \end{aligned}$$

If $\beta_{[I]}^{\intercal}\beta_{[I]}$ is singular, we need replace the matrix inverse above by the Moore-Penrose inverse.

Assumption A.6. *The following conditions hold for U, V, β, and any non-random subset $I \subset [N]$:*

$$\begin{aligned} &(i) \quad \left\|U_{[I]}V^{\intercal}\right\| \lesssim_p |I|^{1/2}T^{1/2}, \quad \left\|U_{[I]}V^{\intercal}\right\|_{\mathrm{MAX}} \lesssim_p (\log N)^{1/2}T^{1/2}. \\ &(ii) \quad \left\|(\beta_{[I]}^{\intercal}\beta_{[I]})^{-1/2}\beta_{[I]}^{\intercal}U_{[I]}V^{\intercal}\right\| \lesssim_p T^{1/2}. \end{aligned}$$

Assumption A.7. *The following conditions hold for U, Z, β, and any non-random subset $I \subset [N]$:*

$$\begin{aligned} &(i) \quad \left\|U_{[I]}Z^{\intercal}\right\| \lesssim_p |I|^{1/2}T^{1/2}, \quad \left\|U_{[I]}Z^{\intercal}\right\|_{\mathrm{MAX}} \lesssim_p (\log N)^{1/2}T^{1/2}. \\ &(ii) \quad \left\|(\beta_{[I]}^{\intercal}\beta_{[I]})^{-1/2}\beta_{[I]}^{\intercal}U_{[I]}Z^{\intercal}\right\| \lesssim_p T^{1/2}. \end{aligned}$$

Assumptions A.5 - A.7 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 $I \subset [N]$. 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. For simplicity, we consider the case $\Sigma_v = \mathbb{I}_p$ here. In general case, replace β and η by $\beta^{'} = \beta\Sigma_v^{1/2}$ and $\eta^{'} = \eta\Sigma_v^{1/2}$ in the following definiiton. In detail, we start with $a_i^{(1)} := \left\|\beta_{[i]}\eta^{\intercal}\right\|_{\mathrm{MAX}}$ 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 $a_i^{(k)} := \left\|\beta_{[i]}\prod_{j<k}\mathbb{M}_{b_j}\eta^{\intercal}\right\|_{\mathrm{MAX}}$, $I_k := \{a_i^{(k)} \geq c_{qN}^{(k)}\}$ 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 $\tilde{p}$ 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 A.8. *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 A.8 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 A.9. *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 A.9 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 fairly standard result of a central limit theory for mixing processes.

In the same vein, we make an assumption on the central limit result between $ZV^\intercal$ and $Z\iota_T$, which we use for inference on α_g.

Assumption A.10. *As $T \to \infty$, the following joint central limit theorem holds:*

$$T^{1/2}\begin{pmatrix} T^{-1}\mathrm{vec}(ZV^\intercal) \\ \bar{z} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Pi_{11} & \Pi_{13} \\ \Pi_{13}^\intercal & \Pi_{33} \end{pmatrix}\right),$$

where Π_{13} and Π_{33} are $dp \times d$, and $d \times d$ matrices, respectively, defined as:

$$\Pi_{13} = \lim_{T\to\infty} \frac{1}{T}\mathrm{E}\left(\mathrm{vec}(ZV^\intercal)\iota_T^\intercal Z^\intercal\right),$$
$$\Pi_{33} = \lim_{T\to\infty} \frac{1}{T}\mathrm{E}\left(Z\iota_T\iota_T^\intercal Z^\intercal\right).$$

Blow we introduce assumptions needed for the SDF estimation. Assumption A.11 ensures that the SDF concept is well defined. Assumption A.12 again can be shown by some large deviation result and certain central limit theorem.

Assumption A.11. *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 A.12. *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:*

$$(1) \quad \left\| T^{-1}\sum_{t=1}^{T}(r_t - \bar{r}_t)(m_t - \bar{m}_t) - \mathrm{Cov}(r_t, m_t) \right\|_{\mathrm{MAX}} \lesssim_p (\log N)^{1/2}T^{-1/2}.$$

$$(2) \quad \left\| T^{-1}\sum_{t=1}^{T}(r_t - \bar{r}_t)(r_t - \bar{r}_t)^\intercal - \mathrm{Cov}(r_t) \right\|_{\mathrm{MAX}} \lesssim_p (\log N)^{1/2}T^{-1/2}.$$

$$(3) \quad \left| T^{-1}\sum_{t=1}^{T} m_t - \mathrm{E}(m_t) \right| \lesssim_p T^{-1/2}.$$

$$(4) \quad \left\| T^{-1}\sum_{t=1}^{T} r_t - \mathrm{E}(r_t) \right\|_{\mathrm{MAX}} \lesssim_p (\log N)^{1/2}T^{-1/2}.$$

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 A.8. A similar assumption has been adopted by, e.g., Wang and Fan (2017).

Assumption A.13. *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 $\beta\Sigma_v\beta^\intercal$.

B Mathematical Proofs

B.1 Proof of Proposition 1

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 1, 2, and 3. 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:

$$\bar{R} = \beta\bar{V} + \bar{U} = \begin{pmatrix} \sqrt{\lambda}\bar{V} + \bar{U}_1 \\ \bar{U}_2 \end{pmatrix}, \tag{B.1}$$

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 $\varsigma = (\varsigma_1, \varsigma_2^\intercal)^\intercal$, where ς_1 is the first element of ς and ς_2 is a vector of the remaining $N-1$ entries of ς. Recall that in Algorithm 1, we denote ξ and ς 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

$$\varsigma_1 = \frac{(\sqrt{\lambda}\bar{V} + \bar{U}_1)\xi}{\sqrt{T\widehat{\lambda}}}, \quad \varsigma_2 = \frac{\bar{U}_2\xi}{\sqrt{T\widehat{\lambda}}}. \tag{B.2}$$

Since the entries of U and V are i.i.d $\mathcal{N}(0,1)$, we have

$$|T^{-1}\bar{V}\bar{V}^\intercal - 1| = |T^{-1}V(\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)V^\intercal - 1| \leq |T^{-1}VV^\intercal - 1| + |\bar{v}|^2 \lesssim_p T^{-1/2},$$

where we use large deviation results $|T^{-1}VV^\intercal - 1| \lesssim_p T^{-1/2}$ and $|\bar{v}| \lesssim_P T^{-1/2}$ in the last equation. This equation also implies that $\|\bar{V}\| - \sqrt{T} \lesssim_p 1$.

Similarly, we can get $|T^{-1}\bar{U}_1\bar{U}_1^\intercal - 1| \lesssim_p T^{-1/2}$ and $\|\bar{U}_1\| - \sqrt{T} \lesssim_p 1$.

In addition, by Lemma A.1 in Wang and Fan (2017), we have $\|N^{-1}U^\intercal U - \mathbb{I}_T\| \lesssim_p \sqrt{T/N}$, which leads to

$$\|N^{-1}\bar{U}^\intercal\bar{U} - (\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)\| = \|(\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)(N^{-1}U^\intercal U - \mathbb{I}_T)(\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)\| \lesssim_p \sqrt{T/N}.$$

Next, by direct calculation using the above inequalities we obtain

$$\left\| \frac{\bar{V}^\intercal\bar{U}_1 + \bar{U}_1^\intercal\bar{V}}{T\sqrt{\lambda}} + \frac{\bar{U}^\intercal\bar{U} - N(\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)}{T\lambda} \right\| \lesssim_p \frac{1}{\sqrt{\lambda}} + \frac{\sqrt{NT}}{T\lambda} \lesssim_p \frac{1}{\sqrt{\lambda}}.$$

Together with (B.1), we have

$$\left\| \frac{\bar{R}^\intercal\bar{R}}{T\lambda} - \frac{\bar{V}^\intercal\bar{V}}{T} - \frac{N(\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)}{T\lambda} \right\| \lesssim_p \frac{1}{\sqrt{\lambda}}. \tag{B.3}$$

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:

$$\lambda_i = \begin{cases} T^{-1}\bar{V}\bar{V}^\intercal + \tilde{B} & i = 1; \\ \tilde{B} & 2 \le i \le T-1; \\ 0 & i = T. \end{cases}, \tag{B.4}$$

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 ξ, Weyl's theorem yields that

$$\frac{\widehat{\lambda}}{\lambda} = \frac{\bar{V}\bar{V}^\intercal}{T} + \tilde{B} + O_p\left(\frac{1}{\sqrt{\lambda}}\right) = 1 + \tilde{B} + O_p\left(\frac{1}{\sqrt{\lambda}} + \frac{1}{\sqrt{T}}\right), \tag{B.5}$$

and the sin-theta theorem in Davis and Kahan (1970) implies that

$$\|\mathbb{P}_{\bar{V}^\intercal} - \mathbb{P}_\xi\| = \left\|\bar{V}^\intercal(\bar{V}\bar{V}^\intercal)^{-1}\bar{V} - \xi\xi^\intercal\right\| \lesssim_p \frac{1}{\sqrt{\lambda}}, \tag{B.6}$$

which implies that $(\bar{V}\bar{V})^{-1}(\bar{V}\xi)^2 = \xi^\intercal\bar{V}^\intercal(\bar{V}\bar{V})^{-1}\bar{V}\xi = 1 + O_p(\lambda^{-1/2} + T^{-1/2})$. Together with $|T^{-1}\bar{V}\bar{V}^\intercal - 1| \lesssim T^{-1/2}$, we have

$$\frac{|\bar{V}\xi|}{\sqrt{T}} = 1 + O_p\left(\frac{1}{\sqrt{\lambda}} + \frac{1}{\sqrt{T}}\right). \tag{B.7}$$

It is easy to observe that the sign of ξ plays no role in the estimator $\widehat{\gamma}_g^{PCA}$, we can choose ξ such that

$$\frac{\bar{V}\xi}{\sqrt{T}} = 1 + O_p\left(\frac{1}{\sqrt{\lambda}} + \frac{1}{\sqrt{T}}\right). \tag{B.8}$$

Recall that the risk premium estimator is $\widehat{\gamma}_g^{PCA} = \widehat{\eta}\widehat{\gamma}$, where

$$\widehat{\eta} = \frac{\bar{G}\xi}{\sqrt{T}} \quad \text{and } \widehat{\gamma} = \frac{\varsigma^\intercal \bar{r}}{\sqrt{\widehat{\lambda}}}. \tag{B.9}$$

Using $\bar{G} = \eta\bar{V}$ and (B.8), we have

$$\widehat{\eta} = \eta + O_p\left(\frac{1}{\sqrt{\lambda}} + \frac{1}{\sqrt{T}}\right). \tag{B.10}$$

Write

$$\widehat{\gamma}=\frac{\varsigma^{\intercal}\bar{r}}{\sqrt{\widehat{\lambda}}}=\frac{\varsigma^{\intercal}\beta(\gamma+\bar{v})}{\sqrt{\widehat{\lambda}}}+\frac{\varsigma^{\intercal}\bar{u}}{\sqrt{\widehat{\lambda}}}=\frac{\sqrt{\lambda}\varsigma_1}{\sqrt{\widehat{\lambda}}}(\gamma+\bar{v})+\frac{\varsigma^{\intercal}\bar{u}}{\sqrt{\widehat{\lambda}}}, \tag{B.11}$$

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 (B.11) one by one. For the first term, using (B.2), we have

$$\frac{\sqrt{\lambda}\varsigma_1}{\sqrt{\widehat{\lambda}}}=\frac{\lambda}{\widehat{\lambda}}\frac{(\bar{V}+\lambda^{-1/2}\bar{U}_1)\xi}{\sqrt{T}}=\frac{\lambda}{\widehat{\lambda}}\left(\frac{\bar{V}\xi}{\sqrt{T}}+\frac{\bar{U}_1\xi}{\sqrt{T\lambda}}\right).$$

Using (B.5) and (B.8) and $\left\|\bar{U}_1\right\|\lesssim_p\sqrt{T}$, it follows that

$$\frac{\sqrt{\lambda}\varsigma_1}{\sqrt{\widehat{\lambda}}}=\frac{1}{1+\tilde{B}}+O_p\left(\frac{1}{\sqrt{\lambda}}+\frac{1}{\sqrt{T}}\right). \tag{B.12}$$

For the second term in (B.11), using (B.2) again, we can write

$$\frac{\varsigma^{\intercal}\bar{u}}{\sqrt{\widehat{\lambda}}}=\frac{\varsigma_1 U_1\iota_T}{T\sqrt{\widehat{\lambda}}}+\frac{\varsigma_2^{\intercal}U_2\iota_T}{T\sqrt{\widehat{\lambda}}}=\frac{\varsigma_1 U_1\iota_T}{T\sqrt{\widehat{\lambda}}}+\frac{\xi^{\intercal}(\mathbb{I}_T-T^{-1}\iota_T\iota_T^{\intercal})U_2^{\intercal}U_2\iota_T}{T^{3/2}\widehat{\lambda}}. \tag{B.13}$$

The condition that entries of U are independent $\mathcal{N}(0,1)$ implies that $\|U_1\iota_T\|\lesssim_p\sqrt{T}$, with $\widehat{\lambda}/\lambda\xrightarrow{p}1+B$ as shown in (B.5), the first term in (B.13) is of order $O_p(T^{-1/2}\lambda^{-1/2})$. For the second term in (B.13), using $\left\|(N-1)^{-1}U_2^{\intercal}U_2-\mathbb{I}_T\right\|\lesssim_p\sqrt{T/N}$, we have

$$\begin{aligned}|\frac{\xi^{\intercal}(\mathbb{I}_T-T^{-1}\iota_T\iota_T^{\intercal})U_2^{\intercal}U_2\iota_T}{T^{3/2}\widehat{\lambda}}|\leq&|\frac{(N-1)\xi^{\intercal}(\mathbb{I}_T-T^{-1}\iota_T\iota_T^{\intercal})\iota_T}{T^{3/2}\widehat{\lambda}}|+\frac{N-1}{T\widehat{\lambda}}\left\|(N-1)^{-1}U_2^{\intercal}U_2-\mathbb{I}_T\right\|\\=&\frac{N-1}{T\widehat{\lambda}}\left\|(N-1)^{-1}U_2^{\intercal}U_2-\mathbb{I}_T\right\|\lesssim_p\frac{1}{\sqrt{\lambda}},\end{aligned}$$

which leads to $|\widehat{\lambda}^{-1/2}\varsigma^{\intercal}\bar{u}|\lesssim_p\lambda^{-1/2}$. Plugging this and (B.12) into (B.11), we obtain

$$\widehat{\gamma}=\frac{\varsigma^{\intercal}\bar{r}}{\sqrt{\widehat{\lambda}}}=\frac{\gamma}{1+\tilde{B}}+O_p\left(\frac{1}{\sqrt{\lambda}}+\frac{1}{\sqrt{T}}\right), \tag{B.14}$$

and thus $\widehat{\gamma}_g^{PCA}\xrightarrow{p}(1+B)^{-1}\eta\gamma$ by (B.10), (B.14) and $\tilde{B}\to B$. □

B.2 Proof of Proposition 2

Proof. Recall that in Section 2.3.2, we have

$$\widehat{\gamma}_g^{PLS}=\left\|\bar{G}\bar{R}^{\intercal}\bar{R}\right\|^{-2}\bar{G}\bar{R}^{\intercal}\bar{R}\bar{G}^{\intercal}\bar{G}\bar{R}^{\intercal}\bar{r}. \tag{B.15}$$

We analyze $\|\bar{G}\bar{R}^\intercal\bar{R}\|$, $\bar{G}\bar{R}^\intercal\bar{R}\bar{G}^\intercal$ and $\bar{G}\bar{R}^\intercal\bar{r}$ separately. Recall that from (B.3), we have

$$\left\|\frac{\bar{R}^\intercal\bar{R}}{T\lambda} - \frac{\bar{V}^\intercal\bar{V}}{T} - \tilde{B}(\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)\right\| \lesssim_p \frac{1}{\sqrt{\lambda}},$$

where $\tilde{B} = N/(T\lambda)$ satisfies $\tilde{B} \to B$. Together with $\bar{G} = \eta\bar{V}$ and $\|\bar{G}\| \lesssim_p \sqrt{T}$, we have

$$\begin{aligned}\frac{1}{T\lambda\sqrt{T}}\left\|\bar{G}\bar{R}^\intercal\bar{R}\right\| &= \frac{1}{\sqrt{T}}\left\|\bar{G}\left(\frac{\bar{V}^\intercal\bar{V}}{T} + \tilde{B}(\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)\right)\right\| + O_p\left(\frac{1}{\sqrt{\lambda}}\right)\\ &= \frac{\eta}{\sqrt{T}}\left\|\frac{\bar{V}\bar{V}^\intercal\bar{V}}{T} + \tilde{B}\bar{V}\right\| + O_p\left(\frac{1}{\sqrt{\lambda}}\right) \xrightarrow{p} \eta(1+B),\end{aligned} \tag{B.16}$$

where we use $|T^{-1}\bar{V}\bar{V}^\intercal - 1| \lesssim_p T^{-1/2}$ and $\|\bar{V}\| - \sqrt{T} \lesssim_p 1$ in the last equation. For the same reason, by direct calculation we have

$$\begin{aligned}\frac{1}{T^2\lambda}\bar{G}\bar{R}^\intercal\bar{R}\bar{G}^\intercal &= \frac{1}{T}\bar{G}\left(\frac{\bar{V}^\intercal\bar{V}}{T} + \tilde{B}(\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)\right)\bar{G}^\intercal + O_p\left(\frac{1}{\sqrt{\lambda}}\right)\\ &= \eta^2\frac{\bar{V}\bar{V}^\intercal\bar{V}\bar{V}^\intercal}{T^2} + \eta^2\tilde{B}\frac{\bar{V}\bar{V}^\intercal}{T} + O_p\left(\frac{1}{\sqrt{\lambda}}\right) \xrightarrow{p} \eta^2(1+B).\end{aligned} \tag{B.17}$$

Next, we write

$$\frac{1}{T\lambda}\bar{G}\bar{R}^\intercal\bar{r} = \frac{1}{T\lambda}\bar{G}\bar{R}^\intercal\beta(\gamma+\bar{v}) + \frac{1}{T\lambda}\bar{G}\bar{R}^\intercal\bar{u}. \tag{B.18}$$

We analyze these two terms in (B.18) separately. For the first term, we can write $\bar{R}$ in the form of (B.1) as in the proof of Proposition 1. Then, using $\|\bar{U}_1\| \lesssim_p \sqrt{T}$ we have

$$\frac{1}{T\lambda}\bar{G}\bar{R}^\intercal\beta = \eta\frac{\bar{V}\bar{V}^\intercal}{T} + \eta\frac{\bar{V}\bar{U}_1^\intercal}{T\sqrt{\lambda}} = \eta\frac{\bar{V}\bar{V}^\intercal}{T} + O_p\left(\frac{1}{\sqrt{\lambda}}\right). \tag{B.19}$$

For the second term in (B.18), we have

$$\begin{aligned}\frac{1}{T\lambda}\bar{G}\bar{R}^\intercal\bar{u} &= \eta\frac{1}{T^2\sqrt{\lambda}}\bar{V}\bar{V}^\intercal\bar{U}_1\iota_T + \eta\frac{1}{T^2\lambda}\bar{V}\bar{U}^\intercal U\iota_T = \eta\frac{1}{\sqrt{\lambda}}\frac{\bar{V}\bar{V}^\intercal}{T}\frac{\bar{U}_1\iota_T}{T} + \eta\frac{1}{T^2\lambda}\bar{V}U^\intercal U\iota_T\\ &= O_p\left(\frac{1}{\sqrt{T\lambda}}\right) + \eta\frac{N}{T^2\lambda}\bar{V}\left(N^{-1}U^\intercal U - \mathbb{I}_T\right)\iota_T + \eta\frac{N}{T^2\lambda}\bar{V}\iota_T = O_p\left(\frac{1}{\sqrt{T\lambda}}\right) + O_p\left(\frac{1}{\sqrt{\lambda}}\right),\end{aligned} \tag{B.20}$$

where we use $\|N^{-1}U^\intercal U - \mathbb{I}_T\| \lesssim_p \sqrt{T/N}$ and $\bar{V}\iota_T = 0$ in the last equation. Plugging (B.19) and (B.20) into (B.18), we have

$$\frac{1}{T\lambda}\bar{G}\bar{R}^\intercal\bar{r} = \eta\frac{\bar{V}\bar{V}^\intercal}{T}(\gamma+\bar{v}) + O_p\left(\frac{1}{\sqrt{\lambda}}\right) \xrightarrow{p} \eta\gamma. \tag{B.21}$$

Plug (B.16), (B.17), (B.21) into (B.15), we have

$$\widehat{\gamma}_g^{PLS} \xrightarrow{p} \frac{1}{\eta^2(1+B)^2}\eta^2(1+B)\eta\gamma = \frac{1}{1+B}\eta\gamma.$$

□

B.3 Proof of Proposition 3

Proof. Since $\operatorname{Rank}(\bar{R}) \leq \min\{N, T-1\}$, and the assumptions of the proposition imply that $N/T \to \infty$, we thereby have a condensed SVD of $\bar{R}$ as

$$\bar{R} = \sqrt{T}(\varsigma, \varsigma_*)\widehat{\Lambda}^{1/2}(\xi, \xi_*)^\intercal = \sqrt{T}\varsigma\widehat{\lambda}^{1/2}\xi^\intercal + \sqrt{T}\varsigma_*\widehat{\Lambda}_*^{1/2}\xi_*^\intercal,$$

where $\widehat{\Lambda}^{1/2}$ is the diagonal matrix of $T-1$ singular values, ς, ξ 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, $\varsigma_* \in \mathbb{R}^{N\times(T-2)}$ and $\xi_* \in \mathbb{R}^{T\times(T-2)}$ 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

$$\sqrt{T}\bar{R}^\intercal\left(\bar{R}\bar{R}^\intercal + \mu I\right)^{-1} = (\xi, \xi_*)\widehat{\Lambda}^{1/2}(\widehat{\Lambda} + T^{-1}\mu I)^{-1}(\varsigma, \varsigma_*)^\intercal = \frac{\widehat{\lambda}^{1/2}}{\widehat{\lambda} + T^{-1}\mu}\xi\varsigma^\intercal + \xi_*\widehat{\Lambda}_*^{1/2}\left(\widehat{\Lambda}_* + T^{-1}\mu I\right)^{-1}\varsigma_*^\intercal,$$

and thus, with $\bar{G} = \eta\bar{V}$, the Ridge estimator can be written as

$$\begin{aligned}\widehat{\gamma}_g^{Ridge} = \bar{G}\bar{R}^\intercal\left(\bar{R}\bar{R}^\intercal + \mu I\right)^{-1}\bar{r} &= \frac{\widehat{\lambda}}{\widehat{\lambda} + T^{-1}\mu}\frac{\eta\bar{V}\xi}{\sqrt{T}}\frac{\varsigma^\intercal\bar{r}}{\sqrt{\widehat{\lambda}}} + \frac{\eta\bar{V}\xi_*}{\sqrt{T}}\widehat{\Lambda}_*^{1/2}\left(\widehat{\Lambda}_* + T^{-1}\mu\right)^{-1}\varsigma_*^\intercal\bar{r} \\ &= \frac{\widehat{\lambda}}{\widehat{\lambda} + T^{-1}\mu}\widehat{\gamma}_g^{PCA} + \frac{\eta\bar{V}\xi_*}{\sqrt{T}}\widehat{\Lambda}_*^{1/2}\left(\widehat{\Lambda}_* + T^{-1}\mu\right)^{-1}\varsigma_*^\intercal\bar{r}. \qquad \text{(B.22)}\end{aligned}$$

