

Quantitative Finance

ISSN: 1469-7688 (Print) 1469-7696 (Online) Journal homepage: www.tandfonline.com/journals/rquf20

A regularized instrumented principal components analysis

Konstantinos Baltas, Gustavo Fruet Dias & Artiom Rumiancev

To cite this article: Konstantinos Baltas, Gustavo Fruet Dias & Artiom Rumiancev (05 Nov 2026): A regularized instrumented principal components analysis, Quantitative Finance, DOI: 10.1080/14697688.2026.2740081

To link to this article: https://doi.org/10.1080/14697688.2026.2740081

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 05 Nov 2026.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=rquf20

Quantitative Finance, 2026
https://doi.org/10.1080/14697688.2026.2740081

Routledge
Taylor & Francis Group
Check for updates

A regularized instrumented principal components analysis

KONSTANTINOS BALTAS *†, GUSTAVO FRUET DIAS ‡ and ARTIOM RUMIANCEV§

†Essex Business School, University of Essex, Colchester, UK
‡School of Economics, University of East Anglia, Norwich, UK
§Department of Mathematics, King's College London, London, UK

(*Received 10 August 2025; accepted 18 September 2026*)

This paper examines instrumented principal components analysis (IPCA), one of the newly developed machine learning methods to model asset returns. It proposes the implementation of a support vector regression (SVR) into IPCA to better capture the relationship between firm characteristics and returns. Using a set of 9689 equities and 46 characteristics we show that IPCA(SVR) has better out-of-sample performance compared to the classic IPCA and multi-factor models in the majority of cases. The results imply that the least squares approach of IPCA overfits the characteristic sensitivity matrix, resulting in better in-sample results but worse out-of-sample fit than regularized IPCA. Although the regularized specifications exhibit superior predictive performance, the null hypothesis of no pricing errors is rejected in several testing environments, highlighting the distinction between predictive accuracy and asset-pricing validity.

Keywords: IPCA; Support vector regression; Asset pricing; Machine learning

1. Introduction

More than half a century has passed since the pioneering works of Treynor (1961, 1962), Sharpe (1964) and later Lintner (1965a, 1965b), and Mossin (1966), building on the modern portfolio theory (Markowitz 1952), laid the foundation to the equilibrium theory of asset pricing (AP henceforth). Independently, their developments suggested that market-wide risk is the core driver of stock returns and influenced the creation of the capital asset pricing model (CAPM henceforth), which despite its conceptual breakthrough, did not yield the power to explain the complex risk-return relationship. As such, numerous attempts to grasp the market-wide risk were made, leading to findings of dozens of return drivers from economically important characteristics.

The key stream of literature that laid the foundation to modern AP originated from the infamous APT framework published by Ross (1976), suggesting that distinct types of systematic risk drive asset compensation. Conversely to the CAPM, APT considered the stochastic discount factor (SDF henceforth) to be a linear combination of various return factors, which prompted the investigation into the drivers of asset returns outside the overall market risk, where Basu (1977), Stattman (1980), and Banz (1981) were among the first to recognize the impact that company growth and size has on returns. Subsequently, a research domain, engaged with identifying financial and economic forces that are associated with asset compensation was established. As a result, propositions of static multi factor models to account for systemic risk were, among many, made by Chen *et al.* (1986) (macroeconomic drivers), Fama and French (1993) (FF henceforth) (company characteristics), Carhart (1997) (persistence in returns), and Fama and French (2015) (investment and profitability risks). Furthermore, within APT, evidence for the spacial interaction between assets has also been identified (Kou *et al.* 2018).

The central concept that these approaches used to account for the unobservable return drivers, in modelling terms, was to construct characteristic-sorted portfolios (CPs henceforth), which would act as proxies for identifying common risk factors. However, along with the growing amount of the multi-factor literature, questions about the importance of certain proxies above others became prevalent. For instance, some frameworks as BARRA (1998) housed tens of characteristics to explain the cross-section of assets, which led to discussions of whether all proposed factors are equally important for modelling returns, dubbed by Cochrane (2011) as the 'factor zoo' problem. Furthermore, research has shown that loadings are both dynamic (e.g. Hansen and Richard 1987, Jagannathan

*Corresponding author. Email: k.baltas@essex.ac.uk

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

and Wang 1996) and depend on latent factors (Chamberlain and Rothschild 1983).

In order to address these limitations, recent studies began employing machine learning (ML henceforth) techniques that are intended to process large datasets and can account for time-variation. Some of the late methodologies were proposed by Kelly *et al.* (2019) (KPS henceforth), Chen *et al.* (2024) (CPZ henceforth), and Gu *et al.* (2021) (GKX henceforth), which relied on ML techniques and neural network structures to map large sets of characteristics onto a small number of explanatory variables. Their results suggested that under traditional economic assumptions, ML is applicable for both pricing and prediction, and can outperform econometric models in both cross-sectional and time-series contexts. More specifically, instrumented principal component analysis (IPCA henceforth) Kelly *et al.* (2019) is one of these ML approaches which uses dynamic PCA to estimate time-varying loadings and constructs latent factors.

Conceptually, IPCA extends the works of Chamberlain and Rothschild (1983) and Connor and Korajczyk (1986, 1988) who first introduced PCA into asset pricing. Stand-alone PCA is rarely employed in practice as it only incorporates returns data and produces unstable principal components, which change when using a contrasting set of test assets (Bray 1994). Kelly *et al.* (2019) overcome these limitations by allowing interaction between characteristics and a return-sensitivity matrix. Consequently, IPCA can handle unbalanced datasets and produces competitive in-sample (IS henceforth) and out-of-sample (OOS henceforth) results compared to traditional multi-factor models (Kelly *et al.* 2019). Because IPCA is a latent factor model, it requires an inspection of factor mappings in order to determine the economic meaning behind the pricing result. Beyond equities, IPCA has also been applied in option pricing (Büchner and Kelly 2022).

Nonetheless, IPCA has some limitations. Firstly, IPCA is a linear parametric model. Newer approaches by Chen *et al.* (2024) and Gu *et al.* (2021) demonstrate that both nonlinear characteristic mapping and semi-parametric AP is crucial. Chen *et al.* (2024) propose a GAN-GMM system with 46 firm characteristics and 178 macroeconomic variables, which produces less errors than IPCA. Gu *et al.* (2021), on the other hand, replace the dynamic PCA optimization of Kelly *et al.* (2019) with an auto-encoder, which results in higher factor Sharpe ratios compared to IPCA. Because of the large characteristic selection pool, IPCA could also be prone to overfitting with multiple feature mapping onto limited number of factors (Kozak *et al.* 2020). Furthermore, the high correlation between some characteristics, as evidenced from figure A1, could result in multicollinearity. Bybee *et al.* (2023) attempt to address this issue by introducing group lasso regularization into the IPCA framework (Sparse IPCA) within a narrative news text-based asset pricing setting. They find that the regularized IPCA approach filters out narratives that are irrelevant for explaining asset price dynamics and improves out-of-sample asset pricing performance. Finally, IPCA, has only been tested using equities and characteristic-sorted portfolios. As argued by Lewellen *et al.* (2010), Daniel and Titman (1997, 2012), and Daniel (2018), characteristics-based models will intuitively price same-characteristic sorted portfolios.

This paper contributes to the literature on estimation and testing within the IPCA framework, which extends the traditional Fama-MacBeth two-pass methodology to settings in which the underlying state variables are not directly observed. More specifically, we focus on two issues that are particularly relevant in empirical asset pricing: overfitting and test-asset choice. First, we propose the use of linear Support Vector Regression (SVR) as an alternative approach for estimating the characteristic sensitivity matrix within the IPCA framework. We refer to this methodology as IPCA(SVR). The standard IPCA model is estimated using the Alternating Least Squares (ALS) algorithm of Zhou *et al.* (2008), which solves the optimization problem without an explicit regularization penalty. In contrast, our approach introduces an L_2 regularization term through linear SVR. Support Vector Regression extends the Support Vector Machine (SVM) framework developed by Boser *et al.* (1992) and Cortes and Vapnik (1995) from classification to regression problems. Unlike ordinary least squares estimation, SVR combines an explicit L_2 penalty with an ε-insensitive loss function, allowing the estimation procedure to control model complexity while reducing the impact of small prediction errors. By introducing regularization into the estimation of the characteristic sensitivity matrix, IPCA(SVR) aims to mitigate multicollinearity, reduce overfitting, and improve out-of-sample asset pricing performance. Second, we address the test-asset choice concern highlighted by Daniel (2018). To do so, we compare the performance of IPCA(SVR), standard IPCA, the Carhart (1997) four-factor model (FFC4), and the Fama and French (2015) five-factor model (FF5) across three test-asset universes: characteristic-sorted portfolios, industry-sorted portfolios, and individual equities. Finally, Fan *et al.* (2016) show that industry portfolios explain a larger proportion of the market covariance structure than characteristic-sorted portfolios. Motivated by this finding, we provide an additional comparison between IPCA and IPCA(SVR) using industry portfolios as test assets, thereby offering a complementary assessment of the robustness and economic relevance of the proposed regularization framework.

The empirical results provide strong evidence that regularization plays an important role in improving the out-of-sample performance of machine learning-based asset pricing models while mitigating overfitting. They also highlight the importance of test asset selection when assessing whether a model can consistently price different assets and portfolios. More specifically, the results indicate that regularization improves IPCA's goodness of fit with the out-of-sample Total R^2 for the panel of equities being higher than that of the traditional IPCA model. The positive statistically significant difference between the OOS and IS Total and Predictive R^2 also indicates that the model exhibits superior OOS performance as compared to other approaches. Furthermore, IPCA(SVR) produces latent factors that generally have higher Sharpe Ratios than those of the classic IPCA. Moreover, IPCA(SVR)'s Predictive R^2 for CPs and industry-sorted portfolios outperforms those of FFC4 and FF5 models. Nonetheless, IPCA(SVR) produces, significant pricing errors in most periods in both portfolio and individual asset pricing tests. The results, however, imply that IPCA(SVR) is more suited for OOS estimation.

In relation to economic importance, we find dividend to market equity ratio and high 5-year beta variables, which correspond to the 'value' of assets, to be the core cross-sectional drivers of price variation. Another important group of variables that are integral to explaining the common variability of returns are long fixed costs to sales and market equity versus short total assets as well as book-to-market equity. These characteristics, captured by the second factor, determine the assets' 'profitability'.

