Sparse Macro Factors

David E. Rapach
Washington University in St. Louis and
Saint Louis University
david.rapach@slu.edu

Guofu Zhou*
Washington University in St. Louis and CAFR
zhou@wustl.edu

Initial Version: September 4, 2018
Current Version: January 31, 2021

*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, Emory University, 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.

Sparse Macro Factors

Abstract

We use machine-learning techniques to estimate sparse principal components (PCs) for 120 monthly macroeconomic variables from the FRED-MD database. Each sparse PC is a sparse linear combination of the underlying macroeconomic variables, allowing for their economic interpretation. Innovations to the sparse PCs constitute a set of sparse macro factors. Robust tests indicate that sparse macro factors corresponding to yields and housing earn statistically and economically significant risk premia. A three-factor model comprised of the market factor and mimicking portfolio returns for the yields and housing factors performs well compared to leading multifactor models in explaining numerous anomalies.

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

Keywords: Sparse principal component analysis, FRED-MD, Risk premia, Factor-mimicking portfolio, Three-pass regression, Multifactor models

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) predicts that innovations to state variables that affect the "investment opportunity set" command risk premia for their exposure of investors to reinvestment risk. 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, p. 712). More generally, innovations to state variables that affect an investor's marginal utility of consumption potentially constitute systematic risk factors that require risk premia. In an unconditional form and with the absence of arbitrage, Ross (1976) shows that expected excess returns are linear functions of risk premia related to systematic risk factors.

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 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 series generate significant risk premia in cross-sectional equity returns (e.g., Cochrane 1996; Jagannathan and Wang 1996; Lettau and Ludvigson 2001; Vassalou 2003; Lustig and Van Nieurwerburgh 2005; Parker and Julliard 2005; Yogo 2006; Malloy, Moskowitz, and Vissing-Jørgensen 2009). However, the evidence that existing "macro factors" generate significant risk premia does not appear very robust, especially after accounting for potential model misspecification (e.g., Shanken and Weinstein 2006; Kan, Robotti, and Shanken 2013; Giglio and Xiu forthcoming). As the literature now stands, there is substantial uncertainty surrounding which macro factors are the most relevant for the cross section of equity returns, and models with macro factors are known to perform worse than leading characteristic-based factor models (e.g., Carhart 1997; Fama and French 2015; Hou, Xue, and Zhang 2015).

In this paper, we utilize advances in machine learning to compute *sparse macro factors* from a large macroeconomic database ("big data"). Existing macro factors from the literature are typically constructed using only one or two macroeconomic variables. However, a larger number of variables are likely relevant for measuring something like labor or housing market conditions, so that reliance on one or two individual variables is potentially arbitrary and ignores pertinent information. To incorporate information from numerous macroeconomic variables in diverse categories, we use a comprehensive set of monthly macroeconomic variables from the extensive FRED-MD database (McCracken and Ng 2016). We consider 120 variables representing manifold measures of a broad array of potentially relevant macroeconomic risks in categories such as output and income; the labor market; housing; consumption, orders, and inventories; money and credit; yields and exchange rates; and inflation. Employing machine-learning tools, we use *sparse principal component analysis* (e.g., Jolliffe, Trendafilov, and Uddin 2003; Zou, Hastie, and Tibshirani 2006; Shen and Huang 2008; Witten, Tibshirani, and Hastie 2009; Croux, Filzmoser, and Fritz 2013; Hubert et al. 2016; Erichson et al. 2020) to capture the systematic risks represented by the entire set of macroeconomic variables.

Conventional principal component analysis (PCA) owes its popularity to its ability to succinctly capture much of the information in a large number of variables. However, this comes at the cost of interpretability, as conventional principal components (PCs) are linear combinations of all of the underlying variables. Sparse PCA restricts the cardinality of the weight vectors, so that the PCs are sparse linear combinations of the underlying variables. By setting many weights to zero, sparse PCA facilitates interpretation of the PCs. The aim of sparse PCA is to improve interpretability via sparsity without unduly sacrificing explanatory ability. Sparse PCA has proven valuable for analyzing and interpreting data for diverse phenomena, such as gene expression and biological outcomes (e.g., Shen and Huang 2008; Witten, Tibshirani, and Hastie 2009) and the El Niño and La Niña warming events (e.g., Erichson et al. 2020). We show that sparse PCA is also valuable for analyzing and interpreting macroeconomic data in the context of cross-sectional asset pricing.

For comparison, we begin by computing the first nine conventional PCs for the 120 macroeconomic variables.[1] Each conventional PC is a linear combination of all 120 of the macroeconomic variables, which makes the individual PCs difficult to interpret. We then use the *penalized matrix decomposition* methodology of Witten, Tibshirani, and Hastie (WTH, 2009) to extract the first nine sparse PCs from the 120 macroeconomic variables. We impose substantive sparsity on the weight vectors: only 108 of the $120 \times 9 = 1{,}080$ weights (10%) are permitted to be nonzero (i.e., *active*). The high degree of sparsity facilitates economic interpretation of the sparse PCs; specifically, we interpret the individual sparse PCs as yields, production, inflation, housing, spreads, employment, costs, money, and credit. Based on vector autoregressions fitted to the conventional and sparse PCs (in turn), innovations to the conventional (sparse) PCs constitute a set of conventional (sparse) macro factors.

We estimate risk premia for the conventional and sparse macro factors using the robust three-pass methodology of Giglio and Xiu (forthcoming), which accounts for the omission of relevant risk factors and measurement error. In estimating the risk premia, we employ a large number (345) of equity portfolios formed on a variety of firm characteristics as test assets. Based on conventional PCA, none of the conventional macro factors earns a significant risk premium at the 5% level.[2] In contrast, based on sparse PCA, two of the macro factors—corresponding to yields and housing—earn significant risk premia at the 1% level. The three-pass methodology can also be used to compute mimicking portfolio returns for the sparse macro factors. Mimicking portfolios for the yields and housing factors deliver sizable Sharpe ratios in magnitude compared to well-known risk factors from the literature.

[1]Nine PCs are selected by the Bai and Ng (2002) modified information criterion.

[2]One of the conventional macro factors earns a significant risk premium at the 10% level.

A three-factor model comprised of the market factor and mimicking portfolio returns for the yields and housing factors performs on par with or better than the well-known multifactor models of Carhart (1997), Fama and French (2015), and Hou, Xue, and Zhang (2015) in explaining 315 anomaly portfolio returns from Chen and Zimmermann (2020). We also estimate regressions linking the yields and housing factors to the factors appearing in these well-known models. The results indicate that popular size, value, momentum, profitability, and return-on-equity factors reflect macroeconomic risks related to yields and housing. Based on the approach of Barillas and Shanken (2017), the sparse macro three-factor model is clearly favored over the Carhart (1997) four-factor model. The approach does not select either the sparse macro three-factor or Fama and French (2015) five-factor model over the other, and the situation is similar with respect to the Hou, Xue, and Zhang (2015) q-factor model.

Our new sparse macro three-factor model appears to represent a major advance in the macro-finance literature. It is the first multifactor model to combine the market factor with factors derived from macroeconomic variables that performs on par with leading multifactor models from the literature. The latter use security information alone to explain cross-sectional equity returns, while the yields and housing factors in the sparse macro three-factor model rely on macroeconomic data. Moreover, the sparse macro three-factor model is more parsimonious than leading multifactor models and has a straightforward macroeconomic interpretation. In sum, our new model constitutes a valuable benchmark that can be used to assess systematic risks in a variety of applications, in the same way as leading multifactor models from the literature.

The sparse macro three-factor model is rooted in PCA, which dates to Pearson (1901) and is the most widely used dimension-reduction technique (Hastie, Tibshirani, and Friedman 2009). Connor and Koraczyk (1986, 1988), Kelly, Pruitt, and Su (2020), Lettau and Pelger (2020a,b), Giglio and Xiu (forthcoming), Kim, Korajczyk, and Neuhierl (forthcoming), and Pelger and Xiong (forthcoming) are examples of extending PCA and applying it to finance. Our new sparse macro three-factor model combines sparse PCA and the Giglio and Xiu (forthcoming) three-pass methodology to identify the systematic risk factors emerging from a comprehensive set of macroeconomic variables. By simultaneously analyzing a large number of macroeconomic variables in a unified framework, the type of data-snooping bias highlighted by Harvey, Liu, and Zhu (2016) is less of a concern in our context.

Our results indicate that sparse PCA is a valuable machine-learning technique for extracting relevant asset pricing information from a large set of macroeconomic variables. In our application, compared to conventional PCA, sparse PCA provides substantial gains in terms of economic interpretation and has relatively little cost in terms of explaining the total variation in the macroeconomic variables. Moreover, sparse PCA provides important gains with respect to asset pricing: while none of the conventional macro factors earn a significant risk premium in the cross section of equity returns, the yields and housing sparse macro factors generate statistically and economically significant risk premia, even though data from cross-sectional returns are not directly incorporated into the construction of the sparse PCs. Sparse PCA thus appears better able than conventional PCA to filter the noise in macroeconomic variables to more reliably detect relevant signals for asset pricing.

The rest of the paper is organized as follows. Section 2 describes sparse PCA and reports results for conventional and sparse PCA applied to 120 macroeconomic variables from FRED-MD. Section 3 reports risk premium estimates based on the three-pass methodology. Section 4 compares the sparse macro three-factor model to leading multifactor models from the literature. Section 5 concludes.

2. Sparse Principal Component Analysis

Conventional PCA is a popular dimension-reduction technique.[3] Suppose that we are interested in reducing the dimension of a $T \times P$ data matrix, $\boldsymbol{X} = [\ \boldsymbol{x}_1 \ \cdots \ \boldsymbol{x}_P\ ]$, whose

[3] See Hastie, Tibshirani, and Wainwright (2015) for a textbook treatment of conventional and sparse PCA.

columns contain observations for P variables over T periods, where $\boldsymbol{x}_p = [\ x_{p,1} \ \ \cdots \ \ x_{p,T}\ ]'$ for $p = 1, \ldots, P$ and $\text{rank}(\boldsymbol{X}) = P < T$. We assume that the variables in $\boldsymbol{X}$ are standardized, so that each column of $\boldsymbol{X}$ has zero mean and unit variance. The PCs can be straightforwardly computed via the singular value decomposition of $\boldsymbol{X}$:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}', \tag{2.1}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are $T \times P$ and $P \times P$ orthonormal matrices, respectively (i.e., $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}_T$ and $\boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}_P$) and $\boldsymbol{D}$ is a $P \times P$ diagonal matrix with the singular values ($d_1 \geq d_2 \geq \cdots \geq d_P > 0$) along its main diagonal. The matrices of left and right singular vectors are given by $\boldsymbol{U} = [\ \boldsymbol{u}_1 \ \ \cdots \ \ \boldsymbol{u}_P\ ]$ and $\boldsymbol{V} = [\ \boldsymbol{v}_1 \ \ \cdots \ \ \boldsymbol{v}_P\ ]$, respectively, where $\boldsymbol{u}_p = [\ u_{p,1} \ \ \cdots \ \ u_{p,T}\ ]'$ and $\boldsymbol{v}_p = [\ v_{p,1} \ \ \cdots \ \ v_{p,P}\ ]'$ for $p = 1, \ldots, P$. Each PC is a linear combination of the variables in $\boldsymbol{X}$, with the weights given by the pth column of $\boldsymbol{V}$: $\boldsymbol{z}_p = [\ z_{p,1} \ \ \cdots \ \ z_{p,T}\ ]' = \boldsymbol{X}\boldsymbol{v}_p$ denotes the pth PC for $p = 1, \ldots, P$. The matrix of PCs, $\boldsymbol{Z} = [\ \boldsymbol{z}_1 \ \ \cdots \ \ \boldsymbol{z}_P\ ]$, can be expressed as $\boldsymbol{U}\boldsymbol{D}$, the matrix of scaled left singular vectors.

Following the exposition of WTH, it is well known that the following holds for any $\hat{p} \leq P$:

$$\sum_{p=1}^{\hat{p}} d_p \boldsymbol{u}_p \boldsymbol{v}_p' = \underset{\hat{\boldsymbol{X}} \in \boldsymbol{M}(\hat{p})}{\arg\min} \ \|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_F^2, \tag{2.2}$$

where $\|\cdot\|_F$ is the Frobenius norm and $\boldsymbol{M}(\hat{p})$ is the set of $T \times \hat{p}$ matrices with $\text{rank}(\hat{p})$. Equation (2.2) indicates that the first $\hat{p}$ columns of $\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}$ provide the best rank-$\hat{p}$ approximation to $\boldsymbol{X}$ in terms of the Frobenius norm. Intuitively, PCA uses the first $\hat{p}$ columns of $\boldsymbol{Z}$ to reduce the dimension of the data from P to $\hat{p} \ll P$, while capturing as much of the variation in the data as possible. Although conventional PCA is a valuable device for dimension reduction, the elements of $\boldsymbol{v}_p$ for $p = 1, \ldots, \hat{p}$ are all nonzero, so that the individual PCs can be difficult to interpret.[4] Sparse PCA harnesses machine-learning techniques to induce sparsity in the

[4] We can rotate the PCs using any full rank $\hat{p} \times \hat{p}$ matrix $\boldsymbol{H}$, and the rotated PCs and weights will explain the same variation in $\boldsymbol{X}$ as the original PCs and weights. However, the rotated PCs can still be difficult to interpret (e.g., Jolliffe 1995).

weight vectors. The goal is to facilitate the interpretability of the PCs without excessively sacrificing the ability of the PCs to capture the variation in the data.

