Economically Guided Sparse Factors

Qihui Chen* Lin William Cong† Chunyu Qi‡

December 16, 2025

Abstract

We propose an economically guided and "doubly sparse" factor framework that combines the strengths of latent and observable factor models in asset pricing. By imposing economic objectives (including constraints) as well as sparsity on PCA, we construct a small set of factors that capture return comovements, while being economically interpretable and close to the efficient frontier. These factors entail only a few characteristics each and correspond to broad concepts of market, momentum, liquidity-profitability, etc., and jointly outperform traditional latent and observable factor models (e.g., PCA-based and Fama-French factor models) in pricing the cross-sectional returns and investment profit (e.g., Sharpe ratio) in U.S. equity test samples. Our framework also nests model-based methods such as Risk-Premium-PCA, reduces overfitting in both factor construction and risk premium estimation, and offers a unified and economically meaningful representation of stochastic discount factors.

Keywords: Asset Pricing, Characteristics, Economic Interpretability, Latent Factors, Machine Learning, Risk Premia, Sparsity.

*School of Management and Economics, The Chinese University of Hong Kong, Shenzhen; qihuichen@cuhk.edu.cn

†Cornell Johnson Graduate School of Management, ABFER, IC3, and NBER; will.cong@cornell.edu

‡School of Management and Economics, The Chinese University of Hong Kong, Shenzhen; chunyuqi@link.cuhk.edu.cn

1 Introduction

Empirical asset pricing has historically oscillated between latent and observable representations of systematic risks. Ross (1976)'s Arbitrage Pricing Theory (APT) formalizes that a set of risk factors govern return comovements and explain the cross-section of expected returns. Early empirical implementations, such as Connor and Korajczyk (1986, 1988), extract a few latent risk factors using PCA on individual stock returns. These statistically derived factors offer little economic interpretation, hampering their use for portfolio construction and for explaining the cross-section of expected returns. Partly in reaction to these deficits, the literature embraced observable factors constructed from firm characteristics and portfolio sorts—most prominently the size, value, profitability, and investment factors of Fama and French (1993, 2015) and Carhart (1997)'s momentum factor. Because these characteristic-sorted factors are tradable, economically interpretable, and effective in explaining the cross-section of expected returns, they quickly became workhorses for empirical tests and for the "risk-versus-mispricing" debate. Yet the proliferation of such factors ("factor zoo") revealed two limitations of the observable factor approach: high dimensionality makes it difficult to isolate priced sources of risk, and naive sorts often lie far from the mean-variance efficient frontier.

Advances in computation and high-dimensional statistics have recently drawn the field back toward latent representations. Methods such as instrumented PCA (IPCA, Kelly et al., 2019), risk-premium PCA (RP-PCA, Lettau and Pelger, 2020b), SDF-first approach (Kozak et al., 2020), and regressed-PCA (Chen et al., 2021) recover latent factors that combine information from many characteristics and outperform traditional sorts in spanning the return space. Nevertheless, their very complexity resurrects the old interpretability problem: the estimated factors are dense linear combinations of characteristic-sorted or managed portfolios, making it difficult to attach clear economic meaning. Furthermore, these factors often embed offsetting positions that neutralize each other's premia, sometimes producing "unpriced" factors out-of-sample.

To reconcile these two traditions, we introduce sparsity into the PCA factor construction and guiding it with an economic constraint. Our framework allows us to obtain economically guided sparse (EGS) factors that (i) remain latent in the sense that they are linear combinations of a few basis portfolios, yet (ii) are as interpretable and parsimonious as the classic observable factors. Empirically, the first sparse factor from a set of characteristic-managed portfolios collapses to the equal-weighted market portfolio, while the remaining factors line up with momentum, liquidity-profitability, short-term reversal/dividend omission, value, and R&D themes. These factors lie markedly closer to the efficient frontier compared to both latent-factor methods (e.g., PCA and IPCA) and leading observable-factor models (e.g., Fama-French and Hou-Xue-Zhang factors), deliver higher in- and out-of-sample Sharpe ratios, and explain a substantially larger share of testing assets—with far less characteristic redundancy—which advances the ultimate goal of constructing a low-dimensional, economically meaningful stochastic discount factor representation.

Classical latent-factor estimators such as PCA are designed to maximize the share of covariance they explain, not the expected-return dimension that ultimately matters for asset pricing. As Lettau and Pelger (2020a,b) show, the leading PCA eigenvectors mostly pick up "strong" sources of return comovement; weaker directions that may carry high risk premia are relegated to later components or obscured by noise, so the resulting factors have low Sharpe ratios even when they account for most time-series covariance. Because the optimization ignores factor risk premia, it is free to load on any returns that co-move, including transitory or unpriced shocks. The estimated factor portfolios therefore look impressive in capturing time-series comovement—they soak up volatility—but transmit little information about the stochastic discount factor that prices assets.

Dense, variance-optimal factors are also statistically fragile. In typical panels (hundreds of assets, a few decades of data) the ratio N/T is large, so sample eigenvectors inherit substantial estimation error; random-matrix theory predicts that many of their entries are dominated by noise (Johnstone and Lu, 2009), When those noisy loadings are carried forward into second-pass regressions, they amplify small-sample bias and produce unstable risk-premium estimates—a textbook form of out-of-sample overfitting. Shrinkage techniques confirm the point: Kozak et al. (2020) obtain markedly lower pricing errors once they shrink the dense PCA factors toward a sparser prior, implying that many original loadings were pure estimation noise.

Imposing economically guided sparsity tackles both shortcomings at once. An ℓ_1-type penalty trims away exposures that (i) are weakly motivated by theory and (ii) contribute little to cross-sectional pricing, leaving a compact set of economically interpretable portfolios. By concentrating estimation effort on those few, high-signal coordinates, the procedure lowers variance, mitigates eigenvector instability, and elevates weak-but-priced directions that PCA overlooks—an effect also documented for "sparse proximate factors" in Pelger and Xiong (2022) and Fan et al. (2022). The upshot is a factor model that still captures common variation (because the retained loadings span return space) yet delivers sharper, more robust SDF estimates and superior out-of-sample mean-variance performance. In short, latent methods without structure tend to "explain comovement"; adding economically grounded sparsity turns that movement into "priced risk."

Our sparsity emphasizes economic discipline over mere statistical frugality. We determine the optimal level of sparsity through an economically guided selection, which constrains factor construction to long positions in characteristic-managed portfolios with positive expected returns and short positions in those with negative expected returns.

Such a design ensures that each position contributes positively to the overall factor risk premium, thereby pushing the resulting factors toward the mean-variance efficient frontier. Interestingly, this selection echoes the spirit of Fama and French (1993), who assign positive (negative) weights to sorted portfolios with high (low) expected returns. Both approaches leverage the monotonicity of characteristics' predictive power for returns. It also aligns with recent findings by Lettau and Pelger (2020b) and Bryzgalova et al. (2023), who show that embedding risk premia or other economic targets into the PCA criterion tends to produce factors that predominantly place positive (negative) weights on sorted portfolios with high (low) expected returns.[1]. In contrast, alternative sparse latent factor approaches, such as those of Kozak et al. (2020), Pelger and Xiong (2022), and Fan et al. (2022), primarily emphasize statistical frugality.

Empirical analysis further confirms that our economically guided sparsity enhances the mean-variance efficiency of factors and strengthens their cross-sectional explanatory power. For any given number of factors, the EGS factors consistently outperform PCA, IPCA, and observable factors in terms of mean-variance efficiency. For instance, the first six EGS factors achieve an out-of-sample Sharpe ratio of 2.04, significantly higher than the 0.97 for PCA factors, 1.57 for IPCA factors, and 0.95 for the Fama-French six-factor model. In addition, EGS factors exhibit stronger explanatory power for test assets than other latent and observable factors. When explaining our 53 characteristic-managed portfolios, the first six EGS factors reduce the out-of-sample root-mean-squared (RMS) *alpha* to 0.48%, lower than the 0.58% for PCA, 0.54% for IPCA, and 0.61% for the Fama-French six-factor model. PCA and IPCA factors show poor out-of-sample explanatory power for Fama and French (2018) 100 portfolios and Kozak et al. (2018) 110 portfolios. Specifically, the first six PCA factors explain these two sets of test assets with RMS *alpha*s of 0.76% and 0.68%, respectively, and the first six IPCA factors yield RMS *alpha*s of 0.57% and 0.55%, both substantially higher than the 0.34% and 0.35% for the Fama-French six-factor model. However, our EGS factors avoid this issue. The first six EGS factors reduce the RMS *alpha* for these two sets of test assets to 0.23% and 0.31%, respectively.

Our paper contributes to three strands literature. First, our methodology delivers a unifying lens on latent and observable factors. These two classes of factors can be viewed as nested limiting cases of our sparse factors, corresponding to extreme values of the sparsity hyperparameter. Specifically, both classes can be expressed as linear combinations of basis portfolios. In the sparsest limit, the factors are constructed from only one characteristic-managed portfolio, falling into the category of observable factors (Fama and French (2020)'s cross-sectional factors). In the densest limit, the factors lose sparsity and revert to PCA factors. We show that an intermediate level of sparsity,

[1] Different from their reliance on the relative magnitudes of expected returns, our economically guided selection leverages the sign of expected returns, because characteristic-managed portfolios are zero-cost whereas sorted portfolios are not.

guided by economic content, achieves the best trade-off—yielding factors that combine strong economic interpretability with superior mean-variance efficiency relative to both extremes.

Second, our EGS factors strike a compelling balance among three key dimensions: economic interpretability, the ability to capture time-series comovement, and explanatory power for the cross-section of expected returns. In Figure 1, we compare several factor construction methods along these three dimensions. PCA and IPCA factors are designed primarily to capture comovement, often at the expense of interpretability and cross-sectional explanatory power. In contrast, observable factors prioritize interpretability and cross-sectional performance but are less effective in capturing common variation. Our method achieves an economically optimal trade-off between these two extremes. It delivers economic interpretability comparable to observable factors, comovement capture close to PCA and IPCA factors, and significantly stronger cross-sectional explanatory power than both. Lettau and Pelger (2020b) and Bryzgalova et al. (2023) also pursue a balance between comovement capture and cross-sectional performance by incorporating economic information into factor construction; however, they do not explicitly address the dimension of economic interpretability.

Third, we achieve sparsity both in factor-construction weights and in risk prices. While Kozak et al. (2020) introduce sparsity (and shrinkage) in the estimation of risk prices, their approach relies on pre-specified factors (i.e., PCs of characteristic sorted- or -managed portfolios). As such, their approach addresses overfitting only in the second-stage risk price estimation, without addressing potential overfitting embedded in the construction of the PCA factors themselves. In contrast, we impose sparsity from the ground up: we first construct sparse factors and then estimate sparse risk prices using the Kozak et al. (2020) framework. This joint-sparsity approach mitigates overfitting at both stages.

Moreover, our EGS framework can be generalized by incorporating economic objectives in the factor construction, in line with the spirit of Bryzgalova et al. (2023). By employing an economic objective based on the risk premium, our EGS framework can be integrated with Lettau and Pelger (2020b)' Risk-Premium PCA (RP-PCA), yielding what we refer to as the RP-EGS factors. We find that, although the RP-PCA factors achieve extremely high Sharpe ratios and explain the baseline portfolios effectively, their overfitting to the first moments of the baseline portfolios leads to poor performance in explaining other test assets. In contrast, the RP-EGS factors retain the high Sharpe ratios of the RP-PCA factors while substantially enhancing their ability to explain these additional test assets.

The remainder of the paper is organized as follows. Section 2 outlines the construction of our EGS factors and describes their relation to both PCA and observable factors. Section 3 demonstrates the strong interpretability of the EGS factors, and shows that

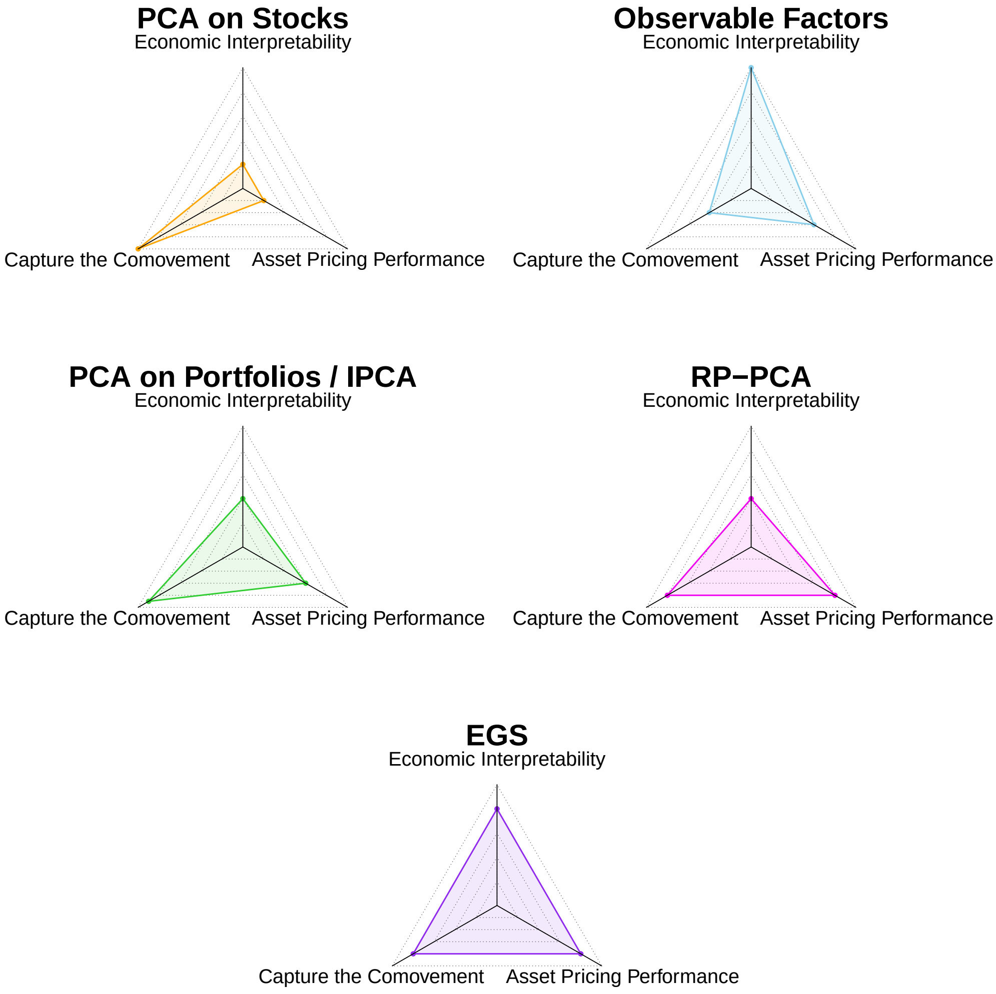


Figure 1. Trade-offs among factor construction approaches

the EGS factors deliver a higher level of mean-variance efficiency and exhibit greater explanatory power for the cross-section of expected returns. Section 4 concludes.