Using (B.5) and the fact that $T^{-1}\lambda^{-1}\mu \to D$ and Proposition 1, we can show that the first term in (B.22) converges to $(1+B+D)^{-1}\eta\gamma$. With respect to the second term, as shown in (B.3), we have

$$\left\|\frac{\bar{R}^\intercal\bar{R}}{T\lambda} - \frac{\bar{V}^\intercal\bar{V}}{T} - \frac{N(\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)}{T\lambda}\right\| \lesssim_p \frac{1}{\sqrt{\lambda}},$$

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 (B.4), it then follows from Weyl's theorem that $\lambda_i(T^{-1}\lambda^{-1}\bar{R}^\intercal\bar{R}) = \tilde{B} + O_p(\lambda^{-1/2})$ for $2 \leq i \leq T-1$. 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

$$\left\|\widehat{\Lambda}_*^{1/2}\left(\widehat{\Lambda}_* + T^{-1}\mu\right)^{-1}\right\| = \max_{1\leq i\leq T-2}\frac{\lambda_{i+1}(T^{-1}\bar{R}^\intercal \bar{R})^{1/2}}{\lambda_{i+1}(T^{-1}\bar{R}^\intercal \bar{R})+T^{-1}\mu} \lesssim_p \frac{1}{\sqrt{\lambda}}. \tag{B.23}$$

Also, with $\|\bar{u}\| \lesssim_p \sqrt{N/T}$, we have

$$\|\varsigma_*^\intercal \bar{r}\| \leq \|\varsigma_*^\intercal \beta(\gamma+\bar{v})\| + \|\varsigma_*^\intercal \bar{u}\| \leq \|\beta(\gamma+\bar{v})\| + \|\bar{u}\| \lesssim_p \sqrt{\lambda} + \sqrt{N/T} \lesssim_p \sqrt{\lambda} \tag{B.24}$$

and

$$\left\|\frac{\bar{V}\xi_*}{\sqrt{T}}\right\|^2 = \left\|\frac{\bar{V}(\xi,\xi_*)}{\sqrt{T}}\right\|^2 - \left\|\frac{\bar{V}\xi}{\sqrt{T}}\right\|^2 \leq \left\|\frac{\bar{V}}{\sqrt{T}}\right\|^2 - \left\|\frac{\bar{V}\xi}{\sqrt{T}}\right\|^2 = 1 + O_p\left(\frac{1}{\sqrt{T}}\right) - \left\|\frac{\bar{V}\xi}{\sqrt{T}}\right\|^2 \lesssim_p \frac{1}{\sqrt{\lambda}} + \frac{1}{\sqrt{T}}, \tag{B.25}$$

where we use (B.8) in the last inequality. Consequently, using (B.23), (B.24) and (B.25), we have

$$|\frac{\eta\bar{V}\xi_*}{\sqrt{T}}\widehat{\Lambda}_*^{1/2}\left(\widehat{\Lambda}_* + T^{-1}\mu\right)^{-1}\varsigma_*^\intercal \bar{r}| \leq \left\|\frac{\eta\bar{V}\xi_*}{\sqrt{T}}\right\|\left\|\widehat{\Lambda}_*^{1/2}\left(\widehat{\Lambda}_* + T^{-1}\mu\right)^{-1}\right\|\|\varsigma_*^\intercal \bar{r}\| \lesssim T^{-1/4} + \lambda^{-1/4}.$$

By comparing this with the limit of the first term in (B.22), we obtain

$$\widehat{\gamma}_g^{Ridge} \xrightarrow{p} \frac{1}{1+B+D}\eta\gamma.$$

□

B.4 Proof of Proposition 4

Proof. By direct calculation, we can write

$$RR^\intercal + T\mu\bar{r}\bar{r}^\intercal = R\left(\mathbb{I}_T + \frac{\mu}{T}\iota_T\iota_T^\intercal\right)R^\intercal = R\left(\mathbb{I}_T + \frac{\tilde{\mu}}{T}\iota_T\iota_T^\intercal\right)^2 R^\intercal, \tag{B.26}$$

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 ς and ξ denote the largest left and right singular vector of $R\left(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^\intercal\right)$. Note that ξ 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 (B.3) in the PCA case, we have the following approximation of $R^\intercal R$

$$\left\| \frac{R^\intercal R}{T\lambda} - \frac{\bar{V}^\intercal \bar{V}}{T} - \gamma \frac{\iota_T \bar{V} + \bar{V}^\intercal \iota_T^\intercal}{T} - \gamma^2 \frac{\iota_T \iota_T^\intercal}{T} - \frac{N}{T\lambda}\mathbb{I}_T \right\| \lesssim_p \frac{1}{\sqrt{T}} + \frac{1}{\sqrt{\lambda}}, \tag{B.27}$$

by $|T^{-1}\bar{V}\bar{V}^\intercal - 1| \lesssim_p T^{-1/2}$, $\|\bar{U}_1\| \lesssim_p T^{1/2}$ and $\|N^{-1}\bar{U}^\intercal \bar{U} - (\mathbb{I}_T - T^{-1}\iota_T \iota_T^\intercal)\| \lesssim_p \sqrt{T/N}$.

Then, with (B.27) and $N/(T\lambda) \to B$, we have

$$\left\| T^{-1}\lambda^{-1}(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^\intercal) R^\intercal R (\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^\intercal) - M^* \right\| = o_p(1) \tag{B.28}$$

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 ξ 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 ξ.

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

$$(x - a_1)(x - a_3) - a_2^2 = 0, \tag{B.29}$$

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 (B.29) is

$$\frac{a_1 + a_3 + \sqrt{(a_1 - a_3)^2 + 4a_2^2}}{2} \geq a_1 > 0 \tag{B.30}$$

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 (B.29), 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 (B.30) 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 $\lambda_1(M^* - B\mathbb{I}_T) - \lambda_2(M^* - B\mathbb{I}_T) \geq \delta > 0$ for some constant $\delta > 0$. Consequently,

$$\lambda_1(M^*) - \lambda_2(M^*) = \lambda_1(M^* - B\mathbb{I}_T) - \lambda_2(M^* - B\mathbb{I}_T) \geq \delta, \tag{B.31}$$

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 (B.29), 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 (B.30) into $\lambda_1(M - B\mathbb{I}_T)\tilde{\xi} = (M - B\mathbb{I}_T)\tilde{\xi}$, we directly get

$$\frac{K_2}{K_1} = \frac{\sqrt{(a_1 - a_3)^2 + 4a_2^2} + a_3 - a_1}{2a_2}, \tag{B.32}$$

which will pin down K_1 and K_2 because we also have $\left\|\tilde{\xi}\right\| = 1$.

Using $\left\|T^{-1}\lambda^{-1}(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^\intercal)R^\intercal R(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^\intercal) - M\right\| = o_p(1)$, (B.31) and sin-theta theorem, we have

$$\left\|\mathbb{P}_\xi - \mathbb{P}_{\tilde{\xi}}\right\| \leq \frac{o_p(1)}{\delta - o_p(1)} = o_p(1),$$

which implies that $|\tilde{\xi}^\intercal\xi| \xrightarrow{p} 1$ and consequently,

$$\left\|\xi - K_1 \left\|\bar{V}\right\|^{-1} \bar{V}^\intercal - K_2 T^{-1/2}\iota_T\right\| = o_p(1) \quad \text{or} \quad \left\|\xi + K_1 \left\|\bar{V}\right\|^{-1} \bar{V}^\intercal + K_2 T^{-1/2}\iota_T\right\| = o_p(1).$$

Since the sign of ξ 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

$$\widehat{F} = \varsigma^\intercal R = \left\|R(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^\intercal)\right\|^{-1} \xi^\intercal(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^\intercal)R^\intercal R. \tag{B.33}$$

With the approximation of $R^\intercal R$ in (B.27), $\bar{V}\iota_T = 0$, $T^{-1}\bar{V}\bar{V}^\intercal = 1 + O_p(T^{-1/2})$ and $N/(T\lambda) \to B$, by direct calculation, we have

$$\left\|\left\|\bar{V}\right\|^{-1} \bar{V}(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^\intercal)R^\intercal R - \lambda T^{1/2}\Big((1+B)\bar{V} + \gamma\iota_T^\intercal\Big)\right\| = o_p(\lambda T), \tag{B.34}$$

and

$$\left\|T^{-1/2}\iota_T^\intercal(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^\intercal)R^\intercal R - \lambda T^{1/2}(1+\tilde{\mu})\Big(\gamma\bar{V} + (\gamma^2 + B)\iota_T^\intercal\Big)\right\| = o_p(\lambda T). \tag{B.35}$$

Plugging (B.34), (B.35) and $\left\|\xi - K_1 \left\|\bar{V}\right\|^{-1} \bar{V}^\intercal + K_2 T^{-1/2}\iota_T\right\| = o_p(1)$ into (B.33) we have

$$\left\|\left\|R(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^\intercal)\right\| \widehat{F} - \lambda T^{1/2}(L_1\bar{V} + L_2\iota_T^\intercal)\right\| = o_p(\lambda T), \tag{B.36}$$

where

$$L_1 = K_1(1+B) + K_2(1+\tilde{\mu})\gamma, \qquad L_2 = K_1\gamma + K_2(1+\tilde{\mu})(\gamma^2 + B). \tag{B.37}$$

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, (B.36) becomes $\left\|\widehat{F}^* - \bar{V} - L_1^{-1}L_2\iota_T^\intercal\right\| = o_p\left(T^{1/2}\right)$. Consequently, $\left\|\widehat{V} - \bar{V}\right\| = \left\|\widehat{F}^*(\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal) - \bar{V}\right\| = o_p\left(T^{1/2}\right)$, $\widehat{\gamma} = T^{-1}\widehat{F}^*\iota_T = L_1^{-1}L_2 + o_p(1)$, and

$$\widehat{\eta} = \bar{G}\widehat{V}^\intercal(\widehat{V}\widehat{V}^\intercal)^{-1} = \eta\bar{V}\widehat{V}^\intercal(\widehat{V}\widehat{V}^\intercal)^{-1} = \eta\Big(\bar{V}\bar{V}^\intercal + o_p(T)\Big)\Big(\bar{V}\bar{V}^\intercal + o_p(T)\Big)^{-1} = \eta + o_p(1),$$

and the estimator $\widehat{\gamma}_g^{rpPCA} = \widehat{\eta}\widehat{\gamma} \xrightarrow{p} \eta L_1^{-1}L_2$, where L_1 and L_2 are defined in (B.37).

In light of that $a_1 \xrightarrow{p} 1$, $a_2 \xrightarrow{p} (1+\tilde{\mu})\gamma$, $\tilde{\mu} = \sqrt{1+\mu} - 1$, $\widehat{\gamma}_g^{rpPCA} \xrightarrow{p} \eta L_2/L_1$, (B.32) and the definitions of L_1 and L_2 in (B.37), we have

$$\widehat{\gamma}_g^{rpPCA} \xrightarrow{p} w(1+B)^{-1}\eta\gamma + (1-w)\eta(\gamma + \gamma^{-1}B),$$

where

$$w = \frac{2+2B}{1+2B+\sqrt{(1-a)^2 + 4(1+\mu)\gamma} + a}, \qquad a = (1+\mu)(\gamma^2 + B) - B.$$

□

B.5 Proof of Proposition 5

Proof. Consider the set $I = \{|\beta_{[i]}| \geq \beta_{\{qN\}}\}$, where $|\beta|_{\{qN\}}$ is the (qN)th largest value in $\left\{|\beta_{[i]}|\right\}_{i\in[N]}$. 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

$$\begin{aligned}\left\|T^{-1}\bar{R}\bar{G}^\intercal - \beta\eta^\intercal\right\|_{\text{MAX}} \lesssim \|\beta\|_{\text{MAX}}\left|T^{-1}\bar{V}\bar{V}^\intercal - 1\right|\|\eta\| + T^{-1}\left\|\bar{U}\bar{V}^\intercal\right\|_{\text{MAX}}\|\eta\| \\ + T^{-1}\|\beta\|_{\text{MAX}}\left\|\bar{V}\bar{Z}^\intercal\right\| + T^{-1}\left\|\bar{U}\bar{Z}^\intercal\right\|_{\text{MAX}} \lesssim_p (\log N)^{1/2}T^{-1/2}.\end{aligned}$$

In other words, the difference between $T^{-1}\bar{R}\bar{G}^\intercal$ and $\beta\eta^\intercal$ for all test assets is bounded by $O_p\left((\log N)^{1/2}T^{-1/2}\right)$, which is $o(1)$ under our assumption.

On the other hand, with the assumption that $\|\beta\|_{\text{MAX}} \lesssim 1$ 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 (B.11), except for an extra term that depends on $\bar{Z}$ and the restriction of $\bar{r}$ on I:

$$\widehat{\gamma}_g^{SPCA} = \frac{\eta \bar{V} \xi}{\sqrt{T}} \frac{\varsigma^\intercal \bar{r}_{[I]}}{\sqrt{\widehat{\lambda}}} + \frac{\bar{Z}\xi}{\sqrt{T}} \frac{\varsigma^\intercal \bar{r}_{[I]}}{\sqrt{\widehat{\lambda}}}. \tag{B.38}$$

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 1, which yields

$$\frac{\eta \bar{V} \xi}{\sqrt{T}} \frac{\varsigma^\intercal \bar{r}_{[I]}}{\sqrt{\widehat{\lambda}}} = \eta\gamma + o_p(1).$$

We now analyze the second term in (B.38). As shown in (B.14), we have

$$\left\| \frac{\varsigma^\intercal \bar{r}_{[I]}}{\sqrt{\widehat{\lambda}}} \right\| \lesssim_p 1,$$

so to prove that SPCA is consistent in this case, it is sufficient to show that $T^{-1/2}\left\|\bar{Z}\xi\right\| \xrightarrow{p} 0$, where ξ is the largest right singular vector of $\bar{R}_{[I]}$. Similar to the proof of (B.6) in Proposition 1, we can show that the difference between projection matrices, $\mathbb{P}_\xi$ and $\mathbb{P}_{\bar{V}^\intercal}$ is small by sin-theta theorem. That is to say, we have $\left\|\xi\xi^\intercal - \bar{V}^\intercal(\bar{V}\bar{V}^\intercal)^{-1}\bar{V}\right\| \xrightarrow{p} 0$. Then, with the fact that

$$\left\|\bar{Z}\bar{V}^\intercal(\bar{V}\bar{V}^\intercal)^{-1}\bar{V}\right\| \leq \left\|\bar{Z}\bar{V}^\intercal\right\| \left\|(\bar{V}\bar{V}^\intercal)^{-1}\right\| \left\|\bar{V}\right\| \lesssim_p T^{1/2} \times T^{-1} \times T^{1/2} \lesssim_p 1,$$

we have $T^{-1/2}\left\|\bar{Z}\xi\xi^\intercal\right\| \xrightarrow{p} 0$. Consequently,

$$T^{-1/2}\left\|\bar{Z}\xi\right\| = T^{-1/2}\left\|\bar{Z}\xi\xi^\intercal\xi\right\| \leq T^{-1/2}\left\|\bar{Z}\xi\xi^\intercal\right\| \left\|\xi\right\| \xrightarrow{p} 0.$$

Hence, z_t does not affect the consistency of the SPCA estimator. This completes the proof. □

B.6 Proof of Theorem 1

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 A.1-A.8 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 1-4, 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 5. 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 $\varsigma_{(k)}$ and $\xi_{(k)}$, while the largest singular value of $\widetilde{R}_{(k)}$ is denoted by $\sqrt{T\widehat{\lambda}_{(k)}}$. As a result, $\widehat{\lambda}_{(k)} = T^{-1}\left\|\widetilde{R}_{(k)}\right\|^2$.

Using the above notation, our estimated factor at k-th step is $\widehat{V}_{(k)} = \sqrt{T}\xi_{(k)}^\intercal \in \mathbb{R}^{1\times T}$, the risk premium of this factor is given by $\widehat{\gamma}_{(k)} = \widehat{\lambda}_{(k)}^{-1/2}\varsigma_{(k)}^\intercal\widetilde{r}_{(k)}$, the loading matrix of R on this factor is $\widehat{\beta}_{(k)} = T^{-1/2}\bar{R}\xi_{(k)}$, and the loading of G on this factor is $\widehat{\eta}_{(k)} = T^{-1/2}\bar{G}\xi_{(k)}$. By footnote 14, we can use $\bar{G}$ instead of $\bar{G}_{(k)}$ in Algorithm 5 and throughout the proof. We denote $\widehat{\eta} = (\widehat{\eta}_{(1)}, \dots, \widehat{\eta}_{(\tilde{p})})$ and $\widehat{\gamma} = (\widehat{\gamma}_{(1)}, \dots, \widehat{\gamma}_{(\tilde{p})})^\intercal$, so the risk premium estimator is $\widehat{\gamma}_g^{SPCA} = \widehat{\eta}\widehat{\gamma}$.

By Lemma 2, we have $\xi_{(i)}^\intercal\xi_{(j)} = 0$ for $i \neq j \leq \tilde{p}$. This suggests that $\widehat{V}_{(k)}$ at each step k are pairwise orthogonal. Using this property and the definition of $\widetilde{R}_{(k)}$, we have

$$\widetilde{R}_{(k)} := \left(\bar{R}_{(k)}\right)_{[\widehat{I}_k]} = \bar{R}_{[\widehat{I}_k]}\prod_{i=1}^{k-1}\mathbb{M}_{\widehat{V}_{(i)}^\intercal} = \bar{R}_{[\widehat{I}_k]}\left(\mathbb{I}_T - \sum_{i=1}^{k-1}\xi_{(i)}\xi_{(i)}^\intercal\right), \tag{B.39}$$

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

$$\widetilde{R}_{(1)} = \bar{R}_{[\widehat{I}_1]} = \beta_{[\widehat{I}_1]}\bar{V} + \bar{U}_{[\widehat{I}_1]}.$$

If we define $\widetilde{\beta}_{(1)} = \beta_{[\widehat{I}_1]}$ and $\widetilde{U}_{(1)} = \bar{U}_{[\widehat{I}_1]}$, then $\widetilde{R}_{(1)}$ can be written in the form $\widetilde{R}_{(1)} = \widetilde{\beta}_{(1)}\bar{V} + \widetilde{U}_{(1)}$. We can iteratively define

$$\widetilde{U}_{(k)} := \bar{U}_{[\widehat{I}_k]} - \sum_{i=1}^{k-1}\frac{\bar{R}_{[\widehat{I}_k]}\xi_{(i)}}{\sqrt{T}}\frac{\varsigma_{(i)}^\intercal\widetilde{U}_{(i)}}{\sqrt{\widehat{\lambda}_{(i)}}} \quad \text{and} \quad \widetilde{\beta}_{(k)} := \beta_{[\widehat{I}_k]} - \sum_{i=1}^{k-1}\frac{\bar{R}_{[\widehat{I}_k]}\xi_{(i)}}{\sqrt{T}}\frac{\varsigma_{(i)}^\intercal\widetilde{\beta}_{(i)}}{\sqrt{\widehat{\lambda}_{(i)}}}. \tag{B.40}$$

Recall that $\xi_{(k)}$ and $\varsigma_{(k)}$ are singular vectors of $\widetilde{R}_{(k)}$, we have

$$\varsigma_{(k)} = \frac{\widetilde{R}_{(k)}\xi_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}, \quad \xi_{(k)} = \frac{\widetilde{R}_{(k)}^\intercal\varsigma_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}. \tag{B.41}$$

Using (B.41), if $\widetilde{R}_{(i)} = \widetilde{\beta}_{(i)}\bar{V} + \widetilde{U}_{(i)}$ for $i < k$, we can write (B.39) as

$$\begin{aligned}
\widetilde{R}_{(k)} = \bar{R}_{[\widehat{I}_k]}\left(\mathbb{I}_T - \sum_{i=1}^{k-1}\xi_{(i)}\xi_{(i)}^\intercal\right) =& \bar{R}_{[\widehat{I}_k]} - \sum_{i=1}^{k-1}\bar{R}_{[\widehat{I}_k]}\xi_{(i)}\frac{\varsigma_{(i)}^\intercal\widetilde{R}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \\
=& \widetilde{\beta}_{[\widehat{I}_k]}\bar{V} + \widetilde{U}_{[\widehat{I}_k]} - \sum_{i=1}^{k-1}\bar{R}_{[\widehat{I}_k]}\xi_{(i)}\frac{\varsigma_{(i)}^\intercal\widetilde{\beta}_{(i)}\bar{V}}{\sqrt{T\widehat{\lambda}_{(i)}}} - \sum_{i=1}^{k-1}\bar{R}_{[\widehat{I}_k]}\xi_{(i)}\frac{\varsigma_{(i)}^\intercal\widetilde{U}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \\
=& \widetilde{\beta}_{(k)}\bar{V} + \widetilde{U}_{(k)}.
\end{aligned}$$

Consequently, by induction, $\widetilde{R}_{(k)} = \widetilde{\beta}_{(k)}\bar{V} + \widetilde{U}_{(k)}$ for $k \leq \tilde{p} + 1$. Similarly, we can write

$$\widetilde{r}_{(k)} = \widetilde{\beta}_{(k)}(\gamma + \bar{v}) + \widetilde{u}_{(k)}, \tag{B.42}$$

where $\widetilde{u}_{(k)}$ is defined by

$$\widetilde{u}_{(k)} := \bar{u}_{[\widehat{I}_k]} - \sum_{i=1}^{k-1} \frac{\bar{R}_{[\widehat{I}_k]}\xi_{(i)}}{\sqrt{T}} \frac{\varsigma_{(i)}^{\intercal}\widetilde{u}_{(i)}}{\sqrt{\widehat{\lambda}_{(i)}}}, \tag{B.43}$$

and $\widetilde{u}_{(1)} = \bar{u}_{[\widehat{I}_1]}$.

Similar representations can be created for $\widetilde{G}_{(k)} := \bar{G}\prod_{i=1}^{k-1}\mathbb{M}_{\widehat{V}_{(i)}^{\intercal}}$. Specifically, we have

$$\begin{aligned}\widetilde{G}_{(k)} := \bar{G}\left(\mathbb{I}_T - \sum_{i=1}^{k-1}\xi_{(i)}\xi_{(i)}^{\intercal}\right) =& \bar{G} - \sum_{i=1}^{k-1}\bar{G}\xi_{(i)}\frac{\varsigma_{(i)}^{\intercal}\widetilde{R}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} = \eta\bar{V} + \bar{Z} - \sum_{i=1}^{k-1}\bar{G}\xi_{(i)}\frac{\varsigma_{(i)}^{\intercal}\widetilde{\beta}_{(i)}\bar{V}}{\sqrt{T\widehat{\lambda}_{(i)}}} - \sum_{i=1}^{k-1}\bar{G}\xi_{(i)}\frac{\varsigma_{(i)}^{\intercal}\widetilde{U}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \\ =& \left(\eta - \sum_{i=1}^{k-1}\bar{G}\xi_{(i)}\frac{\varsigma_{(i)}^{\intercal}\widetilde{\beta}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}\right)\bar{V} + \left(\bar{Z} - \sum_{i=1}^{k-1}\bar{G}\xi_{(i)}\frac{\varsigma_{(i)}^{\intercal}\widetilde{U}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}\right).\end{aligned}$$

Using the following notation

$$\widetilde{\eta}_{(k)} := \eta - \sum_{i=1}^{k-1}\bar{G}\xi_{(i)}\frac{\varsigma_{(i)}^{\intercal}\widetilde{\beta}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}, \quad \text{and} \quad \widetilde{Z}_{(k)} := \bar{Z} - \sum_{i=1}^{k-1}\bar{G}\xi_{(i)}\frac{\varsigma_{(i)}^{\intercal}\widetilde{U}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}, \tag{B.44}$$

$\widetilde{G}_{(k)}$ 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 5, $\widehat{\beta}_{(k)} \in \mathbb{R}^{N\times 1}$ 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 4, we have $\mathrm{P}(\widehat{I}_k = I_k) \to 1$ for $k \le \tilde{p}$ 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 3(ii) and Lemma 4(iii) imply that $\widehat{\lambda}_{(k)} \asymp qN$ and that $|I_k| = qN$. Therefore, the assumptions of Lemmas 6-9 hold.