WTH implement sparse PCA via penalized matrix decomposition. The idea is to introduce a penalty term into the objective function, as in the *least absolute shrinkage and selection operator* (LASSO, Tibshirani 1996). For the rank-one approximation to $\boldsymbol{X}$, consider the following objective function:

$$\min_{d_1,\boldsymbol{u}_1,\boldsymbol{v}_1} 0.5\|\boldsymbol{X} - d_1\boldsymbol{u}_1\boldsymbol{v}_1'\|_F^2 \text{ subject to } \|\boldsymbol{u}_1\|_2^2 = 1,\ \|\boldsymbol{v}_1\|_2^2 = 1,\ \|\boldsymbol{v}_1\|_1 \leq c,\ d \geq 0, \tag{2.3}$$

where $\|\boldsymbol{a}\|_1 = \sum_{i=1}^n |a_i|$ and $\|\boldsymbol{a}\|_2 = [\sum_{i=1}^n a_i^2]^{0.5}$ are the ℓ_1 and ℓ_2 norms, respectively, for a generic n-dimensional vector $\boldsymbol{a}$ and $c \geq 0$ is a shrinkage parameter. As $c \to \infty$, Equation (2.3) approaches Equation (2.2) for $\hat{p} = 1$. WTH show that Equation (2.3) is equivalent to the following rank-one penalized matrix decomposition:

$$\max_{\boldsymbol{u}_1,\boldsymbol{v}_1} \boldsymbol{u}_1'\boldsymbol{X}\boldsymbol{v}_1 \text{ subject to } \|\boldsymbol{u}_1\|_2^2 = 1,\ \|\boldsymbol{v}_1\|_2^2 = 1,\ \|\boldsymbol{v}_1\|_1 \leq c, \tag{2.4}$$

where the optimum value of d_1 is $\boldsymbol{u}_1'\boldsymbol{X}\boldsymbol{v}_1$. To employ methods for solving convex optimization problems, they consider the following biconvex version of Equation (2.4):

$$\max_{\boldsymbol{u}_1,\boldsymbol{v}_1} \boldsymbol{u}_1'\boldsymbol{X}\boldsymbol{v}_1 \text{ subject to } \|\boldsymbol{u}_1\|_2^2 \leq 1,\ \|\boldsymbol{v}_1\|_2^2 \leq 1,\ \|\boldsymbol{v}_1\|_1 \leq c. \tag{2.5}$$

WTH exploit the biconvexity of Equation (2.5) to develop an efficient iterative algorithm for computing $\boldsymbol{u}_1$, $\boldsymbol{v}_1$, and d_1. For sufficiently small c, the ℓ_1 penalty term in Equation (2.5) permits shrinkage to zero for one or more of the elements in $\boldsymbol{v}_1$, thereby generating a PC with a sparse weight vector.

The solution to Equation (2.5) provides $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$ (as well as $d_1 = \boldsymbol{u}_1'\boldsymbol{X}\boldsymbol{v}_1$), which correspond to the first sparse PC. For the remaining sparse PCs ($p = 2, \ldots, \hat{p}$), the iterative algorithm is used to compute the penalized matrix decomposition for $\boldsymbol{X}^{p+1} = \boldsymbol{X}^p - d_p\boldsymbol{u}_p\boldsymbol{v}_p'$, where $\boldsymbol{X}^1 = \boldsymbol{X}$. By appropriately setting the shrinkage parameter, c, we can induce a desired degree of sparsity in the weight vectors, $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{\hat{p}}$.[5]

2.1. Macroeconomic Data

We apply conventional and sparse PCA to 120 macroeconomic variables from the FRED-MD database (McCracken and Ng 2016). FRED-MD is a extensive macroeconomic database of monthly US variables compiled from the popular FRED database hosted by the Federal Reserve Bank of St. Louis.[6] We use 120 variables from the November 2020 vintage of FRED-MD that are available continuously starting in 1960. The variables cover a wide array of categories (output and income; labor market; housing; consumption, orders, and inventories; money and credit; interest and exchange rates; prices); as such, they capture much of the macroeconomic information available to investors.

Table 1 lists the 120 macroeconomic variables as defined by their FRED tickers and provides variable descriptions based on the Updated Appendix for the FRED-MD database.[7] Before conducting sparse PCA, we make two adjustments to the variables. First, where necessary, we transform the variables to render them stationary, as indicated in the second column of Table 1. Second, we adjust the variables for any lags in their reporting. Interest rates, exchange rates, interest rate spreads, and oil prices are reported without delay, so that 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 adjust-

[5] Unlike conventional PCA, sparse PCA does not necessarily produce orthogonal PCs. This is not a problem, as the estimation of risk premia in Section 3 does not require the PCs (or their innovations) to be uncorrelated.

[6] FRED is available at https://fred.stlouisfed.org/, while FRED-MD is available on Michael McCracken's webpage at https://research.stlouisfed.org/econ/mccracken/fred-databases/.

[7] The S&P 500 dividend yield (DIVYLD) is based on data from Robert Shiller's webpage at http://www.econ.yale.edu/~shiller/data.htm.

ments better reflect the flow of information to investors. After making the adjustments, our sample spans 1960:02 to 2019:12 (719 observations).

2.2. Sparse Principal Components

For comparison, we compute conventional and sparse PCs for the 120 macroeconomic variables. Table 2 reports weights for the first nine conventional PCs, while Figure 1 depicts the PCs themselves. Considering a maximum of 20, the Bai and Ng (2002) PC_{p_2} modified information criterion selects nine PCs ($\hat{p} = 9$). The first nine conventional PCs collectively explain 57% of the total variation in the macroeconomic variables. The weights in Table 2 illustrate the difficulty in interpreting conventional PCs. All of the weights are nonzero, and the weights for individual PCs are often sizable for variables across a variety of categories; for example, the first PC evinces relatively large weights (in magnitude) for variables related to employment, housing, yields, yield spreads, and prices. Similarly, the conventional PCs in Figure 1 reflect influences from amalgams of different types of variables. In sum, conventional PCs extracted from the 120 macroeconomic variables are difficult to interpret economically. Of course, conventional PCA maximizes the total variation in the underlying variables explained by the PCs, so that it is not designed to facilitate economic interpretation.

Table 3 and Figure 2 report sparse weights and PCs, respectively, computed via the WTH partial matrix decomposition methodology.[8] We set the shrinkage parameter so that only 10% of the $120 \times 9 = 1{,}080$ weights are active (i.e., nonzero). We thus induce substantive sparsity in the weight vectors, which enables us to intuitively interpret the nine sparse PCs as follows:

1. *Yields.* The second column of Table 3 shows that the first sparse PC is predominantly a linear combination of the nominal interest rates included in FRED-MD, so that we interpret the first sparse PC as "yields." In accord with this interpretation, Panel A

[8]We estimate the sparse weights using the `R` package `PMA` (Witten and Tibshirani 2020).

of Figure 2 shows that the first sparse PC follows well-known fluctuations in nominal interest rates over the postwar era.

2. *Production*. The second sparse PC is primarily a linear combination of most of the industrial production measures in FRED-MD, as well as manufacturing capacity utilization (see the third column of Table 3). Accordingly, we label the second sparse PC as "production," and it exhibits downward spikes during business-cycle recessions in Panel B of Figure 2.

3. *Inflation*. From the fourth column of Table 3, we see that the third sparse PC is essentially a linear combination of various producer and consumer price indices and personal consumption expenditure deflators. We thus label the third sparse PC as "inflation." Panel C of Figure 2 indicates that the second sparse PC displays the well-known "Great Inflation" of the 1970s and early 1980s and subsequent "Great Disinflation."

4. *Housing*. According to the fifth column of Table 3, the fourth sparse PC is a linear combination of the housing start and new private housing permit variables in FRED-MD (as well as real estate loans). We thus call this sparse PC "housing." The fourth sparse PC clearly depicts the housing market cycle in Panel D of Figure 2, 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.

5. *Spreads*. The fifth sparse PC is predominantly a linear combination of the interest rate spreads included in FRED-MD (see the sixth column of Table 3), so that we label this sparse PC as "spreads." The fifth sparse PC in Panel E of Figure 2 displays the distinct countercyclical fluctuations known to characterize yield spreads.

6. *Employment*. As shown in the seventh column of Table 3, the weights for the sixth sparse PC are concentrated in the unemployment and employment variables appear-

ing in FRED-MD. The sixth sparse PC, which we label as "employment," exhibits procyclical behavior in Panel F of Figure 2.

7. *Costs.* We refer to the seventh sparse PC as "costs," as the active elements of its weight vector include the various measures of average hourly earnings found in FRED-MD (see the eighth column of Table 3). More generally, the seventh sparse PC reflects drivers of costs for firms, including price indices for durable goods and medical care. This interpretation is reflected in Panel G of Figure 2, where the seventh sparse PC evinces a secular increase from the late 1960s to the early 1980s, in line with the sharp increases in costs experienced by firms during this period.

8. *Money.* As can be seen from the ninth column of Table 3, the active weights for the eighth sparse PC include all of the money stock measures in FRED-MD. Accordingly, we label this sparse PC as "money." The eighth sparse PC exhibits declines in the early 1980s in Panel H of Figure 2, corresponding to the "Volcker disinflation," as well as sharp increases more recently during the Global Financial Crisis, reflecting "quantitative easing" by the Fed.

9. *Credit.* The final sparse PC has relatively large loadings on credit variables from FRED-MD, including commercial and industrial loans, total nonrevolving credit, the credit-to-income ratio, consumer motor vehicle loans outstanding, and total consumer loans and leases outstanding (see the last column of Table 3). As shown in Panel I of Figure 2, the ninth sparse PC tends to fall during recessions, consistent with a "credit crunch" during economic contractions.

Comparing Tables 2 and 3 and Figures 1 and 2, the substantive sparsity imposed on the weight vectors greatly facilitates economic interpretation of the sparse PCs vis-à-vis the conventional PCs. Fortunately, the increased interpretability of the sparse PCs comes at relatively little cost in terms of explanatory ability: despite the high degree of sparsity, the first nine sparse PCs still explain 46% of the total variation in the 120 macroeconomic variables (compared to 57% for the first nine conventional PCs).

3. Risk Premia

To estimate risk premia, we begin by fitting first-order vector autoregressions to the set of conventional and sparse PCs (in turn), and we use the fitted processes to compute innovations to the conventional and sparse PCs. The innovations to the conventional (sparse) PCs constitute the set of conventional (sparse) macro factors. Table 4 reports correlations for the innovations to the conventional and sparse macro factors. Although the conventional PCs are uncorrelated by construction, the innovations to the conventional PCs in Panel A are correlated. However, most of the correlations are relatively small in magnitude. Similarly, the innovations to the sparse PCs are correlated in Panel B, but many of the correlations are limited in magnitude. The largest correlations (in magnitude) are for production-employment (0.50), housing-employment (0.21), and inflation-costs (0.20). In any event, asset pricing tests do not require uncorrelated macro factors.

We estimate risk premia for the (nontradable) macro factors via the recently developed three-pass methodology of Giglio and Xiu (forthcoming). Their methodology recovers the risk premium for any observable factor, regardless of whether the model includes all of the relevant risk factors, and is robust to measurement error. It also has an "ideal" mimicking portfolio interpretation; indeed, as a byproduct of the three-pass methodology, we can readily compute a mimicking portfolio return series for any nontradable factor.

To understand the three-pass methodology, consider the following linear factor model:

$$\boldsymbol{r}_t = \boldsymbol{\beta}'\boldsymbol{\gamma} + \boldsymbol{\beta}'\boldsymbol{\xi}_t + \boldsymbol{\varepsilon}_t, \tag{3.1}$$

where $\boldsymbol{r}_t = [\ r_{1,t} \quad \cdots \quad r_{N,t}\ ]'$ is the N-vector of period-t excess returns for $t = 1, \ldots, T$; $\boldsymbol{\gamma} = [\ \gamma_1 \quad \cdots \quad \gamma_K\ ]'$ is the K-vector of risk premia for the unobservable fundamental factors; $\boldsymbol{\xi}_t =$

$[\ \xi_{1,t}\ \ \cdots\ \ \xi_{K,t}\ ]'$ is the K-vector of fundamental factor innovations; $\boldsymbol{\beta} = [\ \boldsymbol{\beta}_1\ \ \cdots\ \ \boldsymbol{\beta}_N\ ]$ is the $K \times N$ matrix of fundamental factor exposures; $\boldsymbol{\beta}_i = [\ \beta_{i,1}\ \ \cdots\ \ \beta_{i,K}\ ]'$ for $i = 1, \ldots, N$; $\boldsymbol{\varepsilon}_t = [\ \varepsilon_{1,t}\ \ \cdots\ \ \varepsilon_{N,t}\ ]'$ is the N-vector of idiosyncratic errors; $E(\boldsymbol{\xi}_t) = \mathbf{0}_K$; $E(\boldsymbol{\varepsilon}_t) = \mathbf{0}_N$; and $\mathrm{Cov}(\varepsilon_{i,t}, \xi_{k,t}) = 0$ for $i = 1, \ldots, N$ and $k = 1, \ldots, K$. The D-vector of observable factors, which can be tradable or nontradable, is denoted by $\boldsymbol{g}_t = [\ g_{1,t}\ \ \cdots\ \ g_{D,t}\ ]'$ and governed by

$$\boldsymbol{g}_t = \boldsymbol{\delta} + \boldsymbol{\Theta}'\boldsymbol{\xi}_t + \boldsymbol{\zeta}_t, \tag{3.2}$$

where $\boldsymbol{\delta} = [\ \delta_1\ \ \cdots\ \ \delta_D\ ]'$ is the D-vector of intercept terms; $\boldsymbol{\Theta} = [\ \boldsymbol{\theta}_1\ \ \cdots\ \ \boldsymbol{\theta}_D\ ]$ is the $K \times D$ matrix of loadings for the observable factors on $\boldsymbol{\xi}_t$; $\boldsymbol{\theta}_d = [\ \theta_{d,1}\ \ \cdots\ \ \theta_{d,K}\ ]'$ for $d = 1, \ldots, D$; $\boldsymbol{\zeta}_t = [\ \zeta_{1,t}\ \ \cdots\ \ \zeta_{D,t}\ ]'$; $E(\boldsymbol{\zeta}_t) = \mathbf{0}_D$; and $\mathrm{Cov}(\zeta_{d,t}, \xi_{k,t}) = 0$ for $d = 1, \ldots, D$ and $k = 1, \ldots, K$.

