The Journal of FINANCE
The Journal of THE AMERICAN FINANCE ASSOCIATION

THE JOURNAL OF FINANCE • VOL. LXXX, NO. 1 • FEBRUARY 2025

Test Assets and Weak Factors

STEFANO GIGLIO, DACHENG XIU, and DAKE ZHANG*

ABSTRACT

We show that two important issues in empirical asset pricing—the presence of weak factors and the selection of test assets—are deeply connected. Since weak factors are those to which test assets have limited exposure, an appropriate selection of test assets can improve the strength of factors. Building on this insight, we introduce supervised principal component analysis (SPCA), a methodology that iterates supervised selection, principal-component estimation, and factor projection. It enables risk premia estimation and factor model diagnosis even when weak factors are present and not all factors are observed. We establish SPCA's asymptotic properties and showcase its empirical applications.

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

Another issue that the researcher faces is the potential presence of weak factors. Broadly speaking, the factor of interest to the researcher is one of many factors that potentially drive returns. Some of these factors may be weak. That

*Stefano Giglio is at Yale School of Management, NBER, and CEPR. Dacheng Xiu is at Booth School of Business, University of Chicago and NBER. Dake Zhang is at Antai College of Economics and Management, SJTU-BOC Institute of Technology and Finance, Shanghai Jiao Tong University. We benefited tremendously from discussions with seminar and conference participants at New York University, Yale School of Management, Chicago Booth School of Business, the University of California, Los Angeles, London School of Economics, University of Michigan, Duke University, INSEAD Business School, Cornell University, University of North Carolina at Chapel Hill, University of Miami, Monash University, Temple University, Tilburg University, Rotterdam, Rutgers University, University of Texas at Dallas, ESSEC Business School, Chinese University of Hong Kong, University of Connecticut, Federal Reserve Board, ITAM Business School, Vienna Graduate School Finance, Paris Dauphine University, Goethe University, Durham University, Baruch College, Lund University, AFA 2022 Annual Meeting, 2022 WFA Meeting, 48th EFA Annual Meeting, China International Conference in Finance, and 13th Annual SoFiE Conference. We have read *The Journal of Finance* disclosure policy and have no conflicts of interest to disclose.

Correspondence: Dake Zhang, Antai College of Economics and Management, Shanghai Jiao Tong University, 1954 Huashan Road, Shanghai, China; e-mail: dk.zhang@sjtu.edu.cn.

DOI: 10.1111/jofi.13415

is, the available test assets have little or no exposure to them. This makes it difficult to learn about them using the available assets. The presence of the weak factors also contaminates inference about the entire model as the literature shows that the presence of a weak factor biases the estimation of the risk premia of all factors, including the one of interest to the researcher (whether that factor itself is strong or weak) as well as inference about the model's pricing ability. To make things worse, a weak factor could be latent, so that we may not even know it exists in the first place.

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

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

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

Estimating and testing the risk premium of a factor of interest requires properly controlling for all other factors relevant to investors (whether they are observed or latent) to avoid omitted variable bias (see, e.g., Giglio and Xiu (2021)).[1] Giglio and Xiu (2021) propose doing so by first estimating a latent factor model for the stochastic discount factor (SDF) using principal component analysis (PCA), and then using this model to estimate the risk premium of the factor of interest. This approach eliminates the need for explicit specification of all the control factors, but relies on the assumption that all the latent factors driving the SDF are pervasive (i.e., strong). Our SPCA procedure also uses

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

PCA to extract latent factors while remaining agnostic about the identities of the control factors. However, it exploits correlations with the factor of interest as a guiding criterion for selecting a subset of test assets, before applying PCA. This results in a versatile methodology that remains robust even in scenarios in which certain factors are omitted, including cases in which these omitted factors are weak.

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

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

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

The choice of test assets in the literature mainly follows one of three approaches. The vast majority of the literature adopts a "standard" set of portfolios sorted by a few characteristics, such as size and value, following the seminal work by Fama and French (1993). A second approach employed in more recent work (e.g., Kozak, Nagel, and Santosh (2020)) expands this cross section to include portfolios sorted by a much larger set of characteristics discovered in the last decades, on the order of hundreds of portfolios. Finally, a third approach, (e.g., Ang et al. (2006)) focuses more closely on the factor of interest by sorting assets into portfolios by their estimated exposure to the factor and estimating risk premia using those portfolios expected to be particularly informative about that factor.

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

In this paper, we derive the asymptotic properties of SPCA in a setting that allows for weak factors and for test assets with highly correlated risk exposures. The latter scenario potentially involves the same (asymptotically) rank-deficiency issue as weak factors. We also analyze in this setting alternative estimators proposed in recent literature, that 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, while the Lasso approach is consistent for estimation of the SDF and risk premia, but is generally not as efficient as SPCA. In addition, we perform an extensive set

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

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

Finally, we illustrate the use of SPCA for estimating risk premia of a variety of tradable and nontradable factors proposed in the asset pricing literature, and for diagnosing observable factor models. Using the large cross section of test portfolios produced by Chen and Zimmermann (2022) and Hou, Xue, and Zhang (2020), covering more than 900 and 1,600 portfolios, respectively, for the period 1976 to 2020, we apply SPCA to estimate factor risk premia and evaluate its out-of-sample performance. Almost none of the nontradable factors are priced, except for the intermediary capital factor. We also explore the robustness of SPCA to the weakness of factors by artificially changing the set of test assets used in the estimation. For example, we show that SPCA is able to recover the risk premium for momentum even when momentum assets are removed from the original set of test assets (and therefore the momentum factor is weak in the cross section). Moreover, we illustrate empirically how SPCA can be used to diagnose whether observable factor models are missing important priced factors.

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

[2] Related literature also include Gospodinov, Kan, and Robotti (2013, 2014). However, Pesaran and Smith (2019) investigate the impact of factor strength and pricing error on risk premium estimation and show that the conventional two-pass risk premium estimator converges at a lower rate as the factors become weaker.

The issue of weak factors is particularly important in empirical work in asset pricing because most economically motivated factors (e.g., most macroeconomic factors) do appear to be weak in practice. Moreover, a statistical problem analogous to weak factors arises when betas are collinear, that is, some factors are redundant in terms of explaining the variation in 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, Liu, and Zhu (2016)), many of which are close cousins and do not add any explanatory power (Feng, Giglio, and Xiu (2020)). The weak factor problem appears to be caused by having seemingly more factors than necessary, which is why some suggest eliminating such factors (Bryzgalova (2015)) or shrinking their risk premia estimates (Bryzgalova, Huang, and Julliard (2023)) to improve the estimates for strong factors. We instead argue that the weak factor problem is fundamentally an issue of test asset selection. Since weaker factors may still be priced, our solution is to accommodate them using an adapted procedure with carefully selected test assets.[3]

Several recent papers propose different methodologies to address weak factors. Lettau and Pelger (2020) propose an estimator of the SDF in the presence of weak factors, risk premium PCA (rpPCA), which generalizes PCA with a penalty term that accounts for expected returns. While this estimator features desirable properties as explored by Lettau and Pelger (2020), we show that it is inconsistent for estimating risk premia in the weak-factor setting we consider.[4] Anatolyev and Mikusheva (2022) propose a complementary four-split approach to address weak factors, that is, based on sample-splitting and instrumental variables. This alternative procedure works well to address the weak factor bias, though it does not address omitted priced factors or measurement error in the factors.[5]

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

[4] Lettau and Pelger (2020) focus on the case in which factors are extremely weak—so much so that they are not statistically distinguishable from idiosyncratic noise. In that case, no estimator can be consistent for either risk premia or the SDF. They show that rpPCA does not recover consistently the SDF, but in simulations it correlates with the SDF more than the SDF estimator obtained from standard PCA. Rather than focusing on this extreme case of weak factors, our theory covers a range of factor weaknesses, which include from strong to very weak, and which still permits consistent estimation of factors and risk premia. Formally, we study the case in which 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.

[5] Our paper also relates to a growing strand of econometrics literature on weak factor models. Bai and Ng (2023) show that PCA can recover moderately weak factors at the cost of efficiency. Bai and Ng (2008) and Huang et al. (2022) propose supervised learning methods in the context of factor-based forecasting. Fan, Ke, and Liao (2021) exploit information from observed proxies to improve the estimation of factor models, and Wan et al.

Our paper also relates to a literature that explores different methods to form portfolios to test asset pricing models, like Ahn, Conrad, and Dittmar (2009) or Bryzgalova, Pelger, and Zhu (2020). These methods are useful in helping build or expand the initial cross section for SPCA. In this paper, we use the simpler approach of working with an existing large cross section of portfolios sorted by firm characteristics, as in Chen and Zimmermann (2022) and Hou, Xue, and Zhang (2020).

The concept of SPCA originated from a cancer diagnosis technique applied to DNA microarray data by Bair and Tibshirani (2004), and was later formalized by Bair et al. (2006) in a prediction framework in which some predictors are not correlated with the latent factors that drive the outcome of interest. Bair et al. (2006) suggest a screening step using marginal correlations between predictors and the outcome variable to select the subset of useful predictors before applying the standard PCA to this subset.[6] They prove the consistency of this procedure, but rely on a restrictive identification assumption that any important predictor must also have a substantial marginal correlation with the outcome. We provide several examples of multivariate factor models in which this assumption fails. While the screening step of our SPCA procedure is in similar spirit as theirs (in the sense that their outcome variable is our factor of interest, and their predictors are our test assets), our projection step and the subsequent iteration procedure are new, and are introduced specifically to eliminate the strong identification assumption used in the existing statistics literature. Moreover, our focus is not on prediction per se, but instead on parameter inference.

The remainder of the paper is organized as follows. Section I discusses the SPCA methodology. Section II presents the finite-sample performance of SPCA via simulations. Section III provides our empirical analysis. Finally, Section IV concludes.

I. Methodology

To rigorously address the challenge of weak factors, our approach begins with the specification of a general data-generating process (DGP). We note that within this population model, the concept of weak factors holds no relevance. In population, researchers aiming to identify the risk premium of a factor such as g_t would ideally use all available assets for this purpose.

The real-world (finite-sample) scenario, however, diverges from this idealized population model. In particular, we encounter practical constraints such

(2024) consider moderately weak factors as in Bai and Ng (2023) in this context. Fan and Liao (2022) propose extracting factors by diversifying away idiosyncratic noise directly. Uematsu and Yamagata (2022a) adopt a variant of the sparse PCA algorithm proposed in Uematsu et al. (2019) to estimate a sparsity-induced weak factor model. Uematsu and Yamagata (2022b) provide inference results in that sparse model. Freyaldenhoven (2019) and Bailey, Kapetanios, and Pesaran (2021) adopt a similar framework for estimating factor count and strength.

[6] The screening approach has also been adopted in contexts such as classification and regression. See Fan and Fan (2008) and Fan and Lv (2008).

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

A. Model Setup

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

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

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

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

Note that we model risk exposures (β) as constant: we implicitly assume that the test assets are portfolios sorted so that their factor exposures are modeled as constant, as in Giglio and Xiu (2021). Alternatively, one could work directly with individual stocks (which generally have time-varying risk exposure), combining our procedure with the methodologies of Gagliardini, Ossola, and Scaillet (2016), Kelly, Pruitt, and Su (2019), or Kim, Korajczyk, and Neuhierl (2021) to account for the time variation in betas.

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

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

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

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

where b is an $N \times 1$ vector of SDF loadings that satisfies $\mathrm{E}(r_t) = \Sigma b$, where Σ is the covariance matrix of r_t (see, e.g., Kozak, Nagel, and Santosh (2020)). The

[7] In Section III.B of the Internet Appendix, we discuss the case in which factors are observable, and in Section III.C of the Internet Appendix, we discuss the case in which the zero-beta rate needs to be estimated. The Internet Appendix may be found in the online version of this article.

relationship between the two SDFs depends on the degree of market completeness. As will be shown later, these two forms of the SDF are asymptotically equivalent in the asymptotic scheme we consider, with the number of assets N going to infinity, so that there is no ambiguity with respect to which estimand we consider.

In addition to the SDF, we are also interested in estimating the risk premium of one or more observable factors, summarized in $d \times 1$ vector g_t. It is important to emphasize that g_t is a proxy for some risks, constructed or otherwise chosen by the researcher ex ante, not necessarily tradable, and typically motivated from economic theory. Following Giglio and Xiu (2021), we do not impose that g_t is part of or identical to v_t; instead, we assume that 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.[8] This model clearly nests the classic linear asset pricing model with observable factors only, in which case we can set $\eta = \mathbb{I}_p$ and $z_t = 0$. To price g_t, we can simply use the SDF given by (4), as g_t's risk premium is given by $\gamma_g = -\mathrm{cov}(g_t, m_t) = \eta\gamma$.

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

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

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

[9] We only need u_t to be stationary (so that Σ_u is well defined) when we discuss the SDF in Section I.C. For risk premia estimation, we instead impose a weaker condition, namely, Assumption IA4, which plays a similar role as $\|\Sigma_u\| \lesssim 1$.

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

In the econometrics literature on factor models (e.g., Bai and Ng (2002)), the setup described in (1) is typically complemented by the assumption that $\lambda_i(\beta^\intercal\beta) \asymp N$ for $i = 1, 2, \ldots, p$: all eigenvalues of the matrix $\beta^\intercal\beta$ grow at rate N, so that *all* factors are *pervasive*. Informally, even as the number of test assets N is large, a sufficiently large number of assets are well exposed to each of the risk factors (their β with respect to all factors is nonvanishing for a large number of assets). Under this assumption, as we see below, standard PCA works well to recover the latent factors v_t.

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

How important do we expect these weak factors to be in practice? Consider Figure 4, which provides a scree plot of the eigenvalues of returns from our empirical analysis based on a large cross section of 950 assets. The figure shows that the first one or two eigenvalues are clearly much larger than the others, but the absence of clear gaps among the remaining eigenvalues suggests that several factors beyond the first two may be weak. Despite the large cross section, their eigenvalues remain relatively small and difficult to distinguish from idiosyncratic error.

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

B. Estimating Risk Premia When Factors are Weak

We begin our analysis with risk premia estimation.

B.1. The Benchmark PCA-Based Estimator

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

Algorithm 1 (PCA-based Estimator of Risk Premia): The estimator proceeds as follows:

Inputs: $\bar{R}$ and $\bar{G}$, the matrices of demeaned returns and demeaned g_t, respectively.[10]

S1. Apply singular-value decomposition (SVD) on $\bar{R}$, and write the first p right singular vectors as ξ. The estimated factors are given by $\widehat{V} = \sqrt{T}\xi^{\intercal}$.
S2. Estimate the risk premia of $\widehat{V}$ by $\widehat{\gamma} = (\widehat{\beta}^{\intercal}\widehat{\beta})^{-1}\widehat{\beta}^{\intercal}\bar{r}$, where $\widehat{\beta} = \bar{R}\widehat{V}^{\intercal}(\widehat{V}\widehat{V}^{\intercal})^{-1}$ and $\bar{r}$ is the vector of average excess returns.
S3. Estimate the factor loading of g_t on v_t by $\widehat{\eta} = \bar{G}\widehat{V}^{\intercal}(\widehat{V}\widehat{V}^{\intercal})^{-1}$.

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

As discussed in Giglio and Xiu (2021), one interpretation of this estimator is that it builds a mimicking portfolio for the factor g_t by projecting it onto the first p principal components of the space of returns. A mimicking portfolio would ideally be built directly using *all* possible assets. But when N is large, this can be inefficient or even infeasible (if $N > T$). The three-step estimator effectively regularizes the mimicking portfolio problem by using only p portfolios appropriately constructed as basis assets, that is, the principal components of the returns. Giglio and Xiu (2021) establish the consistency of this estimator and derive its asymptotic inference, in the case that all latent factors are strong. This procedure also recovers the SDF because it consistently estimates all latent factors, $\widehat{v}_t$ (columns of $\widehat{V}$), that drive the SDF, along with their SDF loadings as in (2), $\widehat{\Sigma}_v^{-1}\widehat{\gamma}$.

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

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

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

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

Estimation of the latent factors v_t via PCA involves recovering the matrix of risk exposures β from the covariance matrix of realized returns, $\widehat{\Sigma}$. A successful recovery of β via PCA of realized returns therefore requires a favorable signal-to-noise ratio. If the "signal," as measured by $\|\beta\|$, dominates the "noise," which arises from the idiosyncratic component Σ_u and the estimation error in the sample covariance matrix $\widehat{\Sigma} - \Sigma$, then the first sample eigenvector of $\widehat{\Sigma}$ would (approximately) span the same space spanned by the true β. Thus, using $\widehat{\beta}$, effectively the eigenvector of $\widehat{\Sigma}$, in the cross-sectional regression step (Step S2 of the estimator) would yield a consistent estimator of the risk premium of the estimated latent factor and lead in turn to a consistent estimator of the risk premium of g_t. Otherwise, if the signal $\|\beta\|$ is so weak that the estimation error in $\widehat{\beta}$ dominates, there would be a nonvanishing angle between the space spanned by $\widehat{\beta}$ and that by β. But estimating risk premia requires comparing the average returns of assets with different betas (e.g., computing the slope in a cross-sectional regression). "Measurement" error in the betas thereby induces a bias in the risk premium estimate.

Proposition 1 shows that the PCA-based estimator is consistent only if $N/(\|\beta\|^2 T) \to 0$. This condition formalizes our notion of factor weakness. In a one-factor model, the factor is weak if this condition fails. We generalize this definition for the case of multiple factors later.

PROPOSITION 1: *Suppose that test asset returns follow a single-factor model in the form of (1) with $p = 1$, g_t satisfies (4) with $d = 1$, u_t and v_t are i.i.d. normally distributed and mutually independent, and $z_t = 0$. In addition, suppose that β satisfies $N/(\|\beta\|^2 T) \to B \geq 0$ and $\|\beta\| \to \infty$. We then have that $\widehat{\gamma}_g^{PCA} \xrightarrow{\text{P}} (1 + B)^{-1}\eta\gamma$.*

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

This example also shows that the risk premium estimator could be biased even if we have a consistent estimator of the factors. In fact, the estimated factors in $\widehat{V}$ are consistent under the assumptions of Proposition 1 in the sense that $|\text{corr}(\widehat{V}, V)| \xrightarrow{\text{P}} 1$.[11] However, estimating a large-dimensional vector β given $\widehat{V}$ remains a challenging problem, which also requires $B = 0$ for consistency.

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

B.2. Our Solution: SPCA and Test Asset Selection

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

We formalize the problem by imposing the assumption that there exists a subset $I_0 \subset \langle N \rangle$[12] within which test assets feature a strong factor structure. In other words, there exists a subset of assets that are sufficiently *informative* about latent factors driving test asset returns. To be clear, we do not make any assumption about the remaining test assets in the complement set of I_0—they may or may not be informative. Such a set is thus not uniquely defined. In this regard, this assumption is relatively mild.

To see how this assumption helps, note that in the population model of Proposition 1, the expected excess return of g_t's mimicking portfolio *built only with test assets in* I_0 is

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

[11] We can further establish that a sufficient condition for consistent recovery of factors is $N/(\|\beta\|^4 T) \to 0$, which clearly holds in the setup of Proposition 1.