2 Constructing Economically Guided Sparse Factors

Let X_t be a $p \times 1$ vector of excess returns on characteristic-sorted or managed portfolios in period t, for $t = 1, \ldots, T$. We assume these excess returns follow an approximate factor structure[2]:

$$X_t = (a) + Bf_t + e_t, \quad t = 1, \cdots, T, \tag{1}$$

where f_t is a $K \times 1$ vector of latent factors, B is an unknown $p \times K$ matrix of factor loadings, and e_t is the idiosyncratic component. The $p \times 1$ vector a represents pricing errors. We assume that a is either asymptotically negligible or orthogonal to the factor loadings, i.e., $a'B = 0$.

Given this structure, conventional PCA provides a natural method for factor construction. It estimates the factor loadings as

$$\tilde{B} = \underset{B \in \mathbb{R}^{p \times K}}{\arg\max} \operatorname{tr}(B'\hat{\Sigma}B) \quad \text{s.t. } B'B = I_K, \tag{2}$$

where $\hat{\Sigma} \equiv \sum_{t=1}^{T}(X_t - \bar{X})(X_t - \bar{X})'/T$ and $\bar{X} \equiv \sum_{t=1}^{T} X_t/T$ denote the sample covariance matrix and the sample mean of X_t, respectively. The optimization in (2) can be solved via eigendecomposition of $\hat{\Sigma}$, with the columns of $\tilde{B}$ given by the eigenvectors associated with the K largest eigenvalues. The resulting factors are:

$$\tilde{f}_t = (\tilde{B}'\tilde{B})^{-1}\tilde{B}'X_t = \tilde{B}'X_t, \tag{3}$$

which maximally summarize the variance in X_t and capture the most time-series comovement among the assets. Despite its simplicity and optimality for capturing comovement, conventional PCA yields dense loading matrix $\tilde{B}$, rendering the resulting factors $\tilde{f}_t$ difficult to interpret. Moreover, these factors often involve offsetting long and short positions that cancel out individual portfolio premia (see Figures 3 and 4), thereby undermining their mean-variance efficiency and limiting their ability to explain the cross-section of expected returns.

To improve the interpretability of the estimated factors, we introduce sparsity into the loading matrix B. However, directly imposing a sparsity penalty in the standard PCA formulation (2) yields a nonconvex optimization problem that is computationally challenging. To address this, we adopt the sparse PCA framework of Zou et al. (2006),

[2]In addition to the PCA-based perspective, the EGS factors can also be motivated from conditional factor models; see Appendix A.

which employs two separate loading matrices, B and A, to impose sparsity and orthogonality, respectively.[3] To clarify this approach, we first rewrite the PCA objective (2) in an equivalent form:

$$(\tilde{A}, \tilde{B}) = \underset{A,B \in \mathbb{R}^{p\times K}}{\arg\min} \; -2\mathrm{tr}(A'\hat{\Sigma}B) + \|B\|_F^2 \quad \text{s.t. } A'A = I_K, \tag{4}$$

where the optimal solutions $\tilde{A}$ and $\tilde{B}$ both coincide with the PCA loadings.[4] We then incorporate element-wise sparsity into this reformulated objective function:

$$(\hat{A}, \hat{B}) = \underset{A,B \in \mathbb{R}^{p\times K}}{\arg\min} \; -2\mathrm{tr}(A'\hat{\Sigma}B) + \|B\|_F^2 + \sum_{k=1}^{K} \lambda_k \|w_k \circ b_k\|_1 \quad \text{s.t. } A'A = I_K, \tag{5}$$

where two types of loading matrices are used: B imposes sparsity, while A enforces orthogonality. Sparsity in B is induced via an adaptive lasso penalty. For each column b_k of B ($k = 1, \cdots, K$), w_k is a $p \times 1$ vector of adaptive weights, and $\circ$ denotes element-wise multiplication. The tuning parameters $\lambda_k > 0$ govern the degree of sparsity applied to the k-th column. Compared with the standard lasso penalty employed by Zou et al. (2006), the adaptive lasso penalty yields estimators with favorable asymptotic properties (Leng and Wang, 2009). When $\lambda_1 = \cdots = \lambda_K = 0$, the objective function reduce to (4), and the solutions for both A and B collapse to the conventional PCA loadings.

The optimization problem (5) can be efficiently solved using an alternating minimization algorithm. Given $A = (a_1, \ldots, a_K)'$, the minimizer for each b_k is obtained via soft-thresholding:

$$b_k = \max\left(|\hat{\Sigma}a_k| - \frac{\lambda_k w_k}{2}, 0\right) \cdot \mathrm{sign}(\hat{\Sigma}a_k). \tag{6}$$

Given B, the optimal A is:

$$A = \hat{\Sigma}B(B'\hat{\Sigma}'\hat{\Sigma}B)^{-1/2}. \tag{7}$$

The algorithm iteratively updates A and B until convergence. Initialization uses the standard PCA loadings $A = \tilde{B}$, and each column of B is normalized at the end. Adaptive weights are set as the element-wise inverse of the PCA loadings, $w_k = 1 \oslash \tilde{b}_k$, where $\tilde{b}_k$ denotes the k-th column of $\tilde{B}$ and $\oslash$ denotes element-wise division. Denoting the

[3] Zou et al. (2006) propose two versions of sparse PCA based on different PCA objectives. We adopt the version corresponding to (2), which is computationally more efficient.

[4] Let $f(A, B) = -2\mathrm{tr}(A'\hat{\Sigma}B) + \|B\|_F^2$. This objective is quadratic in B for fixed A, and its minimizer is given by $B(A) = \hat{\Sigma}A$. Substituting this into (4) yields

$$\tilde{A} = \underset{A\in\mathbb{R}^{p\times K}}{\arg\max}\, \mathrm{tr}(A'\hat{\Sigma}^2 A) \quad \text{s.t. } A'A = I_K$$

Thus, $\tilde{A}$ consists of the first K eigenvectors of $\hat{\Sigma}^2$, which are identical to those of $\hat{\Sigma}$. Since $\tilde{B} = \hat{\Sigma}\tilde{A} = \tilde{A}\Theta$ (where $\Theta = \mathrm{diag}(\theta_1, \cdots, \theta_K)$ and $\theta_1, \cdots, \theta_K$ are the K largest eigenvalues of $\hat{\Sigma}$), the columns of B are also the first K eigenvectors of $\hat{\Sigma}$, scaled by the corresponding eigenvalues.

algorithm's output as $\hat{B}$, the estimated factors are

$$\hat{f}_t = (\hat{B}'\hat{B})^{-1}\hat{B}'X_t = \hat{B}'X_t. \tag{8}$$

Because the signs of the loadings and factors are not uniquely determined, we adopt a convention to ensure a consistent interpretation: the signs are chosen so that the sample means of the estimated factors are positive, thereby ensuring positive risk premia.

It remains to choose the hyperparameters in the construction of EGS factors. To address the issue of offsetting positions and "unpriced" factors (details provided in Subsection 2.2), we introduce an economic constraint that guides the degree of sparsity. Specifically, each factor is required to take long positions in characteristic-managed portfolios with positive expected returns (or in characteristic-sorted portfolios with high expected returns) and short positions in those with negative (or low) expected returns. This constraint leverages the monotonic relationship between characteristics and expected returns, echoing the spirit of Fama and French (1993), who construct long-short portfolios based on characteristic sorts. By imposing this constraint, we ensure that all selected portfolios contribute positively to the factor's risk premium, thereby avoiding the inclusion of offsetting or redundant positions that could dilute its mean-variance efficiency[5]. Among all hyperparameter values that satisfy this economic constraint, we choose the one that maximizes the Sharpe ratio of the tangency portfolio formed by all factors.

In our empirical analysis, we adopt characteristic-managed portfolios constructed via Fama-MacBeth regressions. Let Z_t denote an $N \times p$ matrix of p characteristics (which may include a constant term) for N individual stocks in period t, and let R_t denote their excess returns. The characteristic-managed portfolios are constructed as:

$$X_t = (Z'_{t-1}Z_{t-1})^{-1}Z'_{t-1}R_t. \tag{9}$$

These portfolios possess several attractive properties. First, if Z_t includes a constant term (e.g., as its first element) and the nonconstant characteristics are standardized to have a zero mean, then the first element of X_t represents a "level" return. This is the equal-weighted average excess return across all stocks—essentially equal-weighted market portfolio. It has no exposure to any characteristic by construction and costs $1 to implement (i.e., its weights sum to one). Second, as shown by Fama (1976), the remaining elements of X_t are correspond to excess returns on zero-cost portfolios. Each of these portfolios is constructed to have unit exposure to one characteristic and zero exposure to all others. As a result, they isolate the return premium associated with a single characteristic, effectively offering a "pure play" on that characteristic. These can be interpreted as "slope" returns, in contrast to the "level" return represented by the

[5] Campbell and Thompson (2008) employs similar sign constraints to improve equity premium forecasts.

first component. These two properties follow directly from the identity $W_{t-1}Z_{t-1} = I_p$, where $W_{t-1} = (Z'_{t-1}Z_{t-1})^{-1}Z'_{t-1}$ is the weight matrix used to construct X_t. Third, under standard i.i.d. assumptions similar to those in OLS regression, the portfolios in X_t have minimum variance and are often weakly correlated, making them highly diversified in terms of risk. Fourth, as shown by Chen et al. (2021) and Appendix A, performing PCA on X_t provides an econometric method for estimating conditional factor models of individual stock returns. Additional properties of these portfolios are discussed in Hoberg and Welch (2009) and Kirby (2020).

The Fama-MacBeth characteristic-managed portfolios offer several advantages over traditional sorted portfolios. First, sorting requires an arbitrary choice of the number of groups—such as terciles or quintiles—which can significantly influence the resulting portfolios and make findings sensitive to this ex-ante decision. In contrast, Fama-MacBeth regressions do not involve such discretization and instead utilize the full distribution of characteristics. Second, sorting suffers from the curse of dimensionality, making it infeasible to sort on more than four characteristics at a time. This limitation hampers the ability to isolate the unique return effects of each characteristic. In contrast, Fama-MacBeth regressions can easily incorporate a large number of characteristics—such as hundreds—using data on a few thousand individual stocks, enabling richer and more comprehensive analysis. Third, sorting fails to exploit within-group variation in characteristics, as it only uses the differences between groups. In contrast, Fama-MacBeth regressions fully leverage the continuous cross-sectional variation in characteristics, leading to more efficient and precise estimates of characteristic-related return premia.

2.1 A Unifying Lens

Our framework provides a unifying lens through which a subset of latent and observable factors can be understood as two limiting cases of our EGS factors. In the absence of sparsity penalty—i.e., when the hyperparameters are set to zero—the sparse PCA algorithm described in (5)-(7) reduces to the standard PCA and yields the loading matrix $\tilde{B}$, producing the conventional PCA factors. Furthermore, a more general EGS framework, further elaborated in Section 4, generates RP-PCA factors when the sparsity-inducing parameters are set to zero. Conversely, when sparsity is maximized by setting the hyperparameters to large values, each column of B contains only a single nonzero entry,[6] indicating that the resulting factor is constructed from a single characteristic-managed portfolio. This places the factor firmly in the category of observable factors, resembling the cross-sectional factors of Fama and French (2020) and the OLS factors of Kozak and Nagel (2023). The relationship among EGS factors, observable factors, and latent factors is summarized in Figure 2.

[6]While larger hyperparameter values can shrink B to zero, such cases do not yield meaningful factors.

Most latent and observable factors can be viewed as constructed from a common set of basis portfolios—portfolios that effectively summarize the pricing information embedded in individual stocks using firm characteristics. These basis portfolios generally fall into two categories. The first consists of characteristic-sorted portfolios, where stocks are grouped based on a specific characteristic (often in combination with firm size), and stocks within each group form a value- or equal-weighted portfolio. The second category includes characteristic-managed portfolios, which can be constructed either via Fama-MacBeth regressions as $X_t = (Z'_{t-1}Z_{t-1})^{-1}Z'_{t-1}R_t$ (as in our approach), or by directly using characteristics as portfolio weights, as in $X_t = Z'_{t-1}R_t/N$ (e.g., Kozak et al. (2020)).

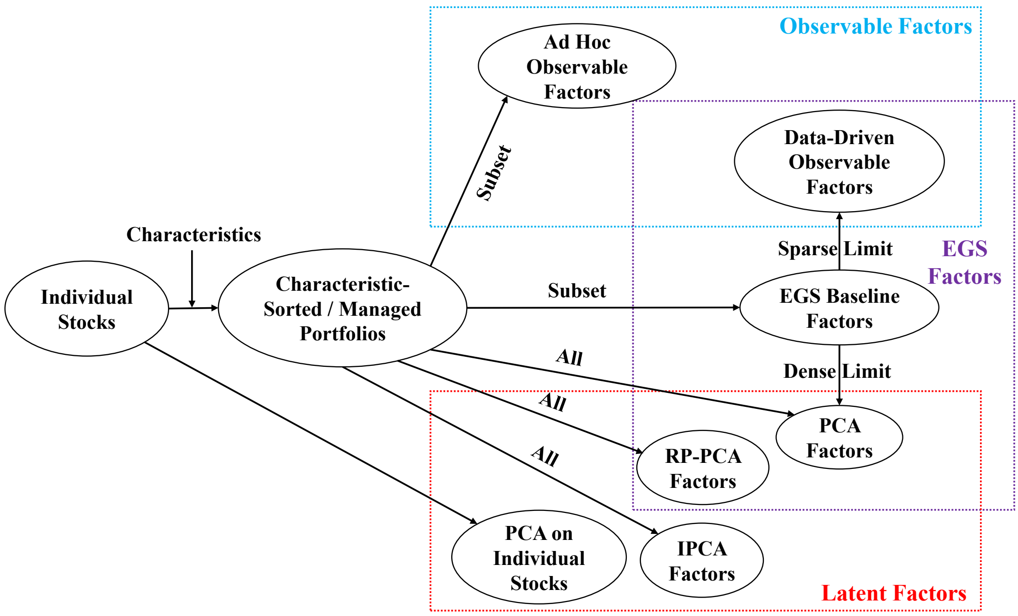


Figure 2. A unifying perspective: latent and observable factors

Observable factors are typically built from a small subset of these basis portfolios that correspond to a specific characteristic. For instance, the long-short factors of Fama and French (1993) take long positions in one tail of the sorted portfolios and short positions in the other, both sorted on the same characteristic (e.g., firms with high or low book-to-market ratios).[7] In contrast, the cross-sectional factors of Fama and French (2020) directly select a single characteristic-managed portfolio as the factor. However, unlike the observable factors derived from our EGS factors, both of these approaches are expertise-driven rather than data-driven—that is, the representative basis portfolios are selected based on researcher judgment rather than being algorithmically determined.