We seek the D-vector of risk premia for the observable factors, $\boldsymbol{\gamma}_g = [\ \gamma_{g,1}\ \ \cdots\ \ \gamma_{g,D}\ ]'$, which, based on Equations (3.1) and (3.2), is given by $\boldsymbol{\gamma}_g = \boldsymbol{\Theta}'\boldsymbol{\gamma}$. Because the fundamental factors are unobservable, we cannot separately identify $\boldsymbol{\Theta}$ and $\boldsymbol{\gamma}$. However, using the property of *rotation invariance* for linear factor models, Giglio and Xiu (forthcoming) show that the product $\boldsymbol{\Theta}'\boldsymbol{\gamma}$ can be identified. Letting $\boldsymbol{H}$ be any full-rank $K \times K$ matrix, we can rewrite Equations (3.1) and (3.2) as

$$\boldsymbol{r}_t = \tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}} + \tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\xi}}_t + \boldsymbol{\varepsilon}_t, \tag{3.3}$$

$$\boldsymbol{g}_t = \boldsymbol{\delta} + \tilde{\boldsymbol{\Theta}}\tilde{\boldsymbol{\xi}}_t + \boldsymbol{\zeta}_t, \tag{3.4}$$

respectively, where $\tilde{\boldsymbol{\beta}} := \boldsymbol{\beta}\boldsymbol{H}^{-1}$, $\tilde{\boldsymbol{\gamma}} := \boldsymbol{H}\boldsymbol{\gamma}$, $\tilde{\boldsymbol{\xi}}_t := \boldsymbol{H}\boldsymbol{\xi}_t$, and $\tilde{\boldsymbol{\Theta}} := \boldsymbol{\Theta}'\boldsymbol{H}^{-1}$. We can apply conventional PCA to the demeaned excess returns to consistently estimate the rotated fundamental factors $\tilde{\boldsymbol{\xi}}_t$ in Equation (3.3) (Bai 2003). Via the familiar cross-sectional regression approach, we can consistently estimate $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\gamma}}$, while we can use time-series regressions to consistently estimate $\tilde{\boldsymbol{\Theta}}$ in Equation (3.4). With consistent estimates of $\tilde{\boldsymbol{\Theta}}$ and $\tilde{\boldsymbol{\gamma}}$ in hand, we can consistently estimate the vector of risk premia for the observable factors by taking the product of the estimates, since $\tilde{\boldsymbol{\Theta}}\tilde{\boldsymbol{\gamma}} = \boldsymbol{\Theta}'\boldsymbol{H}^{-1}\boldsymbol{H}\boldsymbol{\gamma} = \boldsymbol{\Theta}'\boldsymbol{\gamma}$.

In sum, the three-pass approach proceeds as follows:

1. Apply conventional PCA to demeaned excess returns for the N test assets to estimate the rotated fundamental factors $\tilde{\boldsymbol{\xi}}_t$ for $t = 1, \ldots, T$.[9]

2. Run time-series regressions of $r_{i,t}$ on $\tilde{\boldsymbol{\xi}}_t$ for $t = 1, \ldots, T$ and $i = 1, \ldots, N$ to estimate $\tilde{\boldsymbol{\beta}}$; run a cross-sectional regression of $\bar{\boldsymbol{r}} = [\ \bar{r}_1 \ \ \cdots \ \ \bar{r}_N\ ]'$ on the columns of the estimate of $\tilde{\boldsymbol{\beta}}$, where $\bar{r}_i = (1/T)\sum_{t=1}^{T} r_{i,t}$, to estimate $\tilde{\boldsymbol{\gamma}}$.

3. Run time-series regressions of $g_{d,t}$ on $\tilde{\boldsymbol{\xi}}_t$ for $t = 1, \ldots, T$ and $d = 1, \ldots, D$ to estimate $\tilde{\boldsymbol{\Theta}}$.

The product of the estimates of $\tilde{\boldsymbol{\Theta}}$ and $\tilde{\boldsymbol{\gamma}}$ from the third and second steps, respectively, provides the estimate of $\boldsymbol{\gamma}_g$. As a by-product of the three-pass methodology, the third step provides weights for "ideal" mimicking portfolios for the observable factors in $\boldsymbol{g}_t$.

3.1. Test Asset Data

We use 345 (value-weighted) portfolios as test assets, whose monthly returns are available from Kenneth French's Data Library:[10]

- 25 portfolios sorted on size and book-to-market value;
- 25 portfolios sorted on size and operating profitability;
- 25 portfolios sorted on size and investment;
- 25 portfolios sorted on size and momentum;
- 25 portfolios sorted on size and short-term reversal;

[9] Connor and Koraczyk (1986, 1988) pioneered the use of PCA to analyze excess returns.

[10] Available at http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

- 25 portfolios sorted on size and long-term reversal;
- 10 portfolios sorted on earnings-to-price ratio;
- 10 portfolios sorted on cash flow-to-price ratio;
- 10 portfolios sorted on dividend-price ratio;
- 25 portfolios sorted on size and accruals;
- 25 portfolios sorted on size and market beta;
- 35 portfolios sorted on size and net share issuance;
- 25 portfolios sorted on size and variance;
- 25 portfolios sorted on size and residual variance;
- 30 portfolios sorted on industry classification.

We compute the excess return for each portfolio using the CRSP risk-free return. The sample period covers 1963:07 to 2019:12 (678 observations).

3.2. Risk Premium Estimates

To help ensure that the principal components in the first step of the three-pass procedure span the relevant excess return space, we compute the first 15 PCs and corresponding loadings for the 345 test asset excess returns. Considering a maximum of 20, the Bai and Ng (2002) PC_{p_2} modified information criterion selects 15 PCs for the excess returns.

The first five columns of Table 5 report three-pass estimation results for the conventional macro factors. The t-statistics in the third column are based on hoeteroskedasticity- and autocorrelation-robust standard errors (Newey and West 1987). The R_g^2 statistics in the fourth column are the goodness-of-fit measures for the time-series regressions in the third step, while the fifth column reports annualized Sharpe ratios for the factor mimicking portfolios. According to the t-statistics in the third column of Table 5, none of the risk premia for the conventional macro factors are significant at the 5% level. Only one, corresponding to the eighth conventional PC, earns a significant risk premium at the 10% level. Based on the ninth column of Table 2 and Panel H of Figure 1, it is difficult to provide a straightforward economic interpretation of this factor.

Three-pass estimation results for the sparse macro factors are reported in the last five columns of Table 5. There is much stronger evidence of statistically and economically significant risk premia for some of the sparse macro factors. Specifically, based on the t-statistics in the eighth column, the risk premia for the yields and housing factors are significant at the 1% level. The yields and housing factors deliver annualized Sharpe ratios of -0.78 and 1.08, respectively, in the last column of Table 5. These Sharpe ratios are quite sizable relative to those for popular factors from the literature (e.g., market, size, value, and momentum factors).[11]

The significant risk premium earned by the yields factor in Table 5 accords with Merton (1973), who identifies the interest rate as a leading candidate for a state variable in the ICAPM, since changes in the interest rate affect the investment opportunity set. Our results further suggest that common fluctuations in multiple interest rates are relevant for asset pricing. The negative risk premium for the yields factor indicates that an increase in the yields factor corresponds to an increase in the marginal utility of consumption (i.e., a "bad" state of the world). Panel A of Figure 3 shows the log cumulative return for the mimicking portfolio for the yields factor. In accord with the factor's negative risk premium, the cumulative return has a negative trend, while the mid 1970s and early 1980s are characterized by notable increases in the cumulative return relative to its trend.

[11] In the spirit of Chen, Roll, and Ross (1986), we estimated risk premia for innovations to industrial production growth, the term spread (difference in yields between a ten-year Treasury bond and three-month Treasury bill), credit spread (difference in yields between BAA- and AAA-rated corporate bonds), and inflation. (The innovations are computed using fitted first-order autoregressions.) None of these variables earn a significant risk premium based on the three-pass methodology.

Because housing is a substantial component of wealth for many investors and an important part of the macroeconomy (e.g., Leamer 2007), the statistically and economically significant risk premium earned by the housing factor in Table 5 appears quite plausible. Its positive risk premium signals that an increase in the housing factor coincides with a decrease in the marginal utility of consumption (i.e., a "good" state of the world). Panel B of Figure 3 presents the log cumulative return for the housing factor's mimicking portfolio. The cumulative return has a positive trend, in line with its positive risk premium. Coinciding with the housing crisis, the cumulative return "flattens" in the mid 2000s and suffers a "crash" during the Global Financial Crisis and concomitant Great Recession.

In sum, sparse PCA offers two key advantages over conventional PCA in our application. First, by imposing substantive sparsity on the weight vectors, sparse PCA facilitates economic interpretation of the macro factors. Second, sparse PCA better captures the relevant information in macroeconomic variables for asset pricing. Specifically, sparse PCA proves superior to conventional PCA for identifying systematic risk factors in the macroeconomy. This result is striking, as the construction of the sparse weight vectors does not directly incorporate information from cross-sectional equity returns. It also helps to alleviate overfitting concerns when linking macroeconomic data to cross-sectional returns. Our results suggest that sparsity is a powerful machine-learning tool for identifying relevant asset pricing signals in macroeconomic variables.

4. Sparse Macro Three-Factor Model

In this section, we analyze the performance of a sparse macro three-factor model comprised of the market factor and mimicking portfolio returns for the yields and housing factors.

4.1. Pricing Errors

Similarly to Hou, Xue, and Zhang (2015, 2020), we examine the ability of our sparse macro three-factor (SM3) model to account for a plethora of anomalies in cross-sectional equity returns. We compare the performance of the model to three leading multifactor models from the literature: the Carhart (1997) four-factor (C4), Fama and French (2015) five-factor (FF5), and Hou, Xue, and Zhang (2015) q-factor (HXZQ) models. C4 augments the market, size, and value factors from Fama and French (1993) with a momentum factor (Jegadeesh and Titman 1993), while FF5 adds operating profitability and investment factors to the three Fama and French (1993) factors. Finally, HXZQ includes market, size, investment, and return-on-equity factors.

We estimate pricing errors for each multifactor model for 315 long-short anomaly portfolio returns from Chen and Zimmermann (2020).[12] The anomaly portfolio returns are based on nearly all of the characteristics considered in Harvey, Liu, and Zhu (2016), McLean and Pontiff (2016), Green, Hand, and Zhang (2017), and Hou, Xue, and Zhang (2020), so that we consider a comprehensive set of anomalies for measuring pricing errors. For 232 of the anomalies, the sample period is 1967:01 to 2019:12 (636 observations); for the other 83 anomalies, the number of available observations is less than 636.[13]

Table 6 reports means and standard deviations for the absolute values of the alphas (in percent) across the 315 portfolios for each of the multifactor models. The average magnitude of the alphas is largest for C4 (0.529%), followed by HXZQ and FF5 (0.513% and 0.509%, respectively). SM3 produces the smallest alpha in magnitude on average (0.478%). The standard deviations for the alpha magnitudes in the third column range from 0.419% (FF5) to 0.444% (C4). Figure 4 depicts histograms for the absolute values of the alphas for each of

[12]Available at `https://github.com/OpenSourceAP/CrossSection`.

[13]Because we do not have a long enough time series of balanced panel data for the anomaly portfolio returns, we cannot compute the Gibbons, Ross, and Shanken (1989) statistic. We compute this statistic when implementing the Barillas and Shanken (2017) approach in Section 4.2.

the multifactor models. The figure further indicates that SM3 performs comparatively well in accounting for anomalies.

The last four columns of Table 6 report the percent of t-statistics for the alphas that are greater than 1.645, 1.96, 2.58, and three, respectively, in magnitude. SM3 always produces the fewest rejections. For the Harvey, Liu, and Zhu (2016) threshold of three for the absolute value of the t-statistic, approximately a third of the 315 anomalies are significant, while around 50% to 60% are significant for the other multifactor models. Overall, SM3 performs on par with or better than leading multifactor models from the literature with respect to explaining numerous anomalies in cross-sectional equity returns. SM3 has the advantages of being more parsimonious (three factors instead of four or five) and having factors with straightforward macroeconomic interpretations.

The weights for the mimicking portfolios for the yields and housing factors in SM3 in Table 6 and Figure 4 are estimated using data through 2019:12. To investigate if the pricing ability of SM3 is robust to estimating the mimicking portfolio weights excluding data from the sample used to compute the pricing errors, we run the following out-of-sample test. First, we estimate the mimicking portfolio weights for the yields and housing factors using data through 2009:12.[14] Using the fitted mimicking portfolio weights, we then compute out-of-sample returns for the yields and housing factors for 2010:01 to 2019:12. Finally, we compute pricing errors for the 315 anomalies for SM3 for 2010:01 to 2019:12.

Table 7 reports summary statistics for the pricing errors for SM3 constructed using the out-of-sample factors, as well as the three leading multifactor models, for 2010:01 to 2019:12. SM3 continues to perform comparatively well. C4 and FF5 perform quite similarly in Table 7, with means (standard deviations) for the alpha magnitudes of 0.447% and 0.446% (0.442% and 0.445%), respectively. HXZQ exhibits the best overall performance, followed by SM3. The means (standard deviations) for the alpha magnitudes for HZXQ and SM3

[14]The loadings and mimicking portfolio weights for the yields and housing factors are similar for the 1963:07 to 2019:12 and 1963:07 to 2009:12 estimation samples, so that we can readily economically interpret the factors as yields and housing factors for both samples.

are 0.375% and 0.423% (0.398% and 0.443%), respectively. Based on the Harvey, Liu, and Zhu (2016) threshold of three for the absolute value of the t-statistic, HXZQ produces the fewest rejections of the zero-alpha hypothesis (19.37%), followed closely by SM3 (20.32%). The histograms in Figure 5 further indicate that SM3 performs relatively well in accounting for anomalies for the 2010:01 to 2019:12 out-of-sample period. In sum, SM3 continues to perform well compared to leading multifactor models when it comes to explaining anomalies for data not used to estimate the sparse loadings and mimicking portfolio weights for the yields and housing factors.

4.2. Links to Popular Factors

As a final exercise, we estimate regressions relating the mimicking portfolio returns for the yields and housing factors to the factors in C4, FF5, and HXZQ. This exercise has two primary purposes. First, we can investigate links between our new yields and housing factors and popular characteristic-based factors from the literature. Second, Barillas and Shanken (2017) argue that it is more informative to compare asset pricing models by analyzing the ability of the factors in competing models to price each other.

Table 8 reports estimation results for regressions of the mimicking portfolio returns for the yields and housing factors on the factors in C4, FF5, and HXZQ.[15] In general, the size and value factors are significantly negatively (positively) related to the yields (housing) factors. Additionally, momentum, profitability, investment, and return-on-equity factors are significantly positive in regressions for the housing factor. The intercept terms (i.e., alphas) are significant at the 1% level in all of the regressions, and the Gibbons, Ross, and Shanken (1989) statistic rejects the hypothesis that the alphas are jointly zero at the 1% level for the three sets of regressions in Table 8. Hence, C4, FF5, and HXZQ cannot price SM3.

Table 9 reports results for regressing the C4, FF5, and HXZQ factors on the yields and housing factors. The yields factor is significantly negative in regressions for size and

[15]The market excess return appears as a factor in all of the models, so that it is excluded from the regressions in Tables 8 and 9.

value factors, while it is significantly positive in regressions for momentum and profitability factors. Housing is significantly positive in regressions for size, momentum, and return-on-equity factors. With respect to C4, the alphas are individually and jointly insignificant, so that SM3 can price C4. Taking the results in Tables 8 and 9 together, SM3 encompasses C4, while the converse does not hold, so that SM3 is selected over C4 based on Barillas and Shanken (2017).

