Sparse Macro-Finance Factors

David E. Rapach[1] and Guofu Zhou*[2]

[1]Federal Reserve Bank of Atlanta
[2]Washington University in St. Louis

August 28, 2025

Abstract

We estimate sparse principal components from a large set of macro-finance variables. Each component is a sparse linear combination of the underlying variables, enhancing economic interpretability and yielding sharper asset pricing signals. Innovations to the components constitute a set of sparse macro-finance factors. Robust tests show that sparse factors tied to housing, yields, and credit spreads earn significant risk premia. Among the top 20 factor models formed from prominent characteristic-based factors and mimicking portfolios for the housing, yield, and credit spread factors, the latter three play leading roles, highlighting the importance of sparse macro-finance factors for capturing systematic risks.

JEL classifications: C38, C55, C58, E44, G12

Keywords: Sparse principal component analysis, FRED-MD, Risk premia, Factor-mimicking portfolio, Housing, Yields, Credit spreads

*Corresponding author. Send correspondence to Guofu Zhou, Olin School of Business, Washington University in St. Louis, St. Louis, MO 63130; e-mail: zhou@wustl.edu; phone: 314-935-6384. We thank conference and seminar participants at the 2018 Midwest Econometrics Group Meeting, 2018 Conference on Financial Predictability and Data Science, 2019 International Symposium on Forecasting, INQUIRE UK and Europe Spring 2019 Residential Joint Conference, FinTech 2020 Conference, 2021 Frontiers of Factor Investing Conference, 2025 SoFiE Financial Machine Learning Summer School at Yale University, Emory University, Federal Reserve Bank of Atlanta, Georgia State University, Gothenburg University, Indiana University, London Business School, Lund University, Saint Louis University, Southwestern University of Finance and Economics, Stockholm University, University of Bath, University of Illinois at Urbana-Champaign, Washington University in St. Louis, and Zhejiang University, as well as Svetlana Bryzgalova, Victor DeMiguel, Allen Li, Jean Pauphilet, Markus Pelger, Ye Xiaoxia, Dacheng Xiu, and Qianqun Zhou, for helpful comments. The views expressed here are those of the authors and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System. Any remaining errors are the authors' responsibility.

1. Introduction

Modern finance theory posits that securities' expected excess returns are linear functions of risk premia, which represent compensation for exposure to systematic risk factors in the macroeconomy. In addition to the excess return on the market portfolio, the intertemporal capital asset pricing model (ICAPM, Merton 1973) asserts that innovations to state variables that affect the "investment opportunity set" receive risk premia, as they expose investors to reinvestment risk. In an unconditional form and with the absence of arbitrage opportunities, Ross (1976) shows that expected excess returns are linear functions of risk premia related to systematic risk factors in the macroeconomy. Because investors hold jobs and own houses and small businesses, risk premia can also originate from shocks to state variables that reflect labor and housing market conditions as well as the prospects for small businesses (Cochrane 2005). More generally, innovations to state variables that affect an investor's marginal utility of consumption potentially constitute systematic risk factors that require risk premia.

The identification of the relevant systematic risk factors in the macroeconomy is ultimately an empirical issue. In their seminal study, Chen, Roll, and Ross (1986) find evidence of significant risk premia associated with industrial production, default and term spreads, and inflation in the equity market using the Fama and MacBeth (1973) two-pass regression methodology. A number of subsequent studies report that a variety of theoretically motivated state variables derived from individual macroeconomic and financial series generate significant risk premia in cross-sectional equity returns (e.g., Cochrane 1996; Jagannathan and Wang 1996; Lettau and Ludvigson 2001; Pástor and Stambaugh 2003; Vassalou 2003; Lustig and Van Nieurwerburgh 2005; Parker and Julliard 2005; Yogo 2006; Malloy, Moskowitz, and Vissing-Jørgensen 2009; Adrian, Etula, and Muir 2014). However, the evidence that such "macro-finance factors" generate significant risk premia does not appear very robust, especially after accounting for potential model misspecification and weak factors (e.g., Shanken and Weinstein 2006; Kan, Robotti, and Shanken 2013; Gospodinov, Kan, and Robotti 2014; Giglio and Xiu 2021; Giglio, Xiu, and Zhang 2025). As the literature now stands, it is still

unknown which macro-finance factors are relevant for the cross section of equity returns. Moreover, leading multifactor models in the literature all rely on factors formed by sorting on firm characteristics (e.g., Fama and French 1993, 2015, 2018; Carhart 1997; Hou, Xue, and Zhang 2015; Hou et al. 2021), so macro-finance factors are currently overshadowed by characteristic-based factors when it comes to the empirical modeling of systematic risks.

In this paper, we use advances in machine learning to compute *sparse macro-finance factors* for a large macro-finance database. Existing macro-finance factors are typically constructed using one variable. However, a larger number of variables are likely relevant for measuring something like housing or labor market conditions, so reliance on an individual variable can ignore pertinent information. Similarly, there are multiple measures of, for example, the interest rate, term spread, and industrial production. We cannot a priori know which measure is the most relevant, so selecting one can again neglect pertinent information.

We begin by constructing an extensive set of monthly macro-finance variables from the FRED-MD database (McCracken and Ng 2016). FRED-MD is valuable resource and is widely used for research, including for forecasting macro variables and asset returns (e.g., Kotchoni, Leroux, and Stevanovic 2019; Bianchi, Büchner, and Tamoni 2021; Medeiros et al. 2021; Borup and Schütte 2022). We consider 118 variables representing a broad array of potentially relevant economic risks from a variety of categories: output and income; labor market; housing; consumption, orders, and inventories; money and credit; yields; yield spreads; exchange rates; and prices. Then, we use sparse principal component analysis (e.g., Jolliffe, Trendafilov, and Uddin 2003; Zou, Hastie, and Tibshirani 2006; d'Aspremont et al. 2007; Shen and Huang 2008; Witten, Tibshirani, and Hastie 2009; Croux, Filzmoser, and Fritz 2013; Erichson et al. 2020; Chen and Rohe 2023) to perform dimension reduction by extracting twelve sparse components, each of which is a sparse linear combination of the 118 macro-finance variables.

The most popular method for succinctly capturing information in a large number of variables via dimension reduction is conventional principal component analysis (PCA). Why do

we favor sparse over conventional PCA to extract components from the macro-finance variables? Conventional principal components are linear combinations of all the 118 underlying variables. In our context, this has two primary drawbacks. First, the fact that all of the variables receive nonzero weights (or loadings) often makes conventional components difficult to interpret economically.[1] Thus, a given principal component typically does not cleanly align with a specific economic concept (e.g., housing market conditions). Second, because all of the underlying variables receive nonzero loadings, this can dilute the strength of the signal in each component for asset pricing, a manifestation of overfitting. By setting many loadings to zero, sparse PCA can enhance the economic interpretation of each component and at the same time provide sharper signals for asset pricing.

In essence, sparse PCA allows us to find a sweet spot between relying on a single variable or using a linear combination of all the variables (as in conventional PCA) to capture a specific economic concept. This balancing act relates to the well-known bias-variance trade-off. While relying on a single variable reduces the variance, it ignores potentially relevant information in other variables, thereby increasing the bias; while a linear combination of all the variables has greater scope to include information and thereby reduce the bias, the inclusion of all the variables increases the variance (and thus the risk of overfitting).[2] Sparse PCA has proven valuable for interpreting and analyzing data for diverse phenomena, such as gene expression and biological outcomes (e.g., Shen and Huang 2008; Witten, Tibshirani, and Hastie 2009; Chen and Rohe 2023), the El Niño and La Niña warming events (Erichson et al. 2020), and the sparse coding of natural images (Chen and Rohe 2023). We explore its value for interpreting and analyzing macro-finance data in the context of cross-sectional asset pricing.

Our strategy employs sparse component analysis (SCA, Chen and Rohe 2023) in conjunction with the supervised PCA (SPCA, Giglio, Xiu, and Zhang 2025) methodology for

[1] While the components can be rotated, in general, this does not improve their interpretability.

[2] Another approach is to a priori select a group of variables that are believed to be related to an economic concept and somehow combine them. However, it is often difficult to know exactly which variables best capture a concept and how to combine them. Sparse PCA provides a data-driven method for doing so.

estimating risk premia. SCA appears to be the most powerful approach for estimating sparse principal components, as it explains more of the total variation in the underlying variables for a given level of sparsity. SPCA, a refinement of the Giglio and Xiu (2021) three-pass procedure, provides a robust methodology for estimating risk premia that accounts for measurement error as well as omitted and weak factors. Fitting a vector autoregression to the twelve sparse components extracted from the macro-finance variables, we compute innovations to the sparse components, which constitute a set of sparse macro-finance factors. Then, with more than 600 equity portfolios serving as test assets, we use SPCA to construct factor-mimicking portfolios and estimate risk premia for the sparse macro-finance factors. By constructing mimicking portfolios for the sparse factors, we can compare our new sparse macro-finance factors to popular traded factors based on firm characteristics.

Previewing our results, we find that asset pricing data prefer substantial sparsity when it comes to computing the sparse macro-finance factors. Specifically, to implement our strategy, we need to set two hyperparameters. The first governs the degree of sparsity in the twelve sparse components extracted from the macro-finance variables via SCA, while the second determines the number of test asset returns used in the screening process of the SPCA procedure for estimating risk premia. With respect to the latter, as in Giglio, Xiu, and Zhang (2025), we use cross validation applied to the average time-series R^2 statistic for the mimicking-portfolio regressions, which measures the ability of the mimicking portfolios to track the sparse macro-finance factors. For the sparsity hyperparameter, we treat the maximum degree of sparsity permitted by SCA as a benchmark and continue to use the cross-validated average R^2 statistic. Allowing for six to twelve (supervised) principal components in the test asset returns when implementing SPCA, we always find that the cross-validated average R^2 statistic is highest for the benchmark. In other words, imposing the maximum degree of sparsity when computing the sparse macro-finance factors provides the best performance in terms of asset pricing, as measured by the ability of the mimicking portfolios to explain the sparse macro-finance factors. Intuitively, imposing substantial spar-

sity prevents overfitting the components extracted from the macro-finance variables, thereby providing more reliable signals for deciphering systematic risks in the macroeconomy.

Imposing the maximum degree of sparsity, the twelve sparse components extracted from the macro-finance variables via SCA have quite precise economic interpretations. According to the sparse loadings, the twelve components are naturally interpreted as housing, term spreads, inflation, production, credit, currencies, unemployment, yields, wages, tightness, credit spreads, and employment. Time-series plots of the sparse components are in line with the economic interpretations indicated by the sparse loadings. In contrast, the first twelve conventional principal components extracted from the macro-finance variables are difficult to interpret, and the loadings for the conventional components indicate that the individual components are amalgams of influences from diverse categories.

With respect to asset pricing, treating the number of components in the test asset returns as an additional hyperparameter, the cross-validated average R^2 statistic selects six components. In this case, the housing, yield, and credit spread factors command significant risk premia, so the data provide strong support for the relevance of these three factors. The importance of housing, yields, and credit spreads as sources of systematic risk makes economic sense. For many investors, housing is a substantive component of their wealth, making housing market fluctuations an important source of risk. In the aggregate, investors also have substantial bond holdings, which renders investors sensitive to changes in yields and concomitant fluctuations in bond prices. In addition, investors are concerned about having access to credit, especially when a difficult situation or promising opportunity unexpectedly arises, and changes in credit spreads generally correspond to the cost of obtaining credit. The risk premium is positive for the housing factor—so the factor increases during a "good" state of the world (i.e., when the marginal utility of consumption is low)—while the risk premia for the yield and credit spread factors are both negative—so the factors increase during a "bad" state of the world (i.e., when marginal utility is high). In line with the signs of the risk premia, cumulative return plots for the mimicking portfolios show that

the housing factor often performs poorly around business-cycle recessions, while the yield and credit spread factors (especially the latter) tend to perform relatively well around recessions. The magnitudes of the Sharpe ratios for the mimicking portfolios for the housing, yield, and credit spread factors are in line with those for popular characteristic-based factors. The Sharpe ratio for the housing factor is particularly sizable, with an annualized value well above unity.

Because the sparse macro-finance factors are derived from macro-finance data, instead of directly from firm characteristics, it is of interest to compare them to factors in leading characteristic-based multifactor models. Specifically, we consider two well-known models: (1) the Fama and French (2018) six-factor model, which adds a momentum factor to the five factors in Fama and French (2015); (2) the Hou et al. (2021) augmented q-factor model, which augments the Hou, Xue, and Zhang (2015) q-factor model with an expected investment growth factor. We compare the housing, yield, and credit spread factors to the characteristic-based factors using two approaches. The first follows Barillas and Shanken (2017) by examining whether the Fama-French six-factor and augmented q-factor models can price the sparse macro-finance factors in the context of what Hou et al. (2019) call "spanning regressions." Although the Fama-French six-factor and augmented q-factor models can price the yield factor, neither model can price the housing and credit spread factors; we also reject the null hypothesis that either model can jointly price the three sparse factors. Thus, the new sparse macro-finance factors contain information about systematic risks beyond that in leading characteristic-based models.

The second approach employs the Bayesian methodology developed by Chib and Zeng (2020) and Chib, Zeng, and Zhao (2020), which is designed to select relevant risk factors—namely, those appearing in the stochastic discount factor (SDF)—from among a potentially large set of candidate factors. Considering the three sparse macro-finance factors along with the complete set of factors in the Fama-French six-factor and augmented q-factor models, the top-ranked model (according to the posterior marginal likelihood) selects the housing, yield,

and credit spread factors as well as the return on equity and expected investment growth factors from the augmented q model. Indeed, the top 20 models all include the housing and yield factors, while 19 of the top 20 models also include the credit spread factor, so we again find that the sparse macro-finance factors contain important information regarding systematic risks when pitted against popular characteristic-based factors.[3] Posterior distributions indicate that the top-ranked model's SDF loadings for the housing, return on equity, and expected investment growth (yield and credit spread) factors are positive (negative) and significant, so an increase (decrease) in the factors corresponds to a good state of the world. The SDF itself for the top-ranked model appears economically plausible and evinces substantive fluctuations, including sharp upward spikes during recessions and financial crises. Furthermore, the tangency portfolio for the top-ranked model has an annualized Sharpe ratio of 2.08, which is significantly larger than those for the Fama-French six-factor and augmented q-factor models (1.11 and 1.90, respectively). Overall, our new sparse macro-finance factors appear to play a leading role in the empirical modeling of systematic risks, so reliance on characteristic-based factors alone likely neglects important information regarding systematic risks in the macroeconomy.