Latent factors, on the other hand, are typically formed as dense linear combinations of a large number of characteristic-sorted or -managed portfolios. Kozak et al.

[7]In double sorting, we usually focus on the effect of one characteristic while controlling for another. Therefore, the resulting sorted portfolios are primarily associated with the first characteristic.

(2018, 2020) apply classical PCA directly to such portfolios to extract latent factors. Lettau and Pelger (2020b) and Bryzgalova et al. (2023) extend this PCA framework by incorporating pricing error minimization or other economic targets into the factor construction objective. Factors based on conditional factor models can also be expressed in terms of basis portfolios. For example, the IPCA factors of Kelly et al. (2019) can be represented as linear combinations of the characteristic-managed portfolios given by $X_t = Z'_{t-1}R_t/N$[8]; Regressed-PCA factors of Chen et al. (2021) are directly obtained by applying PCA directly to a set of Fama-MacBeth characteristic-managed portfolios.

Our EGS factors lie between these two extremes. The level of sparsity is guided by an economic constraint that favors long (short) positions in portfolios with positive (negative) expected returns. Empirically, we find that EGS factors capture return comovement as effectively as latent factors, offer economic interpretability on par with observable factors, and deliver superior explanatory power for the cross-section of expected returns compared to both. This suggests that neither fully dense (latent) nor fully sparse (observable) factor constructions are optimal. Instead, an economically guided, intermediate approach—such as ours—may provide the most desirable balance.

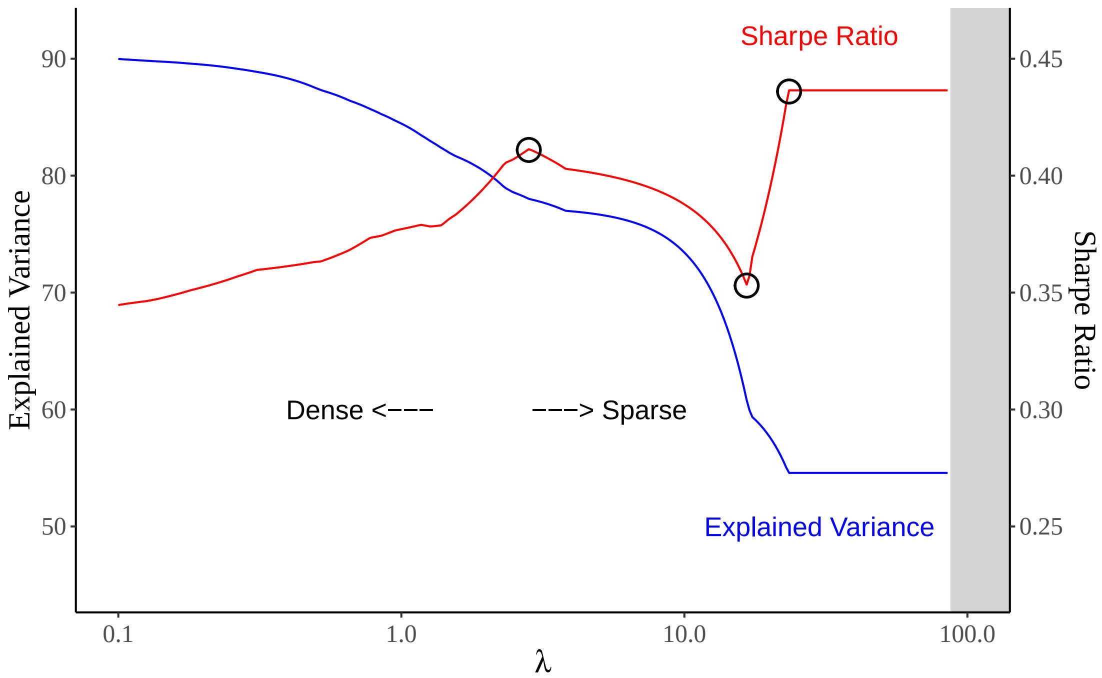
2.2 An Illustration

To illustrate the construction of our EGS factors, we present an empirical analysis of the first factor. Figure 3 displays the Sharpe ratios and explained variances of the first factor across different levels of sparsity. These two metrics reflect the factor's mean-variance efficiency and its ability to capture return comovement, respectively. Our results show that while denser factors tend to explain more comovement, sparser factors deliver higher Sharpe ratios.

A more in-depth analysis of the factor structure reveals the underlying mechanism behind this trade-off. Figure 4 presents the composition of the first at four representative levels of sparsity, ranging from the densest to the sparsest. When no sparsity is imposed (i.e., $\lambda = 0$), the factor takes positions across nearly all characteristic-managed portfolios, which can be categorized into four groups: (i) long positions in portfolios with positive expected returns (red bar); (ii) long positions in portfolios with negative expected returns (blue bar); (iii) short positions in portfolios with positive expected returns; and (iv) short positions in portfolios with negative expected returns[9]. Among these, positions (i) and (iv) contribute positively to the factor's risk premium, while positions (ii) and (iii) detract from it. In the densest case, the contributions of these opposing positions are roughly equal, leading to a significant offsetting effect that reduces

[8] IPCA factors are constructed as $\hat{f}_t = (\check{B}' Z'_{t-1} Z_{t-1} \check{B})^{-1} \check{B}' Z'_{t-1} R_t$, where $\check{B}$ is the output of the IPCA algorithm. Rigourouly speaking, these factors are built from the characteristic-managed portfolios, $X_t = Z'_{t-1} R_t / N$, with time-varying weights, $\check{B}(\check{B}' Z'_{t-1} Z_{t-1} \check{B}/N)^{-1}$.

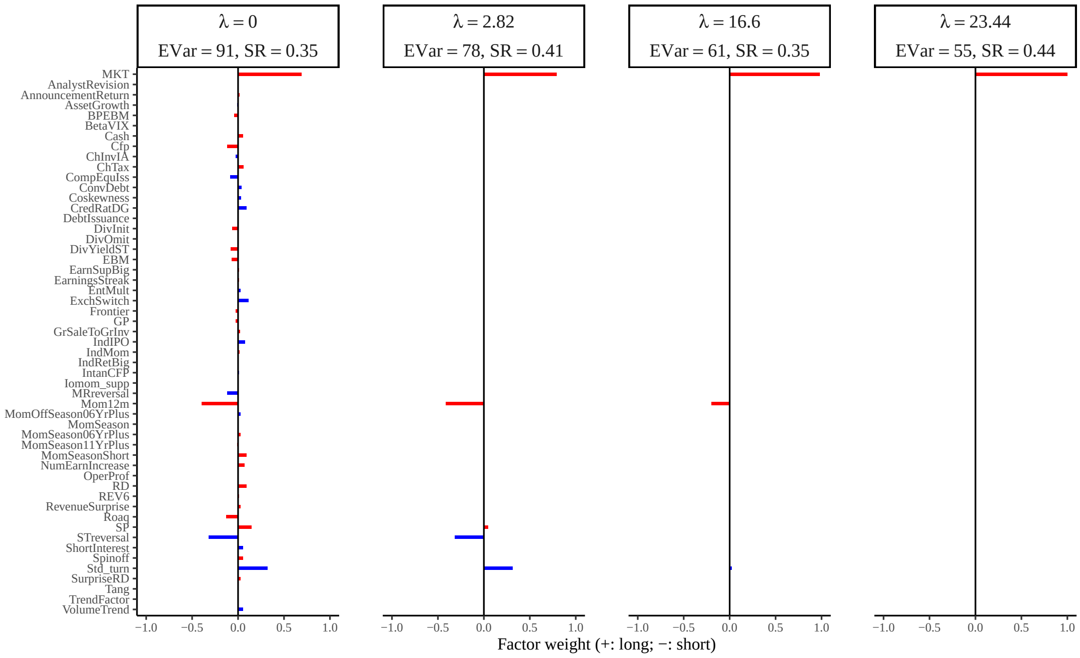
[9] The expected returns of the characteristic-managed portfolios are estimated by their time-series average returns.



Notes: This figure illustrates the relationship between the first Sharpe ratio and explained variance of the first factor as a function of the sparsity level. The horizontal axis is the value of the hyperparameter λ, with higher values corresponding to greater sparsity. The red line plots the Sharpe ratio of the factor, while the blue line shows the proportion of variance it explains. The gray shaded area indicates the region where excessively large values of λ result in a zero factor due to overshrinakge.

Figure 3. Tradeoff between Sharpe ratio and explained variance

the net risk premium. As a result, the factor's explanatory power for the cross-section of expected returns is limited. As sparsity increases, this offsetting effect gradually diminishes. The factor becomes more selective, retaining only those portfolios that contribute positively to the risk premium. This results in improved mean-variance efficiency and stronger cross-sectional explanatory power. Our economically guided selection procedure rules out the first three levels of sparsity and identifies the final level as optimal.



Notes: This figure shows the composition of the first factor at four levels of sparsity. The first subplot corresponds to the fully dense case ($\lambda = 0$). The next three subplots depict the factor structure at the kink points marked in Figure 3. These values represent the weights assigned to each characteristic-managed portfolio for each factor. The colors indicate the sign of the expected returns for the characteristic-managed portfolios: red denotes positive expected returns, while blue denotes negative expected returns.

Figure 4. Composition of the first factor at different levels of sparsity

3 Empirical Findings on U.S. Equities

3.1 Data

Stock return data are sourced from the CRSP database. The sample includes all common stocks listed on the NYSE, AMEX, and Nasdaq, spanning the period from January 1973 to December 2022. This time period is chosen because, beginning in 1973, Nasdaq-listed stocks were incorporated into the CRSP database, significantly increasing the number of stocks available for analysis. After 1973, the number of stocks in each cross-section ranges from 3,500 to 7,500, compared to fewer than 2,500 before that period. This ensures that, when the number of stock characteristics is large, we have sufficient data to construct well-behaved managed portfolios.

Stock characteristics are obtained from Chen and Zimmermann (2022), one of the most comprehensive datasets currently available, which includes 212 characteristics. However, a large proportion of these characteristics suffer from significant missing data. To address this, we exclude those that, on average, cover fewer than one-third of the stocks in each cross-section, resulting in the removal of 53 characteristics. In addition, to ensure that the obtained managed portfolios effectively summarize pricing information, we apply two further filters to the remaining characteristics: (i) eliminating those that are highly correlated with others, as such multicollinearity can inflate the variance of the managed portfolios constructed via cross-sectional regressions; (ii) excluding characteristics that lack (independent) pricing information (Green et al., 2017), as they may generate managed portfolios that primarily capture unpriced risks. Further details are provided in Appendix B. After applying all filters, we retain 52 characteristics (or 53, including the constant). The remaining characteristics are processed following standard procedures: (i) missing values are imputed using the cross-sectional median for each characteristic; (ii) each characteristic is ranked and mapped to the interval $[-0.5, 0.5]$.

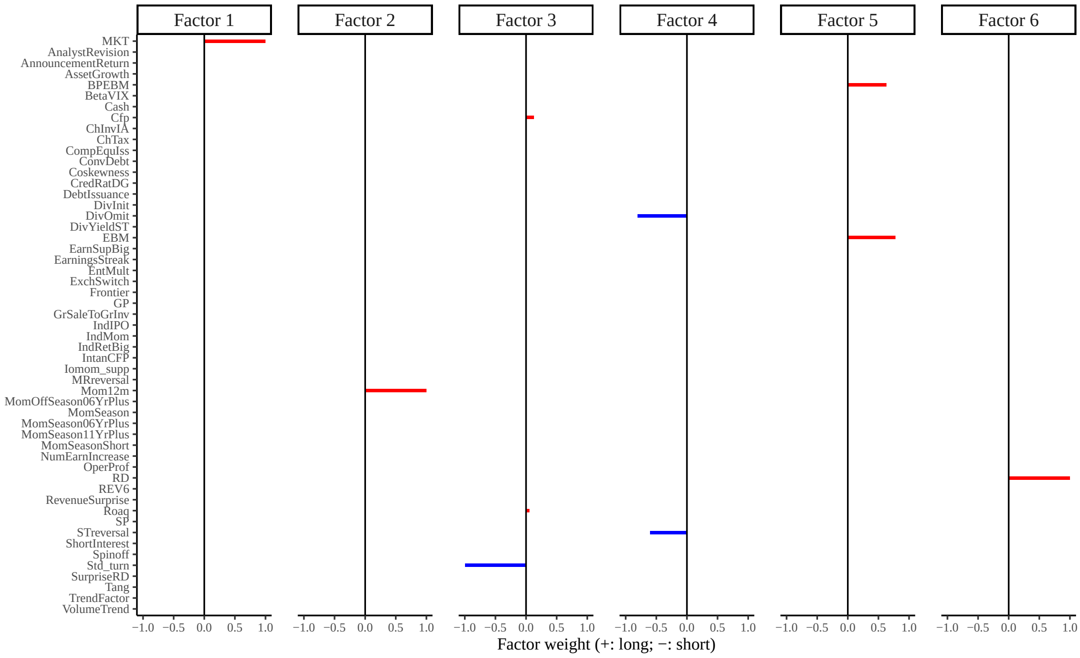
3.2 Economic Interpretability

Economically guided sparsity ensures that each factor is constructed from a limited, distinct subset of characteristic-managed portfolios, thereby enhancing the economic interpretability of the resulting factors. The kth column of $\hat{B}$ (denoted $\hat{b}_k$), which encodes the portfolio weights associated with the kth EGS factor, provides direct insight into the factor's composition, as illustrated in Figure 5. We focus on the case of $K = 6$, as leading factor models typically comprise no more than six factors (e.g., the Fama-French six-factor model).