The alphas are also insignificant in the size regressions for FF5 and HXZQ in Table 9, as well as the value regression for the latter. They are significant in the profitability and investment regressions for FF5, as well as the investment and return-on-equity regressions for HXZQ. According to the Gibbons, Ross, and Shanken (1989) statistics, the alphas are jointly significant for FF5 and HZXQ in Table 9. Together with the results in Table 8, SM3 and FF5 do not encompass each other, so that neither model is selected over the other based on Barillas and Shanken (2017). A similar situation applies to SM3 and HXZQ.

Overall, Tables 8 and 9 indicate the following. The yields factor is significantly related to size, value, and profitability factors from the literature, while the housing factor is significantly related to size, momentum, and return-on-equity factors. Well-known factors thus appear to reflect macroeconomic risks related to yields and housing. Furthermore, according to the Barillas and Shanken (2017) approach, SM3 dominates C4. Neither FF5 nor HXZQ can price SM3, and the latter cannot price the first two models. The Barillas and Shanken (2017) approach thus does not clearly favor either SM3 or FF5; a similar situation holds for SM3 and HXZQ.

5. Conclusion

We extract macro risk factors from a comprehensive set of 120 monthly macroeconomic variables from the FRED-MD database. We apply, for the first time in finance, sparse PCA to obtain macro factors as sparse linear combinations of the underlying macroeconomic variables. Sparse PCA has two advantages relative to conventional PCA. First, it allows for natural economic interpretation of the factors. Second, it reduces the noise in irrelevant variables. Using the robust three-pass methodology of Giglio and Xiu (forthcoming) and 345 portfolios formed from a variety of firm characteristics as test assets, we find two major macroeconomic risk factors—corresponding to yields and housing—in the cross section of US equity returns. The yields and housing factors earn significant risk premia based on the three-pass methodology, and mimicking portfolios for the yields and housing factors deliver large annualized Sharpe ratios (in magnitude). By applying sparse PCA to macroeconomic data, we are better able to uncover risk premia in cross-sectional equity returns, even though the estimation of the sparse PCs does not directly incorporate data from cross-sectional returns.

A sparse macro three-factor model comprised of the market factor and mimicking portfolio returns for the yields and housing factors generally performs as well as or better than leading multifactor models from the literature in accounting for numerous anomalies in cross-sectional equity returns. In light of its parsimony, macroeconomic foundations, and empirical performance, our sparse macro three-factor model constitutes an informative benchmark model for asset pricing. Regressions of popular factors on the yields and housing factors reveal that size, value, momentum, profitability, and return-on-equity factors are linked to macroeconomic risks related to yields and housing.

Intuitively, our use of sparse PCA seeks a balance in utilizing macroeconomic data for cross-sectional asset pricing. At one end of the spectrum, epitomized by Chen, Roll, and Ross (1986), researchers select a single variable to represent a macroeconomic concept. Although this approach permits a sharp interpretation of the risk factor, it potentially ignores relevant information, as multiple variables can reflect fundamental macroeconomic concepts. At the other end of the spectrum, we can use conventional PCA to incorporate the information in a large number of variables in a reasonably parsimonious manner. However, conventional PCs are often difficult to interpret and cleanly identify with a macroeconomic concept. Our use of sparse PCA constitutes a middle ground between these two extremes: by constructing

PCs, we are able incorporate information from a larger number of economic variables that potentially relate to macroeconomic concepts; at the same time, by imposing a high degree of sparsity on the weight vectors, we are better able to economically interpret the PCs. In addition to sharpening the interpretation of PCs extracted from a large number of macroeconomic variables, we find that the sparse PCA approach is better able to identify relevant signals in macroeconomic data for cross-sectional asset pricing.

With respect to future research, our sparse macro factor approach can be applied to mutual funds and hedge funds to understand their macroeconomic risk exposures and evaluate their performance. The impact of other types of economic variables, such as investor sentiment and news, can also be analyzed using sparse PCA. Indeed, the information in any type of "big data" in finance can be conveniently summarized by sparse PCA. Our results suggest that sparse PCA is a promising strategy for identifying relevant information for investors in large data sets.

References

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

Bai, J. and S. Ng (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica* 70:1, 191–221.

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

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

Chen, A. Y. and T. Zimmermann (2020). Open Source Cross-Sectional Asset Pricing. Working Paper.

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

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.

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.

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 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 (forthcoming). Asset Pricing with Omitted Factors. *Journal of Political Economy*.

Green, J., J. R. M. Hand, and X. F. Zhang (2017). The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns. *Review of Financial Studies* 30:12, 4389–4436.

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

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Second Edition. New York: Springer.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. London: CRC Press.

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

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

Hubert, M., T. Reynkens, E. Schmitt, and T. Verdonck (2016). Sparse PCA for High-Dimensional Data With Outliers. *Technometrics* 58:4, 424–434.

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

Jegadeesh, N. and S. Titman (1993). Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency. *Journal of Finance* 48:1, 65–91.

Jolliffe, I. T. (1995). Rotation of Principal Components: Choice of Normalization Constraints. *Journal of Applied Statistics* 22:1, 29–35.

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.

Kelly, B. T., S. Pruitt, and Y. Su (2020). Instrumented Principal Component Analysis. Working Paper.

Kim, S., R. A. Korajczyk, and A. Neuhierl (forthcoming). Arbitrage Portfolios. *Review of Financial Studies*.

Leamer, E. E. (2007). Housing *is* the Business Cycle. In *Housing, Housing Finance, and Monetary Policy*. Federal Reserve Bank of Kansas City, pp. 149–233.

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.

Lettau, M. and M. Pelger (2020a). Estimating Latent Asset-Pricing Factors. *Journal of Econometrics* 218:1, 1–31.

Lettau, M. and M. Pelger (2020b). Factors That Fit the Time Series and Cross-Section of Stock Returns. *Review of Financial Studies* 33:5, 2274–2325.

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.

McLean, R. D. and J. Pontiff (2016). Does Academic Research Destroy Return Predictability? *Journal of Finance* 71:1, 5–32.

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

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.

Pearson, K. (1901). LIII. On Lines and Planes of Closest Fit to Systems of Points in Space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 6:2, 559–572.

Pelger, M. and R. Xiong (forthcoming). State-Varying Factor Models of Large Dimensions. *Journal of Business and Economic Statistics*.

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

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. and R. Tibshirani (2020). `PMA`. `R` Package Version 1.2.1.

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.

Table 1: Macroeconomic variable descriptions

(1) FRED ticker	(2) Transformation	(3) Description
RPI	Log growth	Real Personal Income
W875RX1	Log growth	Real Personal Income Excluding Transfer Receipts
DPCERA3M086SBEA	Log growth	Real Personal Consumption Expenditures
CMRMTSPLx	Log growth	Real Manufacturing and Trade Industries Sales
RETAILx	Log growth	Retail and Food Services Sales
INDPRO	Log growth	Industrial Production Index
IPFPNSS	Log growth	Industrial Production Index: Final Products and Nonindustrial Supplies
IPFINAL	Log growth	Industrial Production Index: Final Products (Market Group)
IPCONGD	Log growth	Industrial Production Index: Consumer Goods
IPDCONGD	Log growth	Industrial Production Index: Durable Consumer Goods
IPNCONGD	Log growth	Industrial Production Index: Nondurable Consumer Goods
IPBUSEQ	Log growth	Industrial Production Index: Business Equipment
IPMAT	Log growth	Industrial Production Index: Materials
IPDMAT	Log growth	Industrial Production Index: Durable Materials
IPNMAT	Log growth	Industrial Production Index: Nondurable Materials
IPMANSICS	Log growth	Industrial Production Index: Manufacturing (SIC)
IPB51222S	Log growth	Industrial Production Index: Residential Utilities
IPFUELS	Log growth	Industrial Production Index: Fuels
CUMFNS	Difference	Capacity Utilization: Manufacturing
HWI	Difference	Help-Wanted Index for United States
HWIURATIO	Difference	Ratio of Help Wanted to Number Unemployed
CLF16OV	Log growth	Civilian Labor Force
CE16OV	Log growth	Civilian Employment
UNRATE	Difference	Civilian Unemployment Rate
UEMPMEAN	Difference	Average Duration of Unemployment (Weeks)
UEMPLT5	Log growth	Civilians Unemployed—Less Than 5 Weeks
UEMP5TO14	Log growth	Civilians Unemployed for 5–14 Weeks
UEMP15OV	Log growth	Civilians Unemployed—15 Weeks and Over
UEMP15T26	Log growth	Civilians Unemployed for 15–26 Weeks
UEMP27OV	Log growth	Civilians Unemployed for 27 Weeks and Over

The table provides descriptions of 120 macroeconomic variables from the FRED-MD database. The first column identifies the variable based on its FRED ticker. (DIVYLD is the S&P 500 dividend yield based on data from Robert Shiller's webpage.) The second column reports how the variable is transformed before computing the principal components; − indicates that the variable is not transformed. The description in the third column is based on the FRED-MD Updated Appendix.

Table 1 (continued)

(1) FRED ticker	(2) Transformation	(3) Description
CLAIMSx	Log growth	Initial Claims
PAYEMS	Log growth	All Employees: Total Nonfarm
USGOOD	Log growth	All Employees: Goods-Producing Industries
CES1021000001	Log growth	All Employees: Mining and Logging: Mining
USCONS	Log growth	All Employees: Construction
MANEMP	Log growth	All Employees: Manufacturing
DMANEMP	Log growth	All Employees: Durable Goods
NDMANEMP	Log growth	All Employees: Nondurable Goods
SRVPRD	Log growth	All Employees: Service-Providing Industries
USTPU	Log growth	All Employees: Trade, Transportation, and Utilities
USWTRADE	Log growth	All Employees: Wholesale Trade
USTRADE	Log growth	All Employees: Retail Trade
USFIRE	Log growth	All Employees: Financial Activities
USGOVT	Log growth	All Employees: Government
CES0600000007	Log growth	Average Weekly Hours: Goods Producing
AWOTMAN	Log growth	Average Weekly Overtime Hours: Manufacturing
AWHMAN	Log growth	Average Weekly Hours: Manufacturing
HOUST	Log	Housing Starts: Total New Privately Owned
HOUSTNE	Log	Housing Starts: Total New Privately Owned, Northeast
HOUSTMW	Log	Housing Starts: Total New Privately Owned, Midwest
HOUSTS	Log	Housing Starts: Total New Privately Owned, South
HOUSTW	Log	Housing Starts: Total New Privately Owned, West
PERMIT	Log	New Private Housing Permits
PERMITNE	Log	New Private Housing Permits, Northeast
PERMITMW	Log	New Private Housing Permits, Midwest
PERMITS	Log	New Private Housing Permits, South
PERMITW	Log	New Private Housing Permits, West
AMDMNOx	Log growth	New Orders for Durable Goods
AMDMUOx	Log growth	Unfilled Orders for Durable Goods
BUSINVx	Log growth	Total Business Inventories

Table 1 (continued)

(1) FRED ticker	(2) Transformation	(3) Description
ISRATIOx	Difference	Total Business: Inventories to Sales Ratio
M1SL	Log growth	M1 Money Stock
M2SL	Log growth	M2 Money Stock
M2REAL	Log growth	Real M2 Money Stock
AMBSL	Log growth	St. Louis Adjusted Monetary Base
TOTRESNS	Log growth	Total Reserves of Depository Institutions
NONBORRES	Log growth	Reserves Of Depository Institutions
BUSLOANS	Log growth	Commercial and Industrial Loans
REALLN	Log growth	Real Estate Loans at All Commercial Banks
NONREVSL	Log growth	Total Nonrevolving Credit
CONSPI	Difference	Ratio of Nonrevolving Consumer Credit to Personal Income
DIVYLD	–	S&P 500 Dividend Yield
FEDFUNDS	–	Effective Federal Funds Rate
CP3Mx	–	3-Month AA Financial Commercial Paper Rate
TB3MS	–	3-Month Treasury Bill Rate
TB6MS	–	6-Month Treasury Bill Rate
GS1	–	1-Year Treasury Rate
GS5	–	5-Year Treasury Rate
GS10	–	10-Year Treasury Rate
AAA	–	Moody's Seasoned Aaa Corporate Bond Yield
BAA	–	Moody's Seasoned Baa Corporate Bond Yield
COMPAPFFx	–	CP3Mx Minus FEDFUNDS
TB3SMFFM	–	TB3MS Minus FEDFUNDS
TB6SMFFM	–	TB6MS Minus FEDFUNDS
T1YFFM	–	GS1 Minus FEDFUNDS
T5YFFM	–	GS5 Minus FEDFUNDS
T10YFFM	–	GS10 Minus FEDFUNDS
AAAFFM	–	AAA Minus FEDFUNDS
BAAFFM	–	BAA Minus FEDFUNDS
EXSZUSx	Log growth	Switzerland/US Foreign Exchange Rate

Table 1 (continued)

(1) FRED ticker	(2) Transformation	(3) Description
EXJPUSx	Log growth	Japan/US Foreign Exchange Rate
EXUSUKx	Log growth	US/UK Foreign Exchange Rate
EXCAUSx	Log growth	Canada/US Foreign Exchange Rate
WPSFD49207	Log growth	Producer Price Index: Finished Goods
WPSFD49502	Log growth	Producer Price Index: Finished Consumer Goods
WPSID61	Log growth	Producer Price Index: Intermediate Materials
WPSID62	Log growth	Producer Price Index: Crude Materials
OILPRICEx	Log growth	Crude Oil, Spliced WTI and Cushing
PPICMM	Log growth	Producer Price Index: Metals and metal products
CPIAUCSL	Log growth	Consumer Price Index: All Items
CPIAPPSL	Log growth	Consumer Price Index: Apparel
CPITRNSL	Log growth	Consumer Price Index: Transportation
CPIMEDSL	Log growth	Consumer Price Index: Medical Care
CUSR0000SAC	Log growth	Consumer Price Index: Commodities
CUSR0000SAD	Log growth	Consumer Price Index: Durables
CUSR0000SAS	Log growth	Consumer Price Index: Services
CPIULFSL	Log growth	Consumer Price Index: All Items Less Food
CUSR0000SA0L2	Log growth	Consumer Price Index: All items Less Shelter
CUSR0000SA0L5	Log growth	Consumer Price Index: All Items Less Medical Care
PCEPI	Log growth	Personal Consumption Expenditures Deflator
DDURRG3M086SBEA	Log growth	Personal Consumption Expenditures Deflator: Durable Goods
DNDGRG3M086SBEA	Log growth	Personal Consumption Expenditures Deflator: Nondurable Goods
DSERRG3M086SBEA	Log growth	Personal Consumption Expenditures Deflator: Services
CES0600000008	Log growth	Average Hourly Earnings: Goods Producing
CES2000000008	Log growth	Average Hourly Earnings: Construction
CES3000000008	Log growth	Average Hourly Earnings: Manufacturing
MZMSL	Log growth	MZM Money Stock
DTCOLNVHFNM	Log growth	Consumer Motor Vehicle Loans Outstanding
DTCTHFNM	Log growth	Total Consumer Loans and Leases Outstanding
INVEST	Log growth	Securities in Bank Credit at All Commercial Banks