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

where $r_{t,[I_0]}$ denotes the vector of returns of test assets in I_0, and $\beta_{[I_0]}$ is their corresponding beta.[13] It can be shown that

$$\text{cov}(g_t, r_{t,[I_0]})\text{cov}(r_{t,[I_0]})^{-1}\text{E}(r_{t,[I_0]}) = \eta\gamma + O(\|\beta_{[I_0]}\|^{-2}) \tag{5}$$

(see the proof in a more general setting in Proposition IA4 of the Internet Appendix). Since test assets in I_0 feature a strong factor structure, $\|\beta_{[I_0]}\|^2 \asymp |I_0| =: N_0$,[14] the approximation error is given by $O(N_0^{-1})$. This result establishes the fact that in population using a smaller number of sufficiently informative assets leads to an asymptotically vanishing error in approximating the risk premium. Moreover, it holds that $N_0/(\|\beta_{[I_0]}\|^2 T) = O(T^{-1})$, that is, factors are pervasive within this subset. Therefore, as long as we locate a subset that satisfies the properties of I_0, we can estimate g_t's risk premium consistently with PCA by only using test assets within this subset.

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

Unfortunately, we do not know ex ante such a set, that is, which assets are informative about the latent factor v_t. Rather than using all assets, the idea of SPCA revolves around selecting the most informative assets based on their covariances with g_t. In the DGP of Proposition 1, the group of assets exhibiting high covariances with g_t comprises those with large magnitudes of β's. Screening via correlation therefore selects a subset of assets satisfying the desirable properties of I_0.

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

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

[13] We use $A_{[I]}$ to denote a submatrix of A whose rows are indexed in I.
[14] For an index set $I \subset \langle N \rangle$, we use $|I|$ to denote its cardinality.

formalizes the insight behind these traditional procedures, the fundamental motivation behind our approach is to circumvent the adoption of arbitrary priors when selecting assets.

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

Algorithm 2 (SPCA-Based Estimator of Risk Premia for a Single-Factor Model ($p = 1$)): The procedure is as follows:

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

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

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

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

SPCA (Algorithm 2) adds the screening step, Step S1, to the PCA-based risk-premium estimation method of Giglio and Xiu (2021) (Algorithm 1). In this step, out of the N assets available, only a subset $\widehat{I}$ is selected, and the three steps of Algorithm 1 are applied to this subset only.

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

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

PROPOSITION 2: *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. The loading matrix* β *satisfies* $\|\beta\|_{\text{MAX}} \lesssim 1$ *and there exists a subset* $I_0 \subset \langle N \rangle$ *such that* $\left\| \beta_{[I_0]} \right\| \asymp \sqrt{N_0}$ *where* $N_0 = |I_0| \to \infty$. *Then, for any choice of* q *in Algorithm 2 such that* $qN/N_0 \to 0$,[16] $qN \to \infty$, *and* $|\beta|_{\{qN+1\}} \leq (1+\delta)^{-1} |\beta|_{\{qN\}}$[17] *for some* $\delta > 0$, *where* $|\beta|_{\{k\}}$ *denotes the* k^{th} *largest value in* $\left\{ |\beta_{[i]}| \right\}_{i \in \langle N \rangle}$, *we have* $\widehat{\gamma}_g^{SPCA} \xrightarrow{\text{P}} \eta\gamma$.

[15] We discuss the case of a multivariate $\bar{G}$ (a $d \times T$ matrix) in Section I.B.4.

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

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

To gain a better understanding of the intuition, we delve into some key steps of the proof, which is detailed in the Internet Appendix. Given a specific choice of the tuning parameter q, we can identify the population counterpart of $\widehat{I}$, denoted by I. This set I consists of the qN largest entries of β in terms of their magnitudes, as specified before Assumption IA7 in the Internet Appendix.[18] The proof of Proposition 2 establishes the consistency of the selected set $\widehat{I}$ (which contains the top qN test assets with the largest sample covariances with g_t) with respect to I in the following sense: $\mathrm{P}(\widehat{I} = I) \to 1$.

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

Second, the condition that there exists I_0 such that $\left\|\beta_{[I_0]}\right\|^2 \asymp N_0$ and $qN/N_0 \to 0$ guarantees the existence of at least qN test assets with nonzero population covariances with g_t. Thus, according to the definition of I, the smallest population covariance with g_t among all test assets in I must be nonzero. This suggests that $\left\|\beta_{[I]}\right\|^2 \asymp |I| = qN$. Furthermore, since we assume a nonvanishing gap between the $(qN)^{\text{th}}$ and $(qN+1)^{\text{th}}$ population covariances, it follows that the set of test assets with largest population covariances must coincide with those having the largest sample covariances because the vanishing estimation error is dominated by this nonvanishing gap in the asymptotic context.

Given that the identified set I can function as I_0 (since $\left\|\beta_{[I]}\right\|^2 \asymp |I|$), then as demonstrated in equation (5), we can directly approximate the risk premium of g_t using its mimicking portfolio built on this subset I of test assets. The consistency of the risk premium estimate thereby follows from the consistency of $\widehat{I}$ in the recovery of I.

Propositions 1 and IA1–IA3 show that in the single-factor case, the consistency of PCA, Ridge, PLS, and rpPCA requires $B = 0$. Suppose $\|\beta\|^2 = N^{\upsilon}$, for some $\upsilon > 0$. Then, $B = 0$ is equivalent to $N^{1-\upsilon}/T \to 0$. The consistency of SPCA, as shown by Proposition 2, nonetheless, requires only $(\log N)/T \to 0$.[19]

B.3. SPCA in the General Case: Selection and Projection

Propositions 1 and 2 focus on an unrealistic single-factor model since they are meant to illustrate the failure of PCA due to the presence of a weak factor as well as the intuition behind our procedure. In general, the DGP of returns

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

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

is likely driven by more than one factor, with these factors generally having different strength in any specific cross section of test assets. Note also that g_t could have more than one dimension in the general setup (4). In this section, we show how to generalize SPCA to the case in which multiple factors of heterogeneous strength are present.

To begin, in the same spirit as Proposition 1, we can show that a general necessary condition for the consistency of PCA in a multifactor model is that

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

If this holds, then even the weakest factor among all p factors in (1) is sufficiently strong that it can be recovered by PCA. In this case, the three-pass estimator of Giglio and Xiu (2021) would properly recover the risk premium of any factor g_t. We therefore define weak factors as those for which test asset exposures fail condition (6). This is a compact formal description of the nonideal finite-sample environment encountered in practice.[20]

As in the single-factor case, in the multifactor case condition (6) can fail if one of the factors is not pervasive. However, in the multifactor case, all factors can be individually strong and condition (6) still fails because the factors' exposures are highly correlated. Consider, for example, a two-factor model in which the beta matrix has the form

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

where β_{11} and β_{12} are $N_0 \times 1$ vectors, β_{21} and β_{22} are $(N - N_0) \times 1$ vectors, and N_0 is small relative to N. Suppose that $\beta_{21} = \beta_{22}$. In this setup, we can identify two groups of test assets. The first is a small group of N_0 test assets, with exposures β_{11} to the first factor and β_{12} to the second factor. The second is a large group of $(N - N_0)$ assets that have the *same* exposure to both factors (since $\beta_{21} = \beta_{22}$). In this case, we can show that condition (6) can fail: even if each factor is strong individually, there is a "rank deficiency" issue in the betas. The reason is that most of the assets (group 2) do not contain information that can separate the risk premia of the two factors because they are equally exposed to the two factors. This loss of information turns out to have exactly the same effect on estimation and inference as the weak factor issue.[21] We therefore need a procedure that consistently estimates risk premia in this case as well.

It is also important to note that in the general case with multiple factors of potentially different strength, a simple extension of Algorithm 2, operating an initial screening (Step S1) and then extracting *multiple* factors via PCA (Step S2) would *not* actually work to recover all factors. To see this, take (7) as an

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

[21] Formally, we can show that $\lambda_{\min}(\beta^\intercal\beta) \leq \|\beta_{11} - \beta_{12}\|^2/2 \lesssim N_0$. As a result, $N/(\lambda_{\min}(\beta^\intercal\beta)T) \gtrsim N/(N_0T)$, which does not necessarily converge to zero if N_0 and T are small, so that condition (6) could fail.

example. Suppose now that $\beta_{21} \neq \beta_{22}$, but $\beta_{22} = 0$, that is, most of the assets have zero exposure to the second factor. In this case, the first factor is strong while the second factor is weak.[22] Now suppose that $\eta = (1, 1)$, implying that the observed factor g is correlated with both factors and, by extension, with all the test assets. The determination of which assets to exclude via screening now hinges on the betas of these test assets. If a majority of the selected assets pertain to the second group, the subsequent application of PCA in Step S2 would only recover the first factor. This would occur if condition (6) fails for the selected assets. In contrast, if many of the selected assets belong to the first group, PCA applied to them has the potential to recover both factors. In this case, the first principal component may capture a linear combination of both the strong and weak factors. This example demonstrates that even though screening assets ensures that the *first* principal component after screening recovers one factor (which could be the strong factor, the weak factor or their mixture on the basis of the original cross section), there is no guarantee that this procedure can solve the weak factor issue in one shot.

We next provide an example that shows that, in some scenarios screening can eliminate too many assets, making a strong factor model become weak or even rank-deficient. Suppose that β has the form

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

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

To address the aforementioned issues, we propose a multistep version of SPCA that iteratively conducts selection and projection. Step S1 of Algorithm 2 described above—valid when there is only one factor—can help identify one strong factor from a selected subset of test assets. In a nutshell, the multistep SPCA, described in Algorithm 3 iteratively applies Algorithm 2 to extract a new factor, with a projection step designed to ensure that each new factor is orthogonal to the estimated factors in the previous steps, similar to the factors extracted by the standard PCA.

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

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

Formally, the algorithm is given as follows.

Algorithm 3 (Selection and Projection): The iterative SPCA procedure for risk premia estimation is as follows:

Inputs: $\bar{R}_{(1)} := \bar{R}$, $\bar{r}_{(1)} := \bar{r}$, and $\bar{G}_{(1)} := \bar{G}$, a $d \times T$ vector.

S1. For $k = 1, 2, \ldots$ iterate the following steps using $\bar{R}_{(k)}$, $\bar{r}_{(k)}$, and $\bar{G}_{(k)}$:
   a. Select an appropriate subset $\widehat{I}_k \subset \langle N \rangle$.
   b. Repeat Steps S1 to S3 of Algorithm 1 with selected return matrix $(\bar{R}_{(k)})_{[\widehat{I}_k]}$ and $\bar{G}_{(k)}$ to extract only the first principle component. Denote the estimates by $\widehat{V}_{(k)}$, $\widehat{\eta}_{(k)}$, $\widehat{\gamma}_{(k)}$.
   c. Estimate the exposure of $\bar{R}_{(k)}$ to $\widehat{V}_{(k)}$ by $\widehat{\beta}_{(k)} = T^{-1}\bar{R}_{(k)}\widehat{V}_{(k)}^{\intercal}$.
   d. Obtain $\bar{R}_{(k+1)} = \bar{R}_{(k)} - \widehat{\beta}_{(k)}\widehat{V}_{(k)}$, $\bar{r}_{(k+1)} = \bar{r}_{(k)} - \widehat{\beta}_{(k)}\widehat{\gamma}_{(k)}$, and $\bar{G}_{(k+1)} = \bar{G}_{(k)} - \widehat{\eta}_{(k)}\widehat{V}_{(k)}$.

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

S2. Estimate risk premia by $\widehat{\gamma}_g^{SPCA} = \sum_{k=1}^{\widehat{p}} \widehat{\eta}_{(k)}\widehat{\gamma}_{(k)}$.

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

Each iteration k of the procedure recovers one latent factor $\widehat{V}_{(k)}$, estimates its risk premium $\widehat{\gamma}_{(k)}$, and estimates the exposure of g_t to that factor, $\widehat{\eta}_{(k)}$. In Step S1, there is first asset selection (Step S1.a). Next, the three-step estimator of risk premia of Giglio and Xiu (2021) is applied using the selected assets (Step S1.b) to recover the k^{th} factor $\widehat{V}_{(k)}$ in addition to $\widehat{\gamma}_{(k)}$ and $\widehat{\eta}_{(k)}$, which are specific to that factor. Then, in Step S1.c, we project the returns of all assets (not just those selected) on the estimated factor $\widehat{V}_{(k)}$, and in Step S1.d we compute the residuals of this projection for returns and the factor g_t itself. Therefore, at the end of Step S1, we have completely eliminated the effect of the k^{th} factor on returns and the target factor g_t. We then repeat Step S1 again, this time using the residuals of returns and g_t, looking for the next factor. Iteration continues for $\widehat{p}$ steps. At the end, Step S2 combines the $\widehat{\gamma}_{(k)}$ and the $\widehat{\eta}_{(k)}$ obtained at each step into an estimator $\widehat{\gamma}_g^{SPCA}$ for the risk premia of g_t.

Algorithm 3 requires an appropriate choice of $\widehat{I}_k$ and a stopping rule. One choice for $\widehat{I}_k$ is[24]

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

[24] Using covariance for screening allows us to replace all $\bar{G}_{(k)}$ in the definition of $\widehat{I}_k$ and Algorithm 3 by $\bar{G}$, that is, only the projections of $\bar{R}_{(k)}$ and $\bar{r}_{(k)}$ are needed, because this replacement would not affect the covariance between $\bar{G}_{(k)}$ and $\bar{R}_{(k)}$, and in turn, the test assets after screening and the estimates of $\widehat{\eta}_{(k)}$. We use this fact in the proofs, which simplifies the notation. We can also use correlation instead of covariance in constructing $\widehat{I}_k$. While this does not affect the asymptotic analysis, we find that correlation screening performs slightly better in finite samples.

Correspondingly, we set the stopping criterion as

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

In other words, we select test assets that have predictive power for at least one variable in g_t and stop when most test assets are uncorrelated with all variables in g_t. With a good choice of tuning parameters, q and c, the iteration stops as soon as most projected residuals of returns appear uncorrelated with the projected residuals of g_t, which implies that all factors that are correlated with g_t are successfully recovered.

It is helpful to revisit the aforementioned examples and understand how the new procedure fixes issues with the one-step SPCA. Recall that in example (7), $\beta_{22} = 0$ and $g_t = v_{1t} + v_{2t}$. As discussed previously, screening will select a subset of q assets that are spread across both groups of assets since they are all correlated with g_t. Consequently, applying PCA to them will identify a factor that is in general spanned by v_{1t} and v_{2t}. Even if this first step only recovers the strong factor v_{1t}, once we project r_t and g_t onto this factor following Algorithm 2, both residuals should depend only on v_{2t}. Subsequently, applying screening to these residuals again will leave us with only the test assets within the first group of assets, to which applying PCA can recover v_{2t}. In cases in which a linear combination of v_{1t} and v_{2t} are recovered in the first step, after projection the residuals feature a strong factor (again a linear combination of v_{1t} and v_{2t} but orthogonal to the first linear combination), since the second group of $N - N_0$ assets have exposure to it. Therefore, a subsequent screening and PCA suffice to recover this factor.

Similarly, in example (8) the second half of the assets will be eliminated in the first step when using $g_t = v_{1t}$ to screen test assets. The returns for the remaining (first half) assets load on $v_{1t} + v_{2t}$ with common loading matrix β_{11}. Applying PCA to these assets thereby finds $(v_{1t} + v_{2t})/\sqrt{2}$ as the first factor (up to a sign, assuming v_{1t} and v_{2t} share the same variance). Following Algorithm 2, we then obtain residuals from projections of r_t and g_t onto this factor. It is easy to see that the residuals of the second half of r_t and the residuals of g_t both load on a single strong factor $(v_{1t} - v_{2t})/\sqrt{2}$ but the first half of the residuals are purely idiosyncratic. Applying screening plus PCA will successfully recover this factor and hence the span of the factor space.

To formally establish the consistency of this estimator, we introduce an assumption akin to the single-factor case. Specifically, we require that a subset of assets, indexed by I_0, satisfies that all factors are strong within this subset. In other words, $\lambda_{\min}(\beta_{[I_0]}^\intercal \beta_{[I_0]}) \asymp N_0$, where $N_0 = |I_0| \to \infty$. Because the number of factors, p, is finite, such a subset I_0 always exists as long as for each factor we can locate a sufficiently large subset, within which this factor can be extracted consistently.[25] Proposition IA4 establishes that test assets in such a subset suffice to serve as basis assets, building on which a mimicking portfolio can

[25] This assumption is weak in that it does not imply that all factors should have identical strength with respect to the entire cross section of assets in r_t. In addition, different groups of assets could be exposed to different factors.

approximate the risk premia of any observable factor. With this identification assumption, along with moment conditions given in the Internet 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 IA1 to IA8 hold. If* $\log(NT)(N_0^{-1} + T^{-1}) \to 0$*, then for any tuning parameters c and q that satisfy*

$$c \to 0, \quad c^{-1}(\log NT)^{1/2}(q^{-1/2}N^{-1/2} + T^{-1/2}) \to 0, \quad qN/N_0 \to 0, \tag{11}$$

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

The screening step in Algorithm 3 ensures that the selected test assets or their residuals must encompass one strong factor, as they have high correlations with g_t. As the SPCA procedure unfolds, each iteration selects a distinct subset of test assets. By amalgamating all such subsets, we obtain a subset of assets within which all factors are potentially strong, given that the number of factors is finite. However, this procedure may not recover all factors that drive returns. The number of factors that SPCA can recover depends on the interplay between η and β as well as the tuning parameters in a complex manner.[26] Some of the factors that SPCA omits might even be strong! Intuitively, only factors correlated with g_t are guaranteed to be recovered. This is the trade-off that arises for using g_t as a supervisory signal.[27] Nonetheless, missing any factors in the SDF that are uncorrelated with g_t does not affect the consistency of the estimate of the risk premia of g_t. This holds true because such factors do not help price g_t. Of course, this result will need to be strengthened if the objective is to recover the entire SDF, a problem we tackle in Section I.C.

The consistency result in Theorem 1 does not rely on Gaussian error assumptions nor on an assumption that all factors have the same strength with respect to all test assets. The assumption on the relative size of N and T is also quite flexible, in contrast to existing results on factor models in the literature, where N cannot grow at a rate exceeding a certain polynomial function of T.

B.4. Asymptotic Inference on Risk Premia

In this section, we develop the asymptotic distribution of the risk premia estimator from Algorithm 3. Naturally, deriving asymptotic inference requires stronger assumptions than those required for consistency discussed above. To consistently estimate the risk premia of g_t, one only needs to recover factors that are correlated with g_t. Nonetheless, if SPCA misses factors that are in the SDF but are not correlated with g_t, consistency is maintained but inference is

[26] We explicitly characterize this number, denoted by $\tilde{p}$, given in the Internet Appendix following Assumption IA7.

[27] In the context of forecasting, Giglio, Xiu, and Zhang (2023) provide the convergence rate of the estimated factor space, spanned by the factors that are correlated with the variables used for supervision in a similar SPCA procedure.

undermined because the omitted factors may contribute a higher-order error that invalidates the central limit result.

More specifically, the conditions in Theorem 1 do not guarantee that $\widehat{\gamma}_g^{SPCA}$ converges to $\eta\gamma$ at the desirable rate $T^{-1/2}$. The major obstacle lies in the recovery of factors not strongly correlated with g_t, which we can explain with the previous single-factor example.