2. Methodology

2.1. IPCA

Kelly *et al.* (2019) formulate the methodology around the classic no-arbitrage model of Ross (1976):

$$r^e_{i,t+1} = \alpha_{i,t} + \boldsymbol{\beta}'_{i,t}\boldsymbol{f}_{t+1} + \epsilon_{i,t+1}, \tag{1}$$

where $r^e_{i,t+1}$ represents the excess return, $\alpha_{i,t}$ is the pricing error (alpha), $\boldsymbol{\beta}'_{i,t} \in \mathbb{R}^{1\times K}$ is the vector of composite sensitivity coefficients (beta) and $\boldsymbol{f}_{t+1} \in \mathbb{R}^{K\times 1}$ are the latent factors. The multi-factor form assumes that $\mathbb{E}_t[\boldsymbol{f}_{t+1}] = \boldsymbol{\lambda}_t$, $Cov_t(\boldsymbol{f}_{t+1}, \epsilon_{i,t+1}) = 0$, and $\mathbb{E}_t[\epsilon_{i,t+1}] = \mathbb{E}_t[\alpha_{i,t}] = 0$. Allowing for multiple drivers of asset returns mapped onto a small number of latent factors, Kelly *et al.* (2019) define alpha and beta of (2) as composite coefficients:

$$\alpha_{i,t} = \boldsymbol{z}'_{i,t}\boldsymbol{\Gamma}_\alpha + \nu_{\alpha,i,t}, \quad \boldsymbol{\beta}_{i,t} = \boldsymbol{z}'_{i,t}\boldsymbol{\Gamma}_\beta + \boldsymbol{\nu}_{\beta,i,t}, \tag{2}$$

where $\boldsymbol{z}_{i,t} \in \mathbb{R}^{L\times 1}$ is a vector of observable asset characteristics, while $\boldsymbol{\Gamma}_\beta \in \mathbb{R}^{L\times K}$ and $\boldsymbol{\Gamma}_\alpha \in (\mathbb{R}^{L\times 1})$ matrices store parameter and pricing error coefficients, with $\boldsymbol{\nu}_{\beta,i,t} \in \mathbb{R}^{1\times K}$ and $\nu_{\alpha,i,t}$ being the factorization errors. This produces a system of T time instances, N assets, L characteristics and K latent factors.

Setting $\boldsymbol{\Gamma}_\alpha = 0$ imposes the restriction that the characteristics in $\boldsymbol{z}_{i,t}$ do not generate characteristic-driven pricing errors and therefore explain expected returns exclusively through their role in determining factor loadings.† Following Kelly *et al.* (2019), this specification is referred to as the *restricted model*. To test for the presence of pricing errors driven by the characteristics in $\boldsymbol{z}_{i,t}$, the restriction on $\boldsymbol{\Gamma}_\alpha$ is relaxed and matrix with the pricing error coefficients is estimated freely. This specification corresponds to the *unrestricted model* of Kelly *et al.* (2019). In matrix form, the unrestricted model reads

$$\boldsymbol{r}^e_{t+1} = \boldsymbol{Z}_t\bar{\boldsymbol{\Gamma}}\bar{\boldsymbol{f}}_{t+1} + \boldsymbol{\epsilon}^*_{t+1}, \tag{3}$$

where $\boldsymbol{r}^e_{t+1}$ and $\boldsymbol{\epsilon}^*_{t+1}$ are both $\mathbb{R}^{N\times 1}$ vectors, with $\boldsymbol{\epsilon}^*_{t+1} = \boldsymbol{\epsilon}_{t+1} + \boldsymbol{\nu}_{\alpha,t} + \boldsymbol{\nu}_{\beta,t}\boldsymbol{f}_{t+1}$ denoting the composite error term. The matrix $\boldsymbol{Z}_t \in \mathbb{R}^{N\times L}$ contains the firm characteristics observed at time t, while $\bar{\boldsymbol{\Gamma}} \equiv [\boldsymbol{\Gamma}_\alpha, \boldsymbol{\Gamma}_\beta] \in \mathbb{R}^{L\times\tilde{K}}$ denotes the α- and β-combined characteristic sensitivity matrix, where $\tilde{K} = K + 1$ includes the K latent factors and the intercept term. Finally, $\bar{\boldsymbol{f}}_{t+1} \equiv [1, \boldsymbol{f}'_{t+1}]' \in \mathbb{R}^{\tilde{K}\times 1}$ augments the factor vector with a constant. The estimation objective is to minimize the sum of composite errors:

$$\mathcal{L}_{\text{ALS}}(\bar{\boldsymbol{\Gamma}}, \bar{\boldsymbol{F}}) := \min_{\bar{\boldsymbol{\Gamma}},\bar{\boldsymbol{F}}} \sum_{t=1}^{T-1} (\boldsymbol{r}^e_{t+1} - \boldsymbol{Z}_t\bar{\boldsymbol{\Gamma}}\bar{\boldsymbol{f}}_{t+1})'(\boldsymbol{r}^e_{t+1} - \boldsymbol{Z}_t\bar{\boldsymbol{\Gamma}}\bar{\boldsymbol{f}}_{t+1}). \tag{4}$$

The derivation of the IPCA estimator and its numerical implementation are provided in Kelly *et al.* (2020). In short, the authors show that the first-order conditions give rise to a system of period-by-period cross-sectional regressions for which no closed-form solution exists. To estimate the model, Kelly *et al.* (2020) propose an Alternating Least Squares (ALS) procedure that jointly updates the latent factors and characteristic sensitivity matrix until convergence is achieved (see Appendix A of Kelly *et al.* 2019, for further details). Specifically, the algorithm iteratively solves the first-order conditions

$$\boldsymbol{f}_{t+1} = \left(\hat{\boldsymbol{\Gamma}}'_\beta\boldsymbol{Z}'_t\boldsymbol{Z}_t\hat{\boldsymbol{\Gamma}}_\beta\right)^{-1}\hat{\boldsymbol{\Gamma}}'_\beta\boldsymbol{Z}'_t\left(\boldsymbol{r}^e_{t+1} - \boldsymbol{Z}_t\hat{\boldsymbol{\Gamma}}_\alpha\right), \tag{5}$$

$$\text{vec}\left(\hat{\bar{\boldsymbol{\Gamma}}}'\right) = \left(\sum_{t=1}^{T-1}\boldsymbol{Z}'_t\boldsymbol{Z}_t \otimes \hat{\bar{\boldsymbol{f}}}_{t+1}\hat{\bar{\boldsymbol{f}}}'_{t+1}\right) \left(\sum_{t=1}^{T-1}\left[\boldsymbol{Z}_t \otimes \hat{\bar{\boldsymbol{f}}}'_{t+1}\right]'\boldsymbol{r}^e_{t+1}\right). \tag{6}$$

Finally, the choice of K must be made *ad hoc* by the researcher. Kelly *et al.* (2019) use a panel of 12,813 firms to investigate the optimal number of factors. Their results show that the IPCA model captures nearly all of the cross-sectional heterogeneity in average stock returns associated with firm characteristics when $K \geq 5$. Moreover, $K = 6$ is the only specification for which the null hypothesis $\boldsymbol{\Gamma}_\alpha = 0$ cannot be rejected at the 5% significance level (p-value = 52%, compared with 2.06% for $K = 5$). This finding suggests that $K = 6$ provides the optimal model specification. Accordingly, throughout this paper, we adopt the best-performing specification identified by Kelly *et al.* (2019) and set the number of latent factors to $K = 6$.

2.2. IPCA with linear SVR

Kelly *et al.* (2019) suggest that each additional IPCA factor requires $T + L$ parameters to be estimated, compared to PCA's $T + N$. For a 300 month period with 46 characteristics, 6 factors and intercept, 2422 parameters need to be computed, which is nonetheless a substantial figure. Considering the similarity of certain characteristics, problems of multicollinearity may arise, with some parameters being redundant. Furthermore, IPCA attempts to maximize explained variance of IS returns, which may lead to overfitting.

To address these issues, we propose estimating $\bar{\boldsymbol{\Gamma}}$ in (3) using linear SVR, thereby introducing an explicit regularization mechanism into the IPCA framework. The estimation procedure can be summarized in two steps. First, assets are aggregated into characteristic-weighted portfolios (CWPs)—also often refferenced as managed portfolios in the asset

† Note that imposing $\boldsymbol{\Gamma}_\alpha = 0$ does not imply that $\alpha_{i,t} = 0$. From the model specification, $\alpha_{i,t}$ may still differ from zero through the idiosyncratic component $\nu_{\alpha,i,t}$.

pricing literature—using all N_{t+1} assets with non-missing observations:

$$\boldsymbol{r}^c_{t+1} = \frac{1}{N_{t+1}} \sum_{i=1}^{N} \boldsymbol{z}'_{i,t} r^e_{i,t+1} = \frac{\boldsymbol{Z}'_t \boldsymbol{r}^e_{t+1}}{N_{t+1}}, \tag{7}$$

where $\boldsymbol{r}^c_{t+1} \in \mathbb{R}^{L\times 1}$ and the ℓ-th element of $\boldsymbol{r}^c_{t+1}$, for $\ell = 1,\ldots,L$, corresponds to the return on a characteristic-weighted portfolio formed using the ℓ-th characteristic as portfolio weights. Thus, each element of $\boldsymbol{r}^c_{t+1}$ is a cross-sectional weighted average of stock returns, where the weights are determined by the corresponding characteristic values observed at time t. Stacking the characteristic-weighted portfolio returns over time yields the matrix $\boldsymbol{R}^c = \left[\boldsymbol{r}^c_2,\ldots,\boldsymbol{r}^c_T\right]' \in \mathbb{R}^{(T-1)\times L}$. Working with characteristic-weighted portfolios (CWPs) instead of the full panel of stock returns offers two important advantages. First, it considerably reduces the dimensionality of the estimation problem from $T \times N$ at the stock level to $T \times L$ in the CWP setting, without sacrificing information relevant for the estimation of the IPCA model (Kelly *et al.* 2019). Second, it naturally avoids the missing-observation problem inherent in the original unbalanced panel of stock returns, since CWPs can be constructed using all available observations at each point in time.