Since our algorithm stops at $\tilde{p}$, it implies that at most $qN - 1$ test assets satisfy $T^{-1}\left\|\left(\bar{R}_{(\tilde{p}+1)}\right)_{[i]}\bar{G}^{\intercal}\right\|_{\mathrm{MAX}} \ge c$. Consider the test assets in I_0, we have

$$T^{-1}\left\|\widetilde{G}_{(\tilde{p}+1)}\bar{R}_{[I_0]}^{\intercal}\right\| = T^{-1}\left\|\left(\bar{R}_{(\tilde{p}+1)}\right)_{[I_0]}\bar{G}^{\intercal}\right\| \lesssim_{\mathrm{P}} q^{1/2}N^{1/2} + cN_0^{1/2} = o\left(N_0^{1/2}\right), \tag{B.45}$$

where we use the the assumptions $c \to 0$ and $qN/N_0 \to 0$ in the last equation.

Write the left hand side of (B.45) as

$$\widetilde{G}_{(\tilde{p}+1)}\bar{R}_{[I_0]}^{\intercal} = \widetilde{\eta}_{(\tilde{p}+1)}\bar{V}\bar{V}^{\intercal}\beta_{[I_0]} + \widetilde{\eta}_{(\tilde{p}+1)}\bar{V}\bar{U}_{[I_0]}^{\intercal} + \widetilde{Z}_{(\tilde{p}+1)}\bar{V}^{\intercal}\beta_{[I_0]} + \widetilde{Z}_{(\tilde{p}+1)}\bar{U}_{[I_0]}^{\intercal}. \tag{B.46}$$

Using (B.45), (B.46) and Lemma 8(i)(ii), we have

$$\left\| \widetilde{\eta}_{(\tilde{p}+1)} \left(\bar{V}\bar{V}^{\intercal}\beta_{[I_0]} + \bar{V}\bar{U}_{[I_0]}^{\intercal} \right) \right\| = o_p\left(N_0^{1/2} T \right). \tag{B.47}$$

Also, using Assumption A.6, Lemma 1(i) and Weyl's theorem, we have

$$|\sigma_p(\bar{V}\bar{V}^{\intercal}\beta_{[I_0]} + \bar{V}\bar{U}_{[I_0]}^{\intercal}) - \sigma_p(T\beta_{[I_0]})| \leq \left\| \bar{V}\bar{U}_{[I_0]}^{\intercal} \right\| + \left\| T^{-1}\bar{V}\bar{V}^{\intercal} - \mathbb{I}_p \right\| \left\| T\beta_{[I_0]} \right\| \lesssim_p N_0^{1/2} T^{1/2}. \tag{B.48}$$

Since Assumption A.3 implies that $\sigma_p(\beta_{[I_0]}) \asymp N_0^{1/2}$, we have $\sigma_p(\bar{V}\bar{V}^{\intercal}\beta_{[I_0]} + \bar{V}\bar{U}_{[I_0]}^{\intercal}) \asymp_p N_0^{1/2} T$. Using this result, (B.47) and the inequality $\left\| \widetilde{\eta}_{(\tilde{p}+1)} \left(\bar{V}\bar{V}^{\intercal}\beta_{[I_0]} + \bar{V}\bar{U}_{[I_0]}^{\intercal} \right) \right\| \geq \sigma_p(\bar{V}\bar{V}^{\intercal}\beta_{[I_0]} + \bar{V}\bar{U}_{[I_0]}^{\intercal}) \left\| \widetilde{\eta}_{(\tilde{p}+1)} \right\|$, we have $\left\| \widetilde{\eta}_{(\tilde{p}+1)} \right\| \xrightarrow{p} 0$. That is, by definition of $\tilde{\eta}_{(\tilde{p}+1)}$ in (B.44),

$$\left\| \eta - \sum_{i=1}^{\tilde{p}} \bar{G}\xi_{(i)} \frac{\varsigma_{(i)}^{\intercal}\widetilde{\beta}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \right\| = o_p(1). \tag{B.49}$$

Multiplying (B.49) by γ from the right-hand side, we have

$$\left\| \eta\gamma - \sum_{i=1}^{\tilde{p}} \bar{G}\xi_{(i)} \frac{\varsigma_{(i)}^{\intercal}\widetilde{\beta}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \gamma \right\| = o_p(1). \tag{B.50}$$

Recall that our final estimator of γ_g is

$$\widehat{\gamma}_g^{SPCA} = \widehat{\eta}\widehat{\gamma} = \sum_{i=1}^{\tilde{p}} \widehat{\eta}_{(i)}\widehat{\gamma}_{(i)} = \sum_{i=1}^{\tilde{p}} \bar{G}\xi_{(i)} \frac{\varsigma_{(i)}^{\intercal}\widetilde{r}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} = \sum_{i=1}^{\tilde{p}} \bar{G}\xi_{(i)} \frac{\varsigma_{(i)}^{\intercal}\widetilde{\beta}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}(\gamma + \bar{v}) + \sum_{i=1}^{\tilde{p}} \bar{G}\xi_{(i)} \frac{\varsigma_{(i)}^{\intercal}\widetilde{u}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}. \tag{B.51}$$

Combining (B.50) and (B.51), we have

$$\|\eta\gamma - \widehat{\eta}\widehat{\gamma}\| \leq \sum_{i=1}^{\tilde{p}} \left\| \bar{G}\xi_{(i)} \frac{\varsigma_{(i)}^{\intercal}\widetilde{\beta}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \bar{v} \right\| + \sum_{i=1}^{\tilde{p}} \left\| \bar{G}\xi_{(i)} \frac{\varsigma_{(i)}^{\intercal}\widetilde{u}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \right\| + o_p(1). \tag{B.52}$$

Using $\left\| \bar{G} \right\| \lesssim_p T^{1/2}$, Lemma 7(ii), Lemma 9(i) and the assumptions that $qN \to \infty$, we have

$$\left\| \bar{G}\xi_{(i)} \frac{\varsigma_{(i)}^{\intercal}\widetilde{\beta}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \bar{v} \right\| \leq \left\| \bar{G}\xi_{(i)} \right\| \left\| \frac{\varsigma_{(i)}^{\intercal}\widetilde{\beta}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \right\| \|\bar{v}\| = o_p(1),$$

and

$$\left\|\bar{G}\xi_{(i)}\frac{\varsigma_{(i)}^{\intercal}\widetilde{u}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}\right\| \leq \|\bar{G}\xi_{(i)}\|\left\|\frac{\varsigma_{(i)}^{\intercal}\widetilde{u}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}\right\| = o_p(1).$$

Plugging them into (B.52) completes the proof.

B.7 Proof of Theorem 2

To derive the asymptotic distribution, we need a more intricate analysis. As in the proof of Theorem 1, we impose that $\widehat{p} = \tilde{p}$ and $\widehat{I}_k = I_k$, since Lemma 4 shows that both events occur with probability approaching 1.

Recall that in Algorithm 5 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 1,

$$\widehat{\eta}_{(k)} = \frac{\bar{G}\xi_{(k)}}{\sqrt{T}} = \frac{\eta\bar{V}\xi_{(k)}}{\sqrt{T}} + \frac{\bar{Z}\xi_{(k)}}{\sqrt{T}}, \qquad \widehat{\gamma}_{(k)} = \frac{\varsigma_{(k)}^{\intercal}\widetilde{r}_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}} = \frac{\varsigma_{(k)}^{\intercal}\widetilde{\beta}_{(k)}(\gamma+\bar{v})}{\sqrt{\widehat{\lambda}_{(k)}}} + \frac{\varsigma_{(k)}^{\intercal}\widetilde{u}_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}}. \tag{B.53}$$

Denote $H_1 = (h_{11}, \ldots, h_{\widehat{p}1})$, $H_2 = (h_{12}, \ldots, h_{\widehat{p}2})$, where

$$h_{k1} = T^{-1/2}\bar{V}\xi_{(k)}, \quad h_{k2} = \widehat{\lambda}_{(k)}^{-1/2}\widetilde{\beta}_{(k)}^{\intercal}\varsigma_{(k)}. \tag{B.54}$$

Therefore, we can write (B.53) as

$$\widehat{\eta}_{(k)} - \eta h_{k1} = \frac{\bar{Z}\xi_{(k)}}{\sqrt{T}}, \quad \widehat{\gamma}_{(k)} - h_{k2}^{\intercal}(\gamma+\bar{v}) = \frac{\varsigma_{(k)}^{\intercal}\widetilde{u}_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}}. \tag{B.55}$$

Since $\xi_{(k)}$ and $\varsigma_{(k)}$ are the largest singular vectors of $\widetilde{R}_{(k)}$ with the singular value $\sqrt{T\widehat{\lambda}_{(k)}}$, we have

$$\varsigma_{(k)} = \frac{\widetilde{R}_{(k)}\xi_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}, \quad \xi_{(k)} = \frac{\widetilde{R}_{(k)}^{\intercal}\varsigma_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}. \tag{B.56}$$

From (B.56), we have

$$\frac{\bar{Z}\xi_{(k)}}{\sqrt{T}} = \frac{\bar{Z}}{\sqrt{T}}\frac{\widetilde{R}_{(k)}^{\intercal}\varsigma_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}} = \frac{\bar{Z}\bar{V}^{\intercal}}{T}\frac{\widetilde{\beta}_{(k)}^{\intercal}\varsigma_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}} + \frac{\bar{Z}\widetilde{U}_{(k)}^{\intercal}\varsigma_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}} = \frac{\bar{Z}\bar{V}^{\intercal}}{T}h_{k2} + \frac{\bar{Z}\widetilde{U}_{(k)}^{\intercal}\varsigma_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}}.$$

Using Lemma 7(ii) and the assumptions on q, we have

$$\left\| \frac{\bar{Z}\widetilde{U}_{(k)}^{\intercal}\varsigma_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}} \right\| = o_p(T^{-1/2}), \qquad \left\| \frac{\varsigma_{(k)}^{\intercal}\widetilde{u}_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}} \right\| = o_p(T^{-1/2}).$$

Then, along with (B.55) and Lemma 1(vi), the above equations lead to

$$\left\| \widehat{\eta} - \eta H_1 - \frac{ZV}{T} H_2 \right\| = o_p(T^{-1/2}), \tag{B.57}$$

and

$$\|\widehat{\gamma} - H_2^{\intercal}\gamma - H_2^{\intercal}\bar{v}\| = o_p(T^{-1/2}). \tag{B.58}$$

Combining (B.57) and (B.58), with $\|H_1\| \lesssim_p 1$, $\|H_2\| \lesssim_p 1$ from Lemma 9 and Assumptions A.1, A.2, we have

$$\left\| \widehat{\eta}\widehat{\gamma} - \eta H_1 H_2^{\intercal}(\gamma + \bar{v}) - \frac{ZV^{\intercal}}{T} H_2 H_2^{\intercal}\gamma \right\| = o_p(T^{-1/2}). \tag{B.59}$$

As shown in Lemma 3(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 (B.59), using Lemma 9 and the assumptions on q, we have

$$\|H_2^{\intercal}H_2 - \mathbb{I}_p\| \leq \|H_1^{\intercal}H_2 - \mathbb{I}_p\| + \|H_1 - H_2\|\,\|H_2\| \lesssim_p T^{-1/2}. \tag{B.60}$$

Then, for the term $H_2 H_2^{\intercal}$, we have

$$\|H_2 H_2^{\intercal} - \mathbb{I}_p\| = \max_{1\leq i\leq p} |\lambda_i(H_2 H_2^{\intercal}) - 1| = \max_{1\leq i\leq p} |\lambda_i(H_2^{\intercal} H_2) - 1| = \|H_2^{\intercal}H_2 - \mathbb{I}_p\| \lesssim_p T^{-1/2} \tag{B.61}$$

since H_2 is a $p \times p$ matrix.

For the term $H_1 H_2^{\intercal}$, by Lemma 9 and the assumptions on q, we have

$$\|H_1^{\intercal}H_2 - \mathbb{I}_p\| = o_p(T^{-1/2}). \tag{B.62}$$

In addition, we have

$$\sigma_p(H_2)\,\|H_2 H_1^{\intercal} - \mathbb{I}_p\| \leq \|(H_2 H_1^{\intercal} - \mathbb{I}_p)H_2\| = \|H_2(H_1^{\intercal}H_2 - \mathbb{I}_p)\| \leq \|H_2\|\,\|H_1^{\intercal}H_2 - \mathbb{I}_p\|. \tag{B.63}$$

Since (B.60) implies that $\sigma_1(H_2)/\sigma_p(H_2) = \lambda_1(H_2 H_2^{\intercal})^{1/2}/\lambda_p(H_2 H_2^{\intercal})^{1/2} \lesssim_p 1$, (B.62) and (B.63) give

$$\|H_1 H_2^{\intercal} - \mathbb{I}_p\| = \|H_2 H_1^{\intercal} - \mathbb{I}_p\| \leq \frac{\sigma_1(H_2)}{\sigma_p(H_2)}\,\|H_1^{\intercal}H_2 - \mathbb{I}_p\| = o_p(T^{-1/2}). \tag{B.64}$$

Combining (B.59), (B.61), and (B.64), we obtain $\left\|\widehat{\eta\gamma} - \eta(\gamma + \bar{v}) - T^{-1}ZV^\intercal\gamma\right\| = o_p(T^{-1/2})$. Using Delta method and Assumption A.9, it is straightforward to obtain: $\sqrt{T}\left(\widehat{\eta\gamma} - \eta\gamma\right) \xrightarrow{d} \mathcal{N}\left(0, \Phi\right)$, where Φ is as defined in Theorem 2. □

B.8 Proof of Theorem 3

Proof. As in the proof of Theorem 2, we have

$$\left\|\widehat{\eta\gamma} - \eta(\gamma + \bar{v}) - T^{-1}ZV^\intercal\gamma\right\| = o_p(T^{-1/2})$$

from (B.59), (B.61), and (B.64). Together with $\widehat{\alpha}_g = \bar{g} - \widehat{\eta\gamma} = \alpha_g + \eta\gamma + \eta\bar{v} + \bar{z} - \widehat{\eta\gamma}$, we have

$$\left\|\widehat{\alpha}_g - \alpha_g - \bar{z} + T^{-1}ZV^\intercal\gamma\right\| = \left\|\widehat{\eta\gamma} - \eta(\gamma + \bar{v}) - T^{-1}ZV^\intercal\gamma\right\| = o_p(T^{-1/2})$$

Using Delta method and the CLT Assumption in Theorem 3 , we have $\sqrt{T}(\widehat{\alpha}_g - \alpha_g) \xrightarrow{d} \mathcal{N}(0, \widetilde{\Phi})$ where $\widetilde{\Phi}$ is as defined in Theorem 3. □

B.9 Proof of Theorem 4

Proof. As shown in the proof of Theorem 2, 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 2 and (B.58), we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}|m_t - \widehat{m}_t|^2 = \frac{1}{T}\left\|\widehat{V}^\intercal\widehat{\gamma} - V^\intercal\gamma\right\|^2 &= \frac{1}{T}\left\|\sqrt{T}\xi(H_2^\intercal\gamma + O_p(T^{-1/2})) - V^\intercal\gamma\right\|^2 \\ &= \frac{1}{T}\left\|\sqrt{T}\xi H_2^\intercal\gamma - \bar{V}^\intercal\gamma\right\|^2 + O_p\left(T^{-1}\right),\end{aligned} \tag{B.65}$$

where $\xi = (\xi_{(1)}, \ldots, \xi_{(p)})$.

Using (B.56), we can write

$$\sqrt{T}\xi_{(k)}h_{k2}^\intercal = \frac{\widetilde{R}_{(k)}^\intercal\varsigma_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}}h_{k2}^\intercal = \frac{\bar{V}^\intercal\widetilde{\beta}_{(k)}^\intercal\varsigma_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}}h_{k2}^\intercal + \frac{\widetilde{U}_{(k)}^\intercal\varsigma_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}}h_{k2}^\intercal. \tag{B.66}$$

Using Lemma 7(i), Lemma 9(i) and $\widehat{\lambda}_{(k)} \asymp_p |I_k|$, $|I_k| = qN$, we can derive from (B.66) that

$$\sqrt{T}\xi_{(k)}h_{k2}^\intercal = \bar{V}^\intercal h_{k2}h_{k2}^\intercal + O_p\left(q^{-1/2}N^{-1/2}T^{1/2} + T^{-1/2}\right).$$

That is,

$$\sqrt{T}\xi H_2^\intercal = \bar{V}^\intercal H_2 H_2^\intercal + O_p\left(q^{-1/2}N^{-1/2}T^{1/2} + T^{-1/2}\right). \tag{B.67}$$

Therefore, using (B.67), (B.61) and the assumptions on q, we have

$$\begin{aligned}
T^{-1/2}\left\|\sqrt{T}\xi H_2^\intercal\gamma-\bar{V}^\intercal\gamma\right\| &\lesssim_p T^{-1/2}\left\|\bar{V}^\intercal H_2H_2^\intercal-\bar{V}^\intercal\right\|\|\gamma\|+q^{-1/2}N^{-1/2}+T^{-1}\\
&\lesssim_p T^{-1/2}\left\|\bar{V}\right\|\|H_2H_2^\intercal-\mathbb{I}_p\|+q^{-1/2}N^{-1/2}+T^{-1}\\
&\lesssim_p q^{-1/2}N^{-1/2}+T^{-1/2}.
\end{aligned}$$

Therefore, it follows from (B.65) that

$$\frac{1}{T}\sum_{t=1}^{T}|m_t-\widehat{m}_t|^2=\frac{1}{T}\left\|\widehat{V}^\intercal\widehat{\gamma}-V^\intercal\gamma\right\|^2\lesssim_p\frac{1}{T}+\frac{1}{qN}.$$

In light of the assumptions on q, we can choose q such that $qN\gtrsim N_0/\log N_0$, which leads to

$$\frac{1}{T}\sum_{t=1}^{T}|m_t-\widehat{m}_t|^2\lesssim_p\frac{1}{T}+\frac{\log N_0}{N_0}.$$

□

B.10 Proof of Proposition 6

Proof. Write $\widetilde{\beta}=\Sigma_u^{-1/2}\beta\Sigma_v^{1/2}$, then by definition $\widetilde{m}_t$ can be written as

$$\widetilde{m}_t=1-\gamma^\intercal\beta^\intercal\Sigma_r^{-1}(\beta v_t+u_t)=1-\gamma^\intercal\Sigma_v^{-1/2}\tilde{\beta}^\intercal\left(\widetilde{\beta}\widetilde{\beta}^\intercal+\mathbb{I}_N\right)^{-1}(\widetilde{\beta}\Sigma_v^{-1/2}v_t+\Sigma_u^{-1/2}u_t),\tag{B.68}$$

or in matrix form

$$\widetilde{M}=1-\gamma^\intercal\beta^\intercal\Sigma_r^{-1}(\beta V+U)=1-\gamma^\intercal\Sigma_v^{-1/2}\tilde{\beta}^\intercal\left(\widetilde{\beta}\widetilde{\beta}^\intercal+\mathbb{I}_N\right)^{-1}(\widetilde{\beta}\Sigma_v^{-1/2}V+\Sigma_u^{-1/2}U),\tag{B.69}$$

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, $\Lambda^{1/2}=\operatorname{diag}(\widetilde{\lambda}_1^{1/2},\cdots,\widetilde{\lambda}_p^{1/2})$ 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

$$\left\|\widetilde{\beta}^\intercal\left(\widetilde{\beta}\widetilde{\beta}^\intercal+\mathbb{I}_N\right)^{-1}\widetilde{\beta}-\mathbb{I}_p\right\|=\left\|\Gamma^\intercal\Lambda^{1/2}(\Lambda+\mathbb{I}_p)^{-1}\Lambda^{1/2}\Gamma-\mathbb{I}_p\right\|=\left\|\Lambda^{1/2}(\Lambda+\mathbb{I}_p)^{-1}\Lambda^{1/2}-\mathbb{I}_p\right\|\lesssim_p\widetilde{\lambda}_p^{-1},\tag{B.70}$$

and

$$\left\|\widetilde{\beta}^\intercal\left(\widetilde{\beta}\widetilde{\beta}^\intercal+\mathbb{I}_N\right)^{-1}\Sigma_u^{-1/2}U\right\| = \left\|\Gamma^\intercal\Lambda^{1/2}(\Lambda+\mathbb{I}_p)^{-1}B^\intercal\Sigma_u^{-1/2}U\right\| \lesssim_p \left(\widetilde{\lambda}_p^{-1/2}\right)\left\|B^\intercal\Sigma_u^{-1/2}U\right\|. \tag{B.71}$$

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

$$\left\|B^\intercal\Sigma_u^{-1/2}U\right\| \le \left\|B^\intercal\Sigma_u^{-1/2}U\right\|_{\mathrm{F}} \lesssim_p T^{1/2}. \tag{B.72}$$

For the same reason, we have $\left\|\Sigma_v^{-1/2}V\right\| \lesssim_p T^{1/2}$. Then, with Assumption A.11, (B.69), (B.70), (B.71), and (B.72), we have

$$\begin{aligned}\sqrt{\sum_{t=1}^T|m_t-\widetilde{m}_t|^2} &\le \left\|\gamma^\intercal\Sigma_v^{-1/2}\left(\widetilde{\beta}^\intercal\left(\widetilde{\beta}\widetilde{\beta}^\intercal+\mathbb{I}_N\right)^{-1}\widetilde{\beta}-\mathbb{I}_p\right)\Sigma_v^{-1/2}V\right\| + \left\|\gamma^\intercal\Sigma_v^{-1}\widetilde{\beta}^\intercal\left(\widetilde{\beta}\widetilde{\beta}^\intercal+\mathbb{I}_N\right)^{-1}\Sigma_u^{-1/2}U\right\| \\ &\lesssim \left\|\widetilde{\beta}^\intercal\left(\widetilde{\beta}\widetilde{\beta}^\intercal+\mathbb{I}_N\right)^{-1}\widetilde{\beta}-\mathbb{I}_p\right\|\left\|\Sigma_v^{-1/2}V\right\| + \left\|\widetilde{\beta}^\intercal\left(\widetilde{\beta}\widetilde{\beta}^\intercal+\mathbb{I}_N\right)^{-1}\Sigma_u^{-1/2}U\right\| \\ &\lesssim_p T^{1/2}\widetilde{\lambda}_p^{-1/2},\end{aligned}$$

which in turn leads to

$$\frac{1}{T}\sum_{t=1}^T|m_t-\widetilde{m}_t|^2 \lesssim_p \widetilde{\lambda}_p^{-1},$$

where

$$\widetilde{\lambda}_p=\lambda_p\left(\Sigma_v^{1/2}\beta^\intercal\Sigma_u^{-1}\beta\Sigma_v^{1/2}\right) \ge \lambda_p(\beta\Sigma_v\beta^\intercal)\lambda_{\min}(\Sigma_u^{-1}) \asymp_p \lambda_p(\beta^\intercal\beta)\lambda_{\max}^{-1}(\Sigma_u) \gtrsim \lambda_p(\beta^\intercal\beta),$$

which concludes the proof. □

B.11 Proof of Theorem 5(a)