Recall that we use the sample correlation/covariance between r_t and g_t to screen test assets. Condition (11) has two key requirements. First, it requires that $c \to 0$, allowing the iteration procedure to continue until the selected r_t exhibit asymptotically diminishing correlations with g_t. At the same time, it requires that $c\sqrt{T} \to \infty$ and $c\sqrt{qN} \to \infty$. In other words, c must be sufficiently large to supersede the estimation error in covariance estimates during the screening step, which is of order $T^{-1/2}$,[28] and to dominate error in the construction of residuals in the projection step when multiple steps are involved, an error of order $T^{-1/2} + (qN)^{-1/2}$. However, for any given threshold, say, $c = T^{-1/4}$, if it happens that $\eta \asymp T^{-1/3} < T^{-1/4}$, then screening based on g_t's correlation with r_t will likely not select any assets, leading in turn to the termination of Algorithm 3 and no discovery of factors. Our procedure thereby gives a risk premium estimate of zero, which is certainly consistent, but the estimation error is of order $T^{-1/3} > T^{-1/2}$, so that the usual central limit theorem (CLT) fails. In general, this problem arises due to the possibility of not identifying all factors in the DGP. Once all factors are recovered, the CLT holds regardless of the magnitude of η. To make correct inference, we thus need a stronger assumption that eliminates scenarios like this.

It appears that if $\eta \in \mathbb{R}^{d\times p}$ meets the condition $\lambda_{\min}(\eta^\intercal\eta) \gtrsim 1$, we can rule out the possibility of missing factors. This condition requires that each latent factor maintains a correlation with at least one of the observable variables within g_t. As a result, this implies that d must be greater than or equal to p, meaning that we require g_t to possess at least the same number of variables as the true number of factors. Meanwhile, our algorithm will not select more factors than needed, as we stop the iteration as soon as $c_q^{(k)}$ is sufficiently small (below c), at which point no common factors are left in the residuals of g_t and r_t. We thus obtain the consistency result on the number of factors, which leads in turn to the CLT result on risk premia. Formally, we have the following theorem.

THEOREM 2: *Under the same assumptions as Theorem 1, if we further have* $T^{-1/2}N_0 \to \infty$, *Assumption IA9, and* $\lambda_{\min}(\eta^\intercal\eta) \gtrsim 1$, *then for any tuning parameters c and q in (9) and (10) satisfying*

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

we have that $\widehat{p}$ *defined in Algorithm 3 satisfies* $\mathrm{P}(\widehat{p} = p) \to 1$, *and that the estimator constructed via Algorithm 3 satisfies* $\widehat{\gamma}_g^{SPCA} - \eta\gamma = O_{\mathrm{P}}(T^{-1/2}) +$

[28] Even if g_t is uncorrelated with the test assets, their sample covariances can be as large as $T^{-1/2}$.

$O_{\mathrm{P}}(q^{-1}N^{-1})$. *Furthermore, we obtain a CLT:*

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

where Φ *is given by*

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

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

$$\Pi_{11} = \lim_{T\to\infty} \frac{1}{T}\mathrm{E}\left(\mathrm{vec}(ZV^\intercal)\mathrm{vec}(ZV^\intercal)^\intercal\right),$$

$$\Pi_{12} = \lim_{T\to\infty} \frac{1}{T}\mathrm{E}\left(\mathrm{vec}(ZV^\intercal)\iota_T^\intercal V^\intercal\right),$$

$$\Pi_{22} = \lim_{T\to\infty} \frac{1}{T}\mathrm{E}\left(V\iota_T\iota_T^\intercal V^\intercal\right).$$

In regard to our theoretical findings, several key points merit attention. First, Theorem 2 hinges on the existence of a tuning parameter, q, that must satisfy two conditions: $q^{-1}N^{-1}T^{1/2} \to 0$ and $qN/N_0 \to 0$. A necessary condition for the existence of such a q is thus $T^{1/2}/N_0 \to 0$.

Second, the estimation error of $\widehat{\gamma}_g^{SPCA} - \eta\gamma$ consists of two components. Part of this error stems from the error accumulation at each step of the iteration in Algorithm 3. This accumulated error is compounded in each step k at most by a factor of $\sqrt{|\widehat{I}_k|/\widehat{\lambda}_{(k)}}$, where $\widehat{\lambda}_{(k)} = \left\|\left(\bar{R}_{(k)}\right)_{[\widehat{I}_k]}\right\|^2/T$. Importantly, the assumption that there exists a subset within which factors are pervasive ensures that $\widehat{\lambda}_{(k)} \asymp_{\mathrm{P}} qN = |\widehat{I}_k|$, implying that the accumulated error is magnified by a constant factor with each iteration of SPCA. Ultimately, our proof establishes that this iterative process results in an overall estimation error in risk premia estimates that is of the order $O_{\mathrm{P}}(T^{-1/2} + q^{-1}N^{-1})$. The condition $q^{-1}N^{-1}T^{1/2} \to 0$ thus guarantees that the $O_{\mathrm{P}}(q^{-1}N^{-1})$ term does not influence the asymptotic distribution. The derivation of the error rate for an iterative procedure is nontrivial, constituting our primary contribution to the econometric literature on factor models.

Third, the estimation error of the factor loading has no impact on the asymptotic variance of risk premia, as the expression of Φ demonstrates. This stands in contrast to the classical Fama-MacBeth regression setting, where Shanken's adjustment term (Shanken (1992)) is crucial. This difference is due to the fact that when dealing with a large cross-sectional dimension ($N \to \infty$), this adjustment term vanishes asymptotically.[29] To make inference feasible, we implement the same Newey-West-type estimator for Φ as in section IV.E of Giglio

[29] For a more detailed discussion on this point, please refer to equation (45) of Giglio, Kelly, and Xiu (2022), and the discussion that follows it.

and Xiu (2021), since each component of Φ can be estimated from the outputs of the SPCA algorithm. These estimates are consistent up to some rotation matrices that cancel each other and yield a consistent estimate of Φ.

Fourth, Theorem 2 suggests that, with probability approaching one, we can expect a perfect recovery of the number of factors p. Yet, in any finite sample, perfect recovery remains challenging. Notably, the assumptions made here are considerably less stringent compared to the prevalent factor assumptions found in the literature (see, e.g., Bai (2003) and Bai and Ng (2002)). In these previous studies, inference theory for factor models also relies on the perfect recovery of the count of (strong) factors. We explore the finite-sample behavior of SPCA through simulations in Section II.

Lastly, in the special case in which the returns of test assets are *exclusively* driven by strong factors, SPCA is asymptotically equivalent to PCA, contingent on the appropriate selection of the tuning parameters c and q. Otherwise, SPCA is less efficient—due to either an excessively small choice of q to the extent that the $O_{\mathrm{P}}(q^{-1}N^{-1})$ term plays a dominant role in the estimation error in finite sample (note that PCA corresponds to the case of $q = 1$) or to the fact that some factors (specifically, those uncorrelated with g_t) may not be recovered by SPCA. The former loss of efficiency can be mitigated through careful tuning parameter selection; the latter typically hinges on the unknown values of β and η, which can be resolved with a multivariate target satisfying $\lambda_{\min}(\eta^{\intercal}\eta) \gtrsim 1$.

B.5. Tuning Parameter Selection

While the enhanced robustness to weak factors is an advantage, it comes at the expense of introducing an additional tuning parameter. To employ the SPCA estimator, we need to make choices regarding the two tuning parameters q and c. The parameter q governs the subset size employed in PCA construction, while c determines the stopping rule and consequently the number of factors, p. In contrast, PCA (and other estimators like PLS) essentially require the selection of p only. We establish in Theorem 2 that we can consistently recover p as long as certain conditions are met by q and c.

In theory, the textbook approach to choosing a tuning parameter for parameter estimation revolves around the analytical minimization of the root-mean-squared error (RMSE) of the estimator.[30] This approach effectively balances the trade-off between bias and variance inherent in the estimation. Regrettably, this method necessitates intricate finite-sample analytical calculations of the RMSE, often relying on strong assumptions regarding the DGP. In our context, assumptions of a normal distribution for returns and certain distributional properties and sparsity conditions for betas are likely necessary. Complicating matters further, our iterative SPCA procedure compounds the

[30] Note that in the realm of machine learning, the prevailing approach involves leaning on the prediction RMSE derived from a validation sample, where the actual values of the prediction target are available. This stands in contrast to the estimation problem, where the true values are never known.

difficulty of this analysis, rendering it practically infeasible. In addition, this RMSE-based criterion primarily hinges on statistical considerations, lacking economic relevance.

In lieu of this, we instead use the R^2 of the hedging portfolio for g_t built by SPCA as a criterion, as it is both simpler to apply and justified from an economic perspective. Recall that any estimator of risk premia for a nontradable factor explicitly or implicitly builds a hedging portfolio, or a factor-mimicking portfolio, for g_t, and computes the risk premium as the average excess return of that portfolio. The empirical R^2 obtained by different estimators then has an economic meaning: it reveals the hedging efficacy of the factor mimicking portfolios constructed (explicitly or implicitly) by any risk premia estimators.[31]

Beyond the economic motivation, the R^2 is a useful criterion from a statistical perspective because attaining an optimal R^2 in a validation sample is a sufficient condition for valid selection of tuning parameters, which in turn guarantees consistency of risk premia estimates. See Proposition IA5 for a rigorous statement.

Furthermore, in practice we can consider directly tuning the parameter p instead of c, as it offers greater interpretability, restricts itself to integer values, and is well informed by the scree plot, providing insights into reasonable ranges for p. Regarding the parameter q, opting for larger values makes SPCA's performance resemble that of PCA, thus reducing its robustness against weak factors. Conversely, smaller values of q raise the risk of overfitting, resulting in a high in-sample R^2 but a low out-of-sample R^2. We suggest tuning $\lfloor qN \rfloor$ instead of q because the former can only take integer values, and multiple choices of the latter may lead to the same integer values of the former.

In our applications, we select tuning parameters based on cross-validation (CV) in a training sample that proceeds as follows. We split the sample into three folds. We then use each of the three folds, for validation while the other two are used for training. We select the optimal tuning parameters according to the average time-series R^2 in the validation folds.

C. Recovery of the SDF

The estimation of risk premia for observable factors g_t, studied in Section I.B, is a natural application of the SPCA approach, since g_t can be used to supervise the latent factor extraction. In this section, we explore another application in which observable factors help extract latent factors, namely, a *diagnostic procedure* for observable factor models.

The asset pricing literature has proposed a variety of models that contain a small number of tradable factors g_t: the CAPM, the Fama-French three- or five-factor models, etc. These models are typically evaluated by computing

[31] To be clear, while comparing R^2's provides an insightful depiction of the empirical performance of the hedging portfolios, this cannot be interpreted as proof of the superiority of one estimator over another (which is instead established based on the theoretical properties, like consistency and efficiency, discussed in the previous sections).

the alphas of a universe of test assets and testing whether these alphas are different from zero. This is clearly a valid test for a model, but it gives only limited insights into why the model is (as is often the case) rejected statistically. Specifically, it does not clarify whether the model's failure is due to the presence of true alphas or to the omission of priced factors. Our SPCA procedure helps shed light on this by recovering strong and weak latent factors that drive the cross section of returns, and evaluating whether those factors are indeed spanned by the observable factor model g_t. This helps us ascertain whether the model is lacking certain factors.

A last point relates to the universe of test assets. The asset pricing literature (e.g., Lewellen, Nagel, and Shanken (2010)) has emphasized that using a large cross section of test assets is important for evaluating asset pricing models, as it can improve the power of the tests. There is, however, a downside in expanding the set of test assets: many of the added assets may have little exposure to some factors, introducing a weak factor problem. The ability of SPCA to handle weak factors frees the researcher from worrying about adding assets to the universe, not only in risk premia estimation, but also in diagnostic tests like the one we conduct in this section.

C.1. Consistency of the SDF Estimator

We first prove that, under certain conditions, SPCA consistently recovers the SDF even in the presence of weak factors. Using the outputs of Algorithm 3, we can estimate the SDF as

$$\widehat{m}_t^{SPCA} = 1 - \widehat{\gamma}^\intercal \widehat{v}_t, \quad \text{where } \widehat{v}_1, \ldots, \widehat{v}_T \text{ are the columns of } \widehat{V}. \tag{12}$$

In the Internet Appendix, we prove the following theorem, which not only shows the consistency of the SDF's recovery, but also derives the rate at which the recovery occurs.

THEOREM 3: *Suppose that the assumptions of Theorem 2 hold. In addition, we have Assumption IA10. Then, the estimator (12) satisfies*

$$\frac{1}{T}\sum_{t=1}^{T} |\widehat{m}_t^{SPCA} - m_t|^2 \lesssim_{\mathrm{P}} \frac{1}{T} + \frac{\log N_0}{N_0}. \tag{13}$$

This theorem shows that consistent estimation of the entire SDF time series is possible in terms of average ℓ_2-distance, but under specific conditions. First, for every weak latent factor in v_t, there must be a sufficiently large subset of assets with exposure to that factor. This condition, reflected in the requirement of a large N_0, is also necessary for the consistent estimation of risk premia.

In addition, for each latent factor in v_t, there must be at least one observable factor in g_t that is correlated with that latent factor. This second assumption is needed not only for asymptotic inference on risk premia but also for SDF recovery here. In cases in which g_t does not correlate with a latent factor, that latent factor can potentially be missed by SPCA, thereby hindering SDF recovery.

C.2. Comparison with Alternative Procedures of SDF Estimation

A number of alternative approaches for SDF estimation with latent factors are proposed in the literature, for example, the selection/shrinkage approach by Kozak, Nagel, and Santosh (2020) and the rpPCA by Lettau and Pelger (2020). In what follows, we provide a theoretical comparison of Lasso- and Ridge-based estimators in our general framework where factors can potentially be weak. The Ridge estimator shares the spirit of PCA-based estimators as shown by Giglio and Xiu (2021) and propositions in previous sections. Examining the asymptotic behavior of these two approaches provides useful insights that may guide their applications in practice. Developing the asymptotic guarantee of these estimators is yet another contribution that we make to existing literature on SDF recovery.

Kozak, Nagel, and Santosh (2020) consider an SDF of the form of (3), whereas we represent it as in (2). Prior to the asymptotic analysis of their estimators, we first establish the asymptotic equivalence of these two definitions in our large-N setting.

PROPOSITION 3: *Suppose that test asset returns in r_t follow (1), and Assumption IA10 holds. Then as $N \to \infty$, we have*

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

Proposition 3 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 $\lambda_{\min}(\beta^\intercal\beta) \gtrsim N_0$, we can replace m_t in the left-hand side of (13) by $\widetilde{m}_t$.

Kozak, Nagel, and Santosh (2020) suggest estimating the SDF by solving the optimization problem

$$\widehat{b} = \arg\min_b \left\{(\bar{r} - \widehat{\Sigma}b)^\intercal \widehat{\Sigma}^{-1}(\bar{r} - \widehat{\Sigma}b) + p_\mu(b)\right\}, \tag{14}$$

which is used to estimate the SDF according to

$$\widehat{m}_t = 1 - \widehat{b}^\intercal(r_t - \bar{r}). \tag{15}$$

In the above, $\widehat{\Sigma}$ is the sample covariance matrix of r_t and $p_\mu(b)$ is a penalty term through which economic priors are imposed. Depending on the penalty function, we denote the resulting estimator of m by $\widehat{m}_t^{Ridge}$ or $\widehat{m}_t^{Lasso}$.

The objective function in (14) appears to require the inverse of $\widehat{\Sigma}$, which is not well defined when $N > T$. Instead, we suggest optimizing an equivalent but different form of (14),

$$\widehat{b} = \arg\min_b \left\{b^\intercal\widehat{\Sigma}b - 2b^\intercal\bar{r} + b^\intercal\widehat{\Sigma}b + p_\mu(b)\right\}, \tag{16}$$

which avoids the calculation of $\widehat{\Sigma}^{-1}$.

The following result sheds light on the asymptotic properties of this estimator in the cases $p_\mu(b) = \mu \|b\|_1$ and $p_\mu(b) = \mu \|b\|^2$.[32]

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 IA4 to IA7 and IA10 to IA12 hold, then we have

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

where λ_p is the p^{th} largest eigenvalue of $\beta\Sigma_v\beta^\intercal$. Since $\lambda_p \asymp \lambda_{\min}(\beta^\intercal\beta)$, we can replace m_t in the above equation by $\widetilde{m}_t$.

(b) Suppose that the true SDF satisfies $\mathrm{E}(\widetilde{m}_t^2) \lesssim 1$. With $p_\mu(b) = \mu \|b\|_1$, if Assumptions IA10 and IA11 hold, then we have

$$\frac{1}{T}\sum_{t=1}^{T} |\widehat{m}_t^{Lasso} - \widetilde{m}_t|^2 \lesssim_{\mathrm{P}} \|b\|_1 \sqrt{\frac{\log N}{T}}. \tag{17}$$

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

$$\frac{1}{T}\sum_{t=1}^{T} |\widehat{m}_t^{Lasso} - \widetilde{m}_t|^2 \lesssim_{\mathrm{P}} \|b\|_0 \frac{\log N}{T}. \tag{18}$$

Interestingly, both the Ridge and Lasso approaches deliver consistent estimates of the SDF, albeit under distinct sets of assumptions.

In the case of Ridge, its convergence rate hinges significantly on the strength of the weakest factor. If condition (6) is not met, the consistency of the SDF is compromised. Failure of this condition is a clear symptom of weak factors, precisely the scenario for which our SPCA estimator is designed.

In contrast, the Lasso approach replaces the explicit factor model assumption on r_t with a sparsity assumption on the vector b. This sparsity assumption dictates that the SDF should be represented as a sparse linear combination of the test assets but imposes no explicit assumptions on the DGP of these test assets. This implies that the Lasso estimator remains consistent regardless of the strength of the factors but converges at a rather slow rate, as indicated in (17), which is $\|b\|_1 \sqrt{\log N/T}$. Consequently, it is not as efficient as our SPCA estimator, which leverages the factor structure to achieve faster convergence. Nevertheless, under a much stronger sparsity assumption where $\|b\|_0^2 \log N/T \to 0$, the Lasso estimator can attain a convergence rate comparable to that of the SPCA. This more stringent notion of sparsity essentially asserts that the set of true factors must be part of the test assets. In contrast, our SPCA estimator

[32] We use $\|\|_0$, $\|\|_1$, and $\|\|$ to denote the ℓ_0-, ℓ_1-, and ℓ_2-norms of a vector, respectively.

allows for the presence of idiosyncratic components in any of the test assets, enhancing its practicality in real-world applications.

We can adapt any SDF estimator to obtain an estimator of risk premia because of the relationship $-\text{cov}(m_t, g_t) = \eta\gamma$. This gives the Lasso-based risk premia estimator[33]

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

Furthermore, the consistency of the SDF estimator translates into the consistency of the resulting risk premia estimator.[34] Deriving a valid inference procedure is possible for the Lasso-based risk premia estimator if we employ an additional debiasing step (see Feng, Giglio, and Xiu (2020)), which is beyond the scope of the current paper.

C.3. Diagnosis of SDF Models using Sharpe Ratios

We now discuss the diagnosis of SDF models that consist of tradable factors exclusively. Recall that the projection of the SDF on the space of returns achieves the highest possible Sharpe ratio. Given that the factors recovered by SPCA are themselves portfolios, as long as SPCA recovers the entire SDF, these factors should achieve the maximal Sharpe ratio. We can then diagnose a model g_t by comparing its Sharpe ratio with that achieved by the estimated SDF supervised by g_t. If g_t contains all the factors that drive the SDF, then the maximal Sharpe ratio achieved by factors in g_t should be on par with the Sharpe ratio of the SDF. Otherwise, if g_t achieves a lower Sharpe ratio, this is a sign that g_t is missing some factors; if g_t's Sharpe ratio is higher than that achieved by SPCA, this indicates that g_t has alpha relative to the entire cross section of test asset returns.