Our paper relates to various studies that employ conventional PCA in asset pricing. For example, a number of studies use conventional PCA to decrease the dimension of a large number of predictors in predictive regressions for aggregate stock or bond returns (e.g., Ludvigson and Ng 2007, 2009; Neely et al. 2014). By applying sparse in lieu of conventional PCA to a large set of macro-finance variables, we improve the economic interpretation of the components and filter more of the noise in the variables to provide even sharper signals for asset pricing. Conventional PCA has also been used to reduce the dimension of a large set of test asset returns when analyzing systematic risk factors (e.g., Chamberlain and Rothschild 1983; Connor and Koraczyk 1986, 1988; Kozak, Nagel, and Santosh 2020; Giglio and Xiu 2021). Intuitively, PCA extracts components that capture substantive comovements in the

[3]The top 20 models all include the return on equity and expected investment growth factors.

test asset returns and thus filters the idiosyncratic components in returns, which theory predicts are not priced. Instead of relying on conventional PCA to extract components from the entire set of test asset returns, SPCA applies screening to select the set of test asset returns used to compute each component. As shown by Giglio, Xiu, and Zhang (2025), screening utilizes the test asset returns that are the most germane for the observable risk factors being analyzed, which makes risk premium estimation robust to weak factors. In sum, our strategy employs sparsity together with screening to refine conventional PCA and thus improve the estimation of macro-finance factors and their risk premia.

Our paper also offers new insights into macro-finance modeling. Traditional asset pricing models, particularly those based on consumption (e.g., Cochrane 2005), struggle to explain variations in expected returns due to the smooth nature of consumption. We find that a factor based on housing, whose market size rivals that of the entire stock market, plays a key role in explaining cross-sectional expected stock returns. This supports macro-finance models that incorporate housing (e.g., Lustig and Van Nieurwerburgh 2005; Piazzesi and Scheider 2016). Additionally, our yield factor reveals a cross-sectional asset pricing role for bond yields that complements the macro-finance implications of the bond-stock correlation (e.g., Chernov, Lochstoer, and Song 2025), while our credit spread factor aligns with the recent findings in Hyland et al. (2025). Importantly, our results indicate that our new housing, yield, and credit spread factors are key macro-finance elements for capturing the systematic risks that affect cross-sectional expected returns.

The rest of the paper is organized as follows. Section 2 describes the estimation of sparse macro-finance factors and their risk premia. Section 3 presents the data and reports the fitted sparse components and risk premium estimates for the sparse macro-finance factors. Section 4 compares the new sparse macro-finance factors to leading characteristic-based factors. Section 5 explains how we will regularly update the sparse macro-finance factors, which are available on the Sparse Macro-Finance Factors website. Section 6 concludes.

2. Methodology

This section outlines the SCA approach for extracting sparse components and the SPCA procedure for robust risk premium estimation as well as our strategy for using SCA in conjunction with SPCA to compute sparse macro-finance factors and estimate their risk premia. In terms of notation, we denote the entry-wise ℓ_p norm for the $m \times n$ matrix $\boldsymbol{A}$ by $\|\boldsymbol{A}\|_p^p \equiv \|\boldsymbol{A}\|_p$, where

$$\|\boldsymbol{A}\|_p = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |A_{i,j}|^p \right)^{1/p}. \tag{1}$$

The Frobenius norm, denoted by $\|\boldsymbol{A}\|_F$, corresponds to $p = 2$ in Equation (1). The Stiefel manifold in $\mathbb{R}^n$ is given by $\mathcal{S}(n, r) = \{\boldsymbol{V} \in \mathbb{R}^{n \times r} \mid \boldsymbol{V}'\boldsymbol{V} = I_r\}$ and its convex hull by $\mathcal{H}(n, r) = \{\boldsymbol{V} \in \mathbb{R}^{n \times r} \mid \boldsymbol{V}'\boldsymbol{V} \preceq I_r\}$.

2.1. Fitting Sparse Components

Suppose that we have observations for $m = 1, \ldots, M$ variables over $t = 1, \ldots, T$ periods. Conventional PCA can be formulated as

$$\max_{\boldsymbol{\Theta}} \|\boldsymbol{X}\boldsymbol{\Theta}\|_F \quad \text{subject to } \boldsymbol{\Theta} \in \mathcal{S}(M, J), \tag{2}$$

where $\boldsymbol{X}$ is a $T \times M$ centered data matrix, $\boldsymbol{\Theta}$ is an $M \times J$ matrix of loadings, and $J \ll M$. A representative row of $\boldsymbol{X}$ contains the M-vector of observations for period t, while a representative column of $\boldsymbol{\Theta}$ contains the M-vector of loadings corresponding to the jth principal component. Each principal component (or score) is a linear combination of the variables in $\boldsymbol{X}$: $\boldsymbol{s}_j = \boldsymbol{X}\boldsymbol{\theta}_j$ for $j = 1, \ldots, J$, where $\boldsymbol{\theta}_j$ is the jth column of $\boldsymbol{\Theta}$. It is straightforward to implement PCA using, for example, a singular value decomposition of $\boldsymbol{X}$ to compute the fitted loading and score matrices, which we denote by $\widehat{\boldsymbol{\Theta}}^{\text{PCA}}$ and $\boldsymbol{X}\widehat{\boldsymbol{\Theta}}^{\text{PCA}}$, respectively.

Although conventional PCA is a canonical statistical technique for dimension reduction, it has two principal drawbacks in our context. First, all of the loadings are generally nonzero, which can make it difficult to economically interpret the components. Second, since all of the underlying variables receive nonzero loadings, this can reduce the signal-to-noise ratio in each component, a manifestation of overfitting in a high-dimensional setting. Sparse PCA mitigates these drawbacks by imposing a sparsity constraint on $\boldsymbol{\Theta}$ to fit components that are sparse linear combinations of the variables in $\boldsymbol{X}$.

We compute sparse components using SCA (Chen and Rohe 2023). SCA can be formulated as the following minimization problem expressed in terms of the matrix reconstruction error for $\boldsymbol{X}$:

$$\min_{\boldsymbol{Z},\boldsymbol{D},\boldsymbol{\Theta}} \|\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{D}\boldsymbol{\Theta}'\|_F \quad \text{subject to } \boldsymbol{Z} \in \mathcal{S}(T, J),\ \boldsymbol{\Theta} \in \mathcal{S}(M, J),\ \|\boldsymbol{\Theta}\|_1 \leq \zeta, \tag{3}$$

where $\boldsymbol{Z}$ and $\boldsymbol{D}$ are $T \times J$ and $J \times J$ matrices, respectively, and $\zeta > 0$ is hyperparameter governing the degree of sparsity in $\boldsymbol{\Theta}$. The $\|\boldsymbol{\Theta}\|_1 \leq \zeta$ constraint in Equation (3) is analogous to that used in a linear regression framework in the seminal least absolute shrinkage and selection operator (LASSO, Tibshirani 1996). Because the constraint involves the ℓ_1 norm, it allows for shrinkage to exactly zero, thereby imposing sparsity on the loadings.[4] To implement SCA, Chen and Rohe (2023) develop a computationally efficient iterative algorithm based on the following objective function, which is a simplified version of an equivalent bilinear form of Equation (3):

$$\max_{\boldsymbol{Z},\boldsymbol{\Theta}} \|\boldsymbol{Z}'\boldsymbol{X}\boldsymbol{\Theta}\|_F \quad \text{subject to } \boldsymbol{Z} \in \mathcal{H}(T, J),\ \boldsymbol{\Theta} \in \mathcal{H}(M, J),\ \|\boldsymbol{\Theta}\|_1 \leq \zeta. \tag{4}$$

The algorithm alternates between updating estimates of $\boldsymbol{Z}$ and $\boldsymbol{\Theta}$, iterating until convergence. We denote the final estimate of $\boldsymbol{\Theta}$ by $\widehat{\boldsymbol{\Theta}}^{\text{SCA}}$ and the fitted sparse components by

[4] Unlike prior sparse PCA procedures, $\boldsymbol{D}$ is not restricted to be diagonal, which enables SCA to explain more of the variation in the data for a given degree of sparsity.

$\boldsymbol{X}\widehat{\boldsymbol{\Theta}}^{\text{SCA}}$. Based on the equivalent bilinear form of Equation (3) that underlies Equation (4), Chen and Rohe (2023) impose bounds on the sparsity hyperparameter: $J \leq \zeta \leq J\sqrt{M}$. Relative to other procedures for constructing sparse principal components, SCA has two advantages. First, as shown by Chen and Rohe (2023), for a given level of sparsity, it explains more of the total variation in the data. Second, it only requires the setting of one hyperparameter (ζ).[5]

With $\widehat{\boldsymbol{\Theta}}^{\text{SCA}}$ in hand, we fit a first-order vector autoregression (VAR(1)) to $\boldsymbol{X}\widehat{\boldsymbol{\Theta}}^{\text{SCA}}$. The fitted residuals for the VAR(1), which represent innovations to the sparse components, constitute the set of J sparse macro-finance factors. We denote the $T \times J$ matrix of sparse macro-finance factors by $\boldsymbol{G} = [\ \boldsymbol{g}_1 \ \cdots \ \boldsymbol{g}_T\ ]'$, where $\boldsymbol{g}_t$ is the J-vector of sparse macro-finance factors for period t.[6]

2.2. Risk Premium Estimation

We estimate risk premia for the sparse macro-finance factors via SPCA (Giglio, Xiu, and Zhang 2025). SPCA posits the following data-generating process (DGP) for the set of N test asset excess returns (indexed by $i = 1, \ldots, N$):

$$\boldsymbol{r}_t = \boldsymbol{B}\boldsymbol{\gamma} + \boldsymbol{B}\boldsymbol{f}_t + \boldsymbol{\varepsilon}_t, \tag{5}$$

where $\boldsymbol{r}_t$ is the N-vector of test asset excess returns for period t, $\boldsymbol{f}_t$ is a P-vector of risk factors, $\boldsymbol{B}$ is an $N \times P$ matrix of factor exposures, $\boldsymbol{\gamma}$ is a P-vector of risk premia for $\boldsymbol{f}_t$, $\boldsymbol{\varepsilon}_t$ is an N-vector of idiosyncratic errors, $E(\boldsymbol{f}_t) = \boldsymbol{0}_P$, $E(\boldsymbol{\varepsilon}_t) = \boldsymbol{0}_N$, and $\text{cov}(\boldsymbol{\varepsilon}_t, \boldsymbol{f}_t') = \boldsymbol{0}_{N\times P}$.

[5] In contrast to conventional PCA, in which the principal components have a natural ordering and can be fit sequentially, SCA fits all of the sparse components simultaneously (meaning that they do not have a natural ordering).

[6] We lose one observation when we fit the VAR(1) to the sparse components in order to compute the sparse macro-finance factors. For notational simplicity, we continue to assume that we have T observations for the sparse macro-finance factors.

Defined in terms of $\boldsymbol{f}_t$, the SDF corresponding to Equation (5) can be expressed as

$$m_t = 1 - \boldsymbol{\gamma}' \boldsymbol{\Sigma}_f^{-1} \boldsymbol{f}_t, \tag{6}$$

where $\boldsymbol{\Sigma}_f$ is the $P \times P$ covariance matrix for $\boldsymbol{f}_t$. The factors in $\boldsymbol{f}_t$ are treated as latent.

SPCA posits an additional DGP that relates a set of observable factors—-in our application, the set of sparse macro-finance factors in $\boldsymbol{g}_t$—to the latent factors in $\boldsymbol{f}_t$:

$$\boldsymbol{g}_t = \boldsymbol{\psi} + \boldsymbol{\Pi} \boldsymbol{f}_t + \boldsymbol{u}_t, \tag{7}$$

where $\boldsymbol{\Pi}$ is a $J \times P$ matrix of observable factor exposures to $\boldsymbol{f}_t$, $\boldsymbol{u}_t$ is a J-vector of errors for the observable factors (potentially reflecting, among other things, measurement error in $\boldsymbol{g}_t$), $E(\boldsymbol{u}_t) = \boldsymbol{0}_J$, $E(\boldsymbol{f}_t, \boldsymbol{u}_t') = \boldsymbol{0}_{P \times J}$, and $E(\boldsymbol{\varepsilon}_t, \boldsymbol{u}_t') = \boldsymbol{0}_{N \times J}$. Based on Equations (5) to (7), the vector of risk premia for $\boldsymbol{g}_t$ is given by

$$\boldsymbol{\gamma}_g = -\text{cov}(\boldsymbol{g}_t, m_t) = -\text{cov}\big(\boldsymbol{\psi} + \boldsymbol{\Pi} \boldsymbol{f}_t + \boldsymbol{u}_t, 1 - \boldsymbol{\gamma}' \boldsymbol{\Sigma}_f^{-1} \boldsymbol{f}_t\big) = \boldsymbol{\Pi} \boldsymbol{\gamma}. \tag{8}$$

Section A.1 in the Internet Appendix provides details for computing $\hat{\boldsymbol{\gamma}}_g^{\text{SPCA}}$, the SPCA estimate of $\boldsymbol{\gamma}_g$ in Equation (8). By applying screening when computing the principal components for the test asset excess returns, SPCA refines the PCA procedure of Giglio and Xiu (2021), which is robust to measurement error and omitted factors, to provide a consistent estimate of $\boldsymbol{\gamma}_g$ that is also robust to weak factors.

2.3. Strategy

Our strategy for estimating sparse macro-finance factors and their risk premia can be summarized as follows:

1. **Sparse PCA.** Extract $J = 12$ sparse components via SCA from a large set of macro-finance variables.

2. **Innovations.** Fit a VAR(1) to the sparse components; the fitted residuals for the VAR(1) constitute the set of sparse macro-finance factors.

3. **Supervised PCA.** Estimate risk premia for the sparse macro-finance factors via SPCA.

Next, we provide details for implementing our strategy.

Mimicking portfolios for the sparse macro-finance factors play a central role in our analysis, including for risk premium estimation via SPCA and for comparing the sparse factors to leading characteristic-based factors in Section 4. A mimicking portfolio transforms a nontraded factor to a traded factor by projecting the nontraded factor onto a set of test asset excess returns (e.g., Huberman, Kandel, and Stambaugh 1987; Breeden, Gibbons, and Litzenberger 1989; Balduzzi and Robotti 2008). Instead of projecting directly onto the test asset excess returns, Giglio and Xiu (2021) and Giglio, Xiu, and Zhang (2025) propose forming a mimicking portfolio by projecting the nontraded factor onto principal components extracted from the excess returns, which effectively constructs a regularized mimicking portfolio. This approach readily accommodates a large number of test asset excess returns, which is needed to circumvent the omitted-variable bias resulting from arbitrarily relying on a smaller number of test assets. At the same time, it improves efficiency via dimension reduction by obviating the need to fit a high-dimensional regression. By including screening, SPCA makes the mimicking portfolio robust to weak factors in the test assets.