Proof. For Ridge SDF estimator $\widehat{m}_t$, we have

$$\frac{1}{T}\sum_{t=1}^T|m_t-\widehat{m}_t|^2=\frac{1}{T}\left\|\bar{R}^\intercal(\widehat{\Sigma}+\mu\mathbb{I}_N)^{-1}\bar{r}-V^\intercal\gamma\right\|^2. \tag{B.73}$$

Recall that in the proof of Proposition 3, we have a condensed form of SVD on $\bar{R}$:

$$\bar{R}=\sqrt{T}\varsigma\widehat{\Lambda}^{1/2}\xi^\intercal+\sqrt{T}\varsigma_*\widehat{\Lambda}_*^{1/2}\xi_*^\intercal,$$

where $\widehat{\Lambda}^{1/2}$ is the diagonal matrix of the first p singular values of $T^{-1/2}\bar{R}$ and $\widehat{\Lambda} = \operatorname{diag}\{\widehat{\lambda}_1, \dots, \widehat{\lambda}_p\}$, ς, ξ are the corresponding left and right singular vectors, and $\varsigma_* \in \mathbb{R}^{N\times K}$, $\xi_* \in \mathbb{R}^{T\times K}$ 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, (B.73) becomes

$$\begin{aligned}\sqrt{\sum_{t=1}^{T}|m_t - \widehat{m}_t|^2} &= \left\|(\bar{V}^\intercal\beta^\intercal + \bar{U}^\intercal)\varsigma(\widehat{\Lambda}+\mu I)^{-1}\varsigma^\intercal\bar{r} - V^\intercal\gamma + (\bar{V}^\intercal\beta^\intercal + \bar{U}^\intercal)\varsigma_*(\widehat{\Lambda}_*+\mu I)^{-1}\varsigma_*^\intercal\bar{r}\right\| \\ &\leq \left\|\bar{V}^\intercal\beta^\intercal\varsigma(\widehat{\Lambda}+\mu I)^{-1}\varsigma^\intercal\beta\gamma - \bar{V}^\intercal\gamma\right\| + \left\|\bar{V}^\intercal\beta^\intercal\varsigma(\widehat{\Lambda}+\mu I)^{-1}\varsigma^\intercal(\beta\bar{v}+\bar{u})\right\| \\ &\quad + \left\|\bar{U}^\intercal\varsigma(\widehat{\Lambda}+\mu I)^{-1}\varsigma^\intercal\bar{r}\right\| + \left\|\bar{V}^\intercal\beta^\intercal\varsigma_*(\widehat{\Lambda}_*+\mu I)^{-1}\varsigma_*^\intercal\bar{r}\right\| \\ &\quad + \left\|\bar{U}^\intercal\varsigma_*(\widehat{\Lambda}_*+\mu I)^{-1}\varsigma_*^\intercal\bar{r}\right\| + \left\|V^\intercal\gamma - \bar{V}^\intercal\gamma\right\| \end{aligned} \tag{B.74}$$

We analyze these terms one-by-one. Firstly, we consider the properties of ς and ξ. Let ς_k and ξ_k denote the kth columns of ς and ξ, respectively. Note that ς_k and ξ_k can be regarded as the $\varsigma_{(k)}$ and $\xi_{(k)}$ in our SPCA procedure with $I_k = [N]$, where ς_k and ξ_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 (B.54), we define

$$\tilde{h}_{k1} = T^{-1/2}\bar{V}\xi_k, \quad \tilde{h}_{k2} = \widehat{\lambda}_k^{-1/2}\beta^\intercal\varsigma_k, \tag{B.75}$$

and $\tilde{H}_1 = (\tilde{h}_{11}, \dots, \tilde{h}_{p1})$, $\tilde{H}_2 = (h_{12}, \dots, \tilde{h}_{p2})$. Using Lemma 14, we can obtain

$$\left\|\tilde{H}_1\tilde{H}_2^\intercal - \mathbb{I}_p\right\| \lesssim_p T^{-1} + \lambda_p^{-1}(T^{-1}N+1), \quad \left\|\tilde{H}_1 - \tilde{H}_2\right\| \lesssim_p T^{-1/2} + \lambda_p^{-1}(T^{-1}N+1). \tag{B.76}$$

Using (B.76) and Lemma 14(i), we have $\left\|\tilde{H}_2\tilde{H}_2^\intercal - \mathbb{I}_p\right\| \leq \left\|\tilde{H}_1\tilde{H}_2^\intercal - \mathbb{I}_p\right\| + \left\|\tilde{H}_1 - \tilde{H}_2\right\| \left\|\tilde{H}_2\right\| \lesssim_p T^{-1/2} + \lambda_p^{-1}(T^{-1}N+1)$, which, by (B.75), is equivalent to

$$\left\|\beta^\intercal\varsigma\widehat{\Lambda}^{-1}\varsigma^\intercal\beta - \mathbb{I}_p\right\| \lesssim_p \frac{1}{\sqrt{T}} + \frac{N+T}{T\lambda_p}. \tag{B.77}$$

Consequently, with Lemma 11 and $\left\|\beta^\intercal\varsigma\widehat{\Lambda}^{-1/2}\right\| = \left\|\tilde{H}_2\right\| \lesssim_p 1$, we have

$$\begin{aligned}\left\|\beta^\intercal\varsigma\left(\widehat{\Lambda}+\mu I\right)^{-1}\varsigma^\intercal\beta - \mathbb{I}_p\right\| &\leq \left\|\beta^\intercal\varsigma\widehat{\Lambda}^{-1/2}\left(\widehat{\Lambda}^{1/2}\left(\widehat{\Lambda}+\mu I\right)^{-1}\widehat{\Lambda}^{1/2} - \mathbb{I}_p\right)\widehat{\Lambda}^{-1/2}\varsigma^\intercal\beta\right\| + \left\|\beta^\intercal\varsigma\widehat{\Lambda}^{-1}\varsigma^\intercal\beta - \mathbb{I}_p\right\| \\ &\leq \left\|\beta^\intercal\varsigma\widehat{\Lambda}^{-1/2}\right\|^2 \left\|\widehat{\Lambda}^{1/2}\left(\widehat{\Lambda}+\mu I\right)^{-1}\widehat{\Lambda}^{1/2} - \mathbb{I}_p\right\| + \left\|\beta^\intercal\varsigma\widehat{\Lambda}^{-1}\varsigma^\intercal\beta - \mathbb{I}_p\right\| \\ &\lesssim_p \frac{1}{\sqrt{T}} + \frac{N+T}{T\lambda_p} + \frac{\mu}{\lambda_p}, \end{aligned} \tag{B.78}$$

where we use $\left\|\widehat{\Lambda}^{1/2}\left(\widehat{\Lambda}+\mu I\right)^{-1}\widehat{\Lambda}^{1/2}-\mathbb{I}_p\right\| = \max_{j\leq p}(\widehat{\lambda}_j+\mu)^{-1}\mu \lesssim_p \lambda_p^{-1}\mu$ in the last step.

With $\left\|\bar{V}\right\| \lesssim_p T^{1/2}$ from Lemma 1, it implies from (B.78) that the first term in (B.74) can be bounded:

$$\left\|\bar{V}^\intercal\beta^\intercal\varsigma(\widehat{\Lambda}+\mu I)^{-1}\varsigma^\intercal\beta\gamma-\bar{V}^\intercal\gamma\right\| \lesssim_p 1+\frac{N+T}{\sqrt{T}\lambda_p}+\frac{\mu\sqrt{T}}{\lambda_p}.$$

For the second term in (B.74), using Lemma 11, we have

$$\left\|\bar{V}^\intercal\beta^\intercal\varsigma(\widehat{\Lambda}+\mu I)^{-1}\varsigma^\intercal(\beta\bar{v}+\bar{u})\right\| \leq \left\|\bar{V}\right\|\left\|\beta^\intercal\varsigma\widehat{\Lambda}^{-1/2}\right\|\left\|\widehat{\Lambda}^{1/2}(\widehat{\Lambda}+\mu I)^{-1}\right\|\left\|\beta\bar{v}+\bar{u}\right\| \lesssim_p \sqrt{\frac{N}{\lambda_p}}. \quad \text{(B.79)}$$

Next, recall that ς_* and ξ_* are singular vectors of $\bar{R}$, we have

$$\bar{V}^\intercal\beta^\intercal\varsigma_*+\bar{U}^\intercal\varsigma_*=\bar{R}^\intercal\varsigma_*=\sqrt{T}\xi_*\widehat{\Lambda}_*^{1/2}. \quad \text{(B.80)}$$

By Weyl's theorem and Assumption A.4, we have

$$|\sigma_j(T^{-1/2}\bar{R})-\sigma_j(T^{-1/2}\beta\bar{V})| \leq T^{-1/2}\left\|\bar{R}-\beta\bar{V}\right\| = T^{-1/2}\left\|\bar{U}\right\| \lesssim_p \sqrt{\frac{N}{T}}+1, \quad \text{(B.81)}$$

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

$$\left\|\widehat{\Lambda}_*^{1/2}\right\| = \sigma_{p+1}(T^{-1/2}\bar{R}) \lesssim_p \sqrt{\frac{N}{T}}+1. \quad \text{(B.82)}$$

Left multiplying (B.80) by $\bar{V}$, we obtain

$$\bar{V}\bar{V}^\intercal\beta^\intercal\varsigma_* = \sqrt{T}\bar{V}\xi_*\widehat{\Lambda}_*^{1/2}-\bar{V}\bar{U}^\intercal\varsigma_*. \quad \text{(B.83)}$$

Together with (B.82) and Assumption A.6, we have

$$\left\|\beta^\intercal\varsigma_*\right\| \leq \left\|(\bar{V}\bar{V}^\intercal)^{-1}\right\|\left(\sqrt{T}\left\|\bar{V}\right\|\left\|\widehat{\Lambda}_*^{1/2}\right\|+\left\|\bar{V}\bar{U}^\intercal\right\|\right) \lesssim_p \sqrt{\frac{N}{T}}+1, \quad \text{(B.84)}$$

and consequently,

$$\left\|\varsigma_*^\intercal\bar{r}\right\| \leq \left\|\varsigma_*^\intercal\beta\right\|\left\|\gamma+\bar{v}\right\|+\left\|\varsigma_*^\intercal\bar{u}\right\| \lesssim_p \sqrt{\frac{N}{T}}+1. \quad \text{(B.85)}$$

Using (B.84), (B.85), Lemma 13(iv) and $\left\|\bar{U}\right\| \lesssim_p N^{1/2}+T^{1/2}$, we have

$$\left\|\beta^\intercal\varsigma_*(\widehat{\Lambda}_*+\mu I)^{-1}\varsigma_*^\intercal\bar{r}\right\| \leq \left\|\beta^\intercal\varsigma_*\right\|\left\|(\widehat{\Lambda}_*+\mu I)^{-1}\right\|\left\|\varsigma_*^\intercal\bar{r}\right\| \lesssim_p \frac{N+T}{\mu T}, \quad \text{(B.86)}$$

and

$$\left\|\bar{U}^\intercal \varsigma_*(\widehat{\Lambda}_* + \mu I)^{-1}\varsigma_*^\intercal \bar{r}\right\| \leq \left\|\bar{U}\right\| \left\|(\widehat{\Lambda}_* + \mu I)^{-1}\right\| \left\|\varsigma_*^\intercal \bar{r}\right\| \lesssim_p \frac{N+T}{\mu\sqrt{T}}. \tag{B.87}$$

Using Lemma 13(iii), we have

$$\left\|\widehat{\Lambda}^{-1/2}\varsigma^\intercal \bar{r}\right\| \lesssim_p \left\|\widehat{\Lambda}^{-1/2}\varsigma^\intercal \beta\right\| + \left\|\widehat{\Lambda}^{-1/2}\varsigma^\intercal \bar{u}\right\| \lesssim_p 1 + \frac{N+T}{T\lambda_p} \lesssim_p 1,$$

where we use $\left\|\widehat{\Lambda}^{-1/2}\varsigma^\intercal \beta\right\| = \left\|\tilde{H}_2\right\| \lesssim_p 1$. Then, with Lemma 13(iv), we have

$$\left\|\bar{U}^\intercal \varsigma(\widehat{\Lambda} + \mu I)^{-1}\varsigma^\intercal \bar{r}\right\| \leq \left\|\bar{U}^\intercal \varsigma\right\| \left\|(\widehat{\Lambda} + \mu I)^{-1}\widehat{\Lambda}^{1/2}\right\| \left\|\widehat{\Lambda}^{-1/2}\varsigma^\intercal \bar{r}\right\| \lesssim_p \sqrt{\frac{T}{\lambda_p}} + \frac{N+T}{\sqrt{T}\lambda_p}. \tag{B.88}$$

Plugging (B.78), (B.79), (B.86), (B.87) and (B.88) into (B.74) and using $\left\|\bar{V} - V\right\| \lesssim_p 1$, we obtain

$$\frac{1}{T}\sum_{t=1}^{T}|m_t - \widehat{m}_t|^2 \lesssim_p \frac{\mu^2}{\lambda_p^2} + \frac{1}{T} + \frac{N+T}{T\lambda_p} + \frac{N^2+T^2}{\mu^2 T^2}.$$

With $\mu^2 \asymp T^{-1}\lambda_p(N+T)$, we achieve the best rate from the above bound:

$$\frac{1}{T}\sum_{t=1}^{T}|m_t - \widehat{m}_t|^2 \lesssim_p \frac{1}{T} + \frac{N+T}{T\lambda_p}.$$

□

B.12 Proof of Theorem 5(b)

Proof. i. (Slow rate) Note that (13) is equivalent to a constrained optimization problem:

$$\widehat{b} = \arg\min_b \left\|\widehat{\Sigma}^{-1/2}\bar{r} - \widehat{\Sigma}^{1/2}b\right\|^2, \quad \text{subject to } \|b\|_1 \leq \mu,$$

for some tuning parameter μ. This implies that the vector of the true SDF loadings, b, satisfies that

$$\left\|\widehat{\Sigma}^{-1/2}\bar{r} - \widehat{\Sigma}^{1/2}\widehat{b}\right\|^2 \leq \left\|\widehat{\Sigma}^{-1/2}\bar{r} - \widehat{\Sigma}^{1/2}b\right\|^2 \quad \text{and} \quad \left\|\widehat{b}\right\|_1 \leq \mu, \ \text{for some}\, \mu \geq s.$$

Equivalently, expanding the left- and right-hand sides of the above we have

$$\widehat{b}^\intercal \widehat{\Sigma}\widehat{b} - b^\intercal \widehat{\Sigma} b \leq 2(\widehat{b} - b)^\intercal \bar{r},$$

which leads to

$$(\widehat{b}-b)^\intercal\widehat{\Sigma}(\widehat{b}-b) \leq 2(\widehat{b}-b)^\intercal(\bar{r}-\widehat{\Sigma}b) \leq 2\left\|\widehat{b}-b\right\|_1\left\|\bar{r}-\widehat{\Sigma}b\right\|_\infty .$$

With a tuning parameter $\mu \asymp s$, we have

$$(\widehat{b}-b)^\intercal\widehat{\Sigma}(\widehat{b}-b) \lesssim s\left\|\bar{r}-\widehat{\Sigma}b\right\|_\infty . \tag{B.89}$$

With Lemma 15, we have

$$\left\|\widehat{\Sigma}^{1/2}(\widehat{b}-b)\right\|^2 \lesssim_p s\sqrt{\frac{\log N}{T}}. \tag{B.90}$$

Therefore, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^T\|\widehat{m}_t-\widetilde{m}_t\|^2 =&\frac{1}{T}\sum_{t=1}^T\left\|\widehat{b}^\intercal(r_t-\bar{r})-b^\intercal(r_t-\mathrm{E}(r_t))\right\|^2\\ \leq&\frac{2}{T}\sum_{t=1}^T\left\|(\widehat{b}-b)^\intercal(r_t-\bar{r})\right\|^2+\frac{2}{T}\sum_{t=1}^T\|b^\intercal(\bar{r}-\mathrm{E}(r_t))\|^2\\ \leq&2\left\|\widehat{\Sigma}^{1/2}(\widehat{b}-b)\right\|^2+2\|b\|_1^2\|\bar{r}-\mathrm{E}(r_t)\|_\infty^2\\ \lesssim_p& s\sqrt{\frac{\log N}{T}}+s^2\frac{\log N}{T}.\end{aligned}$$

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

$$b^\intercal\widehat{\Sigma}b-2b^\intercal\bar{r}+b^\intercal\widehat{\Sigma}b+\mu\|b\|_1 \geq \widehat{b}^\intercal\widehat{\Sigma}\widehat{b}-2\widehat{b}^\intercal\bar{r}+\widehat{b}^\intercal\widehat{\Sigma}\widehat{b}+\mu\|\widehat{b}\|_1. \tag{B.91}$$

Rewrite (B.91) as

$$(\widehat{b}-b)^\intercal\widehat{\Sigma}(\widehat{b}-b) \leq 2(\widehat{b}-b)^\intercal(\bar{r}-\widehat{\Sigma}b)+\mu(\|b\|_1-\|\widehat{b}\|_1). \tag{B.92}$$

If $\mu \geq 4\left\|\bar{r}-\widehat{\Sigma}b\right\|_\infty$, (B.92) becomes

$$\begin{aligned}\left\|\widehat{\Sigma}^{1/2}(\widehat{b}-b)\right\|^2 \leq&2\left\|\widehat{b}-b\right\|_1\left\|\bar{r}-\widehat{\Sigma}b\right\|_\infty+\mu(\|b\|_1-\|\widehat{b}\|_1)\\ \leq&\frac{1}{2}\mu\left\|\widehat{b}-b\right\|_1+\mu(\|b\|_1-\|\widehat{b}\|_1).\end{aligned} \tag{B.93}$$

Let J denote the support of $\widehat{b}$, then (B.93) can be written as

$$\left\|\widehat{\Sigma}^{1/2}(\widehat{b}-b)\right\|^2 \leq\frac{1}{2}\mu\left\|\widehat{b}_J-b_J\right\|_1+\frac{1}{2}\mu\left\|\widehat{b}_{J^c}\right\|_1+\mu\left\|\widehat{b}_J-b_J\right\|_1-\mu\left\|\widehat{b}_{J^c}\right\|_1$$

$$=\frac{3}{2}\mu\left\|\widehat{b}_J-b_J\right\|_1-\frac{1}{2}\mu\left\|\widehat{b}_{J^c}\right\|_1. \tag{B.94}$$

Define $b^*=\widehat{b}-b$, then (B.94) implies that $3\left\|b^*_J\right\|_1\geq\left\|b^*_{J^c}\right\|_1$, and we have

$$\frac{b^{*\intercal}(\Sigma-\widehat{\Sigma})b^*}{\|b^*\|^2}\leq\left\|\Sigma-\widehat{\Sigma}\right\|_{MAX}\frac{\|b^*\|_1^2}{\|b^*\|^2}\lesssim_p\sqrt{\frac{\log N}{T}}\left(\frac{4\left\|b^*_J\right\|_1}{\left\|b^*_J\right\|}\right)^2\lesssim_p|J|\sqrt{\frac{\log N}{T}}.$$

Consequently, with the assumption $|J|\sqrt{\log N/T}\to0$ and $\lambda_{\min}(\Sigma)\gtrsim1$, we have

$$\frac{b^{*\intercal}\widehat{\Sigma}b^*}{\|b^*\|^2}=\frac{b^{*\intercal}\Sigma b^*}{\|b^*\|^2}+\frac{b^{*\intercal}(\Sigma-\widehat{\Sigma})b^*}{\|b^*\|^2}\gtrsim_p1.$$

Therefore, we have

$$\left\|\widehat{\Sigma}^{1/2}(\widehat{b}-b)\right\|^2=b^{*\intercal}\widehat{\Sigma}b^*\gtrsim_p\|b^*\|^2\geq\|b^*_J\|^2\geq|J|^{-1}\|b^*_J\|_1^2=|J|^{-1}\left\|\widehat{b}_J-b_J\right\|_1^2. \tag{B.95}$$

Plugging (B.95) into (B.94), we have

$$\left\|\widehat{\Sigma}^{1/2}(\widehat{b}-b)\right\|^2\leq\frac{3}{2}\mu\left\|\widehat{b}_J-b_J\right\|_1\lesssim_p\mu|J|^{1/2}\left\|\widehat{\Sigma}^{1/2}(\widehat{b}-b)\right\|.$$

Thus,

$$\left\|\widehat{\Sigma}^{1/2}(\widehat{b}-b)\right\|^2\lesssim_p\mu^2|J|. \tag{B.96}$$

Choosing $\mu=4\left\|\bar{r}-\widehat{\Sigma}b\right\|_\infty$ and by Lemma 15, we obtain

$$\left\|\widehat{\Sigma}^{1/2}(\widehat{b}-b)\right\|^2\lesssim_p|J|\frac{\log N}{T}. \tag{B.97}$$

Similar to the slow rate case, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^T\|\widehat{m}_t-\widetilde{m}_t\|^2=&\frac{1}{T}\sum_{t=1}^T\left\|\widehat{b}^\intercal(r_t-\bar{r})-b^\intercal(r_t-\mathrm{E}(r_t))\right\|^2\\ \leq&\frac{2}{T}\sum_{t=1}^T\left\|(\widehat{b}-b)^\intercal(r_t-\bar{r})\right\|^2+\frac{2}{T}\sum_{t=1}^T\|b^\intercal(\bar{r}-\mathrm{E}(r_t))\|^2\\ \leq&2\left\|\widehat{\Sigma}^{1/2}(\widehat{b}-b)\right\|^2+2\left\|b^\intercal(\bar{r}-\mathrm{E}(r_t))\right\|^2\\ \lesssim_p&\|b\|_0\frac{\log N}{T}.\end{aligned}$$

□

B.13 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 4, 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 4.

Lemma 1. *Under Assumptions A.1-A.7, for any $I \subset [N]$, we have the following results:*

(i) $\left\|T^{-1}\bar{V}\bar{V}^\intercal - \mathbb{I}_p\right\| \lesssim_p T^{-1/2}$.

(ii) $\left\|\left(\beta_{[I]}^\intercal\beta_{[I]}\right)^{-\frac{1}{2}}\beta_{[I]}^\intercal\bar{U}_{[I]}\right\| \lesssim_p T^{1/2}$.

(iii) $\left\|\left(\beta_{[I]}^\intercal\beta_{[I]}\right)^{-\frac{1}{2}}\beta_{[I]}^\intercal\bar{U}_{[I]}\bar{V}^\intercal\right\| \lesssim_p T^{1/2}$, $\left\|\left(\beta_{[I]}^\intercal\beta_{[I]}\right)^{-\frac{1}{2}}\beta_{[I]}^\intercal\bar{U}_{[I]}\bar{Z}^\intercal\right\| \lesssim_p T^{1/2}$.

