Working Paper No. 21-04

Test Assets and Weak Factors

Stefano Giglio
Yale University, NBER, and CEPR

Dacheng Xiu
University of Chicago Booth School of Business

Dake Zhang
University of Chicago Booth School of Business

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: January 17, 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 using it to estimate the risk premia of several tradable and nontradable factors.

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

*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 an indispensable component of the empirical work in asset pricing. In this exercise, an essential role is played by the cross-section of assets, namely, test assets, used in the estimation. Yet, little work has been dedicated to investigating rigorously and systematically how test assets should be chosen, and what the pros and cons are of different choices.

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). Another approach, taken more recently, has been to expand this cross-section to include portfolios sorted by a much larger set of characteristics, following the expansion of empirical work in recent decades documenting that many additional characteristics seem to associate with risk premia. Finally, a third approach has been more "targeted" around the specific factor of interest: first sorting assets into portfolios by their exposure to the factor, and then estimating risk premia using only these sorted portfolios, that is, only a small cross-section expected to be particularly informative about that factor.

We argue in this paper that all of these methods to choose test assets have important shortcomings, and propose a new methodology to select test assets and use them to estimate risk premia that addresses them. We refer to our estimator as a *supervised-PCA* (SPCA in short) estimator of risk premia, and prove its consistency and asymptotic properties. Central to understanding the shortcomings of existing methodologies – and the contribution of our procedure – is the concept of *weak factor*. Weak factors – factors that are not well captured in the panel of test assets – have been discussed at length in the literature. Not only identifying the risk premia of weak factors themselves is difficult; the presence of weak factors in a model biases the estimation and inference of the other factors as well, including strong ones.

In this paper, we start from an alternative perspective to think about weak factors. Rather than thinking about weakness of a factor as a property of the factor, we should think about it as a property of the set of test assets chosen for the estimation. *Any factor could be weak or strong depending on which test assets we look at.* As a simple but illustrative 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. Importantly, the choice of test assets determines the strength not only of the factor of interest (e.g., liquidity), but also of *all* other factors that drive the stochastic discount factor. As discussed in the literature (e.g., Jagannathan and Wang (1998) and Giglio and Xiu (2020)), estimating and testing the risk premia of some factors requires properly controlling for the test assets exposure to all the other relevant factors, in order to avoid an omitted variable bias problem. The choice of test assets therefore has both a direct effect on the factor of interest, and on all the other factors in the model (whether they are observed or latent).

The SPCA methodology we propose addresses directly the issue of weak factors, both observable and latent, by combining principal component estimation methods with systematic asset selection; it also addresses the potential for omitted factor bias, whether it is due to strong or weak omitted factors.

In a nutshell, the procedure works recursively in the following way. We start from a large cross-section of assets that is as large as possible. In a first step of the procedure, we compute the univariate correlation

of each asset with the factor of interest. 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. We then compute the first principal component of these portfolios, which will be our first estimated factor. Then, we remove (from both the factor and the returns) the part explained by this first factor, and compute the univariate correlation of the *residuals* of the factor and the *residuals* of the assets. Again, we select among the universe of test assets those for which this correlations is especially high, and compute the principal component of these assets. This will be our second estimated factor. We then further remove (from the factor and the test assets) the part explained by this 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 (strong and weak) in the data or 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. Importantly, the fact that at each step only test assets that are sufficiently correlated with the factor 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. As a consequence, our methodology is able to consistently estimate the risk premium of any factor whether strong or weak in the cross-section of assets, and even in the presence of latent weak factors.

It is useful to contrast the procedure to select test assets with the three standard approaches to select test assets summarized above. Using a standard, small cross-sections (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 premium. Using a large cross-section of test assets (the second approach) may appear, on the surface, to address this issue: a large cross-section contains returns that are exposed to a large number of underlying factors. However, and importantly, if only a *few* assets are exposed to some factor, while most others are not, that factor will be weak in this large cross-section, again yielding biased estimation and inference. 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; so it will generally yield biased estimates (intuitively, exposures with the factor of interest may capture correlated exposures to other risks in the economy).

In fact, one way to interpret the SPCA methodology is that it combines several important strengths of these three choices of testing cross-sections. The selection step, that focuses on assets that covary sufficiently with the factor of interest, inherits the main idea of the "targeted" approach: to learn the most information from a few assets that are particularly informative about the factor, discarding assets that do not contain or contain less information about it. At the same time, SPCA also starts from a large cross-section, exploiting the fact that large cross-sections will contain (somewhere among the many assets) risk exposure to many latent factors, sometimes in strong form, sometimes in weak form. Finally, our method combines these insights with the results of Giglio and Xiu (2020) – that extracting latent factors from the panel of returns and controlling for them can solve the issue of omitted factor bias when estimating risk premia. The main difference between the two papers is that SPCA integrates the selection step at each stage of the estimation of the latent factors to solve the weak factor problem.

A closely related (and essentially the same) problem of weak factors is that of highly correlated factor

exposures. In a multi-factor model, if two factors share similar risk exposures, for instance, two different versions of the liquidity factor are both included in the model, the Fama-MacBeth procedure suffers from the identification failure due to multicolinearity, which results in unstable risk premia estimates. The PCA procedure by Giglio and Xiu (2020) helps resolve this issue because their risk premia estimates for observable factors are constructed separately one at a time. Their assumption, though, requires strong identification of all latent factors. In this paper, we further discuss the case in which latent factors might also have highly correlated exposures to a given set of test assets. Even if all latent factors are strong in the sense that their corresponding betas show considerable variation individually, these betas could be highly correlated, which leads to the same symptoms as that due to weak factors. We show that our SPCA procedure provides a cure in this case as well because an equivalent rotated representation of this model exactly translates this correlated exposures problem to the same problem of weak factors.

In the paper, we formally derive the SPCA approach to estimating factor risk premia and constructing the stochastic discount factor, allowing for weak factors and test assets with highly correlated risk exposures. To justify our approach, we provide asymptotic properties of SPCA as well as alternative estimators 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 SDF estimation (and hence risk premia estimation) but is not as efficient as our SPCA in general. Only when the true factors are part of the test assets can the Lasso estimator achieve a comparable efficiency as our SPCA.

After deriving the theory of SPCA and simulations, we present empirical evidence on the performance of the SPCA estimator for a variety of tradable and nontradable factors studied in the literature. We start from the large cross-section of test portfolios produced by Chen and Zimmermann (2020), covering more than 700 portfolios for the period 1976-2019. We then apply SPCA to estimate the factor risk premia, and compare it with the alternative methodologies, such as PCA of Giglio and Xiu (2020), rpPCA motivated by an SDF estimator of Lettau and Pelger (2020), and a PLS version of Giglio and Xiu (2020) (see Kelly and Pruitt (2013) for introducing PLS in a different asset pricing context). As we discuss further in the paper, each of these methods is effectively building a regularized mimicking portfolio for the factor that exploits the assumptions of the factor structure of the SDF; therefore, a relevant criterion to evaluate the different methodologies is the R^2 of the projection of the factor onto the extracted latent factors. When we compare different models fully out of sample (both based on the risk premia estimates and on the time-series R^2) a few interesting patterns emerge. For the case of strong factors, all methodologies find similar results and perform similarly; the same happens in the case of completely spurious factors. More interestingly, however, different methodologies give different answers for the intermediate case of weak factors. In that case, SPCA shows more robust results, achieving consistently higher out of sample R^2, and in a way that is much more robust to the number of factors p used in the estimation. The robustness with respect to the number of factors used is closely related to the selection step of the methodology: because it selects assets that are highly correlated with the factor of interest, the methodology zooms in quickly on latent factors that are especially informative about the factor. For example, a standard cross-section might have 5 strong factors unrelated to the factor of interest; the 6th principal component is the one that captures the exposure to that factor. Standard PCA will then require 6 factors to properly uncover and estimate the risk premium.

SPCA, on the contrary, might need much fewer factors to do so, because from the very beginning it selects assets that are informative about the factor of interest, discarding assets that do not have exposure to the factor.

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 pioneer contribution of Kan and Zhang (1999) shows that the inference on risk premia estimates from Fama-MacBath regression 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. This issue is quite relevant in practice because many test assets are not very sensitive to macroeconomic variables. Moreover, the same 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)).

To estimate the risk premia of these factors, Giglio and Xiu (2020) suggest an alternative approach to Fama-MacBeth regression, which helps solve the issue of omitted factor bias. The key assumption behind is that the true DGP of returns is driven by latent but strong factors, so that these factors, and in turn the SDF, can be recovered by the principal component analysis (PCA). Therefore, the covariance between any factor and the estimated SDF yields the factor's risk premium. In this paper, we focus on the more challenging (and general) case in which the latent factors are not strong, but can be arbitrarily weak. This is not a minor issue: given the large number of factors and test assets in the literature, it is in fact natural that in any cross-section of test assets, at least some factors will be weak, rather than strong. The weak factor problem is in fact quite pervasive in the data.

Recently, Lettau and Pelger (2020) propose an estimator of the SDF in the presence of weak factors by generalizing the PCA with a penalty term that accounts for pricing errors in expected returns; they refer to the estimator as risk premium PCA, or *rpPCA*. Their paper is among the very first to directly tackle the presence of weak latent factors in the SDF. The SDF estimated using this procedure can then be used to estimate risk premia (since risk premia are covariances with the SDF). While they have shown their estimator of SDF can lead to higher Sharpe ratios empirically, the very assumptions under which rpPCA is derived are so restrictive that they preclude any possibility of consistent estimation of risk premia or the SDF itself. On the contrary, we produce an estimator that we prove is consistent for risk premia in a less restrictive environment, and even when weak factors are present (and in addition, we derive asymptotic inference on risk premia as well). More specifically, we do not require N and T to diverge at the same rate; we do not require the factor loadings to be orthogonal; we do not require the idiosyncratic errors to be multivariate Gaussian; we do not require all factors to be strong or all to be weak. Our key identifying assumption is that 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. This assumption is sufficiently weak to the extent that it allows for weak factors that the standard PCA fails to recover, yet this assumption is just strong enough to ensure identifiability (i.e., separation of factors and idiosyncratic errors), so that consistent estimators exist.

Also related is Pesaran and Smith (2019), who investigate the effect of factor strength and pricing error in

risk premium estimation. They point out that the conventional two-pass risk premium estimator converges at a lower rate as the factors become weaker. However, even if all factors are strong, some factors could be highly correlated, so that the weak factor problem also arises. We propose a new approach that addresses both issues with the classic approach. Anatolyev and Mikusheva (2018) propose a sample-splitting approach to address the issues of weak factors and missing factors. Our assumptions on weak factors are more general, and we also allow for priced missing factors.

We argue the selection of test assets should account for the issues of weak factors and highly correlated risk exposures. We provide asymptotic analysis to justify our approach as well as develop valid statistical inference for our procedure. From a different perspective, Ahn et al. (2009) suggest forming portfolios as test assets by clustering individual securities based on their correlations so that securities within clusters are similar but different across cluster. There is, however, not a clear theoretical rationale behind this proposal. We suggest using correlations between test assets and individual factors, whereas their approach requires correlations among all test assets, which could be computationally expensive, if the number of test assets is large. Moreover, their proposal does not help resolve the aforementioned issues, which is the main focus of our paper.

Bryzgalova et al. (2020) suggest constructing test assets by sequentially sorting characteristics and then select test assets by maximizing the Sharpe ratio of a portfolio comprised of these sorted portfolios. They show that forming test assets this way creates better investment opportunities than the conventional sorting methods, which leads to an estimated SDF with higher Sharpe ratios. Our proposal is agnostic about how one should build a large cross-section of test assets. Our analysis sheds light on the inherent connection between test assets selection and the strength of factors, and provides the missing theoretical rationale on how test assets should be selected to alleviate the problem of weak factors.

There is a growing strand of econometrics literature on weak factor models. Bai and Ng (2008) argue that the properties of idiosyncratic errors should be considered when constructing principal components. Dropping some data, if they are noisy, may improve the forecasting. They compare the performance of hard thresholding, Lasso, the elastic net and Least angle regressions for the selection of subsets for factor estimation. Our SPCA approach shares the spirit of theirs, but is more involved because we allow for multiple selection steps. Our focus is on risk premia estimation instead of forecasting, for which we also provide inference. Similar to our paper, Bailey et al. (2020) also assume a sparse structure on the loading matrix of factor exposure. Under this assumption, they propose a measure of factor strength. Freyaldenhoven (2019) proposes an estimator of the number of factors in the presence of weak factors, though the notion of "weak" factors is somewhat strong because the principal component analysis in his setting can still recover these "weak" factors consistently.

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. 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), but our projection step is new, which is precisely introduced to eliminate the strong identification assumption used before. Also, our focus is not on prediction per se (which resembles the step of building mimicking portfolios in our setting), but instead on inference on parameters (i.e., risk premia) that 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, both $\|a\|$ and $\|a\|_{\mathrm{F}}$ are equal to its Euclidean norm. 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

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.[1] 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. Also, v_t can include latent factors, if any.