Let $\widetilde{\boldsymbol{G}}$ denote the $T \times J$ matrix of decentered observable factors, $\widehat{\boldsymbol{F}}^{\text{SPCA}}$ the $T \times P$ matrix of supervised principal components, and $\widehat{\boldsymbol{B}}^{\text{SPCA}}$ the SPCA estimate of $\boldsymbol{B}$ in Equation (5); see Section A.1 in the Internet Appendix for details for computing $\widehat{\boldsymbol{F}}^{\text{SPCA}}$ and $\widehat{\boldsymbol{B}}^{\text{SPCA}}$. To construct mimicking portfolios for the observable (but nontraded) factors in $\boldsymbol{g}_t$ via SPCA, we run J time-series regressions of the columns of $\widetilde{\boldsymbol{G}}$ on $\widehat{\boldsymbol{F}}^{\text{SPCA}}$, yielding $\hat{\boldsymbol{\Lambda}}^{\text{MP}}$, the $P \times J$ matrix of coefficient estimates. In terms of the test asset excess returns themselves, the $N \times J$ matrix of mimicking portfolio weights is given by $\widehat{\boldsymbol{W}}^{\text{MP}} = \widehat{\boldsymbol{B}}^{\text{SPCA}} \hat{\boldsymbol{\Lambda}}^{\text{MP}}$. Finally, the $T \times J$ matrix of mimicking portfolio excess returns can be straightforwardly computed as

$\boldsymbol{R}^{\text{MP}} = \boldsymbol{R}\widehat{\boldsymbol{W}}^{\text{MP}}$, where $\boldsymbol{R}$ is the $T \times N$ matrix of test asset excess returns. The R^2 statistics for the J regressions of the columns of $\widetilde{\boldsymbol{G}}$ on $\widehat{\boldsymbol{F}}^{\text{SPCA}}$ measure the ability of the mimicking portfolios to track (or hedge) the corresponding nontraded factors.

To implement our strategy, we need to set the sparsity and supervising hyperparameters. Starting with the latter, similarly to Giglio, Xiu, and Zhang (2025), we select the supervising hyperparameter q via cross validation applied to the average R^2 statistic for the J mimicking-portfolio regressions.[7] For a given value of the sparsity hyperparameter ζ used to construct the J sparse components and corresponding sparse macro-finance factors, we consider the following grid of values for q: $\{0.05, 0.1, 0.15, 0.2, 0.25, 0.5, 0.75, 1\}$. We select q from this grid using five-fold cross validation based on five runs. Specifically, we randomly divide the observable factor and test asset excess return samples into five nonoverlapping folds. Reserving the first fold as the initial validation sample, for a given value of q, we use the observable factor and test asset excess return observations from the remaining four folds to compute $\hat{\boldsymbol{\Lambda}}^{\text{MP}}$. Then, we use the $\hat{\boldsymbol{\Lambda}}^{\text{MP}}$ estimate from the four folds to compute the average R^2 statistic for the observations in the validation sample. We repeat this process with each fold serving in turn as the validation sample and take the mean of the average R^2 statistics across the five folds. To make the cross validation more robust, we make five runs and take the mean of the five-fold cross-validated R^2 statistics across the runs. We repeat the entire process for each value of q from the grid and select the value that maximizes the cross-validated R^2 statistic.

With respect to the sparsity hyperparameter ζ, we consider a grid with ten equally spaced elements, starting from the lowest permissible value ($\zeta = J$) and ending with the largest permissible value ($\zeta = J\sqrt{M}$). We treat the smallest permissible value of ζ, which corresponds to the greatest degree of sparsity, as a benchmark. For each value of ζ in the grid, we compute the sparse components and corresponding sparse macro-finance factors, as well as the average R^2 statistic for the value of q selected by cross validation. This allows

[7] Giglio, Xiu, and Zhang (2025) show that tuning q based on the average R^2 statistic in a validation sample is sufficient for valid hyperparameter tuning to obtain consistent risk premium estimates.

us to link the degree of sparsity to the cross-validated average R^2 statistic; in other words, from an out-of-sample perspective, we can examine whether asset pricing data favor more or less sparsity when it comes to constructing the sparse macro-finance factors.[8]

Finally, we also need to specify P, the number of (supervised) principal components extracted from the test asset excess returns for SPCA. We allow for a range of values of P from six to twelve, which appears adequate for capturing the underlying factors in the test asset excess returns potentially related to the observable factors (including potentially weak factors).[9] We can again use the cross-validated average R^2 statistic to guide the selection of P from an out-of-sample standpoint. Specifically, for each value of P, we save the cross-validated average R^2 statistic corresponding to the selected values of ζ and q; then, we select the value of P that maximizes the cross-validated average R^2 statistic across the considered values of P. In sum, we use the cross-validated average R^2 statistic to jointly select ζ, q, and P to estimate sparse macro-finance factors via SCA and their risk premia via SPCA.

3. Empirical Results

This section describes the data, presents the fitted sparse components extracted from the set of macro-finance variables, and reports risk premium estimates for the sparse macro-finance factors.

3.1. Data

The set of macro-finance variables is compiled from the extensive FRED-MD database (McCracken and Ng 2016). Starting with FRED-MD variables that are available from 1963, we construct a set of 118 variables. Table A.1 in the Internet Appendix provides a complete list of the 118 variables, including information on their construction and the abbreviations that

[8] Because the average R^2 statistic is computed via cross validation, it can be viewed as an out-of-sample estimate.

[9] SPCA requires $J \geq P$ for the construction of standard errors for the risk premium estimates (Giglio, Xiu, and Zhang 2025, Section B.4). We set $J = 12$, so the requirement is satisfied for the range of P values.

we use to refer to them. The macro-finance variables can be categorized as follows (with the number of variables in a given category in parentheses): output and income (16); labor market (31); housing (10); consumption, orders, and inventories (5); money and credit (12); yields (10); yield spreads (10); exchange rates (4); prices (20).

In constructing the variables, we employ various transformations as necessary to render them stationary. The transformations are standard. For example, we use log growth rates for measures of industrial production, income, employment, consumption, money, credit, exchange rates, and prices. We use the level of yields and yield spreads, as these variables are typically treated as stationary in finance. We also adjust the variables for any lags in their reporting. Yields, yield spreads, exchange rates, and oil prices are reported without delay, so no timing adjustment is needed for these variables. Nearly all of the remaining variables are reported with a one-month delay. In this case, we lag each observation by one month to account for the reporting delay. A few variables are reported with a two-month delay. Accordingly, we lag each observation by two months for these variables. The timing adjustments better reflect the flow of information to investors.[10] Finally, because the macro-finance variables have different scales, following convention, we standardize each variable to have a mean of zero and standard deviation of one.

For the set of test assets, we use 584 quintile portfolios from the August 2024 vintage of Open Source Asset Pricing (Chen and Zimmermann 2022) and 30 industry portfolios from Kenneth French's Data Library. The quintile portfolios form Open Source Asset Pricing are constructed by sorting on over 200 stock characteristics (considered in turn). We use the 584 quintile portfolios with continuous return data for 1963:07 to 2023:12 (726 observations); return data for all 30 industry portfolios are available for this sample period. Excess returns for the 614 test assets are computed using the monthly risk-free return from French's Data Library. Table A.2 in the Internet Appendix lists the 614 test assets and reports summary statistics for their excess returns.

[10]Table A.1 in the Internet Appendix reports the transformations and timing adjustments for the macro-finance variables.

We use a sample period of 1963:06 to 2023:12 when extracting the sparse components from the 118 macro-finance variables. This relatively long sample should be highly informative in our context, as it spans a number of macro cycles, including eight business-cycle recessions, and financial crises, such as the Asian financial and Long-Term Capital Management crises as well as the global financial crisis. After fitting a VAR(1) to the sparse components and taking the fitted residuals, this leaves a 1963:07 to 2023:12 sample period for the sparse macro-finance factors, matching the sample period for the test asset excess returns.

3.2. Sparse Macro-Finance Factors

We set the number of sparse components to twelve ($J = 12$), which appears sufficient for capturing a broad scope of potential sources of systematic risks in the macroeconomy. To compute the sparse components from the macro-finance variables, we need to determine the degree of sparsity. As explained in Section 2.3, we use asset pricing data—namely, the cross-validated average R^2 statistic for the mimicking-portfolio regressions—to guide the selection of the sparsity hyperparameter ζ.

Figure 1 provides a surface plot for the cross-validated average R^2 statistics (in percent). The horizontal axis shows P, the number of principal components in the test asset excess returns, while the vertical axis shows the sparsity hyperparameter ζ, which ranges from its smallest permissible value ($J = 12$) to its largest ($J\sqrt{M} = 130.35$). As we move down along the vertical axis, the sparsity constraint becomes more binding, corresponding to a greater degree of sparsity. The cross-validated average R^2 statistic generally increases as we move down along the columns in Figure 1. Indeed, for any value of P, the cross-validated R^2 statistic is maximized for $\zeta = 12$. Thus, asset pricing data favor the maximum degree of sparsity when computing sparse components for the set of macro-finance variables. Intuitively, imposing substantial sparsity when constructing the sparse components helps to

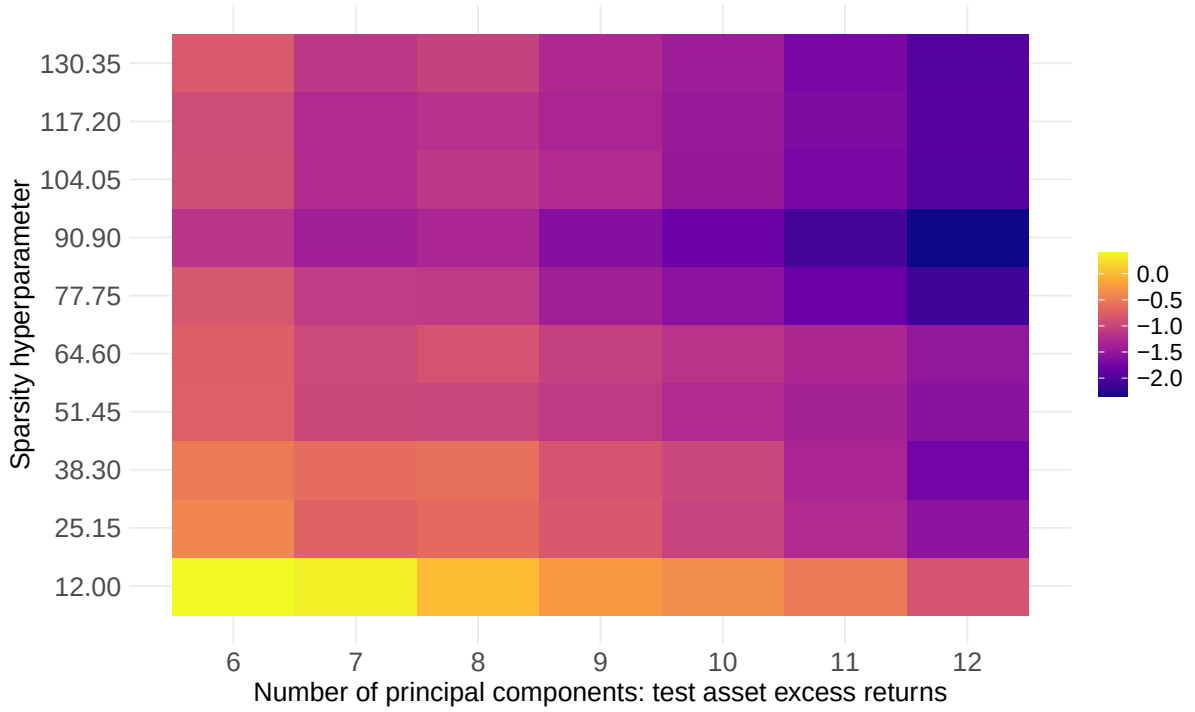


Figure 1. Cross-validated average R^2 statistics. The figure depicts a surface plot of cross-validated average R^2 statistics in percent for mimicking-portfolio regressions. Each cross-validated average R^2 statistic is based on five-fold cross validation and five runs. Moving down along the vertical axis corresponds to a greater degree of sparsity.

guard against overfitting the components, thereby providing more reliable signals for asset pricing.[11]

Table 1 reports the nonzero sparse loadings for each of the twelve sparse components based on $\zeta = 12$, the maximum degree of sparsity, corresponding to the value of ζ selected by the cross-validated average R^2 statistic for each value of P in Figure 1. Figure 2 depicts time-series plots of the sparse components themselves for $\zeta = 12$. Of the $118 \times 12 = 1{,}416$ loadings in total, only 77 (5.44%) are nonzero in Table 1, reflecting substantive sparsity. As reported in Table A.3 in the Internet Appendix, all 1,416 of the loadings for the first twelve conventional principal components are nonzero. The first twelve conventional principal com-

[11]Figure 1 shows that asset pricing data prefer very sparse components over conventional principal components, since the sparse components approach the conventional components as ζ increases (i.e., as the sparsity constraint becomes less binding).

ponents account for 68.54% of the variation in the macro-finance variables. Despite the high degree of sparsity, the sparse components explain 46.67% of the variation in the data.

Table 1. Nonzero sparse loadings

The table reports nonzero loadings for twelve sparse components extracted from 118 macro-finance variables constructed from the FRED-MD database. The complete list of variables and their abbreviations is provided in Table A.1 in the Internet Appendix. The sparse components are estimated using SCA with sparsity hyperparameter $\zeta = 12$. The sample period is 1963:06 to 2023:12.

(1) Variable	(2) Loading	(3) Variable	(4) Loading	(5) Variable	(6) Loading
A. Housing		**B. Term Spreads**		**C. Inflation**	
Starts-Total	0.15	TB6M–TB3M	0.12	PPI-FinGoods	0.10
Starts-MW	0.02	TB1Y–TB3M	0.12	PPI-FinConsGoods	0.10
Starts-S	0.18	TB5Y–TB3M	0.28	PPI-IntMatls	0.09
Starts-W	0.16	TB10Y–TB3M	0.25	CPI-All	0.08
Permits-Total	0.19	TB5Y–TB1Y	0.23	CPI-Transport	0.11
Permits-NE	0.03	TB10Y–TB1Y	0.18	CPI-Commodities	0.13
Permits-MW	0.06			CPI-AllExFood	0.07
Permits-S	0.18			CPI-AllExShelter	0.11
Permits-W	0.17			CPI-AllExMedical	0.09
				PCE-Defl	0.06
				PCE-DeflNondur	0.13
D. Production		**E. Credit**		**F. Currencies**	
IP-Total	0.11	NonrevCredit	0.32	CHF-USD	0.44
IP-Prod	0.12	NonrevConsCreditToPI	0.11	JPY-USD	0.33
IP-FinalProd	0.13	MotorVehLoans	0.34	USD-GBP	−0.36
IP-ConsGoods	0.17	ConsLoansLeases	0.41		
IP-ConsDur	0.07				
IP-ConsNondur	0.10				
IP-BusEqp	0.02				
IP-Matl	0.06				
IP-DurMatl	0.06				
IP-Manuf	0.09				
CapUtil-Manuf	0.11				

Based on their nonzero loadings in Table 1, the twelve sparse components have quite precise economic interpretations:

Housing. As shown in Panel A of Table 1, the first sparse component has nine nonzero loadings, all of which are positive, for measures of housing starts and new private housing permits. These are the two main types of housing-related variables in the

Table 1 (continued)

(1) Variable	(2) Loading	(3) Variable	(4) Loading	(5) Variable	(6) Loading
G. Unemployment		**H. Yields**		**I. Wages**	
Unemp-15Wks&Over	0.44	DivYld	0.04	AvgHrEarn-Goods	0.41
Unemp-15to26Wks	0.48	FedFund	0.08	AvgHrEarn-Construct	0.27
		TBill-3Mo	0.09	AvgHrEarn-Manuf	0.32
		TBill-6Mo	0.09		
		TBond-1Yr	0.09		
		TBond-5Yr	0.10		
		TBond-10Yr	0.10		
		CP-3Mo	0.08		
		CorpBond-AAA	0.11		
		CorpBond-BAA	0.10		
J. Tightness		**K. Credit Spreads**		**L. Employment**	
HelpIndex	0.13	AAA–TB10Y	0.19	CivilianEmp	0.09
HelpIndexToUnemp	0.17	BAA–TB10Y	0.33	UnempRate	−0.09
Unemp-5to14Wks	−0.01	BAA–AAA	0.25	Unemp-<5Wks	−0.01
Manuf&TradeSales	0.35			Emp-Total	0.10
InvtsToSales	−0.36			Emp-GoodsProd	0.01
				Emp-NondurGoods	0.04
				Emp-Services	0.11
				Emp-TTU	0.10
				Emp-WhlslTrade	0.06
				Emp-RetailTrade	0.10

macro-finance data, so the component naturally corresponds to housing. Four of the five housing-start variables in the data have nonzero loadings (the exception is housing starts in the Northeast), while all five housing-permit variables in the data have nonzero loadings.[12] In addition to total starts and permits, the largest loadings are for starts and permits in the South and West, the two regions that experienced the strongest population growth over the sample period. The housing component in Panel A of Figure 2 clearly displays the housing market cycle, including the long bull market from the early 1990s to the mid 2000s, followed by the housing market collapse corresponding to the global financial crisis. The housing component is generally procyclical (e.g.,

[12]Housing starts and permits are transformed using log levels.