(iv) $\|\bar{U}\|_{\text{MAX}} \lesssim_p (\log NT)^{1/2}$, $\|\bar{U}\bar{V}^\intercal\|_{\text{MAX}} \lesssim_p (\log N)^{1/2}T^{1/2}$, $\|\bar{U}\bar{Z}^\intercal\|_{\text{MAX}} \lesssim_p (\log N)^{1/2}T^{1/2}$.

(v) $\|\bar{U}_{[I]}\| \lesssim_p |I|^{1/2} + T^{1/2}$, $\|\bar{U}_{[I]}\bar{V}^\intercal\| \lesssim_p |I|^{1/2}T^{1/2}$, $\|\bar{U}_{[I]}\bar{Z}^\intercal\| \lesssim_p |I|^{1/2}T^{1/2}$.

(vi) $\|\bar{V}\| \lesssim_p T^{1/2}$, $\|\bar{Z}\| \lesssim_p T^{1/2}$, $\|\bar{V}\bar{Z}^\intercal\| \lesssim_p T^{1/2}$, $\|\bar{V}\bar{Z}^\intercal - VZ^\intercal\| \lesssim_p 1$

Proof. (i) Using Assumption A.1, we have

$$\left\|\frac{\bar{V}\bar{V}^\intercal}{T} - \mathbb{I}_p\right\| \le \left\|\frac{VV^\intercal}{T} - \mathbb{I}_p\right\| + \left\|\frac{V\iota_T\iota_T^\intercal V^\intercal}{T^2}\right\| = \left\|\frac{VV^\intercal}{T} - \mathbb{I}_p\right\| + \|\bar{v}\|^2 \lesssim_p T^{-1/2}.$$

(ii) Using Assumption A.5, we have

$$\left\|\left(\beta_{[I]}^\intercal\beta_{[I]}\right)^{-\frac{1}{2}}\beta_{[I]}^\intercal\bar{U}_{[I]}\right\| \le \left\|\left(\beta_{[I]}^\intercal\beta_{[I]}\right)^{-\frac{1}{2}}\beta_{[I]}^\intercal U_{[I]}\right\| + T^{-1}\left\|\left(\beta_{[I]}^\intercal\beta_{[I]}\right)^{-\frac{1}{2}}\beta_{[I]}^\intercal U_{[I]}\iota_T\iota_T^\intercal\right\| \lesssim_p T^{1/2}.$$

(iii) By Assumptions A.1, A.5 and A.6, we have

$$\begin{aligned}\left\|\left(\beta_{[I]}^\intercal\beta_{[I]}\right)^{-\frac{1}{2}}\beta_{[I]}^\intercal\bar{U}_{[I]}\bar{V}^\intercal\right\| &\le \left\|\left(\beta_{[I]}^\intercal\beta_{[I]}\right)^{-\frac{1}{2}}\beta_{[I]}^\intercal U_{[I]}V^\intercal\right\| + T^{-1}\left\|\left(\beta_{[I]}^\intercal\beta_{[I]}\right)^{-\frac{1}{2}}\beta_{[I]}^\intercal U_{[I]}\iota_T\iota_T^\intercal V\right\| \\ &\le \left\|\left(\beta_{[I]}^\intercal\beta_{[I]}\right)^{-\frac{1}{2}}\beta_{[I]}^\intercal U_{[I]}V^\intercal\right\| + \left\|\left(\beta_{[I]}^\intercal\beta_{[I]}\right)^{-\frac{1}{2}}\beta_{[I]}^\intercal U_{[I]}\iota_T\right\|\|\bar{v}\| \lesssim_p T^{1/2}.\end{aligned}$$

Replacing $\bar{V}$ by $\bar{Z}$ in the above proof, with Assumptions A.2, A.5 and A.7, we also have

$$\left\|\left(\beta_{[I]}^\intercal\beta_{[I]}\right)^{-\frac{1}{2}}\beta_{[I]}^\intercal\bar{U}_{[I]}\bar{Z}^\intercal\right\| \lesssim_p T^{1/2}.$$

(iv) Using Assumption A.4, we have

$$\|\bar{U}\|_{\text{MAX}} \le \|U\|_{\text{MAX}} + T^{-1}\left\|U\iota_T\iota_T^\intercal\right\|_{\text{MAX}} \le \|U\|_{\text{MAX}} + \|\bar{u}\|_{\text{MAX}}\|\iota_T\| \lesssim_p (\log N)^{1/2} + (\log T)^{1/2}.$$

Using Assumptions A.1, A.4, A.6, we have

$$\left\|\bar{U}\bar{V}^\intercal\right\|_{\mathrm{MAX}} \le \|UV^\intercal\|_{\mathrm{MAX}} + T^{-1}\left\|U\iota_T\iota_T^\intercal V^\intercal\right\|_{\mathrm{MAX}} \le \|UV^\intercal\|_{\mathrm{MAX}} + T\|\bar{u}\|_{\mathrm{MAX}}\|\bar{v}\| \lesssim_p (\log N)^{1/2}T^{1/2}.$$

Replacing $\bar{V}$ by $\bar{Z}$ in the above proof, with Assumptions A.2, A.4 and A.7, we also have

$$\left\|\bar{U}\bar{Z}^\intercal\right\|_{\mathrm{MAX}} \lesssim_p (\log N)^{1/2}T^{1/2}.$$

(v) Using Assumption A.4 , we have

$$\left\|\bar{U}_{[I]}\right\| \le \left\|U_{[I]}\right\| + T^{-1}\left\|U_{[I]}\iota_T\iota_T^\intercal\right\| \le \left\|U_{[I]}\right\| + \left\|\bar{u}_{[I]}\right\|\|\iota_T\| \lesssim_p |I|^{1/2} + T^{1/2}.$$

Using Assumptions A.1, A.4, A.6, we have

$$\left\|\bar{U}_{[I]}\bar{V}^\intercal\right\| \le \left\|U_{[I]}V^\intercal\right\| + T^{-1}\left\|U_{[I]}\iota_T\iota_T^\intercal V^\intercal\right\| \le \left\|U_{[I]}V^\intercal\right\| + T\left\|\bar{u}_{[I]}\right\|\|\bar{v}\| \lesssim_p |I|^{1/2}T^{1/2}.$$

Replacing $\bar{V}$ by $\bar{Z}$ in the above proof, with Assumptions A.2, A.4 and A.7, we also have

$$\left\|\bar{U}_{[I]}\bar{Z}^\intercal\right\| \lesssim_p |I|^{1/2}T^{1/2}.$$

(vi) Using Assumption A.1, we have

$$\left\|\bar{V}\right\| \le \|V\| + T^{-1}\left\|V\iota_T\iota_T^\intercal\right\| \le \|V\| + \|\bar{v}\|\|\iota_T\| \lesssim_p T^{1/2}.$$

Using Assumption A.2, we have

$$\left\|\bar{Z}\right\| \le \|Z\| + T^{-1}\left\|Z\iota_T\iota_T^\intercal\right\| \le \|Z\| + \|\bar{z}\|\|\iota_T\| \lesssim_p T^{1/2}.$$

Using Assumptions A.1 and A.2, we have

$$\left\|\bar{V}\bar{Z}^\intercal\right\| \le \|VZ\| + T^{-1}\left\|V\iota_T\iota_T^\intercal Z\right\| \le \|V\| + T\|\bar{v}\|\|\bar{z}\| \lesssim_p T^{1/2},$$

and

$$\left\|\bar{V}\bar{Z}^\intercal - VZ^\intercal\right\| = \left\|T^{-1}V\iota_T\iota_T^\intercal Z\right\| = T\|\bar{v}\|\|\bar{z}\| \lesssim_p 1.$$

□

Lemma 2. *The singular vectors $\xi_{(k)}$s we obtain from Algorithm 5 satisfy $\xi_{(j)}^\intercal\xi_{(k)} = \delta_{jk}$ for $j, k \le \widehat{p}$.*

Proof. If $j = k$, this result holds from the definition of $\xi_{(k)}$. If $j < k$, recall that $\widetilde{R}_{(k)}$ is defined in

(B.39) and $\xi_{(k)}$ is the first right singular vector of $\widetilde{R}_{(k)}$, we have

$$\widetilde{R}_{(k)} = \bar{R}_{[I_k]} \prod_{i<k} \left(\mathbb{I}_T - \xi_{(i)} \xi_{(i)}^{\intercal} \right) \qquad \text{and} \qquad \xi_{(k)} = \arg\max_{\alpha} \frac{\left\| \widetilde{R}_{(k)} \alpha \right\|}{\|\alpha\|}.$$

If $\xi_{(k)}^{\intercal} \xi_{(j)} = c_0 \neq 0$ for some $j < k$, then

$$\left\| \widetilde{R}_{(k)} (\xi_{(k)} - c_0 \xi_{(j)}) \right\| = \left\| \widetilde{R}_{(k)} \xi_{(k)} - c_0 \widetilde{R}_{(k)} \xi_{(j)} \right\| = \left\| \widetilde{R}_{(k)} \xi_{(k)} \right\|, \tag{B.98}$$

since the definition of $\widetilde{R}_{(k)}$ implies that $\widetilde{R}_{(k)} \xi_{(j)} = 0$ for $j < k$.

On the other hand, since $\xi_{(k)}^{\intercal} \xi_{(j)} = c_0 \neq 0$, we have $(\xi_{(k)} - c_0 \xi_{(j)})^{\intercal} \xi_{(j)} = 0$, and consequently,

$$\left\| \xi_{(k)} \right\|^2 = \left\| \xi_{(k)} - c_0 \xi_{(j)} \right\|^2 + \left\| c_0 \xi_{(j)} \right\|^2 > \left\| \xi_{(k)} - c_0 \xi_{(j)} \right\|^2. \tag{B.99}$$

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 (B.98) and (B.99), we have

$$\left\| \widetilde{R}_{(k)} \right\| = \frac{\left\| \widetilde{R}_{(k)} \xi_{(k)} \right\|}{\left\| \xi_{(k)} \right\|} \leq \frac{\left\| \widetilde{R}_{(k)} (\xi_{(k)} - c_0 \xi_{(j)}) \right\|}{\left\| \xi_{(k)} - c_0 \xi_{(j)} \right\|},$$

which contradicts with the definition of $\xi_{(k)}$. Therefore, $\xi_{(k)}^{\intercal} \xi_{(j)} = 0$ for $j < k$. This completes the proof. □

Lemma 3. *Under Assumption A.3, if $c \to 0$, $qN/N_0 \to 0$ then b_k, $\beta_{(k)}$ and $\tilde{p}$ defined in Section A 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 2, 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

$$\frac{1}{|I_k|} \sum_{i \in I_k} \left\| \beta_{[i]} \prod_{j<k} \mathbb{M}_{b_j} \eta^{\intercal} \right\|_{\text{MAX}}^2 \geq \frac{1}{N_0} \sum_{i \in I_0} \left\| \beta_{[i]} \prod_{j<k} \mathbb{M}_{b_j} \eta^{\intercal} \right\|_{\text{MAX}}^2. \tag{B.100}$$

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

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

and

$$\|A\|^2 \leq \|A\|_{\text{F}}^2 \leq d \, \|A\|^2 ,$$

we thereby have

$$\|A\|^2 \asymp \sum_{i \in I} \left\| A_{[i]} \right\|_{\text{MAX}}^2 . \tag{B.101}$$

Using this result, (B.100) 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

$$\frac{1}{\sqrt{|I_k|}} \left\| \beta_{(k)} \right\| \left\| \prod_{j<k} \mathbb{M}_{b_j} \eta^\intercal \right\| \geq \frac{1}{\sqrt{|I_k|}} \left\| \beta_{[I_k]} \prod_{j<k} \mathbb{M}_{b_j} \eta^\intercal \right\| \gtrsim \frac{1}{\sqrt{N_0}} \left\| \beta_{[I_0]} \prod_{j<k} \mathbb{M}_{b_j} \eta^\intercal \right\| \geq \frac{1}{\sqrt{N_0}} \sigma_p \left(\beta_{[I_0]} \right) \left\| \prod_{j<k} \mathbb{M}_{b_j} \eta^\intercal \right\| , \tag{B.102}$$

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 A.3, (B.102) leads to $\left\| \beta_{(k)} \right\| \gtrsim |I_k|^{1/2}$. In addition, $\|\beta\|_{\text{MAX}} \lesssim 1$ from Assumption A.3 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 A. 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 (B.101) and the assumptions $c \to 0$, $qN/N_0 \to 0$. With $\sigma_p(\beta_{[I_0]}) \gtrsim \sqrt{N_0}$ from

Assumption A.3, we have

$$\left\|\eta \prod_{j\le\tilde{p}} \mathbb{M}_{b_j}\right\| \le \sigma_p(\beta_{[I_0]})^{-1}\left\|\beta_{[I_0]} \prod_{j\le\tilde{p}} \mathbb{M}_{b_j}\eta^\intercal\right\| = o(1). \tag{B.103}$$

If $\tilde{p} \le p-1$, using (i), we have

$$\eta \prod_{j\le\tilde{p}} \mathbb{M}_{b_j} = \eta - \eta \sum_{j\le\tilde{p}} b_j b_j^\intercal,$$

which implies that

$$\sigma_p(\eta) \le \sigma_1\left(\eta \prod_{j\le\tilde{p}} \mathbb{M}_{b_j}\right) + \sigma_p\left(\eta \sum_{j\le\tilde{p}} b_j b_j^\intercal\right). \tag{B.104}$$

Since

$$\text{Rank}\left(\eta \sum_{j\le\tilde{p}} b_j b_j^\intercal\right) \le \tilde{p} \le p-1, \tag{B.105}$$

we have $\sigma_p\left(\eta \sum_{j\le\tilde{p}} b_j b_j^\intercal\right) \le 0$ and thus (B.104) and (B.103) lead to $\sigma_p(\eta) \lesssim \sigma_1\left(\eta \prod_{j\le\tilde{p}} \mathbb{M}_{b_j}\right) \longrightarrow 0$. This contradicts with the assumption that $\lambda_p(\eta^\intercal\eta) \gtrsim 1$. Therefore, we have $\tilde{p} \ge p$. Together with the result in (iii), we have $\tilde{p} = p$. □

Lemma 4. *Suppose Assumptions A.1-A.8 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 $k \le \tilde{p}$ and for I_k, $\tilde{p}$ and $\beta_{(k)}$ defined in Section A, we have*

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

(ii) $\left\|\widetilde{R}_{(k)} - \beta_{(k)}\bar{V}\right\| \lesssim_p q^{1/2}N^{1/2} + T^{1/2}$.

(iii) $|\widehat{\lambda}_{(k)}^{1/2}/\left\|\beta_{(k)}\right\| - 1| \lesssim_p q^{-1/2}N^{-1/2} + T^{-1/2}$.

(iv) $\left\|\mathbb{P}_{\widehat{V}_{(k)}^\intercal} - T^{-1}\bar{V}^\intercal\mathbb{P}_{b_k}\bar{V}\right\| \lesssim_p q^{-1/2}N^{-1/2} + T^{-1/2}$.

(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 A.1, A.2, A.3, A.6 A.7, we have

$$\left\|T^{-1}\bar{R}\bar{G}^{\intercal}-\beta\eta^{\intercal}\right\|_{\mathrm{MAX}}\lesssim\|\beta\|_{\mathrm{MAX}}\left\|T^{-1}\bar{V}\bar{V}^{\intercal}-\mathbb{I}_p\right\|\|\eta\|+T^{-1}\left\|\bar{U}\bar{V}^{\intercal}\right\|_{\mathrm{MAX}}\|\eta\|$$
$$+T^{-1}\|\beta\|_{\mathrm{MAX}}\left\|\bar{V}\bar{Z}^{\intercal}\right\|+T^{-1}\left\|\bar{U}\bar{Z}^{\intercal}\right\|_{\mathrm{MAX}}\lesssim_p(\log N)^{1/2}T^{-1/2}.$$

From Assumption A.8, we have $c^{(1)}_{qN}-c^{(1)}_{qN+1}\gtrsim c^{(1)}_{qN}$ and the the definition of $\tilde{p}$ implies that $c^{(k)}_{qN}\geq c$ for $k\leq\tilde{p}$. Thus, we have $c^{(1)}_{qN}-c^{(1)}_{qN+1}\gtrsim c$. Define the events

$$\begin{aligned}
A_1:&=\left\{\ \left\|T^{-1}\bar{R}_{[i]}\bar{G}^{\intercal}\right\|_{\mathrm{MAX}}>(c^{(1)}_{qN}+c^{(1)}_{qN+1})/2\text{ for all }i\in I_1\right\},\\
A_2:&=\left\{\ \left\|T^{-1}\bar{R}_{[i]}\bar{G}^{\intercal}\right\|_{\mathrm{MAX}}<(c^{(1)}_{qN}+c^{(1)}_{qN+1})/2\text{ for all }i\in I_1^c\right\},\\
A_3:&=\left\{\ \left\|T^{-1}\bar{R}_{[i]}\bar{G}^{\intercal}-\beta_{[i]}\eta^{\intercal}\right\|_{\mathrm{MAX}}\geq(c^{(1)}_{qN}-c^{(1)}_{qN+1})/2\text{ for some }i\in[N]\right\}.
\end{aligned}\tag{B.106}$$

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 $\left\|\beta_{[i]}\eta^{\intercal}\right\|_{\mathrm{MAX}}\geq c^{(1)}_{qN}$ for all $i\in I_1$ and $\left\|\beta_{[i]}\eta^{\intercal}\right\|_{\mathrm{MAX}}\leq c^{(1)}_{qN+1}$ for all $i\in I_1^c$. Therefore, if A_1^c occurs, we have

$$\left\|T^{-1}\bar{R}_{[i]}\bar{G}^{\intercal}-\beta_{[i]}\eta^{\intercal}\right\|_{\mathrm{MAX}}\geq(c^{(1)}_{qN}-c^{(1)}_{qN+1})/2,$$

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

$$\mathrm{P}(\widehat{I}_1=I_1)\geq\mathrm{P}(A_1\cap A_2)=1-\mathrm{P}(A_1^c\cup A_2^c)\geq1-\mathrm{P}(A_3).\tag{B.107}$$

Using $c^{-1}(\log N)^{1/2}T^{-1/2}\to0$ and $c^{(1)}_{qN}-c^{(1)}_{qN+1}\gtrsim c$, we have $\mathrm{P}(A_3)\to0$ and consequently, $\mathrm{P}(\widehat{I}_1=I_1)\to1$.

(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 A.13 gives $\left\|\widetilde{R}_{(1)}-\beta_{(1)}\bar{V}\right\|=\left\|\bar{U}_{[I_1]}\right\|\lesssim_p q^{1/2}N^{1/2}+T^{1/2}$.

(iii) From Lemma 10, we have $\sigma_j(\beta_{(1)}\bar{V})/\sigma_j(\beta_1)=T^{1/2}+O_p(1)$. The result in (ii) implies that

$$|\left\|\widetilde{R}_{(1)}\right\|-\|\beta_{(1)}\bar{V}\||\leq\left\|\widetilde{R}_{(1)}-\beta_{(1)}\bar{V}\right\|\lesssim_p q^{1/2}N^{1/2}+T^{1/2}.$$

Together with $\left\|\beta_{(1)}\right\|\asymp qN$ from Lemma 3, we have

$$|\frac{\widehat{\lambda}^{1/2}_{(1)}}{\left\|\beta_{(k)}\right\|}-1|=|\frac{\left\|\widetilde{R}_{(1)}\right\|}{T^{1/2}\left\|\beta_{(1)}\right\|}-1|\leq\frac{|\left\|\widetilde{R}_{(1)}\right\|-\|\beta_{(1)}\bar{V}\||}{T^{1/2}\left\|\beta_{(1)}\right\|}+\frac{|\|\beta_{(1)}\bar{V}\|-T^{1/2}\left\|\beta_{(1)}\right\||}{T^{1/2}\left\|\beta_{(1)}\right\|}\lesssim_p q^{-1/2}N^{-1/2}+T^{-1/2}.$$

(iv) Let $\tilde{\xi}_{(1)}\in\mathbb{R}^{T\times1}$ denote the first right singular vector of $\beta_{(1)}\bar{V}$. From Lemma 10, we have

$$\left\|\mathbb{P}_{\tilde{\xi}_{(1)}}-T^{-1}\bar{V}^{\intercal}\mathbb{P}_{b_k}\bar{V}\right\|\lesssim_p T^{-1/2}\tag{B.108}$$

and $\sigma_j(\beta_{(1)}\bar{V})/\sigma_j(\beta_{(1)}) = T^{1/2} + O_p(1)$ for $j \leq p$, which leads to

$$\sigma_1(\beta_{(1)}\bar{V}) - \sigma_2(\beta_{(1)}\bar{V}) = T^{1/2}(\sigma_1(\beta_{(1)}) - \sigma_2(\beta_{(1)})) + O_p(\sigma_1(\beta_{(1)})) \asymp_p T^{1/2}\sigma_1(\beta_{(1)}), \tag{B.109}$$

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

Using $\left\|\widetilde{R}_{(1)} - \beta_{(1)}\bar{V}\right\| \lesssim_p q^{1/2}N^{1/2} + T^{1/2}$ as proved in (ii), (B.109), Lemma 3 and Wedin's sin-theta theorem for singular vectors in Wedin (1972), we have

$$\left\|\mathbb{P}_{\widehat{V}_{(k)}^{\intercal}} - \mathbb{P}_{\tilde{\xi}_{(1)}}\right\| \lesssim_p \frac{q^{1/2}N^{1/2} + T^{1/2}}{\sigma_1(\beta_{(1)}\bar{V}) - \sigma_2(\beta_{(1)}\bar{V})} \lesssim_p q^{-1/2}N^{-1/2} + T^{-1/2}, \tag{B.110}$$

In light of (B.108) and (B.110), 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

$$\begin{aligned}
&\left\|\beta\prod_{j=1}^{k-1}\mathbb{M}_{b_j}\eta^{\intercal} - T^{-1}(\beta\bar{V} + \bar{U})\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}}(\eta\bar{V} + \bar{Z})^{\intercal}\right\|_{\mathrm{MAX}} \\
\leq &\left\|\beta\prod_{j=1}^{k-1}\mathbb{M}_{b_j}\eta^{\intercal} - T^{-1}\beta\bar{V}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}}\bar{V}^{\intercal}\eta^{\intercal}\right\|_{\mathrm{MAX}} + T^{-1}\left\|\beta\bar{V}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}}\bar{Z}^{\intercal}\right\|_{\mathrm{MAX}} \\
&+ T^{-1}\left\|\bar{U}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}}\bar{V}^{\intercal}\eta^{\intercal}\right\|_{\mathrm{MAX}} + T^{-1}\left\|\bar{U}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^{\intercal}}\bar{Z}^{\intercal}\right\|_{\mathrm{MAX}}
\end{aligned} \tag{B.111}$$

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

$$\left\|\sum_{j=1}^{k-1}\mathbb{P}_{\widehat{V}_{(j)}^{\intercal}} - T^{-1}\bar{V}^{\intercal}\sum_{j=1}^{k-1}\mathbb{P}_{b_j}\bar{V}\right\| = \left\|\sum_{j=1}^{k-1}\left(\mathbb{P}_{\widehat{V}_{(j)}^{\intercal}} - T^{-1}\bar{V}^{\intercal}\mathbb{P}_{b_j}\bar{V}\right)\right\| \lesssim_p q^{-1/2}N^{-1/2} + T^{-1/2}. \tag{B.112}$$