Table 2: Conventional principal component weights

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
RPI	0.01	0.08	−0.07	0.05	−0.05	−0.01	−0.01	0.01	−0.21
W875RX1	0.02	0.10	−0.07	0.05	−0.06	−0.03	0.01	−0.01	−0.19
DPCERA3M086SBEA	0.01	0.07	−0.04	0.04	0.03	−0.002	−0.06	0.16	−0.04
CMRMTSPLx	0.02	0.07	0.01	−0.02	0.04	−0.06	0.06	−0.16	−0.01
RETAILx	0.05	0.05	0.05	0.02	0.06	−0.03	−0.02	0.17	0.004
INDPRO	0.06	0.19	0.06	0.17	0.003	0.12	−0.002	−0.01	0.003
IPFPNSS	0.06	0.18	0.05	0.18	0.01	0.14	−0.02	0.02	−0.03
IPFINAL	0.06	0.16	0.05	0.19	0.0002	0.16	−0.04	0.01	−0.04
IPCONGD	0.04	0.14	0.04	0.21	0.04	0.17	−0.08	0.02	−0.03
IPDCONGD	0.03	0.13	0.06	0.19	0.04	0.12	0.001	−0.07	−0.01
IPNCONGD	0.03	0.08	0.01	0.13	0.02	0.15	−0.12	0.11	−0.03
IPBUSEQ	0.07	0.15	0.05	0.10	−0.05	0.10	0.02	0.01	−0.04
IPMAT	0.04	0.17	0.06	0.14	−0.001	0.09	0.02	−0.04	0.02
IPDMAT	0.05	0.17	0.08	0.12	0.02	0.08	0.07	−0.08	−0.01
IPNMAT	0.04	0.12	0.04	0.09	0.03	0.05	−0.02	0.04	0.08
IPMANSICS	0.06	0.19	0.07	0.16	0.02	0.10	0.02	−0.01	0.01
IPB51222S	−0.004	0.005	−0.01	0.05	−0.01	0.13	−0.10	0.02	−0.04
IPFUELS	0.020	0.02	−0.004	0.04	0.01	0.04	−0.05	0.04	0.05
CUMFNS	0.04	0.19	0.08	0.18	0.04	0.09	0.06	−0.02	0.02
HWI	0.01	0.07	0.02	−0.02	−0.004	−0.07	−0.02	−0.04	0.12
HWIURATIO	0.01	0.10	0.02	−0.03	−0.02	−0.07	−0.003	−0.03	0.09
CLF16OV	0.05	0.02	−0.02	−0.03	0.01	−0.10	−0.05	0.15	0.08
CE16OV	0.06	0.10	−0.001	−0.03	−0.01	−0.16	0.03	0.07	0.11
UNRATE	−0.03	−0.13	−0.03	0.004	0.03	0.12	−0.13	0.10	−0.06
UEMPMEAN	−0.02	−0.03	0.03	0.10	0.08	0.05	−0.01	−0.04	−0.11
UEMPLT5	0.003	−0.03	−0.02	−0.03	−0.02	0.01	−0.05	0.10	−0.07
UEMP5TO14	−0.01	−0.07	−0.03	−0.01	−0.004	0.05	−0.06	0.03	0.05
UEMP15OV	−0.02	−0.11	−0.01	0.07	0.09	0.14	−0.14	0.03	−0.10
UEMP15T26	−0.01	−0.08	−0.02	0.01	0.03	0.08	−0.11	0.02	−0.11
UEMP27OV	−0.02	−0.08	0.01	0.09	0.10	0.11	−0.08	0.01	−0.04

The table reports weights for the first nine principal components extracted from 120 macroeconomic variables (listed in Table 1) from the FRED-MD database. The sample period is 1960:02 to 2019:12. The variable name in the first column corresponds to its FRED ticker.

Table 2 (continued)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
CLAIMSx	−0.01	−0.10	−0.02	−0.09	−0.05	0.03	−0.05	0.04	−0.03
PAYEMS	0.10	0.18	0.01	−0.03	−0.08	−0.16	−0.04	0.06	−0.02
USGOOD	0.08	0.19	0.04	−0.01	−0.09	−0.11	0.03	−0.02	0.03
CES1021000001	0.03	0.01	0.05	−0.01	−0.06	−0.02	−0.02	−0.001	−0.02
USCONS	0.05	0.13	−0.01	−0.06	−0.01	−0.10	0.10	0.10	0.07
MANEMP	0.07	0.18	0.06	0.03	−0.10	−0.10	−0.01	−0.07	0.004
DMANEMP	0.07	0.18	0.05	0.04	−0.10	−0.07	−0.01	−0.09	−0.01
NDMANEMP	0.06	0.14	0.06	0.02	−0.08	−0.16	−0.04	0.03	0.05
SRVPRD	0.10	0.13	−0.02	−0.05	−0.06	−0.17	−0.13	0.12	−0.07
USTPU	0.10	0.13	0.01	−0.05	−0.06	−0.21	−0.03	0.09	−0.03
USWTRADE	0.11	0.12	0.01	−0.07	−0.09	−0.15	−0.05	0.04	−0.02
USTRADE	0.09	0.11	−0.01	−0.02	−0.01	−0.21	−0.05	0.08	−0.05
USFIRE	0.11	0.08	−0.08	−0.05	−0.003	−0.09	−0.10	0.01	−0.01
USGOVT	0.03	0.02	−0.05	−0.06	−0.02	−0.01	−0.28	0.10	−0.10
CES0600000007	−0.05	0.12	0.08	−0.10	−0.21	−0.01	0.14	0.04	0.06
AWOTMAN	0.02	0.07	0.05	0.09	0.05	0.07	0.02	−0.06	−0.01
AWHMAN	−0.06	0.12	0.09	−0.12	−0.20	0.01	0.14	0.03	0.05
HOUST	0.12	0.10	−0.17	−0.15	0.13	0.10	−0.01	−0.01	0.04
HOUSTNE	0.11	0.07	−0.14	−0.11	0.09	0.03	−0.19	0.01	−0.05
HOUSTMW	0.11	0.08	−0.13	−0.12	0.12	0.08	−0.15	0.01	−0.02
HOUSTS	0.11	0.09	−0.16	−0.14	0.11	0.10	0.13	−0.03	0.08
HOUSTW	0.11	0.09	−0.16	−0.15	0.15	0.10	0.04	0.0004	0.05
PERMIT	0.10	0.10	−0.16	−0.17	0.14	0.12	0.09	−0.02	0.08
PERMITNE	0.10	0.09	−0.14	−0.13	0.09	0.06	−0.17	0.004	−0.02
PERMITMW	0.10	0.10	−0.12	−0.16	0.14	0.11	−0.11	0.01	0.02
PERMITS	0.06	0.07	−0.12	−0.14	0.09	0.12	0.28	−0.04	0.12
PERMITW	0.11	0.09	−0.16	−0.15	0.15	0.11	0.06	−0.01	0.07
AMDMNOx	0.03	0.07	0.03	0.07	0.02	0.03	0.003	−0.02	0.002
AMDMUOx	0.09	0.07	0.02	−0.08	−0.11	0.002	−0.02	−0.10	0.01
BUSINVx	0.11	0.01	0.04	−0.05	−0.14	−0.06	−0.07	0.10	−0.09

Table 2 (continued)

(1) Variable	(2) PC1	(3) PC2	(4) PC3	(5) PC4	(6) PC5	(7) PC6	(8) PC7	(9) PC8	(10) PC9
ISRATIOx	−0.02	−0.05	−0.05	0.01	−0.11	0.02	−0.07	0.23	−0.03
M1SL	−0.03	−0.02	−0.02	0.07	0.05	−0.18	0.004	−0.24	0.08
M2SL	0.01	−0.003	−0.11	0.06	0.10	−0.14	−0.10	−0.31	0.11
M2REAL	−0.10	0.05	−0.19	0.07	0.03	−0.09	−0.06	−0.22	0.07
AMBSL	−0.05	−0.02	−0.09	0.08	−0.04	−0.07	−0.04	−0.13	−0.05
TOTRESNS	−0.06	−0.03	−0.09	0.05	−0.04	−0.08	−0.07	−0.17	−0.01
NONBORRES	0.001	0.0002	0.002	−0.01	0.005	0.01	0.02	−0.02	0.01
BUSLOANS	0.09	0.03	−0.04	−0.08	−0.17	−0.05	−0.05	−0.09	0.03
REALLN	0.10	0.05	−0.15	−0.08	0.04	0.03	−0.06	−0.05	0.03
NONREVSL	0.04	0.06	−0.05	−0.10	−0.04	−0.05	0.11	−0.16	−0.40
CONSPI	−0.02	0.03	−0.03	−0.09	−0.03	0.001	0.13	−0.13	−0.42
DIVYLD	0.03	0.005	0.02	−0.02	−0.07	−0.05	−0.05	−0.09	0.03
FEDFUNDS	0.18	−0.09	−0.10	0.09	−0.06	−0.01	0.07	0.04	0.01
CP3Mx	0.18	−0.09	−0.10	0.09	−0.04	−0.03	0.04	0.05	0.002
TB3MS	0.18	−0.08	−0.11	0.10	−0.03	−0.03	0.06	0.05	0.004
TB6MS	0.18	−0.08	−0.11	0.10	−0.02	−0.04	0.06	0.06	−0.001
GS1	0.18	−0.08	−0.11	0.11	−0.01	−0.05	0.07	0.07	−0.01
GS5	0.16	−0.08	−0.11	0.13	0.05	−0.08	0.10	0.10	−0.02
GS10	0.15	−0.08	−0.11	0.15	0.07	−0.10	0.13	0.10	−0.03
AAA	0.14	−0.09	−0.11	0.16	0.08	−0.10	0.18	0.09	−0.02
BAA	0.12	−0.10	−0.11	0.17	0.08	−0.10	0.18	0.08	−0.01
COMPAPFFx	−0.10	0.06	0.02	−0.03	0.13	−0.11	−0.26	0.03	−0.06
TB3SMFFM	−0.13	0.11	0.04	−0.04	0.17	−0.10	−0.08	0.03	−0.02
TB6SMFFM	−0.13	0.12	0.03	−0.03	0.19	−0.12	−0.10	0.05	−0.04
T1YFFM	−0.09	0.10	−0.01	0.03	0.24	−0.19	−0.05	0.10	−0.07
T5YFFM	−0.12	0.07	0.01	0.06	0.27	−0.16	0.04	0.11	−0.08
T10YFFM	−0.13	0.06	0.03	0.05	0.25	−0.15	0.07	0.09	−0.07
AAAFFM	−0.15	0.05	0.04	0.04	0.22	−0.11	0.11	0.05	−0.04
BAAFFM	−0.15	0.02	0.03	0.07	0.22	−0.12	0.13	0.04	−0.03
EXSZUSx	0.001	0.002	0.004	0.05	−0.03	−0.05	0.03	−0.12	0.11

Table 2 (continued)

(1) Variable	(2) PC1	(3) PC2	(4) PC3	(5) PC4	(6) PC5	(7) PC6	(8) PC7	(9) PC8	(10) PC9
EXJPUSx	0.01	−0.01	0.03	0.01	−0.04	−0.01	0.03	−0.10	0.07
EXUSUKx	0.01	0.01	0.03	−0.06	0.04	0.10	0.01	0.13	−0.13
EXCAUSx	−0.001	−0.003	−0.02	−0.002	−0.03	−0.14	−0.05	−0.09	0.02
WPSFD49207	0.11	−0.05	0.20	−0.10	0.08	0.003	0.005	−0.07	−0.02
WPSFD49502	0.10	−0.04	0.21	−0.12	0.09	0.01	0.02	−0.05	−0.01
WPSID61	0.12	−0.03	0.20	−0.11	0.06	0.01	0.03	−0.05	−0.02
WPSID62	0.06	0.01	0.17	−0.08	0.09	0.04	0.05	−0.04	−0.02
OILPRICEx	0.02	0.001	0.04	−0.03	0.002	0.08	0.04	0.02	0.01
PPICMM	0.05	0.04	0.09	−0.07	0.05	0.03	0.07	−0.001	−0.04
CPIAUCSL	0.16	−0.08	0.16	−0.04	0.06	−0.01	−0.01	−0.02	0.01
CPIAPPSL	0.08	−0.03	0.05	0.05	−0.01	−0.06	−0.10	−0.04	0.01
CPITRNSL	0.10	−0.04	0.23	−0.08	0.13	−0.01	0.03	−0.01	−0.02
CPIMEDSL	0.10	−0.10	−0.04	0.13	0.02	−0.05	−0.03	0.05	0.01
CUSR0000SAC	0.13	−0.05	0.22	−0.08	0.12	−0.01	−0.01	−0.03	−0.01
CUSR0000SAD	0.12	−0.07	0.04	0.11	0.03	−0.11	−0.10	0.02	−0.04
CUSR0000SAS	0.14	−0.09	= 0.003	0.05	−0.04	−0.02	−0.03	= 0.003	0.0003
CPIULFSL	0.15	−0.09	0.16	−0.02	0.07	−0.02	−0.01	−0.01	0.0003
CUSR0000SA0L2	0.15	−0.07	0.20	−0.05	0.10	−0.01	= 0.005	−0.03	0.004
CUSR0000SA0L5	0.16	−0.08	0.16	−0.05	0.07	−0.02	−0.01	−0.03	0.01
PCEPI	0.17	−0.09	0.13	−0.01	0.08	−0.04	−0.03	−0.03	0.03
DDURRG3M086SBEA	0.12	−0.07	0.01	0.09	0.03	−0.11	−0.14	−0.01	−0.04
DNDGRG3M086SBEA	0.12	−0.05	0.22	−0.10	0.11	0.01	= 0.001	−0.04	0.004
DSERRG3M086SBEA	0.15	−0.09	−0.01	0.07	0.04	−0.05	−0.02	−0.01	0.05
CES0600000008	0.10	−0.03	−0.02	0.07	0.01	0.002	−0.18	−0.24	−0.06
CES2000000008	0.04	−0.05	−0.02	−0.004	−0.003	0.01	−0.17	−0.20	−0.08
CES3000000008	0.10	−0.01	−0.005	0.11	0.02	0.02	−0.13	−0.22	−0.03
MZMSL	−0.03	−0.01	−0.09	0.07	0.14	−0.08	0.06	−0.23	0.13
DTCOLNVHFNM	0.04	0.001	−0.04	−0.02	0.02	−0.01	0.17	−0.11	−0.29
DTCTHFNM	0.05	0.01	−0.06	−0.06	−0.01	−0.002	0.16	−0.09	−0.42
INVEST	−0.01	−0.01	−0.02	0.07	0.14	0.004	−0.01	−0.05	0.09