The second stage of the estimation procedure consists of estimating the characteristic sensitivity matrix $\bar{\boldsymbol{\Gamma}}$ and the latent factor realizations $\boldsymbol{f}_{t+1}$. To do so, we adopt an iterative procedure analogous to the ALS algorithm described above, with the exception that equation (6) is replaced by an SVR-based estimation step. Specifically, given an initial estimate of $\bar{\boldsymbol{\Gamma}}$, factor realizations are obtained as

$$\boldsymbol{f}_{t+1} = \left(\hat{\boldsymbol{\Gamma}}'_\beta \boldsymbol{Z}'_t \boldsymbol{Z}_t \hat{\boldsymbol{\Gamma}}_\beta\right)^{-1} \hat{\boldsymbol{\Gamma}}'_\beta \boldsymbol{Z}'_t \left(\boldsymbol{r}^e_{t+1} - \boldsymbol{Z}_t \hat{\boldsymbol{\Gamma}}_\alpha\right). \tag{8}$$

where $\boldsymbol{Z}_t \widehat{\boldsymbol{\Gamma}}_\beta$ is the dot product of asset characteristics and the α- and β-combined characteristic sensitivity matrix at time t. The updated estimate of $\bar{\boldsymbol{\Gamma}}$ is then obtained by solving the following SVR optimization problem using the coordinate-descent algorithm of Hsieh *et al.* (2008):

$$\begin{aligned}\mathcal{L}_{\text{SVR}}(\widehat{\bar{\boldsymbol{\Gamma}}}, C, \varepsilon) := \min_{\widehat{\bar{\boldsymbol{\Gamma}}}} \frac{1}{2}\text{vec}(\widehat{\bar{\boldsymbol{\Gamma}}})'\text{vec}(\widehat{\bar{\boldsymbol{\Gamma}}}) \\ + C \sum_{i=1}^{N}\sum_{t=1}^{T-1} \max(0, |\xi_{i,t+1}| - \varepsilon),\end{aligned} \tag{9}$$

where $\xi_{i,t+1} = r^e_{i,t+1} - \boldsymbol{z}'_{i,t}\bar{\boldsymbol{\Gamma}}\bar{\boldsymbol{f}}_{t+1}$ denotes the residual associated with the loss function, $\text{vec}(\widehat{\bar{\boldsymbol{\Gamma}}}) \in \mathbb{R}^{P\times 1}$ is the vectorized estimates of the coefficient matrix with dimension $P = \tilde{K}L$, and C and ε are the SVR regularization hyperparameters. The optimization problem is solved using the linear SVR implementation of Fan *et al.* (2008) available through Scikit-learn's `sklearn.svm.LinearSVR` routine (Pedregosa *et al.* 2011).

To initialize the estimation procedure, unstandardized PCA is applied to the matrix of characteristic-weighted portfolios. The resulting K eigenvectors, collected in $\boldsymbol{V} \in \mathbb{R}^{L\times K}$, are used as initial estimates of $\boldsymbol{\Gamma}$, which are then employed to obtain preliminary factor realizations. After obtaining these initial estimates, the steps described in (8)–(9) are iteratively repeated while imposing the identification restrictions on orthogonality and sign convention until convergence is achieved under a pre-specified error tolerance. The pseudocode reported in Algorithm 1, adapted from Büchner and Bybee (2021), summarizes the estimation procedure for the unrestricted IPCA(SVR) model.

An important feature of the ε-insensitive loss function is that it allows for a user-defined error tolerance. In our baseline specification, the hyperparameters C and ε are selected using the data-driven procedure proposed by Cherkassky and Ma (2004). To assess the robustness of the results to the choice of hyperparameters, we additionally implement a Grid Search cross-validation procedure to identify the optimal combination of C and ε. A detailed discussion of both approaches is provided in Section 2.3.

Distinguishing between the *restricted* and *unrestricted* model boils down to formally test whether the expected returns depend on characteristics that are not explained by the latent common factors. The null hypothesis, $H_0 : \boldsymbol{\Gamma}_\alpha = \boldsymbol{0}$, states that firm characteristics explain expected returns exclusively through their role in determining factor loadings and therefore do not generate characteristic-based pricing errors. Under the alternative hypothesis, $\boldsymbol{\Gamma}_\alpha \neq \boldsymbol{0}$, characteristics contribute directly to expected returns, implying the presence of systematic pricing errors that cannot be explained by the latent factor structure. Following Kelly *et al.* (2019), we construct a Wald-type statistic,

$$W_\alpha = \text{vec}(\hat{\boldsymbol{\Gamma}}_\alpha)'\text{vec}(\hat{\boldsymbol{\Gamma}}_\alpha),$$

which measures the distance between the unrestricted estimate and the null restriction $\boldsymbol{\Gamma}_\alpha = \boldsymbol{0}$. Large values of W_α indicate that characteristics contribute significantly to pricing errors after controlling for their role in factor loadings.

To obtain valid finite-sample inference, we adopt the residual bootstrap procedure of Kelly *et al.* (2019). This bootstrap procedure has the advantage of being robust to conditional heteroskedasticity through the use of a wild-bootstrap scheme and generally exhibits superior finite-sample properties relative to asymptotic-based testing procedures. Furthermore, the residual bootstrap relies on relatively weak assumptions on the error process. In particular, the residuals are only required to form a martingale difference sequence with a finite fourth moment and may follow a broad class of conditionally heteroskedastic processes, including GARCH-type models and stochastic volatility specifications (Gonçalves and Kilian 2004). More specifically, the bootstrap procedure work as follows. First, the *unrestricted* model is estimated and both the parameters, $\widehat{\boldsymbol{\Gamma}}_\alpha$ and $\widehat{\boldsymbol{\Gamma}}_\beta$, and the common factor estimates $\{\widehat{\boldsymbol{f}}_t\}_{t=1}^T$ are collected. Using these estimates, the residuals of the characteristic-weighted portfolios are computed as

$$\widehat{\boldsymbol{d}}_{t+1} = \boldsymbol{r}^c_{t+1} - (\boldsymbol{Z}'_t\boldsymbol{Z}_t)\widehat{\boldsymbol{\Gamma}}_\alpha - (\boldsymbol{Z}'_t\boldsymbol{Z}_t)\widehat{\boldsymbol{\Gamma}}_\beta\widehat{\boldsymbol{f}}_{t+1}, \quad t = 1,\ldots,T. \tag{10}$$

Next, bootstrap residuals are generated by sampling with replacement from the set $\{\widehat{\boldsymbol{d}}_{t+1}\}_{t=1}^T$. Let $\widehat{\boldsymbol{d}}_{b,t+1}$ denote the residual selected in bootstrap replication b, where $b = 1,\ldots,B$. To account for potential conditional heteroskedasticity in the

original series, the resampled residuals are multiplied by independent random variables,

$$\widetilde{\boldsymbol{d}}_{b,t+1} = \boldsymbol{q_{b,t+1}}\widehat{\boldsymbol{d}}_{b,t+1},$$

where $\boldsymbol{q_{b,t+1}}$ is a Student-t random variable with zero mean, unit variance, and five degrees of freedom. The b-th bootstrap sample of characteristic-weighted portfolio returns is then generated as

$$\boldsymbol{r}^{c}_{b,t+1} = (\boldsymbol{Z}_t'\boldsymbol{Z}_t)\widehat{\boldsymbol{\Gamma}}_{\beta}\widehat{\boldsymbol{f}}_{t+1} + \widetilde{\boldsymbol{d}}_{b,t+1}, \quad t = 1, \ldots, T. \tag{11}$$

Note that the bootstrap sample $\{\boldsymbol{r}^{c}_{b,t+1}\}_{t=1}^{T}$ is generated under the null hypothesis $\boldsymbol{\Gamma}_{\alpha} = \boldsymbol{0}$. This is achieved by excluding the estimated characteristic-driven pricing-error component from the bootstrap data-generating process while retaining the estimated factor structure and bootstrap residuals. As a result, the empirical distribution of $W_{b,\alpha}$ provides an approximation of the sampling distribution of the test statistic under the null hypothesis. Using the bootstrap sample $\{\boldsymbol{r}^{c}_{b,t+1}\}_{t=1}^{T}$, the unrestricted IPCA(SVR) model is re-estimated and new estimates of $\boldsymbol{\Gamma}_{\alpha}$, $\boldsymbol{\Gamma}_{\beta}$, and $\{\boldsymbol{f}_t\}_{t=1}^{T}$ are obtained. The bootstrap test statistic is then computed as

$$W_{b,\alpha} = \operatorname{vec}\left(\widehat{\boldsymbol{\Gamma}}_{b,\alpha}\right)' \operatorname{vec}\left(\widehat{\boldsymbol{\Gamma}}_{b,\alpha}\right).$$

The entire procedure is repeated $B = 1,000$ times, producing the bootstrap distribution $\{W_{b,\alpha}\}_{b=1}^{B}$. Finally, the bootstrap p-value is computed as the fraction of bootstrap statistics that exceed the value of W_{α} obtained from the original sample.

It is important to highlight that for each bootstrap replication, the IPCA(SVR) model is re-estimated using exactly the same estimation procedure and hyperparameter configuration as in the original sample. This ensures that the bootstrap distribution reflects the additional uncertainty associated with the regularized estimation procedure.

2.3. Meta-parameters

Hyperparameter selection plays a crucial role in the performance of Support Vector Machines (SVMs) and Support Vector Regression (SVR) models (Chapelle *et al.* 2001). However, there is no universally accepted framework for determining the optimal values of the regularization parameter C and the ε-insensitive loss threshold ε. Consequently, a variety of approaches have been proposed in the literature.

Among the most widely used methods are Grid Search procedures (see, e.g. Hsu *et al.* 2003), meta-heuristic optimization algorithms (see, e.g. Talbi 2009), and gradient-based optimization techniques (see, e.g. Keerthi *et al.* 2006). These methods numerically search for hyperparameter combinations that maximize predictive performance. An alternative approach relies on data-driven plug-in estimators that determine hyperparameters directly from the statistical properties of the training sample. While numerical optimization techniques often identify highly competitive hyperparameter configurations, they can become computationally expensive, particularly in iterative estimation procedures such as IPCA(SVR) (Cherkassky and Ma 2004). This issue is especially relevant when the model is estimated using large asset panels and repeated recursively over time.

For this reason, we consider two complementary appro aches. Our baseline specification adopts the analytical plug-in procedure proposed by Cherkassky and Ma (2004), while a Grid Search cross-validation procedure is used as a robustness check.