Using Lemma 2 and Lemma 3(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, (B.112), and $\left\|T^{-1}\bar{V}\bar{V}^\intercal - \mathbb{I}_p\right\| \lesssim_p T^{-1/2}$, we have

$$T^{-1/2}\left\|\bar{V}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^\intercal} - \prod_{j=1}^{k-1}\mathbb{M}_{b_j}\bar{V}\right\| = T^{-1/2}\left\|\bar{V}\sum_{j=1}^{k-1}\mathbb{P}_{\widehat{V}_{(j)}^\intercal} - \sum_{j=1}^{k-1}\mathbb{P}_{b_j}\bar{V}\right\| \lesssim_p q^{-1/2}N^{-1/2} + T^{-1/2}. \tag{B.113}$$

Similarly, right multiplying $\bar{V}^\intercal$ to the term inside the $\|\cdot\|$ of (B.113), we have

$$\left\|T^{-1}\bar{V}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^\intercal}\bar{V}^\intercal - \prod_{j=1}^{k-1}\mathbb{M}_{b_j}\right\| \lesssim_p q^{-1/2}N^{-1/2} + T^{-1/2}. \tag{B.114}$$

Then, we analyze these four terms in (B.111) one by one. For the first term, using (B.114) and Assumption A.3, we have

$$\begin{aligned}\left\|\beta\prod_{j=1}^{k-1}\mathbb{M}_{b_j}\eta^\intercal - T^{-1}\beta\bar{V}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^\intercal}\bar{V}^\intercal\eta^\intercal\right\|_{\text{MAX}} &\lesssim \|\beta\|_{\text{MAX}}\left\|\prod_{j=1}^{k-1}\mathbb{M}_{b_j} - T^{-1}\bar{V}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^\intercal}\bar{V}^\intercal\right\|\|\eta\| \\ &\lesssim_p q^{-1/2}N^{-1/2} + T^{-1/2}.\end{aligned}$$

For the second term, using (B.113), Lemma 1 and Assumptions A.3 and A.2, we have

$$\begin{aligned}T^{-1}\left\|\beta\bar{V}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^\intercal}\bar{Z}^\intercal\right\|_{\text{MAX}} &\lesssim T^{-1}\|\beta\|_{\text{MAX}}\left\|\prod_{j=1}^{k-1}\mathbb{M}_{b_j}\right\|\left\|\bar{V}\bar{Z}^\intercal\right\| + T^{-1}\|\beta\|_{\text{MAX}}\left\|\bar{V}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^\intercal} - \prod_{j=1}^{k-1}\mathbb{M}_{b_j}\bar{V}\right\|\left\|\bar{Z}\right\| \\ &\lesssim_p q^{-1/2}N^{-1/2} + T^{-1/2}.\end{aligned}$$

For the third term, using (B.113) and Lemma 1, we have

$$\begin{aligned}T^{-1}\left\|\bar{U}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^\intercal}\bar{V}^\intercal\eta^\intercal\right\|_{\text{MAX}} \lesssim & T^{-1}\left\|\bar{U}\bar{V}^\intercal\right\|_{\text{MAX}}\left\|\prod_{j=1}^{k-1}\mathbb{M}_{b_j}\right\|\|\eta\| \\ & + T^{-1}\left\|\bar{U}\right\|_{\text{MAX}}T^{1/2}\left\|\bar{V}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^\intercal} - \prod_{j=1}^{k-1}\mathbb{M}_{b_j}\bar{V}\right\|\|\eta\| \\ \lesssim_p & (\log NT)^{1/2}\left(q^{-1/2}N^{-1/2} + T^{-1/2}\right).\end{aligned}$$

For the forth term, using (B.112) and Lemma 1, we have

$$T^{-1}\left\|\bar{U}\prod_{j=1}^{k-1}\mathbb{M}_{\widehat{V}_{(j)}^\intercal}\bar{Z}^\intercal\right\|_{\text{MAX}} \lesssim T^{-1}\left\|\bar{U}\bar{Z}^\intercal\right\|_{\text{MAX}} + T^{-2}\left\|\bar{U}\bar{V}^\intercal\right\|_{\text{MAX}}\left\|\sum_{j=1}^{k-1}\mathbb{P}_{b_j}\right\|\left\|\bar{V}\bar{Z}^\intercal\right\|$$

$$+ T^{-1/2} \left\|\bar{U}\right\|_{\mathrm{MAX}} \left\| T^{-1}\bar{V}^{\intercal} \sum_{j=1}^{k-1} \mathbb{P}_{b_j} \bar{V} - \sum_{j=1}^{k-1} \mathbb{P}_{\widehat{V}^{\intercal}_{(j)}} \right\| \left\|\bar{Z}\right\|$$
$$\lesssim_p (\log NT)^{1/2} \left(q^{-1/2}N^{-1/2} + T^{-1/2}\right).$$

Hence, we have

$$\left\| T^{-1}\bar{R} \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{V}^{\intercal}_{(j)}} \bar{G}^{\intercal} - \beta \prod_{j=1}^{k-1} \mathbb{M}_{b_j} \eta^{\intercal} \right\|_{\mathrm{MAX}} \lesssim_p (\log NT)^{1/2} \left(q^{-1/2}N^{-1/2} + T^{-1/2}\right). \tag{B.115}$$

As in the case of $k = 1$, from Assumption A.8, we have $c_{qN}^{(k)} - c_{qN+1}^{(k)} \gtrsim c_{qN}^{(k)}$. In addition, since the stopping rule for the procedure in Section A 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 (B.106) and (B.107) 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 $\widetilde{R}_{(k)} = \bar{R}_{[I_k]} \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{V}^{\intercal}_{(j)}}$ and thus

$$\widetilde{R}_{(k)} - \beta_{(k)}\bar{V} = \bar{R}_{[I_k]} \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{V}^{\intercal}_{(j)}} - \beta_{(k)}\bar{V} = \bar{\beta}_{[I_k]} \left(\bar{V} \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{V}^{\intercal}_{(j)}} - \prod_{j=1}^{k-1} \mathbb{M}_{b_j} \bar{V} \right) + \bar{U}_{[I_k]} \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{V}^{\intercal}_{(j)}}.$$

Hence, using Assumptions A.3, Lemma 1, and (B.113), we have

$$\left\| \widetilde{R}_{(k)} - \beta_{(k)}\bar{V} \right\| \leq \left\|\beta_{[I_k]}\right\| \left\| \bar{V} \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{V}^{\intercal}_{(j)}} - \prod_{j=1}^{k-1} \mathbb{M}_{b_j} \bar{V} \right\| + \left\|\bar{U}_{[I_k]}\right\| \left\| \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{V}^{\intercal}_{(j)}} \right\| \lesssim_p q^{1/2}N^{1/2} + T^{1/2}.$$

(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 $|\widehat{\lambda}_{(k)}^{1/2} / \left\|\beta_{(k)}\right\| - 1| \lesssim_p q^{-1/2}N^{-1/2} + T^{-1/2}$.

(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 $\left\| \mathbb{M}_{\tilde{\xi}_{(k)}} - T^{-1}\bar{V}^{\intercal}\mathbb{M}_{b_k}\bar{V} \right\| \lesssim_p T^{-1/2}$ from Lemma 10. Since we have $\left\| \widetilde{R}_{(k)} - \beta_{(k)}\bar{V} \right\| \lesssim_p q^{-1/2}N^{-1/2} + T^{-1/2}$ from (ii), using the same proof as in the case $k = 1$, we have

$$\left\| \mathbb{M}_{\widehat{V}^{\intercal}_{(k)}} - \mathbb{M}_{\tilde{\xi}_{(k)}} \right\| \lesssim_p q^{-1/2}N^{-1/2} + T^{-1/2},$$

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 $T^{-1}\bar{R}_{[i]} \prod_{j<k} \mathbb{M}_{\widehat{V}^{\intercal}_{(j)}} \bar{G}^{\intercal}$. Since (iv) holds for $j \leq \tilde{p}$ as shown above, using the same proof for (B.115),

we have

$$\left\|T^{-1}\bar{R}\prod_{j=1}^{\tilde{p}}\mathbb{M}_{\widehat{V}^{\intercal}_{(j)}}\bar{G}^{\intercal}-\beta\prod_{j=1}^{\tilde{p}}\mathbb{M}_{b_j}\eta^{\intercal}\right\|_{\text{MAX}}\lesssim_p(\log NT)^{1/2}\left(q^{-1/2}N^{-1/2}+T^{-1/2}\right). \tag{B.116}$$

The assumption $c_{qN}^{(\tilde{p}+1)}\leq(1+\delta)^{-1}c$ in Assumption A.8 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 (B.106) and (B.107) with events

$$\begin{aligned}
B_1:&=\left\{\left\|T^{-1}\bar{R}_{[i]}\prod_{j=1}^{\tilde{p}}\mathbb{M}_{\widehat{V}^{\intercal}_{(j)}}\bar{G}^{\intercal}\right\|_{\text{MAX}}>(c+c_{qN}^{(\tilde{p}+1)})/2\text{ for at most }qN-1\text{ rows }i\in[N]\right\},\\
B_2:&=\left\{\left\|T^{-1}\bar{R}_{[i]}\prod_{j=1}^{\tilde{p}}\mathbb{M}_{\widehat{V}^{\intercal}_{(j)}}\bar{G}^{\intercal}-\beta_{[i]}\prod_{j=1}^{\tilde{p}}\mathbb{M}_{b_j}\eta^{\intercal}\right\|_{\text{MAX}}\geq(c-c_{qN}^{(\tilde{p}+1)})/2\text{ for some }i\in[N]\right\},
\end{aligned} \tag{B.117}$$

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 5. *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*

$$\begin{pmatrix}s^1_{(k)}\\ s^2_{(k)}\end{pmatrix}:=\Gamma_{(k)}\varsigma_{(k)}\quad\text{and}\quad\begin{pmatrix}\widetilde{U}^1_{(k)}\\ \widetilde{U}^2_{(k)}\end{pmatrix}:=\Gamma_{(k)}\bar{U}_{[I_k]},$$

where $s^1_{(k)}\in\mathbb{R}^{p\times 1}$ and $\widetilde{U}^1_{(k)}\in\mathbb{R}^{p\times T}$ are the first p rows of $\Gamma_{(k)}\varsigma_{(k)}$ and $\Gamma_{(k)}\bar{U}_{[I_k]}$, respectively. Then, under Assumptions A.1-A.8, we have

(i) $\left\|s^2_{(k)}\right\|\lesssim_p T^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}(|I_k|^{1/2}+T^{1/2})$.

(ii) $\left\|\widetilde{U}^1_{(k)}\right\|\lesssim_p T^{1/2}$, $\left\|\widetilde{U}^1_{(k)}\bar{V}^{\intercal}\right\|\lesssim_p T^{1/2}$, $\left\|\widetilde{U}^1_{(k)}\bar{Z}^{\intercal}\right\|\lesssim_p T^{1/2}$.

Proof. (i) The assumption $\widehat{I}_k=I_k$ and the definition (B.39) of $\widetilde{R}_{(k)}$ together lead to

$$\widetilde{R}_{(k)}=\bar{R}_{[I_k]}\prod_{i<k}\left(\mathbb{I}_T-\xi_{(i)}\xi_{(i)}^{\intercal}\right).$$

Then, with (B.56) and Lemma 2, we have $\varsigma_{(k)}=\bar{R}_{[I_k]}\xi_{(k)}/\sqrt{T\widehat{\lambda}_{(k)}}$. From the construction of $\Gamma_{(k)}$, we have

$$\Gamma_{(k)}\bar{R}_{(k)}=\begin{pmatrix}\left(\beta_{[I_k]}^{\intercal}\beta_{[I_k]}\right)^{\frac{1}{2}}\bar{V}+\widetilde{U}^1_{(k)}\\ \widetilde{U}^2_{(k)}\end{pmatrix},$$

which in turn gives

$$\begin{pmatrix} s^1_{(k)} \\ s^2_{(k)} \end{pmatrix} = \Gamma_{(k)}\varsigma_{(k)} = \frac{1}{\sqrt{T\widehat{\lambda}_{(k)}}} \begin{pmatrix} \left(\beta^{\intercal}_{[I_k]}\beta_{[I_k]}\right)^{\frac{1}{2}} \bar{V} + \widetilde{U}^1_{(k)} \\ \widetilde{U}^2_{(k)} \end{pmatrix} \xi_{(k)}.$$

With Lemma 1(v), we have

$$\left\| s^2_{(k)} \right\| = \left\| \frac{\widetilde{U}^2_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}} \right\| \leq \left\| \frac{\bar{U}_{[I_k]}}{\sqrt{T\widehat{\lambda}_{(k)}}} \right\| \lesssim_p T^{-1/2}\widehat{\lambda}^{-1/2}_{(k)}(|I_k|^{1/2} + T^{1/2}).$$

(ii) With Lemma 1(ii)(iii) and the definition of $\Gamma_{(k)}$, these results follow immediately. □

Lemma 6. *Under Assumptions A.1-A.8, if $\widehat{\lambda}_{(j)} \asymp_p |I_j|$ and $|I_j| \asymp qN$ for $j \leq \tilde{p}$, then for $k \leq \tilde{p}$, we have*

(i) $\left\| \frac{\bar{U}^{\intercal}_{[I_k]}\varsigma_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}} \right\| \lesssim_p q^{-1/2}N^{-1/2} + T^{-1}.$

(ii) $\left\| \frac{\bar{V}\bar{U}^{\intercal}_{[I_k]}\varsigma_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}} \right\| \lesssim_p q^{-1}N^{-1} + T^{-1}$, $\left\| \frac{\bar{Z}\bar{U}^{\intercal}_{[I_k]}\varsigma_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}} \right\| \lesssim_p q^{-1}N^{-1} + T^{-1}$, $|\frac{\varsigma^{\intercal}_{(k)}\bar{u}_{[I_k]}}{\sqrt{\widehat{\lambda}_{(k)}}}| \lesssim_p q^{-1}N^{-1} + T^{-1}$.

Proof. (i) Using the equation $\varsigma^{\intercal}_{(k)}\bar{U}_{[I_k]} = (s^1_{(k)})^{\intercal}\widetilde{U}^1_{(k)} + (s^2_{(k)})^{\intercal}\widetilde{U}^2_{(k)}$ and Lemma 5, we have

$$\begin{aligned} \left\| \varsigma^{\intercal}_{(k)}\bar{U}_{[I_k]} \right\| &\leq \left\| s^1_{(k)} \right\| \left\| \widetilde{U}^1_{(k)} \right\| + \left\| s^2_{(k)} \right\| \left\| \widetilde{U}^2_{(k)} \right\| \leq \left\| s^1_{(k)} \right\| \left\| \widetilde{U}^1_{(k)} \right\| + \left\| s^2_{(k)} \right\| \left\| \bar{U}_{[I_k]} \right\| \\ &\lesssim_p \sqrt{T} + \frac{|I_k| + T}{\sqrt{T\widehat{\lambda}_{(k)}}}, \end{aligned} \tag{B.118}$$

which leads to

$$\left\| \frac{\bar{U}^{\intercal}_{[I_k]}\varsigma_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}} \right\| \lesssim_p \frac{1}{\sqrt{\widehat{\lambda}_{(k)}}} + \frac{|I_k| + T}{T\widehat{\lambda}_{(k)}} \lesssim_p q^{-1/2}N^{-1/2} + T^{-1}.$$

(ii) From Lemmas 1 and 5, we have

$$\begin{aligned} \left\| \bar{V}\bar{U}^{\intercal}_{[I_k]}\varsigma^{(k)} \right\| \leq \left\| \bar{V}\left(\widetilde{U}^1_{(k)}\right)^{\intercal} s^1_{(k)} \right\| + \left\| \bar{V}\left(\widetilde{U}^2_{(k)}\right)^{\intercal} s^2_{(k)} \right\| &\leq \left\| \bar{V}\left(\widetilde{U}^1_{(k)}\right)^{\intercal} \right\| + \left\| \bar{V}\bar{U}^{\intercal}_{[I_k]} \right\| \left\| s^2_{(k)} \right\| \\ &\lesssim_p \sqrt{T} + \frac{|I_k| + T}{\sqrt{\widehat{\lambda}_{(k)}}}, \end{aligned}$$

which leads to

$$\left\| \frac{\bar{V}\bar{U}^{\intercal}_{[I_k]}\varsigma_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}} \right\| \lesssim_p \frac{1}{\sqrt{T\widehat{\lambda}_{(k)}}} + \frac{|I_k| + T}{T\widehat{\lambda}_{(k)}} \lesssim_p q^{-1}N^{-1} + T^{-1}.$$

Replacing $\bar{V}$ by $\bar{Z}$ and $\iota_T^\top$ in the above proof and using Lemmas 1 and 5, we have similar results:

$$\left\| \frac{\bar{Z}\bar{U}_{[I_k]}^\top \varsigma_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}} \right\| \lesssim_p q^{-1}N^{-1} + T^{-1}, \qquad \text{and} \qquad |\frac{\bar{u}_{[I_k]}^\top \varsigma_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}}| \lesssim_p q^{-1}N^{-1} + T^{-1}. \tag{B.119}$$

□

Lemma 7. *Under Assumptions A.1-A.8, if $\widehat{\lambda}_{(j)} \asymp_p |I_j|$ and $|I_j| \asymp qN$ for $j \leq \tilde{p}$, then for $k, l \leq \tilde{p}$, we have*

(i) $\quad \left\| \frac{\widetilde{U}_{(k)}^\top \varsigma_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}} \right\| \lesssim_p q^{-1/2}N^{-1/2} + T^{-1}, \qquad \left\| \frac{\widetilde{U}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}} \right\| \lesssim_p q^{-1/2}N^{-1/2} + T^{-1/2}.$

(ii) $\quad \left\| \frac{\bar{V}\widetilde{U}_{(k)}^\top \varsigma_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}} \right\| \lesssim_p q^{-1}N^{-1} + T^{-1}, \qquad \left\| \frac{\bar{Z}\widetilde{U}_{(k)}^\top \varsigma_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}} \right\| \lesssim_p q^{-1}N^{-1} + T^{-1}, \qquad |\frac{\varsigma_{(k)}^\top \widetilde{u}_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}}| \lesssim_p q^{-1}N^{-1} + T^{-1}.$

(iii) $\quad |\frac{\xi_{(l)}^\top \widetilde{U}_{(k)}^\top \varsigma_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}| \lesssim_p q^{-1}N^{-1} + T^{-1}.$

Proof. (i) Recall that in the definition of $U_{(k)}$ in (B.40), we have

$$\widetilde{U}_{(k)} = \bar{U}_{[I_k]} - \sum_{i=1}^{k-1} \frac{\bar{R}_{[I_k]}\xi_{(i)}}{\sqrt{T}} \frac{\varsigma_{(i)}^\top \widetilde{U}_{(i)}}{\sqrt{\widehat{\lambda}_{(i)}}}. \tag{B.120}$$

Then, a direct multiplication of $\varsigma_{(k)}^\top / \sqrt{T\widehat{\lambda}_{(k)}}$ from the left side of (B.120) leads to

$$\frac{\varsigma_{(k)}^\top \widetilde{U}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}} = \frac{\varsigma_{(k)}^\top \bar{U}_{[I_k]}}{\sqrt{T\widehat{\lambda}_{(k)}}} - \sum_{i=1}^{k-1} \frac{\varsigma_{(k)}^\top \bar{R}_{[I_k]}\xi_{(i)}}{\sqrt{T\widehat{\lambda}_{(k)}}} \frac{\varsigma_{(i)}^\top \widetilde{U}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}.$$

Consequently, with Lemma 6(i) we have

$$\begin{aligned} \left\| \frac{\varsigma_{(k)}^\top \widetilde{U}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}} \right\| &\leq \left\| \frac{\varsigma_{(k)}^\top \bar{U}_{[I_k]}}{\sqrt{T\widehat{\lambda}_{(k)}}} \right\| + \sum_{i=1}^{k-1} \left\| \frac{\bar{R}_{[I_k]}}{\sqrt{T\widehat{\lambda}_{(k)}}} \right\| \left\| \frac{\varsigma_{(i)}^\top \widetilde{U}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \right\| \lesssim_p \frac{1}{\sqrt{\widehat{\lambda}_{(k)}}} + \frac{|I_k| + T}{T\widehat{\lambda}_{(k)}} + \sqrt{\frac{|I_k|}{\widehat{\lambda}_{(k)}}} \sum_{i=1}^{k-1} \left\| \frac{\varsigma_{(i)}^\top \widetilde{U}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \right\| \\ &\lesssim_p q^{-1/2}N^{-1/2} + T^{-1} + \sum_{i=1}^{k-1} \left\| \frac{\varsigma_{(i)}^\top \widetilde{U}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}} \right\|. \end{aligned} \tag{B.121}$$

If $\left\| T^{-1/2}\widehat{\lambda}_{(i)}^{-1/2} \varsigma_{(i)}^\top \widetilde{U}_{(i)} \right\| \lesssim_p q^{-1/2}N^{-1/2} + T^{-1}$ holds for $i \leq k-1$, then (B.121) 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 6(i). Therefore, we have $\left\| T^{-1/2}\widehat{\lambda}_{(k)}^{-1/2} \varsigma_{(k)}^\top \widetilde{U}_{(k)} \right\| \lesssim_p q^{-1/2}N^{-1/2} + T^{-1}$ for $k \leq \tilde{p}$ by induction.

Using (B.120) again, with Assumption A.4, we have

$$\left\|\frac{\widetilde{U}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}\right\| \leq \left\|\frac{\bar{U}_{[I_k]}}{\sqrt{T\widehat{\lambda}_{(k)}}}\right\| + \sum_{i=1}^{k-1}\left\|\frac{\bar{R}_{[I_k]}}{\sqrt{T\widehat{\lambda}_{(k)}}}\right\|\left\|\frac{\widetilde{U}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}\right\| \lesssim_p q^{-1/2}N^{-1/2} + T^{-1/2} + \sum_{i=1}^{k-1}\left\|\frac{\widetilde{U}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}\right\|. \tag{B.122}$$

When $k=1$, Assumption A.4 implies that $\left\|T^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}\widetilde{U}_{(k)}\right\| \lesssim_p q^{-1/2}N^{-1/2} + T^{-1/2}$. Then, using the same induction argument with (B.122), we have this ineqaulity holds for $k \leq \tilde{p}$.