Table 3: Sparse principal component weights

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	SPC1	SPC2	SPC3	SPC4	SPC5	SPC6	SPC7	SPC8	SPC9
Variable	Yields	Production	Inflation	Housing	Spreads	Employment	Costs	Money	Credit
RPI	–	–	–	–	–	–	–	–	0.04
W875RX1	–	–	–	–	–	–	–	–	0.09
DPCERA3M086SBEA	–	–	–	–	–	–	–	–	–
CMRMTSPLx	–	–	–	–	–	–	–	–	–
RETAILx	–	–	–	–	–	–	–	–	–
INDPRO	–	0.41	–	–	–	–	–	–	–
IPFPNSS	–	0.40	–	–	–	–	–	–	–
IPFINAL	–	0.36	–	–	–	–	–	–	–
IPCONGD	–	0.23	–	–	–	–	–	–	–
IPDCONGD	–	0.18	–	–	–	–	–	–	–
IPNCONGD	–	–	–	–	–	–	–	–	–
IPBUSEQ	–	0.15	–	–	–	–	–	–	–
IPMAT	–	0.20	–	–	–	–	–	–	–
IPDMAT	–	0.23	–	–	–	–	–	–	–
IPNMAT	–	–	–	–	–	–	–	–	–
IPMANSICS	–	0.42	–	–	–	–	–	–	–
IPB51222S	–	–	–	–	–	–	–	–	–
IPFUELS	–	–	–	–	–	–	–	–	–
CUMFNS	–	0.40	–	–	–	–	–	–	–
HWI	–	–	–	–	–	–	–	–	–
HWIURATIO	–	–	–	–	–	–	–	–	–
CLF16OV	–	–	–	–	–	–	–	–	–
CE16OV	–	–	–	–	–	–	–	–	0.07
UNRATE	–	–	–	–	–	–	–	–	−0.15
UEMPMEAN	–	–	–	–	–	−0.001	–	–	–
UEMPLT5	–	–	–	–	–	–	–	–	–
UEMP5TO14	–	–	–	–	–	–	–	–	–
UEMP15OV	–	–	–	–	–	–	–	–	−0.18
UEMP15T26	–	–	–	–	–	–	–	–	–
UEMP27OV	–	–	–	–	–	–	–	–	−0.19

The table reports nonzero weights for the first nine sparse principal components extracted from 120 macroeconomic variables (listed in Table 1) from the FRED-MD database. The sample period is 1960:02 to 2019:12. The variable name in the first column corresponds to its FRED ticker. The column headings provide descriptions of the sparse principal components based on the active elements of their weight vectors. The components are computed using the Witten, Tibshirani, and Hastie (2009) penalized matrix decomposition methodology.

Table 3 (continued)

(1)	(2)	(4)	(3)	(5)	(7)	(6)	(8)	(9)	(10)
	SPC1	SPC2	SPC3	SPC4	SPC5	SPC6	SPC7	SPC8	SPC9
Variable	Yields	Production	Inflation	Housing	Spreads	Employment	Costs	Money	Credit
CLAIMSx	–	–	–	–	–	-	–	–	–
PAYEMS	–	–	–	–	–	0.45	–	–	–
USGOOD	–	–	–	–	–	0.41	–	–	–
CES1021000001	–	–	–	–	–	-	–	–	–
USCONS	–	–	–	–	–	0.05	–	–	0.02
MANEMP	–	0.01	–	–	–	0.37	–	–	–
DMANEMP	–	0.003	–	–	–	0.34	–	–	–
NDMANEMP	–	–	–	–	–	0.25	–	–	–
SRVPRD	–	–	–	–	–	0.30	–	–	–
USTPU	–	–	–	–	–	0.33	–	–	–
USWTRADE	–	–	–	–	–	0.27	–	–	–
USTRADE	–	–	–	–	–	0.21	–	–	–
USFIRE	–	–	–	–	–	0.03	–	–	0.10
USGOVT	–	–	–	–	–	–	–	–	–
CES0600000007	–	–	–	–	–	–	−0.36	–	–
AWOTMAN	–	–	–	–	–	–	–	–	–
AWHMAN	–	–	–	–	–	–	−0.39	–	–
HOUST	–	–	–	0.43	–	–	–	–	–
HOUSTNE	–	–	–	0.15	–	–	–	–	–
HOUSTMW	–	–	–	0.27	–	–	–	–	–
HOUSTS	–	–	–	0.33	–	–	–	–	–
HOUSTW	–	–	–	0.38	–	–	–	–	–
PERMIT	–	–	–	0.41	–	–	–	–	–
PERMITNE	–	–	–	0.21	–	–	–	–	–
PERMITMW	–	–	–	0.32	–	–	–	–	–
PERMITS	–	–	–	0.12	–	–	–	–	–
PERMITW	–	–	–	0.38	–	–	–	–	–
AMDMNOx	–	–	–	–	–	–	–	–	–
AMDMUOx	–	–	–	–	–	–	–	–	0.13
BUSINVx	–	–	–	–	−0.01	−0.01	–	−0.12	0.04

Table 3 (continued)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	SPC1	SPC2	SPC3	SPC4	SPC5	SPC6	SPC7	SPC8	SPC9
Variable	Yields	Production	Inflation	Housing	Spreads	Employment	Costs	Money	Credit
ISRATIOx	–	–	–	–	–	–	–	–	–
M1SL	–	–	–	–	–	–	–	0.26	–
M2SL	–	–	–	–	–	–	–	0.44	–
M2REAL	–	–	–	–	–	–	–	0.58	–
AMBSL	–	–	–	–	–	–	–	0.13	–
TOTRESNS	–	–	–	–	–	–	–	0.15	–
NONBORRES	–	–	–	–	–	–	–	–	–
BUSLOANS	–	–	–	–	−0.002	–	–	–	0.13
REALLN	–	–	–	0.01	–	–	–	–	0.07
NONREVSL	–	–	–	–	–	–	–	–	0.58
CONSPI	–	–	–	–	–	–	–	–	0.45
DIVYLD	–	–	–	–	–	–	–	–	–
FEDFUNDS	0.33	–	–	–	–	–	–	–	–
CP3Mx	0.34	–	–	–	–	–	–	–	–
TB3MS	0.34	–	–	–	–	–	–	–	–
TB6MS	0.35	–	–	–	–	–	–	–	–
GS1	0.35	–	–	–	–	–	–	–	–
GS5	0.35	–	–	–	–	–	–	–	–
GS10	0.34	–	–	–	–	–	–	–	–
AAA	0.30	–	–	–	–	–	–	–	–
BAA	0.28	–	–	–	–	–	–	–	–
COMPAPFFx	–	–	–	–	0.21	–	–	–	–
TB3SMFFM	–	–	–	–	0.35	–	–	–	–
TB6SMFFM	–	–	–	–	0.37	–	–	–	–
T1YFFM	–	–	–	–	0.35	–	–	–	–
T5YFFM	–	–	–	–	0.39	–	–	–	–
T10YFFM	–	–	–	–	0.39	–	–	–	–
AAAFFM	–	–	–	–	0.36	–	–	–	–
BAAFFM	–	–	–	–	0.32	–	–	–	–
EXSZUSx	–	–	–	–	–	–	–	–	–

Table 3 (continued)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	SPC1	SPC2	SPC3	SPC4	SPC5	SPC6	SPC7	SPC8	SPC9
Variable	Yields	Production	Inflation	Housing	Spreads	Employment	Costs	Money	Credit
EXJPUSx	–	–	–	–	–	–	–	–	–
EXUSUKx	–	–	–	–	–	–	–	–	–
EXCAUSx	–	–	–	–	–	–	–	–	–
WPSFD49207	–	–	0.11	–	–	–	–	−0.16	–
WPSFD49502	–	–	0.08	–	–	–	–	−0.21	–
WPSID61	–	–	0.06	–	–	–	–	−0.27	–
WPSID62	–	–	–	–	–	–	–	−0.20	–
OILPRICEx	–	–	–	–	–	–	–	–	–
PPICMM	–	–	–	–	–	–	–	−0.09	–
CPIAUCSL	–	–	0.38	–	–	–	–	–	–
CPIAPPSL	–	–	–	–	–	–	0.05	–	–
CPITRNSL	–	–	0.21	–	–	–	–	–	–
CPIMEDSL	–	–	–	–	−0.01	–	0.34	–	–
CUSR0000SAC	–	–	0.39	–	–	–	–	–	–
CUSR0000SAD	–	–	–	–	–	–	0.39	–	–
CUSR0000SAS	–	–	–	–	−0.16	–	0.18	–	–
CPIULFSL	–	–	0.32	–	–	–	–	–	–
CUSR0000SA0L2	–	–	0.41	–	–	–	–	–	–
CUSR0000SA0L5	–	–	0.38	–	–	–	–	–	–
PCEPI	–	–	0.30	–	–	–	–	–	–
DDURRG3M086SBEA	–	–	–	–	–	–	0.40	–	–
DNDGRG3M086SBEA	–	–	0.35	–	–	–	–	–	–
DSERRG3M086SBEA	0.01	–	–	–	−0.08	–	0.32	–	–
CES0600000008	–	–	–	–	–	–	0.29	–	–
CES2000000008	–	–	–	–	–	–	0.02	–	–
CES3000000008	–	–	–	–	–	–	0.27	–	–
MZMSL	–	–	–	–	–	–	–	0.39	–
DTCOLNVHFNM	–	–	–	–	–	–	–	–	0.28
DTCTHFNM	–	–	–	–	–	–	–	–	0.47
INVEST	–	–	–	–	–	–	–	–	–

Table 4: Innovation correlations

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Innovations to conventional PCs (conventional macro factors)								
PC	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
PC2	0.38							
PC3	0.80	0.08						
PC4	0.11	0.68	−0.08					
PC5	0.23	−0.26	0.51	−0.24				
PC6	−0.03	0.07	0.13	0.50	−0.01			
PC7	−0.02	0.09	0.02	−0.17	−0.16	−0.09		
PC8	−0.01	0.17	0.05	−0.07	−0.13	0.01	−0.08	
PC9	0.01	−0.01	0.11	0.05	0.08	−0.08	0.07	0.03
Panel B: Innovations to sparse PCs (sparse macro factors)								
Description	Yields	Production	Inflation	Housing	Spreads	Employment	Costs	Money
Production	0.12							
Inflation	0.17	0.01						
Housing	0.14	0.17	0.06					
Spreads	−0.11	−0.08	−0.03	0.17				
Employment	0.18	0.50	0.09	0.21	−0.05			
Costs	0.17	−0.03	0.20	−0.03	−0.02	−0.04		
Money	−0.03	−0.05	−0.50	0.05	0.09	−0.09	0.01	
Credit	0.04	0.03	−0.04	0.05	−0.03	0.14	0.01	0.04

The table reports correlations for innovations to conventional and sparse principal components extracted from 120 macroeconomic variables (listed in Table 1) from the FRED-MD database. The innovations are computed by fitting a first-order vector autoregression to each set of principal components. The sample period is 1960:03 to 2019:12. The row and column headings in Panel B provide descriptions of the sparse principal components based on the active elements of their weight vectors. The sparse principal components are computed using the Witten, Tibshirani, and Hastie (2009) penalized matrix decomposition methodology

Table 5: Risk premia

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Innovations to conventional PCs (conventional macro factors)					Innovations to sparse PCs (sparse macro factors)				
Factor	$\hat{\gamma}_g$	t-stat.	R_g^2	Sharpe	Factor	$\hat{\gamma}_g$	t-stat.	R_g^2	Sharpe
PC1	−0.002	−0.05	2.75%	−0.02	Yields	−0.033	-3.70^{***}	14.68%	−0.78
PC2	0.032	0.64	3.87%	0.24	Production	0.006	0.09	3.59%	0.03
PC3	0.009	0.22	2.22%	0.10	Inflation	0.004	0.09	2.34%	0.03
PC4	−0.043	−1.35	3.79%	−0.45	Housing	0.047	3.64^{***}	4.56%	1.08
PC5	0.023	1.12	2.37%	0.41	Spreads	0.017	0.79	1.49%	0.36
PC6	0.024	0.80	5.57%	0.24	Employment	0.024	0.47	2.59%	0.21
PC7	0.021	1.24	1.46%	0.58	Cost	−0.035	−1.33	2.31%	−0.45
PC8	0.068	1.95^{*}	4.84%	0.66	Money	−0.061	−0.82	3.38%	−0.38
PC9	−0.017	−0.64	1.27%	−0.34	Credit	0.030	0.67	2.18%	0.22
R_f^2	72.39%								

The table reports three-pass regression results for innovations to conventional and sparse principal components extracted from 120 macroeconomic variables (listed in Table 1) from the FRED-MD database. The names in the sixth column provide descriptions of the sparse principal components based on the active elements of their weight vectors. The sample period is 1963:07 to 2019:12. The risk premium estimates are computed via the Giglio and Xiu (forthcoming) three-pass methodology using 345 equity portfolios as test assets. The first step estimates 15 principal components for the excess return observations for the test assets. The second step estimates risk premia for the 15 principal components from the first step via a cross-sectional regression that relates the average excess returns of the test assets to the betas for the 15 principal components; R_f^2 is the goodness-of-fit measure for the cross-sectional regression. The third step estimates time-series regressions that relate each innovation to the 15 principal components from the first step; R_g^2 is the goodness-of-fit measure for the time-series regression. $\hat{\gamma}_g$ is the estimated risk premium for the innovation, which is based on the estimated cross-sectional and time-series slope coefficients in the second and third steps, respectively; t-statistics are computed using heteroskedasticity- and autocorrelation-robust standard errors (Newey and West 1987); *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. "Sharpe" is the annualized Sharpe ratio for the factor mimicking portfolio for the innovation.