For this purpose, it is more convenient to rewrite our SPCA estimator of the SDF given by equation (12) in the form of portfolio returns as in (15), so that we can directly evaluate its Sharpe ratio. In other words, we need an SPCA-based estimate of b in the definition of SDF given by equation (3). Formally, we provide the following algorithm.[35]

[33] The SDF-induced Ridge estimator is numerically equivalent to (IA1), so we do not introduce it again.

[34] By Assumption IA11(1), Cauchy-Schwartz and triangle inequalities, we have

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

[35] The effectiveness of this procedure stems from the fact that the SPCA estimates of $\widehat{V}$ can be written as a rotation of $B^\intercal\bar{R}$. Given that b is invariant to rotations of factors, we can exploit this invariance property to construct a convenient estimator $\widehat{b}$. To elaborate, if we use $B^\intercal\bar{R}$ as the factors, denoted by, $\widetilde{V}$, with their risk premia and covariance denoted by $\widetilde{\gamma}$ and $\widetilde{\Sigma}$, respectively, we can express the SDF as $m_t = 1 - \widehat{\gamma}^\intercal(\widehat{\Sigma}_v)^{-1}\widehat{v}_t = 1 - \widetilde{\gamma}^\intercal(\widetilde{\Sigma}_v)^{-1}\widetilde{v}_t = 1 - \widetilde{\gamma}^\intercal(\widetilde{\Sigma}_v)^{-1}B^\intercal(r_t - \bar{r})$. Consequently,

Algorithm 4: The SPCA-based procedure for estimating SDF loadings is as follows:

Inputs: $\bar{R}_{(1)} := \bar{R}$, $\bar{r}_{(1)} := \bar{r}$, and $\bar{G}_{(1)} := \bar{G}$, a $d \times T$ vector.

S1. For $k = 1, 2, \ldots$ iterate the following steps using $\bar{R}_{(k)}$, $\bar{r}_{(k)}$, and $\bar{G}_{(k)}$ and construct an $N \times p$ matrix B:
   a. Run S1.a of Algorithm 3 to obtain $\widehat{I}_k$
   b. Run S1 to S3 of Algorithm 1 with selected return matrix $\left(\bar{R}_{(k)}\right)_{[\widehat{I}_k]}$ and $\bar{G}_{(k)}$. Construct the k^{th} column of B as $B_{[\widehat{I}_k],k} = \varsigma_{(k)}$ and $B_{[\widehat{I}_k^c],k} = 0$, where $\varsigma_{(k)}$ is the left singular vector of $\left(\bar{R}_{(k)}\right)_{[\widehat{I}_k]}$. Also, obtain $\widehat{V}_{(k)}$ and $\widehat{\eta}_{(k)}$.
   c. Run S1.c of Algorithm 3 to obtain $\widehat{\beta}_{(k)}$.
   d. Run S1.d of Algorithm 3 to obtain $\bar{R}_{(k+1)}$ and $\bar{G}_{(k+1)}$.

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

S2. Estimate the SDF loading b as

$$\widehat{b}^{SPCA} = TB\left(B^\intercal \bar{R}\bar{R}^\intercal B\right)^{-1} B^\intercal \bar{r}. \tag{19}$$

Outputs: $\widehat{b}^{SPCA}$.

Similarly, we can construct estimates of b using PCA and PLS.[36] With $\widehat{b}$ it is convenient to build SDFs (optimal portfolios) and evaluate their Sharpe ratio.

THEOREM 5: *Under the same assumptions as Theorem 1, if Assumption IA10 holds, then the Sharpe ratio of the optimal portfolio constructed by $\widehat{b}^{SPCA}$ in (19) satisfies*

$$\sqrt{\gamma^\intercal \Sigma_v^{-1} \gamma} \geq \lim_{N,T\to\infty} \frac{\widehat{b}^{SPCA\intercal} \mathrm{E}(r_t)}{\sqrt{\widehat{b}^{SPCA\intercal} \Sigma \widehat{b}^{SPCA}}} \geq \sqrt{\gamma^\intercal \eta^\intercal (\eta \Sigma_v \eta^\intercal)^\dagger \eta \gamma}, \tag{20}$$

where † denotes the Moore-Penrose inverse of a matrix.

In the inequality (20), the upper bound corresponds to the optimal Sharpe ratio of the SDF, while the middle term represents the optimal Sharpe ratio achieved by the SPCA estimator. Meanwhile, the lower bound corresponds to the optimal Sharpe ratio achieved by $\eta(v_t + \gamma)$. This lower bound also matches the bound attained by g_t, except for any undiversified idiosyncratic errors that may persist in g_t. These errors would further reduce the Sharpe ratio, but for the sake of our discussion exclusively on observable factor models in the literature, we follow the convention and assume that g_t comprises well-diversified portfolios, so we can ignore this aspect in this section. A sufficient condition

we can deduce that

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

[36] For PCA, the k^{th} column of B can be chosen as the left singular vectors of $\bar{R}$. Equation (19) then yields the standard PCA-based SDF loadings. For PLS, B is a similar weight matrix given by the iterative procedure. We compare these SDF estimators in simulations.

for the upper and lower bounds to be equal is that $\lambda_{\min}(\eta^\intercal \eta) \gtrsim 1$. In this case, the SPCA-based SDF estimator also achieves the optimal Sharpe ratio. This result is not surprising, especially considering the consistency result outlined in Theorem 3.

Theorem 5 serves as the basis for diagnosing SDF models. We do not observe the left side of the equation (the true maximal Sharpe ratio), but we can estimate and compare the middle term (Sharpe ratio obtained by the SPCA-recovered SDF) and the right term (Sharpe ratio of g_t). If we find in the data that the Sharpe ratio from SPCA is higher, then we learn that g_t must be missing a factor. If instead we find that the Sharpe ratio from g_t is higher, then this means that there are factors in g_t that are insufficiently represented in r_t (e.g., if none of the assets in r_t has exposure to those factors): this points to an insufficiently rich set of test assets r_t.[37]

II. Simulations

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

A. Results on Risk Premia

We implement a number of risk premia estimators for comparison, some of which are robust to omitted or weak factors, including PCA and its related estimators (Ridge, PLS, and rpPCA), Lasso, as well as the four-split estimator by Anatolyev and Mikusheva (2022).[38] Both the standard two-pass and four-split methods directly use g_t as if they were the true factors in their regressions. The PCA, rpPCA, Ridge, and Lasso effectively construct the SDF first without knowledge of g_t, and then estimate the risk premia of g_t factor by factor, using the covariance between each factor and the resulting SDF. PLS and SPCA use all variables in g_t to supervise the estimation procedure.

To implement the SPCA estimator, we select the tuning parameters p and $\lfloor qN \rfloor$ by CV using the procedure detailed in Section I.B.5. To ensure a conservative basis for comparison, all methods except SPCA use optimal (albeit infeasible) tuning parameters. Specifically, for PCA, PLS, and rpPCA, we make use of the true number of factors, $p = 4$, even though it is difficult to obtain a consistent estimator of p in the regime of weak factors. The tuning parameter μ of the Ridge estimator is determined via maximum likelihood estimation, with perfect knowledge of $\Sigma = \text{cov}(r_t)$ and $\text{E}(r)$. The second tuning parameter of rpPCA is selected by maximizing the theoretical Sharpe ratio of the estimated SDF,

[37] Of course, it can also be the case that the two Sharpe ratios are the same. In that case, g_t and the latent factor model recovered by SPCA are equivalent in terms of their pricing ability.

[38] The four-split estimator, which does not rely on dimension reduction, selection, or shrinkage techniques, is valid in the presence of weak observable factors and strong omitted factors that are *not* priced. However, it does not have asymptotic guarantees against omitted and priced strong/weak factors, or measurement error in the observed factors.

again, using perfect knowledge of Σ and $\mathrm{E}(r)$. Due to limited sample size, estimating the sample mean and sample covariances in a separate validation sample is rather challenging, which would further deteriorate their performance.

To demonstrate and compare the performance of different estimators, we consider various DGPs of returns and/or the observed variables in g_t. We start with the benchmark scenario a), in which all factors are strong and observed. Specifically, we consider a four-factor DGP as given by equation (1), where the first three factors are calibrated to match the three Fama-French factors (RmRf, SMB, and HML) as in Giglio and Xiu (2021), and the last factor is a potentially weak factor, denoted by V. We calibrate the parameters such that the monthly Sharpe ratio for the optimal portfolio out of these factors is about 0.256. The process generating u_t is modeled as a vector autoregressive process: $u_t = 0.8u_{t-1} + \epsilon_t$, where ϵ_t is drawn from a Gaussian distribution with a diagonal covariance matrix.[39] The standard deviation of u_t is calibrated at 0.04. For comparison, the standard deviations of the four factors are calibrated at 0.04, 0.03, 0.03, and 0.02. The loadings of RmRf are generated independently from $\mathcal{N}(1, 1)$ and the loadings of SMB and HML are generated independently from $\mathcal{N}(0, 1)$. We generate the exposure to the fourth factor V, $\beta_{i,V}$, independently from a Gaussian mixture distribution, with probability a from $\mathcal{N}(0, 1)$ and $1 - a$ from $\mathcal{N}(0, 0.1^2)$. Our calibration suggests that $a = 0.5$ ensures the factor V is sufficiently strong with respect to the cross section of assets in simulations. g_t includes exactly these four factors in the DGP (RmRF, SMB, HML, and V), and we set $\eta = \mathbb{I}_4$, and measurement error is absent.

In scenario (b), we choose $a = 0.1$ so that V is weak in that for almost all test assets their factor loadings to V are tiny: only 10% of the assets have nontrivial exposure to this factor. In scenario (c), the DGP is the same as that of the benchmark case, except we add Gaussian measurement error, z_t, to each of the factors in g_t. In scenario (d), we simulate β for V according to $\beta_{i,V} = -\beta_{i,HML} + e_i$ instead, where e_i's are generated independently from the same mixture Gaussian distribution as above with $a = 0.1$. This nearly results in a rank deficiency in the factor loading matrix due to their correlated exposures. The variable g_t contains all four factors with no measurement error. In scenario (e), we consider the same DGP of returns as in scenario (d), but in g_t we omit the HML factor. Finally, in scenario (f), we further add measurement error to scenario (d).

For each of these six scenarios (including the benchmark), we plot in Figure 1 histograms of the estimated risk premium of V (one entry in g_t) for all estimators.[40] If an estimator is consistent, then the histogram is expected to be centered on the true risk premium of V, whose value is represented by a vertical dashed line. This is indeed the case for SPCA in *all* scenarios. It is also the case for almost all estimators in the benchmark scenario (a), when factors

[39] Although it is conceivable to employ a more complex covariance matrix for u_t, calibrating such a model can be challenging. We therefore simulate u_t's that are cross-sectionally uncorrelated for simplicity.

[40] Panels A to F in Figure 1 correspond to scenarios (a) to (f).

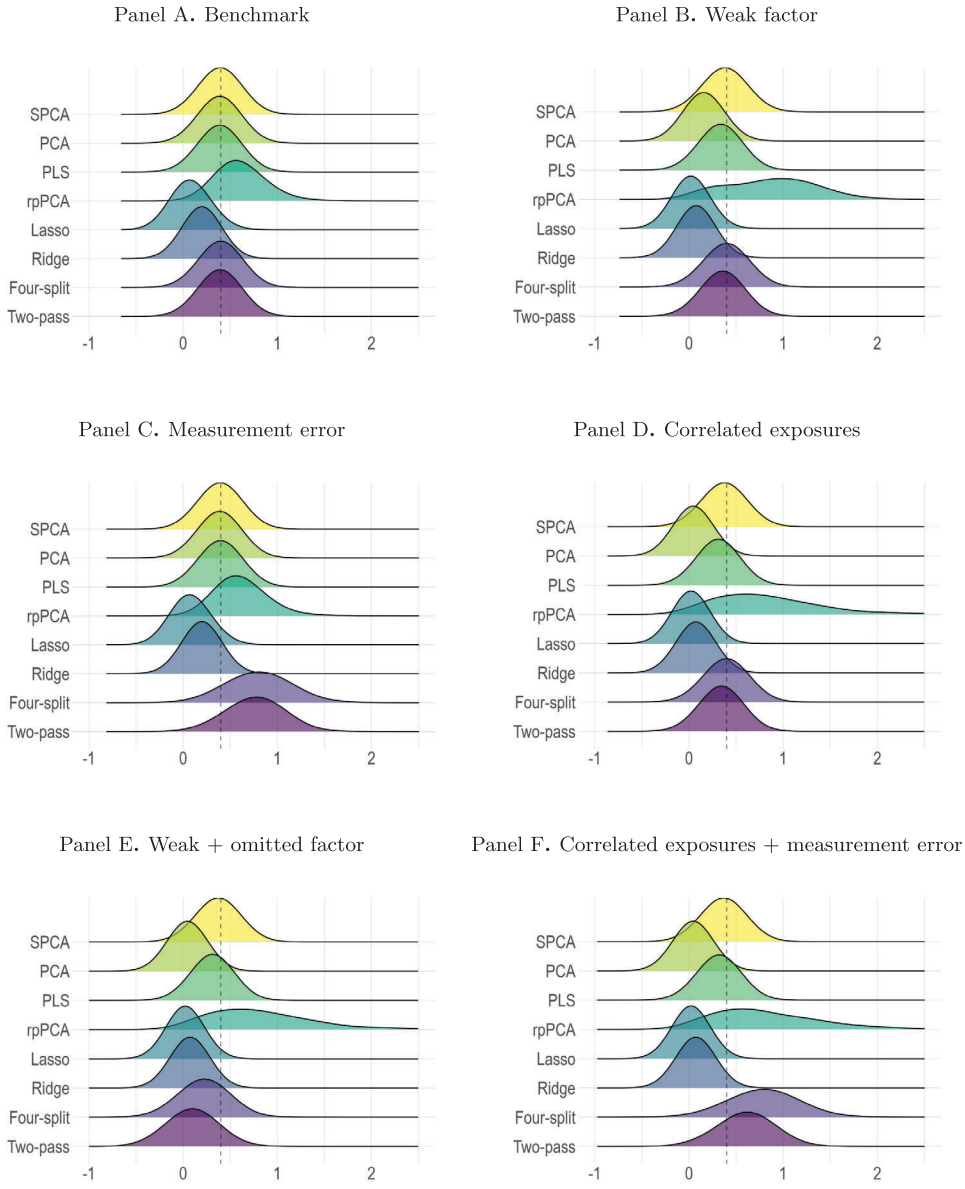

Figure 1. Histogram of risk premium estimates of V. The figure provides histograms of the risk premium estimates in six scenarios for eight estimators we compare, including SPCA, PCA, PLS, rpPCA, Lasso, Ridge, four-split, and the standard two-pass estimator. We simulate the models with $N = 1,000$ and $T = 240$. The number of Monte Carlo repetitions is 1,000. Values reported are percentages. (Color figure can be viewed at wileyonlinelibrary.com)

are strong (except for Lasso and Ridge, which have a large shrinkage bias). This suggests that the latter two estimators are not suitable for *inference* on risk premia. Furthermore, in scenario (b), when weak factors are present, only SPCA and four-split are consistent. The same is true for scenario (d) in which a similar rank-deficiency issue arises. In scenario (c), the four-split estimator becomes inconsistent due to measurement error, and it is also ill-behaved in scenario (e) because the omitted variable, HML, is priced. The PCA and PLS estimators are consistent in scenario (c) but also fail in (e) because they are robust to measurement error but not to omitted weak factors. The standard two-pass estimator is consistent only in the benchmark scenario. Overall, the simulation evidence is in agreement with our theoretical predictions.

We next focus on the last scenario (f), which includes the case of weak factors as well as measurement error. For this case, we report in Table I the bias and RMSE of all estimators for various sample sizes T. The four rows in each panel provide the results of risk premia estimation for RmRf, SMB, HML, and the weak factor V, respectively. We find that our SPCA approach has smaller biases for the weak factors, whereas the remaining estimators have larger biases and RMSEs, which agrees with our theoretical analysis and Figure 1. Notably, PLS ranks second. All estimators perform better in terms of RMSE as T increases.

In the Internet Appendix, we also report a scenario similar to scenario (c) except the last factor is pure noise. In other words, the DGP is driven by the first three factors, but econometricians, lacking knowledge of the true model, include these three factors alongside this pure noise variable in their attempt to estimate risk premia. This scenario closely resembles that extensively discussed by Kan and Zhang (1999) and Kleibergen (2009). For the sake of comparison, PLS and SPCA incorporate this pure noise variable along with the aforementioned three factors into g_t. The histograms corresponding to the risk premium estimates associated with the noise factor suggest that SPCA, PCA, PLS, rpPCA, Lasso, and Ridge remain consistent and cluster around zero. The consistency stems from the fact that none of these methods involves a cross-sectional regression on the estimated beta for the noise factor. In contrast, the four-split and two-pass methods seem to exhibit considerable variances.

We next investigate the finite-sample performance of the inference result developed in Theorem 2. Figure 2 plots histograms of the standardized risk premia estimators using the estimated asymptotic standard errors for SPCA and PCA, respectively, using the DGP in scenario (f) as an example. The histograms of PCA deviate from the standard Gaussian distribution for the two highly correlated factors, V and HML. In contrast, the histograms corresponding to SPCA closely align with the standard Gaussian distribution, showcasing significantly reduced bias for these two factors. A portion of this small bias stems from the population-level approximation as demonstrated in (5) (see also Proposition IA4). This phenomenon thereby likely persists irrespective of the value of T. Finally, we also investigate the statistical power of SPCA in strong and weak cases, and we draw a comparative analysis with PCA. We report these results in the Internet Appendix.

Table I
Simulation Results for Risk Premia Estimators

In this table, we report the bias (column "Bias") and the root-mean-squared error (column "RMSE") of the risk premia estimates using the SPCA, PCA, rpPCA, Lasso, PLS, Ridge, four-split, and the standard two-pass regression approaches. The true data-generating process, given by scenario (f), has four factors, driven by RmRf, SMB, HML, and V, whereas we estimate the risk premia for noisy versions of these four factors. Their true risk premia are provided in column "True." We fix $N = 1,000$ while varying $T = 120$, 240, and 480 in this experiment. All values reported are in basis points.