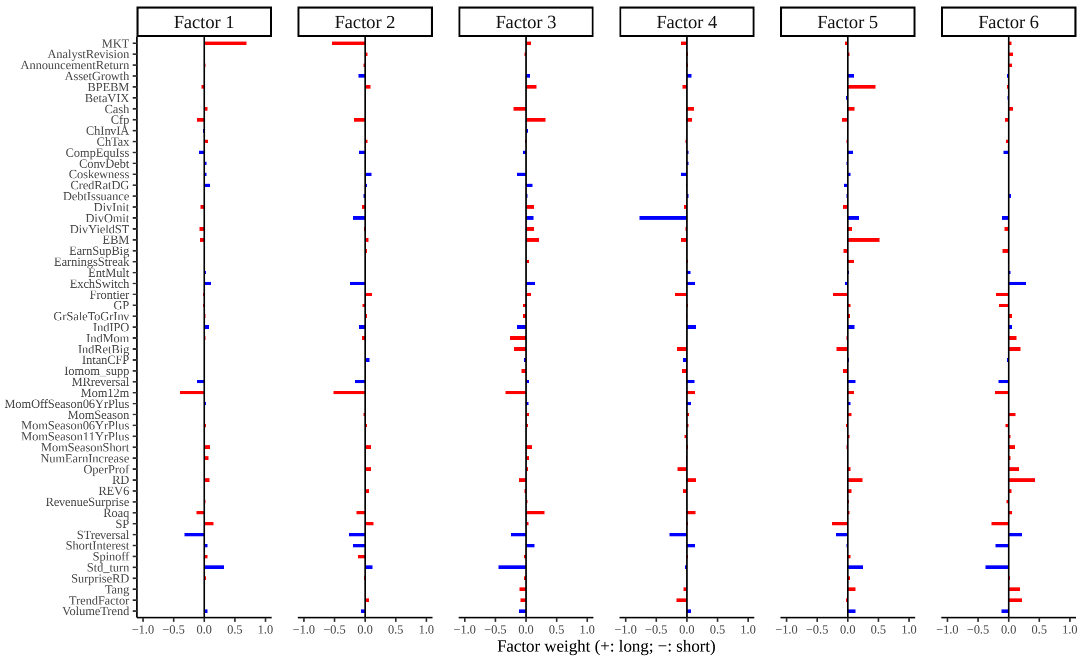
As shown in the figure, Factor 1 corresponds to the managed portfolio constructed from the intercept of the Fama-MacBeth regressions. Given that all characteristics have been normalized, this managed portfolio represents the equal-weighted average excess return across all individual stocks, commonly referred to as the "equal-weighted market factor" or the "level factor." This finding aligns with both the CAPM and leading observable factor models — including the Fama-French three- and five-factor models as well as the Hou-Xue-Zhang q-factor model — which similarly emphasize the market factor as the first (or sole) factor.

Factor 2 represents the momentum factor. While a body of literature views the momentum effect as a result of mispricing (e.g., Jegadeesh and Titman (1993); Hong and Stein (1999)), other studies emphasize risk-based explanations. For instance, Pástor and Stambaugh (2003) and Asness et al. (2013) argue that the momentum effect is associated with liquidity risk. Liu and Zhang (2008) demonstrated that the momentum effect can largely be explained by the growth rate of industrial production. Additionally, Ehsani and Linnainmaa (2022) posited that momentum may originates from the

Panel A: EGS Factors



Panel B: PCA Factors



Notes: the figure illustrates the values of $\hat{b}_k$ (the kth column of $\hat{B}$ for $k = 1, \cdots, 6$) corresponding to the EGS, PCA, and IPCA factors. These values represent the weights assigned to each characteristic-managed portfolio for each factor. The colors indicate the sign of the expected returns for the characteristic-managed portfolios: red denotes positive expected returns, while blue denotes negative expected returns.

Figure 5. Composition of EGS, PCA, and IPCA factors

Panel C: IPCA Factors

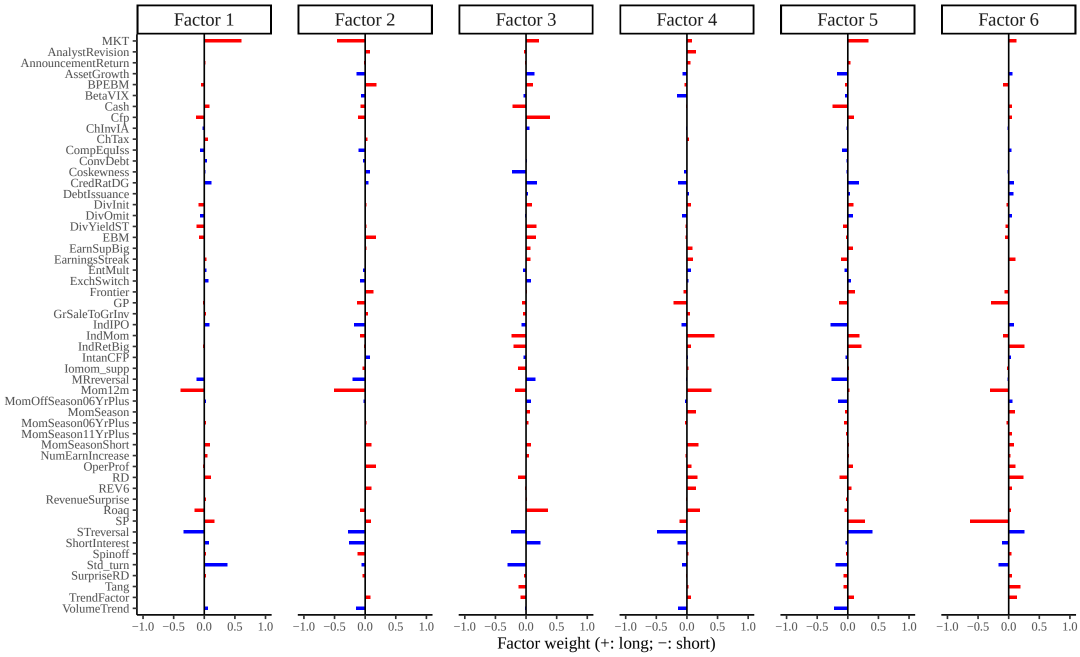


Figure 5. Composition of EGS, PCA, and IPCA factors (continued)

autocorrelations of other risk factors.

Factor 3 captures the joint effect of a stock's liquidity and the firm's profitability. Liquidity is proxied by trading volume volatility, while profitability is captured by the operating cash flow-to-price ratio and return on assets (ROA). The interaction between liquidity and firm fundamentals has also been highlighted by Lee and Swaminathan (2000), who document that firms with high (low) trading volumes often exhibit glamour (value) characteristics and tend to earn lower (higher) future returns.

Factor 4 reflects the joint effect of short-term reversal and dividend omission. As noted by Nagel (2012) and Da et al. (2014), short-term reversal can be partially attributed to the risk compensation received by market makers for providing liquidity. Our study indicates that dividend omission may interact with short-term reversal in driving stock returns, though this joint effect has been overlooked in the literature.

Factor 5 corresponds to the value factor. Penman et al. (2007) decomposed the book-to-market ratio into two distinct components: an enterprise component reflecting operating risk and a leverage component reflecting financing risk. However, in Factor 5, these components are recombined, once again capturing the overall value effect as in Fama and French (1993).

Factor 6 represents the R&D factor. This factor may be associated with the highly unpredictable nature of R&D project outcomes, as indicated by Chan et al. (2001).

Our results provide statistical insights into the ongoing "risk versus mispricing" debate on stock characteristics. Compared to other characteristics, those related to extracted factors behave more like risk, as they more effectively capture return comovement. However, caution is warranted when interpreting these characteristics as proxies for risk, and further research is essential to fully understand their underlying mechanisms.

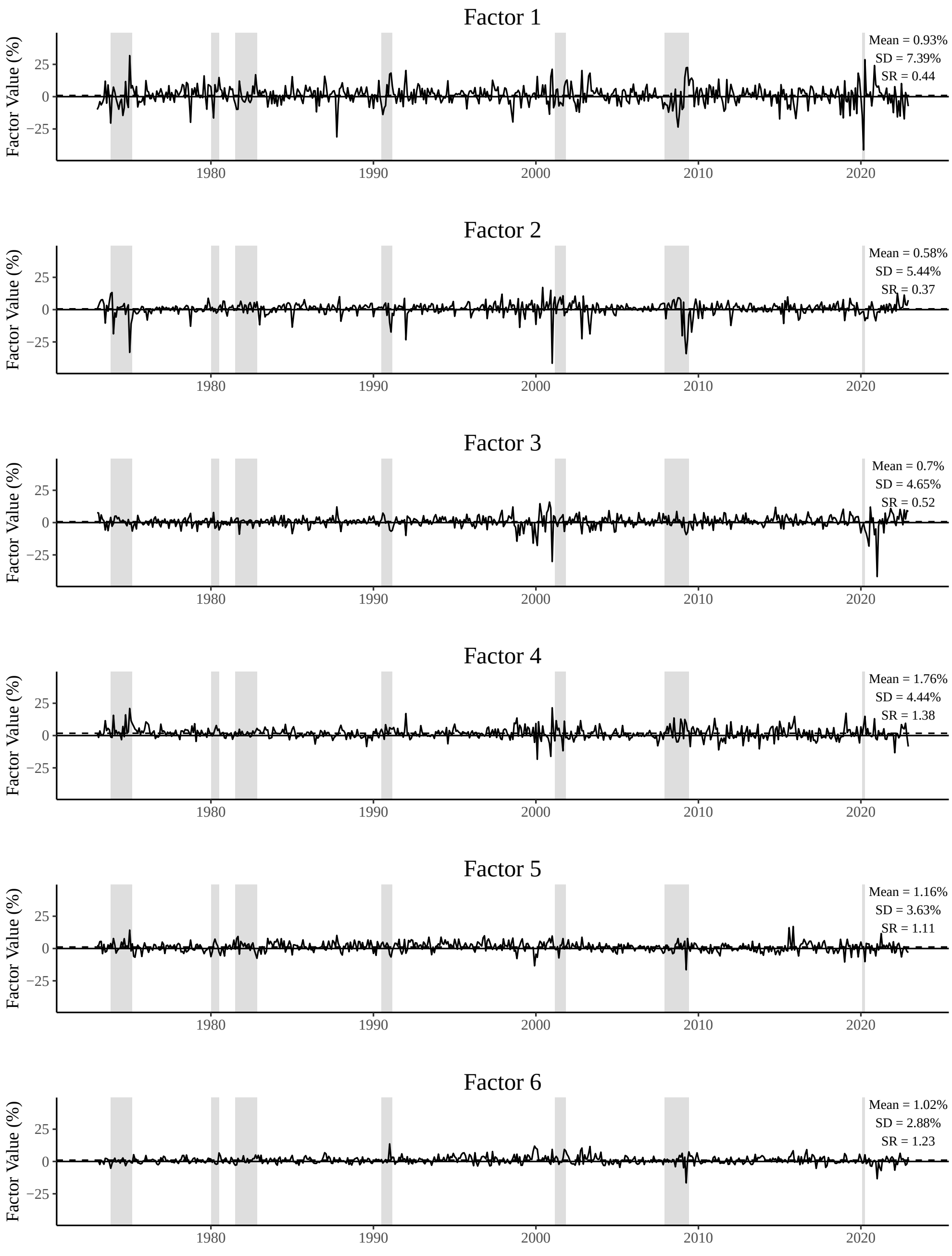
For comparison, we also present results for the PCA and IPCA factors (note that in the remainder of the paper, the PCA factors refer to PCs of the characteristic-managed portfolios). Panel B and C illustrate the structure of $\hat{b}_k$ these two types of non-sparse factors. In both cases, nearly all managed portfolios receive non-negligible weights, which complicates economic interpretation to some extent. Notably, the managed portfolios identified by the EGS factors also feature prominently in the PCA an IPCA factors, as indicated by the prominent bars in these figures. However, distinguishing their contributions from those of other portfolios remains challenging, underscoring the value of sparsity for improving interpretability.

The dynamics of the six EGS latent factors are depicted in the first six charts of Figure 6. A clear, progressive decline in variance is observed from Factor 1 to Factor 6, which aligns with the principles of PCA. The gray vertical bars in the figure mark the periods of recession as defined by the NBER. We find that these latent factors respond differently to economic recessions. For instance, during the 2007-2008 Great Recession, Factors 1, 2, 5, and 6 experienced substantial negative returns. In contrast, during the COVID-19 pandemic around 2020, only Factors 1 and 3 were notably impacted. This finding supports our argument that these latent factors capture different dimensions of risk.

3.3 Mean-Variance Efficiency

As discussed in Subsection 2.2, introducing a certain degree of sparsity improves the mean-variance efficiency and the ability to explain the cross-section of expected returns, albeit at the cost of some explanatory power for return comovements. In this subsection, we demonstrate that imposing economic-guided sparsity enhances the mean-variance efficiency of the factors.

Figures 5 provide visual evidence supporting this mechanism. For most non-sparse latent factors, positions with opposing contributions to the factor's risk premium each account for roughly half of the total. The effect of these opposing positions tends to offset each other, potentially limiting the factor's mean-variance efficiency. Some of these non-sparse factors may even be "unpriced": while they capture return comovements well, they fail to explain the cross-section of expected returns. By incorporating sparsity and economic sign-constraint, our EGS factors ensure that all managed portfolios within the factor contribute positively to the factor's risk premium. This structural alignment



Notes: the figure illustrates the dynamics of the six EGS factors. The dashed lines represent the sample mean levels of these factors. The gray vertical bars indicate the periods of economic recession as defined by the NBER. Additionally, the figure presents the values of the mean, standard deviation, and Sharpe ratio for each factor over the sample period.

Figure 6. Factor dynamics

facilitates an improvement in the factor's mean-variance efficiency.

To validate our argument, we compare our EGS factors with two types of non-sparse latent factors (PCA and IPCA factors) as well as typical observable factors. We evaluate the mean-variance efficiency of these factors by assessing their in-sample and out-of-sample Sharpe ratios.

3.3.1 Individual Factor Efficiency

A direct way to compare the mean-variance efficiency of different factors is by examining their Sharpe ratios. We begin by comparing the annualized in-sample Sharpe ratios of the EGS factors with those of other typical latent and observable factors. For the EGS and other latent factors, we focus on the factors extracted in the $K = 6$ case, consistent with the previous analysis. For observable factors, we consider the factors from the Fama-French three-, five-, and six-factor models (FF3, FF5, and FF6), as well as those from the Hou-Xue-Zhang q- and q^5-factor models (Q4 and Q5).

Table I summarizes the results. Compared to two types of non-sparse latent factors, our EGS factors generally exhibit a higher level of mean-variance efficiency. Specifically, across Factors 1 through 6, the EGS factors consistently achieve higher Sharpe ratios than both PCA and IPCA factors. The only exception is IPCA Factor 4, which has an in-sample Sharpe ratio of 2.93, significantly higher than that of EGS Factor 4 (1.37) and PCA Factor 4 (0.32). Figure 5 illustrates the reason for this occasional high Sharpe ratio: most of the managed portfolios in IPCA Factor 4 contribute positively to the factor's risk premium (i.e., position types (i) and (iv) dominate the factor).

In addition, PCA and IPCA sometimes extract unpriced factors, as evidenced by their extremely low Sharpe ratios. In the $K = 6$ case, both PCA Factor 2 and IPCA Factor 6 are typical examples of unpriced factors, with annualized Sharpe ratios of 0.03 and 0.17, respectively. Theoretically, such factors have little explanatory power for the cross-section of expected returns. By incorporating the economic sign-constraint, our EGS method ensures that all positions within the factor contribute positively to its risk premium, thus avoiding the extraction of unpriced factors. In the $K = 6$ case, all EGS factors exhibit relatively high Sharpe ratios, with the lowest being 0.37.