The approach of Cherkassky and Ma (2004) determines C and ε directly from the statistical properties of the input data. The authors demonstrate that the resulting hyperparameters exhibit good generalization performance across a wide range of datasets, and the method has been successfully applied in financial time-series applications (see, e.g. Lu *et al.* 2009). Following Cherkassky and Ma (2004), the hyperparameters used in the IPCA(SVR) framework are computed as

$$C = \max(|\bar{\boldsymbol{R}}^{e} + 3\sigma_{\bar{\boldsymbol{R}}^{e}}|, |\bar{\boldsymbol{R}}^{e} - 3\sigma_{\bar{\boldsymbol{R}}^{e}}|), \tag{12}$$

$$\varepsilon = \tau\sigma_{\bar{\boldsymbol{R}}^{e}}\sqrt{\frac{\ln J}{J}}, \tag{13}$$

where $\bar{\boldsymbol{R}}^{e}$ denotes the mean excess return across all assets and periods, $\sigma_{\bar{\boldsymbol{R}}^{e}}$ denotes the corresponding standard deviation, and τ is set equal to three.

To assess the sensitivity of our findings to the choice of hyperparameters, we additionally implement a Grid Search procedure combined with walk-forward cross-validation. Specifically, for each estimation window, we construct a multiplicative grid over the SVR penalty parameter C and the ε-insensitive loss threshold ε, centred on the baseline values implied by Cherkassky and Ma (2004). The grid consists of five multiplicative scaling factors, $\{0.5, 1.0, 1.5, 2.0, 2.5\}$, applied to each hyperparameter, resulting in a search space of $5^2 = 25$ candidate combinations.

Hyperparameter selection is performed using a three-fold walk-forward cross-validation scheme. The estimation window expands sequentially through time, while validation is conducted exclusively on the immediately subsequent out-of-sample period, thereby preventing look-ahead bias. For each candidate pair (C^*, ε^*), the SVR model is estimated using the PCA-based initialization described in Section 2.2, and predictive performance is evaluated on the validation sample using the mean squared error (MSE). The hyperparameter pair that minimizes the average validation MSE across all folds is subsequently used to construct the final IPCA(SVR) estimator and is held fixed throughout the remaining iterations of the optimization problem in (9).

2.4. Multi-factor models

In order to evaluate the benefits of introducing SVR to IPCA, common risk factor models FFC4 and FF5 are used as benchmarks to further contrast the IPCA method. These models are fitted as follows:

$$r^{e}_{i,t} = \alpha_{i,t} + \beta^{1}_{i}RMRF_t + \beta^{2}_{i}SMB^{1}_{t} + \beta^{3}_{i}HML_t + \beta^{4}_{i}WML_t + \epsilon_{i,t}, \tag{14}$$

$$r^{e}_{i,t} = \alpha_{i,t} + \beta^{1}_{i}RMRF_t + \beta^{5}_{i}SMB^{2}_{t} + \beta^{3}_{i}HML_t + \beta^{6}_{i}RMW_t + \beta^{7}_{i}CMA_t + \epsilon_{i,t}. \tag{15}$$

where $RMRF_t$, HML_t, RMW_t, and CMA_t are the Fama and French (1993, 2015) factor mimicking portfolios for market, growth, profitability, and investment style, respectively. WML_t is the momentum proxy (Carhart 1997). SMB_t^1 is the Fama and French (1993) size factor, while SMB_t^2 is its revised version (Fama and French 2015). In line with Cochrane (2009, Chapter 12.3), a rolling 5-year (Fama and MacBeth 1973) (FMB henceforth) procedure is implemented using (14) and (15) to obtain loadings and risk premia at each time-step. In the first stage, least squares betas are estimated for each stock:

$$r_{i,t}^e = \alpha_i + \boldsymbol{\beta}_i' \boldsymbol{f}_t + \epsilon_{i,t}, \quad i = 1, \ldots, N. \tag{16}$$

The second stage consists of estimating the cross-sectional regression

$$\boldsymbol{r}_{t+1}^e = \lambda_t^c + \hat{\boldsymbol{\beta}}' \boldsymbol{\lambda}_t + \boldsymbol{\nu}_{t+1}, \tag{17}$$

where $\mathbb{E}_t[\alpha_{i,t}] = \lambda_t^c$ is the cross-sectional intercept, while $\boldsymbol{\nu}_{t+1}$ is a $N \times 1$ vector collecting the pricing errors across assets. In line with the FMB framework, we estimate λ^c, $\boldsymbol{\lambda}$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_N)'$ as

$$\bar{\lambda}^k = \frac{1}{T}\sum_{t=1}^{T} \hat{\lambda}_t^k, \quad \widehat{\boldsymbol{\lambda}} = \frac{1}{T}\sum_{t=1}^{T} \widehat{\boldsymbol{\lambda}}_t, \quad \text{and} \quad \widehat{\boldsymbol{\nu}} = \frac{1}{T}\sum_{t=1}^{T} \widehat{\boldsymbol{\nu}}_t. \tag{18}$$

To account for the error-in-variables (EIV) problem, we employ the standard error correction proposed by Shanken (1992) for the Fama-MacBeth procedure. Inference is then conducted using either the joint significance test of Cochrane (2009, Chapter 12.3) for portfolio test assets or a diagonal Wald test for panels of individual assets. Both procedures are closely related to the classical Gibbons *et al.* (1989) framework. For portfolio test assets, we compute the following test statistic:

$$W_p = \boldsymbol{\nu}' \hat{\boldsymbol{\Sigma}}_{\nu}^{-1} \boldsymbol{\nu} \tag{19}$$

where $\hat{\boldsymbol{\Sigma}}_{\nu}$ denotes a consistent estimator of the variance-covariance matrix of pricing errors given by $\frac{1}{T^2}\sum_{t=1}^{T}(\widehat{\boldsymbol{\nu}}_t - \widehat{\boldsymbol{\nu}})(\widehat{\boldsymbol{\nu}}_t - \widehat{\boldsymbol{\nu}})'$, which is valid under the assumption that the time series of pricing errors, $\{\widehat{\boldsymbol{\nu}}_t\}_{t=1}^T$, is serially uncorrelated. Under the null hypothesis of zero pricing errors, $W_p \xrightarrow{d} \chi^2_{N-1}$, where $\xrightarrow{d}$ denotes convergence in distribution and χ^2_{N-1} is a chi-squared distribution with $N-1$ degrees of freedom. Note that this limiting distribution holds under the traditional assumption that the factors and the errors are i.i.d. over time and errors are independent of the factors. Additionally, the asymptotic validity of this test relies on the traditional asset-pricing framework in which the number of test assets N remains fixed while the time-series dimension satisfies $T \to \infty$, implying that $N \ll T$ and ensuring consistent estimation of the pricing-error covariance matrix. For panels of individual equities, we employ the Gibbons *et al.* (1989) diagonal representation of the Wald test:

$$W_d = \left(1 + \boldsymbol{\mu}_f' \hat{\boldsymbol{\Sigma}}_f^{-1} \boldsymbol{\mu}_f\right)^{-1} \sum_{i=1}^{N} \left(\frac{\hat{\alpha}_i^2}{\hat{\sigma}_{ii}^2}\right) \tag{20}$$

where $\boldsymbol{\mu}_f$ denotes the vector of factor risk premia, $\hat{\sigma}_{ii}^2$ is the variance of the pricing residual associated with asset i, and $\hat{\boldsymbol{\Sigma}}_f$ is a consistent estimator of the factor variance-covariance matrix. Under the null hypothesis of zero pricing errors, errors are i.i.d. over time homoskedastic and independent of the factors, it follows that $W_d \xrightarrow{d} \chi^2_{N-1}$. The quantities $p(W_p)$ and $p(W_d)$ denote the p-values of the portfolio and diagonal Wald tests, respectively.

2.5. Comparable tests

Consistent with Kelly *et al.* (2019) and Chen *et al.* (2024), we evaluate the alternative pricing models using three performance measures: the ex-post Sharpe ratio, Total R^2, and Predictive R^2. The ex-post Sharpe ratio is computed as follows:

$$SR_f = \sqrt{\boldsymbol{\mu}_f' \boldsymbol{\Sigma}_f^{-1} \boldsymbol{\mu}_f}. \tag{21}$$

Following Kelly *et al.* (2019) and Chen *et al.* (2024), we assess model fit using Total R^2 and Predictive R^2. Although both studies adopt closely related measures, we implement the approach of Kelly *et al.* (2019), modified using a mean-period adjustment to ensure compatibility with the rolling Fama-MacBeth framework. The Total R^2 is defined as

$$\begin{aligned} TR^2 &= 1 - \underbrace{\frac{\sum_{i,t}\left(r_{i,t+1}^e - \left[\hat{\lambda}_t^c + \hat{\boldsymbol{\beta}}'\boldsymbol{\lambda}_t\right]\right)^2}{\sum_{i,t}\left(r_{i,t+1}^e\right)^2}}_{FMB} \\ &= 1 - \underbrace{\frac{\sum_{i,t}\left(r_{i,t+1}^e - \boldsymbol{z}_{i,t}'\left[\hat{\boldsymbol{\Gamma}}_{\alpha} + \hat{\boldsymbol{\Gamma}}_{\beta}\hat{\boldsymbol{\lambda}}_t\right]\right)^2}{\sum_{i,t}\left(r_{i,t+1}^e\right)^2}}_{IPCA} \end{aligned} \tag{22}$$

while the Predictive R^2 is given by

$$\begin{aligned} PR^2 &= 1 - \underbrace{\frac{\sum_{i,t}\left(r_{i,t+1}^e - \left[\bar{\lambda}_t^c + \hat{\boldsymbol{\beta}}'\bar{\boldsymbol{\lambda}}_t\right]\right)^2}{\sum_{i,t}\left(r_{i,t+1}^e\right)^2}}_{FMB} \\ &= 1 - \underbrace{\frac{\sum_{i,t}\left(r_{i,t+1}^e - \boldsymbol{z}_{i,t}'\left[\hat{\boldsymbol{\Gamma}}_{\alpha} + \hat{\boldsymbol{\Gamma}}_{\beta}\bar{\boldsymbol{\lambda}}_t\right]\right)^2}{\sum_{i,t}\left(r_{i,t+1}^e\right)^2}}_{IPCA} \end{aligned} \tag{23}$$

Finally, for individual securities, both measures are reported separately for the in-sample (IS) and out-of-sample (OOS) periods.