Prior to the discussion of factor strength, we need a well-defined 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 $1 \lesssim_p \lambda_{\min}(\Sigma_v) \lesssim_p \lambda_{\max}(\Sigma_v) \lesssim_p 1$. This assumption is rather weak as it only rules out factors whose risks are (asymptotically) negligible. We also maintain the assumption that $\|\Sigma_u\| \lesssim_p 1$, so that there exists no factor structure in the residuals u_t. This condition is needed for identification purpose, which ensures that all factors, regardless of their strength, must be distinguishable from the idiosyncratic errors.

In this setting, a factor's strength is entirely determined by test assets' exposures to it, since all factors have non-negligible 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 might diversify the exposure of the momentum factor.

The concern on factor strength has come to light since Kan and Zhang (1999), who discussed the failure of two-pass cross-sectional regressions in an extreme case where some factors are purely useless, to which test assets have zero-exposures. At the other extreme, the most prevalent assumption adopted by the literature on factor models, e.g., Bai and Ng (2002), is that all factors are strong or pervasive, that is, $\|\beta\| \asymp \sqrt{N}$, which dominates the strength of the idiosyncratic component, as measured by $\|\Sigma_u\|$. In contrast, our focus is on the regime of weak factors, in which the norm of columns of β is allowed to diverge at different and slower rates, which will be made more precisely later. The fact that weak factors are relevant in practice can be illustrated from the scree plot of eigenvalues of returns. The eigen gap between weak factors and the idiosyncratic components may not appear as clear-cut as the gap between the pervasive factors and idiosyncratic components.

We anchor our discussion on weak factors in two intriguing asset pricing exercises: the estimation of risk premia and the recovery of pricing kernel.

In this model, the stochastic discount factor (SDF) can be represented 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 an SDF 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. As will be shown later, these two forms of the SDF are asymptotically equivalent in the asymptotic

[1] 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).

scheme we consider, so that there is no ambiguity with respect to which estimand we consider.

We are also interested in risk premia of some observable factor proxies, summarized in a $d \times 1$ vector, g_t. Following Giglio and Xiu (2020), we do not assume 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. The risk premia of g_t is $\eta\gamma$, which is our parameter of interest. 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, these observable proxies in g_t may therefore be weak. That said, their risk exposures (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 the sample size.

2.3 Inconsistency of Existing Estimators

In what follows, we revisit a number of existing procedures for risk premia estimates, and then demonstrate their failure using a simple model with a single weak factor.

2.3.1 PCA

Giglio and Xiu (2020) 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, $\widehat{v}_t$;[2] 2) use Fama-MacBeth regression 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 $\widehat{\eta}$. A combination of these estimates 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 largest* p *left and right singular vectors as* ς *and xi and the corresponding singular values as* $\sqrt{T}\widehat{\lambda}^{1/2}$*. The estimated factors are given by* $\widehat{V} = \sqrt{T}\xi^{\intercal}$.

S2. Estimate the risk premia of $\widehat{V}$ *by* $\widehat{\gamma} = \widehat{\lambda}^{-1/2}\varsigma^{\intercal}\bar{r}$.

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

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

Giglio and Xiu (2020) establish the consistency of this estimator and 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). Unfortunately, this estimator 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

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

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 the beta as defined by $\widehat{\beta} = T^{-1}\bar{R}\widehat{V}^\intercal = \widehat{\varsigma}\widehat{\lambda}^{1/2}$ in the cross-sectional regression would yield a consistent estimator of risk premium of the estimated latent factor, which in turn leads to a consistent estimator of the risk premium of g. Otherwise, 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 $N/(\|\beta\|^2 T) \to 0$.

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.

2.3.2 PLS

Giglio and Xiu (2020) 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 is amount 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, they might be misled by large idiosyncratic errors in r_t when the signal is not sufficiently strong. In contrast with PCA, the 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 variations of returns that are relevant to the target factor of interest. The key difference is that PCA seeks for linear combinations of $\bar{R}$ that maximizes variation, ignoring information from the target $\bar{G}$, whereas PLS seeks for linear combinations that have the largest covariance with $\bar{G}$. In the case that $\bar{G}$ is a $1 \times T$ vector, the weight vector of the first PLS component is proportional to $\bar{R}\bar{G}^\intercal$, so the first factor is spanned by $\bar{G}\bar{R}^\intercal\bar{R}$. We can normalized this factor to have a unit norm, and then continue the procedure with the residuals from a projection of $\bar{R}$ onto the first factor. We formulate the PLS-based algorithm for a general $d \times T$ matrix of $\bar{G}$ 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* $\bar{G}$.

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 $\bar{R}_{(k)}\bar{G}^\intercal$.*
   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 $\widehat{\eta} = T^{-1} \bar{G} \widehat{V}^\intercal$ by $\widehat{V} = (\widehat{V}_{(1)}^\intercal, \cdots, \widehat{V}_{(p)}^\intercal)^\intercal$ and the estimated risk premium is $\widehat{\gamma} = (\widehat{\gamma}_{(1)}, \cdots, \widehat{\gamma}_{(p)})^\intercal$.

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.

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:

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

where $\mu > 0$ is some tuning parameter. In the case of pervasive factors, Giglio and Xiu (2020) 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 impact 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 $RR^\intercal + \mu\bar{r}\bar{r}^\intercal$*, where* μ *is a tuning parameter, and write the largest* 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* $\widehat{\eta} = T^{-1}\bar{G}\widehat{V}^\intercal(\widehat{V}\widehat{V}^\intercal)^{-1}$.

Outputs: $\widehat{\gamma}_g^{rpPCA} = \widehat{\eta}\widehat{\gamma}$.

The standard PCA is applied to the covariance matrix of returns, that is $RR^\intercal - \bar{r}\bar{r}^\intercal$. 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. They derive asymptotic properties of this estimator in a setting where all factors are weak and N and T increase to infinity at the same rate. In their setting, the strength of weak factors remains indistinguishable from that of idiosyncratic errors as N and T increase, which precludes the consistency of any estimators. We discuss here an induced risk premium estimator based on their SDF estimator, but in a setting where a single factor can be weak yet its strength increases asymptotically. This setting is more informative for comparing different approaches, under which a consistent estimation procedure also exists.

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 also inconsistent, with a more involved bias term compared to the above estimators.

2.4 Our Solution: Test Asset Selection

Results in the previous section shed light on the limitation of dimension reduction or shrinkage estimators, when factors are not pervasive. In practice, test assets are constructed by characteristics-sorted portfolios, e.g., by size or value. If the majority of the test assets have zero or little exposure to some of the factors, say, momentum, in the data generating process of returns, the weak factor problem arises.

One potential solution is to screen test assets and only keep those that have nontrivial exposure to the factor of interest. This factor is likely strong within this smaller set of test assets, so 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 practice in the empirical asset pricing literature. Very often, test assets are formulated using the exact characteristics-sorted portfolios that the factor of interest is born from. For instance, Fama and French (1993) use size and value double-sorted portfolios as test assets when estimating a factor model that include size and value as factors. This choice of test assets ensures considerable exposure to these factors, therefore the model with such factors is useful in explaining the cross-section of these test assets. While this strategy might appear ad hoc, we formalize this intuition and make this procedure rigorous.

We start with a simple one factor setting as discussed in the previous propositions, which helps illustrate the intuition behind our proposal and facilitates 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 empirical researchers who decide which test assets to employ in their study. Assuming that a strong factor structure exists at least within a subset of test assets seems 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.*[3]

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

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

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

[3] We discuss the case of a multivariate $\bar{G}$ in Section 2.6.

other. The loading matrix β *satisfies* $\|\beta\|_{\text{MAX}} \lesssim 1$ *and there exists a subset* $I_0 \subset [N]$ *such that* $\left\|\beta_{[I_0]}\right\| \asymp \sqrt{N_0}$ *where* $N_0 = |I_0| \to \infty$*. In addition, suppose that* $\beta_{\{qN\}}$ *and* $|\beta|_{\{qN+1\}}$ *are distinct in the sense that* $|\beta|_{\{qN+1\}} \leq (1+\delta)^{-1} |\beta|_{\{qN\}}$ *for some* $\delta > 0$*, where* $|\beta|_{\{k\}}$ *denotes the* k*th largest value in* $\left\{\left|\beta_{[i]}\right|\right\}_{i\in[N]}$*. Then, for any choice of* q *in Algorithm 4 such that* $qN/N_0 \to 0$ *and* $qN \to \infty$*, we have* $\widehat{\gamma}_g^{SPCA} \xrightarrow{p} \eta\gamma$*.*

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

$$N/(\lambda_{\min}(\beta^\intercal\beta)T) \to 0. \tag{5}$$

Intuitively, this condition requires that the weakest one among all p factors in (1) is sufficiently strong so that it can be recovered by PCA. Once again, we consider below more challenging regimes in which the condition (5) 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:

$$\beta = \left[\begin{array}{c|c} \beta_{11} & \beta_{12} \\ \hline \beta_{21} & \beta_{22} \end{array}\right]. \tag{6}$$

where β_{11} and β_{12} are $N_0 \times 1$ vectors and β_{21} and β_{22} are $(N - N_0) \times 1$ vectors. 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, $N/(\lambda_{\min}(\beta^\intercal\beta)T) \gtrsim N/(N_0T)$, which does not necessarily converge to 0 if N_0 and T are small, so that the condition (5) could fail. In this example, while both factors are strong, since they have highly correlated exposures, the same "weak factor" issue arises.

Also, applying the screening approach alone would not work in a general multi-factor model. Take (6) 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 would occur in this example. Needless to say, it is the second factor that is weak since most of test assets' exposure to it is zero. Now suppose that $\eta = (1, 1)$, then it implies that the observed factor g is correlated with both factors and hence with all test assets, so that screening would not eliminate any of them, and yet PCA with all test assets will not recover the weak factor, if $N/(N_0T)$ does not vanish. This example demonstrates that even though pre-screening assets ensures that the first PC after screening is strong, there is no guarantee that this procedure can solve the weak factor issue in one step.

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 as illustrated in Giglio and Xiu (2020), both the risk premia and the SDF estimands remain the same.

The above examples illustrate that the screening step may not eliminate any test assets to the extent that the weak factor problem remains. We provide another example that shows screening can eliminate too many assets so that a strong factor model becomes a weak or even rank-deficient one. For example, suppose β has the following form:

$$\beta = \left[\begin{array}{c|c} \beta_{11} & \beta_{11} \\ \hline 0 & \beta_{22} \end{array}\right], \tag{7}$$

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 that 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 after screening with g_t. These remaining test assets, however, have perfectly correlated exposures to both factors, so that only one factor, $v_{1t} + v_{2t}$, is left. This example shows the supervised procedure (screening plus PCA) proposed by Bair et al. (2006), may be counterproductive in a multi-factor setting.

To resolve the issue of weak factors and avoid the excessive screening trap, 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 following-up screening step. 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 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 that of PCA. It is also 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 $k = 1, 2, \ldots$ *repeat 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* $\left(\bar{R}_{(k)}\right)_{[\widehat{I}_k]}$ *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* $\bar{G}_{(k+1)} = \bar{G}_{(k)} - \widehat{\eta}_{(k)}\widehat{V}_{(k)}$.

S2. Stop at $k = \widehat{p}$, where $\widehat{p}$ is chosen based on some proper stopping rule.

S3. Estimate the risk premium by $\widehat{\gamma}_g^{SPCA} = \sum_{k=1}^{\widehat{p}} \widehat{\eta}_{(k)}\widehat{\gamma}_{(k)}$.

Output: $\widehat{\gamma}_g^{SPCA}$, $\widehat{\eta} = (\widehat{\eta}_{(1)}^\intercal, \cdots, \widehat{\eta}_{(\widehat{p})}^\intercal)^\intercal$, $\widehat{\gamma} = (\widehat{\gamma}_{(1)}, \cdots, \widehat{\gamma}_{(\widehat{p})})^\intercal$, $\widehat{V} = (\widehat{V}_{(1)}^\intercal, \cdots, \widehat{V}_{(\widehat{p})}^\intercal)^\intercal$ *and* $\widehat{\beta} = (\widehat{\beta}_{(1)}, \cdots, \widehat{\beta}_{(\widehat{p})})$.

In Algorithm 5, we recover one 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 $j \leq k$, so similar to PCA, the factors we obtain are orthogonal with each other.