			SPCA		PCA		rpPCA		PLS	
T	Param	True	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
	RmRf	53.7	0.2	39.2	0.4	38.9	1.8	66.4	0.2	39.1
120	SMB	21.7	−0.0	29.0	0.6	28.4	1.7	65.1	0.4	28.7
	HML	25.4	−6.7	29.3	−38.0	43.9	114.6	205.8	−15.7	30.6
	V	40.0	−6.6	20.9	−37.0	38.9	109.9	195.8	−15.7	22.6
	RmRf	53.7	0.7	29.7	0.6	29.6	1.3	36.4	0.7	29.7
240	SMB	21.7	0.2	20.1	0.6	19.5	1.2	27.8	0.4	19.8
	HML	25.4	−3.3	19.7	−36.3	39.3	64.1	111.9	−8.0	20.1
	V	40.0	−3.4	14.6	−35.5	36.5	63.0	109.0	−8.2	15.4
	RmRf	53.7	−0.1	20.2	0.0	20.2	0.2	20.7	0.0	20.2
480	SMB	21.7	−0.3	14.2	−0.2	14.0	−0.2	14.7	−0.2	14.1
	HML	25.4	−2.6	14.6	−13.4	18.6	22.3	34.6	−4.1	14.5
	V	40.0	−3.1	10.3	−13.7	16.1	20.7	32.7	−4.7	10.6

			Lasso		Ridge		Four-Split		Two-Pass	
T	Param	True	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
	RmRf	53.7	−27.6	37.0	−8.1	32.4	12.4	52.0	11.5	48.1
120	SMB	21.7	−12.6	16.5	−5.1	16.9	4.9	47.2	5.4	41.8
	HML	25.4	−30.6	31.6	−33.4	36.2	12.9	50.5	−6.1	40.1
	V	40.0	−38.3	38.6	−36.0	36.8	32.3	58.6	9.1	32.6
	RmRf	53.7	−31.6	37.4	−4.2	25.8	13.4	40.1	12.4	37.9
240	SMB	21.7	−14.0	16.3	−3.0	13.9	6.1	33.3	5.9	29.5
	HML	25.4	−29.9	30.7	−31.5	33.7	16.2	37.3	2.5	27.4
	V	40.0	−37.6	37.9	−32.7	33.4	38.8	51.2	20.7	32.1
	RmRf	53.7	−18.5	24.7	−1.7	19.1	12.6	29.5	11.9	27.3
480	SMB	21.7	−9.0	11.9	−1.5	12.0	4.3	24.0	4.7	20.9
	HML	25.4	−32.8	33.5	−29.1	30.9	16.6	29.4	8.3	22.1
	V	40.0	−36.8	37.1	−29.5	30.1	38.6	45.6	28.0	33.5

B. Results on SDF Recovery

We next study the finite-sample behavior of the SDF estimators. We compare the performance of SPCA, PCA, rpPCA, Lasso, and Ridge estimators in scenario (f). We report in Table II the MSE of the SDF estimators where the true SDF is defined by equation (3). We also include the tuned number of factors determined by our SPCA approach. In addition, in Table III we report the out-of-sample Sharpe ratios of different methods, given by $\widehat{b}^\top \mathrm{E}(r)/\sqrt{\widehat{b}^\top \Sigma \widehat{b}}$,

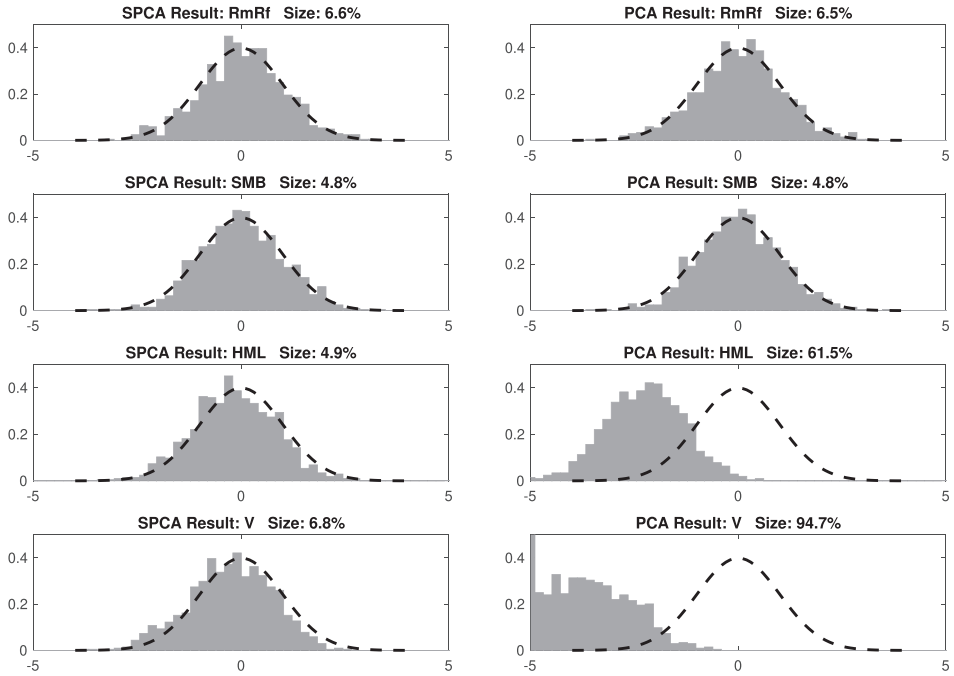

Figure 2. Histogram of the standardized estimates in simulations. The left panels provide the histograms of the standardized SPCA estimates as in Algorithm 3 with asymptotic standard errors given by Theorem 2, while the right panels provide those of the standardized PCA-based risk premia estimates as in Algorithm 1. We simulate the model in scenario (f)) with $N = 1,000$ and $T = 240$. The number of Monte Carlo repetitions is 1,000. These standardized statistics serve as the basis for testing the null hypotheses that the risk premia are equal to their true values. The sizes of these t-tests at the 5% level are reported in the figure subtitles, allowing us to assess the tail behavior of our asymptotic approximations.

where $\mathrm{E}(r)$ and Σ are the true mean and covariance of all test assets and $\widehat{b}$ is the estimated SDF loading using each method. Overall, we find that SPCA outperforms all other methods. PLS ranks second, while rpPCA performs the worst. rpPCA is competitive only in terms of the out-of-sample Sharpe ratio. For risk premia estimation, the underperformance of rpPCA can be attributed to not only its inherent bias but also to its tuning parameters, which are oriented primarily toward maximizing the Sharpe ratio. Last but not least, the tuning parameter $\widehat{p}$ is found to be in close proximity to the true value four.

Finally, in Figure 3 we investigate the pattern of out-of-sample Sharpe ratios for various models g_t. The setting resembles scenario (f), except that we consider different models g_t to examine the role of g_t in supervising the procedure. We report Sharpe ratios as a function of number of factors $\widehat{p}$ used in the PCA and SPCA procedure. For SPCA, we select $\lfloor qN \rfloor$ via CV using the time-series R^2 for each given $\widehat{p}$. The sample size T is fixed at 240. The theoretical value of the optimal Sharpe ratio is 0.256, as shown in Table III, though in finite sample the maximum Sharpe ratio achieved by SPCA is around 0.226.

Table II
Simulation Results for SDF Estimators

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 the SPCA, PCA, rpPCA, PLS, Lasso, and Ridge approaches. The reported MSEs are the sample average over 1,000 Monte Carlo repetitions and their standard deviations are reported in brackets. We also report the mean and standard deviation of the estimated number of factors $\widehat{p}$ using the SPCA approach. The true data-generating process, given by scenario (f), has four factors, driven by RmRf, SMB, HML, and a weak factor V, whereas we estimate the SDF using a vector of factor proxies, g_t, that includes noisy versions of the four factors. We compare three scenarios with $T = 120$, 240, and 480, where $N = 1,000$ is fixed.

	SPCA		PCA	rpPCA	PLS	Lasso	Ridge
T	$\widehat{p}$	MSE	MSE	MSE	MSE	MSE	MSE
120	4.186	0.044	0.074	9.200	0.050	0.056	0.054
	(0.389)	(0.030)	(0.026)	(11.332)	(0.026)	(0.010)	(0.013)
240	4.011	0.021	0.058	1.901	0.025	0.055	0.045
	(0.104)	(0.014)	(0.013)	(3.313)	(0.013)	(0.009)	(0.010)
480	4.004	0.010	0.018	0.087	0.012	0.050	0.036
	(0.063)	(0.007)	(0.007)	(0.083)	(0.007)	(0.007)	(0.008)

Table III
Simulation Results for Out-of-Sample Sharpe Ratios of Optimal Portfolios

In this table, we report the mean and standard deviation of the out-of-sample Sharpe ratios for various optimal portfolios constructed by the SPCA, PCA, rpPCA, PLS, Lasso, and Ridge approaches. The true data-generating process, given by scenario (f), has four factors, driven by RmRf, SMB, HML, and a weak factor V, whereas we estimate the SDF using a vector of factor proxies, g_t, that includes noisy versions of the four factors. The reported Sharpe ratios are the sample average over 1,000 Monte Carlo repetitions and their standard errors are reported in the brackets. The column "'Theoretical Value" provides the benchmark Sharpe ratio calculated by $b^\intercal \mathrm{E}(r)/\sqrt{b'\Sigma b}$ using true parameter values. We compare three scenarios with $T = 120$, 240, and 480, where $N = 1,000$ is fixed.

T	SPCA	PCA	rpPCA	PLS	Lasso	Ridge	Theoretical Value
120	0.193	0.084	0.134	0.164	0.113	0.109	0.256
	(0.049)	(0.046)	(0.035)	(0.051)	(0.024)	(0.046)	
240	0.226	0.110	0.192	0.214	0.122	0.137	0.256
	(0.026)	(0.036)	(0.033)	(0.031)	(0.019)	(0.032)	
480	0.241	0.227	0.242	0.238	0.127	0.162	0.256
	(0.012)	(0.019)	(0.008)	(0.015)	(0.021)	(0.019)	

We consider four cases of $g_t = \eta v_t + z_t$. In case (a), we set $\eta = \mathbb{I}_4$, so all factors are included in g_t to supervise the procedure. In case (b), only the factor V and HML are included in g_t. In case (c), we fix $\eta = (1, 0, 0, 0)$, that is, g_t only includes the (strong) market factor. Finally, in case (d), we let $\eta = \gamma^\intercal \Sigma_v^{-1}$, so that g_t is a noisy measure of the SDF. In light of Theorem 5, SPCA should achieve the maximal out-of-sample Sharpe ratio in cases (a) and (d), provided

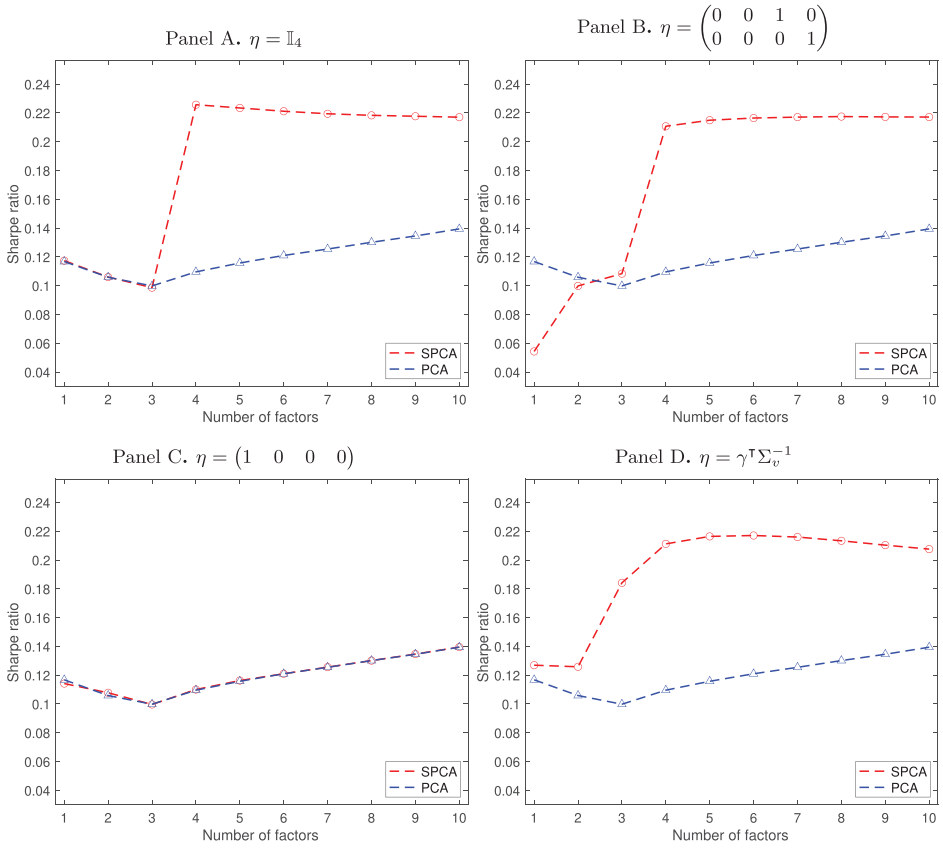

Figure 3. Out-of-sample Sharpe ratio patterns with different models of g_t. Each panel reports out-of-sample Sharpe ratios for PCA (blue) and SPCA (red) as a function of number of factors, p, for a specific model of $g_t = \eta v_t + z_t$. (Color figure can be viewed at wileyonlinelibrary.com)

appropriate tuning parameters. Figure 3 confirms this result.[41] In case (a), SPCA reaches its highest Sharpe ratio out-of-sample at $\widehat{p} = 4$, and the Sharpe ratio declines slightly as $\widehat{p}$ increases beyond four, since these additional factors only add noise. Case (d) exhibits a similar pattern. In contrast, the PCA approach cannot achieve the maximal Sharpe ratio, even as $\widehat{p}$ increases to 10, because PCA cannot recover the weak factor, which contributes to the SDF. In case (b), SPCA is supervised by two factors with highly correlated loadings, so it can recover the part of the SDF spanned by the weak factors. With a large enough p, we force the procedure of SPCA to continue, and it will also extract the strong factors and achieve the maximal Sharpe ratio. In case (c), however, SPCA and PCA provide similar results—neither achieves the optimum—because g_t only includes the market factor, which does not help SPCA recover the missing weak factor.

[41] Panels A to D in Figure 3 correspond to cases (a) to (d).

III. Empirical Analysis

In this section, we apply SPCA to estimate the risk premia of a variety of observable factors, and to diagnose observable factor models.

A. Data

Our main data set is the Chen and Zimmermann (2022) data, which include a large number of equity portfolios sorted by characteristics. Specifically, we employ the April 2021 release of the data. For each characteristic considered, Chen and Zimmermann (2022) construct a variable number of portfolios (as many as are used in the original papers that introduced the anomaly in the literature, typically 2, 5, or 10). Not all test assets are available for the entire time period; for our analysis, we study the period 1976m3 to 2020m12, for which 901 test portfolios are available without missing values. To these sorted portfolios, we add 49 industry portfolios from Ken French's website. All of our results are at the monthly frequency.[42]

We also consider an alternative data set, proposed by Hou, Xue, and Zhang (2020), that for the same period includes 1,672 portfolios sorted by characteristics without missing values. Hou, Xue, and Zhang (2020) classify their portfolios into six groups: momentum, value, investment, profitability, intangibles, and frictions. These two data sets are similar and yield comparable results. Rather than producing two versions of each result using the two data sets, we choose Chen and Zimmermann (2022) to be our main data set and report the robustness of the main results using the Hou, Xue, and Zhang (2020) data (see Section III.B.6). What both data sets have in common is that they capture a wide universe of anomaly equity portfolios discovered in the last four decades of asset pricing research.

We consider both tradable and nontradable factors in our analysis, focusing on the best-known ones from the literature. The tradable factors are: the market (in excess of the risk-free rate); size (SMB); value (HML); profitability (RMW); investment (CMA); momentum (MOM); betting-against-beta (BAB, from Frazzini and Pedersen (2014)); and quality-minus-junk (QMJ, from Asness, Frazzini, and Pedersen (2013)). The nontradable factors are: the liquidity factor from Pástor and Stambaugh (2003); the intermediary capital factor from He, Kelly, and Manela (2017); AR(1) innovations in industrial production growth (IP); VAR(1) innovations in the first three principal components of 279 macrofinance variables from Ludvigson and Ng (2010); AR(1) innovations in the three uncertainty indexes of Jurado, Ludvigson, and Ng (2015), representing financial uncertainty, macroeconomic uncertainty, and real uncertainty; AR(1) innovations in the term spread, the credit spread, and the unemployment rate; AR(1) innovations in two sentiment indexes, one from

[42] The theory is silent on the correct frequency of the data to study. Here, we follow the literature and focus on the monthly frequency. We leave for future work a more comprehensive study and comparison across frequencies.

Huang et al. (2015) and one from Baker and Wurgler (2006); oil price growth AR(1) innovations; and consumption growth AR(1) innovations.[43]

B. Estimation of Risk Premia using SPCA

In this section, we estimate the risk premia of a variety of tradable and nontradable factors. We begin by discussing details of the implementation of the estimator.

B.1. Choice of Tuning Parameters and Implementation Details

To apply SPCA to the estimation of the risk premia and to evaluate its out-of-sample performance, we split the sample period into two equal-sized subsamples. The first half of the sample (training period) is used to choose the tuning parameters and produce the risk premium estimate. The second half of the sample (evaluation period) is used to evaluate the out-of-sample performance of the estimator and the choice of tuning parameter.

For ease of presentation, we select only one tuning parameter q (or, equivalently, the number of assets selected $\lfloor qN \rfloor$) for each plausible choice of p (the number of factors) in our analysis. This approach reduces the number of tuning parameters to only one, and also conveniently serves as a robustness check with respect to the number of factors.

To determine reasonable candidates for p, we examine the factor structure of the panel of test asset returns. Figure 4 provides the scree plot of the log of the first 25 eigenvalues. There appear to be at least three strong factors. In addition, it appears that factors 4 to 11 might also be relevant, but *weak*. Motivated by the scree plot, in the empirical study below we highlight results for p equal to 3, 5, 7, and 11, therefore showing robustness of our results to a wide range of model dimensions.

To choose the tuning parameter q, we adopt the same R^2 criterion as in simulations to evaluate the estimator's out-of-sample performance, namely, the hedging ability of the portfolio built by SPCA for g_t. Guided by this *statistical* justification, in our empirical work we choose q by threefold CV(100 runs) within the training sample, maximizing the hedging R^2 for g_t. Section V of the Internet Appendix describes in detail the steps for the CV. After we tune q, we use it to compute the SPCA risk premium estimate for g_t.

[43] The market factor, SMB, HML, RMW, CMA, and MOM are from Ken French's website. BAB and QMJ are from AQR's website. The liquidity factor is from Lubos Pastor's website. The intermediary capital factor is from Asaf Manela's website. The macro principal components and the uncertainty indexes are from Sydney Ludvigson's website. Industrial production, the credit spread, the unemployment rate, the term spread, and oil price are from Fred-MD. The Huang et al. (2015) sentiment index is from Huang's webpage. The Baker and Wurgler (2006) sentiment index is from Wurgler's website. The consumption factor was built from national income and product accounts (NIPA) data using the methodology of Schorfheide, Song, and Yaron (2018).

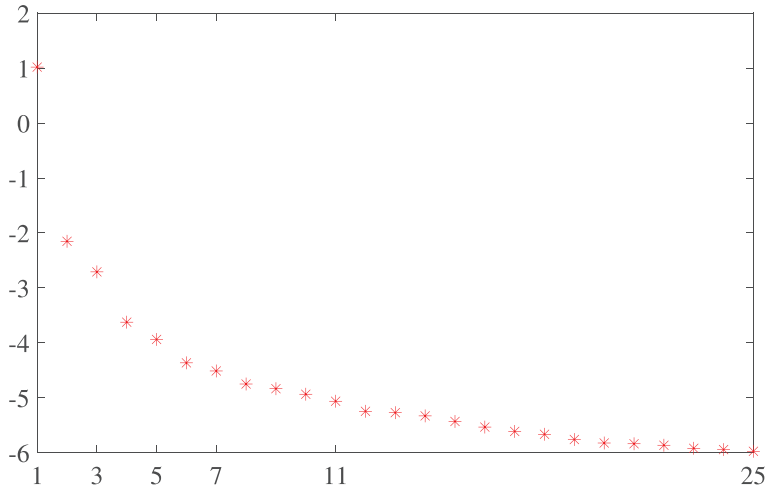

Figure 4. Logarithm of the first 25 eigenvalues in the Chen-Zimmerman data. The figure plots the logarithm of the first 25 eigenvalues of the data, obtained from Chen and Zimmermann (2022) plus 49 industry portfolios, covering the period 1976 to 2020. (Color figure can be viewed at wileyonlinelibrary.com)

B.2. Results: Estimation of Risk Premia and Out-of-Sample Evaluation

We report the main empirical results in Table IV and Figures 5 and 6. Each row of Table IV corresponds to one factor; the first eight are tradable, the rest are nontradable. For tradable factors, the first two columns show the average excess return of the factor separately for the training sample and the evaluation sample; these numbers correspond to model-free estimates of the risk premia of tradable factors, and can be directly compared with the model-based estimate obtained from SPCA.