A comparison with several well-known observable factors further supports the superior mean-variance efficiency of our EGS factors. Most of these observable factors have Sharpe ratios ranging from 0.3 to 0.7, with only a few exceptions. Among our six EGS factors, the first three fall within this range, while the last three exhibit higher Sharpe ratios. In contrast, both PCA and IPCA occasionally produce unpriced factors with Sharpe ratios significantly below this range.

Table I. In-sample Sharpe ratios for individual factors

Panel A: Latent factors ($K = 6$)						
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
EGS	0.44	0.37	0.52	1.37	1.11	1.23
PCA	0.35	0.17	0.46	0.32	0.60	0.79
IPCA	0.38	0.33	0.34	2.93	0.44	0.03
Panel B: Observable factors						
	MKT	SMB	HML	RMW	CMA	MOM
FF3	0.43	0.17	0.36			
FF5	0.43	0.22	0.36	0.45	0.57	
FF6	0.43	0.22	0.36	0.45	0.57	0.50
	MKT	ME	IA	ROE	EG	
Q4	0.43	0.29	0.66	0.69		
Q5	0.43	0.29	0.66	0.69	1.30	

Notes: the table reports the in-sample (annualized) Sharpe ratios for individual factors, reflecting the mean-variance efficiency of each factor. The comparison includes: (i) EGS, PCA, and IPCA factors, each with $K = 6$; and (ii) observable factors from the Fama-French three-, five-, and six-factor models (FF3, FF5, and FF6), as well as the Hou-Xue-Zhang q- and q^5-factor models (Q4 and Q5).

3.3.2 Factor Tangency Portfolio

In addition to analyzing the in-sample Sharpe ratios of individual factors, we also evaluate the overall mean-variance efficiency of different sets of factors out-of-sample. For each set of factors, we construct an out-of-sample factor tangency portfolio and assess the overall mean-variance efficiency based on the portfolio's profitability, including its Sharpe ratio and maximum drawdown.

We construct the factor tangency portfolio on a monthly basis using a 10-year rolling-window. Specifically, at the end of each month t (for $t = 120, \cdots, T-1$), we apply our EGS method to the managed portfolios from months $t-119$ to t to reconstruct the factors. The rolling-window EGS factors and the corresponding factor weights are denoted as $\{\hat{f}_s^{(t)}\}_{s=t-119}^{t}$ and $\hat{B}_t$ (where the subscript or superscript t indicates that the rolling window extends up to time t). The $K \times 1$ weight vector for the factor tangency portfolio is then given by:

$$w_{t+1}^{TP} = \frac{\hat{\Sigma}_t^{-1}\hat{\mu}_t}{1'\hat{\Sigma}_t^{-1}\hat{\mu}_t}, \tag{10}$$

where $\hat{\mu}_t$ and $\hat{\Sigma}_t$ represent the mean vector and covariance matrix of the rolling-window EGS factors $\{\hat{f}_s^{(t)}\}_{s=t-119}^{t}$. This weight vector optimally allocates each factor within the tangency portfolio according to mean-variance principles. The return of this tangency portfolio at month $t+1$ is:

$$R_{t+1}^{TP} = \frac{\hat{f}_{t,t+1}'\hat{\Sigma}_t^{-1}\hat{\mu}_t}{1'\hat{\Sigma}_t^{-1}\hat{\mu}_t}. \tag{11}$$

where $\hat{f}_{t,t+1} := \left(\hat{B}_t'\hat{B}_t\right)^{-1}\hat{B}_t'\tilde{R}_{t+1}$ represents the realized factors, which are the time-$(t+1)$ realized values of EGS factors that are pre-constructed within the rolling window.

Table II presents the Sharpe ratios and maximum drawdowns for out-of-sample factor tangency portfolios constructed from different sets of factors. These factor sets include: (i) the EGS factors and other latent factors, across various values of K; and (ii) observable factors from several well-known factor models, including the CAPM, FF3, Q4, FF5, and FF6 models, corresponding to $K = 1, 3, 4, 5, 6$, respectively. To ensure comparability, all portfolio returns are scaled to match the standard deviation of the value-weighted market portfolio.

For latent factors, the out-of-sample Sharpe ratio generally increases with the number of factors K, indicating that the majority of these factors provide incremental explanatory power for the cross-section of expected returns. Among the different types of factors, EGS factors consistently outperforms the others. Across nearly all values of K, the EGS tangency portfolios achieve significantly higher Sharpe ratios and lower maximum drawdowns compared to PCA and IPCA tangency portfolios. Notably, when

$K \leq 3$, the out-of-sample Sharpe ratios of the PCA and IPCA tangency portfolios are disappointingly low, ranging from 0.19 to 0.28 for PCA and from 0.22 to 0.34 for IPCA, with exceptionally high maximum drawdowns. These results suggests that the first few factors extracted by PCA and IPCA tend to be unpriced out-of-sample.

Table II. Out-of-sample profitability of factor tangency portfolios

	K									
	1	2	3	4	5	6	7	8	9	10
Panel A: Latent factors										
EGS										
SR	0.30	0.70	1.08	1.34	1.91	2.04	2.46	2.67	2.94	3.30
MDD (%)	80.7	60.4	42.3	48.7	35.3	37.5	37.8	34.6	23.2	27.9
PCA										
SR	0.23	0.28	0.19	0.58	0.75	0.97	1.21	1.35	1.67	2.29
MDD (%)	112.1	64.7	83.4	68.4	82.1	79.4	94.8	87.2	65.2	89.4
IPCA										
SR	0.26	0.22	0.34	0.59	0.78	1.57	1.91	2.49	2.53	2.71
MDD (%)	85.6	91.9	70.7	54.4	78.2	51.3	62.4	63.0	66.7	51.9
Panel B: Observable factors										
	MKT		**FF3**	**Q4**	**FF5**	**FF6**				
SR	0.54		0.15	0.89	0.69	0.95				
MDD (%)	68.1		16.5	78.9	38.0	31.5				

Notes: the table reports the out-of-sample performance of tangency portfolios constructed from different types of factors. Out-of-sample performance is evaluated using the following indicators: (i) the out-of-sample annualized Sharpe ratio (SR) and (ii) the maximum drawdown (MDD, %). The comparison includes: (i) EGS, PCA, and IPCA factors, across various number of factors (K); and (ii) observable factors, including the CAPM, FF3, Q4, FF5, and FF6 models, corresponding to $K = 1, 3, 4, 5, 6$, respectively. To compute the maximum drawdown, all returns are scaled to match the standard deviation of the value-weighted market portfolio.

A comparison with observable factors provides additional insights. Since latent factors are constructed using a richer set of firm characteristics, their associated tangency portfolios are generally expected to achieve higher Sharpe ratios than those constructed from observable factors. This expectation is validated by EGS, as its tangency portfolios consistently exhibit higher Sharpe ratios than those from observable factors with the same K (with the exception of the $K = 1$ case[10]). For example, under a four-factor

[10]For $K = 1$, the CAPM employs the value-weighted market factor, whereas the EGS factor corresponds to the equal-weighted market factor. The higher Sharpe ratio of the CAPM tangency portfolio in this example reflects differences in factor construction rather than in factor structure.

specification, the EGS tangency portfolio achieves a Sharpe ratio of 1.34, compared to 0.89 for the Q4 tangency portfolio. In contrast, the PCA and IPCA tangency portfolios underperform observable factors in this case, with Sharpe ratios of 0.58 and 0.59, respectively.

3.4 Explaining the Cross-Section of Expected Returns

Factors with a high level of mean-variance efficiency should also exhibit a stronger ability to explain the cross-section of expected returns. In this subsection, we demonstrate the superior cross-sectional explanatory power of the EGS factors from three perspectives: 1) cross-sectional R^2 for managed portfolios; 2) the ability to price individual stocks; and 3) the ability to price test portfolios.

3.4.1 Time-series and Cross-Sectional Fit

As discussed in Subsection 2.2, our EGS factors sacrifice some explanatory power for time-series comovement to enhance their ability to explain the cross-section of expected returns. In this subsection, we examine this trade-off from the perspective of model fit. We employ two model fit metrics: the time-series R^2 and the cross-sectional R^2, as defined below:

$$\text{Time-series } R^2 = 1 - \frac{1}{p}\sum_{j=1}^{p}\frac{\sum_{t=1}^{T}\left(x_{t,j} - \hat{b}_j'\hat{f}_t\right)^2}{\sum_{t=1}^{T} x_{t,j}^2} \tag{12}$$

$$\text{Cross-sectional } R^2 = 1 - \frac{\sum_{j=1}^{p}\left(\frac{1}{T}\sum_{t=1}^{T}\left(x_{t,j} - \hat{b}_j'\hat{f}_t\right)\right)^2}{\sum_{j=1}^{p}\left(\frac{1}{T}\sum_{t=1}^{T} x_{t,j}\right)^2} \tag{13}$$

These two metrics evaluate the factors' ability to capture the time-series comovement of managed portfolios and their ability to explain the cross-section of expected returns, respectively. For EGS, PCA, and IPCA factors, the factor loadings correspond to the j-th row of the estimates $\hat{B}$, which are obtained from the respective algorithm. For observable factors, the factor loadings are estimated from time-series regressions of portfolios on the factors. We compute both in-sample and out-of-sample versions of these metrics, with the out-of-sample case involving realized factors and factor loadings estimated from the rolling window.

Table III presents our results. PCA factors exhibit the highest time-series R^2, followed by IPCA factors. Our EGS factors are weaker in this regard compared to the other two, reflecting the necessary trade-off for imposing sparsity. As for the cross-sectional R^2, our EGS factors outperform other latent and observable factors out-of-sample. This aligns with our argument: the imposition of sparsity sacrifices some ability to capture time-series comovement, while enhancing the ability to explain the cross-section of expected returns. Furthermore, compared to latent factors, observable factors have both lower time-series and cross-sectional R^2s.

3.4.2 Pricing Individual Stocks

Our EGS factors can also be described from the perspective of conditional factor models and, consequently, can be used to price individual stocks (see Appendix A). A set of factors with stronger explanatory power for individual stock returns will, in aggregate, generate smaller *alpha*s. To evaluate the out-of-sample pricing performance of these factors at the individual stock level, we construct a pure-*alpha* portfolio for each specification. The weight vector for this portfolio is:

$$w_{t+1}^{PA} = Z_t(Z_t'Z_t)^{-1}\hat{a}_t, \tag{14}$$

where $\hat{a}_t$ denotes the rolling window estimate of a (see Appendix A for details on the estimation of a). This portfolio allocates weights to individual stocks in proportion to their conditional expected returns in excess of risk compensation. The return of this pure-*alpha* portfolio at month $t+1$ is then given by:

$$R_{t+1}^{PA} = \hat{a}_t'(Z_t'Z_t)^{-1}Z_t'R_{t+1} = \hat{a}_t'X_{t+1}. \tag{15}$$

Thus, the pure-*alpha* portfolio can also be interpreted as a linear combination of the managed portfolios X_{t+1}, with weights given by $\hat{a}_t$.

Table IV compares the out-of-sample profitability of pure-*alpha* portfolios for different sets of factors. As the number of factors increases, the out-of-sample *alpha*s decrease, as reflected by the declining Sharpe ratios of the corresponding pure-*alpha* portfolios. Nevertheless, for $K \leq 10$, the Sharpe ratios of pure-*alpha* portfolios across all factor types remain above 3, indicating that a non-negligible portion of the conditional expected returns is consistently classified as mispricing. Compared to PCA and IPCA factors, our EGS factors yields considerably smaller out-of-sample *alpha*s, as evidenced by lower Sharpe ratios and higher maximum drawdowns. Furthermore, the *alpha*s associated with EGS factors diminish more rapidly as K increases, compared to those associated with PCA or IPCA factors. These results underscore the superior pricing ability of EGS factors at the individual stock level.

3.4.3 Pricing Test Portfolios

An efficient set of factors should effectively price not only individual stocks but also test portfolios. Test portfolios, which are typically diversified combinations of stocks designed to capture a broad cross-section of expected returns, provide a robust benchmark for

Table III. Time-series and cross-sectional R^2 (%)

	K									
	1	2	3	4	5	6	7	8	9	10
Panel A: Time-series R^2 (%), in-sample										
EGS	1.89	3.77	5.66	8.29	11.75	13.61	15.33	17.18	19.25	21.28
PCA	8.64	13.09	17.64	20.56	24.12	27.17	29.18	31.95	35.06	38.49
IPCA	3.68	8.17	13.73	17.35	20.24	22.89	26.17	28.60	30.48	33.17
Observable	4.41		9.15	10.07	11.28	13.36				
Panel B: Time-series R^2 (%), out-of-sample										
EGS	2.08	4.26	6.73	9.10	11.30	13.43	15.07	16.97	18.70	20.57
PCA	7.72	11.21	14.73	17.57	19.78	22.06	24.05	25.90	27.87	29.61
IPCA	2.51	7.53	11.00	12.88	15.06	16.89	18.68	20.57	22.46	24.37
Observable	3.70		5.36	4.69	4.34	4.91				
Panel C: Cross-sectional R^2 (%), in-sample										
EGS	3.87	5.36	7.13	21.04	27.36	32.12	42.84	44.07	45.20	61.11
PCA	4.12	4.49	6.22	6.90	8.84	11.78	12.68	12.78	28.82	61.03
IPCA	1.30	1.43	2.61	3.00	33.68	41.86	58.62	61.31	60.74	62.72
Observable	2.92		5.09	5.80	6.84	4.41				
Panel D: Cross-sectional R^2 (%), out-of-sample										
EGS	3.90	8.57	18.07	24.13	34.84	40.44	44.59	48.83	54.39	57.80
PCA	3.87	5.08	6.13	11.06	13.04	15.54	20.27	23.95	28.12	39.08
IPCA	1.63	5.19	6.91	10.64	14.72	25.90	33.28	41.88	46.56	50.11
Observable	0.64		1.83	1.93	4.43	5.23				

Notes: the table presents the time-series and cross-sectional R^2s for different types of factors, considering both in-sample and out-of-sample cases. The comparison includes: (i) EGS, PCA, and IPCA factors, across various number of factors (K); and (ii) observable factors, including the CAPM, FF3, Q4, FF5, and FF6 models, corresponding to $K = 1, 3, 4, 5, 6$, respectively.