3. Data

3.1. In-sample

In order to conduct the analysis, monthly equity returns, accounting and performance characteristics, and indicators are retrieved from CSRP/Compustat databases, at Wharton Research Data Services (2021), as well as from French (2021). FFC4 and FF5 factor proxies are collected from French (2021) library, with their descriptive statistics presented in table A3. Based on the data, only expected values of RMRF, RMW, and WML are statistically different from zero at 1%, 5%, and 10%, respectively. Furthermore, only returns of HML are normally distributed and it also has the highest correlation among factors (not including SMB(1) with SMB(2)) with CMA. It is also worth noting that WML has a downturn of -34.4%, which is the largest among all factors. For returns and characteristics data, the WRDS SAS Studio code generating 406 characteristics in 93 countries, provided by Jensen *et al.* (2023), is utilized and restricted to U.S. Equities with 46 non-missing firm characteristics. These firm-level variables are listed either on French (2021) or applied in Freyberger *et al.* (2020), with their descriptions depicted in table A1. Overall, most characteristics are in line with those used by Chen *et al.* (2024).

The working sample covers a 30-year and 7-month period from the beginning of June 1990 until the end of December 2020. Since characteristic values are of various magnitude, which may complicate optimization, as in Kelly *et al.* (2019), they are cross-sectionally transformed. Period-by-period, using stocks' ranks for each characteristic, values are divided by the number of observations, minus 0.5, to map them on a $[-0.5, 0.5]$ interval. Year-per-year, the sample is smaller than that of Kelly *et al.* (2019), who only utilize 36 characteristics, whilst also replacing missing values with their cross-sectional median. As stated by Chen *et al.* (2024), this approach can introduce additional assumption error and ignore the dependency structure of characteristics. Consequently, the current sample is best compared to that of Chen *et al.* (2024). The comparison is provided in figure A2, with the current sample, although smaller, closely resembling that of Chen *et al.* (2024). Overall, the dataset is ample to form portfolios.

In addition to individual equities, the methods are compared based on both characteristic-sorted and industry-sorted portfolios.The purpose of including these lower-dimensional portfolio datasets is not to maximize the cross-sectional dimension available for factor estimation. Rather, they serve as standard benchmark test assets in the asset-pricing literature, facilitating comparisons with previous studies and allowing us to assess the robustness of our findings across alternative testing environments. The portfolios are constructed to mirror those available in the French (2021) library and comprise the equal-weighted 25 Size and Book-to-Market, 25 Size and Momentum, and value-weighted 17 Industry portfolios. Using the documentation provided by French (2021), along with size, book-to-market, momentum breakpoints and SIC industry codes as well as GVKEY, PERMNO, and SIC links from Compustat, the portfolios are formed at the end of June 1990 and rebalanced annually. Analogically to equities, these portfolios yield characteristics, which are computed period-by-period as the cross-sectional median of their constituting assets' characteristics. The high similarity of these portfolios with the originals (average correlation above 94%) is presented in the Appendix in table A2, while tables A4, A5, and A6, pages 75, 79, and 83 respectively, depict their descriptive statistics.

Overall, the final datasets accommodate 30 years of return and characteristic entries for stocks and managed portfolios, starting from January 1991 up to and including December 2020. For stocks, this results in a sample of 813,695 observations and 9,689 unique equities. Furthermore, the data for both portfolios and individual stocks is also divided into 5 sub-periods of 5 years, to examine pricing consistency.

3.2. Out-of-sample

The second asset pricing scenario evaluates whether risk premia estimated from one cross-section of equities can successfully price a different cross-section of equities out of sample. Specifically, within each rolling 60-month estimation window, the cross-section of stocks is divided into 80% in-sample (IS) and 20% out-of-sample (OOS) securities. The IS securities are used to estimate factor betas (first pass) and risk premia (second pass) for the FFC4 and FF5 models, as well as the characteristic sensitivity matrix $\bar{\mathbf{\Gamma}}$ for the IPCA and IPCA(SVR) models. The estimated parameters are then applied to the OOS securities.

The OOS securities are re-selected each month as the rolling estimation window advances and firm characteristics evolve over time. To avoid concentrating the OOS sample within a particular size category, the 20% holdout sample is drawn proportionally from Small, Medium, and Large capitalization groups defined using the 30th and 70th percentile market-capitalization breakpoints, following Fama and French (1993). Within each size group, securities are partitioned into approximately equal sections and the stock whose market capitalization is closest to the section median is selected, thereby ensuring broad cross-sectional representation.

Importantly, this exercise is designed to evaluate the cross-sectional generalization ability of the competing asset-pricing models rather than their time-series forecasting performance. In other words, it assesses whether risk premia estimated on one set of securities can accurately price previously unseen securities observed over the same time period.

The resulting OOS sample contains 2302 unique equities and 37 826 stock-month observations, which is approximately 4.3 times smaller than the nominal 20% split because of the data requirements imposed by the Fama-MacBeth estimation procedure. If a model generalizes well across securities, the in-sample and out-of-sample Total R^2 and Predictive R^2 measures should be similar. We therefore employ Welch's t-test (Welch 1947) to evaluate the statistical significance of the differences between the IS and OOS goodness-of-fit measures.

Table 1. 25 size and book-to-market portfolios.

		1996–2000	2001–2005	2006–2010	2011–2015	2016–2020	1996–2020
FFC4	TR^2	0.888	0.922	0.955	0.932	0.948	0.931
	PR^2	0.016	0.038	0.015	0.032	0.027	0.023
	$W_p - stat.$	1.539	1.300	0.589	0.602	0.582	0.202
	$p(W_p)$	1.000	1.000	1.000	1.000	1.000	1.000
FF5	TR^2	0.896	0.931	0.957	0.938	0.949	0.936
	PR^2	0.017	0.039	0.012	0.034	0.031	0.023
	$W_p - stat.$	1.104	1.560	0.530	0.998	0.686	0.134
	$p(W_p)$	1.000	1.000	1.000	1.000	1.000	1.000
IPCA	TR^2	0.947	0.967	0.977	0.959	0.969	0.926
	PR^2	− 0.023	0.039	0.002	0.027	− 0.380	0.003
	$p(W_\alpha)$	1.000	1.000	1.000	1.000	1.000	1.000
IPCA (SVR)	TR^2	0.847	0.880	0.914	0.845	0.881	0.896
	PR^2	0.024	0.051	0.004	0.033	0.040	0.022
	$p(W_\alpha)$	0.000	0.000	0.260	0.000	0.010	0.020
IPCA (SVR)-GS	TR^2	0.887	0.883	0.929	0.858	0.901	0.886
	PR^2	0.024	0.046	0.006	0.033	0.039	0.024
	$p(W_\alpha)$	0.000	0.000	0.104	0.000	0.000	0.021

Notes: The table reports average cross-sectional risk premia and performance metrics of 25 Size and Book-to-Market portfolios for FFC4, FF5, IPCA, IPCA(SVR) and IPCA(SVR)-GS models in 6 periods. TR^2 and PR^2 are total and predictive in-sample goodness of fit measures. $p(W_p)$ and $p(W_a)$ are the pricing error test' p-values.

Table 2. 25 size and momentum portfolios.

		1996–2000	2001–2005	2006–2010	2011–2015	2016–2020	1996–2020
FFC4	TR^2	0.714	0.761	0.851	0.830	0.834	0.799
	PR^2	0.047	0.039	0.029	0.057	0.067	0.038
	$W_p - stat.$	0.684	0.088	0.054	0.396	0.757	0.312
	$p(W_p)$	1.000	1.000	1.000	1.000	1.000	1.000
FF5	TR^2	0.775	0.773	0.858	0.846	0.833	0.818
	PR^2	0.052	0.039	0.030	0.057	0.068	0.040
	$W_p - stat.$	0.666	0.124	0.410	0.344	0.539	0.097
	$p(W_p)$	1.000	1.000	1.000	1.000	1.000	1.000
IPCA	TR^2	0.958	0.965	0.979	0.973	0.976	0.900
	PR^2	− 0.023	− 0.248	− 0.051	− 0.226	− 0.432	0.023
	$p(W_\alpha)$	1.000	1.000	1.000	1.000	1.000	1.000
IPCA (SVR)	TR^2	0.816	0.895	0.913	0.840	0.836	0.885
	PR^2	0.023	0.047	0.011	0.026	0.045	0.024
	$p(W_\alpha)$	0.000	0.120	0.000	0.000	0.020	0.000
IPCA (SVR)-GS	TR^2	0.842	0.874	0.929	0.832	0.889	0.873
	PR^2	0.020	0.041	0.014	0.026	0.038	0.023
	$p(W_\alpha)$	0.000	0.000	0.000	0.000	0.001	0.000

Notes: The table reports average cross-sectional risk premia and performance metrics of 25 size and momentum portfolios for FFC4, FF5, IPCA, IPCA(SVR) and IPCA(SVR)-GS models in 6 periods. TR^2 and PR^2 are total and predictive in-sample goodness of fit measures. $p(W_p)$ and $p(W_a)$ are the pricing error test' p-values).

4. Empirical results

4.1. A comparison of pricing performance

The sample is divided into five subperiods corresponding to several economically significant events. Specifically, 1996–2000 covers the Dot-com bubble, 2001–2005 corresponds to the pre-crisis expansion period, 2006–2010 spans the Global Financial Crisis, 2011–2015 captures the post-crisis recovery, and 2016–2020 includes both the US-China trade conflict and the COVID-19 market crash. Chen *et al.* (2024) argue that asset-pricing model performance may vary substantially across economic cycles. Accordingly, the subperiod analysis is used to assess the consistency and robustness of the competing pricing models across different market environments.

Table 3. 17 Fama-French industry portfolios.