Algorithm 5 yields a consistent estimator of γ_g as long as an appropriate choice of I_k and a stopping rule are adopted. One possible choice for I_k is:[4]

$$\widehat{I}_k = \left\{ i \middle| T^{-1} \left\| (\bar{R}_{(k)})_{[i]} \bar{G}_{(k)}^\intercal \right\|_{\text{MAX}} \geq c_q^{(k)} \right\}, \\ \text{where } c_q^{(k)} \text{ is the } (1-q)th\text{-quantile of } \left\{ T^{-1} \left\| (\bar{R}_{(k)})_{[i]} \bar{G}_{(k)}^\intercal \right\|_{\text{MAX}} \right\}_{i \in [N]}. \tag{8}$$

Correspondingly, we set the stopping criterion as:

$$c_q^{(k)} < c, \quad \text{for some threshold } c. \tag{9}$$

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 good tuning of q, the iteration stops as soon as most of the rows of the projected residuals of returns appears 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 within which this factor is strong. With this identification assumption, as well as 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 $\widehat{\gamma}_g^{SPCA} \xrightarrow{p} \eta\gamma$.

[4] 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 screening instead of covariance in $\widehat{I}_k$. Despite this does not affect the asymptotic analysis, we find correlation screening performs slightly better in finite samples.

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 $\widehat{\gamma}_g$ 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 test assets, their sample correlation can be as large as $T^{-1/2}\log T$. Therefore, the threshold needs no smaller than $T^{-1/2}$. However, if $\eta \asymp T^{-1/3}$, it suggests that g_t is not too much different from random noise, so its correlation with r_t will likely not lead to any discovery of strong factors. Our procedure will give a risk premium estimate of 0, which is certainly consistent, but the estimation error is of an order $T^{-1/3}$, so that the CLT fails. Generally speaking, this issue arises because of 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 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. 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. We provide the CLT result below and investigate its finite sample behavior 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 (8) and (9) satisfying*

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

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

$$\Phi = \gamma^\intercal\Sigma_v^{-1}\Pi_{11}\Sigma_v^{-1}\gamma + \gamma^\intercal\Sigma_v^{-1}\Pi_{12}\eta^\intercal + \eta\Pi_{12}^\intercal\Sigma_v^{-1}\gamma + \eta\Pi_{22}\eta^\intercal,$$

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 (2020), 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.

2.7 The Case of Observable Factors

The previous discussion does not assume the perfect knowledge of the factors v_t in (1). If these factors were known, 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.

If all factors in v_t are known and tradable, and that 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 them or not tradable, then a simple time series regression of g_t onto the factors v_t can recover its loading, η, which along with the risk premia estimates of v_t by their averages, give rise to the risk premium estimate of g. These scenarios are simple, which do not require cross-sectional regressions.

If some of the factors are not tradable, say, GDP growth is part of v_t, 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 procedure in Algorithm 5, to supervise the construction of mimicking portfolios for each of the observed non-tradable factors, while using residuals from the projection of test asset returns onto tradable factors as new test assets.

2.8 Recovery of the Stochastic Discount Factor

The previous sections focus on estimating the risk premia. We have shown that in order to construct valid asymptotic inference, we must recover all factors that drive the SDF. Once all these factors are recovered, we can also reconstruct the SDF. More specifically, from the outputs of Algorithm 5, we can estimate the SDF by:

$$\widehat{m}_t^{SPCA} = 1 - \widehat{\gamma}^\intercal \widehat{v}_t, \qquad \text{where } \widehat{v}_1, \cdots, \widehat{v}_T \text{ are the columns of } \widehat{V}. \tag{10}$$

Theorem 3. *Suppose the same assumptions as in Theorem 2 hold. In addition, we have Assumption A.10. Then the estimator (10) satisfies*

$$\frac{1}{T}\sum_{t=1}^{T} \left|\widehat{m}_t^{SPCA} - m_t\right|^2 \lesssim_p \frac{1}{T} + \frac{\log N_0}{N_0}. \tag{11}$$

There are a number of alternative approaches for SDF estimation proposed in the literature, e.g., the shrinkage approach by Kozak et al. (2020) and the risk premia PCA by Lettau and Pelger (2020). In what follows, we provide theoretical comparison of these estimators in the general weak factor framework.

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 estimator, we first establish the asymptotic equivalence of these two definitions:

Proposition 6. *Suppose that test asset returns r_t follows (1), and Assumption A.10 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 3, and that $\lambda_{\min}(\beta^\intercal\beta) \asymp N_0$, we can replace m_t in the left-hand side

of (11) by $\widetilde{m}_t$.

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

$$\widehat{b} = \arg\min_b \left\{ (\bar{r} - \widehat{\Sigma} b)^\intercal \widehat{\Sigma}^{-1} (\bar{r} - \widehat{\Sigma} b) + p_\mu(b) \right\}, \tag{12}$$

with which the estimated pricing kernel is given by

$$\widehat{m}_t = 1 - \widehat{b}^\intercal (r_t - \bar{r}). \tag{13}$$

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 (12) 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 (12):

$$\widehat{b} = \arg\min_b \left\{ b^\intercal \widehat{\Sigma} b - 2 b^\intercal \bar{r} + b^\intercal \widehat{\Sigma} b + p_\mu(b) \right\}, \tag{14}$$

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 4. *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.10-A.12 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 $p_\mu(b) = \mu \|b\|_1$, if Assumptions A.10, A.11 hold, we have

$$\frac{1}{T} \sum_{t=1}^{T} |\widehat{m}_t^{Lasso} - \widetilde{m}_t|^2 \lesssim_p \|b\|_1 \sqrt{\frac{\log N}{T}}. \tag{15}$$

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

$$\frac{1}{T} \sum_{t=1}^{T} |\widehat{m}_t^{Lasso} - \widetilde{m}_t|^2 \lesssim_p \|b\|_0 \frac{\log N}{T}. \tag{16}$$

Interestingly, both 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 (5) fails, then the SDF is not consistent. Furthermore, this estimator may not converge in the regime that $N/(\lambda_p T) \to \infty$, which is precisely the issue caused by weak factors which our SPCA estimator can tackle with.

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. In this case, the Lasso estimator remains consistent, but converge at a rather slow rate, $\|b\|_1 \sqrt{\log N/T}$ as shown in (15), so it is not as efficient as our SPCA estimator. Nonetheless, under a much stronger sparsity assumption that $\|b\|_0^2 \log N/T \to 0$, the Lasso estimator can achieve a comparable rate to that of the SPCA. This stronger sparsity assumption 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.

The SPCA estimator given by equation (10) can also be rewritten in the form of (13), 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 $\left(\bar{R}_{(k)}\right)_{[I_k]}$ in Step S1.b. 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 $m_t = 1 - \widehat{\gamma}^\intercal (\widehat{\Sigma}_v)^{-1} \widehat{v}_t = 1 - \widetilde{\gamma}^\intercal (\widetilde{\Sigma}_v)^{-1} \widetilde{v}_t = 1 - \widetilde{\gamma}^\intercal (\widetilde{\Sigma}_v)^{-1} B^\intercal (r_t - \bar{r})$, 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 optimal portfolios, which we investigate in simulations and empirical studies.

3 Simulations

In this section, we study the finite sample performance of our SPCA procedure using Monte Carlo simulations. We consider a 4-factor DGP as given by equation (1), where the first three factors are calibrated to match the de-noised three Fama-French factors (RmRf, SMB, HML) as in Giglio and Xiu (2020), and the last one is a potentially weak factor, denoted by V_1. The risk premium and variance of V_1 are selected such that this factor achieves a 0.35 annualized Sharpe ratio. The realizations of u_t and z_t are generated independently from a Gaussian distribution with mean 0 and standard deviation σ_u and σ_z, respectively. We set $\sigma_z = 0.5\bar{\sigma}_v$ and $\sigma_u = 2\bar{\sigma}_v$, where $\bar{\sigma}_v^2$ denotes the average volatility of these 4 factors.

The loadings of RmRf are generated independently from $\mathcal{N}(1,1)$ and the loadings of SMB and HML are generated independently from $\mathcal{N}(0,1)$. For the fourth factor, we simulate two cases of its loadings β_{V_1}, in which the weak factor problem arises.

Case 1: We generate β_{i,V_1} independently from a Gaussian mixture distribution, with probability a from

$\mathcal{N}(0,1)$ and $1-a$ from $\mathcal{N}(0,0.1^2)$. We use $a=0.1$ in the simulations so that for most test assets their factor loadings are very tiny.

Case 2: We simulate according to $\beta_{i,V_1} = \beta_{i,HML} + e_i$, where e_is are generated independently from the same mixture Gaussian distribution as above. In this case, the loading matrices of the factor V_1 and HML are very similar, which (almost) leads to a rank deficient factor loading matrix.

First of all, we study the behavior of the aforementioned risk premia estimators, including PCA, Ridge, PLS and our SPCA, for some observable factor proxy vector g_t that includes noisy versions of the four factors in the DGP plus a useless factor V_2 whose $\eta = 0$.

			SPCA		PCA		rpPCA		PLS		Lasso		Ridge	
T	Param	True	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
	Panel A: A Single Weak Factor													
	RmRf	53.7	-0.2	41.4	-0.1	41.2	0.3	43.4	0.0	41.4	-16.6	26.8	-17.0	26.9
	SMB	21.7	0.2	28.0	0.2	27.3	0.9	32.0	0.3	27.6	-7.6	16.4	-8.1	15.4
120	HML	25.4	-2.0	26.3	-2.8	25.9	-0.0	30.7	-2.3	26.2	-18.2	23.0	-17.4	22.4
	V_1	80.0	-18.5	26.4	-38.6	40.7	65.1	92.2	-29.9	33.5	-67.6	68.7	-72.7	73.3
	V_2	0.0	0.1	6.0	-0.0	4.7	-0.1	13.2	0.0	5.1	-0.1	2.8	-0.0	2.1
	RmRf	53.7	-0.3	29.3	-0.3	29.3	0.3	30.1	-0.2	29.3	-11.4	21.2	-11.5	20.8
	SMB	21.7	-1.2	19.1	-1.1	19.0	-0.8	20.3	-1.1	19.1	-5.7	13.7	-6.5	13.6
240	HML	25.4	-1.1	18.3	-1.4	18.1	0.3	19.5	-1.1	18.2	-13.3	18.9	-12.4	18.8
	V_1	80.0	-13.8	19.0	-26.5	28.7	36.5	49.6	-19.0	22.5	-58.2	60.1	-66.1	67.6
	V_2	0.0	0.1	4.0	0.1	3.5	0.2	6.5	0.1	3.8	0.0	2.4	0.0	2.0
	RmRf	53.7	0.1	23.4	0.1	23.4	0.6	23.7	0.1	23.4	-8.2	17.5	-8.6	17.1
	SMB	21.7	-0.9	15.7	-0.8	15.6	-0.6	16.4	-0.8	15.7	-4.0	11.7	-4.9	11.8
360	HML	25.4	-0.0	14.5	-0.2	14.4	1.2	15.3	-0.0	14.5	-10.8	16.0	-9.8	16.2
	V_1	80.0	-11.5	15.9	-20.0	22.3	24.4	33.9	-13.8	17.3	-50.9	53.5	-60.4	62.5
	V_2	0.0	0.1	3.3	0.0	3.0	0.1	4.7	0.1	3.2	0.1	2.4	0.0	1.8
	Panel B: Two Strong Factors with Highly Correlated Loadings													
	RmRf	53.7	0.1	42.1	0.2	42.0	1.3	44.1	0.3	42.0	-10.9	30.6	-10.0	31.6
	SMB	21.7	-0.3	26.6	-0.4	26.3	0.7	29.7	-0.4	26.6	-5.2	18.5	-5.4	17.7
120	HML	25.4	5.0	25.2	12.3	25.9	-14.8	60.2	9.8	25.5	5.6	19.0	8.3	18.9
	V_1	80.0	-8.7	22.6	-17.4	25.1	14.3	54.7	-13.6	23.4	-46.9	49.6	-46.0	47.8
	V_2	0.0	-0.3	6.4	-0.2	6.0	-0.5	13.2	-0.3	6.1	-0.2	4.3	-0.2	4.0
	RmRf	53.7	0.5	29.1	0.6	29.0	1.0	29.4	0.7	29.1	-6.1	24.1	-5.2	24.5
	SMB	21.7	0.7	19.0	0.6	18.8	1.2	19.5	0.6	19.0	-2.5	15.1	-2.9	14.3
240	HML	25.4	3.4	19.0	8.2	19.6	-5.9	33.2	5.8	19.2	9.5	17.8	13.4	19.3
	V_1	80.0	-5.5	15.3	-10.8	17.0	5.7	27.1	-7.8	15.7	-36.9	39.4	-37.9	39.3
	V_2	0.0	-0.2	4.3	-0.1	4.2	-0.0	5.6	-0.1	4.2	-0.1	3.5	-0.1	3.2
	RmRf	53.7	0.7	23.3	0.7	23.2	1.0	23.4	0.8	23.2	-4.1	20.1	-3.6	20.3
	SMB	21.7	0.2	15.7	0.1	15.6	0.4	16.0	0.2	15.7	-1.7	13.0	-2.2	12.5
360	HML	25.4	3.7	15.6	6.8	16.1	-1.9	20.4	4.9	15.8	10.2	16.6	14.8	18.8
	V_1	80.0	-4.5	12.6	-7.9	13.7	2.2	17.6	-5.6	12.8	-32.0	34.0	-33.6	34.8
	V_2	0.0	-0.1	3.6	-0.1	3.5	-0.1	4.0	-0.1	3.5	-0.0	3.0	-0.1	2.8

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, and Ridge approaches, respectively. The true data-generating process has four factors, driven by RmRf, SMB, HML, and V_1, whereas we estimate the risk premia for noisy versions of these four factors, as well as a useless noise factor V_2. Their true risk premia are provided in Column "True." We fix $N = 1,000$ while varying $T = 120, 240$, and 360 in this experiment. Panel A reports result for the case of a single weak factor V_1, and Panel B the case of two strong factors (HML and V_1) with highly correlated exposures.