Table IV. Out-of-sample profitability of pure-*alpha* portfolios

	K									
	1	2	3	4	5	6	7	8	9	10
EGS										
SR	4.09	4.04	3.91	3.88	3.87	3.76	3.69	3.59	3.32	3.24
MDD (%)	26.3	31.3	39.8	42.2	34.0	15.0	15.8	17.0	26.9	20.5
PCA										
SR	4.54	4.51	4.60	4.60	4.57	4.51	4.46	4.42	4.37	4.17
MDD (%)	25.3	34.0	18.0	18.0	9.8	7.2	7.3	6.6	10.0	6.8
IPCA										
SR	4.45	4.42	4.48	4.42	4.37	4.35	4.19	4.04	4.02	4.02
MDD (%)	22.4	10.9	9.1	11.3	14.8	12.5	12.7	9.0	8.6	8.9

Notes: the table reports the out-of-sample performance of pure-*alpha* portfolios. Out-of-sample performance is evaluated using the following indicators: (i) the out-of-sample annualized Sharpe ratio (SR) and (ii) the maximum drawdown (MDD, %). The comparison includes pure-*alpha* portfolios corresponding to EGS, PCA, and IPCA factors, across various values of K. To compute the maximum drawdown, all returns are scaled to match the standard deviation of the value-weighted market portfolio.

assessing the generalization power of factor models. In this subsection, we compare the ability of our EGS factors to explain test portfolios relative to other types of factors. Specifically, we consider three sets of test portfolios: (i) the 53 managed portfolios from our study; (ii) Fama and French (2018) 100 sorted portfolios, including 5×5 portfolios sorted by size/book-to-market, size/profitability, size/investment, and size/momentum, respectively; and (iii) Kozak et al. (2018) 110 sorted portfolios, including decile-1 and decile-10 portfolios based on 55 characteristics.

Following Lettau and Pelger (2020b), we employ the in-sample and out-of-sample root-mean-squared (RMS) *alpha* (RMS_α^{IS} and RMS_α^{OOS}) to evaluate the overall pricing errors. For this purpose, we first calculate the in-sample and out-of-sample *alpha* for each test portfolio m ($m = 1, \cdots, M$). The in-sample *alpha* α_m^{IS} is computed as the intercept from the regression of test portfolio m on the selected factors. The out-of-sample *alpha* α_m^{OOS} is computed in three steps: (i) within a rolling window (from months $t - 119$ to t), we construct the EGS factors and compute the loadings of test portfolio m on these factors; (ii) we calculate the realized value of the factors at month $t+1$, and use them to calculate the out-of-sample residual for test portfolio m at month $t+1$; and (iii) the out-of-sample *alpha* is then computed as the time-series mean of the out-of-sample residuals. Finally, we compute the in-sample and out-of-sample RMS *alpha* for the cross-section of test portfolios, which represents the overall pricing error.

Table V compares the explanatory power of different factors for three types of test portfolios from an in-sample perspective. The relative performance of these factors varies with the number of factors K. When $K \leq 5$, the four types of factors perform similarly in explaining our managed portfolios; for the sorted portfolios, the EGS factors exhibit the strongest explanatory power, followed by observable factors, IPCA factors, and PCA factors. However, when $K \geq 6$, IPCA factors show a sharp improvement in explanatory power, surpassing the other three types of factors. This may be because, when $K \geq 6$, IPCA extracts a few factors with extremely high Sharpe ratios, as discussed in Subsection 3.3[11]

From the out-of-sample perspective, our EGS factors demonstrate a clear advantage. In explaining our managed portfolios, the EGS factors perform best, followed by IPCA factors, PCA factors, and finally observable factors. For the two types of sorted portfolios, the advantage of EGS factors is even more pronounced: their out-of-sample RMS *alpha* is roughly half, or even less, of that of PCA and IPCA factors. Observable factors lie in between. Interestingly, the advantage of IPCA factors in the in-sample analysis does not persist out-of-sample—IPCA significantly underperforms both EGS and observable factors in explaining the sorted portfolios. This suggests that the extremely high Sharpe ratio factors occasionally extracted by IPCA do not generalize well out-of-sample.

4 General EGS Framework

We extend the baseline EGS factors to a more general EGS framework that incorporates economic objectives, in line with the spirit of Bryzgalova et al. (2023):

$$\min_{A,B\in\mathbf{R}^{p\times K}} -\frac{2}{T}\text{tr}\left[A'X'(M_T+\gamma W)XB\right] + \|B\|_F^2 + \sum_{k=1}^{K}\lambda_k\|b_k\|_1 \text{ s.t. } A'A = I_K. \quad (16)$$

In this specification, the covariance matrix $\hat{\Sigma} = X'M_TX/T$ in (5) is generalized to $X'(M_T+\gamma W)X/T$. The matrix W modifies the covariance structure to reflect specific economic objectives, and γ determines the weight placed on these objectives. As before, the parameters $\lambda_1, \ldots, \lambda_K$ govern the degree of sparsity, and we continue to impose the economic sign constraint to guide their selection. Setting $W = 0$ recovers the baseline EGS criterion in (5). Different choices of W integrate the EGS framework with different economic objectives. For instance, $W = 1_T 1_T'/T$ integrates EGS with the Risk-Premium PCA (RP-PCA) of Lettau and Pelger (2020b), while $W = G(G'G)^{-1}G'$, where G is a $T \times L$ matrix of macroeconomic variables, emphasizes return components associated

[11]Note that the IPCA factors vary considerably under different specifications. For example, in the $K = 6$ case reported in Table I, IPCA Factor 4 exhibits an extremely high Sharpe ratio, whereas this is not observed in the $K = 4$ case.

Table V. In-sample RMS *alpha* for test portfolios

	K									
	1	2	3	4	5	6	7	8	9	10
Panel A: Our 53 Managed Portfolios (%)										
EGS	0.64	0.63	0.63	0.58	0.55	0.54	0.49	0.49	0.48	0.41
PCA	0.64	0.64	0.63	0.63	0.62	0.61	0.61	0.61	0.55	0.41
IPCA	0.64	0.64	0.64	0.64	0.52	0.49	0.40	0.39	0.39	0.38
Observable	0.64		0.63	0.63	0.63	0.64				
Panel B: FF 100 Sorted Portfolios (%)										
EGS	0.57	0.49	0.58	0.24	0.26	0.26	0.39	0.37	0.21	0.18
PCA	0.72	0.82	0.87	0.79	0.86	0.95	0.98	0.98	1.03	0.48
IPCA	0.63	0.62	0.66	0.58	0.35	0.18	0.31	0.25	0.28	0.22
Observable	0.54		0.42	0.39	0.40	0.40				
Panel C: KNS 110 Sorted Portfolios (%)										
EGS	0.46	0.39	0.51	0.35	0.28	0.27	0.33	0.34	0.30	0.25
PCA	0.63	0.66	0.74	0.64	0.68	0.76	0.79	0.78	0.86	0.45
IPCA	0.57	0.56	0.60	0.58	0.45	0.26	0.28	0.27	0.30	0.28
Observable	0.46		0.43	0.45	0.43	0.44				

Notes: the table reports the in-sample root-mean-squared (RMS) *alpha* when different types of factors are used to price various test portfolios. Three sets of test portfolios are considered: (i) the 53 managed portfolios from our study; (ii) Fama and French (2018) 100 sorted portfolios, including 5×5 portfolios sorted by size/book-to-market, size/profitability, size/investment, and size/momentum, respectively; and (iii) Kozak et al. (2018) 110 sorted portfolios, including decile-1 and decile-10 portfolios based on 55 characteristics. The comparison includes: (i) EGS, PCA, and IPCA factors, across various number of factors (K); and (ii) observable factors, including the CAPM, FF3, Q4, FF5, and FF6 models, corresponding to $K = 1, 3, 4, 5, 6$, respectively.

Table VI. Out-of-sample RMS *alpha* for test portfolios

	K									
	1	2	3	4	5	6	7	8	9	10
Panel A: Our 53 Managed Portfolios (%)										
EGS	0.61	0.60	0.57	0.55	0.51	0.48	0.47	0.45	0.42	0.41
PCA	0.61	0.61	0.61	0.59	0.58	0.58	0.56	0.55	0.53	0.49
IPCA	0.62	0.61	0.60	0.59	0.58	0.54	0.52	0.48	0.46	0.45
Observable	0.62		0.62	0.62	0.61	0.61				
Panel B: FF 100 Sorted Portfolios (%)										
EGS	0.63	0.42	0.31	0.27	0.26	0.23	0.23	0.22	0.23	0.25
PCA	0.80	0.87	0.83	0.79	0.81	0.76	0.73	0.74	0.70	0.71
IPCA	0.60	0.67	0.74	0.67	0.64	0.57	0.53	0.47	0.48	0.45
Observable	0.37		0.35	0.35	0.33	0.34				
Panel C: KNS 110 Sorted Portfolios (%)										
EGS	0.53	0.36	0.31	0.30	0.31	0.31	0.33	0.34	0.35	0.37
PCA	0.69	0.80	0.75	0.73	0.74	0.68	0.67	0.68	0.67	0.71
IPCA	0.54	0.64	0.67	0.64	0.61	0.55	0.49	0.41	0.44	0.41
Observable	0.36		0.36	0.37	0.35	0.35				

Notes: the table reports the out-of-sample root-mean-squared (RMS) *alpha* when different types of factors are used to price various test portfolios. Three sets of test portfolios are considered: (i) the 53 managed portfolios from our study; (ii) Fama and French (2018) 100 sorted portfolios, including 5×5 portfolios sorted by size/book-to-market, size/profitability, size/investment, and size/momentum, respectively; and (iii) Kozak et al. (2018) 110 sorted portfolios, including decile-1 and decile-10 portfolios based on 55 characteristics. The comparison includes: (i) EGS, PCA, and IPCA factors, across various number of factors (K); and (ii) observable factors, including the CAPM, FF3, Q4, FF5, and FF6 models, corresponding to $K = 1, 3, 4, 5, 6$, respectively.

with macroeconomic shocks.

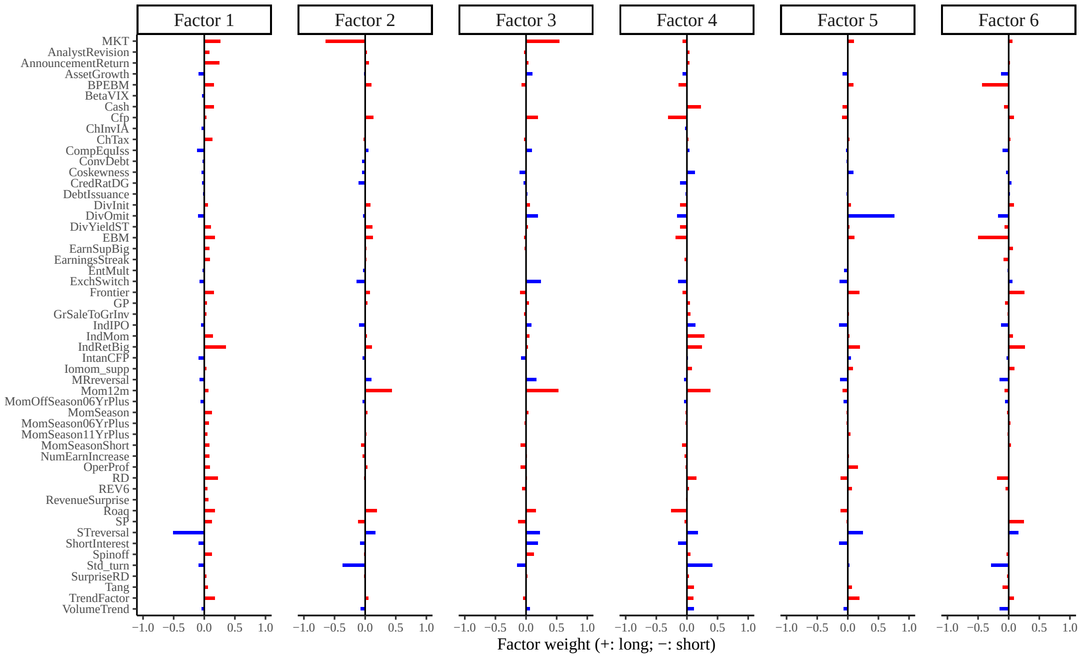
As a primary application, we integrate EGS with RP-PCA and refer to the resulting framework as RP-EGS. While traditional PCA relies solely on second-moment information (the covariance matrix), RP-PCA incorporates both first- and second-moment information by including expected returns. This extension yields factors with improved mean-variance efficiency and stronger explanatory power for the cross-section of expected returns. We examine whether integrating EGS into the RP-PCA framework further enhances factor performance. Consistent with Lettau and Pelger (2020b), we set $\gamma = 10$.

Figure 7 illustrates the composition of the RP-PCA and RP-EGS factors. RP-PCA Factor 1 takes long positions in all managed portfolios with positive expected returns and short positions in those with negative expected returns; in other words, every managed portfolio contributes positively to the factor risk premium. This pattern reflects the incorporation of first-moment information. Consequently, the EGS framework imposes minimal sparsity on this factor, so the composition of RP-EGS Factor 1 remains dense and closely mirrors that of RP-PCA Factor 1. By contrast, RP-PCA Factors 2-6 exhibit roughly equal numbers of portfolios contributing positively and negatively to the risk premium, indicating that first-moment information is relatively limited in these factors. Our economically guided sparsity primarily affects these factors, rendering RP-EGS Factors 2-6 sparse. The economic interpretations of RP-EGS Factors 2-6 closely align with those of EGS Factors 1-5, underscoring the robustness of our EGS approach.