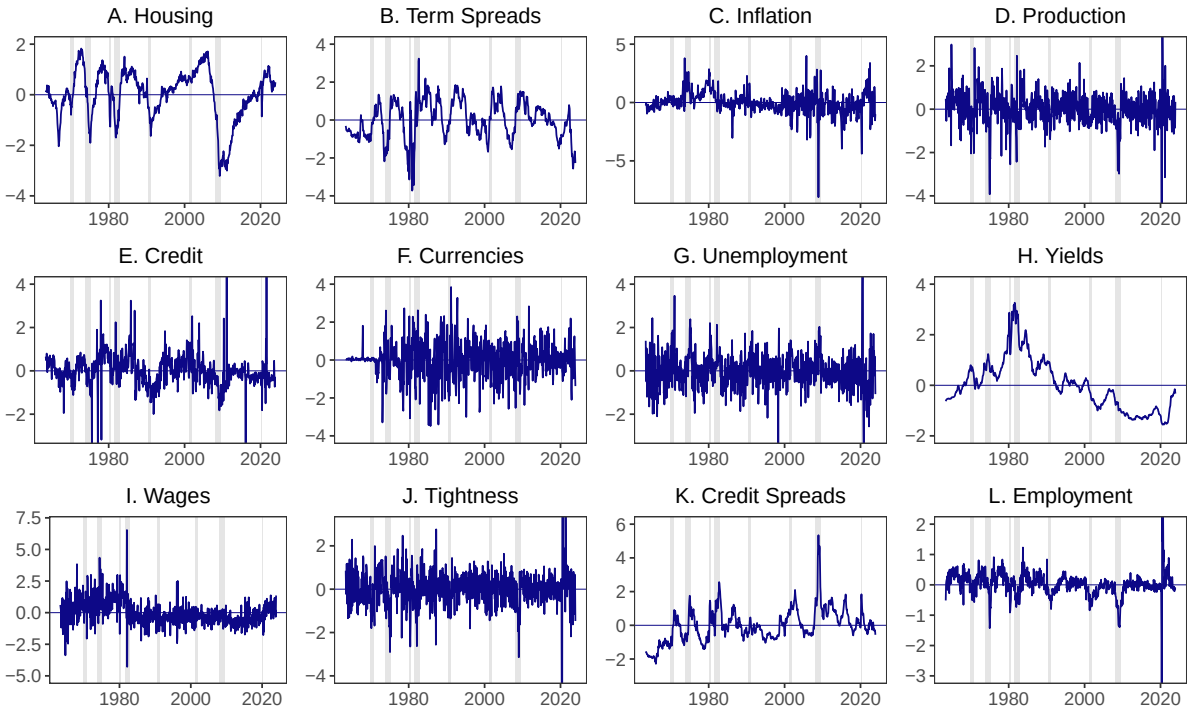


Figure 2. Sparse components. The figure depicts twelve sparse components extracted from 118 macro-finance variables constructed from the FRED-MD database. The components are standardized to have zero mean and unit variance. Vertical bars delineate business-cycle recessions as dated by the National Bureau of Economic Research.

Leamer 2007), evincing an especially sharp decline during the Great Recession of 2007 to 2009.

Term Spreads. The loadings in Panel B of Table 1 for the second sparse component are nonzero for the six term-spread measures in the macro-finance data, so we straightforwardly interpret this component as reflecting term spreads. All of the nonzero loadings are positive, and they are larger for measures involving bonds with longer maturities in the long end of the spread. The term spread component in Panel B of Figure 2 shows the well-known tendency of term spreads to markedly decline near the onset of a recession (e.g., Estrella and Hardouvelis 1991; Rudebusch and Williams 2009).

Inflation. According to Panel C of Table 1, the third sparse component reflects inflation. The loadings are nonzero for three producer price indices, six consumer price indices, and two personal consumption expenditure deflators, where the price indices and de-

flators are transformed using log growth rates. All of the nonzero loadings are positive. Well-known inflationary episodes are evident for the inflation component in Panel C of Figure 2, including surges in inflation during the 1970s, the early 1980s, and more recently in the wake of the COVID-19 crisis.

Production. In line with the nonzero loadings in Panel D of Table 1, the fourth sparse component corresponds to production. Specifically, the loadings are nonzero, and all are positive, for ten measures of industrial production in the macro-finance data (all of which are transformed using log growth rates). The capacity utilization rate for manufacturing (transformed via first differences) also receives a nonzero weight for this component. In line with its interpretation, the production component in Panel D of Figure 2 exhibits procyclical behavior, with sizable declines during the mid-1970s recession, the double-dip recession of the early 1980s, the Great Recession, and especially the recession in early 2020 corresponding to the COVID-19 pandemic.

Credit. The fifth sparse component in Panel E of Table 1 has nonzero loadings, all of which are positive, for four credit-related variables: total nonrevolving credit, the nonrevoling credit to personal income ratio, consumer motor vehicle loans outstanding, and total consumer loans and leases outstanding.[13] The credit component in Panel E of Figure 2 reveals well-known aspects of the credit cycle, including a marked decline during the global financial crisis and concomitant Great Recession followed by a sharp increase in response to historically unprecedented actions by the Federal Reserve.

Currencies. The sixth sparse component in Panel F of Table 1 corresponds to currencies, as it has nonzero loadings for three of the four exchange rates in the macro-finance data (all of which are transformed using log growth rates). The loadings are positive for the Swiss francs per US dollar and Japanese yen per US dollar exchange rates, while the loading is negative for the US dollars per British pound exchange rate.

[13]The first, third, and fourth (second) variables are transformed using log growth rates (first differences).

Thus, an increase in the currency component corresponds to a stronger US dollar. The currency component in Panel F of Figure 2 displays limited volatility at the beginning of the sample and subsequently becomes much more volatile with the breakdown of the Bretton Woods system in the early 1970s.

Unemployment. Based on the nonzero loadings in Panel G of Table 1, the seventh sparse component corresponds to long-term unemployment. Specifically the loadings are positive and nearly of equal size for the number of unemployed for 15 weeks and over and the number of unemployed for 15 to 26 weeks, both of which are transformed via log growth rates. Upward spikes in the unemployment factor during recessions are readily evident in Panel G of Figure 2, most notably during the recession in early 2020 associated with the COVID-19 pandemic.

Yields. For the eighth sparse component in Panel H of Table 1, the loadings are nonzero for the ten yield measures in the macro-finance data. All of the loadings are positive, and they tend to be slightly higher for bonds with longer maturities. The yield component in Panel H of Figure 2 exhibits well-known patterns in yields, such as a general increase during the early 1980s and a subsequent decline, low levels in the aftermath of the global financial crisis, and an increase at the end of the sample.

Wages. We associate the ninth sparse component in Panel I of Table 1 with wages, as it has nonzero loadings for the three measures of average hourly earnings (goods producing, construction, and manufacturing) in the macro-finance data; each of the three measures is transformed using log growth rates. The nonzero loadings are positive, and average hourly earnings in the goods-producing sector receive the largest loading. Elevated wage growth in the late 1960s through the early 1980s, as well as at the end of the sample, can be seen for the wage component in Panel I of Figure 2.

Tightness. Based on the loadings in Panel J of Table 1, we interpret the tenth sparse component as a measure of tightness in the macroeconomy. This component has five nonzero

loadings. The loadings are positive (negative) for the help-wanted index, the ratio of the help-wanted index to the number of unemployed, and real manufacturing and trade sales (the number of unemployed for five to 14 weeks and the ratio of inventories to sales).[14] The tightness component falls sharply during severe recessions, especially the recession in early 2020. Shortly after this recession, the tightness measures jumps, reflecting the efforts of many businesses to restore production in the aftermath of the COVID-19 pandemic.

Credit Spreads. The eleventh sparse component in Panel K of Table 1 has nonzero loadings, all of which are positive, for three of the four credit spreads in the macro-finance data. The three credit spreads are the differences between AAA corporate and ten-year Treasury bond yields, BAA corporate and ten-year Treasury bond yields, and BAA and AAA corporate bond yields.[15] The difference between BAA corporate and ten-year Treasury bond yields receives the largest loading. The credit spread component tends to rise during recessions, especially during the Great Recession corresponding to the global financial crisis.

Employment. From Panel L of Table 1, we see that the last sparse component has nonzero loadings on eight employment variables from the macro-finance data: civilian employment from the household survey and seven employment measures from the establishment survey (including total nonfarm employment), where the employment variables are transformed using log growth rates. The loadings on the employment variables are all positive. In contrast, the two other nonzero loadings in Panel L are negative, corresponding to the unemployment rate and the number of unemployed for less than five weeks, where the former (latter) is transformed via first differences (log growth

[14]The help-wanted index, the ratio of the help-wanted index to the number of unemployed, and the ratio of inventories to sales (the number of unemployed for five to 14 weeks and real manufacturing and trade sales) are transformed using first differences (log growth rates).

[15]The excluded credit spread is the difference between three-month commercial paper and three-month Treasury bill yields.

rates). As expected, the employment component is strongly procyclical in Panel L of Figure 2.

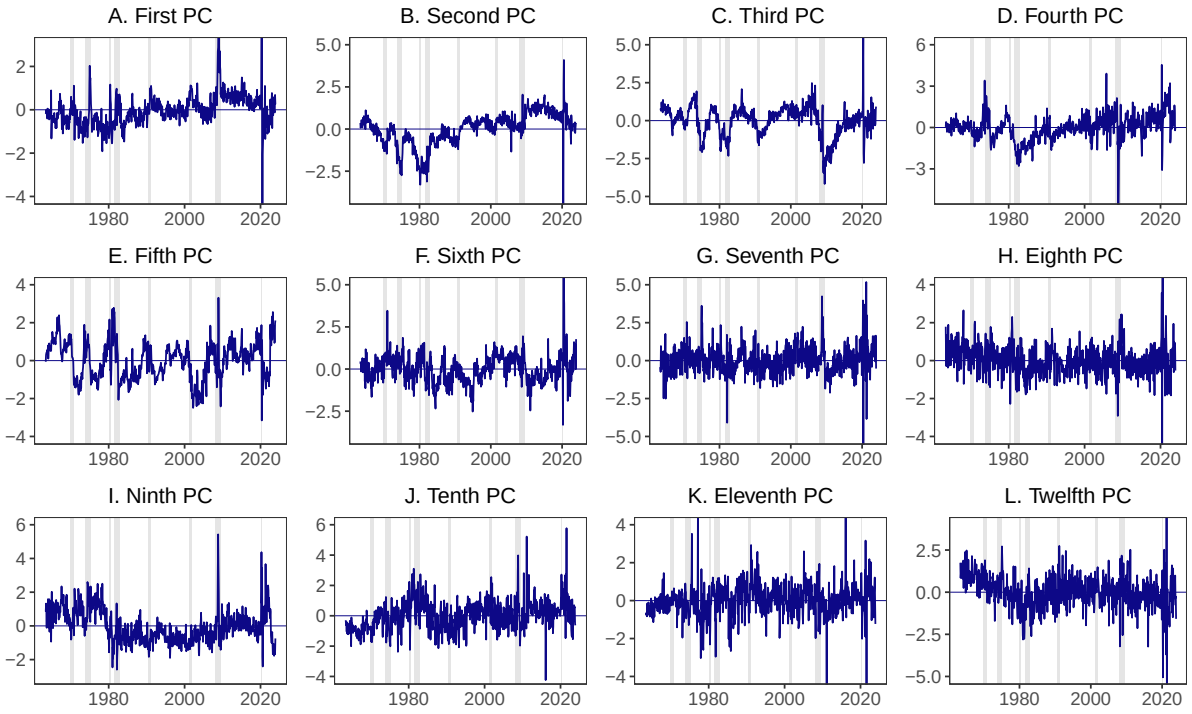


Figure 3. Conventional components. The figure depicts the first twelve conventional principal components extracted from 118 macro-finance variables constructed from the FRED-MD database. The components are standardized to have zero mean and unit variance. Vertical bars delineate business-cycle recessions as dated by the National Bureau of Economic Research.

By comparison, the loadings for the first twelve conventional principal components extracted from the macro-finance variables, which are reported in Table A.3 in the Internet Appendix, make it considerably more difficult to economically interpret the conventional components. Specifically, each conventional component has loadings that are relatively large in magnitude for variables from different economic categories, so each component appears to reflect an amalgam of economic concepts. Figure 3, which depicts time-series plots of the first twelve conventional principal components, reiterates the lack of precise economic interpretation for the conventional components vis-à-vis their sparse counterparts. This is not necessarily surprising, as conventional PCA is not designed to promote interpretation per se. In contrast, the raison d'être for sparse PCA is the construction of components

that are easier to interpret (e.g., Jolliffe, Trendafilov, and Uddin 2003), and, as noted in the Section 1, sparse PCA has been used to improve the interpretation of components in a variety of domains. We find that it is also useful for interpreting components extracted from a large set of macro-finance variables. As previously indicated, despite the substantial degree of sparsity imposed—which significantly facilitates economic interpretation—the twelve sparse components fitted via SCA still account for approximately 47% of the variation in the macro-finance variables (compared to 69% for conventional PCA). Thus, we are reasonably close to having our cake and eating it too when it comes to interpreting the components and accounting for the variation in the data.