Table 6: Pricing errors

(1)	(2)	(3)	(4)	(5)	(6)	(7)
			$\lvert t\text{-statistic}\rvert \geq$			
Factor model	Mean	Std. dev.	1.645	1.96	2.58	3
Carhart (1997) 4-factor model	0.529	0.444	80.32	75.56	65.40	59.68
Fama and French (2015) 5-factor model	0.509	0.419	79.05	74.29	65.40	56.83
Hou, Xue, and Zhang (2015) q-factor model	0.513	0.424	74.60	67.94	58.41	49.52
Sparse macro 3-factor model	0.478	0.421	64.13	60.32	43.49	33.97

The table reports summary statistics for pricing errors (in percent) for 315 long-short anomaly portfolio returns from Chen and Zimmermann (2020) based on the multifactor model in the first column. For 232 of the anomalies, the sample period is 1967:01 to 2019:12 (636 observations); for 83 of the anomalies, the number of available observations is less than 636. The second (third) column reports the mean (standard deviation) of the absolute values of estimated intercept terms across 315 multiple regressions, each of which regresses a long-short anomaly portfolio return on an intercept term and the set of factors in the first column. The last four columns report the percent of t-statistics for the intercept terms that are larger (in absolute value) than 1.645, 1.96, 2.58, and 3, respectively; t-statistics are computed using heteroskedasticity- and autocorrelation-robust standard errors (Newey and West 1987). The Carhart (1997) 4-factor model includes market, size, value, and momentum factors. The Fama and French (2015) 5-factor includes market, size, value, profitability, and investment factors. The Hou, Xue, and Zhang (2015) q-factor model includes market, size, investment, and return-on-equity factors. The sparse macro 3-factor model includes market, yields, and housing factors. The yields and housing factors are constructed by first applying sparse principal component analysis to 120 macroeconomic variables (listed in Table 1) from the FRED-MD database, where yields and housing correspond to innovations to sparse principal components whose active loadings allow them to be interpreted as yields and housing. The yields and housing factors in the sparse macro 3-factor model are mimicking portfolio returns computed using the Giglio and Xiu (forthcoming) three-pass methodology.

Table 7: **Out-of-sample pricing errors**

(1)	(2)	(3)	(4)	(5)	(6)	(7)
			$\lvert t\text{-statistic}\rvert \geq$			
Factor model	Mean	Std. dev.	1.645	1.96	2.58	3
Carhart (1997) 4-factor model	0.447	0.442	50.48	43.81	31.43	23.49
Fama and French (2015) 5-factor model	0.446	0.445	54.92	45.40	32.70	25.40
Hou, Xue, and Zhang (2015) q-factor model	0.375	0.398	41.90	32.38	23.17	19.37
Sparse macro 3-factor model	0.423	0.443	41.27	35.24	24.13	20.32

The table reports summary statistics for pricing errors (in percent) for 315 long-short anomaly portfolio returns from Chen and Zimmermann (2020) based on the multifactor model in the first column. The sample period is 2010:01 to 2019:12 (120 observations); for one anomaly, the number of available observations is 24. The second (third) column reports the mean (standard deviation) of the absolute values of estimated intercept terms across 315 multiple regressions, each of which regresses a long-short anomaly portfolio return on an intercept term and the set of factors in the first column. The last four columns report the percent of t-statistics for the intercept terms that are larger (in absolute value) than 1.645, 1.96, 2.58, and 3, respectively; t-statistics are computed using heteroskedasticity- and autocorrelation-robust standard errors (Newey and West 1987). The Carhart (1997) 4-factor model includes market, size, value, and momentum factors. The Fama and French (2015) 5-factor includes market, size, value, profitability, and investment factors. The Hou, Xue, and Zhang (2015) q-factor model includes market, size, investment, and return-on-equity factors. The sparse macro 3-factor model includes market, yields, and housing factors. The yields and housing factors are constructed by first applying sparse principal component analysis to 120 macroeconomic variables (listed in Table 1) from the FRED-MD database, where yields and housing correspond to innovations to sparse principal components whose active loadings allow them to be interpreted as yields and housing. The yields and housing factors in the sparse macro 3-factor model are mimicking portfolio returns computed using the Giglio and Xiu (forthcoming) three-pass methodology. The weights for constructing the mimicking portfolio returns are based on data through 2009:12.

Table 8: Explaining yields and housing factors

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Carhart (1997) four-factor model			Fama and French (2015) five-factor model			Hou, Xue, and Zhang (2015) q-factor model		
Factor	Yields	Housing	Factor	Yields	Housing	Factor	Yields	Housing
Alpha (%)	−3.313***	3.082***	alpha (%)	−3.294***	3.942***	alpha (%)	−3.166***	3.126***
SMB	−0.711***	0.823**	SMB	−0.558*	0.932***	ME	−0.794***	1.223**
HML	−0.757**	1.373***	HML	−0.759*	0.867*	I/A	−0.680	0.945***
MOM	0.291	1.545***	RMW	0.603*	0.976**	ROE	0.266	1.501***
			CMA	−0.027	−0.208			
R^2 (%)	4.83	22.22%	R^2 (%)	4.71%	5.90%	R^2 (%)	3.74%	9.25%
GRS	29.45***		GRS	35.28***		GRS	25.37***	

The table reports ordinary least squares estimation results for regressions of sparse macro yields and housing factors on factors from the Carhart (1997) four-factor, Fama and French (2015) five-factor, and Hou, Xue, and Zhang (2015) q-factor models. The sample period is 1967:01 to 2019:12. SMB, HML, and MOM are the Carhart (1997) size, value, and momentum factors, respectively; SMB, HML, RMW, and CMA are the Fama and French (2015) size, value, profitability, and investment factors, respectively; ME, I/A, and ROE are the Hou, Xue, and Zhang (2015) size, investment, and return-on-equity factors, respectively. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively, based on t-statistics computed using heteroskedasticity- and autocorrelation-robust standard errors (Newey and West 1987). GRS is the Gibbons, Ross, and Shanken (1989) statistic for testing the null hypothesis that the intercepts (alphas) are jointly zero in the yields and housing factor regressions. The yields and housing factors are constructed by first applying sparse principal component analysis to 120 macroeconomic variables (listed in Table 1) from the FRED-MD database, where yields and housing correspond to innovations to sparse principal components whose active loadings allow them to be interpreted as yields and housing. The yields and housing factors are mimicking portfolio returns computed using the Giglio and Xiu (forthcoming) three-pass methodology.

Table 9: Explaining leading factors

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Carhart (1997) four-factor model			Fama and French (2015) five-factor model				Hou, Xue, and Zhang (2015) q-factor model		
Factor	SMB	HML	MOM	SMB	HML	RMW	CMA	ME	I/A	ROE
Alpha (%)	0.003	0.105	0.264	0.011	0.105	0.274**	0.209**	0.031	0.287***	0.463***
Yields	−0.024**	−0.024**	0.045**	−0.026**	−0.024**	0.019***	−0.013	−0.032***	−0.007	0.019
Housing	0.018	0.026*	0.112***	0.023**	0.026*	0.015	0.008	0.028**	0.011	0.031***
R^2 (%)	2.32	3.63	16.95	3.24	3.63	2.45	1.34	4.76	1.18	12.77
GRS	1.51			4.96***				12.77***		

The table reports ordinary least squares estimation results for regressions of factors from the Carhart (1997) four-factor, Fama and French (2015) five-factor, and Hou, Xue, and Zhang (2015) q-factor models on sparse macro yields and housing factors. The sample period is 1967:01 to 2019:12. SMB, HML, and MOM are the Carhart (1997) size, value, and momentum factors, respectively; SMB, HML, RMW, and CMA are the Fama and French (2015) size, value, profitability, and investment factors, respectively; ME, I/A, and ROE are the Hou, Xue, and Zhang (2015) size, investment, and return-on-equity factors, respectively. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively, based on t-statistics computed using heteroskedasticity- and autocorrelation-robust standard errors (Newey and West 1987). GRS is the Gibbons, Ross, and Shanken (1989) statistic for testing the null hypothesis that the intercepts (alphas) are jointly zero in regressions for the group of factors. The yields and housing factors are constructed by first applying sparse principal component analysis to 120 macroeconomic variables (listed in Table 1) from the FRED-MD database, where yields and housing correspond to innovations to sparse principal components whose active loadings allow them to be interpreted as yields and housing. The yields and housing factors are mimicking portfolio returns computed using the Giglio and Xiu (forthcoming) three-pass methodology.

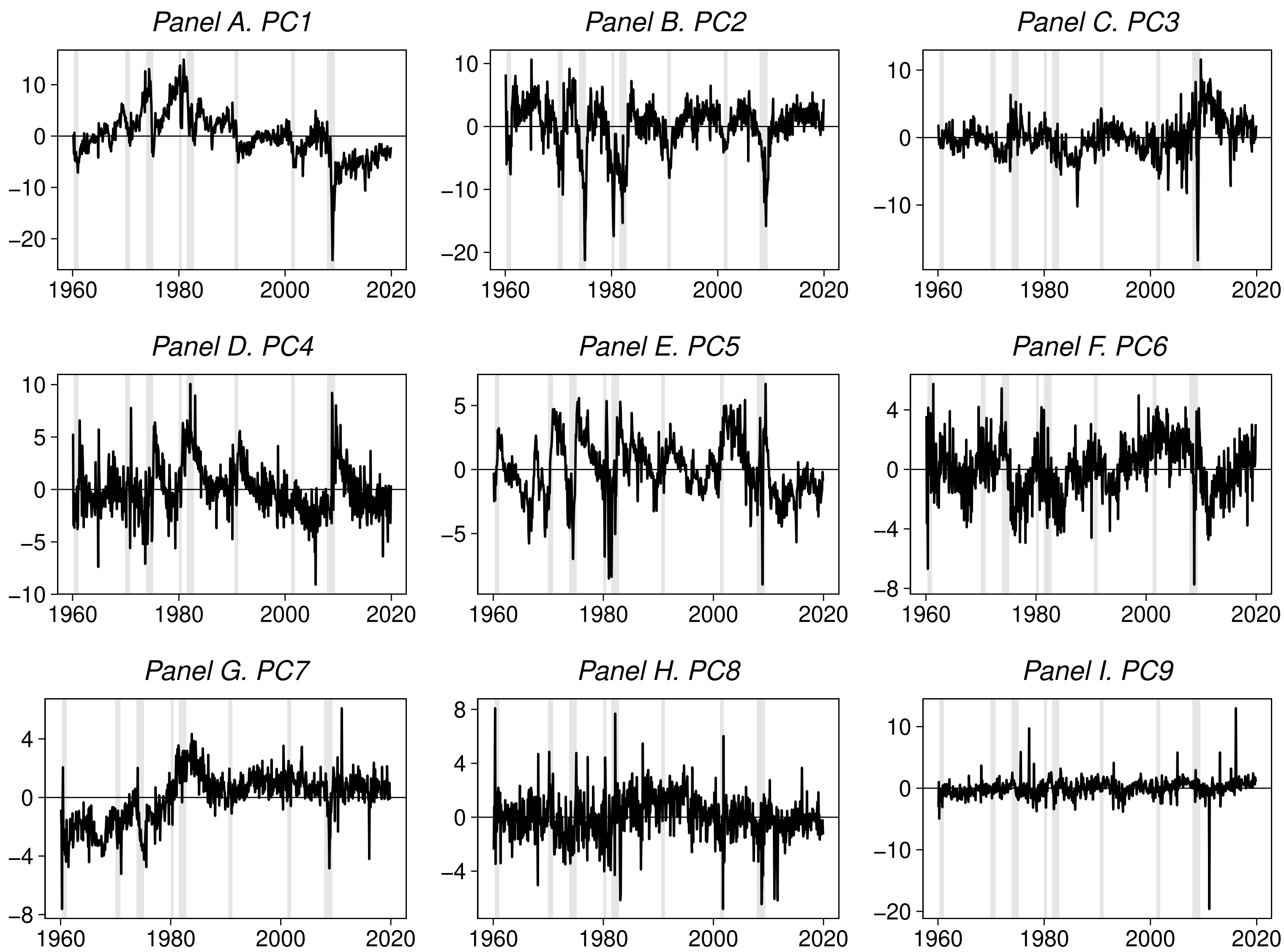

Figure 1: Conventional principal components

The figure depicts the first nine conventional principal components extracted from 120 macroeconomic variables (listed in Table 1) from the FRED-MD database. The sample period is 1960:02 to 2019:12. Vertical bars delineate recessions as dated by the National Bureau of Economic Research.

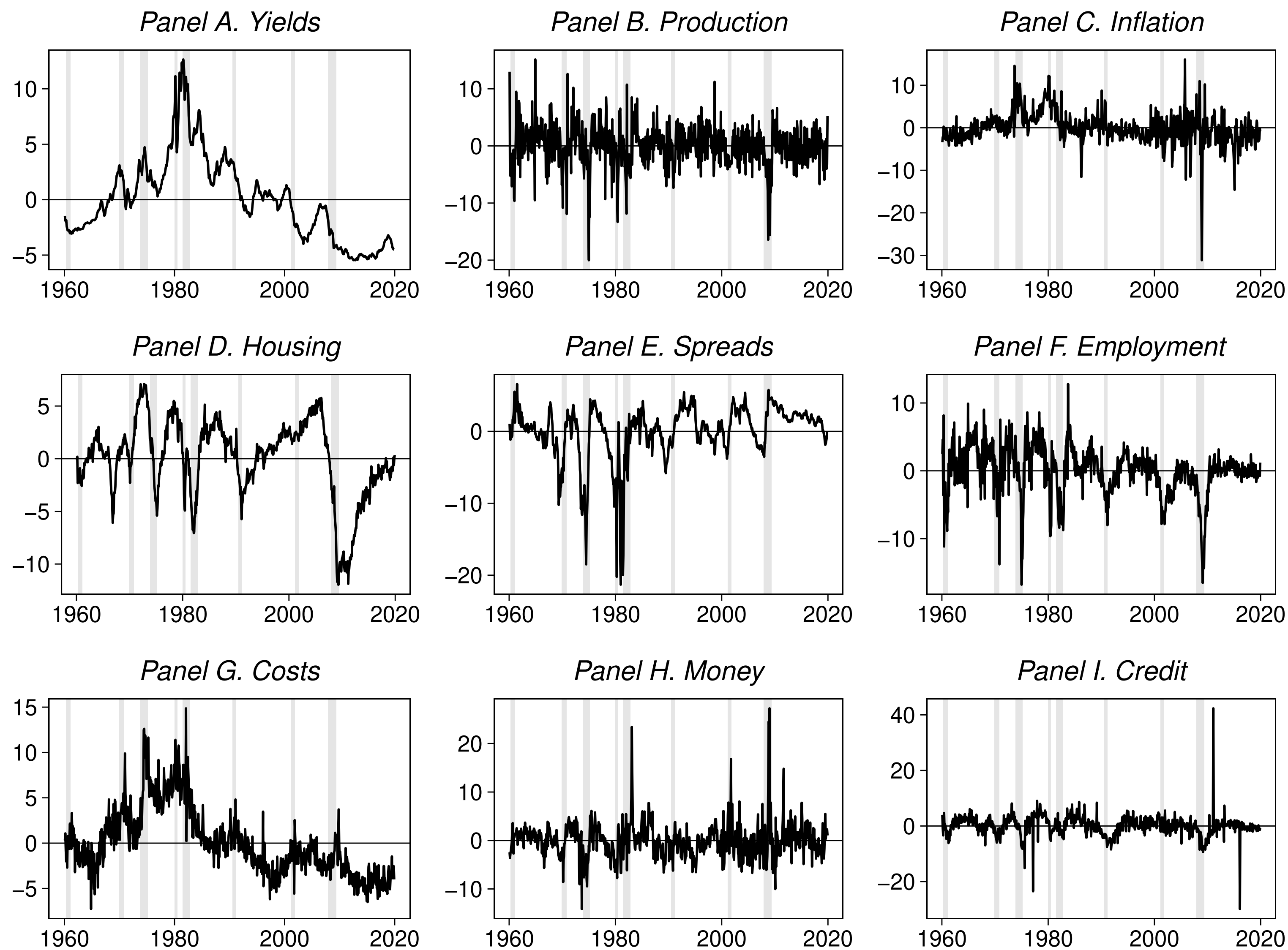

Figure 2: Sparse principal components

The figure depicts the first nine sparse principal components extracted from 120 macroeconomic variables (listed in Table 1) from the FRED-MD database. The panel headings provide descriptions of the sparse principal components based on the active elements of their weight vectors. The components are computed using the Witten, Tibshirani, and Hastie (2009) penalized matrix decomposition methodology. The sample period is 1960:02 to 2019:12. Vertical bars delineate recessions as dated by the National Bureau of Economic Research.