The next columns of the table show the SPCA results in four groups of columns, corresponding to the number of latent factors $p = 3,\ 5,\ 7$, and 11. For each choice of p, we report the risk-premium estimate (obtained in the training sample, in basis points [bps] per month), the number of assets selected by SPCA (determined by q), and the out-of-sample R^2 obtained in the evaluation period. These estimates are obtained factor by factor, that is, in each case, g_t contains one factor and the asset selection is driven by that factor only. In the last two columns of the table, we repeat the exercise (with $p = 11$) but estimate all risk premia simultaneously: g_t contains all the factors and the selection of the assets is based on all of them simultaneously (so that $d \geq p$ as opposed to $d = 1$). In theory, both approaches are consistent. In practice, estimating risk premia factor by factor has the advantage that the latent factors zoom in immediately on the assets relevant for each factor. In contrast, the joint estimation is required for the CLT of Section I.B.4.

Consider first the market portfolio (first row of the table), a strong factor in this data set. The average return of the market in the training sample is

Table IV

Risk Premia Estimates

In this table, we report the estimation results for tradable and nontradable factors using SPCA. The first two columns report the average excess returns for tradable factors, both in the training sample (first half of the sample period) and in the evaluation sample (second half of the sample period). The remaining columns report, for different values of the number of factors p, the risk premia estimates (in basis points per month, computed in the training period), the number of assets selected by SPCA (governed by the parameter q), and the out-of-sample R^2 of the implied hedging portfolio. The last two columns report risk premia estimates and standard errors including all factors in g_t simultaneously, with $p = 11$. The sample comprises the Chen and Zimmermann (2022) test portfolios plus 49 industry portfolios over the period 1976 to 2020.

	Avg. Ret.	Avg. Ret.	3 Latent Factors			5 Latent Factors			7 Latent Factors			11 Latent Factors			Joint Estim, 11 factors	
	(train.)	(eval.)	RP	# Assets	R^2	RP	# Assets	R^2	RP	# Assets	R^2	RP	# Assets	R^2	RP	Stderr
Market	74	62	68	100	0.98	70	100	0.98	72	100	0.99	74	100	0.99	73	26
HML	39	−7	50	100	0.70	37	100	0.79	39	150	0.78	44	250	0.79	54	18
SMB	12	25	15	100	0.82	5	100	0.85	10	100	0.85	10	100	0.85	7	18
RMW	37	28	−8	100	−0.18	40	100	0.56	33	100	0.61	27	150	0.66	23	9
CMA	26	19	36	250	0.41	40	100	0.55	27	200	0.55	31	350	0.53	34	11
Momentum	91	30	67	100	0.79	86	100	0.87	102	100	0.87	101	100	0.88	96	23
BAB	126	56	112	100	0.43	120	100	0.38	112	150	0.35	128	150	0.45	93	20
QMJ	41	39	−9	100	0.43	28	100	0.81	31	100	0.80	36	150	0.78	20	10
Liquidity			70	550	0.01	85	650	0.02	83	700	0.04	95	900	0.03	105	25
Intermed. Cap.			112	100	0.59	101	100	0.56	121	150	0.55	116	350	0.52	109	41
IP growth			−4	950	−0.01	−4	950	−0.02	−5	950	−0.03	−2	950	0.00	−2	3
LN 1			225	550	−0.28	202	650	−0.19	150	700	−0.11	54	950	−0.12	35	146
LN 2			−70	950	−0.05	−79	950	−0.12	−24	950	−0.16	−29	950	−0.17	−53	82
LN 3			96	400	0.03	86	650	0.06	16	700	0.06	−21	850	0.05	−92	78

(*Continued*)

Table IV—*Continued*

	Avg. Ret.	Avg. Ret.	3 Latent Factors			5 Latent Factors			7 Latent Factors			11 Latent Factors			Joint Estim, 11 factors	
	(train.)	(eval.)	RP	# Assets	R^2	RP	# Assets	R^2	RP	# Assets	R^2	RP	# Assets	R^2	RP	Stderr
Consumption			2	950	−0.01	3	950	0.00	3	950	−0.01	2	950	−0.01	2	2
Fin. Unc.			−61	350	−0.08	−48	750	0.00	−40	850	0.09	−41	950	0.10	−46	17
Real Unc.			−6	950	0.05	−7	950	0.04	−9	950	0.04	−11	950	0.06	−17	12
Macro Unc.			−7	950	0.08	−10	950	0.08	−10	950	0.08	−16	950	0.09	−19	10
Term			229	950	−0.11	81	950	−0.36	−57	950	−0.54	262	950	−0.59	384	372
Credit			41	950	−0.03	62	950	−0.03	41	950	−0.02	−43	950	−0.03	−32	77
Unempl.			65	950	0.00	109	950	−0.01	112	950	−0.01	110	950	0.00	45	108
Sentiment HJTZ			−24	950	0.01	−27	950	−0.03	−18	950	−0.06	−40	950	−0.07	−34	76
Sentiment BW			57	950	0.00	64	950	0.00	50	950	0.01	16	950	−0.02	44	71
Oil			−37	950	−0.05	−62	950	−0.02	−42	950	−0.03	−20	950	−0.02	−9	41

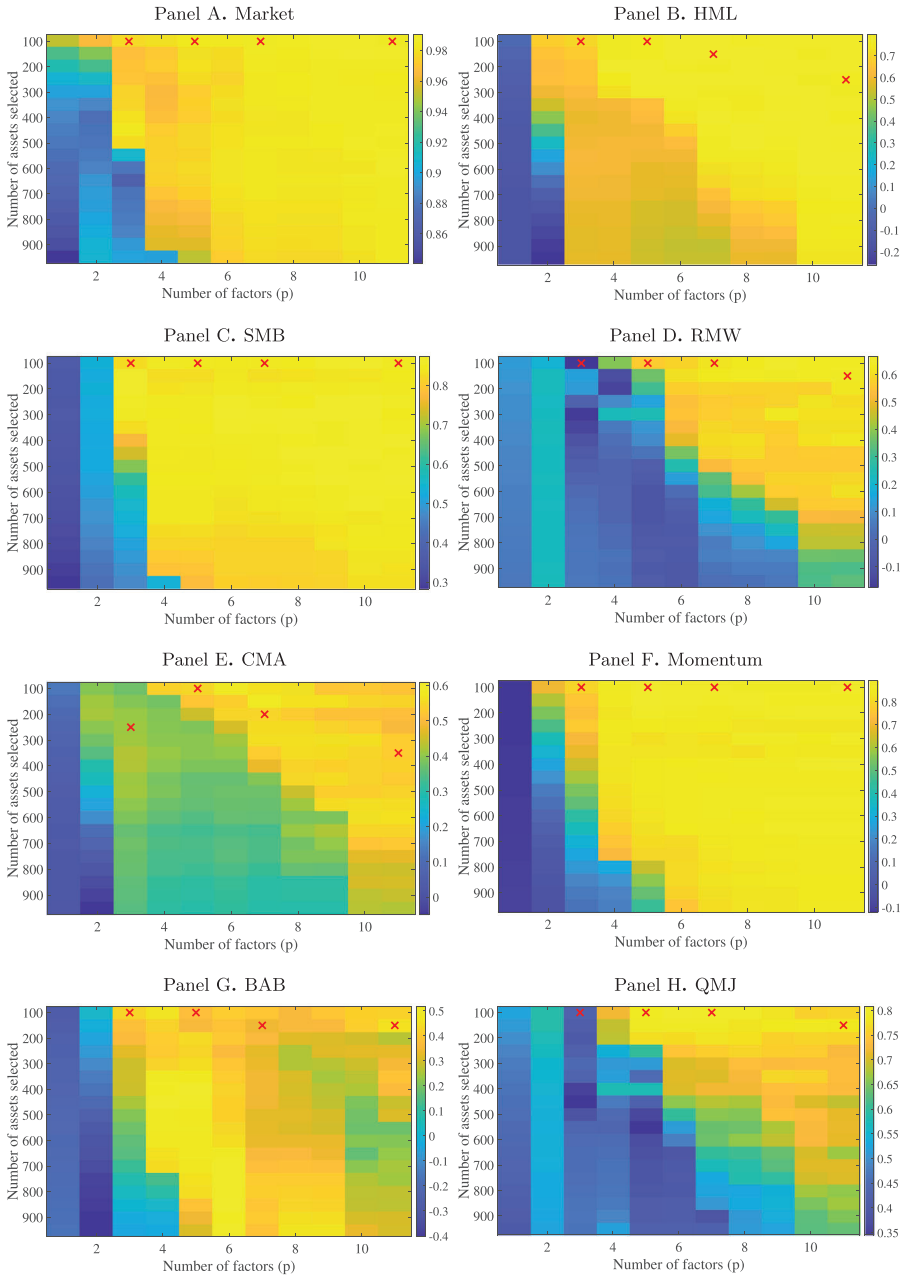

Figure 5. Out-of-sample R^2 heatmaps, tradable factors. Each panel reports the out-of-sample R^2 heatmap for a different factor. The x-axis reports p. The y-axis reports the number of assets selected, governed by q. The colors in the heatmap correspond to the out-of-sample R^2 of the SPCA-implied hedging portfolio for the factor g_t; this R^2 is computed entirely in the evaluation period. The red marks are the points chosen by CV within the training sample. (Color figure can be viewed at wileyonlinelibrary.com)

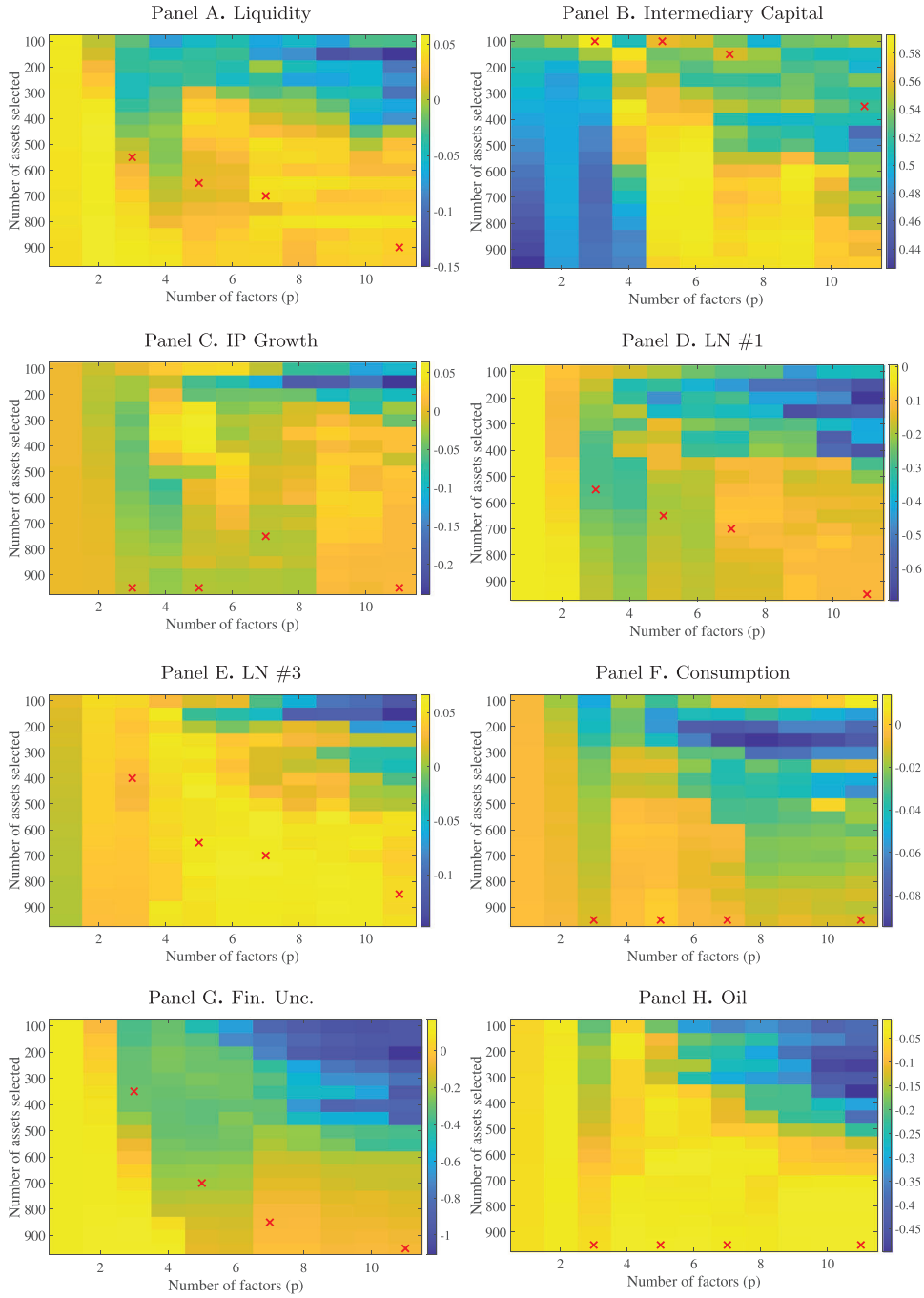

Figure 6. Out-of-sample R^2 heatmaps, nontradable factors. Same as Figure 5, but for a subset of nontradable factors. (Color figure can be viewed at wileyonlinelibrary.com)

74 bps per month, and 62 bps in the evaluation period. The SPCA estimates of the market risk premium, for the four chosen values of p, are 68, 70, 72, and 74 bps per month, respectively, all close to the average excess return. To obtain these estimates, SPCA estimates the latent factors picking, in each iteration, 100 assets out of the total of 950. Finally, the portfolio that SPCA builds to hedge the market achieves a very high out-of-sample R^2 above 0.98 for all p.

To better understand the performance of the estimator and the tuning parameter choice, we can examine the heatmap in Figure 5, Panel A, which focuses on the market factor. In the heatmap, the x-axis corresponds to the number of factors p and the y-axis to the number of test assets selected by SPCA (determined in turn by q); for each combination of p and q, the heatmap shows the out-of-sample R^2 of the hedging portfolio built by SPCA.

Panel A shows that for all combinations of p and q, out-of-sample R^2's are overall very high for the market portfolio, above 85%. However, there appears to be a subset of the parameter space in which hedging performance is especially good: combinations with high p and low q. The red marks in the heatmap correspond to the values of q chosen by CV in the training sample (one for each value of p considered in the table: 3, 5, 7, and 11). Ideally, the values of q chosen by CV in the training sample would yield a hedging portfolio that performs well out of sample, that is, the marks should lie in areas in the heatmap with high out-of-sample R^2's. This is indeed the case, as the figure shows, indicating good out-of-sample performance of the tuning parameter selection procedure.

Consider now another tradable factor, CMA, in the fifth row of Table IV. As for the market, the estimated risk premium for CMA is not significantly different from the average excess return of the factor. The number of assets selected by SPCA ranges between 100 and 350, and the out-of-sample R^2 is above 50%, indicating that the hedge portfolio built by our latent factor model is able to capture the majority of the variation on CMA out of sample.[44]

The heatmap of the out-of-sample R^2 for the hedging portfolio of this factor is provided in Panel E of Figure 5. The figure shows that for the case of CMA, different combinations of p and q yield very different out-of-sample hedging performance, with R^2's ranging from above 50% to below zero. Ideally, if the tuning parameter were chosen properly, we would see that the hedging portfolio also does well out of sample. The red marks in the figure show that this is indeed the case, especially for $p = 5$ and above.

These heatmaps also allow us to compare the results with the PCA-based estimator of Giglio and Xiu (2021). This is because the last row of the heatmap corresponds to the case $q = 1$, that is, all assets are used to estimate the factors, so PCA corresponds to a particular choice for the tuning parameter. Looking

[44] Given that the universe of test assets includes portfolios sorted by the same characteristics used to construct the tradable factors like CMA, one may wonder why an out-of-sample R^2 of 100% is not always obtained for tradable factors. The reason is that SPCA attempts to build a hedging portfolio for the target g_t with factors that must also explain covariation among the universe of test assets. An advantage of our approach is that the hedging portfolio is able to avoid fitting the "measurement error" component in g_t, which, as discussed above, can be thought of as nondiversified idiosyncratic error for tradable factors, or more literally measurement error for nontradables.

across the various panels of Figure 5, it is clear that while for some factors (such as the market) similar R^2 can be obtained by PCA and SPCA, for other factors (such as CMA and RMW) the out-of-sample R^2's obtained by SPCA are substantially higher than those by PCA. This is not surprising given that the scree plot shows the presence of several weak factors in the data.

One additional advantage of SPCA that is clearly visible in the heatmaps is that SPCA often manages to achieve the same (or better) R^2 than PCA, while estimating a much smaller number of factors. For example, consider the momentum factor in Panel F. The last row of the heatmap shows that extracting factors via PCA achieves an R^2 above 70% only when at least six factors are included; SPCA gets there with three factors. The reason is intuitive: SPCA focuses on the test assets most informative about g_t, and therefore can zoom in quickly on the most relevant latent factors.

For nontradable factors, we cannot compare the risk premium estimate from SPCA with the average excess return; beyond relying on the theory and simulations, we can look at the out-of-sample R^2 for suggestive evidence about the empirical performance of the estimator. Note that it is well known in the literature that it is difficult to hedge nontradable factors, like consumption or IP growth, in equity markets. We show, however, that SPCA gives a hedging portfolio that successfully hedges at least part of the variation in many nontradable factors.

Consider first the liquidity factor of Pástor and Stambaugh (2003), in row 9 of Table IV and Panel A of Figure 6. The out-of-sample R^2 achieved by SPCA is above zero (up to 4%), and the estimated risk premium appears to be high (between 70 and 95 bps per month). Panel A of Figure 6 shows how strongly this R^2 depends on p and q. Among all combinations of parameters, a large fraction delivers a negative out-of-sample R^2. This result highlights how difficult it is to hedge this factor (like most macro factors) using equity markets, and indicates again the relatively good performance of SPCA as tuned in the training sample.

The remainder of the table and of the two figures shows the results for all of the other factors (due to space considerations, the heatmaps report only a subset of the factors, while the table reports them all). A few interesting patterns emerge. First, for tradable factors, SPCA gives risk premia estimates that are always close to the model-free estimates obtained from average excess returns: the two are never statistically different at the 5% level (with the only exception of QMJ with $p = 3$). Second, confirming the previous literature, nontradable factors are much harder to hedge than tradable factors; in fact, for several factors—like the first two JLN macro factors—we do not get positive R^2 at all. For those factors, there is so little exposure in equity returns that SPCA cannot build a proper hedging portfolio. However, SPCA is able to hedge out of sample at least part of the variation of many factors, like the third LN factor, the three uncertainty measures, the liquidity factor, and the intermediary capital factor (for which it achieves an R^2 above 50%). Third, the risk premia estimated by SPCA—for those factors for which SPCA can actually hedge some of the variation—make economic sense. For example, the

liquidity and intermediary factors command significantly positive risk premia, whereas the three uncertainty measures command negative risk premia.