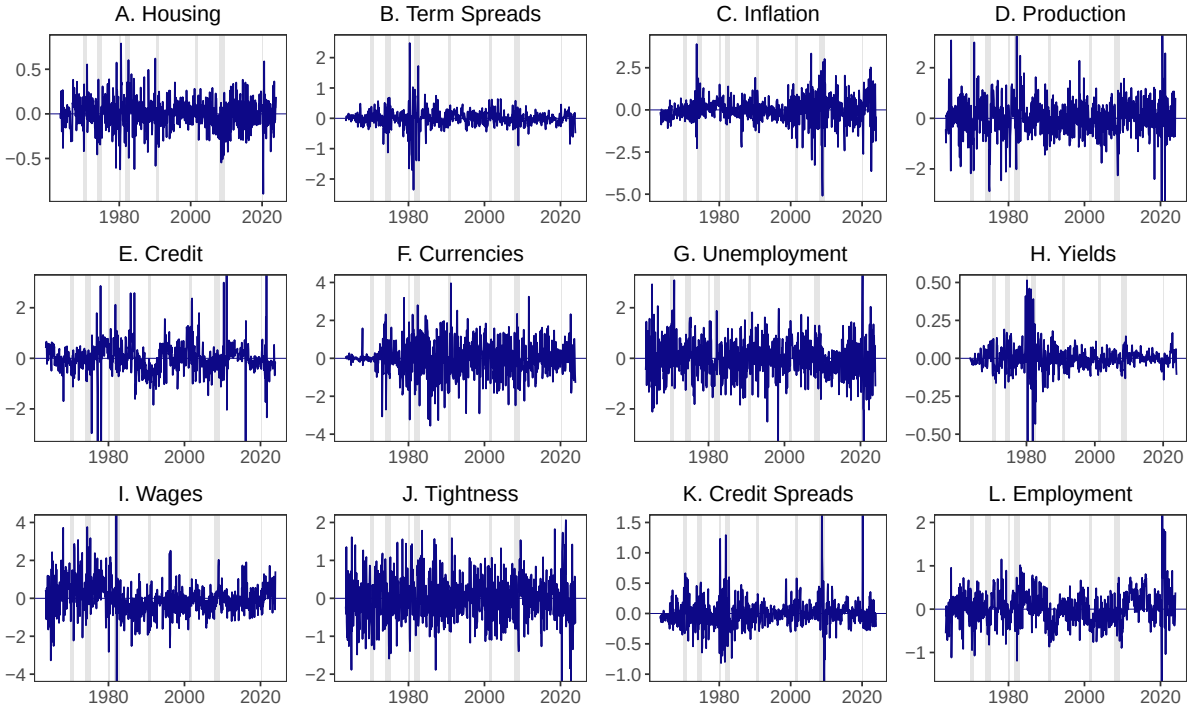


Figure 4. Sparse macro-finance factors. The figure depicts twelve sparse macro-finance factors, which are computed by fitting a VAR(1) to the sparse components in Figure 2 and taking the fitted residuals. Vertical bars delineate business-cycle recessions as dated by the National Bureau of Economic Research.

The twelve sparse components can be viewed as potential state variables distilled from a large set of macro-finance variables. For estimating risk premia, we compute innovations to the state variables by fitting a VAR(1) to the sparse components, with the fitted residuals

comprising the set of sparse macro-finance factors. Figure 4 provides time-series plots of the twelve sparse factors.

3.3. Risk Premia

Next, we estimate risk premia for the twelve sparse macro-finance factors via SPCA. To implement SPCA, we also need to select P, the number of (supervised) principal components in the test asset returns. As discussed in Section 2.3, we use the cross-validated average R^2 statistic for the mimicking-portfolio regressions to jointly select the sparsity and supervising hyperparameters (ζ and q, respectively) as well as P. Looking back to Figure 1, the cross-validated average R^2 statistic selects $P = 6$ (corresponding to the lower-left corner of Figure 1).[16]

Table 2 reports the risk premium estimates. The estimated risk premium for the housing and yield factors are significant at the 1% level, while that for the credit spread factor is significant at the 5% level. Thus, we find evidence of significant risk premia for three of the sparse macro-finance factors at conventional levels.[17] The statistically significant risk premium estimates are also economically significant according to the magnitudes of the annualized Sharpe ratios for their mimicking portfolios in the third column. The Sharpe ratios are sizable in the context of popular characteristic-based factors, such as the Fama and French (1993) size and value factors, as discussed further in Section 4. The annualized Sharpe ratio is especially large for the housing factor (1.23). According to the R^2 statistics in the fourth column of Table 2, the mimicking portfolios track the housing, yield, and credit spread factors quite well in light of the well-known difficulty that mimicking portfolios often have tracking nontraded factors. The R^2 statistics are particularly sizable for the yield and credit spread factors in the fourth column, with values of 19.55% and 14.24%, respectively. The cross-validated R^2 statistics for these two factors are also relatively large (11.31% and

[16]The corresponding value for the supervising hyperparameter for SPCA is $q = 0.1$. Taking the three hyperparameters together, the cross-validated average R^2 statistic jointly selects $\zeta = 12$, $q = 0.1$, and $P = 6$.

[17]Using a χ^2-statistic, we reject the null hypothesis that the the twelve risk premia are jointly zero at the 1% level.

Table 2. Risk premia

The table reports risk premium estimates computed via SPCA for twelve sparse macro-finance factors and 614 test assets. Twelve sparse components are first extracted from 118 macro-finance variables constructed from the FRED-MD database. The sparse components are estimated using SCA with sparsity hyperparameter $\zeta = 12$. The sparse macro-finance factors are computed by fitting a VAR(1) to the sparse components and taking the fitted residuals. SPCA estimation is based on a supervising hyperparameter of $q = 0.1$ and $P = 6$, where the latter is the number of (supervised) principal components in the test asset excess returns. The hyperparameters ζ, q, and P are selected using the cross-validated average R^2 statistic for the twelve mimicking-portfolio regressions. Brackets in the second column report t-statistics based on robust standard errors (Newey and West 1987); *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. The third column reports the annualized Sharpe ratio for the sparse factor's mimicking portfolio. The fourth and fifth columns report the in-sample and cross-validated R^2 statistics for the sparse factor's mimicking-portfolio regressions. The sample period is 1963:07 to 2023:12.

(1)	(2)	(3)	(4)	(5)
			Mimicking-portfolio R^2 (%)	
Sparse factor	Risk premium (%)	Ann. Sharpe ratio	In sample	Cross validated
Housing	1.56 [4.04]***	1.23	5.43	0.61
Term Spreads	0.28 [0.47]	0.29	1.08	−1.86
Inflation	1.79 [1.32]	0.69	1.31	−0.96
Production	1.97 [1.13]	0.59	1.69	−2.91
Credit	0.89 [0.73]	0.35	1.01	−1.00
Currency	0.53 [0.36]	0.09	5.20	2.08
Unemployment	1.14 [0.86]	0.41	1.39	−3.22
Yields	−0.52 [−2.71]***	−0.52	19.55	11.31
Wages	−0.96 [−0.67]	−0.35	1.44	−0.66
Tightness	1.03 [1.22]	0.55	1.59	−0.83
Credit Spreads	−0.89 [−2.00]**	−0.50	14.24	6.55
Employment	0.47 [0.71]	0.31	0.77	−4.29

6.55%, respectively) in the last column of Table 2.[18] We examine the housing, yield, and credit spread factors in further detail in the next section.

[18]The cross-validated R^2 statistic tends to understate the true out-of-sample R^2 statistic (e.g., James et al. 2021, Chapter 5), so the latter likely lies between the values in the fourth and fifth columns of Table 2.

4. Factor Comparison

In this section, we economically interpret the housing, yield, and credit spread factors and compare their mimicking portfolios to leading characteristic-based factors.

4.1. Factor Properties

Panel A of Table 3 reports annualized summary statistics for the housing, yield, and credit spread factors; Panels B and C report summary statistics for the factors constituting two popular characteristic-based multifactor models, namely, the Fama and French (2018) six-factor (FF6) and Hou et al. (2021) augmented q-factor (Q5) models, respectively. The former includes the five factors from Fama and French (2015)—market (MKT), size (SMB), value (HML), profitability (RMW), and investment (CMA)—along with a momentum (MOM) factor; the latter augments the four factors from the Hou, Xue, and Zhang (2015) q-factor model—market, size (ME), investment (IA), and return on equity (ROE)—with an expected investment growth (EG) factor. The sample period in Panels A and B is 1963:07 to 2023:12; because data for the Q5 factors begin later, the sample period is 1967:01 to 2023:12 in Panel C.[19]

The housing factor earns an annualized average return of 18.77% and an annualized Sharpe ratio of 1.23, while the average return and Sharpe ratio for the yield (credit spread) factor are −6.23% and −0.52 (−10.67% and −0.50), respectively.[20] Compared to the Sharpe ratios for the characteristic-based factors in Panels B and C, those for the housing, yield, and credit spread factors in Panel A are economically sizable. The Sharpe ratio for the housing factor is particularly large, and it is nearly equal to the largest Sharpe ratio among the characteristic-based factors (1.30 for the expected investment growth factor in the Q5 model). The Sharpe ratios for the yield and credit spread factors are larger in magnitude

[19]Data for the FF6 model are from French's Data Library. Data for the Q5 model are from Lu Zhang's website.

[20]The annualized average returns for the housing, yield, and credit spread factors in Panel A of Table 3 are computed by annualizing the corresponding risk premium estimates in the second column of Table 2; the annualized Sharpe ratios in Panel A of Table 3 are the corresponding values in the third column of Table 2.

Table 3. Annualized summary statistics

The table reports annualized summary statistics for sparse macro-finance and characteristic-based factors. Housing, Yields, and Credit Spreads in Panel A are factor-mimicking portfolios derived from sparse principal components corresponding to housing, yields, and credit spreads, respectively. For the Fama-French six-factor model (Fama and French 2018) in Panel B, MKT = market factor, SMB = small-minus-big size factor, HML = high-minus-low value factor, RMW = robust-minus-weak profitability factor, CMA = conservative-minus-aggressive investment factor, and MOM = momentum factor. For the augmented q-factor model (Hou et al. 2021) in Panel C, MKT = market factor, ME = size factor, IA = investment factor, ROE = return on equity factor, and EG = expected investment growth factor. The sample period in Panels A and B (C) is 1963:07 to 2023:12 (1967:01 to 2023:12).

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Factor	Mean (%)	Vol. (%)	Sharpe ratio	Factor	Mean (%)	Vol. (%)	Sharpe ratio
A. Sparse macro-finance factors				**B. Fama-French six-factor model**			
Housing	18.77	15.29	1.23	MKT	6.86	15.57	0.44
Yields	−6.23	12.06	−0.52	SMB	2.52	10.46	0.24
Credit Spreads	−10.67	21.41	−0.50	HML	3.56	10.29	0.35
				RMW	3.36	7.72	0.44
				CMA	3.20	7.15	0.45
				MOM	7.09	14.58	0.49
C. Augmented q-factor model							
MKT	6.99	15.88	0.44				
ME	3.14	10.60	0.30				
IA	4.32	7.16	0.60				
ROE	6.40	9.07	0.71				
EG	9.37	7.20	1.30				

than those for all of the factors for the FF6 model in Panel B and the market and size factors for the Q5 model in Panel C.

Like the market factor, the housing factor commands a positive risk premium in Table 2 and Panel A of Table 3, so an increase in the housing factor corresponds to a "good" state of the world when an investor's marginal utility of consumption is low. This makes economic sense, as housing is a substantial component of wealth for many investors. As with the market factor, investors require a positive risk premium to hold stocks that are positively exposed to the housing factor, as a fall in the factor correlates with an increase in marginal

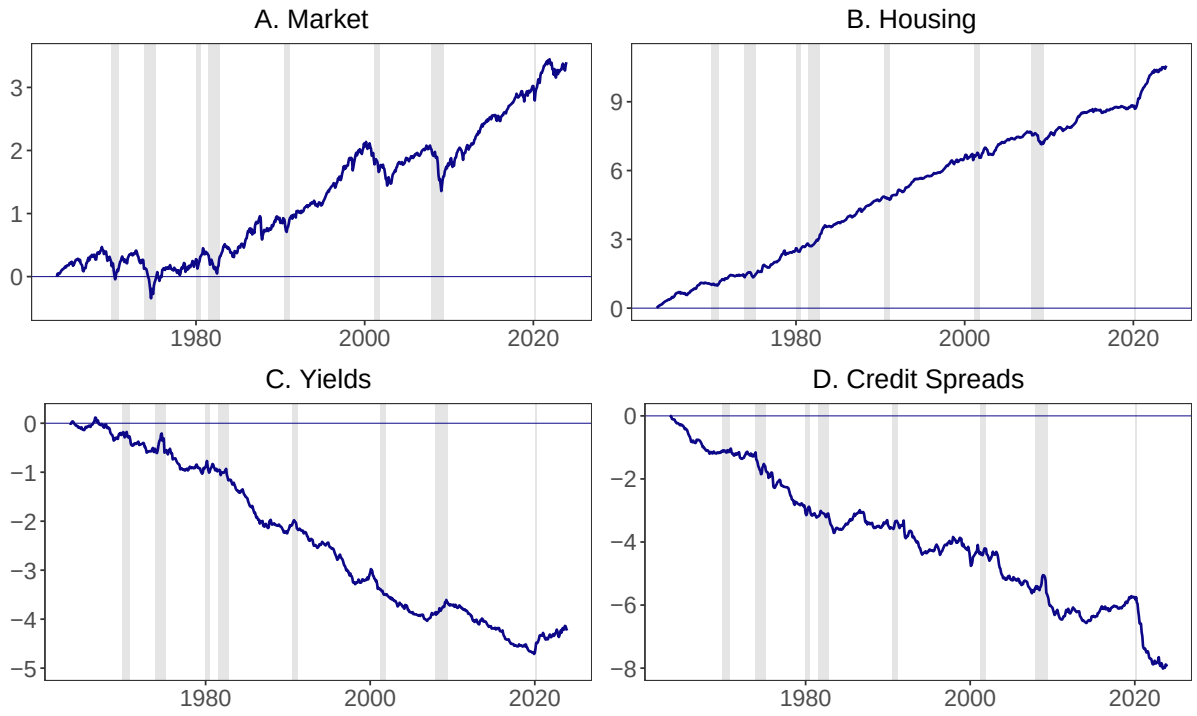


Figure 5. Log cumulative returns. The figure depicts log cumulative returns for the market, housing, yield, and credit spread factors. Vertical bars delineate business-cycle recessions as dated by the National Bureau of Economic Research.

utility (i.e., a "bad" state of the world). Panels A and B of Figure 5 portray log cumulative returns for the market and housing factors, respectively. In line with their positive risk premia, both factors tend to fall—often sharply—during recessions.

In contrast to the market and housing factors, the risk premia for the yield and credit spread factors are negative, as shown in Table 2 and Panel A of Table 3, so an increase in the yield and credit spread factors is associated with a bad state of the world. Accordingly, stocks with positive exposures to these factors provide a hedge against bad states, so they sell at a premium (corresponding to a negative risk premium in terms of expected return). Again in contrast to the market and housing factors, the yield and credit spread factors typically increase during recessions in Panels C and D, respectively, of Figure 5, indicative of their ability to hedge against bad states. The relevance of the yield and credit spread factors for the systematic risks faced by investors is economically intuitive. Overall, investors have considerable bond holdings, so changes in yields can lead to sizable changes in wealth

as bond prices fluctuate. Changes in credit spreads affect the cost of accessing credit for many investors, with an increase in the cost of obtaining credit making it more difficult for an investor to access credit if they find themselves facing a burdensome situation or wish to take advantage of a promising investment opportunity.