We report in Table 1 the bias and the RMSE (root-mean-square error) of the estimates. To construct

the SPCA, we use all factors in g_t to supervise the procedure. The parameter c and q in SPCA are tuned by cross-validation using the time series R^2 of the mimicking portfolios for g_t as the criterion. This means that the tuning parameters are selected such that in the validation sample, the mimicking portfolios are the best for tradable proxies for g_t. Except for SPCA, all the remaining methods use optimal yet 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 five rows in each panel provide the results of risk premia estimation for the RmRf, SMB, HML, the weak factor V_1, and the useless factor V_2, respectively.

We find that our SPCA approach has smaller biases for the weak factors, whereas PCA, Ridge and PLS estimates have substantial biases, which agrees with our theoretical analysis. Nonetheless, if the true risk premium is small and N_0, T are not very large, PCA, Ridge and PLS can be better than SPCA in the sense of RMSE since the variance term is the main portion of RMSE in that case.

We next investigate the finite sample performance of the inference result developed in Theorem 2. Figure 1 plots histograms of the standardized risk premia estimators using the estimated asymptotic standard errors for SPCA and PCA, respectively, using the DGP in Case 2 as an example. The histograms of PCA deviates from the standard normal distribution for HML and the weak factor. In contrast, the histograms corresponding to the SPCA match the normal distribution well, which verifies our central limit results.

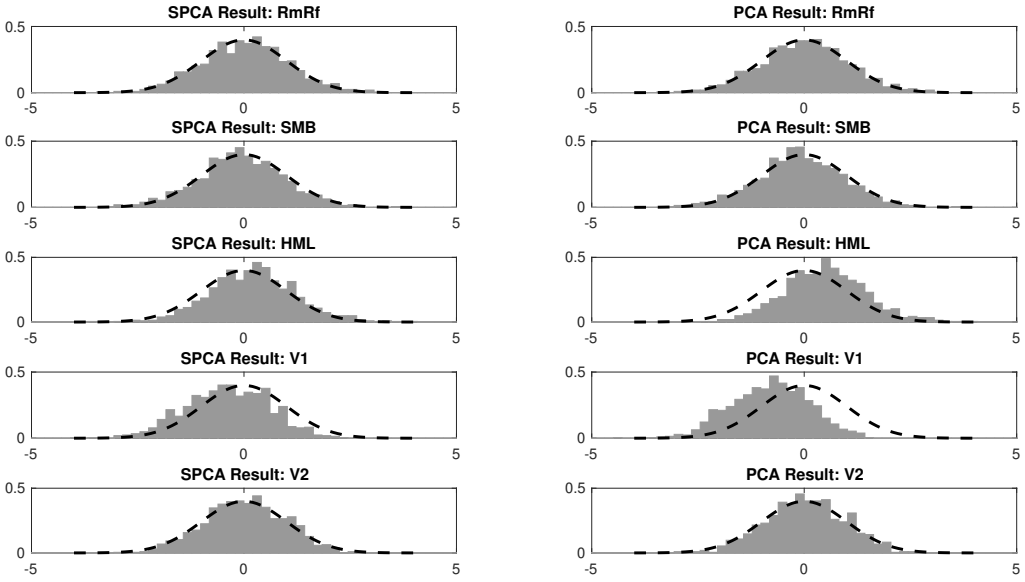

Figure 1: Histogram of the Standardized Estimates in Simulations

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 models with $N = 1,000$ and $T = 240$. The number of Monte Carlo repetitions is 1,000.

Finally, we study the finite sample behavior of the SDF estimators. We compare the performance of

	Panel A: A Single Weak Factor						
	SPCA		PCA	rpPCA	PLS	Lasso	Ridge
T	$\widehat{p}$	MSE	MSE	MSE	MSE	MSE	MSE
120	4.054	0.044	0.055	0.401	0.059	0.110	0.112
	(0.278)	(0.028)	(0.030)	(0.780)	(0.029)	(0.027)	(0.028)
240	4.008	0.022	0.026	0.105	0.029	0.090	0.096
	(0.089)	(0.014)	(0.015)	(0.118)	(0.014)	(0.022)	(0.028)
360	4.004	0.014	0.016	0.052	0.018	0.084	0.084
	(0.063)	(0.008)	(0.009)	(0.049)	(0.009)	(0.031)	(0.028)
	Panel B: Two Strong Factors with Highly Correlated Loadings						
120	4.038	0.039	0.039	0.531	0.040	0.078	0.055
	(0.242)	(0.027)	(0.027)	(1.460)	(0.027)	(0.025)	(0.026)
240	4.003	0.019	0.019	0.109	0.020	0.061	0.040
	(0.055)	(0.013)	(0.013)	(0.350)	(0.013)	(0.017)	(0.016)
360	4.012	0.013	0.013	0.042	0.014	0.056	0.034
	(0.109)	(0.009)	(0.009)	(0.136)	(0.009)	(0.030)	(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 has four factors, driven by RmRf, SMB, HML, and a weak factor V_1, whereas we estimate the SDF using a vector of factor proxies, g_t, that includes noisy versions of the four factors, as well as a useless pure noise factor V_2. We compare three scenarios with $T = 120$, 240, and 360, where $N = 1,000$ is fixed. In Case 1, there is a single weak factor, V_1, whereas in Case 2 HML and V_1 are highly correlated.

SPCA, PCA, rpPCA, Lasso and Ridge in the aforementioned two cases. 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 MSE, SPCA outperforms all other methods, which agrees with our theoretical prediction, and that rpPCA is on par with SPCA, both dominating PCA, Ridge, and Lasso when it comes to the out-of-sample Sharpe ratio. Last but not least, SPCA produces a good estimator of $\widehat{p}$ when T is large.

	Panel A: A Single Weak Factor						
T	SPCA	PCA	rpPCA	PLS	Lasso	Ridge	Theoretical Value
120	0.326	0.279	0.331	0.278	0.207	0.187	0.385
	(0.040)	(0.049)	(0.050)	(0.049)	(0.052)	(0.067)	
240	0.358	0.341	0.361	0.340	0.252	0.231	0.385
	(0.019)	(0.026)	(0.017)	(0.025)	(0.039)	(0.065)	
360	0.368	0.360	0.369	0.360	0.272	0.257	0.385
	(0.012)	(0.015)	(0.009)	(0.015)	(0.041)	(0.062)	
	Panel B: Two Strong Factors with Highly Correlated Loadings						
120	0.357	0.347	0.335	0.346	0.285	0.319	0.393
	(0.029)	(0.030)	(0.040)	(0.030)	(0.037)	(0.043)	
240	0.374	0.370	0.363	0.370	0.311	0.341	0.393
	(0.015)	(0.015)	(0.026)	(0.015)	(0.024)	(0.024)	
360	0.380	0.378	0.375	0.378	0.323	0.350	0.393
	(0.010)	(0.011)	(0.018)	(0.011)	(0.031)	(0.018)	

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 has four factors, driven by RmRf, SMB, HML, and a weak factor V_1, whereas we estimate the SDF using a vector of factor proxies, g_t, that includes noisy versions of the four factors, as well as a useless noise factor V_2. 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 $b^\intercal \mathrm{E}(r)/\sqrt{b'\Sigma^{-1}b}$ using true parameter values. We compare three scenarios with $T = 120$, 240, and 360, where $N = 1,000$ is fixed. In Case 1, there is a single weak factor, V_1, whereas in Case 2 HML and V_1 are highly correlated.

4 Empirical Analysis

In this section we apply our SPCA methodology to estimate the risk premium of several factors, both tradable and non-tradable, and we compare them with those obtained using alternative methodologies, including the PCA-based method of Giglio and Xiu (2020), PLS, and rpPCA.

4.1 Data

Our main dataset is the Chen and Zimmermann (2020) data, that is composed of a large number of equity portfolios sorted by characteristics. Specifically, we employ version 0.1.2 of the data. For each anomaly considered, Chen and Zimmermann (2020) construct a variable number of portfolios (as many as used in the original papers that introduced the anomaly in the literature: 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 2019m9, for which 772 test portfolios are available. All of our results are at the monthly frequency.

We study the risk premium of both tradable and nontradable factors. Among the hundreds of possible factors, we focus on a few representative ones to illustrate our methodology. 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), and VAR(1) innovations in the first three principal components of 279 macro-finance variables from Ludvigson and Ng (2010).

4.2 Latent Factors in the Returns Data

We start by examining the factor structure of the panel of returns. This is important because the methodologies studied in this paper (with the exception of Lasso and Ridge) require taking a stand on the number of factors. The plot of eigenvalues we present here will offer some guidance in choosing the baseline number of latent factors.

Figure 2 plots the log of the first 15 eigenvalues. There appear to be at least three strong factors. In addition, based on this figure, it appears that factors 4-10 might also be relevant, though weaker. In discussing the empirical results, we will present all of the results under different choices for the number of factors, p. Motivated by the scree plot, we will report results for p equal to 1, 3, 4, 5, 8, 11, and 13, therefore showing the robustness of our results to a wide range of model dimensions.

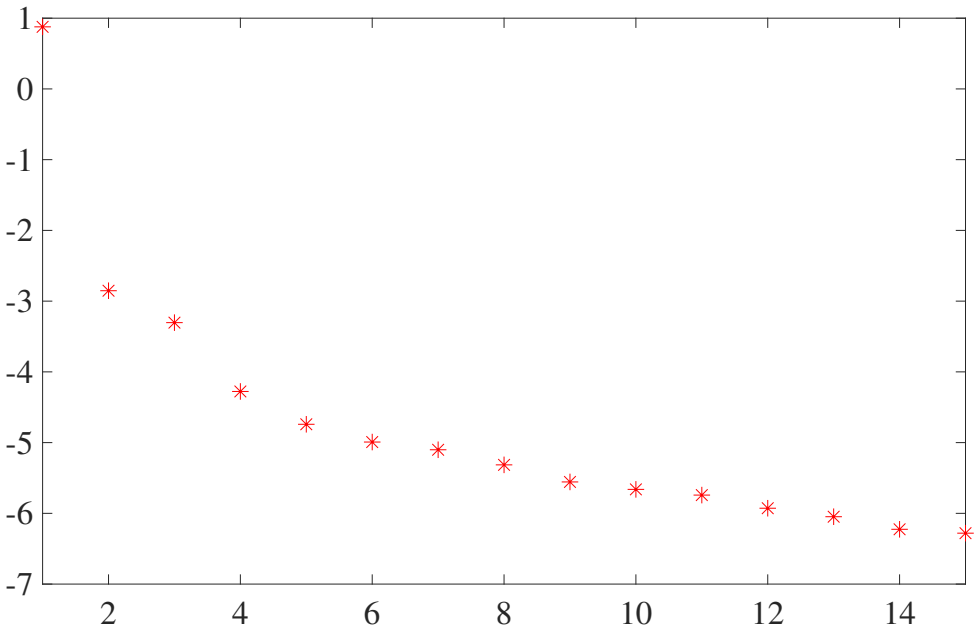

Figure 2: Logarithm of the First 15 Eigenvalues in the Chen-Zimmerman data

Note: The figure plots the log of the first 15 eigenvalues of the data, obtained from Chen and Zimmermann (2020), covering the period 1976-2019.

4.3 Estimation and Out-of-sample Evaluation

For our empirical analysis, we split the sample period into two equal-sized subsamples. The first half sample is used for the estimation of the model and the selection of the tuning parameter q (using cross-validation, as explained below). The latter half sample is used for evaluation of the results, to guarantee an entirely

out-of-sample evaluation. We slightly abuse the notation and call the first sample training sample (which is also the validation sample), and the second half testing sample.

To estimate the risk premia, we proceed as follows. We first choose the number of factors p in the model (from 1 to 13). In the training sample, we then run 3-fold cross-validation 100 times. In each cross-validation run, the tuning parameter is chosen to maximize the time-series R^2 of the mimicking portfolio implied by our procedure. This gives us 100 choices for the tuning parameter. We then select the median value across those 100 as our choice of q. The risk-premium estimate is the one that corresponds to that chosen q.