B.3. Asset Selection

To better understand how SPCA estimates risk premia, we can study which assets are selected when extracting the latent factors. Table V shows, for four representative factors (two tradables—Momentum and RMW, and two nontradables—liquidity and intermediary capital), the top 10 test assets (by absolute value of correlation) selected at each step. The names of the portfolios follow Chen and Zimmermann (2022), with the numbers indicating the quintile or decile of the characteristic.

Consider momentum in the first set of rows. To extract the first factor, SPCA selects the assets with the highest correlation with the momentum factor. The table indicates that the highest correlation, at 0.44, is with IntMom09, an intermediate momentum portfolio. The other assets with high correlation are all momentum-related, not surprisingly. In the next columns, the table shows the assets selected at the second iteration of SPCA, after orthogonalizing g_t and the test assets to the first factor. Interestingly, the correlations among these residuals are even higher, up to 0.79 for a different momentum sort (Mom12mOffSeason, momentum without the seasonal component). This suggests that the first factor captures some of the asset variation that is not exclusively specific to momentum (e.g., part of the market factor), which the projection step of SPCA removes.

The remainder of the table shows which assets are selected at the different iterations for RMW, Liquidity, and Intermediary Capital. For RMW (a profitability factor), the selected assets are often based on accounting measures, like asset growth, accruals, leverage, and operating profits. For liquidity, portfolios sorted by payout yield and beta seem to play an important role in hedging the risk. Finally, for intermediary capital, the portfolios selected by SPCA relate to idiosyncratic volatility, liquidity, as well as two industry portfolios (not surprisingly, banking and financials).

The selection of particularly informative assets is the central mechanism through which SPCA addresses the issue of weak factors. It is also responsible for the parsimony of SPCA to the number of factors used, since SPCA zooms in on the most informative assets.

B.4. Strength of the Factors

We next report the strength of the factors extracted by SPCA at each step. To make the results comparable across iterations of SPCA, and between SPCA and PCA, we compute the strength of a latent factor as the eigenvalue of the factor normalized by the number of assets used to extract it. Figure 7 reports, in each panel, the log-normalized eigenvalues for the factors extracted from PCA (dashed line) and for the factors extracted by SPCA, grouped across panels for the various targets (since the factors extracted by SPCA are different

Table V

Assets Selected by SPCA

For each factor (one per panel), the table shows the top 10 assets selected by SPCA in extracting the latent factors. Assets are sorted by absolute value of the correlation. For each of the top three factors, the table reports the names of the portfolios selected and the absolute value of the correlation with g_t. Naming convention for the portfolios follows Chen and Zimmermann (2022).

	Factor #1		Factor #2		Factor #3	
	Asset	| Corr |	Asset	| Corr |	Asset	| Corr |
Mom	IntMom09	0.44	Mom12mOffSeason02	0.79	Mom12m08	0.64
	IntMom10	0.4	Mom12mOffSeason03	0.76	BMdec05	0.63
	MomVol10	0.37	Size01	0.74	IntMom03	0.63
	MomVol09	0.36	ResidualMomentum01	0.73	SP05	0.62
	IntMom08	0.36	ResidualMomentum02	0.73	ShareIss5Y05	0.62
	Mom12m10	0.36	NumEarnIncrease01	0.72	BookLeverage02	0.62
	FirmAgeMom05	0.35	ShareIss5Y01	0.7	cfp05	0.61
	Mom12mOffSeason10	0.34	MomVol03	0.69	BMdec04	0.61
	Mom12mOffSeason09	0.33	CompEquIss01	0.68	ShareIss1Y05	0.6
	Mom12m09	0.33	Mom12m03	0.68	LRreversal04	0.6
RMW	Industry:Gold	0.27	OperProf05	0.54	OperProfRD01	0.53
	MomOffSeason10	0.27	OperProfRD09	0.53	RoE01	0.47
	AccrualsBM02	0.27	CBOperProf09	0.5	GP01	0.45
	DelEqu05	0.27	RoE05	0.49	CBOperProf02	0.45
	LRreversal05	0.27	CBOperProf10	0.49	DolVol01	0.44
	roaq01	0.26	Leverage02	0.49	OperProfRD02	0.44
	AssetGrowth10	0.26	OperProfRD08	0.49	CBOperProf01	0.43
	DolVol05	0.25	realestate03	0.49	OperProf01	0.41
	ChEQ05	0.25	GP05	0.49	RoE02	0.4
	Price05	0.25	GP04	0.48	VolMkt02	0.4

(*Continued*)

Table V—*Continued*

	Factor #1		Factor #2		Factor #3	
	Asset	| Corr |	Asset	| Corr |	Asset	| Corr |
Liq.	InvGrowth06	0.47	InvGrowth06	0.28	InvGrowth06	0.3
	NetPayoutYield07	0.47	BetaFP09	0.26	DolVol01	0.27
	PayoutYield05	0.46	EntMult06	0.25	XFIN08	0.26
	PayoutYield07	0.46	NetPayoutYield07	0.24	MeanRankRevGrowth01	0.26
	BetaFP03	0.46	PayoutYield07	0.24	BetaFP03	0.25
	DelLTI02	0.46	PayoutYield05	0.24	ShortInterest01	0.25
	IntanBM03	0.46	cfp04	0.23	BetaFP09	0.24
	EntMult06	0.46	BetaFP10	0.23	EntMult06	0.24
	VolMkt04	0.46	XFIN08	0.23	PayoutYield07	0.24
	PayoutYield06	0.46	ShortInterest01	0.22	ChEQ04	0.23
Interm.	Industry:Banks	0.9	Industry:banks	0.76	Industry:banks	0.7
	Industry:Fin	0.84	Industry:Fin	0.56	Industry:Fin	0.47
	IntMom05	0.8	DelEqu02	0.46	DebtIssuance02	0.38
	EquityDuration04	0.8	grcapx3y02	0.44	NOA10	0.36
	IdioVolAHT05	0.8	OScore02	0.43	ChAssetTurnover04	0.35
	IdioVol3F05	0.79	GrLTNOA10	0.43	HerfAsset05	0.35
	MaxRet08	0.79	ChAssetTurnover04	0.43	ShareRepurchase01	0.35
	Illiquidity01	0.79	IntMom05	0.43	HerfBE05	0.35
	IdioRisk05	0.79	IdioVolAHT05	0.42	DelEqu05	0.32
	CBOperProf03	0.78	Tax01	0.42	Beta05	0.32

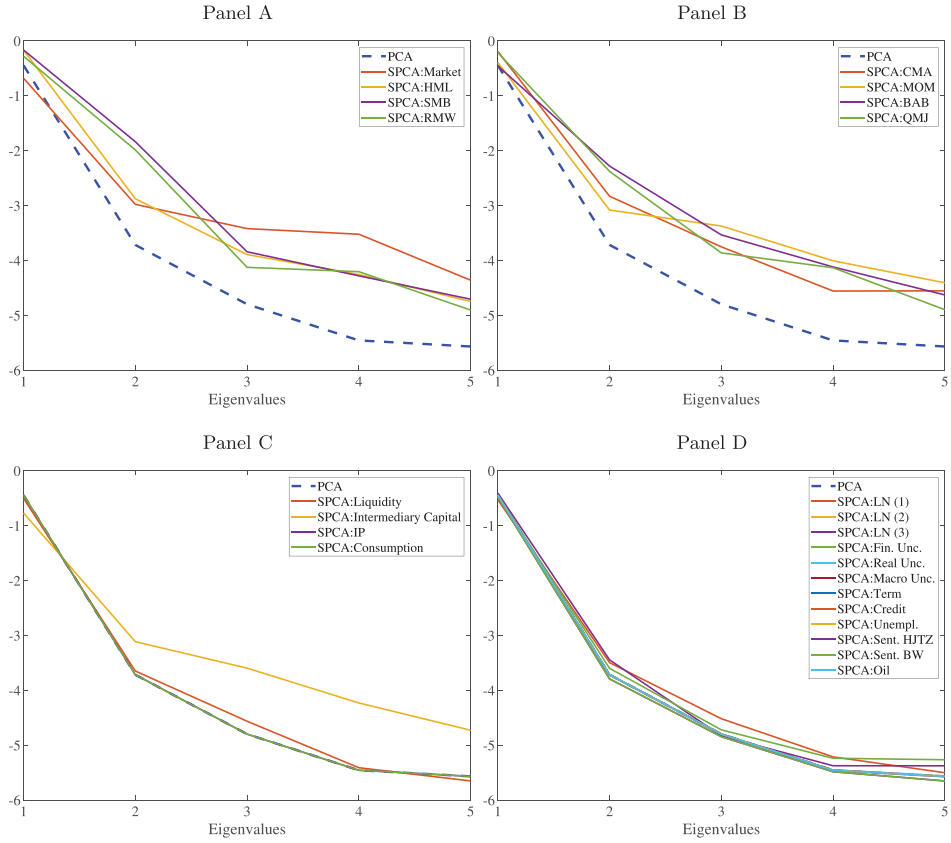

Figure 7. Strength of the latent factors. Each panel of the figure shows the log eigenvalues extracted by PCA from the universe of all assets in the training sample, as well as the log eigenvalues extracted by SPCA at each iteration (for the first five factors), for the tuning parameter selected by CV. All eigenvalues are normalized by the number of assets used, which is a measure of the strength of the factor that is directly comparable. Panels A and B study two groups of tradable factors, Panel C a selection of the nontradables, and Panel D the remaining nontradables. (Color figure can be viewed at wileyonlinelibrary.com)

for different targets g_t): Panels A and B show the factors extracted when the targets are tradable factors, Panel C focuses on a subset of nontradables, and Panel D corresponds to the remaining nontradables. The figure shows eigenvalues corresponding to the first five factors.

As expected, the log eigenvalues for PCA decrease as lower-variance factors are extracted. This is mostly (but not always) the case for SPCA, though we see a large difference across factors. For some factors (like most nontradables, which, as discussed above, are mostly noise factors), SPCA chooses a large number of assets, so the results look very similar to PCA (e.g., see Panel D). For factors where SPCA chooses a small number of assets (e.g., intermediary capital and many tradables) we see that the strength of the extracted factor is higher than with PCA. This effect is strongest for the first eigenvalue (the

log scale hides it somewhat), but is there for subsequent factors as well. In general, it appears that SPCA does indeed strengthen the factor extracted from the cross section, compared to PCA—especially so when fewer assets are selected.[45]

B.5. SPCA and the Universe of Test Assets

The fact that SPCA estimates the latent factors using the most informative assets also makes it particularly robust to the universe of test assets used in the estimation. We explore this in more detail here by considering three factors, value, momentum, and profitability, for which we can easily identify test assets informative about them. Specifically, we consider (for this section only) the data set from Hou, Xue, and Zhang (2020), which, as discussed in Section III.A, collects test portfolios by characteristics in six groups, among which one is labeled "value vs. growth," one "momentum," and one "profitability." We can then ask how SPCA performs in estimating the value risk premium if we exclude the value and growth sorts from the universe. Similarly, we can examine how it performs in estimating the momentum and profitability risk premia if momentum and profitability test assets, respectively, are removed. Once the sorted portfolios are removed, the corresponding factors naturally become weaker. However, we expect SPCA to continue to perform well, provided sufficient exposure to the factor is present in the remaining test assets. In contrast, we expect PCA's performance to deteriorate more sharply.

We again look at the performance of SPCA through the lens of the hedging portfolio R^2. Figure 8 provides the out-of-sample time-series R^2 heatmap for the three factors: value, momentum, and profitability. On the left of each row, we can see the R^2 obtained using all assets from the Hou, Xue, and Zhang (2020) data set; on the right we can see the results excluding the test assets corresponding to each factor. By looking at the last row of each heatmap, which corresponds to the PCA estimate with no selection, it is clear that the hedging performance of a portfolio built via PCA deteriorates significantly when the most informative assets are removed. Consider, for example, the case $p = 9$. For value, the PCA hedging portfolio's out-of-sample R^2 decreases from 64% to 47%, as value and growth assets are removed; SPCA's R^2 decreases by substantially less, from 74% to 62%. In the case of momentum, the R^2 decreases from 76% to 48% for PCA, but from only 86% to 77% for SPCA. Finally, for profitability, the R^2 decreases from 41% to 14% for PCA, but from only 71% to 60% for SPCA. In all cases, the SPCA portfolio hedging ability deteriorates little when the relative sorts are removed and the factor is made weaker, whereas the deterioration is much larger for PCA.

[45] One caveat is that once the main factors are extracted, and mostly noise is left in the cross section, noise itself could lead to higher normalized eigenvalues. This is why the criterion for tuning the parameter q of SPCA is the out-of-sample R^2 of the hedging portfolio, and not this measure of factor strength.

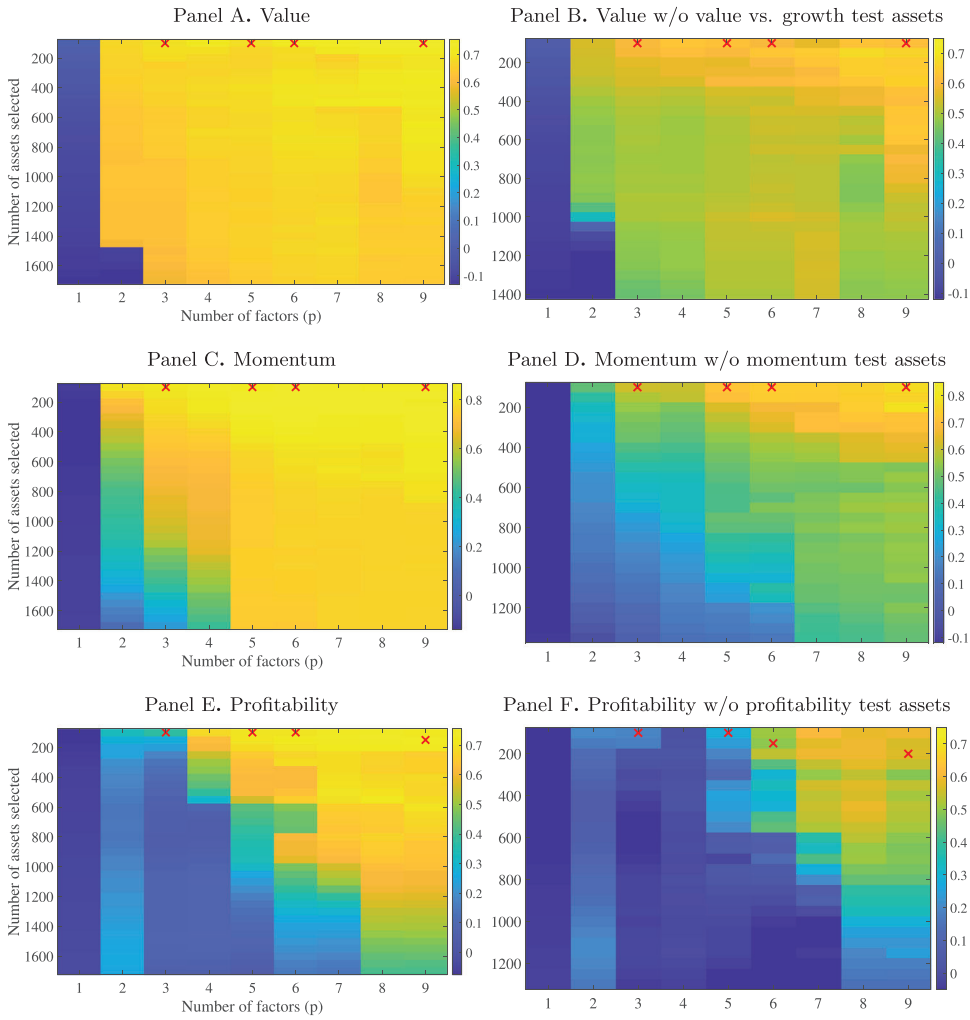

Figure 8. Varying the universe of test assets. For value, momentum, and RMW (profitability), the figure provides out-of-sample R^2 heatmaps when all the test assets from Hou, Xue, and Zhang (2020) are used in the estimation (left), and when value portfolios, momentum portfolios, or profitability portfolios, respectively, are excluded (right). (Color figure can be viewed at wileyonlinelibrary.com)

To summarize, these empirical results mirror the simulations in Section II, which show that SPCA performs well even when the factor of interest is weak in the universe of test assets considered.

B.6. Robustness

We conclude by reporting in Table VI a version of Table IV obtained using the Hou, Xue, and Zhang (2020) data set instead of the Chen and Zimmermann

Table VI

Risk Premia Estimates, Hou, Xue, and Zhang (2020) Data

Same as Figure IV, but using the characteristic-sorted portfolios from Hou, Xue, and Zhang (2020) instead of those from Chen and Zimmermann (2022).

	Avg. Ret.	Avg. Ret.	Three Latent Factors			Five Latent Factors			Six Latent Factors			Nine Latent Factors			Joint Estim, Nine factors	
	(train.)	(eval.)	RP	# Assets	R^2	RP	# Assets	R^2	RP	# Assets	R^2	RP	# Assets	R^2	RP	Stderr
Market	74	62	72	100	0.98	74	100	0.99	74	100	0.99	74	100	0.99	71	26
HML	39	−7	22	100	0.69	20	100	0.69	16	100	0.71	18	100	0.74	24	16
SMB	12	25	−12	100	0.74	−13	100	0.72	−16	100	0.74	−15	100	0.79	−9	18
RMW	37	28	12	100	0.38	26	100	0.71	25	100	0.71	36	150	0.76	18	9
CMA	26	19	8	100	0.70	11	100	0.65	12	100	0.66	4	300	0.65	8	11
Momentum	91	30	68	100	0.81	60	100	0.86	57	100	0.85	55	100	0.86	58	20
BAB	126	56	47	100	0.08	37	100	0.07	31	100	0.08	27	200	0.04	37	12
QMJ	41	39	−3	150	0.68	15	100	0.82	15	100	0.83	17	150	0.80	6	10
Liquidity			28	1,700	0.05	35	1,700	0.06	42	1,700	0.06	35	1,700	0.06	37	18
Intermed. Cap.			107	100	0.49	98	100	0.46	91	100	0.51	63	100	0.43	90	37
IP growth			−2	1,700	0.01	−4	1,700	−0.02	−3	1,700	−0.01	−3	1,700	−0.01	−1	2
LN 1			171	1,200	−0.11	215	1,650	−0.15	151	1,700	−0.11	169	1,700	−0.08	110	93
LN 2			−19	1,700	−0.08	−17	1,700	−0.08	−13	1,700	−0.08	−4	1,700	−0.12	−83	55
LN 3			16	1,000	0.03	69	1,550	0.04	26	1,700	0.02	15	1,700	0.03	11	62

(*Continued*)

Table VI—*Continued*

	Avg. Ret.	Avg. Ret.	Three Latent Factors			Five Latent Factors			Six Latent Factors			Nine Latent Factors			Joint Estim, Nine factors	
	(train.)	(eval.)	RP	# Assets	R^2	RP	# Assets	R^2	RP	# Assets	R^2	RP	# Assets	R^2	RP	Stderr
Consumption			0	1,700	0.00	0	1,700	0.00	1	1,700	0.00	0	1,700	0.00	1	1
Fin. Unc.			−5	1,600	0.18	−15	1,700	0.16	−15	1,700	0.16	−9	1,700	0.14	−18	11
Real Unc.			−4	1,700	0.02	−5	1,700	0.02	−8	1,700	0.02	−6	1,700	−0.03	−11	6
Macro Unc.			−2	1,700	0.05	−4	1,700	0.05	−6	1,700	0.05	−4	1,700	0.04	−9	5
Term			−11	1,700	−0.11	24	1,700	−0.10	77	1,700	−0.08	24	1,700	−0.14	261	240
Credit			24	1,700	−0.02	29	1,700	−0.03	0	1,700	−0.06	8	1,700	−0.09	16	40
Unempl.			42	1,700	0.00	116	1,700	−0.01	112	1,700	−0.01	101	1,700	−0.02	89	61
Sentiment HJTZ			−44	1,700	0.01	−39	1,700	0.01	−22	1,700	0.02	−20	1,700	0.02	−39	44
Sentiment BW			−29	1,700	0.03	−31	1,700	0.02	−21	1,700	0.02	−25	1,700	−0.01	9	43
Oil			−8	1,600	−0.03	−39	1,500	0.00	−35	1,600	0.00	−26	1,550	−0.01	−47	28

(2022) data. The results are qualitatively similar to those obtained using the Chen and Zimmermann (2022) data, and, with a few exceptions, not statistically different. This confirms that the results do not depend on using one particular universe of test assets. That said, the results also suggest some differences between these two universes of test assets, which our analysis in the next section sheds some light on.