4.2. Spanning Regressions

Next, we follow the approach of Barillas and Shanken (2017) and analyze the ability of the the FF6 and Q5 factor models to price the housing, yield, and credit spread factors. For the FF6 (Q5) model, this entails estimating multiple regressions for the housing, yield, and credit spread factors (considered in turn) on the FF6 (Q5) factors. Table 4 reports the results.

The results in Panel A show that the FF6 factors can price the yield factor; however, they are unable to price the housing and credit spread factors. Specifically, although the alpha in the yield regression in the third column is insignificant, the alphas in the housing and credit spread regressions in the second and fourth columns, respectively, are significant at the 1% level. Furthermore, the null hypothesis that the three alphas are jointly zero is rejected at the 1% level based on the Gibbons, Ross, and Shanken (1989, GRS) statistic, so the FF6 factors cannot price the housing, yield, and credit spread factors in an overall sense. The average magnitude of the three annualized alphas in Panel A is 6.04%, which is economically sizable. The magnitudes of the alphas for the housing and credit spread factors are particularly large (8.56% and −8.65%, respectively).

The results are similar for the Q5 model in Panel B of Table 4. While the alpha in the yield regression in the seventh column is insignificant, those in the housing and credit spread regressions in the sixth and eighth columns, respectively, are significant at the 1% level. Based on the GRS statistic, we also reject the null that the three alphas are jointly zero at the 1% level. The average magnitude of the alphas is economically large (6.13%), primarily driven by the considerable magnitudes of the alphas for the housing and credit spread factors

Table 4. Spanning regressions

The table reports spanning regression estimation results regarding the ability of the Fama-French six-factor model (Fama and French 2018, Panel A) and the augmented q-factor model (Hou et al. 2021, Panel B) to price the housing, yield, and credit spread factors. The dependent variable is the factor in the column heading. Brackets report t-statistics based on robust standard errors (Newey and West 1987); *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. MKT = market factor, SMB = small-minus-big size factor, HML = high-minus-low value factor, RMW = robust-minus-weak profitability factor, CMA = conservative-minus-aggressive investment factor; MOM = momentum factor, ME = size factor, IA = investment factor, ROE = return on equity factor, EG = expected investment growth factor. GRS is the Gibbons, Ross, and Shanken (1989) test of the null hypothesis that the three alphas are jointly zero. $\overline{|\alpha|}$ is the average absolute value of the three alphas. The sample period is 1963:07 to 2023:12 (1967:01 to 2023:12) in Panel A (B).

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Factor	Housing	Yields	Credit Spreads	Factor	Housing	Yields	Credit Spreads
A. Fama-French six-factor model				**B. Augmented q-factor model**			
Ann. α	8.56%	−0.90%	−8.65%	Ann. α	7.29%	−0.95%	−10.15%
	[4.45]***	[−0.67]	[−2.89]***		[2.72]***	[−0.57]	[−3.19]***
MKT	0.44	−0.49	0.01	MKT	0.41	−0.48	0.03
	[10.27]***	[−12.52]***	[0.09]		[8.61]***	[−12.00]***	[0.41]
SMB	0.51	0.14	−1.00	ME	0.59	0.11	−0.86
	[8.82]***	[2.80]***	[−10.63]***		[9.18]***	[2.01]**	[−8.66]***
HML	0.26	−0.38	−0.01	IA	0.46	−0.58	−0.13
	[3.54]***	[−5.36]***	[−0.07]		[4.08]***	[−7.42]***	[−0.89]
RMW	0.40	−0.07	0.16	ROE	0.21	0.01	0.66
	[5.07]***	[−1.19]	[1.46]		[2.10]**	[0.13]	[4.79]***
CMA	0.31	−0.24	−0.24	EG	0.36	−0.03	−0.11
	[2.51]**	[−2.65]***	[−1.39]		[2.99]***	[−0.38]	[−0.70]
MOM	0.37	0.005	0.10				
	[6.57]***	[0.11]	[1.17]				
R^2	41.01%	42.04%	26.99%	R^2	30.73%	33.89%	29.42%
GRS	20.11***			GRS	12.14***		
$\overline{\lvert\alpha\rvert}$	6.04%			$\overline{\lvert\alpha\rvert}$	6.13%		

(7.29% and −10.15%, respectively). In sum, the spanning regression results indicate that the sparse macro-finance factors contain information regarding systematic risks beyond that in the FF6 and Q5 models.

4.3. Bayesian Model Selection

To further compare the sparse macro-finance factors to the characteristic-based factors, we employ the Bayesian risk factor selection methodology developed by Chib and Zeng (2020) and Chib, Zeng, and Zhao (2020). Their elegant framework allows us to analyze the set of sparse macro-finance and characteristic-based factors together in a unified framework.

We begin by collecting the entire set of factors in the D-dimensional vector $\boldsymbol{f}_t$. In our application, after collecting the three sparse macro-finance factors along with all of the factors in the FF6 and Q5 models, $D = 13$. The factors are then divided into risk and nonrisk categories, with the division defining a given model. The risk factors in a given model are the factors that appear in the SDF. Considering all possible divisions of the 13 factors yields $Q = 2^D - 1 = 8{,}191$ potential models (since the model with no risk factors is excluded). Indexing a given model by q (for $q = 1, \ldots, Q$), $\boldsymbol{f}_t$ is partitioned as $\boldsymbol{f}_t = [\ \boldsymbol{z}'_{q,t} \quad \boldsymbol{w}'_{q,t}\ ]'$, where $\boldsymbol{z}_{q,t}$ ($\boldsymbol{w}_{q,t}$) is the $d_{z,q}$ ($d_{w,q}$) -dimensional vector of risk (nonrisk) factors for model q, and $d_{z,q} + d_{w,q} = D$. The factors are assumed to follow a multivariate Student-t distribution:

$$\boldsymbol{f}_t \sim t_D(\boldsymbol{\mu}_q, \boldsymbol{\Omega}_q, \nu_f), \tag{9}$$

where $\boldsymbol{\mu}_q = [\ \boldsymbol{\mu}'_{z,q} \quad \boldsymbol{\mu}'_{w,q}\ ]'$ is the D-dimensional mean vector for $\boldsymbol{f}_t$, $\boldsymbol{\mu}_{z,q}$ ($\boldsymbol{\mu}_{w,q}$) is the $d_{z,q}$ ($d_{w,q}$) -dimensional mean vector for $\boldsymbol{z}_{q,t}$ ($\boldsymbol{w}_{q,t}$), $\boldsymbol{\Omega}_q$ is the $D \times D$ scale matrix, and $\nu_f > 2$ is the degrees of freedom. Relative to a Gaussian distribution, the Student-t distribution allows for fatter tails in $\boldsymbol{f}_t$, in line with factor data. Letting $m_{q,t}$ denote the SDF for model q, after factoring Equation (9) into the marginal distribution of $\boldsymbol{z}_{q,t}$ and the distribution of $\boldsymbol{w}_{q,t}$ conditional on $\boldsymbol{z}_{q,t}$, the vector of asset pricing restrictions, $E(m_{q,t}\boldsymbol{f}_t) = \boldsymbol{0}_D$, implies that the marginal and conditional distributions are given by

$$\boldsymbol{z}_{q,t} = \boldsymbol{\gamma}_{z,q} + \boldsymbol{\varepsilon}_{z,q,t}, \tag{10}$$

$$\boldsymbol{w}_{q,t} = \boldsymbol{\Psi}_{w,q}\boldsymbol{z}_{q,t} + \boldsymbol{\varepsilon}_{w,q,t}, \tag{11}$$

respectively, where $\boldsymbol{\gamma}_{z,q}$ is the $d_{z,q}$-vector of risk premia for the risk factors in $\boldsymbol{z}_{q,t}$, $\boldsymbol{\Psi}_{w,q} = \boldsymbol{\Omega}'_{zw,q}\boldsymbol{\Omega}^{-1}_{z,q}$, $\boldsymbol{\Omega}_{zw,q}$ ($\boldsymbol{\Omega}_{z,q}$) is the $d_{z,q} \times d_{w,q}$ ($d_{z,q} \times d_{z,q}$) submatrix in the upper-right (left) part of $\boldsymbol{\Omega}_q$,

$$\boldsymbol{\varepsilon}_{q,t} = \begin{bmatrix} \boldsymbol{\varepsilon}'_{z,q,t} & \boldsymbol{\varepsilon}'_{w,q,t} \end{bmatrix}' \sim t_D(\mathbf{0}_D, \boldsymbol{\Omega}_{\varepsilon,q}, \nu_f), \tag{12}$$

$\boldsymbol{\Omega}_{\varepsilon,q}$ is a $D \times D$ block-diagonal scale matrix whose nonzero elements depend on the elements in $\boldsymbol{\Omega}_q$, and the SDF defined in terms of $\boldsymbol{z}_{q,t}$ is

$$m_{q,t} = 1 - [(\nu_f - 2)/\nu_f]\boldsymbol{\gamma}'_{z,q}\boldsymbol{\Omega}^{-1}_{z,q}(\boldsymbol{z}_{q,t} - \boldsymbol{\mu}_{z,q}). \tag{13}$$

Chib and Zeng (2020, Section 2.1) provides details for deriving Equations (10) to (12).[21]

Chib and Zeng (2020) propose a set of proper priors for the parameters that are comparable across the different models indexed by q as well as an efficient Markov chain Monte Carlo (MCMC) algorithm to simulate sampling the parameters from the posterior distribution. They use methods from Chib (1995) and Chib and Jeliazkov (2001) to compute the posterior marginal likelihood for each of the Q models, which can be used for model selection and comparison. The priors are such that they play a minimal role in affecting the marginal likelihood so that differences in marginal likelihoods primarily reflect differences in the ability of the models to explain the data. The marginal likelihood implicitly penalizes model complexity, meaning that it balances fit and complexity. In addition to helping guard against overfitting, this is an attractive feature, as researchers typically view relatively parsimonious models of systematic risks in the macroeconomy as more plausible.[22]

[21] In terms of the mean vectors for the risk and nonrisk factors, the asset pricing restrictions imply that $\boldsymbol{\mu}_{z,q} = \boldsymbol{\gamma}_{z,q}$ and $\boldsymbol{\mu}_{w,q} = \boldsymbol{\Psi}_{w,q}\boldsymbol{\gamma}_{z,q}$, respectively.

[22] For the Barillas and Shanken (2017) and Chib and Zeng (2020) approaches, the set of test assets does not matter for comparing factor models with traded factors (like the market and characteristic-based factors). For nontraded factors (like the sparse macro-finance factors) that need to be converted to traded factors, the test assets matter insofar as different sets of test assets can generate different mimicking portfolios. Because we use an extensive set of test assets to construct the mimicking portfolios, this should play a limited role in our analysis.

Table 5. Top 20 risk factor models

The table reports the risk factors selected by the Chib and Zeng (2020) Bayesian procedure for the top 20 models ranked according to their posterior log marginal likelihoods. ★ indicates that the factor is included as a risk factor in the model. All possible models formed by combining 13 factors are considered. The total number of model considered is $2^{13} - 1 = 8{,}191$ (since the model with no risk factors is not considered). The 13 factors are three sparse macro-finance factors and factors from the Fama-French six-factor (Fama and French 2018) and augmented q-factor (Hou et al. 2021) models: HSG = housing factor, YLD = yield factor, CS = credit spread factor, MKT = market factor, SMB = small-minus-big size factor, HML = high-minus-low value factor, RMW = robust-minus-weak profitability factor, CMA = conservative-minus-aggressive investment factor; MOM = momentum factor, ME = size factor, IA = investment factor, ROE = return on equity factor, EG = expected investment growth factor. Log(ML) is the log marginal likelihood. The sample period is 1967:01 to 2023:12, where the first 10% of observations are reserved as a training sample for constructing the prior distributions. The results are based on 10,000 MCMC draws after 1,000 burn-in iterations.

(1) Rank	(2) HSG	(3) YLD	(4) CS	(5) MKT	(6) SMB	(7) HML	(8) RMW	(9) CMA	(10) MOM	(11) ME	(12) IA	(13) ROE	(14) EG	(15) Log(ML)
1	★	★	★									★	★	20,125.02
2	★	★	★				★					★	★	20,124.42
3	★	★	★		★							★	★	20,124.25
4	★	★	★							★		★	★	20,123.75
5	★	★	★	★								★	★	20,123.60
6	★	★	★		★		★					★	★	20,123.33
7	★	★	★						★			★	★	20,123.26
8	★	★	★	★			★					★	★	20,123.26
9	★	★	★								★	★	★	20,123.18
10	★	★	★					★				★	★	20,123.15
11	★	★	★			★						★	★	20,123.09
12	★	★	★		★					★		★	★	20,123.02
13	★	★			★							★	★	20,122.95
14	★	★	★		★		★			★		★	★	20,122.90
15	★	★	★	★	★							★	★	20,122.84
16	★	★	★				★			★		★	★	20,122.80
17	★	★	★				★	★				★	★	20,122.59
18	★	★	★				★				★	★	★	20,122.58
19	★	★	★				★		★			★	★	20,122.56
20	★	★	★			★	★					★	★	20,122.52

Table 5 reports the selected risk factors for the top 20 models ranked according to their posterior marginal likelihoods, where $\nu_f = 5$.[23] The top-ranked model includes five risk factors: the three sparse macro-finance factors—housing, yields, and credit spreads—as well as the return on equity and expected investment growth factors from the Q5 model. Thus, the "best" model is a blend of sparse macro-finance and characteristic-based factors. Remarkably, all 20 of the top-ranked models select the housing, yield, return on equity, and expected

[23]The initial 10% of observations (1967:01 to 1972:07) are reserved for constructing the prior distributions. Details for constructing the priors are provided in Chib and Zeng (2020, Section 2.3).

investment growth factors, while 19 of the 20 top-ranked models also select the credit spread factor. Apart from these five factors, the factor that is selected most frequently among the top 20 models is the profitability factor from the FF6 model, which is selected nine times. Interestingly, the market factor is only selected three times among the top 20 models, despite its canonical status. The top 20 models are all fairly parsimonious, selecting between four and seven of the 13 factors as risk factors, with a mean of 5.45 selected factors.

Table 6. Bayes factors

The table reports Bayes factors for different risk factor models based on posterior log marginal likelihoods computed using the Chib and Zeng (2020) Bayesian procedure. Top ranked (Top ranked plus MKT) is the model ranked first (fifth) in Table 5. FF6 is the Fama-French six-factor model (Fama and French 2018) that includes market, size, value, profitability, investment, and momentum factors. Q5 is the augmented q-factor model (Hou et al. 2021) that includes market, size, investment, return on equity, and expected investment growth factors. Parenthesis report the log marginal likelihood for the model. BF_{12} is the Bayes factor for the model in the first column relative to the model in the second column. The interpretation in the fourth column is based on Jeffreys (1939). The sample period is 1967:01 to 2023:12, where the first 10% of observations are reserved as a training sample for constructing the prior distributions. The results are based on 10,000 MCMC draws after 1,000 burn-in iterations.