(ii) Similarly, by simple multiplication of $\bar{V}^\top$ from the right side of (B.120), we have

$$\frac{\varsigma_{(k)}^\top \widetilde{U}_{(k)}\bar{V}^\top}{T\sqrt{\widehat{\lambda}_{(k)}}} = \frac{\varsigma_{(k)}^\top \bar{U}_{[I_k]}\bar{V}^\top}{T\sqrt{\widehat{\lambda}_{(k)}}} - \sum_{i=1}^{k-1}\frac{\varsigma_{(k)}^\top \bar{R}_{[I_k]}\xi_{(i)}}{\sqrt{T\widehat{\lambda}_{(k)}}}\frac{\varsigma_{(i)}^\top \widetilde{U}_{(i)}\bar{V}^\top}{T\sqrt{\widehat{\lambda}_{(i)}}}.$$

Consequently, we have

$$\begin{aligned}\left\|\frac{\varsigma_{(k)}^\top \widetilde{U}_{(k)}\bar{V}^\top}{T\sqrt{\widehat{\lambda}_{(k)}}}\right\| &\leq \left\|\frac{\varsigma_{(k)}^\top \bar{U}_{[I_k]}\bar{V}^\top}{T\sqrt{\widehat{\lambda}_{(k)}}}\right\| + \sum_{i=1}^{k-1}\left\|\frac{\bar{R}_{[I_k]}}{\sqrt{T\widehat{\lambda}_{(k)}}}\right\|\left\|\frac{\varsigma_{(i)}^\top \widetilde{U}_{(i)}\bar{V}^\top}{T\sqrt{\widehat{\lambda}_{(i)}}}\right\| \\ &\lesssim_p \frac{1}{\sqrt{T\widehat{\lambda}_{(k)}}} + \frac{|I_k|+T}{T\widehat{\lambda}_{(k)}} + \sqrt{\frac{|I_k|}{\widehat{\lambda}_{(k)}}}\sum_{i=1}^{k-1}\left\|\frac{\varsigma_{(i)}^\top \widetilde{U}_{(i)}\bar{V}^\top}{\sqrt{T\widehat{\lambda}_{(i)}}}\right\| \\ &\lesssim_p q^{-1}N^{-1} + T^{-1} + \sum_{i=1}^{k-1}\left\|\frac{\varsigma_{(i)}^\top \widetilde{U}_{(i)}\bar{V}^\top}{\sqrt{T\widehat{\lambda}_{(i)}}}\right\|. \end{aligned} \tag{B.123}$$

When $k=1$, $\left\|T^{-1}\widehat{\lambda}_{(k)}^{-1/2}\varsigma_{(k)}^\top \widetilde{U}_{(k)}\bar{V}^\top\right\| \lesssim_p q^{-1}N^{-1} + T^{-1}$ is a result of Lemma 6(ii). Then, a direct induction argument using (B.123) leads to this inequality for $k \leq \tilde{p}$.

Replacing $\bar{V}$ by $\bar{Z}$ and $\iota_T^\top$ in the above proof, and using Lemma 6(ii), we have the following results:

$$\left\|\frac{\bar{Z}\widetilde{U}_{(k)}^\top \varsigma_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}}\right\| \lesssim_p q^{-1}N^{-1} + T^{-1} \quad \text{and} \quad |\frac{\widetilde{u}_{(k)}^\top \varsigma_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}}| \lesssim_p q^{-1}N^{-1} + T^{-1}.$$

(iii) Recall that $\widetilde{R}_{(k)} = \widetilde{\beta}_{(k)}\bar{V} + \widetilde{U}_{(k)}$ as defined in (B.39), we have

$$|\varsigma_{(l)}^\top \widetilde{R}_{(l)}\widetilde{U}_{(k)}^\top \varsigma_{(k)}| \leq |\varsigma_{(l)}^\top \widetilde{\beta}_{(l)}\bar{V}\widetilde{U}_{(k)}^\top \varsigma_{(k)}| + |\varsigma_{(l)}^\top \widetilde{U}_{(l)}\widetilde{U}_{(k)}^\top \varsigma_{(k)}| \leq \left\|\varsigma_{(l)}^\top \widetilde{\beta}_{(l)}\right\|\left\|\bar{V}\widetilde{U}_{(k)}^\top \varsigma_{(k)}\right\| + \left\|\varsigma_{(l)}^\top \widetilde{U}_{(l)}\right\|\left\|\widetilde{U}_{(k)}^\top \varsigma_{(k)}\right\|.$$

Using (B.56), we have

$$|\frac{\xi_{(k)}^{\intercal}\widetilde{U}_{(k)}^{\intercal}\varsigma_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}| = |\frac{\varsigma_{(l)}^{\intercal}\widetilde{R}_{(l)}\widetilde{U}_{(k)}^{\intercal}\varsigma_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}\widehat{\lambda}_{(l)}}}| \leq \left\|\frac{\varsigma_{(l)}^{\intercal}\widetilde{\beta}_{(l)}}{\sqrt{\widehat{\lambda}_{(l)}}}\right\| \left\|\frac{\bar{V}\widetilde{U}_{(k)}^{\intercal}\varsigma_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}}\right\| + \left\|\frac{\widetilde{U}_{(k)}^{\intercal}\varsigma_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}\right\| \left\|\frac{\widetilde{U}_{(l)}^{\intercal}\varsigma_{(l)}}{\sqrt{T\widehat{\lambda}_{(l)}}}\right\|. \tag{B.124}$$

With Lemma 1 and (i), we have

$$T^{1/2}\left\|\widetilde{\beta}_{(k)}\right\| \lesssim_p \sigma_p(\bar{V})\left\|\widetilde{\beta}_{(k)}\right\| \leq \left\|\widetilde{\beta}_{(k)}\bar{V}\right\| \leq \left\|\widetilde{U}_{(k)}\right\| + \left\|\widetilde{R}_{(k)}\right\| \leq \left\|\widetilde{U}_{(k)}\right\| + \left\|\bar{R}_{[I_k]}\right\| \lesssim_p T^{1/2}q^{1/2}N^{1/2}, \tag{B.125}$$

which leads to $\left\|\widehat{\lambda}_{(k)}^{-1/2}\varsigma_{(k)}^{\intercal}\widetilde{\beta}_{(k)}\right\| \lesssim_p q^{-1/2}N^{-1/2}\left\|\widetilde{\beta}_{(k)}\right\| \lesssim_p 1$. Using this inequality and results of (i) and (ii) in (B.124) completes the proof. □

Lemma 8. *Under Assumptions A.1-A.8, if $\widehat{\lambda}_{(j)} \asymp_p |I_j|$ and $|I_j| \asymp qN$ for $j \leq \tilde{p}$, then for $k \leq \tilde{p}+1$, we have*

(i) $\left\|\widetilde{Z}_{(k)}\bar{V}^{\intercal}\right\| \lesssim_p T^{1/2} + Tq^{-1}N^{-1}$.

(ii) $\left\|\widetilde{Z}_{(k)}\bar{U}_{[I_0]}^{\intercal}\right\| \lesssim_p N_0^{1/2}T^{1/2} + Tq^{-1/2}N^{-1/2}$.

Proof. (i) From the definition (B.44) of $\widetilde{Z}_{(k)}$, we have

$$\widetilde{Z}_{(k)}\bar{V}^{\intercal} = \bar{Z}\bar{V}^{\intercal} - \sum_{i=1}^{k-1}\bar{G}\xi_{(i)}\frac{\varsigma_{(i)}^{\intercal}\widetilde{U}_{(i)}\bar{V}^{\intercal}}{\sqrt{T\widehat{\lambda}_{(i)}}}.$$

Then, with Lemma 7(ii), we have

$$\left\|\widetilde{Z}_{(k)}\bar{V}^{\intercal}\right\| \leq \left\|\bar{Z}\bar{V}^{\intercal}\right\| + \sum_{i=1}^{k-1}\left\|\bar{G}\xi_{(i)}\right\|\left\|\frac{\varsigma_{(i)}^{\intercal}\widetilde{U}_{(i)}\bar{V}^{\intercal}}{\sqrt{T\widehat{\lambda}_{(i)}}}\right\| \lesssim_p T^{1/2} + T\left(q^{-1}N^{-1} + T^{-1}\right) \lesssim_p T^{1/2} + Tq^{-1}N^{-1}.$$

(ii) With (B.44) again, we have

$$\widetilde{Z}_{(k)}\bar{U}_{[I_0]}^{\intercal} = \bar{Z}\bar{U}_{[I_0]}^{\intercal} - \sum_{i=1}^{k-1}\bar{G}\xi_{(i)}\frac{\varsigma_{(i)}^{\intercal}\widetilde{U}_{(i)}\bar{U}_{[I_0]}^{\intercal}}{\sqrt{T\widehat{\lambda}_{(i)}}},$$

which, along with Lemma 7(i) and the assumptions on q, lead to

$$\begin{aligned}\left\|\widetilde{Z}_{(k)}\bar{U}_{[I_0]}^{\intercal}\right\| &\leq \left\|\bar{Z}\bar{U}_{[I_0]}^{\intercal}\right\| + \sum_{i=1}^{k-1}\left\|\bar{G}\xi_{(i)}\right\|\left\|\frac{\varsigma_{(i)}^{\intercal}\widetilde{U}_{(i)}}{\sqrt{T\widehat{\lambda}_{(i)}}}\right\|\left\|\bar{U}_{[I_0]}\right\| \\ &\lesssim_p N_0^{1/2}T^{1/2} + \left(q^{-1/2}N^{-1/2} + T^{-1}\right)\left(N_0^{1/2}T^{1/2} + T\right)\end{aligned}$$

$$\lesssim_p N_0^{1/2} T^{1/2} + Tq^{-1/2} N^{-1/2}.$$

□

Lemma 9. *Suppose that Assumptions A.1-A.8 hold. If* $\widehat{\lambda}_{(j)} \asymp_p |I_j|$ *and* $|I_j| \asymp qN$ *for* $j \leq \tilde{p}$*, then* H_1*,* H_2 *defined by* (B.54) *satisfy*

(i) $\|H_1\| \lesssim_p 1$*,* $\|H_2\| \lesssim_p 1$.

(ii) $\|H_1^\intercal H_2 - \mathbb{I}_{\tilde{p}}\| \lesssim_p T^{-1} + q^{-1}N^{-1}$.

(iii) $\|H_1 - H_2\| \lesssim_p T^{-1/2} + q^{-1}N^{-1}$.

Proof. (i) Using the definition (B.54) of H_1 and Lemma 1, we have

$$\|h_{k1}\| = \left\| \frac{\bar{V}\xi_{(k)}}{\sqrt{T}} \right\| \leq T^{-1/2} \left\| \bar{V} \right\| \lesssim_p 1,$$

which leads to $\|H_1\| \lesssim_p 1$.

Using the definition (B.54) of H_2, we have

$$\|h_{k2}\| = \left\| \frac{\widetilde{\beta}_{(k)}^\intercal \varsigma_{(k)}}{\sqrt{\widehat{\lambda}_{(k)}}} \right\| \leq q^{-1/2}N^{-1/2} \left\| \widetilde{\beta}_{(k)} \right\|. \tag{B.126}$$

With Lemma 1 and Lemma 7(i), we have

$$T^{1/2} \left\| \widetilde{\beta}_{(k)} \right\| \lesssim_p \sigma_p(\bar{V}) \left\| \widetilde{\beta}_{(k)} \right\| \leq \left\| \widetilde{\beta}_{(k)} \bar{V} \right\| \leq \left\| \widetilde{U}_{(k)} \right\| + \left\| \widetilde{R}_{(k)} \right\| \leq \left\| \widetilde{U}_{(k)} \right\| + \left\| \bar{R}_{[I_k]} \right\| \lesssim_p T^{1/2} q^{1/2} N^{1/2}. \tag{B.127}$$

Combining (B.126) and (B.127), we have $\|h_{k2}\| \lesssim_p 1$ and thus $\|H_2\| \lesssim_p 1$.

(ii) By (B.56) and Lemma 2, we have

$$\delta_{lk} = \xi_{(l)}^\intercal \xi_{(k)} = \frac{\xi_{(l)}^\intercal \bar{V}^\intercal \widetilde{\beta}_{(k)}^\intercal \varsigma_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}} + \frac{\xi_{(l)}^\intercal \widetilde{U}_{(k)}^\intercal \varsigma_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}} = h_{l1}^\intercal h_{k2} + \frac{\xi_{(l)}^\intercal \widetilde{U}_{(k)}^\intercal \varsigma_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}.$$

By Lemma 7(iii), we have

$$|h_{l1}^\intercal h_{k2} - \delta_{lk}| \lesssim_p q^{-1}N^{-1} + T^{-1},$$

and thus $\|H_1^\intercal H_2 - \mathbb{I}_{\tilde{p}}\| \lesssim_p q^{-1}N^{-1} + T^{-1}$.

(iii) Using (B.56), we have

$$\bar{V}\xi_{(k)} = \frac{\bar{V}\bar{V}^{\intercal}\widetilde{\beta}^{\intercal}_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}\varsigma_{(k)} + \frac{\bar{V}\widetilde{U}^{\intercal}_{(k)}\varsigma_{(k)}}{\sqrt{T\widehat{\lambda}_{(k)}}}.$$

With the definition of h_{k1} and h_{k2}, it becomes

$$h_{k1} = \frac{\bar{V}\bar{V}^{\intercal}}{T}h_{k2} + \frac{\bar{V}\widetilde{U}^{\intercal}_{(k)}\varsigma_{(k)}}{T\sqrt{\widehat{\lambda}_{(k)}}}. \tag{B.128}$$

With $\|h_{k2}\| \lesssim_p 1$, Lemma 1 and Lemma 7(ii), (B.128) leads to

$$h_{k1} - h_{k2} \lesssim_p T^{-1/2} + q^{-1}N^{-1}.$$

This completes the proof. □

Lemma 10. *For any $N \times p$ matrix β, if $\left\|T^{-1}\bar{V}\bar{V}^{\intercal} - \mathbb{I}_p\right\| \lesssim_p T^{-1/2}$, we have*

(i) $\sigma_j(\beta\bar{V})/\sigma_j(\beta) = T^{1/2} + O_p(1)$ for $j \leq p$.

(ii) If $\sigma_1(\beta) - \sigma_2(\beta) \asymp \sigma_1(\beta)$, then $\left\|\mathbb{P}_{\tilde{\xi}} - T^{-1}\bar{V}^{\intercal}\mathbb{P}_b\bar{V}\right\| \lesssim_p T^{-1/2}$, 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 $\left\|T^{-1}\bar{V}\bar{V} - \mathbb{I}_p\right\| \lesssim_p T^{-1/2}$, we have $T^{-1/2}\sigma_j(\beta\bar{V})/\sigma_j(\beta) = 1 + O_p\left(T^{-1/2}\right)$ by sin-theta theorem.

(ii) Let ς and $\tilde{\varsigma}$ be the first singular vectors of β and $\beta\bar{V}$, respectively. Equivalently, ς and $\tilde{\varsigma}$ are the eigenvectors of $\beta\beta^{\intercal}$ and $T^{-1}\beta\bar{V}\bar{V}^{\intercal}\beta^{\intercal}$. Since $\left\|\beta\beta^{\intercal} - T^{-1}\beta\bar{V}\bar{V}^{\intercal}\beta^{\intercal}\right\| \leq \|\beta\|^2\left\|T^{-1}\bar{V}\bar{V}^{\intercal} - \mathbb{I}_p\right\| \lesssim_p \sigma_1(\beta)^2T^{-1/2}$ and $\sigma_1(\beta) - \sigma_2(\beta) \asymp \sigma_1(\beta)$, by sin-theta theorem we have

$$\|\varsigma\varsigma^{\intercal} - \tilde{\varsigma}\tilde{\varsigma}^{\intercal}\| \lesssim \frac{\left\|\beta\beta^{\intercal} - T^{-1}\beta\bar{V}\bar{V}^{\intercal}\beta^{\intercal}\right\|}{\sigma_1(\beta)^2 - \sigma_2(\beta)^2 - O(\left\|\beta\beta^{\intercal} - T^{-1}\beta\bar{V}\bar{V}^{\intercal}\beta^{\intercal}\right\|)} \lesssim_p T^{-1/2}.$$

Using the relationship between left and right singular vectors, we have

$$b^{\intercal} = \frac{\varsigma^{\intercal}\beta}{\sigma_1(\beta)}, \quad \tilde{\xi}^{\intercal} = \frac{\tilde{\varsigma}^{\intercal}\beta\bar{V}}{\left\|\beta\bar{V}\right\|}.$$

Therefore,

$$\left\| \mathbb{P}_{\tilde{\xi}} - \frac{\sigma_1(\beta)^2}{\left\|\beta\bar{V}\right\|^2} \bar{V}^\intercal \mathbb{P}_b \bar{V} \right\| = \left\| \tilde{\xi}\tilde{\xi}^\intercal - \frac{\bar{V}^\intercal\beta^\intercal\varsigma\varsigma^\intercal\beta\bar{V}}{\left\|\beta\bar{V}\right\|^2} \right\| = \left\| \frac{\bar{V}^\intercal\beta^\intercal\tilde{\varsigma}\tilde{\varsigma}^\intercal\beta\bar{V}}{\left\|\beta\bar{V}\right\|^2} - \frac{\bar{V}^\intercal\beta^\intercal\varsigma\varsigma^\intercal\beta\bar{V}}{\left\|\beta\bar{V}\right\|^2} \right\| \lesssim_p T^{-1/2}. \quad \text{(B.129)}$$

By Weyl's inequality, we have $T^{-1}\left\|\beta\bar{V}\right\|^2 = \lambda_1(T^{-1}\beta\bar{V}\bar{V}^\intercal\beta^\intercal) = \lambda_1(\beta\beta^\intercal) + O_p(\sigma_1(\beta)^2 T^{-1/2}) = \sigma_1(\beta)^2 + O_p(\sigma_1(\beta)^2 T^{-1/2})$. Plugging this result into (B.129), we have $\left\|\mathbb{P}_{\tilde{\xi}} - T^{-1}\bar{V}^\intercal\mathbb{P}_b\bar{V}\right\| \lesssim_p T^{-1/2}$. □

Lemmas 11-13 below are concerned with the singular values and singular vectors of $T^{-1/2}\bar{R}$. We use ς_j, ξ_j and $\widehat{\lambda}_j^{1/2}$, $j \le p$ to denote them throughout Lemmas 11-13.

Lemma 11. *Under the assumptions of Theorem 5(a), we have*

$$\frac{\widehat{\lambda}_j}{\lambda_j} - 1 \lesssim_p \lambda_j^{-1/2}(T^{-1/2}N^{1/2} + 1) + T^{-1/2},$$

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

$$\lambda_j\left(\beta^\intercal\beta\right)\lambda_p\left(\frac{\bar{V}\bar{V}^\intercal}{T}\right) \le \frac{\lambda_j\left(\beta^\intercal\beta\bar{V}\bar{V}^\intercal\right)}{T} \le \lambda_j\left(\beta^\intercal\beta\right)\lambda_1\left(\frac{\bar{V}\bar{V}^\intercal}{T}\right). \quad \text{(B.130)}$$

By Lemma 1(i) and Weyl's inequality, we have $\lambda_j\left(T^{-1}\bar{V}\bar{V}^\intercal\right) - 1 \lesssim_p T^{-1/2}$ for $j \le p$. Then, (B.130) becomes

$$\frac{\lambda_j\left(\beta\bar{V}\bar{V}^\intercal\beta^\intercal\right)}{T\lambda_j(\beta^\intercal\beta)} - 1 \lesssim_p T^{-1/2},$$

which is equivalent to

$$\frac{\sigma_j\left(\beta\bar{V}\right)}{\sqrt{T}\sigma_j(\beta)} - 1 \lesssim_p T^{-1/2}. \quad \text{(B.131)}$$

Using Weyl's inequality again, we have $|\sigma_j\left(\bar{R}\right) - \sigma_j\left(\beta\bar{V}\right)| \le \left\|\bar{U}\right\| \lesssim_p N^{1/2} + T^{1/2}$, which is equivalent to

$$\frac{\widehat{\lambda}_j^{1/2}}{\lambda_j^{1/2}} - \frac{\sigma_j(\beta\bar{V})}{\sqrt{T}\sigma_j(\beta)} \lesssim_p \frac{1}{\sqrt{T}} + \frac{\sqrt{N}+\sqrt{T}}{\sqrt{T\lambda_j}}. \quad \text{(B.132)}$$

Combine (B.131) and (B.132), we complete the proof. □

Lemma 12. *Suppose that the SVD of β is given by:*

$$\beta = \Gamma^\intercal \begin{pmatrix} \Lambda^{\frac{1}{2}} \\ 0 \end{pmatrix} H, \tag{B.133}$$

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 $\Gamma\varsigma_j = (s_{j1}^\intercal, s_{j2}^\intercal)^\intercal$, where $s_{j1} \in \mathbb{R}^p$, $s_{j2} \in \mathbb{R}^{N-p}$. Then under the assumptions of Theorem 5(a), we have

(i) $\left\|(\Lambda/\lambda_j)^{1/2}\left(s_{j1} - \langle s_{j1}, e_{j1}\rangle e_{j1}\right)\right\| \lesssim_p \lambda_j^{-1/2}(T^{-1/2}N^{1/2}+1)$, *where e_{i1} is a $p\times 1$ unit vector with the ith entry being equal to 1.*

(ii) $\|s_{j1} - \langle s_{j1}, e_{j1}\rangle e_{j1}\| \lesssim_p \lambda_j^{-1/2}(T^{-1/2}N^{1/2}+1)$.

(iii) $\left\|(\Lambda/\lambda_j)^{1/2} s_{j1}\right\| \lesssim_p 1$.

(iv) $\|s_{j2}\| \lesssim_p \lambda_j^{-1/2}(T^{-1/2}N^{1/2}+1)$.