In addition to estimating the risk premium of the factor, our procedure also obtains (still in the training sample) the portfolio weights for the mimicking-portfolio of the factor, based on the estimated factors. The testing sample can then be used to evaluate how closely the mimicking portfolio is able to track the factor out of sample. In what follows, therefore, we will report these two key numbers: the estimated risk premium of the factor, and the out-of-sample R^2 of the factor using our model.

4.3.1 Out-of-sample R^2

Table 4 shows the out-of-sample R^2 obtained by SPCA. Each panel corresponds to a different factor; each row within a panel corresponds to a different choice for p. SPCA is the first column in each panel.

The table shows large heterogeneity in the OOS R^2 for different factors. For example, the market has an R^2 of 89% even when using just one latent factor, which rises to 97% using 13 factors. This should not be surprising: the market effectively corresponds to the first latent factor in the panel of returns; so even with only 1 factor, SPCA's first factor corresponds closely to it. Other tradable factors also display high OOS R^2, though not as high as the market. For example, the R^2s for SMB and Momentum reach above 80%, and that for HML close to 70%.

On the other hand, many of the factors (especially nontradable ones) have effectively zero R^2. This indicates that these factors are not spanned by the large panel of test assets; they appear spurious (in fact, out-of-sample R^2 are often negative, precisely because the factor that the procedure is aiming to fit is effectively noise.). This is the case of the three LN factors, IP growth, and liquidity.

There is however an intermediate case, which is the focus of this paper: there are factors, both tradables and nontradables, that are not clearly strong, in that their time-series R^2s are significantly below 1, but at the same time they are not spurious: the R^2 one can obtain is in the range .3-.6. These are likely weak factors, whose risk premium our procedure helps estimate. These weak factors include both tradables (like RMW, CMA and BAB) and nontradables (like the intermediary factor of He et al. (2017)).

There are two things worth reemphasizing here. First, strength of a factor is not really a discrete statement, but rather there is a continuum of strength from spurious to strong. Weak factors as the set of factors whose strength falls somewhere in between the two extremes; but an exact categorization of the factors in strong or weak is not important for practical purposes, since our methodology recovers the correct risk premium for all levels of strengths. Second, strength and weakness of factors are not inherent properties of the factors; they simply reflect how pervasive a factor is *in the chosen cross-section of test assets.*

Figure 3 reports the time-series R^2 information in graphical form, for all 14 factors (using 3 to 13 latent factors with SPCA). The distinction between strong factors (at the top), weak factors (in the middle) and spurious factors (at the bottom) appears quite clear from this figure.

	Market				SMB				HML			
# factors	SPCA	PCA	PLS	rpPCA	SPCA	PCA	PLS	rpPCA	SPCA	PCA	PLS	rpPCA
1	0.89	0.76	0.77	0.76	0.37	0.33	0.34	0.33	-0.01	-0.07	-0.07	-0.07
3	0.95	0.88	0.92	0.87	0.81	0.41	0.77	0.30	0.64	0.47	0.59	-0.02
4	0.96	0.90	0.97	0.87	0.75	0.50	0.78	0.43	0.73	0.50	0.68	0.58
5	0.95	0.89	0.97	0.89	0.82	0.51	0.84	0.48	0.70	0.53	0.67	0.59
8	0.96	0.95	0.98	0.94	0.86	0.63	0.85	0.64	0.64	0.64	0.67	0.61
11	0.97	0.96	0.98	0.96	0.86	0.76	0.85	0.79	0.67	0.66	0.59	0.64
13	0.97	0.96	0.98	0.96	0.85	0.80	0.84	0.82	0.66	0.65	0.59	0.67

	Momentum				CMA				RMW			
# factors	SPCA	PCA	PLS	rpPCA	SPCA	PCA	PLS	rpPCA	SPCA	PCA	PLS	rpPCA
1	-0.12	-0.09	-0.10	-0.09	0.12	0.05	0.06	0.05	0.14	0.07	0.10	0.07
3	0.75	-0.04	0.79	0.05	0.41	0.35	0.40	0.25	0.40	0.05	0.35	0.30
4	0.76	0.01	0.77	0.57	0.47	0.36	0.46	0.37	0.68	0.11	0.68	0.18
5	0.81	0.50	0.83	0.57	0.45	0.33	0.49	0.37	0.65	0.24	0.64	0.21
8	0.82	0.77	0.83	0.69	0.43	0.42	0.21	0.42	0.68	0.23	0.69	0.24
11	0.81	0.80	0.83	0.80	0.52	0.47	0.09	0.45	0.65	0.48	0.72	0.33
13	0.84	0.79	0.81	0.79	0.52	0.48	0.07	0.50	0.69	0.51	0.71	0.54

	BAB				QMJ				Liquidity			
# factors	SPCA	PCA	PLS	rpPCA	SPCA	PCA	PLS	rpPCA	SPCA	PCA	PLS	rpPCA
1	-0.34	-0.25	-0.26	-0.25	0.51	0.43	0.44	0.43	0.02	0.02	0.02	0.02
3	0.45	0.17	0.47	0.41	0.81	0.44	0.72	0.61	-0.06	0.02	-0.12	0.00
4	0.62	0.20	0.67	0.40	0.81	0.49	0.80	0.63	-0.05	0.02	-0.07	0.01
5	0.41	0.54	0.48	0.41	0.77	0.55	0.73	0.65	-0.03	0.01	-0.16	0.01
8	0.53	0.53	0.72	0.55	0.73	0.71	0.77	0.70	-0.05	-0.02	-0.39	-0.01
11	0.50	0.50	0.75	0.55	0.74	0.75	0.75	0.70	-0.05	-0.01	-0.51	0.00
13	0.54	0.36	0.76	0.33	0.73	0.71	0.71	0.72	-0.03	-0.01	-0.69	0.02

	LN1				LN2				LN3			
# factors	SPCA	PCA	PLS	rpPCA	SPCA	PCA	PLS	rpPCA	SPCA	PCA	PLS	rpPCA
1	0.00	0.00	-0.01	0.00	-0.04	-0.03	-0.03	-0.03	0.04	0.02	0.03	0.02
3	-0.43	-0.28	-0.41	-0.12	-0.01	-0.01	-0.16	0.00	0.05	0.05	0.06	0.05
4	-0.39	-0.25	-0.57	-0.29	-0.01	-0.01	-0.15	-0.01	0.06	0.06	0.03	0.05
5	-0.21	-0.19	-0.33	-0.27	-0.01	0.00	-0.28	-0.01	0.05	0.06	-0.04	0.06
8	-0.22	-0.18	-0.95	-0.17	-0.01	-0.01	-0.44	0.01	0.06	0.06	-0.19	0.06
11	-0.26	-0.22	-1.78	-0.22	-0.01	-0.01	-0.50	-0.01	0.06	0.06	-0.29	0.07
13	-0.11	-0.11	-1.84	-0.10	-0.01	-0.01	-0.56	-0.02	0.06	0.06	-0.36	0.07

	Intermediary				IP growth			
# factors	SPCA	PCA	PLS	rpPCA	SPCA	PCA	PLS	rpPCA
1	0.48	0.38	0.39	0.38	0.01	0.01	0.01	0.01
3	0.59	0.45	0.51	0.39	-0.03	-0.02	-0.16	-0.02
4	0.52	0.47	0.56	0.47	-0.03	-0.02	-0.24	-0.02
5	0.58	0.52	0.51	0.48	0.00	0.00	-0.31	-0.02
8	0.56	0.56	0.47	0.55	-0.01	-0.01	-0.89	0.00
11	0.53	0.55	0.36	0.55	-0.02	-0.02	-1.09	-0.02
13	0.56	0.56	0.37	0.55	-0.03	-0.02	-1.18	-0.01

Table 4: Out-of-sample R^2

Note: In this table, we report the out-of-sample time-series R^2 achieved with different methods (SPCA, PCA, PLS, rpPCA). Each panel corresponds to a different factor. Rows correspond to a different choice for the number of factors p. Sample is the Chen-Zimmerman data for the period 1976-2019.

Market

# factors	SPCA	PCA	PLS	rpPCA	SPCA (many)	Std. err.
1	79	66	66	69	75	26
3	87	81	96	117	84	26
4	85	79	88	105	64	26
5	93	85	90	111	88	26
8	80	108	68	127	94	26
11	71	97	66	69	83	26
13	68	99	63	66	70	26
Avg ret	72					

SMB

SPCA	PCA	PLS	rpPCA	SPCA (many)	Std. err.
28	29	29	30	25	9
12	10	-3	-32	-1	16
25	14	12	14	50	17
18	16	6	3	15	17
33	-21	28	-48	-1	18
48	-4	32	42	26	18
47	-5	34	35	47	17
19					

HML

SPCA	PCA	PLS	rpPCA	SPCA (many)	Std. err.
-20	-18	-18	-19	-21	9
43	33	23	154	54	16
13	33	-14	-37	28	16
16	29	4	-36	57	16
37	-10	39	-36	32	18
5	-12	49	-4	7	19
38	-9	62	20	14	20
40					

Momentum

# factors	SPCA	PCA	PLS	rpPCA	SPCA (many)	Std. err.
1	13	8	9	9	10	6
3	97	-10	94	108	-19	14
4	129	-8	140	374	78	15
5	125	72	130	371	100	20
8	110	153	85	298	114	21
11	95	146	85	89	139	23
13	113	144	87	74	115	23
Avg ret	86					

CMA

SPCA	PCA	PLS	rpPCA	SPCA (many)	Std. err.
-12	-11	-11	-12	-13	5
28	16	18	98	27	9
15	16	8	13	20	9
26	19	12	15	37	9
-10	13	32	13	32	11
10	0	38	12	22	13
18	-2	49	30	7	15
27					

RMW

SPCA	PCA	PLS	rpPCA	SPCA (many)	Std. err.
-5	-3	-4	-3	-1	2
10	-6	27	21	-11	6
34	-5	55	81	4	7
22	6	41	79	0	8
29	31	22	76	16	8
28	44	30	48	24	9
35	44	19	23	26	11
37					

BAB

# factors	SPCA	PCA	PLS	rpPCA	SPCA (many)	Std. err.
1	23	15	17	16	12	5
3	105	59	96	254	69	15
4	109	60	119	202	92	18
5	103	93	113	200	116	18
8	94	92	113	133	119	20
11	116	84	102	147	114	22
13	123	94	112	164	118	21
Avg ret	125					

QMJ

SPCA	PCA	PLS	rpPCA	SPCA (many)	Std. err.
-14	-15	-15	-16	-13	5
25	-17	28	22	-25	8
36	-16	50	90	-5	8
22	-10	38	87	-14	9
41	34	12	92	17	9
38	35	24	68	17	10
30	41	9	45	16	11
41					

Liquidity

SPCA	PCA	PLS	rpPCA	SPCA (many)	Std. err.
42	34	34	36	36	14
84	58	84	133	56	23
96	58	39	87	85	27
84	76	-32	85	82	31
75	77	-131	14	92	35
96	44	-123	8	85	40
43	52	-162	-3	78	32

LN1

# factors	SPCA	PCA	PLS	rpPCA	SPCA (many)	Std. err.
1	11	11	29	12	28	31
3	208	169	202	443	238	84
4	220	154	-63	-160	32	100
5	17	46	199	-119	159	102
8	182	110	716	196	213	143
11	59	49	842	-352	-34	171
13	50	53	937	115	-199	259

LN2

SPCA	PCA	PLS	rpPCA	SPCA (many)	Std. err.
45	27	31	29	41	28
102	106	200	292	96	59
108	109	80	75	118	70
99	84	248	66	93	84
65	-3	-93	88	19	97
112	114	-309	351	83	105
113	118	-378	293	253	115

LN3

SPCA	PCA	PLS	rpPCA	SPCA (many)	Std. err.
30	24	25	25	16	20
19	19	-183	-98	55	42
12	12	-289	-245	4	68
47	39	-62	-224	46	70
-215	-94	-40	-374	-40	90
-259	-226	141	13	-68	93
-230	-215	46	32	-75	138

Intermediary

# factors	SPCA	PCA	PLS	rpPCA	SPCA (many)	Std. err.
1	90	72	73	76	85	31
3	122	122	136	244	143	35
4	88	121	116	96	73	39
5	136	90	164	98	116	40
8	207	137	23	210	144	42
11	171	148	32	113	113	44
13	179	170	61	127	99	51

IP growth

SPCA	PCA	PLS	rpPCA	SPCA (many)	Std. err.
0	0	0	0	0	0
-1	-1	-3	-4	-2	1
-1	-1	-2	-1	-1	1
0	-1	-4	-1	-1	1
-2	-2	-10	-4	-3	2
-2	-2	-12	1	0	2
-2	-2	-15	-2	-1	3

Table 5: Risk Premium

Note: In this table, we report the risk premium in bp estimated using different methods (SPCA, PCA, PLS, rpPCA). The last two columns report the SPCA estimate using all 14 factors jointly as well as the standard error estimates. Each panel corresponds to a different factor. Rows correspond to a different choice for the number of factors p. Sample is the Chen-Zimmerman data for the period 1976-2019.