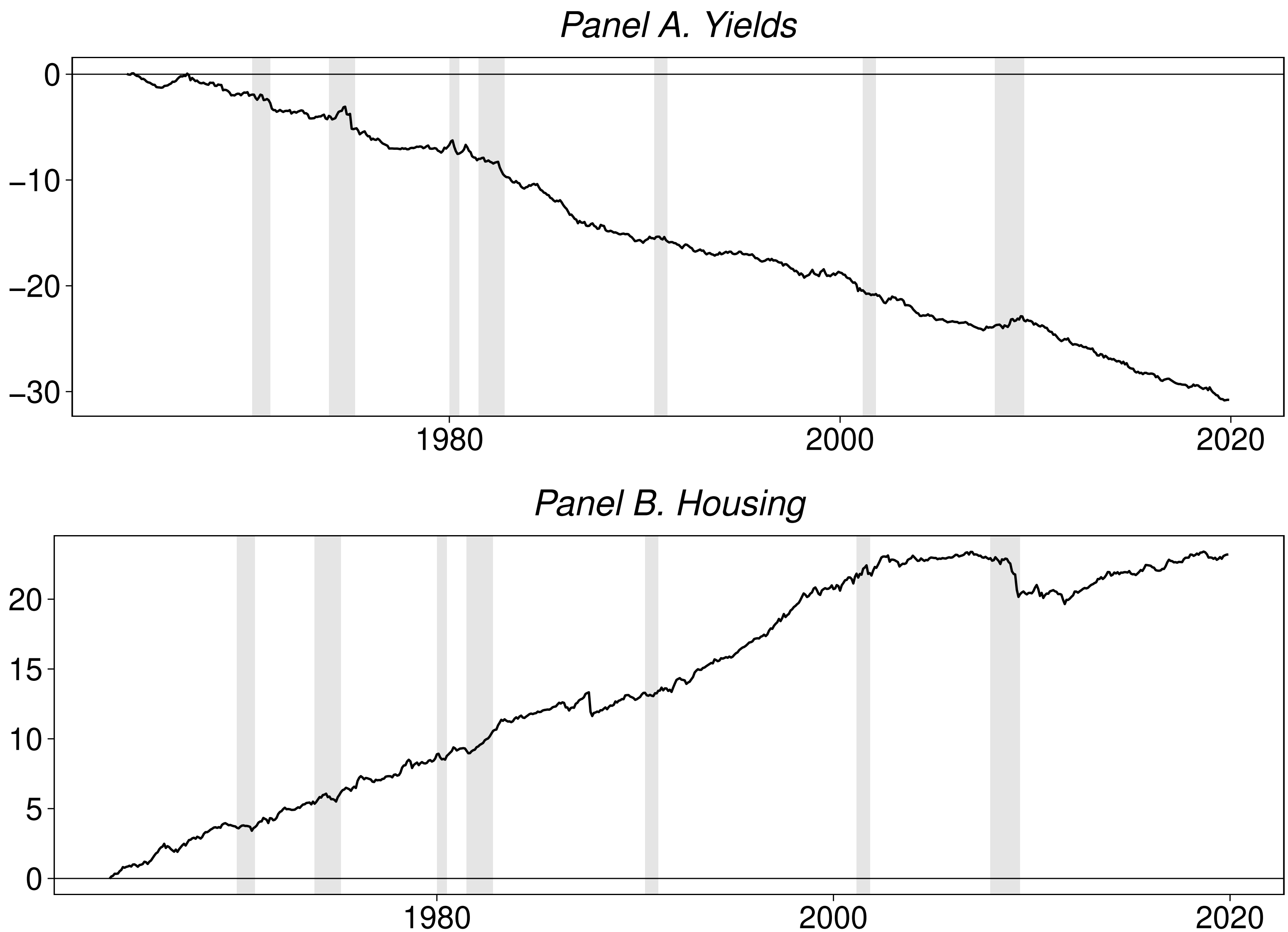

Figure 3: Log cumulative returns for factor mimicking portfolios

The figure depicts log cumulative returns for mimicking portfolios for the yields and housing factors. The factor mimicking portfolios are constructed using the Giglio and Xiu (forthcoming) three-pass methodology. The sample period is 1963:07 to 2019:12. Vertical bars delineate recessions as dated by the National Bureau of Economic Research.

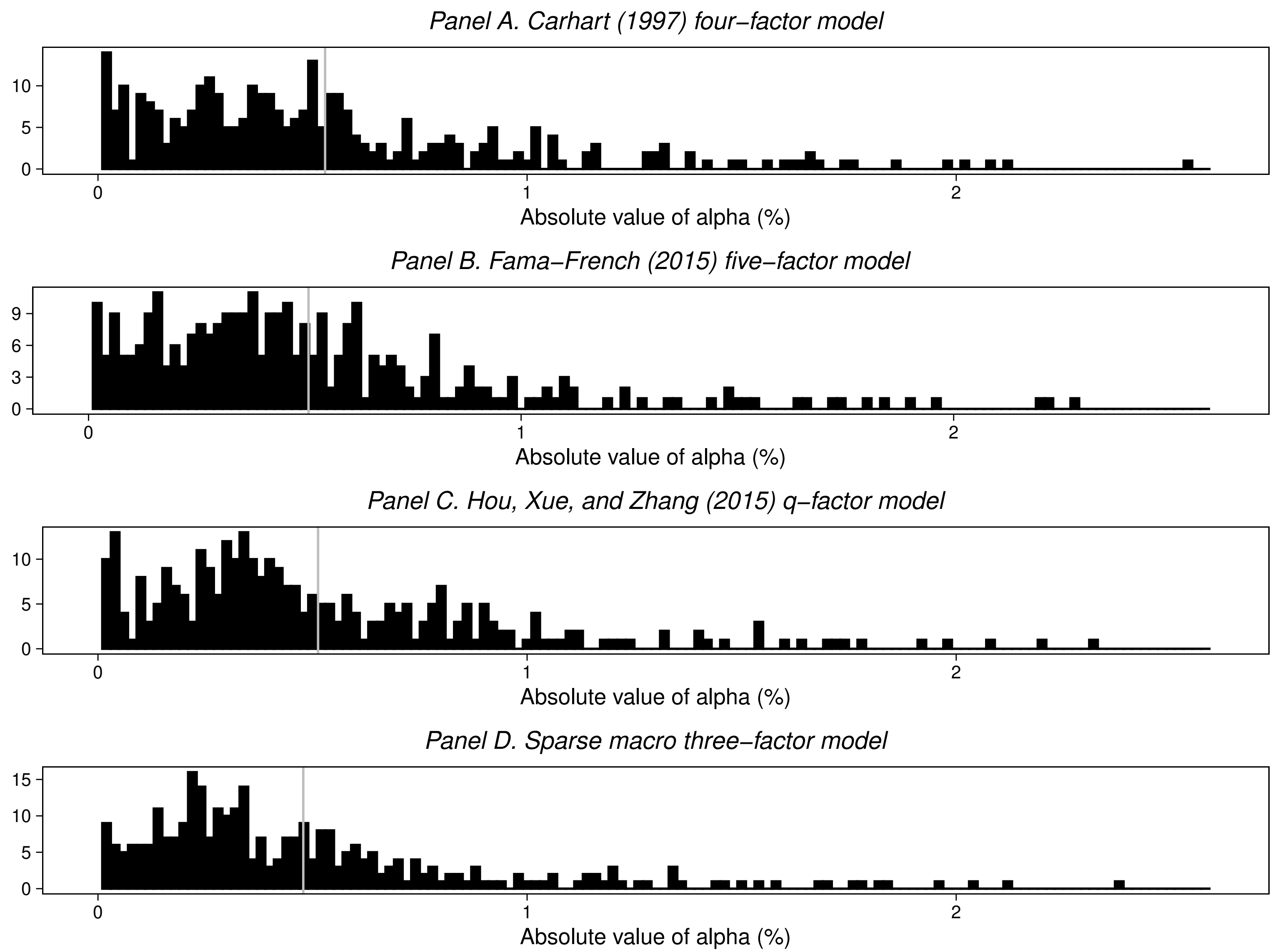

Figure 4: Histograms for pricing errors

The figure depicts histograms for the absolute values of pricing errors (in percent) for four multifactor models. The pricing errors are for 315 long-short anomaly portfolio returns from Chen and Zimmermann (2020). The sample period is 1967:01 to 2019:12. The vertical gray line delineates the mean.

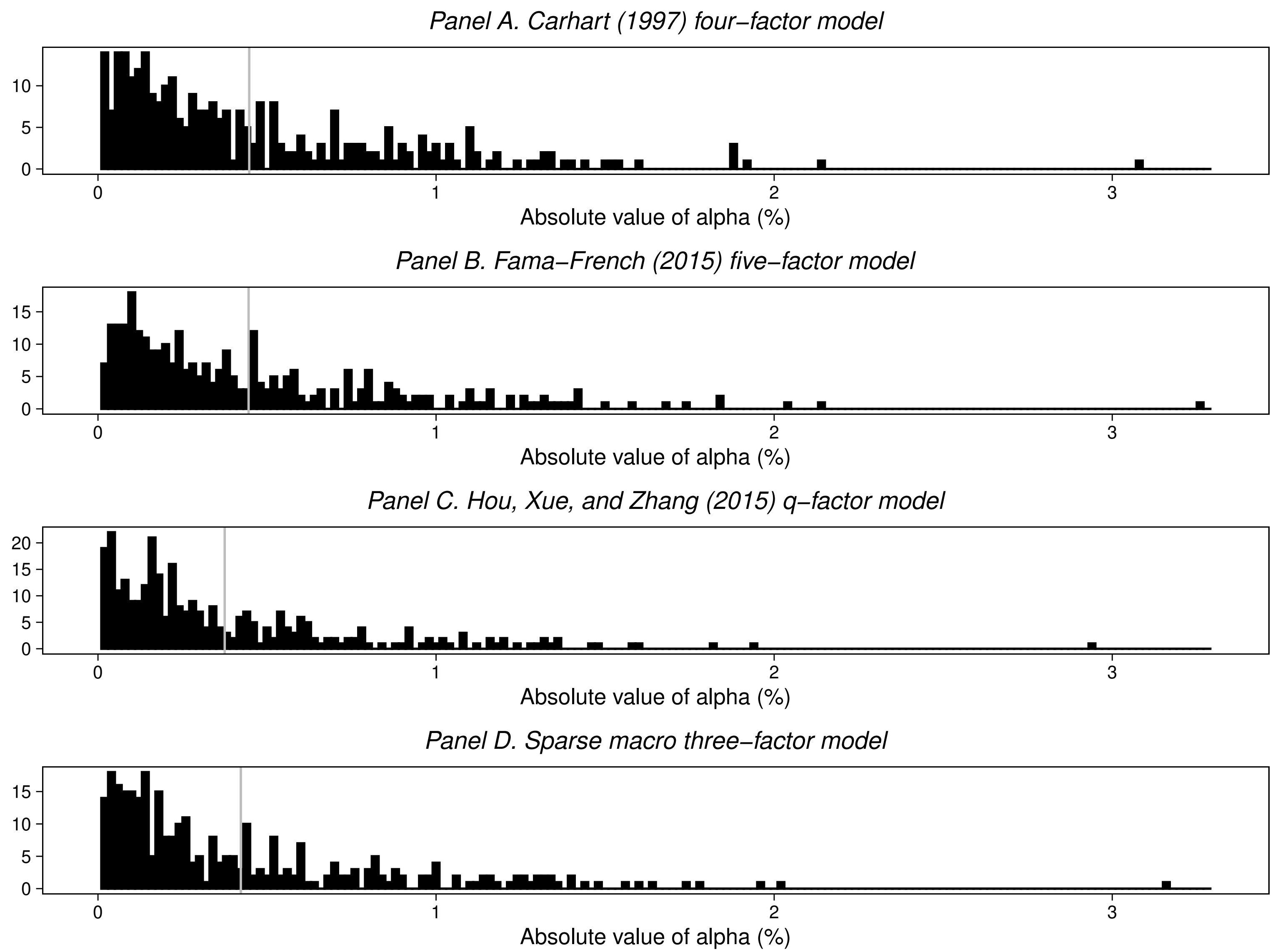

Figure 5: Histograms for out-of-sample pricing errors

The figure depicts histograms for the absolute values of pricing errors (in percent) for four multifactor models. The pricing errors are for 315 long-short anomaly portfolio returns from Chen and Zimmermann (2020). The sample period is 2010:01 to 2019:12. The weights for the mimicking portfolio returns for the yields and housing factors in the sparse macro three-factor model are computed using data through 2009:12. The vertical gray line delineates the mean.