		1996–2000	2001–2005	2006–2010	2011–2015	2016–2020	1996–2020
FFC4	TR^2	0.872	0.929	0.942	0.932	0.946	0.927
	PR^2	0.016	0.046	0.011	0.030	0.033	0.023
	$W_p - stat.$	0.566	0.654	0.826	1.189	0.321	0.129
	$p(W_p)$	1.000	1.000	1.000	1.000	1.000	1.000
FF5	TR^2	0.876	0.926	0.945	0.939	0.948	0.930
	PR^2	0.014	0.044	0.018	0.038	0.035	0.024
	$W_p - stat.$	0.582	0.494	1.360	1.418	0.267	0.094
	$p(W_p)$	1.000	1.000	1.000	1.000	1.000	1.000
IPCA	TR^2	0.938	0.970	0.977	0.957	0.962	0.918
	PR^2	− 0.325	0.029	0.007	0.021	0.005	0.021
	$p(W_\alpha)$	1.000	1.000	1.000	1.000	1.000	1.000
IPCA (SVR)	TR^2	0.723	0.747	0.840	0.776	0.800	0.845
	PR^2	0.087	0.058	0.040	0.075	0.077	0.043
	$p(W_\alpha)$	0.000	0.000	0.030	0.000	0.000	0.000
IPCA (SVR)-GS	TR^2	0.545	0.665	0.782	0.691	0.854	0.805
	PR^2	0.020	0.041	0.014	0.026	0.038	0.023
	$p(W_\alpha)$	0.000	0.000	0.000	0.000	0.001	0.000

Notes: The table reports average cross-sectional risk premia and performance metrics of 25 size and momentum portfolios for FFC4, FF5, IPCA, IPCA(SVR), and IPCA(SVR)-GS models in 6 periods. TR^2 and PR^2 are total and predictive in-sample goodness of fit measures. $p(W_p)$ and $p(W_a)$ are the pricing error test' p-values.

Tables 1 and 2 report the pricing performance of the 25 Size and Book-to-Market (25SBM) and 25 Size and Momentum (25SM) portfolios, respectively. We compare the Carhart (1997) four-factor model (FFC4), the Fama-French (2015) five-factor model (FF5), the standard IPCA framework of Kelly *et al.* (2019), and two regularized IPCA specifications proposed in this paper. The first, denoted IPCA(SVR), employs the data-driven hyperparameter selection procedure of Cherkassky and Ma (2004), while the second, denoted IPCA(SVR)-GS, selects the SVR hyperparameters C and ε using a Grid Search cross-validation procedure.

It is important to acknowledge a potential limitation of applying the IPCA and IPCA(SVR) models to datasets with a relatively small cross-sectional dimension (N). The asymptotic results in Kelly *et al.* (2020) rely on the assumption that $N, T \to \infty$. However, the estimator of the characteristic sensitivity matrix $\bar{\mathbf{\Gamma}}$ converges at rate $\sqrt{NT}$, which is faster that the $\sqrt{N}$ rate obtained for leading estimates in standard PCA analysis. In contrast, the estimated latent factors converge at rate $\sqrt{N}$, as in standard PCA. Therefore, a sufficiently long time series can partially compensate for a smaller cross-section in the estimation procedure. Nevertheless, the evidence obtained from the portfolio datasets should be interpreted with caution, as a relatively small N may result in greater estimation uncertainty for the latent factors.

Because the 25SBM and 25SM portfolios produce qualitatively similar results, we focus primarily on the former. Standard IPCA delivers the strongest in-sample fit, with TR^2 exceeding 94.7% in every subperiod, but its predictive performance is generally weaker and, in some cases, negative. In contrast, the regularized specifications, IPCA(SVR) and IPCA(SVR)-GS, achieve superior predictive performance and produce remarkably similar results, indicating that the findings are robust to the choice of hyperparameter selection procedure. These results are consistent with the argument of Daniel (2018) that characteristic-sorted portfolios may mechanically favour models constructed using the same characteristics. While standard IPCA appears to overfit the cross section, regularization improves predictive performance by mitigating this tendency. Nevertheless, because factor extraction relies on a relatively small number of test assets, greater uncertainty in the estimated latent factors and risk premia may reduce the power of the pricing-error tests, potentially explaining why the regularized specifications are frequently rejected despite their stronger predictive performance.

It is also noteworthy that all models achieve their highest TR^2 values during the Global Financial Crisis, potentially reflecting the greater prominence of common risk factors during periods of heightened market volatility. Despite its superior predictive performance, IPCA(SVR) and IPCA(SVR)-GS are the only models that fail the pricing-error test in the subperiods in which they outperform the competing specifications. In the remaining subperiods, both models satisfy the test at the 1% significance level. Furthermore, the regularized specifications are the only models whose in-sample goodness-of-fit falls below 90% in most subperiods. A similar pattern emerges for the 25SM portfolios, where IPCA(SVR) and IPCA(SVR)-GS generally outperform the competing models in terms of PR^2 but are rejected by the pricing-error test in the majority of cases.

Table 3 reports the results for the 17 industry-sorted portfolios (17IP). The findings broadly mirror those obtained for the 25SBM and 25SM portfolios. Standard IPCA again delivers the strongest in-sample fit, while the regularized IPCA

Table 4. Cross-section of excess stock returns.

		1996–2000	2001–2005	2006–2010	2011–2015	2016–2020	1996–2020
FFC4	TR^2	0.266	0.306	0.305	0.209	0.238	0.256
	OTR^2	0.196	0.258	0.257	0.176	0.285	0.244
	ΔTR^2	−4.065***	−2.081**	−0.622	−0.590	−0.232	−3.751***
	$p(\Delta TR^2)$	0.000	0.040	0.535	0.556	0.817	0.000
	PR^2	0.004	0.010	0.005	0.008	0.007	0.006
	OPR^2	0.008	0.008	0.006	0.006	0.008	0.006
	ΔPR^2	0.608	−0.657	−0.076	0.043	−0.108	−0.127
	$p(\Delta PR^2)$	0.544	0.512	0.939	0.966	0.914	0.899
	$W_d - stat.$	173.448	1201.401	11470.002	22752.377	24777.429	12074.931
	$p(W_d)$	1.000	1.000	0.000	0.000	0.000	0.000
FF5	TR^2	0.279	0.309	0.307	0.212	0.239	0.258
	OTR^2	0.176	0.246	0.255	0.174	0.276	0.239
	ΔTR^2	−4.607***	−2.555**	−0.815	−0.792	−0.499	−4.541***
	$p(\Delta TR^2)$	0.000	0.012	0.417	0.430	0.619	0.000
	PR^2	0.003	0.012	0.006	0.008	0.008	0.006
	OPR^2	0.009	0.009	0.008	0.007	0.009	0.006
	ΔPR^2	0.563	−0.591	−0.072	0.043	−0.082	−0.095
	$p(\Delta PR^2)$	0.574	0.556	0.942	0.966	0.935	0.924
	$W_d - stat.$	218.627	1459.805	12460.169	27055.512	29287.877	14096.398
	$p(W_d)$	1.000	1.000	0.000	0.000	0.000	0.000
IPCA	TR^2	0.154	0.183	0.185	0.120	0.130	0.152
	OTR^2	0.173	0.249	0.227	0.144	0.163	0.186
	ΔTR^2	−1.070	1.252	1.608	0.959	0.424	1.586
	$p(\Delta TR^2)$	0.288	0.214	0.111	0.340	0.672	0.113
	PR^2	0.006	0.013	0.004	0.007	0.004	0.005
	OPR^2	0.007	0.006	0.007	0.008	0.006	0.004
	ΔPR^2	−0.321	−1.088	−0.585	0.871	−0.140	−0.908
	$p(\Delta PR^2)$	0.750	0.279	0.560	0.387	0.889	0.365
	$p(W_\alpha)$	0.050	0.010	0.050	0.570	0.290	0.000
IPCA (SVR)	TR^2	0.141	0.173	0.171	0.110	0.107	0.143
	OTR^2	0.196	0.258	0.243	0.148	0.195	0.200
	ΔTR^2	0.416	2.049**	2.476**	1.908*	2.586***	3.668***
	$p(\Delta TR^2)$	0.679	0.044	0.015	0.059	0.011	0.000
	PR^2	0.004	0.011	0.002	0.005	0.001	0.004
	OPR^2	0.005	0.008	0.006	0.007	0.007	0.005
	ΔPR^2	0.337	−1.027	−0.559	0.971	0.270	−0.345
	$p(\Delta PR^2)$	0.737	0.307	0.577	0.335	0.788	0.730
	$p(W_\alpha)$	0.000	0.000	0.000	0.000	0.000	0.000
IPCA (SVR)-GS	TR^2	0.141	0.172	0.172	0.110	0.108	0.142
	OTR^2	0.198	0.257	0.242	0.148	0.198	0.199
	ΔTR^2	0.489	2.036**	2.443***	1.991**	2.611***	3.661***
	$p(\Delta TR^2)$	0.626	0.045	0.016	0.049	0.010	0.000
	PR^2	0.003	0.011	0.003	0.005	0.001	0.004
	OPR^2	0.007	0.007	0.007	0.007	0.007	0.004
	ΔPR^2	0.562	−1.179	−0.580	0.996	0.318	−0.418
	$p(\Delta PR^2)$	0.576	0.241	0.563	0.322	0.751	0.676
	$p(W_\alpha)$	0.000	0.000	0.000	0.000	0.000	0.000

Notes: The table reports average cross-sectional risk premia and performance metrics of individual equities using FFC4, FF5, IPCA, IPCA(SVR) and IPCA(SVR)-GS models in 6 periods. ΔTR^2 and ΔPR^2 are the difference t-tests between out-of-sample and in-sample fit, with $p(\Delta TR^2)$ and $p(\Delta PR^2)$ being their p-values. $p(W_d)$ and $p(W_\alpha)$ are the pricing error test' p-values. t-statistics are in parenthesis. *, **, *** represent values that are statistically significant at 10%, 5%, and 1% significance level, respectively.

Table 5. Ex-post sharpe ratios of IPCAs.