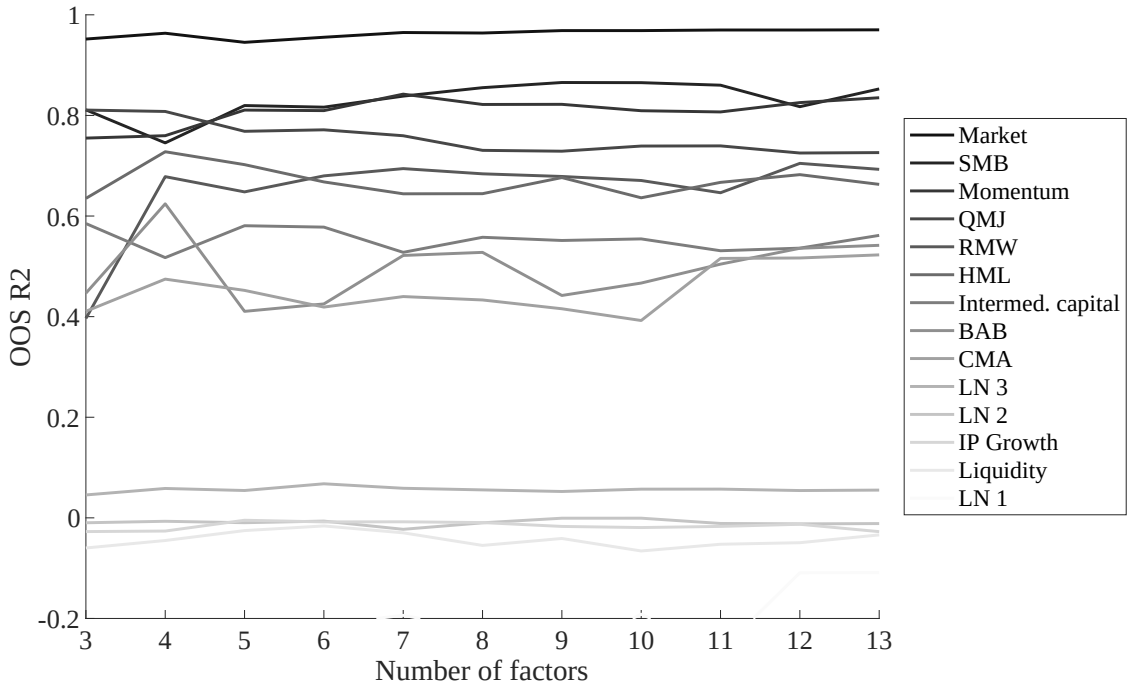

Figure 3: Out-of-sample R^2s of Different Factors

Note: The figure plots the time-series out of sample R^2 for 14 different factors, using SPCA. The number of latent factors extracted with SPCA is reported in the x axis. The data is the panel from Chen and Zimmermann (2020), covering the period 1976-2019.

4.3.2 Risk Premia Estimates and Asset Selection

Table 5 reports (in the first column of each panel) the estimated risk premium of the factors, for each choice of p. While naturally the estimates vary somewhat with the number of factors p, they are broadly stable when p is greater than 1. The table also reports for comparison the average return of the factor in the training and testing samples, for tradable factors.

Ideally, for tradable factors the point estimate using SPCA should be close to the average realized return of the factor; therefore, the comparison between the two reveals how well SPCA can estimate this risk premium.

Given that SPCA selects assets that have sufficiently high correlation with the factor, one may wonder whether we should expect SPCA to perfectly recover the risk premium of the tradable factor; just like a standard mimicking portfolio estimate should perfectly mimic the tradable factor if the tradable factor itself is among the test assets. The reason why this is not necessarily the case here, so that the risk premium may differ from the average realized return and the R^2 may be less than 100%, is that SPCA is *not* simply trying to find the best mimicking portfolio for the factor. Rather, it is imposing the assumed factor structure, and using it to recover the latent factors that drive the SDF (or at least the part that is relevant for the factor of interest), using them to estimate the factor risk premium. This has two advantages. For tradable factors, it allows to account for potential measurement error in the factor; for example, if the tradable factor is exposed to a true latent risk factor but also contains undiversified idiosyncratic risk, standard mimicking portfolio regressions would not yield efficient estimates of the risk premium (because they would "mimick" the entire portfolio, including the undiversified risk). SPCA instead aims to extract only the part of the observable factor that, according to the model, relates to the fundamental factors driving returns. For nontradable factors, a mimicking portfolio approach would be inefficient or even infeasible (as in our empirical application) as the number of assets is greater than the available time series.

Each of the estimates in column (1) of these panels reports the risk premia of each factor estimated individually. While this provides a consistent estimate of the risk premium (as discussed in Section 2.5), it does not allow us to do inference, because of the possibility of omitted weak factors. Under the assumption that all relevant weak latent factors can be captured when considering all 14 observable factors only (that is, that for each weak latent factor in the panel of returns, there is at least one of the 14 observable factors that loads on it), we can then also do inference. Columns (5) and (6) report the risk premia estimates and standard errors obtained using all factors simultaneously.

Table 5 shows that for tradable factors, the estimated risk premia are close to the realized average returns of the factors, both statistically and economically, and the results are robust to the choice of p.

To gain a better understanding of how SPCA works, it is useful to look into the specific assets that are selected to build the latent factors. We consider three factors for illustration purposes: one strong (momentum), one weak tradable factor (RMW), and the intermediary capital factor, a weak nontradable factor.

To estimate the risk premium for momentum, SPCA chooses $qN = 100$ assets to build each latent factor. For the first latent factor, the assets selected have correlation with momentum as high as .37, and a large fraction of them relate to different versions of momentum (specifically, the long portfolios that enter the

momentum factor).[5] The second latent factor is built by considering the correlation between the part of momentum not explained by the first factor and the residuals of the returns. Interestingly, these correlations are even higher, and are as high as .75 in absolute value. To build the second factor, SPCA picks assets on the short side of the momentum strategies.[6] This simple analysis shows how SPCA is able to zoom immediately into portfolios that are relevant to explain the time variation in the Momentum factor. The out-of-sample time-series R^2 of the factor using SPCA is 75% even when using only 3 factors.

RMW is a profitability portfolio. It should therefore not be a surprise that the first factor selected by SPCA combines returns of portfolios sorted by various accounting variables, like ROA, EPS, and book equity (though also other variables are selected, like volatility); further factors add more accounting information.

For the case of intermediary capital, the assets selected by SPCA are portfolios with high idiosyncratic volatility and low size for the first two factors (with correlations with the factor as high as .8), whereas the further factors include sorts by leverage and profitability.

From all these results, it is clear that SPCA does *not* simply choose to build the first factor as the market; rather it selects assets that capture relevant information for estimating the risk premium of the factor of interest.

4.3.3 Comparison with Other Estimators

We now compare the results obtained using SPCA with those obtained using alternative estimators, namely: the PCA-based estimator as in Giglio and Xiu (2020), a PLS version of this estimator, and rpPCA motivated from the SDF estimator by Lettau and Pelger (2020). In both Table 4 and Table 5, the risk premia and out-of-sample R^2 are reported in each panel in columns (2) to (4).

Starting from the OOS R^2 (Table 4), several patterns emerge. First, for strong factors (e.g., the market), all estimators perform very similarly, irrespective of p. For completely spurious factors (e.g., IP growth), all methods work as expected (with zero or negative R^2). Negative out-of-sample R^2s for what are clearly spurious factors arise from overfitting, something that all methodologies appear equally subject to.

More interesting are the cases with the intermediate factors. There are two main patterns to note. First, SPCA often obtains a high R^2 with significantly fewer factors than the alternative methods. For example, in the case of momentum, SPCA achieves a 75% R^2 with 3 factors only, whereas standard PCA approach requires 8 factors to get to the same R^2, and rpPCA 11 factors (PLS behaves similarly to SPCA in this case, but as we will see later, it is less robust in general).

The second pattern is that for all other factors, where R^2 of the order 20%-60% are achieved, SPCA does as well or better than all other methodologies very consistently, and in a way that is most robust to the choice of p. Specifically, PLS has the highest variability, obtaining in a few cases very good out-of-sample performance (e.g., for BAB), but in many cases performing disastrously (e.g., dramatically overfitting spurious factors). PCA does almost as well as SPCA in most cases, but its performance depends more strongly on the choice of p (and for given p, its performance is still typically below that of SPCA). rpPCA also

[5] Specifically, the list of the top 10 assets by correlation is: Mom12m05, FirmAgeMom0, IntMom05, retMomVol05, MomVol04, FirmAgeMom04, ResidualMomentum11m05, Mom12m04, High5205, RIO_BM01, where the name of the sorting characteristic follows Chen and Zimmermann (2020), and the last two digit indicate the number of the portfolio.

[6] Specifically, the list of the top 10 assets by correlation is: ResidualMomentum11m01, Mom12m02, MomVol01, DownForecast01, UpForecast01, Mom6m02, ResidualMomentum6m01, IntMom02, sfe01, MomSeasAlt1n02.

performs almost as well as SPCA, but underperforms for some of the weak factors, sometimes significantly (e.g. for RMW).

These similarities and differences across estimators are visible also when comparing risk premia (Table 5). In those cases where the estimators achieve similar out-of-sample R^2, the risk premia estimates are also close; but in several interesting cases, the risk-premia estimates differ significantly. For example, for the intermediary capital factor, SPCA tends to show a significantly higher risk premium than the other methodologies.

Overall, the empirical results show that all these methodologies perform relatively similarly in capturing the time variation and risk premium of both tradable and nontradable factors: especially so for strong factors, much less so for weak factors. SPCA shows two advantages: first, it consistently performs as well or better than all other estimators, across all factor strengths. Second, it does so in a way that appears very robust to the choice of factors p, much more so than the other methods. Its flexibility allows it to handle well all ranges of factor strengths.

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 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.

In the paper, we propose some empirical applications of SPCA, and compare its performance to alternative methodologies to estimate the SDF and risk premia. While the road to a full understanding of risk and risk premia in financial markets is still long, we believe that properly selecting the cross-section of test portfolios and addressing weak and strong omitted factors are important steps in this direction.

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 (2018). Factor models with many assets: strong factors, weak factors, and the two-pass procedure. *arXiv preprint arXiv:1807.04094*.

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

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

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

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

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

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

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

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*.

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

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.

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

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. 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. W. and D. Xiu (2020). Asset pricing with omitted factors. *Journal of Political Economy, forthcoming*.

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.

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.

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.

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.

Appendix

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:*

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

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

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

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

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

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

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*

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

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:*

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

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:*

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

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* $I \subset [N]$*:*

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

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

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

Assumptions A.5 - A.7 resemble Assumptions A.7, A.9, and A.10 of Giglio and Xiu (2020), 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 $a_i^{(1)} := \left\|\beta_{[i]}\eta^\intercal\right\|_{\text{MAX}}$ and define $I_1 := \{a_i^{(1)} \geq c_{qN}^{(1)}\}$, where $c_{qN}^{(1)}$ is the (qN)th largest value in $\left\{a_i^{(1)}\right\}_{i=1,\ldots,N}$. Then, we denote the largest right singular vector of $\beta_{(1)} := \beta_{[I_1]}$ by b_1. For $k > 1$, we obtain $a_i^{(k)} := \left\|\beta_{[i]} \prod_{j<k} \mathbb{M}_{b_j} \eta^\intercal\right\|_{\text{MAX}}$, $I_k := \{a_i^{(k)} \geq c_{qN}^{(k)}\}$ and b_k is the largest right singular vector of $\beta_{(k)} := \beta_{[I_k]} \prod_{j<k} \mathbb{M}_{b_j}$. This procedure is stopped at step $\tilde{p}$ (for some $\tilde{p}$ 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:*

$$T^{1/2}\begin{pmatrix} T^{-1}\mathrm{vec}(VZ^\intercal) \\ \bar{v} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{12}^\intercal & \Pi_{22} \end{pmatrix}\right),$$

where Π_{11}*,* Π_{12}*,* Π_{22} *are* $dp \times dp$*,* $dp \times p$*, and* $p \times p$ *matrices, respectively, defined as:*

$$\begin{aligned}
\Pi_{11} &= \lim_{T\to\infty} \frac{1}{T}\mathrm{E}\left(VZ^\intercal ZV^\intercal\right), \\
\Pi_{12} &= \lim_{T\to\infty} \frac{1}{T}\mathrm{E}\left(VZ^\intercal \iota_T^\intercal V^\intercal\right), \\
\Pi_{22} &= \lim_{T\to\infty} \frac{1}{T}\mathrm{E}\left(V\iota_T\iota_T^\intercal V^\intercal\right).
\end{aligned}$$

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.