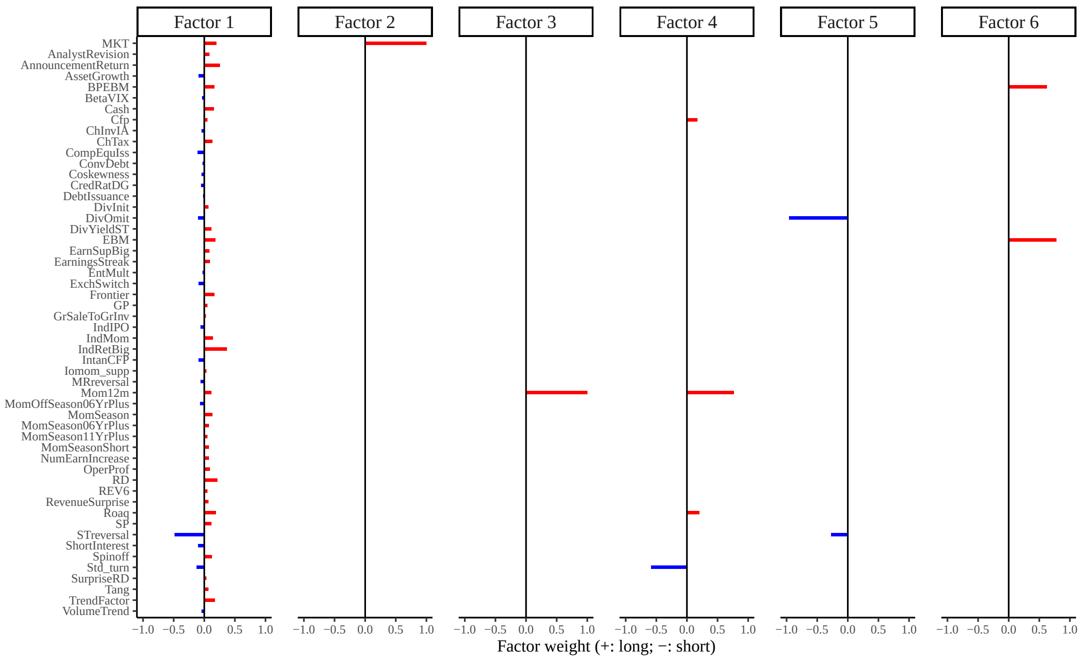
We further evaluate the asset pricing performance of RP-PCA and RP-EGS factors by examining their out-of-sample Sharpe ratios, cross-sectional R^2, and their explanatory power for the test assets. The results are reported in Tables VII, VIII, and IX. By effectively exploiting first-moment information, the RP-PCA factors achieve exceptionally high Sharpe ratios, with the first two factors alone reaching a Sharpe ratio of 4.61. The strong fit on the first moment also translates into a remarkable ability to explain the cross-section of managed portfolios: starting from the first factor, the out-of-sample cross-sectional R^2 exceeds 95%, and the root-mean-square *alpha* remains below 0.15%. These results substantially outperform those of PCA factors and baseline EGS factors. The RP-EGS factors preserve these advantages of the RP-PCA factors.

However, the RP-PCA factors perform poorly when explaining other test assets. Although RP-PCA Factor 1 exhibits an extremely high Sharpe ratio and plays a dominant role in explaining managed portfolios, it negatively contributes to the explanation of other test assets. Specifically, the expected returns of Fama and French (2018) 100 sorted portfolios and Kozak et al. (2018) 110 sorted portfolios are 1.08% and 1.01%, respectively, yet using RP-PCA Factor 1 to explain these returns yields *alphas* exceeding 2%. This indicates that the RP-PCA factors overfit the baseline assets' expected returns. Even after incorporating additional factors, the performance of RP-PCA factors

Panel A: RP-PCA Factors



Panel B: RP-EGS Factors



Notes: the figure illustrates the values of $\hat{b}_k$ (the kth column of $\hat{B}$ for $k = 1, \cdots, 6$) corresponding to the RP-PCA and RP-EGS factors. These values represent the weights assigned to each characteristic-managed portfolio for each factor. The colors indicate the sign of the expected returns for the characteristic-managed portfolios: red denotes positive expected returns, while blue denotes negative expected returns.

Figure 7. Composition of RP-PCA and RP-EGS factors

Table VII. RP-PCA + EGS: Out-of-sample Portfolio Profitability

	K									
	1	2	3	4	5	6	7	8	9	10
PCA										
SR	0.23	0.28	0.19	0.58	0.75	0.97	1.21	1.35	1.67	2.29
MDD (%)	112.1	64.7	83.4	68.4	82.1	79.4	94.8	87.2	65.2	89.4
EGS										
SR	0.30	0.70	1.08	1.34	1.91	2.04	2.46	2.67	2.94	3.30
MDD (%)	80.7	60.4	42.3	48.7	35.3	37.5	37.8	34.6	23.2	27.9
RP-PCA ($\gamma = 10$)										
SR	3.20	4.61	4.52	4.75	4.77	4.80	4.86	4.83	4.81	4.79
MDD (%)	85.8	43.9	55.8	35.9	37.5	31.0	23.3	20.1	17.1	20.5
RP-EGS ($\gamma = 10$)										
SR	4.08	4.13	4.19	4.16	4.24	4.52	4.43	4.52	4.64	4.65
MDD (%)	50.4	41.3	50.5	47.5	52.2	31.7	32.6	18.9	18.5	16.5

Notes: the table reports the out-of-sample performance of tangency portfolios constructed from different types of factors. Out-of-sample performance is evaluated using the following indicators: (i) the out-of-sample annualized Sharpe ratio (SR) and (ii) the maximum drawdown (MDD, %). The comparison includes PCA factors, RP-PCA factors, and their combinations with the EGS framework (EGS and RP-EGS factors). Following Lettau and Pelger (2020b), we set the parameter $\gamma = 10$. To compute the maximum drawdown, all returns are scaled to match the standard deviation of the value-weighted market portfolio.

Table VIII. RP-PCA + EGS: Time-series and cross-sectional R^2 (%)

	K									
	1	2	3	4	5	6	7	8	9	10
Panel A: Time-series R^2 (%), in-sample										
PCA	8.64	13.09	17.64	20.56	24.12	27.17	29.18	31.95	35.06	38.49
EGS	1.89	3.77	5.66	8.29	11.75	13.61	15.33	17.18	19.25	21.28
RP-PCA	6.74	15.98	20.32	24.72	27.83	31.16	33.71	35.93	38.70	40.79
RP-EGS	7.01	9.27	11.10	14.16	16.26	19.33	20.89	21.16	23.00	24.87
Panel B: Time-series R^2 (%), out-of-sample										
PCA	7.72	11.21	14.73	17.57	19.78	22.06	24.05	25.90	27.87	29.61
EGS	2.08	4.26	6.73	9.10	11.30	13.43	15.07	16.97	18.70	20.57
RP-PCA	6.60	14.85	18.17	21.70	24.21	26.36	28.84	30.55	32.24	34.18
RP-EGS	6.77	9.14	10.96	13.55	15.91	17.80	19.54	20.73	22.79	24.52
Panel C: Cross-sectional R^2 (%), in-sample										
PCA	4.12	4.49	6.22	6.90	8.84	11.78	12.68	12.78	28.82	61.03
EGS	3.87	5.36	7.13	21.04	27.36	32.12	42.84	44.07	45.20	61.11
RP-PCA	98.80	99.97	99.97	99.98	99.98	99.98	99.99	99.99	99.99	99.99
RP-EGS	99.99	100.00	100.00	100.00	99.98	99.98	99.99	99.93	99.97	99.93
Panel D: Cross-sectional R^2 (%), out-of-sample										
PCA	3.87	5.08	6.13	11.06	13.04	15.54	20.27	23.95	28.12	39.08
EGS	3.90	8.57	18.07	24.13	34.84	40.44	44.59	48.83	54.39	57.80
RP-PCA	94.86	97.61	97.56	97.66	97.77	97.82	97.77	97.76	97.97	97.99
RP-EGS	96.95	97.45	97.38	97.94	97.79	97.72	97.91	97.87	98.04	97.88

Notes: the table presents the time-series and cross-sectional R^2s for different types of factors, considering both in-sample and out-of-sample cases. The comparison includes PCA factors, RP-PCA factors, and their combinations with the EGS framework (EGS and RP-EGS factors). Following Lettau and Pelger (2020b), we set the parameter $\gamma = 10$.

Table IX. RP-PCA + EGS: Out-of-sample RMS *alpha* for test portfolios

	K									
	1	2	3	4	5	6	7	8	9	10
	Panel A: Our 53 Managed Portfolios (%)									
PCA	0.61	0.61	0.61	0.59	0.58	0.58	0.56	0.55	0.53	0.49
EGS	0.61	0.60	0.57	0.55	0.51	0.48	0.47	0.45	0.42	0.41
RP-PCA	0.14	0.10	0.10	0.10	0.09	0.09	0.09	0.09	0.09	0.09
RP-EGS	0.11	0.10	0.10	0.09	0.09	0.09	0.09	0.09	0.09	0.09
	Panel B: FF 100 Sorted Portfolios (%)									
PCA	0.80	0.87	0.83	0.79	0.81	0.76	0.73	0.74	0.70	0.71
EGS	0.63	0.42	0.31	0.27	0.26	0.23	0.23	0.22	0.23	0.25
RP-PCA	2.15	1.03	0.65	0.60	0.64	0.63	0.63	0.67	0.66	0.62
RP-EGS	2.24	0.52	0.37	0.45	0.43	0.39	0.46	0.52	0.52	0.58
	Panel C: KNS 110 Sorted Portfolios (%)									
PCA	0.69	0.80	0.75	0.73	0.74	0.68	0.67	0.68	0.67	0.71
EGS	0.53	0.36	0.31	0.30	0.31	0.31	0.33	0.34	0.35	0.37
RP-PCA	2.36	1.13	0.77	0.69	0.68	0.66	0.65	0.70	0.68	0.66
RP-EGS	2.41	0.61	0.41	0.54	0.48	0.45	0.54	0.61	0.62	0.69

Notes: the table reports the out-of-sample root-mean-squared (RMS) *alpha* when different types of factors are used to price various test portfolios. Three sets of test portfolios are considered: (i) the 53 managed portfolios from our study; (ii) Fama and French (2018) 100 sorted portfolios, including 5×5 portfolios sorted by size/book-to-market, size/profitability, size/investment, and size/momentum, respectively; and (iii) Kozak et al. (2018) 110 sorted portfolios, including decile-1 and decile-10 portfolios based on 55 characteristics. The comparison includes PCA factors, RP-PCA factors, and their combinations with the EGS framework (EGS and RP-EGS factors). Following Lettau and Pelger (2020b), we set the parameter $\gamma = 10$.

on other test assets is only comparable to that of PCA factors and remains well below that of EGS factors. By contrast, the use of RP-EGS factors significantly improves the explanatory power of RP-PCA factors for other test assets.

5 Conclusion

We develop an EGS factor framework that bridges the gap between latent and observable factor approaches in empirical asset pricing. While traditional latent factor approaches such as PCA and IPCA excel at capturing return comovement, they often lack economic interpretability and deliver “unpriced” factors. Observable factors, in contrast, offer more economic interpretability but may fail to span the full return space and suffer from high dimensionality. By introducing sparsity into PCA with economic guidance—favoring long (short) positions in characteristic-managed portfolios with positive (negative) expected returns—we construct factors that are both interpretable and statistically robust. Our EGS factors strike a compelling balance among three key dimensions: economic interpretability, the ability to capture time-series co-movement, and explanatory power for the cross-section of expected returns.

Empirically, in the U.S. equity market, the first EGS factor corresponds to the equal-weighted market portfolio, while the remaining factors align with momentum, liquidity-profitability, short-term reversal/dividend omission, value, and R&D themes. They lie closer to the mean-variance efficient frontier, deliver higher Sharpe ratios, and explain a larger share of cross-sectional variation with fewer, more meaningful characteristics. In addition, these factors exhibit differential sensitivities to economic recessions, suggesting that they capture distinct dimensions of systematic risk. Our EGS factors produce significantly lower pricing errors in out-of-sample evaluations, whether applied to individual stocks or classic test portfolios.

Our approach mitigates overfitting in both factor construction and risk premium estimation, offering a unified and economically disciplined lens on the stochastic discount factor. This joint-sparsity framework advances the goal of building a low-dimensional, economically meaningful representation of asset pricing.

References

Asness, C. S., T. J. Moskowitz, and L. H. Pedersen (2013): "Value and momentum everywhere," *The journal of finance*, 68, 929–985.

Bryzgalova, S., V. DeMiguel, S. Li, and M. Pelger (2023): "Asset-Pricing Factors with Economic Targets," *Available at SSRN 4344837*.

Campbell, J. Y. and S. B. Thompson (2008): "Predicting excess stock returns out of sample: Can anything beat the historical average?" *The Review of Financial Studies*, 21, 1509–1531.

Carhart, M. M. (1997): "On persistence in mutual fund performance," *The Journal of finance*, 52, 57–82.

Chan, L. K., J. Lakonishok, and T. Sougiannis (2001): "The stock market valuation of research and development expenditures," *The Journal of finance*, 56, 2431–2456.

Chen, A. Y. and T. Zimmermann (2022): "Open Source Cross-Sectional Asset Pricing," *Critical Finance Review*, 11, 207–264.

Chen, Q., N. Roussanov, and X. Wang (2021): "Semiparametric Conditional Factor Models: Estimation and Inference," Tech. rep., arXiv preprint arXiv:2112.07121.

Cochrane, J. H. (2011): "Presidential address: Discount rates," *The Journal of finance*, 66, 1047–1108.

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

——— (1988): "Risk and return in an equilibrium APT: Application of a new test methodology," *Journal of financial economics*, 21, 255–289.

Da, Z., Q. Liu, and E. Schaumburg (2014): "A closer look at the short-term return reversal," *Management science*, 60, 658–674.

Daniel, K. and S. Titman (1997): "Evidence on the characteristics of cross sectional variation in stock returns," *the Journal of Finance*, 52, 1–33.

Ehsani, S. and J. T. Linnainmaa (2022): "Factor momentum and the momentum factor," *The Journal of Finance*, 77, 1877–1919.

Fama, E. F. (1976): *Foundations of Finance*, New York: Basic Books.

Fama, E. F. and K. R. French (1992): "The cross-section of expected stock returns," *the Journal of Finance*, 47, 427–465.

——— (1993): "Common risk factors in the returns on stocks and bonds," *Journal of financial economics*, 33, 3–56.

——— (2015): "A five-factor asset pricing model," *Journal of financial economics*, 116, 1–22.

——— (2018): "Choosing factors," *Journal of financial economics*, 128, 234–252.

——— (2020): "Comparing cross-section and time-series factor models," *The Review of*

Financial Studies, 33, 1891–1926.

Fan, J., Z. T. Ke, Y. Liao, and A. Neuhierl (2022): "Structural deep learning in conditional asset pricing," *Available at SSRN 4117882*.

Green, J., J. R. Hand, and X. F. Zhang (2017): "The characteristics that provide independent information about average US monthly stock returns," *The Review of Financial Studies*, 30, 4389–4436.

Hoberg, G. and I. Welch (2009): "Optimized vs. sort-based portfolios," Tech. rep., AFA 2010 Atlanta Meetings Paper.

Hong, H. and J. C. Stein (1999): "A unified theory of underreaction, momentum trading, and overreaction in asset markets," *The Journal of finance*, 54, 2143–2184.

Jegadeesh, N. and S. Titman (1993): "Returns to buying winners and selling losers: Implications for stock market efficiency," *The Journal of finance*, 48, 65–91.

Johnstone, I. M. and A. Y. Lu (2009): "On consistency and sparsity for principal components analysis in high dimensions," *Journal of the American Statistical Association*, 104, 682–693.

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

Kirby, C. (2020): "Firm characteristics, cross-sectional regression estimates, and asset pricing tests," *The Review of Asset Pricing Studies*, 10, 290–334.