(1) Model 1	(2) Model 2	(3) BF_{12}	(4) Interpretation
Top ranked (20,125.02)	Top ranked plus MKT (20,123.60)	4.14	Moderate evidence for Model 1
Top ranked (20,125.02)	FF6 (20,039.29)	> 100	Extreme evidence for Model 1
Top ranked (20,125.02)	Q5 (20,093.40)	> 100	Extreme evidence for Model 1
FF6 (20,039.29)	Q5 (20,093.40)	< 0.01	Extreme evidence for Model 2

Next, we compare various models using Bayes factors. The Bayes factor for comparing a pair of models is the marginal likelihood of one model divided by that of the other model. Table 6 reports the Bayes factor for the model in the first column vis-à-vis the model in the second column. Given the prominence of the market factor, we begin by comparing the top-ranked model in Table 5 with the model that includes the same factors plus the market factor (the fifth-ranked model). The Bayes factor is 4.14 in Table 6, which, according to

the Jeffreys (1939) interpretive scale, provides moderate evidence in favor of the top-ranked model.[24] When comparing the top-ranked model with the FF6 and Q5 models, the Bayes factors are both greater than 100, so there is extreme evidence in favor of the top-ranked model relative to the two competing models. Comparing the FF6 and Q5 models, the Bayes factor provides strong support for the latter, in line with the spanning regression results in Hou et al. (2019). While the Q5 model dominates the FF6 model, both models are dominated by the top-ranked model.

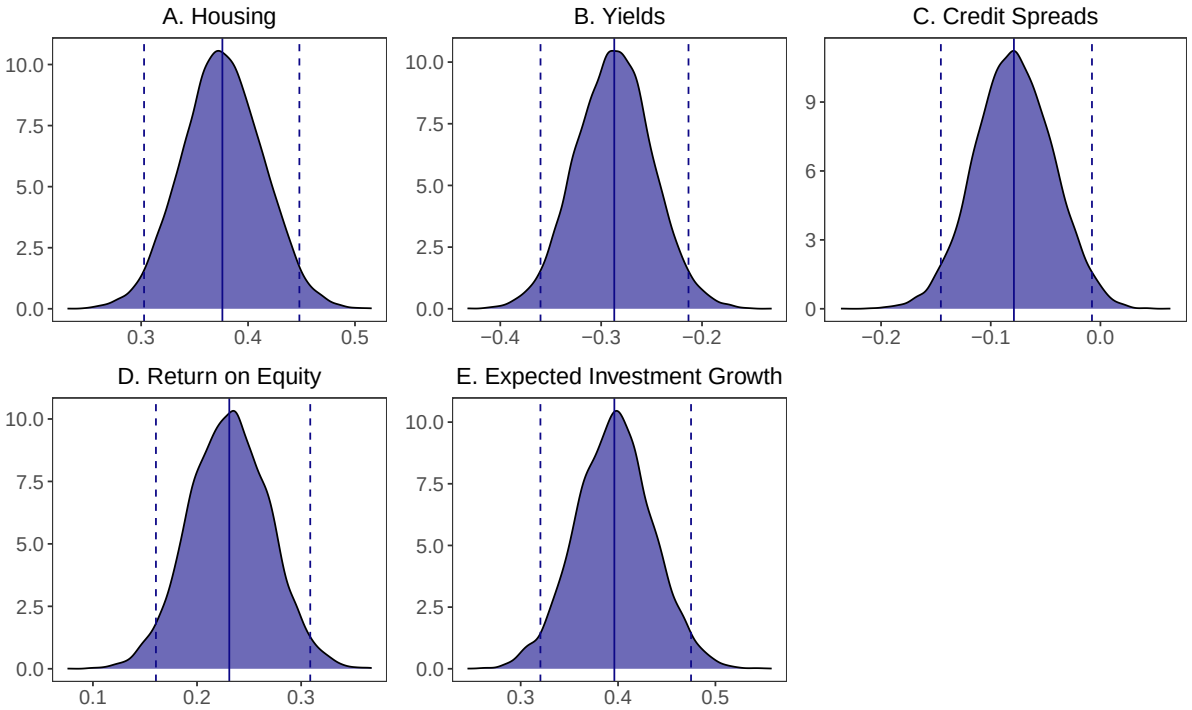


Figure 6. SDF loadings for the top-ranked model. Each panel depicts the posterior distribution for the SDF loading for the risk factor in the panel heading. Vertical solid (dashed) lines delineate the posterior mean (95% highest-density intervals). The results are based on 10,000 MCMC draws after 1,000 burn-in iterations.

Overall, three new sparse macro-finance factors—housing, yields, and credit spreads—and two characteristic-based factors from the Q5 model—return on equity and expected investment growth—stand out as relevant risk factors in Tables 5 and 6, including consti-

[24] Letting ML_1 (ML_2) denote the marginal likelihood for model 1 (2), it is straightforward to compute the Bayes factor for model 1 relative to model 2 from the difference in log marginal likelihoods using $\exp[\log(\text{ML}_1) - \log(\text{ML}_2)] = \exp[\log(\text{ML}_1/\text{ML}_2)] = \text{ML}_1/\text{ML}_2$.

tuting the five risk factors in the top-ranked model. We take a closer look at the top-ranked model through the lens of its SDF in Equation (13). Figure 6 depicts posterior distributions for the elements of $\boldsymbol{b}_q = \boldsymbol{\Omega}_{z,q}^{-1}\boldsymbol{\gamma}_{z,q}$, the vector of SDF loadings in Equation (13). The SDF loadings are all significant in the sense that none of the 95% highest-density intervals cover zero. The housing, return on equity, and expected investment growth factors have positive SDF loadings in Panels A, D, and E, respectively, of Figure 6. According to Equation (13), a decrease in each of these factors leads to an increase in the SDF (corresponding to a bad state of the world), in line with the positive risk premium estimates for these factors in the second column of Table 3. In contrast, the SDF loadings are negative for the yield and credit spread factors in Panels B and C, respectively, of Figure 6. Thus, an increase in each of these two factors contributes to an increase in the SDF, in agreement with the negative risk premium estimates for these factors in Table 3.[25]

Figure 7 plots the posterior distribution for the (log) SDF itself in Equation (13) for the top-ranked model for 1972:08 to 2023:12.[26] The figure paints a highly plausible picture of the SDF. The SDF evinces sharp upward spikes at business-cycle peaks that mark the start of a recession, while it exhibits downward spikes near business-cycle troughs that define the beginning of a new expansion. A generally elevated SDF is evident around 1998, coinciding with the Asian financial and Long-Term Capital Management crises. The SDF typically increases from about 2005 to 2008 in conjunction with the global financial crisis, and it spikes upward in late 2008 during the worst phase of the crisis. Toward the end of the sample, the SDF evinces a sizable decline from mid 2020 through 2021 as the economy recovered from the COVID-19 crisis.

[25]With respect to $\boldsymbol{\gamma}_{z,q}$, the posterior means for the annualized risk premia are as follows: housing, 19.78%; yields, −7.66%; credit spreads, −5.62%; return on equity, 7.69%; expected investment growth, 8.96%. The posteriors means for the annualized Sharpe ratios for the factors are as follows: housing, 1.33; yields, −0.63; credit spreads, −0.26; return on equity, 0.98; expected investment growth, 1.34. All of the estimated risk premia and Sharpe ratios for the factors are significant according to their 95% highest-density intervals.

[26]Recall that the initial 10% of observations are reserved for constructing the prior distributions; see Footnote 23.

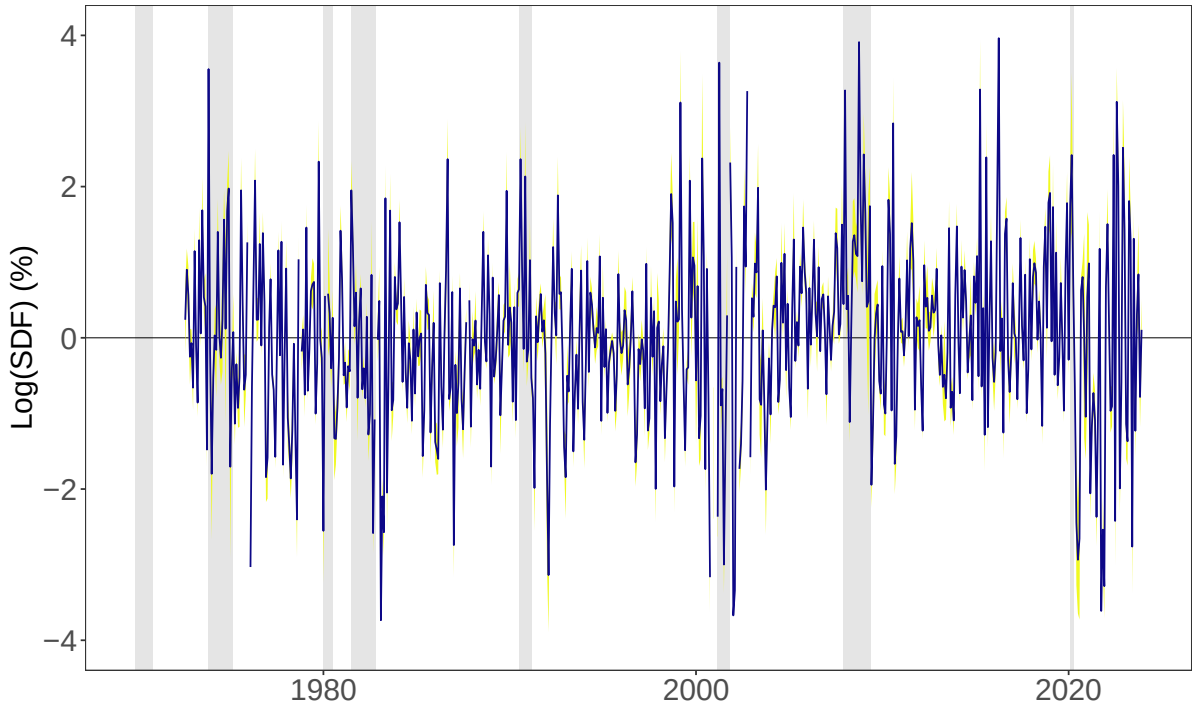


Figure 7. SDF for the top-ranked model. The figure depicts the mean (dark line) and 95% highest-density intervals (light bands) for the posterior distribution for the log SDF for the top-ranked model. The results are based on 10,000 MCMC draws after 1,000 burn-in iterations. Vertical bars delineate business-cycle recessions as dated by the National Bureau of Economic Research.

Finally, Figure 8 depicts posterior distributions for the tangency portfolio annualized Sharpe ratios for the top-ranked model and the FF6 and Q5 models for 1972:08 to 2023:12. For model q, the $d_{z,q}$-dimensional vector of tangency portfolio weights is given by

$$\boldsymbol{\omega}_{z,q} = \frac{\boldsymbol{\Omega}_{z,q}^{-1}\boldsymbol{\gamma}_{z,q}}{\boldsymbol{\iota}_{d_{z,q}}'\boldsymbol{\Omega}_{z,q}^{-1}\boldsymbol{\gamma}_{z,q}}, \tag{14}$$

where $\boldsymbol{\iota}_{d_{z,q}}$ is a $d_{z,q}$-dimensional vector of ones. The posterior mean for the tangency portfolio Sharpe ratio is 2.08 for the top-ranked model, while those for the FF6 and Q5 models are 1.11 and 1.90, respectively. According to the 95% highest-density intervals, the tangency portfolio Sharpe ratio for the top-ranked model is significantly larger than those for the other

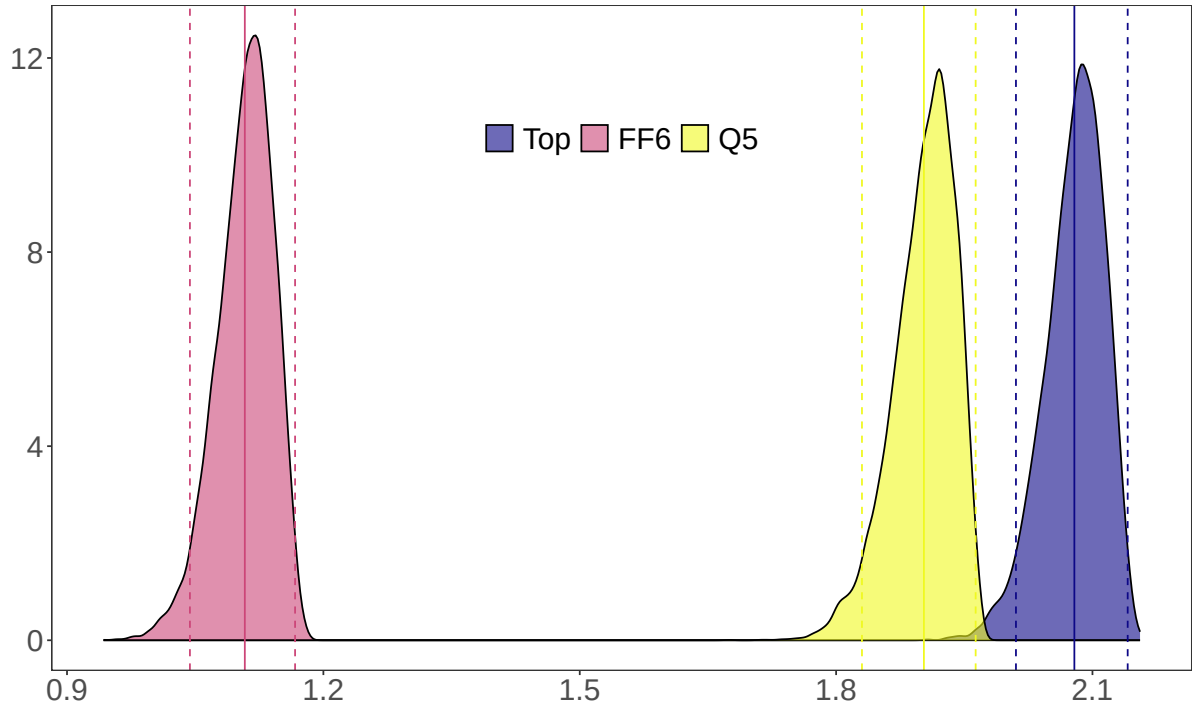


Figure 8. Tangency portfolio annualized Sharpe ratios. The figure depicts posterior distributions for the tangency portfolio annualized Sharpe ratios for the top-ranked model (Top), the Fama-French six-factor model (FF6), and the augmented q-factor model (Q5). Vertical solid (dashed) lines delineate the posterior mean (95% highest-density intervals). Each posterior distribution is based on 10,000 MCMC draws after 1,000 burn-in iterations.

two models, providing further evidence of the relevance of the sparse macro-finance factors for modeling systematic risks.[27]

5. Updating the Sparse Macro-Finance Factors

The sparse macro-finance factors presented in this paper, as well as the data and code used to generate them, are available on the Sparse Macro-Finance Factors website. We plan to update the sparse factors annually on the website. Updating the sparse factors is more complicated than updating characteristic-based factors, as subsequently explained.