Blow we introduce assumptions needed for the SDF estimation. Assumption A.10 ensures that the SDF concept is well defined. Assumption A.11 again can be shown by some large deviation result and certain central limit theorem.

Assumption A.10. *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.11. *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.12. *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

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

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

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

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

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

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:

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

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 $\left|T^{-1}\bar{V}\bar{V}^\intercal - 1\right| \lesssim T^{-1/2}$, we have

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

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

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

which leads to $\left|\widehat{\lambda}^{-1/2}\varsigma^\intercal\bar{u}\right| \lesssim_p \lambda^{-1/2}$. 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 $\left\|\bar{G}\bar{R}^\intercal\bar{R}\right\|$, $\bar{G}\bar{R}^\intercal\bar{R}\bar{G}^\intercal$ and $\bar{G}\bar{R}^\intercal\bar{r}$ separately. Recall that from (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 $\left\|\bar{G}\right\| \lesssim_p \sqrt{T}$, 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 $\left|T^{-1}\bar{V}\bar{V}^\intercal - 1\right| \lesssim_p T^{-1/2}$ and $\left\|\bar{V}\right\| - \sqrt{T} \lesssim_p 1$ 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 $\left\|\bar{U}_1\right\| \lesssim_p \sqrt{T}$ 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 $\left\|N^{-1}U^\intercal U - \mathbb{I}_T\right\| \lesssim_p \sqrt{T/N}$ 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 $\text{Rank}(\bar{R}) \leq \min\{N, T-1\}$, and the assumptions of the proposition imply that $N/T \to \infty$, we thereby have a condensed SVD of $\bar{R}$ as

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

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

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

$$= \frac{\widehat{\lambda}}{\widehat{\lambda} + T^{-1}\mu} \widehat{\gamma}_g^{PCA} + \frac{\eta \bar{V} \xi_*}{\sqrt{T}} \widehat{\Lambda}_*^{1/2} \left(\widehat{\Lambda}_* + T^{-1}\mu \right)^{-1} \varsigma_*^{\intercal} \bar{r}. \tag{B.22}$$

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

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

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

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

where $\tilde{\mu} = \sqrt{\mu+1} - 1$. Hence, the eigenvectors of $RR^\intercal + T^{-1}\mu\bar{r}\bar{r}^\intercal$ 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 $\left|T^{-1}\bar{V}\bar{V}^\intercal - 1\right| \lesssim_p T^{-1/2}$, $\left\|\bar{U}_1\right\| \lesssim_p T^{1/2}$ and $\left\|N^{-1}\bar{U}^\intercal\bar{U} - (\mathbb{I}_T - T^{-1}\iota_T\iota_T^\intercal)\right\| \lesssim_p \sqrt{T/N}$.

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

$$\left\|T^{-1}\lambda^{-1}(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^\intercal)R^\intercal R(\mathbb{I}_T + T^{-1}\tilde{\mu}\iota_T\iota_T^\intercal) - M^*\right\| = 0_p(1) \tag{B.28}$$

where the matrix M^* here is defined by

$$M^* := B\mathbb{I}_T + T^{-1}\bar{V}^\intercal\bar{V} + T^{-1}(1+\tilde{\mu})\gamma(\iota_T\bar{V} + \bar{V}^\intercal\iota_T^\intercal) + T^{-1}\Big((1+\tilde{\mu})^2\gamma^2 + \tilde{\mu}^2B + 2\tilde{\mu}B\Big)\iota_T\iota_T^\intercal.$$

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 $a_3 = (1+\tilde{\mu})^2\gamma^2 + \tilde{\mu}^2B + 2\tilde{\mu}B$. 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 $\left|\tilde{\xi}^\intercal \xi\right| \xrightarrow{p} 1$ 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 $\{|\beta_{[i]}|\}_{i\in[N]}$. 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 $\left\|\beta_{[I_0]}\right\|^2 \lesssim qN + (N_0 - qN)|\beta|_{\{qN\}}^2$. Together with the assumption that $qN/N_0 \to 0$ and $\left\|\beta_{[I_0]}\right\| \asymp \sqrt{N_0}$, it leads to $|\beta|_{\{qN\}}^2 \gtrsim \left\|\beta_{I_0}\right\|^2/N_0 \asymp 1$. 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,

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

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 1-3, 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 $\widehat{\gamma}_{(k)} = \widehat{\lambda}_{(k)}^{-1/2}\varsigma_{(k)}^{\intercal} r_{(k)}$, 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 4, we can use $\bar{G}$ instead of $\bar{G}_{(k)}$ in Algorithm 5 and throughout the proof. We denote $\widehat{\eta} = (\widehat{\eta}_{(1)}, \ldots, \widehat{\eta}_{(\tilde{p})})$ and $\widehat{\gamma} = (\widehat{\gamma}_{(1)}, \ldots, \widehat{\gamma}_{(\tilde{p})})^{\intercal}$, so the risk premium estimator is $\widehat{\gamma}_g^{SPCA} = \widehat{\eta}\widehat{\gamma}$.

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

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

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

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

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

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

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

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

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

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

Since Assumption A.3 implies that $\sigma_p(\beta_{[I_0]})\asymp N_0^{1/2}$, we have $\sigma_p(\bar{V}\bar{V}^\intercal\beta_{[I_0]}+\bar{V}\bar{U}_{[I_0]}^\intercal)\asymp N_0^{1/2}T$. 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

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

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

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

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

□

B.8 Proof of Theorem 3

Proof. As 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

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

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,

$$\sqrt{T}\xi H_2^\intercal=\bar{V}^\intercal H_2H_2^\intercal+O_p\left(q^{-1/2}N^{-1/2}T^{1/2}+T^{-1/2}\right). \tag{B.67}$$

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.9 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, $\Lambda^{1/2} = \text{diag}(\widetilde{\lambda}_1^{1/2}, \cdots, \widetilde{\lambda}_p^{1/2})$ 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 $\text{Cov}(B^\intercal\Sigma_u^{-1/2}u_t) = \mathbb{I}_p$, we have $\text{E}\left(\left\|B^\intercal\Sigma_u^{-1/2}U\right\|_{\text{F}}^2\right) = pT$, which leads to

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

For the same reason, we have $\left\|\Sigma_v^{-1/2}V\right\| \lesssim_p T^{1/2}$. Then, with Assumption A.10, (B.69), (B.70), (B.71), and (B.72), we have

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

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

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

which concludes the proof. □

B.10 Proof of Theorem 4(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 $\widehat{\Lambda}=\mathrm{diag}\{\widehat{\lambda}_1,\ldots,\widehat{\lambda}_p\}$, 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 $\tilde{H}_1 = (\tilde{h}_{11}, \ldots, \tilde{h}_{p1})$, $\tilde{H}_2 = (h_{12}, \ldots, \tilde{h}_{p2})$. 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 $\|\bar{V}\| \lesssim_p T^{1/2}$ 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

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

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

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

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

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

for $j \leq \min\{N, T\}$. Since $\text{Rank}(T^{-1/2}\beta\bar{V}) \leq p$, we have $\sigma_j(T^{-1/2}\beta\bar{V}) = 0$ for $j > p$ and thus

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

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

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

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

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

and consequently,

$$\|\varsigma_*^\intercal\bar{r}\| \leq \|\varsigma_*^\intercal\beta\| \|\gamma + \bar{v}\| + \|\varsigma_*^\intercal\bar{u}\| \lesssim_p \sqrt{\frac{N}{T}} + 1. \tag{B.85}$$

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

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

and

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

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

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

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.11 Proof of Theorem 4(b)

Proof. i. (Slow rate) Note that (12) 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

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

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

$$\begin{aligned}\left\|\widehat{\Sigma}^{1/2}(\widehat{b}-b)\right\|^2 \leq&\frac{1}{2}\mu\left\|\widehat{b}_J-b_J\right\|_1+\frac{1}{2}\mu\left\|\widehat{b}_{J^c}\right\|_1+\mu\left\|\widehat{b}_J-b_J\right\|_1-\mu\left\|\widehat{b}_{J^c}\right\|_1\\ =&\frac{3}{2}\mu\left\|\widehat{b}_J-b_J\right\|_1-\frac{1}{2}\mu\left\|\widehat{b}_{J^c}\right\|_1.\end{aligned} \tag{B.94}$$

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 $|J|\sqrt{\log N/T}\to 0$ and $\lambda_{\min}(\Sigma)\gtrsim 1$, we have

$$\frac{b^{*\intercal}\widehat{\Sigma}b^*}{\|b^*\|^2}=\frac{b^{*\intercal}\Sigma b^*}{\|b^*\|^2}+\frac{b^{*\intercal}(\Sigma-\widehat{\Sigma})b^*}{\|b^*\|^2}\gtrsim_p 1.$$

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

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

□

B.12 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) $\|\bar{U}\|_{\mathrm{MAX}}\lesssim_p(\log NT)^{1/2}$, $\|\bar{U}\bar{V}^\intercal\|_{\mathrm{MAX}}\lesssim_p(\log N)^{1/2}T^{1/2}$, $\|\bar{U}\bar{Z}^\intercal\|_{\mathrm{MAX}}\lesssim_p(\log N)^{1/2}T^{1/2}$.

(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

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

(ii) Using Assumption A.5, we have

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

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

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

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

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

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

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

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

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

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

(v) Using Assumption A.4 , we have

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

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

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

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

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

Using Assumption A.2, we have

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

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

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

and

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

□

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

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

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

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

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

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

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

and

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

we thereby have

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

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, $\|\beta\|_{\mathrm{MAX}}\lesssim 1$ 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 $\left\|\beta_{[i]}\prod_{j\leq\tilde{p}}\mathbb{M}_{b_j}\eta^\intercal\right\|_{\mathrm{MAX}}\geq c$, 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

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

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

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

which implies that

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

Since

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

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

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

(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

$$\left\|T^{-1}\bar{R}\bar{G}^\intercal-\beta\eta^\intercal\right\|_{\text{MAX}}\lesssim\|\beta\|_{\text{MAX}}\left\|T^{-1}\bar{V}\bar{V}^\intercal-\mathbb{I}_p\right\|\|\eta\|+T^{-1}\left\|\bar{U}\bar{V}^\intercal\right\|_{\text{MAX}}\|\eta\|$$

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

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

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

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

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

for some $i \in I_1$, which implies $A_1^c \subset A_3$. Similarly, we have $A_2^c \subset A_3$. Using $\{\widehat{I}_1 = I_1\} \supset A_1 \cap A_2$ and $A_1^c \cup A_2^c \subset A_3$, we have

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

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

(ii) Since $\widehat{I}_1 = I_1$ with high probability, we impose $\widehat{I}_1 = I_1$ below. Then, we have $\widetilde{R}_{(1)} = \bar{R}_{[I_1]}$ and Assumption A.12 gives $\left\|\widetilde{R}_{(1)} - \beta_{(1)}\bar{V}\right\| = \|\bar{U}_{[I_1]}\| \lesssim_p q^{1/2}N^{1/2} + T^{1/2}$.

(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

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

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

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

(iv) Let $\tilde{\xi}_{(1)} \in \mathbb{R}^{T\times 1}$ 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 $j \le k-1$, 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

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

Since (iv) holds for $j \le k-1$, 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

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

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

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

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

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

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

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

Hence, we have

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

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 $\widetilde{R}_{(k)} = \bar{R}_{[I_k]} \prod_{j=1}^{k-1} \mathbb{M}_{\widehat{V}_{(j)}^\intercal}$ and thus

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

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

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

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

(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

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

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

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

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

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

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*

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

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

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

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

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

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

which in turn gives

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

With Lemma 1(v), we have

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

(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* $j \le \tilde{p}$*, then for* $k \le \tilde{p}$*, we have*

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

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

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

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

which leads to

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

(ii) From Lemmas 1 and 5, we have

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

which leads to

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

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

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

□

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

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

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

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

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

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

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

Consequently, with Lemma 6(i) we have

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

If $\left\|T^{-1/2}\widehat{\lambda}_{(i)}^{-1/2}\varsigma_{(i)}^{\intercal}\widetilde{U}_{(i)}\right\| \lesssim_p q^{-1/2}N^{-1/2} + T^{-1}$ 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 $\left\|T^{-1/2}\widehat{\lambda}_{(k)}^{-1/2}\varsigma_{(k)}^{\intercal}\widetilde{U}_{(k)}\right\| \lesssim_p q^{-1/2}N^{-1/2} + T^{-1}$ 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 $\bar{V}^{\intercal}$ from the right side of (B.120), we have

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

Consequently, we have

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

When $k=1$, $\left\|T^{-1}\widehat{\lambda}_{(k)}^{-1/2}\varsigma_{(k)}^{\intercal}\widetilde{U}_{(k)}\bar{V}^{\intercal}\right\| \lesssim_p q^{-1}N^{-1} + T^{-1}$ 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 $\iota_T^\intercal$ in the above proof, and using Lemma 6(ii), we have the following results:

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

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

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

Using (B.56), we have

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

With Lemma 1 and (i), we have

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

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

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

(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

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

□

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

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

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

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

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

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

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

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

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 $j \le p$.