	Period	IPCA	IPCA(SVR)	IPCA(SVR)-GS	FFC4	FF5
25SBM	1996–2000	1.349	1.697	1.057	–	–
	2001–2005	1.223	3.753	2.869	–	–
	2006–2010	0.739	1.281	1.233	–	–
	2011–2015	1.434	2.063	2.632	–	–
	2016–2020	1.105	1.424	1.906	–	–
	1996–2020	0.696	1.336	1.107	–	–
25SM	1996–2000	1.420	1.446	1.360	–	–
	2001–2005	1.204	1.974	3.550	–	–
	2006–2010	1.110	2.022	1.065	–	–
	2011–2015	0.802	1.582	1.428	–	–
	2016–2020	0.958	3.974	3.743	–	–
	1996–2020	0.795	0.993	0.970	–	–
17IP	1996–2000	1.654	2.145	3.200	–	–
	2001–2005	1.470	1.307	1.266	–	–
	2006–2010	2.171	1.241	1.710	–	–
	2011–2015	1.016	3.222	3.544	–	–
	2016–2020	1.271	1.886	2.789	–	–
	1996–2020	1.217	1.058	1.507	–	–
CSS	1996–2000	2.635	1.601	1.723	2.055	2.439
	2001–2005	1.108	2.525	2.440	1.876	1.716
	2006–2010	1.033	1.678	1.569	1.448	1.681
	2011–2015	2.780	2.697	2.529	0.656	1.674
	2016–2020	1.776	2.672	2.797	1.582	1.363
	1996–2020	1.600	1.425	1.334	0.862	1.199

Notes: The table presents annualized Sharpe ratios of IPCA, IPCA(SVR), and IPCA(SVR)-GS specifications, applied to 25 Size and Book-to-Market (25SBM), 25 Size and Momentum (25SM), 17 Industry (17IP) portfolios, and the cross-section of stocks (CSS), alongside the annualized Sharpe ratios for the FFC4 and FF5 models computed from the CSS dataset.

Table 6. IPCA(SVR) principal component correlation with common risk factors.

		IPCA(SVR) Factor					
		K1	K2	K3	K4	K5	K6
Factors	RMRF	0.90	0.02	0.07	−0.20	−0.01	−0.07
	SMB^1	0.54	0.21	−0.25	0.11	−0.29	0.17
	SMB^2	0.56	0.14	−0.18	0.11	−0.19	0.19
	HML	−0.01	−0.39	0.26	0.01	0.50	0.10
	RMW	−0.47	−0.33	0.33	−0.15	0.35	0.10
	CMA	−0.23	−0.13	0.17	0.15	0.39	0.08
	WML	−0.42	0.31	0.15	−0.27	−0.45	0.36

Notes: The table reports the correlation between the six factors ($K1, K2, \ldots, K6$) estimated using the IPCA(SVR) model and the factor-mimicking portfolios of Fama and French (1993, 2015). Specifically, *RMRF*, *HML*, *RMW*, *CMA*, *WML*, SMB^1, and SMB^2 denote the market, value, profitability, investment, momentum, original size, and revised size factors, respectively.

specifications produce superior predictive performance. Relative to standard IPCA, the predictive fit of IPCA(SVR) and IPCA(SVR)-GS improves by as much as seven percentage points over the full sample. Traditional multifactor models exhibit the most stable combination of TR^2 and PR^2 across subperiods and remain competitive overall. Nevertheless, as in the previous portfolio exercises, the regularized specifications are frequently rejected by the pricing-error tests despite their stronger predictive performance.

The next exercise applies the regularized IPCA specifications to the cross-section of stocks (CSS). An important advantage of this setting is that it more closely satisfies the asymptotic requirements underlying the IPCA framework, namely a large cross-sectional dimension (N). Consequently, the latent factors and associated risk premia can be estimated more precisely than in the portfolio-based applications.

Table 4 shows that either the IPCA(SVR) or IPCA(SVR)-GS specification achieves the highest out-of-sample Total R^2 and Predictive R^2 in every evaluation period. Relative to standard IPCA, the regularized specifications improve overall OOS Total R^2 by approximately 20% and OOS Predictive R^2 by approximately 0.5%. However, their performance remains broadly comparable to that of the FFC4 and FF5 benchmark models. More importantly, IPCA(SVR) and IPCA(SVR)-GS are the only specifications for which the difference between IS and OOS goodness-of-fit measures is positive and statistically significant at the 5% level across most sampling periods.

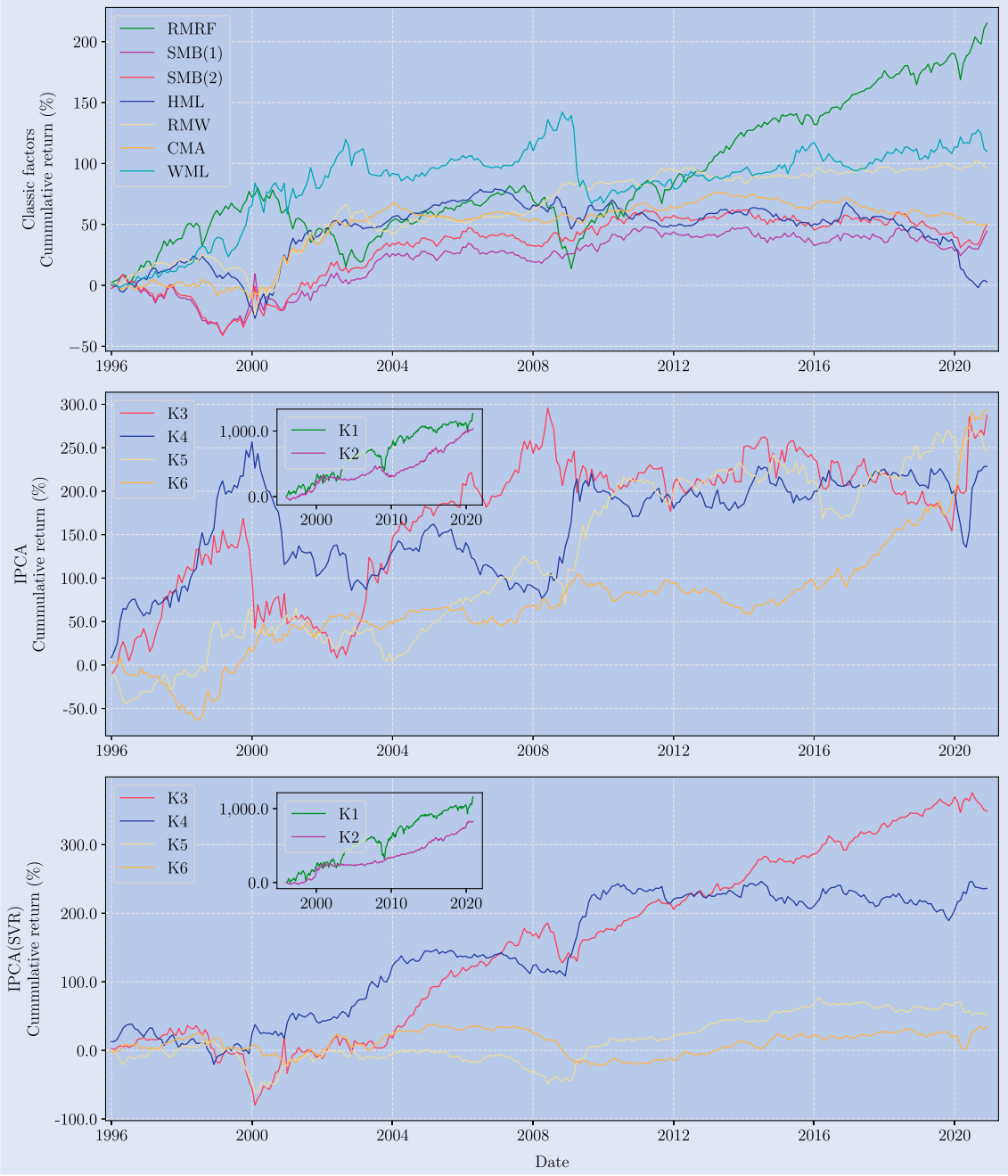


Figure 1. Cummulative returns of factors.
Notes: The figure demonstrates classic factor portfolios' and principal components' (IPCA factors) cumulative returns. RMRF is the excess value-weighted market return. SMB(1) and HML are Fama and French (1993) size and growth proxies, while SMB(2) is the revised size proxy (Fama and French 2015). RMW and CMA are Fama and French (2015) profitability and investment factors. WML is the Carhart (1997) momentum factor. $K(n)$ is the n^{th} principal component. IPCA and IPCA(SVR) plots include a separate sub-plot for the principal components, which performance exceeds other factors by more than 5 times, leading to disproportional illustration.

This result suggests that regularization produces a more parsimonious characteristic sensitivity matrix, leading to improved out-of-sample performance and indicating that the standard IPCA specification may overfit the cross-sectional structure of returns.

In contrast to the managed-portfolio results, FF5 delivers the strongest overall TR^2 and PR^2 performance in the stock-level analysis. Furthermore, FFC4, FF5, IPCA(SVR), and IPCA(SVR)-GS fail the pricing-error tests in all periods, whereas standard IPCA passes five out of six tests at the 1% significance level. Importantly, the failure of the pricing-error tests does not appear to be driven by the choice of SVR hyperparameters. Despite employing a formal Grid Search and walk-forward cross-validation procedure, the IPCA(SVR)-GS specification produces pricing-error test results that are qualitatively identical to those of the baseline IPCA(SVR) model. Taken together, these findings suggest that regularization primarily improves the predictive performance of the IPCA framework rather than its ability to satisfy pricing restrictions. More broadly, the stronger performance of the IPCA-class models in the stock-level analysis supports the argument that these methods are better suited to environments with a large