Proof. With the orthogonal matrix Γ defined above, we can write

$$\widetilde{U} = \Gamma\bar{U} = \begin{pmatrix} \widetilde{U}_{1_{p\times T}} \\ \widetilde{U}_{2_{(N-p)\times T}} \end{pmatrix}, \tag{B.134}$$

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 ς_j and ξ_j can be written as

$$\Gamma\varsigma_j = \frac{\left(\Gamma\bar{R}\right)\xi_j}{\sqrt{T\widehat{\lambda}_j}}, \quad \xi_j = \frac{\left(\Gamma\bar{R}\right)^\intercal\left(\Gamma\varsigma_j\right)}{\sqrt{T\widehat{\lambda}_j}}. \tag{B.135}$$

Specifically, we have

$$s_{j1} = \frac{\left(\Lambda^{\frac{1}{2}}\bar{V} + \widetilde{U}_1\right)\xi_j}{\sqrt{T\widehat{\lambda}_j}}, \quad s_{j2} = \frac{\widetilde{U}_2\xi_j}{\sqrt{T\widehat{\lambda}_j}}, \quad \xi_j = \frac{\left(\Lambda^{\frac{1}{2}}\bar{V} + \widetilde{U}_1\right)^\intercal s_{j1} + \widetilde{U}_2^\intercal s_{j2}}{\sqrt{T\widehat{\lambda}_j}}. \tag{B.136}$$

From (B.136), we have

$$\left(\Lambda^{\frac{1}{2}}\bar{V} + \widetilde{U}_1\right)\left(\Lambda^{\frac{1}{2}}\bar{V} + \widetilde{U}_1\right)^\intercal s_{j1} + \left(\Lambda^{\frac{1}{2}}\bar{V} + \widetilde{U}_1\right)\widetilde{U}_2^\intercal s_{j2} = T\widehat{\lambda}_j s_{j1}. \tag{B.137}$$

We can rewrite (B.137) as

$$\left(\mathbb{I}_p - \frac{\Lambda}{\lambda_j}\right) s_{j1} = \frac{1}{T\lambda_j}\left(\Lambda^{\frac{1}{2}}\bar{V} + \widetilde{U}_1\right)\widetilde{U}_2^\intercal s_{j2} + \frac{1}{\lambda_j}\Lambda^{\frac{1}{2}}\left(\frac{\bar{V}\bar{V}^\intercal}{T} - I\right)\Lambda^{\frac{1}{2}} s_{j1} + \frac{\Lambda^{\frac{1}{2}}\bar{V}\widetilde{U}_1^\intercal}{T\lambda_j}s_{j1}$$
$$+ \frac{\widetilde{U}_1\bar{V}^\intercal\Lambda^{\frac{1}{2}}}{T\lambda_j}s_{j1} + \frac{\widetilde{U}_1\widetilde{U}_1^\intercal}{T\lambda_j}s_{j1} - \left(\frac{\widehat{\lambda}_j}{\lambda_j} - 1\right)s_{j1}. \tag{B.138}$$

Define $L = \operatorname{diag}(l_1, \ldots, l_p)$, 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 (B.138), we have

$$s_{j1} - \langle s_{j1}, e_{j1}\rangle e_{j1} = \frac{1}{T\lambda_j}L\Lambda^{\frac{1}{2}}\bar{V}\frac{\widetilde{U}_2^\intercal\widetilde{U}_2}{\sqrt{T\widehat{\lambda}_j}}\xi_j + \frac{1}{T\lambda_j}L\widetilde{U}_1\frac{\widetilde{U}_2^\intercal\widetilde{U}_2}{\sqrt{T\widehat{\lambda}_j}}\xi_j + \frac{1}{\lambda_j}L\Lambda^{\frac{1}{2}}\left(\frac{\bar{V}\bar{V}^\intercal}{T} - \mathbb{I}_p\right)\Lambda^{\frac{1}{2}}s_{j1}$$
$$+ \frac{L\Lambda^{\frac{1}{2}}\bar{V}\widetilde{U}_1^\intercal}{T\lambda_j}s_{j1} + L\frac{\widetilde{U}_1\bar{V}^\intercal\Lambda^{\frac{1}{2}}}{T\lambda_j}s_{j1} + L\frac{\widetilde{U}_1\widetilde{U}_1^\intercal}{T\lambda_j}s_{j1} - \left(\frac{\widehat{\lambda}_j}{\lambda_j} - 1\right)Ls_{j1}. \tag{B.139}$$

Now left multiplying $\left(\frac{\Lambda}{\lambda_j}\right)^{\frac{1}{2}}$ again, we have

$$\left(\frac{\Lambda}{\lambda_j}\right)^{\frac{1}{2}}(s_{j1} - \langle s_{j1}, e_{j1}\rangle e_{j1}) = \frac{1}{T\lambda_j^{3/2}}\Lambda^{\frac{1}{2}}L\Lambda^{\frac{1}{2}}\bar{V}\frac{\widetilde{U}_2^\intercal\widetilde{U}_2}{\sqrt{T\widehat{\lambda}_j}}\xi_j + \frac{1}{T\lambda_j^{3/2}}\Lambda^{\frac{1}{2}}L\widetilde{U}_1\frac{\widetilde{U}_2^\intercal\widetilde{U}_2}{\sqrt{T\widehat{\lambda}_j}}\xi_j$$
$$+ \frac{1}{\lambda_j}\Lambda^{\frac{1}{2}}L\Lambda^{\frac{1}{2}}\left(\frac{\bar{V}\bar{V}^\intercal}{T} - \mathbb{I}_p\right)\left(\frac{\Lambda}{\lambda_j}\right)^{\frac{1}{2}}s_{j1} + \Lambda^{\frac{1}{2}}L\Lambda^{\frac{1}{2}}\frac{\bar{V}\widetilde{U}_1^\intercal}{T\lambda_j^{3/2}}s_{j1}$$
$$+ \Lambda^{\frac{1}{2}}L\frac{\widetilde{U}_1\bar{V}^\intercal}{T\lambda_j}\left(\frac{\Lambda}{\lambda_j}\right)^{\frac{1}{2}}s_{j1} + \Lambda^{\frac{1}{2}}L\frac{\widetilde{U}_1\widetilde{U}_1^\intercal}{T\lambda_j^{3/2}}s_{j1} - \left(\frac{\widehat{\lambda}_j}{\lambda_j} - 1\right)\left(\frac{\Lambda}{\lambda_j}\right)^{\frac{1}{2}}Ls_{j1}$$
$$= K_1 + K_2 + K_3 + K_4 + K_5 + K_6 + K_7. \tag{B.140}$$

Before we analyze these seven terms in (B.140), we first analyze $\|L\|$, $\left\|L\Lambda^{1/2}\right\|$ and $\|L\Lambda\|$. Since L and Λ are diagonal matrices, by Assumption A.13 we can easily show that

$$\|L\| \lesssim 1, \quad \left\|L\Lambda^{1/2}\right\| \lesssim \lambda_j^{1/2}, \quad \|L\Lambda\| \lesssim \lambda_j. \tag{B.141}$$

In addition, Lemma 1(ii)(iii)(v) imply that

$$\left\|\widetilde{U}_1\right\| = \left\|(\beta^\intercal\beta)^{-1/2}\beta^\intercal\bar{U}\right\| \lesssim_p T^{1/2}, \quad \left\|\widetilde{U}_1\bar{V}^\intercal\right\| = \left\|(\beta^\intercal\beta)^{-1/2}\beta^\intercal\bar{U}\bar{V}^\intercal\right\| \lesssim_p T^{1/2}, \quad \left\|\widetilde{U}_2\right\| \leq \left\|\bar{U}\right\| \lesssim_p N^{1/2} + T^{1/2}. \tag{B.142}$$

Using Lemma 1(i)(vi), Lemma 11, (B.141) and (B.142), we analyze these seven terms in (B.140) one by one. For the first term, we have

$$\|K_1\| \leq T^{-3/2}\lambda_j^{-3/2}\widehat{\lambda}_j^{-1/2}\left\|L\Lambda\right\|\left\|\bar{V}\right\|\left\|\widetilde{U}_2^{\intercal}\widetilde{U}_2\right\|\|\xi_j\| \lesssim_p \lambda_j^{-1}(T^{-1}N+1),$$

where we also use $\left\|\widetilde{U}_2^{\intercal}\widetilde{U}_2\right\| \leq \left\|\bar{U}^{\intercal}\bar{U}\right\| \lesssim_p N+T$ in the last equation. For the second term, we have

$$\|K_2\| \leq T^{-3/2}\lambda_j^{-3/2}\widehat{\lambda}_j^{-1/2}\left\|\Lambda^{1/2}L\right\|\left\|\widetilde{U}_1\right\|\left\|\widetilde{U}_2^{\intercal}\widetilde{U}_2\right\|\|\xi_j\| \lesssim_p \lambda_j^{-3/2}(T^{-1}N+1).$$

For the third term, we have

$$\|K_3\| \leq \lambda_j^{-1}\left\|L\Lambda\right\|\left\|T^{-1}\bar{V}\bar{V}^{\intercal}-\mathbb{I}_p\right\|\left\|(\Lambda/\lambda_j)^{1/2}s_{j1}\right\| \lesssim_p T^{-1/2}\left\|(\Lambda/\lambda_j)^{1/2}s_{j1}\right\|.$$

For the forth term, we have

$$\|K_4\| \leq T^{-1}\lambda_j^{-3/2}\left\|L\Lambda\right\|\left\|\bar{V}\widetilde{U}_1^{\intercal}\right\| \lesssim_p \lambda_j^{-1/2}T^{-1/2},$$

where we use $\left\|\bar{V}\widetilde{U}_1^{\intercal}\right\| \lesssim_p T^{1/2}$ from Lemma 1. For the fifth term, we have

$$\|K_5\| \leq T^{-1}\lambda_j^{-1}\left\|L\Lambda^{1/2}\right\|\left\|\widetilde{U}_1\bar{V}^{\intercal}\right\|\left\|(\Lambda/\lambda_j)^{1/2}s_{j1}\right\| \lesssim_p \lambda_j^{-1/2}T^{-1/2}\left\|(\Lambda/\lambda_j)^{1/2}s_{j1}\right\|.$$

For the sixth term, we have

$$\|K_6\| \leq T^{-1}\lambda_j^{-3/2}\left\|L\Lambda^{1/2}\right\|\left\|\widetilde{U}_1\widetilde{U}_1^{\intercal}\right\| \lesssim_p \lambda_j^{-1},$$

where we use $\left\|\widetilde{U}_1\widetilde{U}_1^{\intercal}\right\| \lesssim_p T$ as shown in Lemma 1. For the last term, we have

$$\|K_7\| \leq \lambda_j^{-2}|\widehat{\lambda}_j-\lambda_j|\left\|L\Lambda^{1/2}\right\| \lesssim_p \lambda_j^{-1/2}(T^{-1/2}N^{1/2}+1)+T^{-1/2}.$$

To sum up, (B.140) gives

$$\left\|(\Lambda/\lambda_j)^{1/2}\left(s_{j1}-\langle s_{j1},e_{j1}\rangle e_{j1}\right)\right\| \lesssim_p \lambda_j^{-1/2}(T^{-1/2}N^{1/2}+1)+T^{-1/2}+T^{-1/2}\left\|(\Lambda/\lambda_j)^{1/2}s_{j1}\right\|. \tag{B.143}$$

Note that

$$\begin{aligned}\left\|(\Lambda/\lambda_j)^{1/2}s_{j1}\right\| &\leq \left\|(\Lambda/\lambda_j)^{1/2}\left(s_{j1}-\langle s_{j1},e_{j1}\rangle e_{j1}\right)\right\| + \left\|(\Lambda/\lambda_j)^{1/2}\langle s_{j1},e_{j1}\rangle e_{j1}\right\| \\ &\leq \left\|(\Lambda/\lambda_j)^{1/2}\left(s_{j1}-\langle s_{j1},e_{j1}\rangle e_{j1}\right)\right\| + |\langle s_{j1},e_{j1}\rangle|\sqrt{\lambda_j^{-1}e_{j1}^{\intercal}\Lambda e_{j1}} \\ &= \left\|(\Lambda/\lambda_j)^{1/2}\left(s_{j1}-\langle s_{j1},e_{j1}\rangle e_{j1}\right)\right\| + O_p(1).\end{aligned}$$

Plugging this into (B.143), we have

$$\left\|(\Lambda/\lambda_j)^{1/2}\left(s_{j1}-\langle s_{j1},e_{j1}\rangle e_{j1}\right)\right\| \lesssim_p \lambda_j^{-1/2}(T^{-1/2}N^{1/2}+1)+T^{-1/2}, \tag{B.144}$$

which in turn leads to $\left\|(\Lambda/\lambda_j)^{1/2}s_{j1}\right\| \lesssim_p 1$ as by assumption $\lambda_j^{-1/2}(T^{-1/2}N^{1/2}+1)\to 0$. Similarly, we can analyze corresponding terms in (B.139), and obtain

$$\|s_{j1}-\langle s_{j1},e_{j1}\rangle e_{j1}\| \lesssim_p T^{-1/2}\left\|(\Lambda/\lambda_j)^{1/2}s_{j1}\right\| + \lambda_j^{-1/2}(T^{-1/2}N^{1/2}+1) \lesssim_p \lambda_j^{-1/2}(T^{-1/2}N^{1/2}+1)+T^{-1/2}.$$

From (B.136), we have

$$\|s_{j2}\| \le \left\|\frac{\widetilde{U}_2}{\sqrt{T\lambda_j}}\right\| \left\|\left(\frac{\lambda_j}{\widehat{\lambda}_j}\right)^{\frac{1}{2}}\right\| \|\xi_j\| \lesssim_p \lambda_j^{-1/2}(T^{-1/2}N^{1/2}+1). \tag{B.145}$$

This concludes the proof. □

Lemma 13. *Under the assumptions of Theorem 5(a), we have*

(i) $\left\|\frac{\xi_i^\intercal \bar{U}^\intercal \varsigma_j}{\sqrt{T\widehat{\lambda}_j}}\right\| \lesssim_p \frac{1}{T}+\frac{N+T}{T\lambda_i}+\frac{N+T}{T\lambda_j}$.

(ii) $\left\|\frac{\bar{V}\bar{U}^\intercal \varsigma_i}{T\sqrt{\widehat{\lambda}_i}}\right\| \lesssim_p \frac{1}{T}+\frac{N+T}{T\lambda_i}$, $\quad |\frac{\varsigma_i^\intercal \bar{u}}{\sqrt{\widehat{\lambda}_i}}| \lesssim_p \frac{1}{T}+\frac{N+T}{T\lambda_i}$.

(iv) $\left\|\frac{\varsigma_i^\intercal \bar{U}}{\sqrt{T\widehat{\lambda}_i}}\right\| \lesssim_p \frac{1}{\sqrt{\lambda_i}}+\frac{N+T}{T\lambda_i}$.

Proof. (i) From (B.135), we have

$$\frac{\xi_i^\intercal \bar{U}^\intercal \varsigma_j}{\sqrt{T\widehat{\lambda}_j}} = \frac{\varsigma_i^\intercal \bar{R}\bar{U}^\intercal \varsigma_j}{T\sqrt{\widehat{\lambda}_i\widehat{\lambda}_j}}.$$

Using the orthogonal matrix Γ and the notations in Lemma 11 and Lemma 12, we have

$$\begin{aligned}\varsigma_i^\intercal \bar{R}\bar{U}^\intercal\varsigma_j = s_i^\intercal\left(\Gamma\beta\bar{V}+\widetilde{U}\right)\widetilde{U}^\intercal s_j =& s_{i1}^\intercal\left(\Lambda^{\frac{1}{2}}\bar{V}+\widetilde{U}_1\right)\widetilde{U}_1^\intercal s_{j1} + s_{i2}^\intercal\widetilde{U}_2\widetilde{U}_1^\intercal s_{j1}\\ &+ s_{i1}^\intercal\left(\Lambda^{\frac{1}{2}}\bar{V}+\widetilde{U}_1\right)\widetilde{U}_2^\intercal s_{j2} + s_{i2}^\intercal\widetilde{U}_2\widetilde{U}_2^\intercal s_{j2}\\ =& K_1+K_2+K_3+K_4.\end{aligned}$$

Recall that from Lemma 12, we have $\left\|(\Lambda/\lambda_j)^{1/2}s_{j1}\right\| \lesssim_p 1$. Using this result and Lemma 1, we analyze these four terms one by one. For the first term, we have

$$\|K_1\| \le \left\|s_{i1}^\intercal\Lambda^{\frac{1}{2}}\right\| \left\|\bar{V}\widetilde{U}_1^\intercal\right\| \|s_{j1}\| + \|s_{i1}\| \left\|\widetilde{U}_1\widetilde{U}_1^\intercal\right\| \|s_{j1}\| \lesssim_p \sqrt{\lambda_i T}+T.$$

For the second term, we have

$$\|K_2\| \leq \|s_{i2}\| \left\|\widetilde{U}_2\right\| \left\|\widetilde{U}_1\right\| \lesssim_p \sqrt{\frac{N+T}{T\lambda_i}} \left(\sqrt{N}+\sqrt{T}\right)\sqrt{T} \lesssim_p \lambda_i^{-1/2}(N+T).$$

For the third term, we have

$$\|K_3\| \leq \left(\left\|s_{i1}^\intercal \Lambda^{\frac{1}{2}}\right\| \left\|\bar{V}\right\| + \left\|\widetilde{U}_1\right\|\right) \left\|\widetilde{U}_2\right\| \|s_{j2}\| \lesssim_p \sqrt{\lambda_i T}\left(\sqrt{N}+\sqrt{T}\right)\sqrt{\frac{N+T}{T\lambda_j}} = \lambda_j^{-1/2}\lambda_i^{1/2}(N+T).$$

For the last term, we have

$$\|K_4\| \leq \left\|\widetilde{U}_2\widetilde{U}_2^\intercal\right\| \|s_{i2}\| \|s_{j2}\| \lesssim_p \lambda_i^{-1/2}\lambda_j^{-1/2}T^{-1}(N+T)^2.$$

Using above equations and Lemma 11, we get

$$\left\|\frac{\xi_i^\intercal \bar{U}^\intercal \varsigma_j}{\sqrt{T\widehat{\lambda}_j}}\right\| = \left\|\frac{\varsigma_i^\intercal \bar{R}\bar{U}^\intercal \varsigma_j}{T\sqrt{\widehat{\lambda}_i\widehat{\lambda}_j}}\right\| \lesssim_p \frac{1}{T} + \frac{N+T}{T\lambda_i} + \frac{N+T}{T\lambda_j}.$$

(ii) Using $\bar{U}^\intercal\varsigma_i = \widetilde{U}_1^\intercal s_{i1} + \widetilde{U}_2^\intercal s_{i2}$ and (B.142), we have

$$\left\|\bar{V}\bar{U}^\intercal\varsigma_i\right\| \leq \left\|\bar{V}\widetilde{U}_1^\intercal s_{i1}\right\| + \left\|\bar{V}\widetilde{U}_2^\intercal s_{i2}\right\| \leq \left\|\bar{V}\widetilde{U}_1^\intercal\right\| + \left\|\bar{V}\right\| \left\|\bar{U}\right\| \|s_{i2}\| \lesssim_p \sqrt{T} + \frac{N+T}{\sqrt{\lambda_i}}.$$

Then, with Lemma 11, we have $\left\|T^{-1}\widehat{\lambda}_i^{-1/2}\bar{V}\bar{U}^\intercal\varsigma_i\right\| \lesssim_p T^{-1} + \lambda_i^{-1}(T^{-1}N+1)$.

Replace $\bar{V}$ in the above proof by $\iota_T^\intercal$, we can get $\left\|\widehat{\lambda}_i^{-1/2}\bar{u}^\intercal\varsigma_i\right\| \lesssim_p T^{-1} + \lambda_i^{-1}(T^{-1}N+1)$.
(iii) Using $\bar{U}^\intercal\varsigma_i = \widetilde{U}_1^\intercal s_{i1} + \widetilde{U}_2^\intercal s_{i2}$ and (B.142), we have

$$\left\|\varsigma_i^\intercal \bar{U}\right\| \leq \left\|s_{i1}^\intercal \widetilde{U}_1\right\| + \left\|s_{i2}^\intercal \widetilde{U}_2\right\| \leq \left\|\widetilde{U}_1\right\| + \left\|\bar{U}\right\| \lesssim_p \sqrt{T} + \frac{N+T}{\sqrt{T\lambda_i}}.$$

Applying Lemma 11 again completes the proof. □

Lemma 14. *Under the assumptions of Theorem 5(a), $\tilde{H}_1$, $\tilde{H}_2$ defined by (B.75) satisfy*

(i) $\left\|\tilde{H}_1\right\| \lesssim_p 1$, $\left\|\tilde{H}_2\right\| \lesssim_p 1$.

(ii) $\left\|\tilde{H}_1^\intercal\tilde{H}_2 - \mathbb{I}_{\tilde{p}}\right\| \lesssim_p T^{-1} + \lambda_p^{-1}(T^{-1}N+1)$.

(iii) $\left\|\tilde{H}_1 - \tilde{H}_2\right\| \lesssim_p T^{-1/2} + \lambda_p^{-1}(T^{-1}N+1)$.

Proof. (i) Using the definition of $\tilde{H}_1$ in (B.75) and Lemma 1, we have

$$\left\|\tilde{h}_{k1}\right\| = \left\|\frac{\bar{V}\xi_k}{\sqrt{T}}\right\| \leq T^{-1/2}\left\|\bar{V}\right\| \lesssim_p 1,$$

which leads to $\left\|\tilde{H}_1\right\| \lesssim_p 1$.

Using $\Gamma\varsigma_k = (s_{k1}^\intercal, s_{k2}^\intercal)^\intercal$, the SVD of β in (B.133), the definition of $\tilde{H}_2$ in (B.75), Lemma 11 and Lemma 12(iii), we have

$$\left\|\tilde{h}_{k2}\right\| = \left\|\frac{\beta^\intercal\varsigma_k}{\sqrt{\widehat{\lambda}_k}}\right\| = \left\|\frac{\Lambda^{1/2}s_{k1}}{\sqrt{\widehat{\lambda}_k}}\right\| \lesssim_p 1, \tag{B.146}$$

which leads to $\left\|\tilde{H}_2\right\| \lesssim_p 1$.

(ii) By (B.135) and Lemma 2, for $l, k \leq p$, we have

$$\delta_{lk} = \xi_l^\intercal\xi_k = \frac{\xi_l^\intercal\bar{V}^\intercal\beta^\intercal\varsigma_k}{\sqrt{T\widehat{\lambda}_k}} + \frac{\xi_l^\intercal\bar{U}^\intercal\varsigma_k}{\sqrt{T\widehat{\lambda}_k}} = \tilde{h}_{l1}^\intercal\tilde{h}_{k2} + \frac{\xi_l^\intercal\bar{U}^\intercal\varsigma_k}{\sqrt{T\widehat{\lambda}_k}}.$$

By Lemma 13(i), we have

$$|\tilde{h}_{l1}^\intercal\tilde{h}_{k2} - \delta_{lk}| \lesssim_p \frac{1}{T} + \frac{N+T}{T\min\{\lambda_l, \lambda_k\}} \leq \frac{1}{T} + \frac{N+T}{T\lambda_p},$$

and thus $\left\|\tilde{H}_1^\intercal\tilde{H}_2 - \mathbb{I}_p\right\| \lesssim_p T^{-1} + \lambda_p^{-1}(T^{-1}N + 1)$.

(iii) Using (B.135), we have

$$\bar{V}\xi_k = \frac{\bar{V}\bar{V}^\intercal\beta^\intercal}{\sqrt{T\widehat{\lambda}_k}}\varsigma_k + \frac{\bar{V}\bar{U}^\intercal\varsigma_k}{\sqrt{T\widehat{\lambda}_k}}.$$

With the definition of h_{k1} and h_{k2}, it becomes

$$\tilde{h}_{k1} = \frac{\bar{V}\bar{V}^\intercal}{T}\tilde{h}_{k2} + \frac{\bar{V}\bar{U}^\intercal\varsigma_k}{T\sqrt{\widehat{\lambda}_k}}. \tag{B.147}$$

With $\left\|\tilde{h}_{k2}\right\| \lesssim_p 1$, Lemma 1 and Lemma 13(ii), (B.147) leads to

$$\left\|\tilde{h}_{k1} - \tilde{h}_{k2}\right\| \leq \left\|T^{-1}\bar{V}\bar{V}^\intercal - \mathbb{I}_p\right\|\left\|\tilde{h}_{k2}\right\| + \left\|\frac{\bar{V}\bar{U}^\intercal\varsigma_k}{T\sqrt{\widehat{\lambda}_k}}\right\| \lesssim_p T^{-1/2} + \lambda_p^{-1}(T^{-1}N + 1),$$

which concludes the proof of (iii). □

Lemma 15. *Under Assumption A.13, we have*

$$\left\|\bar{r} - \widehat{\Sigma}b\right\|_\infty \lesssim_p \sqrt{\frac{\log N}{T}}, \qquad \|b^\intercal(\bar{r} - \mathrm{E}(r_t))\| \lesssim_p \frac{1}{\sqrt{T}}.$$

Proof. For the first inequality, we have

$$\left\|\bar{r}-\widehat{\Sigma}b\right\|_\infty \leq \|\bar{r}-\mathrm{E}(r)\|_\infty + \left\|\Sigma b-\widehat{\Sigma}b\right\|_\infty \lesssim_p \sqrt{\frac{\log N}{T}},$$

where we use large deviation inequalities in Assumption A.12:

$$\|\bar{r}-\mathrm{E}(r_t)\|_\infty \lesssim_p \sqrt{\frac{\log N}{T}}, \quad \text{and} \quad \left\|\Sigma b-\widehat{\Sigma}b\right\|_\infty = \left\|\frac{1}{T}\bar{R}\bar{R}^\intercal b - \mathrm{Cov}(r_t, r_t^\intercal b)\right\|_\infty \lesssim_p \sqrt{\frac{\log N}{T}}.$$

The second inequality follows immediately from Assumption A.12:

$$\|b^\intercal(\bar{r}-\mathrm{E}(r_t))\| = |\frac{1}{T}\sum_{t=1}^{T} m_t - \mathrm{E}(m_t)| \lesssim_p \frac{1}{\sqrt{T}}.$$

□

References

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

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

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

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

Wang, W. and J. Fan (2017, 06). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Ann. Statist. 45*(3), 1342–1374.

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