To compute the sparse macro-finance factors and their mimicking portfolios, we need to extract the sparse components from the 118 macro-finance variables, fit a VAR(1) to

[27]Based on the 95% highest-density intervals, the tangency portfolio Sharpe ratio for the Q5 model is significant larger than that for the FF6 model.

the sparse components to compute the sparse factors, and estimate the mimicking-portfolio weights using the sparse factors and test asset excess returns. This means that the entire time series of the sparse factors and their mimicking portfolios can change when we estimate them with updated data, as the fitted nonzero sparse loadings and mimicking-portfolio weights can change. However, since we use a relatively long sample period in this paper, the fitted nonzero loadings and mimicking-portfolio weights should not change very much as we include twelve-month increments of additional data in the updates. For each update (or vintage), we will provide the entire time series of mimicking-portfolio returns for the sparse macro-finance factors. Section A.2 in the Internet Appendix provides evidence indicating that the mimicking-portfolio returns for the sparse factors should be quite stable for overlapping dates as we add twelve-month increments of data.

6. Conclusion

We develop a new strategy for estimating tradable macro-finance factors for asset pricing that, for the first time, rival the characteristic-based factors in leading models, including the Fama and French (2018) six-factor and Hou et al. (2021) augmented q-factor models. We begin by extracting twelve sparse components via SCA (Chen and Rohe 2023) from a large set of macro-finance variables. The sparse components can be viewed as potential state variables, and we compute innovations to the state variables by fitting a VAR(1) to the sparse components and taking the fitted residuals, providing us with a set of sparse macro-finance factors. We subsequently apply SPCA (Giglio, Xiu, and Zhang 2025) to a large set of test assets to estimate risk premia and mimicking portfolios for the sparse factors.

To implement our strategy, we need to select a number of hyperparameters, including the degree of sparsity to impose for SCA. Using cross validation based on the average R^2 statistic for the mimicking-portfolio regressions, we find that the asset pricing data favor a high degree of sparsity. The high degree of sparsity has keen advantages, as it substantially

improves the economic interpretation of the sparse components extracted from the macro-finance variables and provides sharper signals for asset pricing by helping to guard against overfitting. We find that sparse macro-finance factors corresponding to housing, yields, and credit spreads earn significant risk premia in the equity market. The importance of these factors is economically intuitive, as housing and bonds are substantial components of wealth for many investors, and investors value access to credit to take advantage of a new investment opportunity or weather a difficult financial situation. The housing factor commands a positive risk premium, while the yield and credit spread factors earn negative premia; consequently, stocks with positive exposures to the housing (yield and credit spread) factor tend to perform poorly (well) during bad times, so they receive a positive (negative) risk premium.

We compare the new housing, yield, and credit spread factors to leading characteristics-based factors appearing in the Fama and French (2018) six-factor and Hou et al. (2021) augmented q-factor models. Using the Barillas and Shanken (2017) approach, we find that neither the Fama-French six-factor nor the augmented q-factor model can price the sparse macro-finance factors, signifying that the sparse factors are informative for measuring systematic risks relative to characteristic-based factors. Furthermore, we use the powerful Bayesian methodology of Chib and Zeng (2020) and Chib, Zeng, and Zhao (2020) to analyze the entire set of sparse macro-finance and characteristic-based factors taken together. The top-ranked model according to the posterior marginal likelihood includes the housing, yield, and credit spread factors together with the return on equity and expected investment growth factors from the augmented-q model. Strikingly, the top 20 models always select the housing and yield factors (as well as the return on equity and expected investment growth factors) as relevant risk factors, and 19 of the top 20 models also select the credit spread factor. Bayes factors strongly favor the top-ranked model over both the Fama-French six-factor and augmented q-factor models. The SDF for the top-ranked model appears highly plausible,

evincing marked increases during recessions and financial crises, and the tangency portfolio annualized Sharpe ratio for the top-ranked model is above two.

Major advances have been made in the development of characteristic-based multifactor models, including the Fama-French six-factor and augmented q-factor models, and such models now predominate empirical models of systematic risks. We present strong evidence that our new housing, yield, and credit spread factors are relevant for capturing systematic risks. As researchers continue the quest for the "ultimate" empirical model of systematic risks in the macroeconomy, our results indicate that including sparse macro-finance factors along with characteristic-based factors offers a promising path forward.

References

Adrian, T., E. Etula, and T. Muir (2014). "Financial Intermediaries and the Cross-Section of Asset Returns." *Journal of Finance* 69:6, 2557–2596.

Balduzzi, P. and C. Robotti (2008). "Mimicking Portfolios, Economic Risk Premia, and Tests of Multi-Beta Models." *Journal of Business & Economic Statistics* 26:3, 354–368.

Barillas, F. and J. Shanken (2017). "Which Alpha?" *Review of Financial Studies* 30:4, 1316–1338.

Bianchi, D., M. Büchner, and A. Tamoni (2021). "Bond Risk Premiums with Machine Learning." *Review of Financial Studies* 34:2, 1046–1089.

Borup, D. and E. C. M. Schütte (2022). "In Search of a Job: Forecasting Employment Growth Using Google Trends." *Journal of Business & Economic Statistics* 40:1, 186–200.

Breeden, D. T., M. R. Gibbons, and R. H. Litzenberger (1989). "Empirical Tests of the Consumption-Oriented CAPM." *Journal of Finance* 44:2, 231–262.

Carhart, M. M. (1997). "On Persistence in Mutual Fund Performance." *Journal of Finance* 52:1, 57–82.

Chamberlain, G. and M. Rothschild (1983). "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets." *Econometrica* 51:5, 1281–1304.

Chen, A. Y. and T. Zimmermann (2022). "Open Source Cross-Sectional Asset Pricing." *Critical Finance Review* 27:2, 207–264.

Chen, F. and K. Rohe (2023). "A New Basis for Sparse Principal Component Analysis." *Journal of Computational and Graphical Statistics* 33:2, 421–434.

Chen, N.-F., R. Roll, and S. A. Ross (1986). "Economic Forces and the Stock Market." *Journal of Business* 59:3, 383–403.

Chernov, M., L. A. Lochstoer, and D. Song (2025). "The Real Channel for Nominal Bond-Stock Puzzles." *Journal of Finance*, forthcoming.

Chib, S. (1995). "Marginal Likelihood from the Gibbs Output." *Journal of the American Statistical Association* 90:432, 1313–1321.

Chib, S. and I. Jeliazkov (2001). "Marginal Likelihood From the Metropolis–Hastings Output." *Journal of the American Statistical Association* 96:453, 270–281.

Chib, S. and X. Zeng (2020). "Which Factors Are Risk Factors in Asset Pricing? A Model Scan Framework." *Journal of Business and Economic Statistics* 38:4, 771–783.

Chib, S., X. Zeng, and L. Zhao (2020). "On Comparing Asset Pricing Models." *Journal of Finance* 75:1, 551–577.

Cochrane, J. H. (1996). "A Cross-Sectional Test of an Investment-Based Asset Pricing Model." *Journal of Political Economy* 104:3, 572–621.

Cochrane, J. H. (2005). *Asset Pricing*. Revised Edition. Princeton, NJ: Princeton University Press.

Connor, G. and R. A. Koraczyk (1986). "Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis." *Journal of Financial Economics* 15:3, 373–394.

Connor, G. and R. A. Koraczyk (1988). "Risk and Return in an Equilibrium APT: Application of a New Test Methodology." *Journal of Financial Economics* 21:2, 255–289.

Croux, C., P. Filzmoser, and H. Fritz (2013). "Robust Sparse Principal Component Analysis." *Technometrics* 55:2, 202–214.

d'Aspremont, A., L. El Ghaoui, M. I. Jordan, and G. R. G. Lanckriet (2007). "A Direct Formulation for Sparse PCA Using Semidefinite Programming." *SIAM Review* 49:3, 434–448.

Erichson, N. B., P. Zeng, K. Manohar, S. L. Brunton, J. N. Kutz, and A. Y. Aravkin (2020). "Sparse Principal Component Analysis Via Variable Projection." *SIAM Journal of Applied Mathematics* 80:2, 977–1002.

Estrella, A. and G. A. Hardouvelis (1991). "The Term Structure as a Predictor of Real Economic Activity." *Journal of Finance* 26:2, 555–576.

Fama, E. F. and K. R. French (1993). "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics* 33:1, 3–56.

Fama, E. F. and K. R. French (2015). "A Five-Factor Asset Pricing Model." *Journal of Financial Economics* 116:1, 1–22.

Fama, E. F. and K. R. French (2018). "Choosing Factors." *Journal of Financial Economics* 128:2, 234–252.

Fama, E. F. and J. D. MacBeth (1973). "Risk, Return, and Equilibrium: Empirical Tests." *Journal of Political Economy* 81:3, 607–636.

Gibbons, M. R., S. A. Ross, and J. Shanken (1989). "A Test of the Efficiency of a Given Portfolio." *Econometrica* 57:5, 1121–1152.

Giglio, S. and D. Xiu (2021). "Asset Pricing with Omitted Factors." *Journal of Political Economy* 129:7, 1947–1990.

Giglio, S., D. Xiu, and D. Zhang (2025). "Test Assets and Weak Factors." *Journal of Finance* 80:1, 259–319.

Gospodinov, N., R. Kan, and C. Robotti (2014). "Misspecification-Robust Inference in Linear Asset-Pricing Models with Irrelevant Risk Factors." *Review of Financial Studies* 27:7, 2139–2170.

Hou, K., H. Mo, C. Xue, and L. Zhang (2019). "Which Factors?" *Review of Finance* 23:1, 1–35.

Hou, K., H. Mo, C. Xue, and L. Zhang (2021). "An Augmented q-Factor Model with Expected Growth." *Review of Finance* 25:1, 1–41.

Hou, K., C. Xue, and L. Zhang (2015). "Digesting Anomalies: An Investment Approach." *Review of Financial Studies* 28:3, 650–703.

Huberman, G., S. Kandel, and R. F. Stambaugh (1987). "Mimicking Portfolios and Exact Arbitrage Pricing." *Journal of Finance* 42:1, 1–9.

Hyland, C. C., T. Papamichalis, N. Romanidis, and D. P. Tsomocos (2025). "TA Macro-Finance Model of Credit Spreads." SSRN Working Paper No. 5313973.

Jagannathan, R. and Z. Wang (1996). "The Conditional CAPM and the Cross-Section of Expected Returns." *Journal of Finance* 51:1, 3–53.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2021). *An Introduction to Statistical Learning*. Second Edition. New York: Springer.

Jeffreys, H. (1939). *Theory of Probability*. Oxford University Press.

Jolliffe, I. T., N. T. Trendafilov, and M. Uddin (2003). "A Modified Principal Component Technique Based on the LASSO." *Journal of Computational and Graphical Statistics* 12:3, 531–537.

Kan, R., C. Robotti, and J. Shanken (2013). "Pricing Model Performance and the Two-Pass Cross-Sectional Regression Methodology." *Journal of Finance* 68:6, 2617–2649.

Kotchoni, R., M. Leroux, and D. Stevanovic (2019). "Macroeconomic Forecast Accuracy in a Data-Rich Environment." *Journal of Applied Econometrics* 34:7, 1050–1072.

Kozak, S., S. Nagel, and S. Santosh (2020). "Shrinking the Cross-Section." *Journal of Financial Economics* 135:2, 271–292.

Leamer, E. E. (2007). "Housing IS the Business Cycle." NBER Working Paper No. 13428.

Lettau, M. and S. Ludvigson (2001). "Resurrecting the (C)CAPM: A Cross-Sectional Test When Risk Premia Are Time-Varying." *Journal of Political Economy* 109:6, 1238–1287.

Ludvigson, S. C. and S. Ng (2007). "The Empirical Risk–Return Relation: A Factor Analysis Approach." *Journal of Financial Economics* 83:1, 171–222.

Ludvigson, S. C. and S. Ng (2009). "Macro Factors in Bond Risk Premia." *Review of Financial Studies* 22:12, 5027–5067.

Lustig, H. and S. G. Van Nieurwerburgh (2005). "Housing Collateral, Consumption Insurance and Risk Premia: An Empirical Perspective." *Journal of Finance* 60:3, 1167–1219.

Malloy, C. J., T. J. Moskowitz, and A. Vissing-Jørgensen (2009). "Long-Run Stockholder Consumption Risk and Asset Returns." *Journal of Finance* 64:6, 2427–2479.

McCracken, M. W. and S. Ng (2016). "FRED-MD: A Monthly Database for Macroeconomic Research." *Journal of Business and Economic Statistics* 34:4, 574–589.

Medeiros, M. C., G. F. R. Vasconcelos, Á. Veiga, and E. Zilberman (2021). "Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods." *Journal of Business & Economic Statistics* 39:1, 98–119.

Merton, R. C. (1973). "An Intertemporal Capital Asset Pricing Model." *Econometrica* 41:5, 867–887.

Neely, C. J., D. E. Rapach, J. Tu, and G. Zhou (2014). "Forecasting the Equity Risk Premium: The Role of Technical Indicators." *Management Science* 60:7, 1772–1791.

Newey, W. K. and K. D. West (1987). "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55:3, 703–708.

Parker, J. A. and C. Julliard (2005). "Consumption Risk and the Cross Section of Expected Returns." *Journal of Political Economy* 113:1, 185–222.

Pástor, L. and R. F. Stambaugh (2003). "Liquidity Risk and the Expected Stock Returns." *Journal of Political Economy* 111:3, 642–685.

Piazzesi, M. and M. Scheider (2016). "Housing and Macroeconomics." In: *Handbook of Macroeconomics*. Vol. 2. New York: Elsevier, pp. 1547–1640.

Ross, S. A. (1976). "The Arbitrage Theory of Capital Asset Pricing." *Journal of Economic Theory* 13:3, 341–360.

Rudebusch, G. D. and J. C. Williams (2009). "Forecasting Recessions: The Puzzle of the Enduring Power of the Yield Curve." *Journal of Business & Economic Statistics* 27:4, 493–503.

Shanken, J. and M. I. Weinstein (2006). "Economic Forces and the Stock Market Revisited." *Journal of Empirical Finance* 13:2, 129–144.

Shen, H. and J. Z. Huang (2008). "Sparse Principal Component Analysis Via Regularized Low Rank Matrix Approximation." *Journal of Multivariate Analysis* 99:6, 1015–1034.

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the LASSO." *Journal of the Royal Statistical Society. Series B (Methodological)* 58:1, 267–288.

Vassalou, M. (2003). "News Related to Future GDP Growth as a Risk Factor in Equity Returns." *Journal of Financial Economics* 68:1, 47–73.

Witten, D. M., R. Tibshirani, and T. Hastie (2009). "A Penalized Matrix Decomposition, with Applications to Sparse Principal Components and Canonical Correlation Analysis." *Biostatistics* 10:3, 515–534.

Yogo, M. (2006). "A Consumption-Based Explanation of Expected Stock Returns." *Journal of Finance* 61:2, 539–580.

Zou, H., T. Hastie, and R. Tibshirani (2006). "Sparse Principal Component Analysis." *Journal of Computational and Graphical Statistics* 15:2, 265–286.