Kozak, S. and S. Nagel (2023): "When do cross-sectional asset pricing factors span the stochastic discount factor?" Tech. rep., National Bureau of Economic Research.

Kozak, S., S. Nagel, and S. Santosh (2018): "Interpreting factor models," *The Journal of Finance*, 73, 1183–1223.

——— (2020): "Shrinking the cross-section," *Journal of Financial Economics*, 135, 271–292.

Lee, C. M. and B. Swaminathan (2000): "Price momentum and trading volume," *the Journal of Finance*, 55, 2017–2069.

Leng, C. and H. Wang (2009): "On general adaptive sparse principal component analysis," *Journal of Computational and Graphical Statistics*, 18, 201–215.

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

——— (2020b): "Factors that fit the time series and cross-section of stock returns," *The Review of Financial Studies*, 33, 2274–2325.

Liu, L. X. and L. Zhang (2008): "Momentum profits, factor pricing, and macroeconomic risk," *The Review of Financial Studies*, 21, 2417–2448.

Nagel, S. (2012): "Evaporating liquidity," *The Review of Financial Studies*, 25, 2005–2039.

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

Pelger, M. and R. Xiong (2022): "Interpretable sparse proximate factors for large dimensions," *Journal of Business & Economic Statistics*, 40, 1642–1664.

Penman, S. H., S. A. Richardson, and I. Tuna (2007): "The book-to-price effect in stock returns: accounting for leverage," *Journal of accounting research*, 45, 427–467.

Ross, S. (1976): "The arbitrage pricing theory," *Journal of Economic Theory*, 13, 341–360.

Zou, H., T. Hastie, and R. Tibshirani (2006): "Sparse principal component analysis," *Journal of computational and graphical statistics*, 15, 265–286.

Appendix A A Conditional Factor Model Perspective

We can also describe the EGS factors from the perspective of a conditional factor model. Consider the following conditional latent factor model for individual stock returns:

$$r_{i,t} = \alpha_{i,t-1} + \beta'_{i,t-1} f_t + e_{i,t},\ i = 1, \dots, N, t = 1, \dots, T, \tag{A.1}$$

$$\alpha_{i,t-1} = a' z_{i,t-1} + e_{\alpha,i,t-1},\ \beta_{i,t-1} = B' z_{i,t-1} + e_{\beta,i,t-1}, \tag{A.2}$$

where $r_{i,t}$ is the excess return of stock i in month t, and $z_{i,t-1}$ is a $p \times 1$ vector of characteristics which includes a constant term. The model describes a conditional factor structure of the return by leveraging characteristics. Specifically, (A.1) decomposes the return into three components: i) the time-varying pricing error $\alpha_{i,t-1}$; ii) the time-varying risk compensation $\beta_{i,t-1} f_t$, where f_t is a $K \times 1$ vector of latent risk factors and $\beta_{i,t-1}$ is a $K \times 1$ vector of risk exposures; iii) the idiosyncratic noise $e_{i,t}$. (A.2) decomposes the pricing error and risk exposures into two parts: i) the informative parts $a' z_{i,t-1}$ and $B' z_{i,t-1}$ that are explained by the characteristics, where a and B are $p \times 1$ vector and $p \times K$ matrix of unknown coefficients; ii) the noisy parts $e_{\alpha,i,t-1}$ and $e_{\beta,i,t-1}$ that are not accounted for by the characteristics, i.e., they are orthogonal to the characteristics. Consequently, the model allows for distinguishing between the "risk" and "mispricing" explanations of the role of characteristics in predicting returns, thereby resolving the ongoing "characteristics versus covariance" debate (Daniel and Titman, 1997).

By substituting (A.2) into (A.1), we obtain the following expression:

$$r_{i,t} = z'_{i,t-1} a + z'_{i,t-1} B f_t + \varepsilon_{i,t}. \tag{A.3}$$

Here, $\varepsilon_{i,t} \equiv e_{i,t} + e_{\alpha,i,t-1} + e'_{\beta,i,t-1} f_t$ represents the composite noise. Let $R_t \equiv (r_{1,t}, \cdots, r_{N,t})'$, $Z_{t-1} \equiv (z_{1,t-1}, \cdots, z_{N,t-1})'$, and $\mathcal{E}_t \equiv (\varepsilon_{1,t}, \cdots, \varepsilon_{N,t})'$ be the stacked vector/martrix of returns, characteristics, and composite noises of all stocks in month t, which collectively form one cross-section. Then we may write (A.3) in a matrix form:

$$R_t = Z_{t-1} a + Z_{t-1} B f_t + \mathcal{E}_t. \tag{A.4}$$

Applying Fama-MacBeth regressions to (A.4) allows us to transform the conditional factor model into an unconditional form:

$$X_t = (Z'_{t-1} Z_{t-1})^{-1} Z'_{t-1} R_t = a + B f_t + \mathcal{E}^*_{t+1}, \tag{A.5}$$

where $\mathcal{E}^*_{t+1} := (Z'_t Z_t)^{-1} Z'_t \mathcal{E}_{t+1}$. Since $\mathcal{E}^*_{t+1}$ is negligible due to the law of large numbers, performing PCA on X_t provides an estimation of the conditional factor model. This in turn allows us to enhance the estimation by using our EGS factors.

From the perspective of conditional factor models, in addition to estimating the EGS factors $f_1, \cdots, f_T$ and factor weights $\hat{B}$ (see Section 2), we can also estimate the pricing error coefficient a:

Since $\bar{X} \approx a + B\bar{f}$ where $\bar{f} = \sum_{t=1}^{T} f_t/T$, imposing $a'B = 0$ allows us to estimate a as:

$$\hat{a} = (I_p - \hat{B}\hat{B}')\bar{X}. \tag{A.6}$$

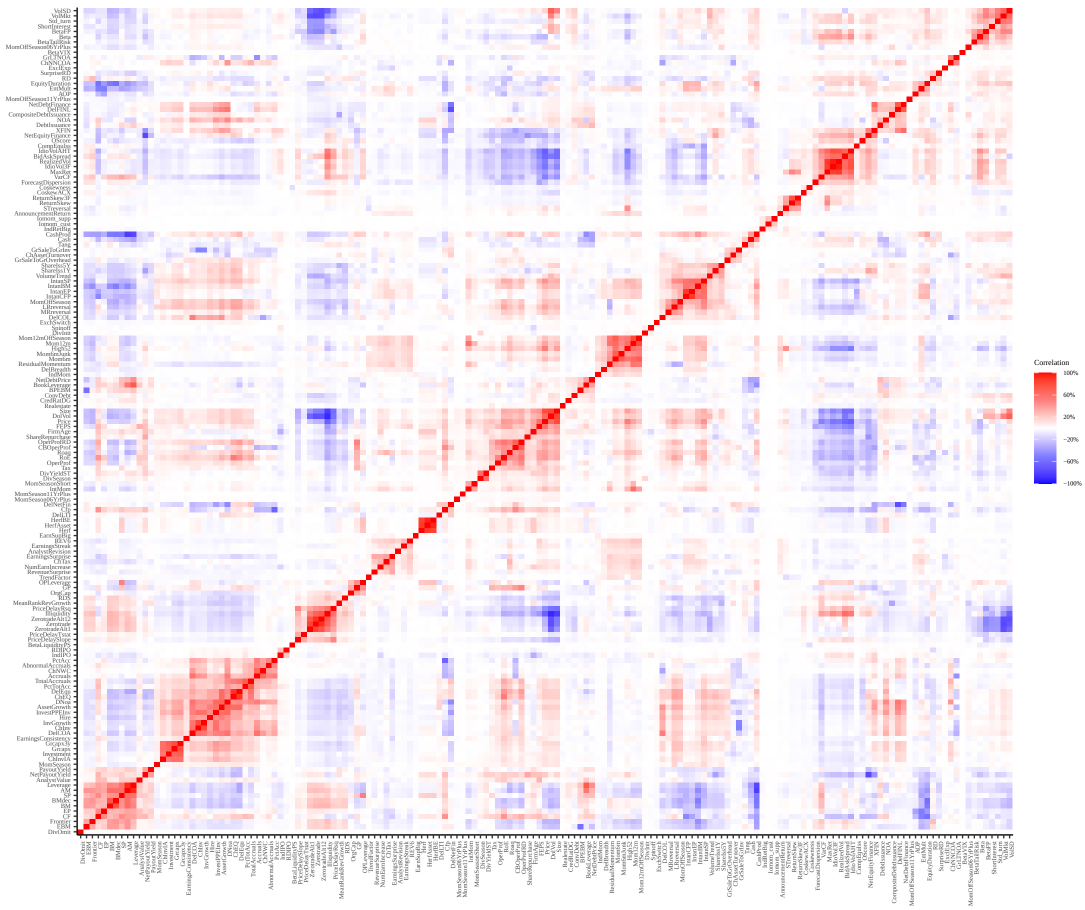
Appendix B Dataset of Characteristics

We require a set of stock characteristics to construct the managed portfolios. The initial dataset is sourced from Chen and Zimmermann (2022). As discussed in the main text, we first remove 53 characteristics with severe missing values. However, directly inputting the remaining 159 characteristics into the cross-sectional regressions is still impractical for two main reasons. First, many of these characteristics lack (independent) pricing information and may capture unpriced risk. Including a large number of such variables in the regression would introduce substantial noise, potentially obscuring the true pricing factors. Second, many of these characteristics exhibit high correlation with one another, which could lead to multicollinearity and inflate the variance of the managed portfolios. Both issues pose challenges for the construction of the EGS and other latent factors. To address these challenges, we implement the following three-step procedure to filter out irrelevant stock characteristics:

First, we exclude characteristics that lack pricing information. Specifically, We run simple Fama-MacBeth regressions on each characteristic and remove those that are statistically insignificant ($t < 1.96$). Including a large number of such characteristics in the cross-sectional regressions would add substantial noise to managed portfolios.

Second, we eliminate characteristics that cause multicollinearity. Many characteristics derive from related financial or trading items and convey similar firm-specific information, leading to high inter-correlations. This can induce multicollinearity in the cross-sectional regressions, inflating the variance of the managed portfolios. Figure B.1 illustrates the pairwise correlations among the 159 characteristics, highlighting several clusters of highly correlated variables. Within each cluster, we run multivariate Fama-MacBeth regressions. If only one characteristic demonstrates statistical significance and preserves its sign, we retain it and remove the others. This phenomenon, where a characteristic "absorbs" the effects of others in the cluster, is also documented by Fama and French (1992), which shows that the book-to-market ratio absorbs the effects of other valuation ratios. If no characteristic dominates, we select the most significant variable from the simple Fama-MacBeth regressions, representing the strongest signal.

Third, we remove characteristics that fail to provide independent information. As noted by Cochrane (2011) and Green et al. (2017), many of the characteristics in the socalled "zoo" capture overlapping and redundant information. Including a large number of such characteristics in multivariate regressions would also introduce substantial noise. To mitigate this issue, we employ multivariate Fama-MacBeth regressions in a stepwise manner, systematically eliminating redundant characteristics. At each step, we follow a clear process: (i) remove any characteristic that fails to preserve its sign, as this suggests misalignment with economic meaning, or (ii) if no such characteristic is identified, eliminate the least statistically significant variable. This process is repeated until all remaining characteristics are both consistent with economic meaning and statistically significant.



Notes: The figure illustrates the pairwise correlation among 159 characteristics. Color intensity reflects the strength of these correlations: redder grids represent more positive correlations, while bluer grids signify more negative correlations. The characteristics have undergone agglomerative hierarchical clustering to reveal underlying patterns.

Figure B.1. Correlations among Characteristics

Appendix C Factor Predictability

We assess the predictive ability of the different types of factors using the out-of-sample predictive R^2. The out-of-sample predictive R^2 for managed portfolios is defined as follows:

$$\text{Out-of-sample predictive } R^2 = 1 - \frac{\sum_{j=1}^{p}\sum_{t=121}^{T}\left(x_{t,j} - \hat{b}_j^{(t-1)\prime}\hat{\mu}_{t-1}\right)^2}{\sum_{j=1}^{p}\sum_{t=121}^{T} x_{t,j}^2} \tag{C.1}$$

where $\hat{\mu}_{t-1} = \sum_{\tau=t-120}^{t-1} f_\tau$ represents the time-series average of the factors, representing the factors' risk premium.. Since EGS, PCA, and IPCA factors can also be viewed as extracted from conditional factor models, they are capable of explaining individual stocks as well. Therefore, we also measure their out-of-sample predictive R^2 for individual stocks:

$$\text{Out-of-sample predictive } R^2 = 1 - \frac{\sum_{i=1}^{N}\sum_{t=121}^{T}\left(r_{i,t} - \hat{\beta}_{i,t-1}^{(t-1)\prime}\hat{\mu}_{t-1}\right)^2}{\sum_{i=1}^{N}\sum_{t=121}^{T} r_{i,t}^2} \tag{C.2}$$

where $\hat{\beta}_{i,t-1}^{(t-1)\prime} = \hat{B}^{(t-1)\prime} z_{i,t-1}$ represents the dynamic loadings of each stock on the factors. To mitigate the impact of extreme values, we include only stocks with out-of-sample data points exceeding 120 months. The results are presented in Table C.I. Not surprisingly, our EGS factors outperform other latent and observable factors in terms of out-of-sample predictive ability, whether for predicting managed portfolios or for predicting individual stocks.

Table C.I. Out-of-sample predictive R^2 (%)

	K									
	1	2	3	4	5	6	7	8	9	10
	Panel A: Managed portfolios									
EGS	0.08	0.22	0.48	0.71	1.10	1.26	1.51	1.69	1.91	2.15
PCA	0.00	-0.01	-0.10	0.13	0.11	0.13	0.22	0.30	0.46	0.84
IPCA	-0.32	-0.25	-0.20	-0.11	0.01	0.35	0.66	1.04	1.12	1.32
Observable	-0.34		-0.18	-0.64	-0.24	-0.40				
	Panel B: Individual stocks									
EGS	0.09	0.30	0.26	0.22	0.32	0.31	0.33	0.36	0.36	0.39
PCA	0.24	0.22	0.19	0.18	0.16	0.17	0.16	0.14	0.13	0.10
IPCA	0.25	0.19	0.18	0.17	0.19	0.16	0.13	0.19	0.18	0.20

Notes: the table presents the out-of-sample predictive R^2s for different types of factors. The comparison for predicting managed portfolio returns includes: (i) EGS, PCA, and IPCA factors, across various number of factors (K); and (ii) observable factors, including the CAPM, FF3, Q4, FF5, and FF6 models, corresponding to $K = 1, 3, 4, 5, 6$, respectively. The comparison for predicting individual stock returns includes only (i), as these factors can be viewed as derived from conditional factor models, which are specifically designed for explaining individual stocks.