(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 $j \le p$, $\sigma_j(\beta\bar{V})^2 = \lambda_j(\beta\bar{V}\bar{V}^\intercal\beta^\intercal) = \lambda_j(\beta^\intercal\beta\bar{V}\bar{V}^\intercal)$ which implies

$$\lambda_j(\beta^\intercal\beta)\lambda_p(\bar{V}\bar{V}^\intercal) \le \sigma_j(\beta\bar{V})^2 \le \lambda_j(\beta^\intercal\beta)\lambda_1(\bar{V}\bar{V}^\intercal).$$

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 $\left\|\beta\beta^\intercal - T^{-1}\beta\bar{V}\bar{V}^\intercal\beta^\intercal\right\| \le \|\beta\|^2\left\|T^{-1}\bar{V}\bar{V}^\intercal - \mathbb{I}_p\right\| \lesssim_p \sigma_1(\beta)^2T^{-1/2}$ 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,

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

By Weyl's inequality, we have $T^{-1}\left\|\beta\bar{V}\right\|^2 = \lambda_1(T^{-1}\beta\bar{V}\bar{V}^\intercal\beta^\intercal) = \lambda_1(\beta\beta^\intercal) + O_p(\sigma_1(\beta)^2T^{-1/2}) = \sigma_1(\beta)^2 + O_p(\sigma_1(\beta)^2T^{-1/2})$. 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 $\widehat{\lambda}_j^{1/2}$, $j \leq p$ to denote them throughout Lemmas 11-13.

Lemma 11. *Under the assumptions of Theorem 4(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

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

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 $j \leq p$. 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

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

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

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

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 4(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* i*th 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

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

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

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

Define $L = \mathrm{diag}(l_1, \ldots, l_p)$, where l_i is equal to $\lambda_j/(\lambda_j - \lambda_i)$ if $i \neq j$ and 0 otherwise.

By left multiplying L to both sides of (B.138), we have

$$\begin{aligned}s_{j1} - \langle s_{j1}, e_{j1}\rangle e_{j1} =& \frac{1}{T\lambda_j}L\Lambda^{\frac{1}{2}}\bar{V}\frac{\widetilde{U}_2^\intercal\widetilde{U}_2}{\sqrt{T\widehat{\lambda}_j}}\xi_j + \frac{1}{T\lambda_j}L\widetilde{U}_1\frac{\widetilde{U}_2^\intercal\widetilde{U}_2}{\sqrt{T\widehat{\lambda}_j}}\xi_j + \frac{1}{\lambda_j}L\Lambda^{\frac{1}{2}}\left(\frac{\bar{V}\bar{V}^\intercal}{T} - \mathbb{I}_p\right)\Lambda^{\frac{1}{2}}s_{j1} \\ &+ \frac{L\Lambda^{\frac{1}{2}}\bar{V}\widetilde{U}_1^\intercal}{T\lambda_j}s_{j1} + L\frac{\widetilde{U}_1\bar{V}^\intercal\Lambda^{\frac{1}{2}}}{T\lambda_j}s_{j1} + L\frac{\widetilde{U}_1\widetilde{U}_1^\intercal}{T\lambda_j}s_{j1} - \left(\frac{\widehat{\lambda}_j}{\lambda_j} - 1\right)Ls_{j1}.\end{aligned} \tag{B.139}$$

Now left multiplying $\left(\frac{\Lambda}{\lambda_j}\right)^{\frac{1}{2}}$ again, we have

$$\begin{aligned}\left(\frac{\Lambda}{\lambda_j}\right)^{\frac{1}{2}}(s_{j1}-\langle s_{j1},e_{j1}\rangle e_{j1}) =&\frac{1}{T\lambda_j^{3/2}}\Lambda^{\frac{1}{2}}L\Lambda^{\frac{1}{2}}\bar{V}\frac{\widetilde{U}_2^\intercal\widetilde{U}_2}{\sqrt{T\widehat{\lambda}_j}}\xi_j+\frac{1}{T\lambda_j^{3/2}}\Lambda^{\frac{1}{2}}L\widetilde{U}_1\frac{\widetilde{U}_2^\intercal\widetilde{U}_2}{\sqrt{T\widehat{\lambda}_j}}\xi_j\\&+\frac{1}{\lambda_j}\Lambda^{\frac{1}{2}}L\Lambda^{\frac{1}{2}}\left(\frac{\bar{V}\bar{V}^\intercal}{T}-\mathbb{I}_p\right)\left(\frac{\Lambda}{\lambda_j}\right)^{\frac{1}{2}}s_{j1}+\Lambda^{\frac{1}{2}}L\Lambda^{\frac{1}{2}}\frac{\bar{V}\widetilde{U}_1^\intercal}{T\lambda_j^{3/2}}s_{j1}\\&+\Lambda^{\frac{1}{2}}L\frac{\widetilde{U}_1\bar{V}^\intercal}{T\lambda_j}\left(\frac{\Lambda}{\lambda_j}\right)^{\frac{1}{2}}s_{j1}+\Lambda^{\frac{1}{2}}L\frac{\widetilde{U}_1\widetilde{U}_1^\intercal}{T\lambda_j^{3/2}}s_{j1}-\left(\frac{\widehat{\lambda}_j}{\lambda_j}-1\right)\left(\frac{\Lambda}{\lambda_j}\right)^{\frac{1}{2}}Ls_{j1}\\=&K_1+K_2+K_3+K_4+K_5+K_6+K_7.\end{aligned}\tag{B.140}$$

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.12 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

$$\left\|\widetilde{U}_1\right\|=\left\|(\beta^\intercal\beta)^{-1/2}\beta^\intercal\bar{U}\right\|\lesssim_p T^{1/2},\quad \left\|\widetilde{U}_1\bar{V}^\intercal\right\|=\left\|(\beta^\intercal\beta)^{-1/2}\beta^\intercal\bar{U}\bar{V}^\intercal\right\|\lesssim_p T^{1/2},\quad \left\|\widetilde{U}_2\right\|\le\|\bar{U}\|\lesssim_p N^{1/2}+T^{1/2}.\tag{B.142}$$

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

$$\|K_1\|\le T^{-3/2}\lambda_j^{-3/2}\widehat{\lambda}_j^{-1/2}\|L\Lambda\|\left\|\bar{V}\right\|\left\|\widetilde{U}_2^\intercal\widetilde{U}_2\right\|\|\xi_j\|\lesssim_p\lambda_j^{-1}(T^{-1}N+1),$$

where we also use $\left\|\widetilde{U}_2^\intercal\widetilde{U}_2\right\|\le\left\|\bar{U}^\intercal\bar{U}\right\|\lesssim_p N+T$ in the last equation. For the second term, we have

$$\|K_2\|\le T^{-3/2}\lambda_j^{-3/2}\widehat{\lambda}_j^{-1/2}\left\|\Lambda^{1/2}L\right\|\left\|\widetilde{U}_1\right\|\left\|\widetilde{U}_2^\intercal\widetilde{U}_2\right\|\|\xi_j\|\lesssim_p\lambda_j^{-3/2}(T^{-1}N+1).$$

For the third term, we have

$$\|K_3\|\le\lambda_j^{-1}\|L\Lambda\|\left\|T^{-1}\bar{V}\bar{V}^\intercal-\mathbb{I}_p\right\|\left\|(\Lambda/\lambda_j)^{1/2}s_{j1}\right\|\lesssim_p T^{-1/2}\left\|(\Lambda/\lambda_j)^{1/2}s_{j1}\right\|.$$

For the forth term, we have

$$\|K_4\|\le T^{-1}\lambda_j^{-3/2}\|L\Lambda\|\left\|\bar{V}\widetilde{U}_1^\intercal\right\|\lesssim_p\lambda_j^{-1/2}T^{-1/2},$$

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

$$\|K_5\|\le T^{-1}\lambda_j^{-1}\left\|L\Lambda^{1/2}\right\|\left\|\widetilde{U}_1\bar{V}^\intercal\right\|\left\|(\Lambda/\lambda_j)^{1/2}s_{j1}\right\|\lesssim_p\lambda_j^{-1/2}T^{-1/2}\left\|(\Lambda/\lambda_j)^{1/2}s_{j1}\right\|.$$

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

$$\|K_7\| \leq \lambda_j^{-2}\left|\widehat{\lambda}_j - \lambda_j\right|\left\|L\Lambda^{1/2}\right\| \lesssim_p \lambda_j^{-1/2}(T^{-1/2}N^{1/2}+1) + T^{-1/2}.$$

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

$$\|s_{j2}\| \leq \left\|\frac{\widetilde{U}_2}{\sqrt{T\lambda_j}}\right\|\left\|\left(\frac{\lambda_j}{\widehat{\lambda}_j}\right)^{\frac{1}{2}}\right\|\|\xi_j\| \lesssim_p \lambda_j^{-1/2}(T^{-1/2}N^{1/2}+1). \tag{B.145}$$

This concludes the proof. □

Lemma 13. *Under the assumptions of Theorem 4(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) $\left\|\frac{\bar{V}\bar{U}^\intercal\varsigma_i}{T\sqrt{\widehat{\lambda}_i}}\right\| \lesssim_p \frac{1}{T} + \frac{N+T}{T\lambda_i}, \quad \left|\frac{\varsigma_i^\intercal\bar{u}}{\sqrt{\widehat{\lambda}_i}}\right| \lesssim_p \frac{1}{T} + \frac{N+T}{T\lambda_i}$.

(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

$$\|K_2\| \le \|s_{i2}\| \left\|\widetilde{U}_2\right\| \left\|\widetilde{U}_1\right\| \lesssim_p \sqrt{\frac{N+T}{T\lambda_i}}\left(\sqrt{N}+\sqrt{T}\right)\sqrt{T} \lesssim_p \lambda_i^{-1/2}(N+T).$$

For the third term, we have

$$\|K_3\| \le \left(\left\|s_{i1}^\intercal \Lambda^{\frac{1}{2}}\right\| \|\bar{V}\| + \left\|\widetilde{U}_1\right\|\right)\left\|\widetilde{U}_2\right\| \|s_{j2}\| \lesssim_p \sqrt{\lambda_i T}\left(\sqrt{N}+\sqrt{T}\right)\sqrt{\frac{N+T}{T\lambda_j}} = \lambda_j^{-1/2}\lambda_i^{1/2}(N+T).$$

For the last term, we have

$$\|K_4\| \le \left\|\widetilde{U}_2\widetilde{U}_2^\intercal\right\| \|s_{i2}\| \|s_{j2}\| \lesssim_p \lambda_i^{-1/2}\lambda_j^{-1/2}T^{-1}(N+T)^2.$$

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

$$\left\|\bar{V}\bar{U}^\intercal\varsigma_i\right\| \le \left\|\bar{V}\widetilde{U}_1^\intercal s_{i1}\right\| + \left\|\bar{V}\widetilde{U}_2^\intercal s_{i2}\right\| \le \left\|\bar{V}\widetilde{U}_1^\intercal\right\| + \|\bar{V}\|\,\|\bar{U}\|\,\|s_{i2}\| \lesssim_p \sqrt{T} + \frac{N+T}{\sqrt{\lambda_i}}.$$

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 4(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

$$\left|\tilde{h}_{l1}^\intercal \tilde{h}_{k2} - \delta_{lk}\right| \lesssim_p \frac{1}{T} + \frac{N+T}{T\min\{\lambda_l, \lambda_k\}} \leq \frac{1}{T} + \frac{N+T}{T\lambda_p},$$

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.12, 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.11:

$$\|\bar{r} - \mathrm{E}(r_t)\|_\infty \lesssim_p \sqrt{\frac{\log N}{T}}, \quad \text{and} \quad \left\|\Sigma b - \widehat{\Sigma}b\right\|_\infty = \left\|\frac{1}{T}\bar{R}\bar{R}^\intercal b - \mathrm{Cov}(r_t, r_t^\intercal b)\right\|_\infty \lesssim_p \sqrt{\frac{\log N}{T}}.$$

The second inequality follows immediately from Assumption A.11:

$$\|b^\intercal(\bar{r} - \mathrm{E}(r_t))\| = \left|\frac{1}{T}\sum_{t=1}^{T} m_t - \mathrm{E}(m_t)\right| \lesssim_p \frac{1}{\sqrt{T}}.$$

□