C. Diagnosing Factor Models via SPCA

In the previous section, we apply SPCA to the estimation of risk premia. In this section, we illustrate the use of SPCA to diagnose missing factors in observable factor models, applying the theory developed in Section I.C. Recall that given an observable factor model g_t, and a set of test assets r_t, we can use SPCA to recover the latent factor SDF (using g_t to supervise the extraction of weak factors). If we find that the Sharpe ratio achieved by the latent factors recovered by SPCA is higher than that achieved by g_t, we can conclude that the factor model using g_t to span the SDF is missing some factor. This is not just a test of whether g_t explains r_t; rather, it sheds light on *why* a model may be rejected in the data.

We consider five observable factor models: the CAPM, the Fama-French three-factor model (FF3), the Fama-French five-factor model (FF5), and finally two richer models: one with the FF5 factors plus momentum, and one with FF5 plus momentum, BAB, and QMJ. We diagnose these models using both the CZ and the HXZ data sets.

We divide the sample into two parts as in Section I.B, and use the first half for training (and selection of the tuning parameter) and the second half for out-of-sample evaluation. Maximal Sharpe ratios achieved using the factors in g_t and using the factors from SPCA are calculated out of sample.

Figure 9 reports the results. Each panel corresponds to a different model. The x-axis in each figure corresponds to the number of factors extracted via SPCA. The y-axis corresponds to the out-of-sample Sharpe ratio. The Sharpe ratio achieved by g_t is represented by a dashed solid line, which naturally does not depend on the number of latent factors. In each graph, we overlay the SPCA results with the HXZ and CZ data, respectively, using different markers (blue triangles for HXZ and red circles for CZ). Not surprisingly, the out-of-sample Sharpe ratios are somewhat noisy; we also plot fitted lines using raw estimates to help visualize the trend.

Consider Panel A, in which g_t is just the market. The market in our out-of-sample period achieves a Sharpe ratio of 0.46 (dashed line). SPCA factors extracted using g_t achieve significantly higher Sharpe ratios, in both the HXZ and CZ data. The Sharpe ratio increases with the number of factors, indicating that the CAPM misses several sources of risk. Results for the FF3 and FF5 models (Panels B and C) are similar: for both, once the number of factors is sufficiently large, SPCA produces a Sharpe ratio that is superior to either model. When momentum is included (Panel D), the model performs as well as SPCA in the HXZ data. This result suggests that relative to the universe of

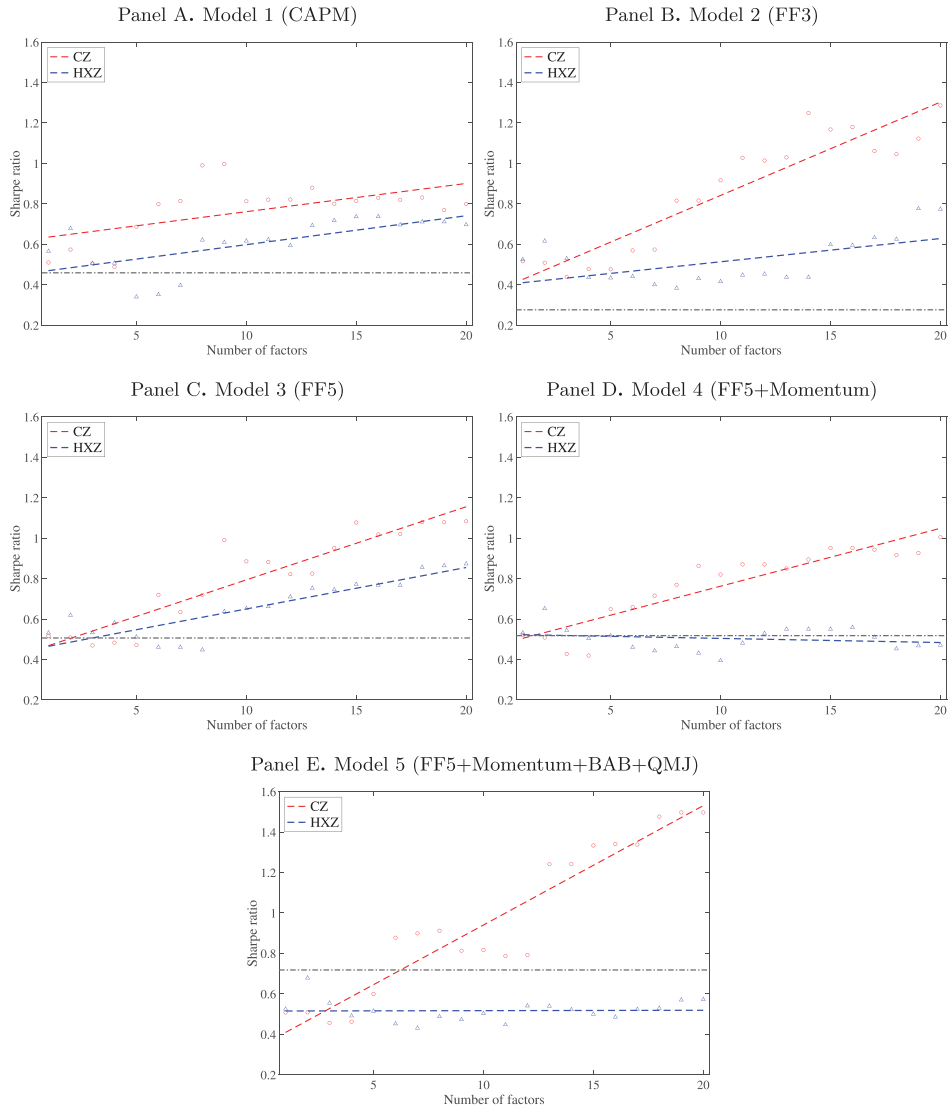

Figure 9. Out-of-sample Sharpe ratios of different factor models. Each panel reports the out-of-sample Sharpe ratio of an observable factor model g_t (dashed line), together with the out-of-sample Sharpe ratio obtained from the factors recovered using SPCA, using both the HXZ data (triangles) and CZ (circles). The x-axis corresponds to the number of factors used in SPCA (p). (Color figure can be viewed at wileyonlinelibrary.com)

test assets in the HXZ data set, this model (FF5+momentum) appears to be spanned by almost all sources of risk driving this data set (but not so in the CZ data set).

As more observable factors are added to these models (Panel E that includes BAB and QMJ), we should expect the Sharpe ratio of the model to increase, as

long as more latent factors adds risk factors and not noise. We indeed find that this is the case. Overall, this suggests that these richer models do a better job in capturing the fundamental sources of risk in these data set, although some amount of misspecification remains visible in the CZ data set.

The differences between the results using the HXZ and CZ data sets also emphasize the importance of the choice of test assets. Ideally, to have as powerful tests as possible, we would want to have a large and varied universe of test assets. The number of assets in a data set, however, is not a perfect proxy for the richness of the universe in terms of risk exposures. In fact, as we have noted in this paper, a universe with large N but low exposures to some factors can introduce a weak factor problem. Here, we see another case in which the size of the data set does not necessarily translate into richer risk exposure: HXZ contains more assets than CZ, yet the results in this section show that using the test assets, r_t, from CZ, SPCA diagnoses additional factors compared to those diagnosed using HXZ (this could reflect, e.g., a different construction of the portfolios in the different data sets, or different selection of characteristics).

Overall, these results illustrate that the ability of SPCA to recover weak latent factors can prove useful as a diagnostic tool for observable factor models, and again highlights the importance of the choice of test assets in performing asset pricing tests.

IV. 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 characteristics gives us hope that the returns of these portfolios can help us uncover and identify the pricing of various dimensions of risk, including those that are not well captured by standard cross sections. On the other hand, the large dimensionality goes hand in hand with the weak factor issue: a factor may well be captured by *some* assets within the large cross section, but if most assets do not have exposure to that factor, it will be weak and inference will be incorrect.

Traditional methodologies take the cross section of assets as given. In this paper, we present a new methodology, SPCA, that instead actively selects assets to estimate risk premia of factors of interest, whether they are strong or weak, and at the same time addresses the issue of potentially omitted factors, again regardless of whether they are strong or weak. In addition, SPCA can exploit its ability to recover weak latent factors to help diagnose omitted factors in observable factor models. The paper confirms the good performance of SPCA for both of these tasks in a variety of simulations, and illustrates the application of the methodology in various empirical contexts.

While the road to a full understanding of risk and risk premia in financial markets is still long, we believe that systematically tackling weak factors in empirical asset pricing is an important step forward that opens the door to the

study of factors that, while important to investors, may be not pervasive in either the standard cross sections or the recently developed large universes of test assets.

Two pressing issues on the debates related to the factor zoo are economic interpretability and the overwhelming amount of degrees of freedom in empirical asset pricing research. The central challenge we address in this paper is to evaluate factors motivated by economic theories. Our proposal eliminates two critical degrees of freedom altogether from this exercise: the choice of control factors when estimating risk premia of economically motivated factors, and the choice of test assets used for estimation and testing. Our study thereby contributes to a promising agenda developing a fusion of asset pricing theory and machine learning. It does so by using the factor structure as a main theoretical foundation, and applying to it tools and results from machine learning, in order to exploit these statistical advances while maintaining economic interpretability.

Initial submission: June 24, 2022; Accepted: December 21, 2023
Editors: Stefan Nagel, Philip Bond, Amit Seru, and Wei Xiong

REFERENCES

Ahn, Dong-Hyun, Jennifer Conrad, and Robert F. Dittmar, 2009, Basis assets, *Review of Financial Studies* 22, 5133–5174.

Anatolyev, Stanislav, and Anna Mikusheva, 2022, Factor models with many assets: strong factors, weak factors, and the two-pass procedure, *Journal of Econometrics* 229, 103–126.

Ang, Andrew, Robert Hodrick, Yuhang Xing, and Xiaoyan Zhang, 2006, The cross-section of volatility and expected returns, *Journal of Finance* 61, 259–299.

Asness, Clifford S., Andrea Frazzini, and Lasse Heje Pedersen, 2013, Quality minus junk, Technical report, AQR.

Bai, Jushan, 2003, Inferential theory for factor models of large dimensions, *Econometrica* 71, 135–171.

Bai, Jushan, and Serena Ng, 2002, Determining the number of factors in approximate factor models, *Econometrica* 70, 191–221.

Bai, Jushan, and Serena Ng, 2008, Forecasting economic time series using targeted predictors, *Journal of Econometrics* 146, 304–317.

Bai, Jushan, and Serena Ng, 2023, Approximate factor models with weaker loadings, *Journal of Econometrics* 235, 1893–1916.

Bailey, Natalia, George Kapetanios, and M. Hashem Pesaran, 2021, Measurement of factor strength: Theory and practice, *Journal of Applied Econometrics* 36, 587–613.

Bair, Eric, Trevor Hastie, Debashis Paul, and Robert Tibshirani, 2006, Prediction by supervised principal components, *Journal of the American Statistical Association* 101, 119–137.

Bair, Eric, and Robert Tibshirani, 2004, Semi-supervised methods to predict patient survival from gene expression data, *PLoS Biology* 2, 511–522.

Baker, Malcolm, and Jeffrey Wurgler, 2006, Investor sentiment and the cross-section of stock returns, *Journal of Finance* 61, 1645–1680.

Bryzgalova, Svetlana, 2015, Spurious factors in linear asset pricing models, Technical report, Stanford University.

Bryzgalova, Svetlana, Jiantao Huang, and Christian Julliard, 2023, Bayesian solutions for the factor zoo: We just ran two quadrillion models, *Journal of Finance* 78, 487–557.

Bryzgalova, Svetlana, Markus Pelger, and Jason Zhu, 2020, Forest through the trees: Building cross-sections of asset returns, Technical report, London School of Business and Stanford University.

Chamberlain, Gary, and Michael Rothschild, 1983, Arbitrage, factor structure, and mean-variance analysis on large asset markets, *Econometrica* 51, 1281–1304.

Chen, Andrew Y., and Tom Zimmermann, 2022, Open source cross-sectional asset pricing, *Critical Finance Review* 11, 207–264.

Fama, Eugene F., and Kenneth R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.

Fan, Jianqing, and Yingying Fan, 2008, High dimensional classification using features annealed independence rules, *Annals of Statistics* 36, 2605–2637.

Fan, Jianqing, Yuan Ke, and Yuan Liao, 2021, Augmented factor models with applications to validating market risk factors and forecasting bond risk premia, *Journal of Econometrics* 222, 269–294.

Fan, Jianqing, and Yuan Liao, 2022, Learning latent factors from diversified projections and its applications to over-estimated and weak factors, *Journal of the American Statistical Association* 117, 909–924.

Fan, Jianqing, and Jinchi Lv, 2008, Sure independence screening for ultrahigh dimensional feature space, *Journal of the Royal Statistical Society, B* 70, 849–911.

Feng, Guanhao, Stefano Giglio, and Dacheng Xiu, 2020, Taming the factor zoo: A test of new factors, *Journal of Finance* 75, 1327–1370.

Frazzini, Andrea, and Lasse Heje Pedersen, 2014, Betting against beta, *Journal of Financial Economics* 111, 1–25.

Freyaldenhoven, Simon, 2019, A generalized factor model with local factors, Technical report, FRB of Philadelphia Working paper.

Gagliardini, Patrick, Elisa Ossola, and Olivier Scaillet, 2016, Time-varying risk premium in large cross-sectional equity datasets, *Econometrica* 84, 985–1046.

Giglio, Stefano, Bryan Kelly, and Dacheng Xiu, 2022, Factor models, machine learning, and asset pricing, *Annual Review of Financial Economics* 14, 337–368.

Giglio, Stefano, and Dacheng Xiu, 2021, Asset pricing with omitted factors, *Journal of Political Economy* 129, 1947–1990.

Giglio, Stefano, Dacheng Xiu, and Dake Zhang, 2023, Prediction when factors are weak, Technical report, University of Chicago.

Gospodinov, Nikolay, Raymond Kan, and Cesare Robotti, 2013, Chi-squared tests for evaluation and comparison of asset pricing models, *Journal of Econometrics* 173, 108–125.

Gospodinov, Nikolay, Raymond Kan, and Cesare Robotti, 2014, Misspecification-robust inference in linear asset-pricing models with irrelevant risk factors, *Review of Financial Studies* 27, 2139–2170.

Harvey, Campbell R., Yan Liu, and Heqing Zhu, 2016, …and the cross-section of expected returns, *Review of Financial Studies* 29, 5–68.

He, Zhiguo, Bryan Kelly, and Asaf Manela, 2017, Intermediary asset pricing: New evidence from many asset classes, *Journal of Financial Economics* 126, 1–35.

Hou, Kewei, Chen Xue, and Lu Zhang, 2020, Replicating anomalies, *Review of Financial Studies* 33, 2019–2133.

Huang, Dashan, Fuwei Jiang, Kunpeng Li, Guoshi Tong, and Guofu Zhou, 2022, Scaled PCA: A new approach to dimension reduction, *Management Science* 68, 1678–1695.

Huang, Dashan, Fuwei Jiang, Jun Tu, and Guofu Zhou, 2015, Investor sentiment aligned: A powerful predictor of stock returns, *Review of Financial Studies* 28, 791–837.

Jurado, Kyle, Sydney C. Ludvigson, and Serena Ng, 2015, Measuring uncertainty, *American Economic Review* 105, 1177–1216.

Kan, Raymond, and Chu Zhang, 1999, Two-pass tests of asset pricing models with useless factors, *Journal of Finance* 54, 203–235.

Kelly, Bryan, Seth Pruitt, and Yinan Su, 2019, Characteristics are covariances: A unified model of risk and return, *Journal of Financial Economics* 134, 501–524.

Kim, Soohun, Robert A. Korajczyk, and Andreas Neuhierl, 2021, Arbitrage portfolios, *Review of Financial Studies* 34, 2813–2856.
Kleibergen, Frank, 2009, Tests of risk premia in linear factor models, *Journal of Econometrics* 149, 149–173.
Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh, 2020, Shrinking the cross-section, *Journal of Financial Economics* 135, 271–292.
Lettau, Martin, and Markus Pelger, 2020, Estimating latent asset-pricing factors, *Journal of Econometrics* 218, 1–31.
Lewellen, Jonathan, Stefan Nagel, and Jay Shanken, 2010, A skeptical appraisal of asset pricing tests, *Journal of Financial Economics* 96, 175–194.
Ludvigson, Sydney C., and Serena Ng, 2010, A factor analysis of bond risk premia, in Aman Ulah, and David E. A. Giles, eds.: *Handbook of Empirical Economics and Finance*, Vol. 1, 313–372 (Chapman and Hall, Boca Raton, FL).
Pástor, Luboš, and Robert F. Stambaugh, 2003, Liquidity risk and expected stock returns, *Journal of Political Economy* 111, 642–685.
Pesaran, M. Hashem, and Ron P. Smith, 2019, The role of factor strength and pricing errors for estimation and inference in asset pricing models, Technical report, CESifo Working paper.
Ross, Stephen A., 1976, The arbitrage theory of capital asset pricing, *Journal of Economics Theory* 13, 341–360.
Schorfheide, Frank, Dongho Song, and Amir Yaron, 2018, Identifying long-run risks: A Bayesian mixed-frequency approach, *Econometrica* 86, 617–654.
Shanken, Jay, 1992, On the estimation of beta pricing models, *Review of Financial Studies* 5, 1–33.
Uematsu, Yoshimasa, Yingying Fan, Kun Chen, Jinchi Lv, and Wei Lin, 2019, Sofar: Large-scale association network learning, *IEEE Transactions on Information Theory* 65, 4924–4939.
Uematsu, Yoshimasa, and Takashi Yamagata, 2022a, Estimation of sparsity-induced weak factor models, *Journal of Business & Economic Statistics* 41, 213–227.
Uematsu, Yoshimasa, and Takashi Yamagata, 2022b, Inference in sparsity-induced weak factor models, *Journal of Business & Economic Statistics* 41, 126–139.
Wan, Runzhe, Yingying Li, Wenbin Lu, and Rui Song, 2024, Mining the factor zoo: Estimation of latent factor models with sufficient proxies, *Journal of Econometrics* 239, 105386.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1: Internet Appendix.
Replication Code.