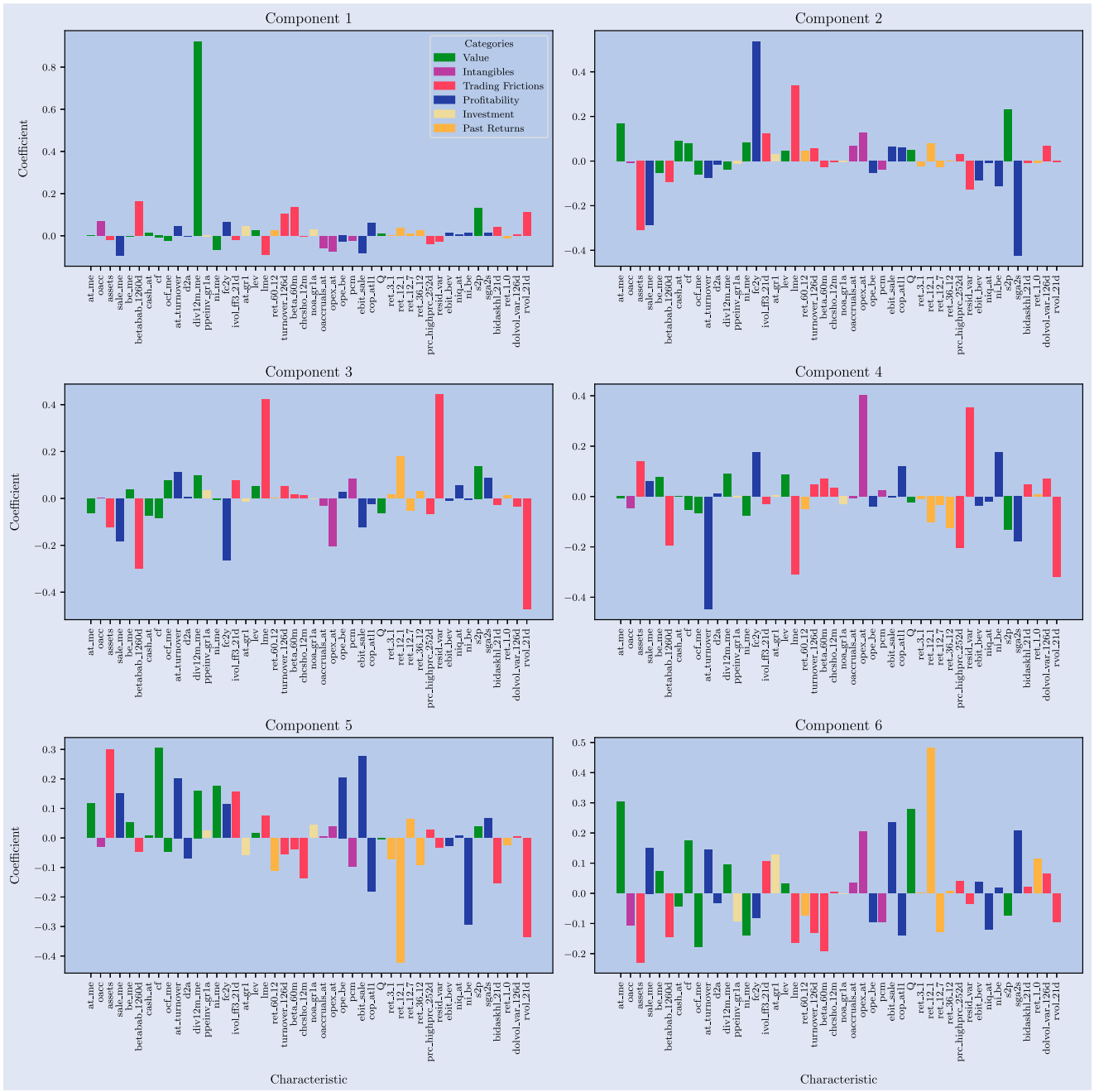


Figure 2. Coefficients of IPCA(SVR) chracteristics.
Notes: The figure depicts the IPCA(SVR) $\mathbf{\Gamma}_\beta$ matrix composition for each of the 6 principal components. Characteristics are groupped into 6 categories similarly to those of Chen *et al.* (2024): Value, Intangibles, Trading Frictions, Profitability, Investment, and Past Returns.

cross-sectional dimension, where latent factors can be estimated more precisely. Consistent with Kelly *et al.* (2019), the results therefore indicate that the IPCA framework is generally more robust when applied to individual-stock pricing than to small sets of test portfolios.

4.2. Factors and alphas

Table 5 presents the ex-post Sharpe ratios (SRs) for the competing specifications. Overall, the regularized models, particularly IPCA(SVR) and IPCA(SVR)-GS, frequently generate higher factor SRs than standard IPCA and remain broadly competitive with the benchmark multifactor models. However, their advantage is not uniform across all test assets and subperiods, with the relative ranking of the models varying over time. These findings suggest that regularization can improve the efficiency of the extracted latent factors, although no single specification consistently dominates across all settings. For the CSS dataset, the IPCA-class models generally produce higher SRs than FFC4 and remain competitive with FF5, further supporting the view that the IPCA framework performs best in environments with a large cross-sectional dimension.

As the latent factors and economic interpretations obtained from the IPCA(SVR)-GS specification are qualitatively similar to those of IPCA(SVR), the remainder of this section

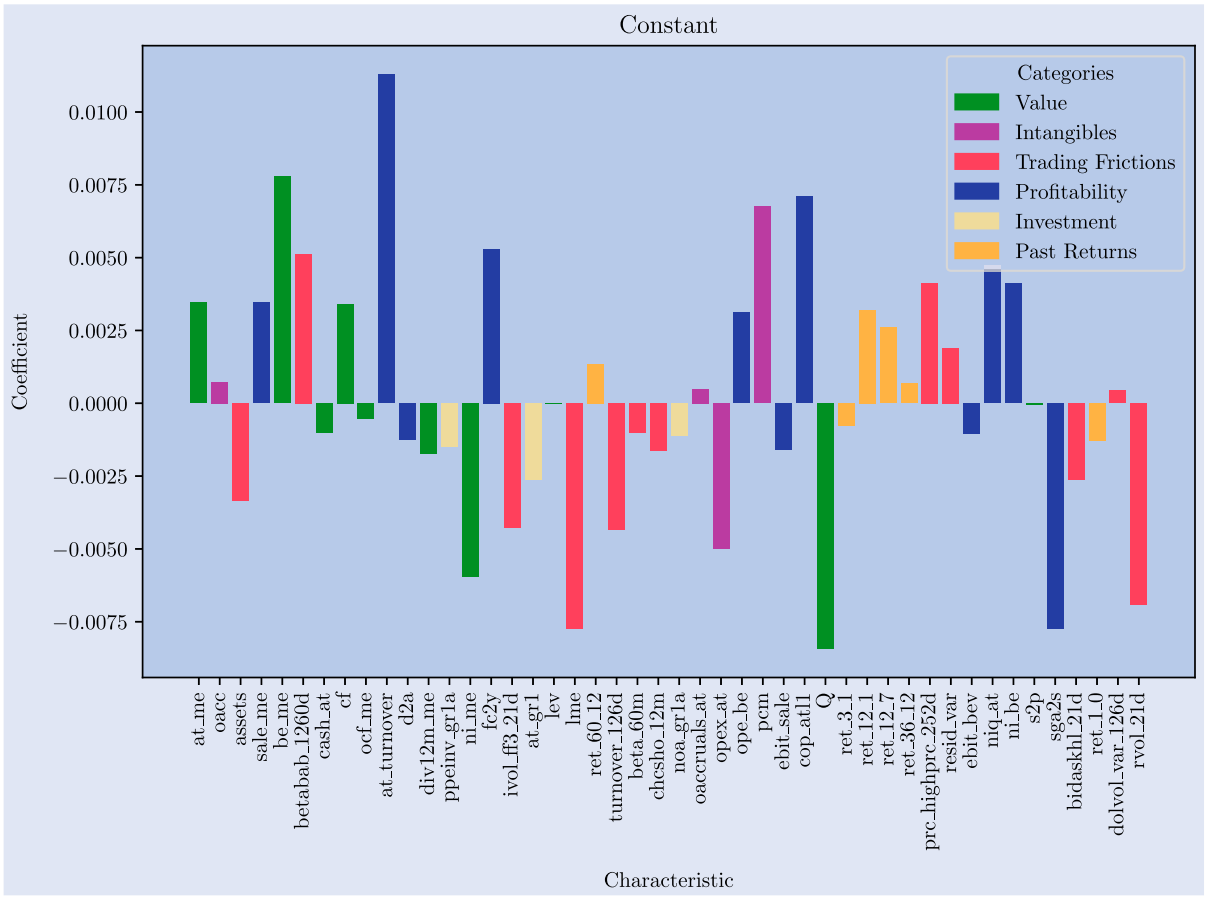


Figure 3. Coefficients of IPCA(SVR) alpha.
Notes: The figure depicts IPCA(SVR) $\mathbf{\Gamma}_\alpha$ coefficients. Characteristics are groupped into 6 categories similarly to those of Chen *et al.* (2024): Value, Intangibles, Trading Frictions, Profitability, Investment, and Past Returns.

focuses on the latter. The first three components account for most of the economically interpretable variation in the IPCA(SVR) factor structure. As shown in figure 1 and table 6, the first factor is strongly related to the market factor, while the second and third factors load predominantly on characteristics associated with profitability, growth, and trading frictions. The remaining factors display weaker links with conventional factor proxies and appear to capture more specialized dimensions of cross-sectional variation. Taken together, the results suggest that IPCA(SVR) largely recovers familiar systematic risk dimensions while simultaneously identifying additional sources of return variation that are not fully captured by traditional factor models.

Considering the factor loadings shown in figure 2, the first component places substantial weight on firms with high dividend-to-market equity ratios (`div12m_me`) and high five-year betas (`betabab_1260d`), effectively loading on both value and systematic risk. This combination gives rise to a leveraged market factor (K1), as illustrated in figure 1. Overall, this component can be interpreted as a value-oriented factor with a strong market-risk dimension.

The second factor loads positively on fixed costs relative to sales (`fc2y`) and market equity (`lme`), while loading negatively on total assets (`assets`), book-to-market equity (`be_me`), operating expenses relative to sales (`sga2s`), and sales-to-price (`s2p`). Since market equity is a more timely measure of firm size than total assets, this component appears to identify firms experiencing rapid market-value growth relative to their accounting size, thereby resembling a growth-related factor. Although its construction is partly linked to size characteristics, it exhibits only a weak negative correlation with HML. Consequently, this component is more closely associated with firms' profitability and growth prospects than with traditional value effects.

The remaining components combine characteristics related to trading frictions, profitability, and return persistence. Although they exhibit weaker correlations with traditional factor proxies, they contribute to capturing additional dimensions of cross-sectional return variation beyond those represented by standard benchmark models.

ptThe $\mathbf{\Gamma}_\alpha$ component shown in figure 3 exhibits coefficients that are, on average, more than ten times smaller in magnitude than those of $\mathbf{\Gamma}_\beta$. Nevertheless, the constant is a mix of multiple characteristics. Most notably, it overweights asset turnover (`at_turnover`) and under-weights Tobin's Q (`Q`) to identify undervalued equities and adjust their return upwards. Hence, the model sometimes underprices equities that have higher asset turnover and low market value. This is further affirmed by the overweighted book-to-market equity (`be_me`) and under-weighted (`lme`), suggesting that important size characteristic(s) are missing to fully explain the mispricing in relation between market and intrinsic value.

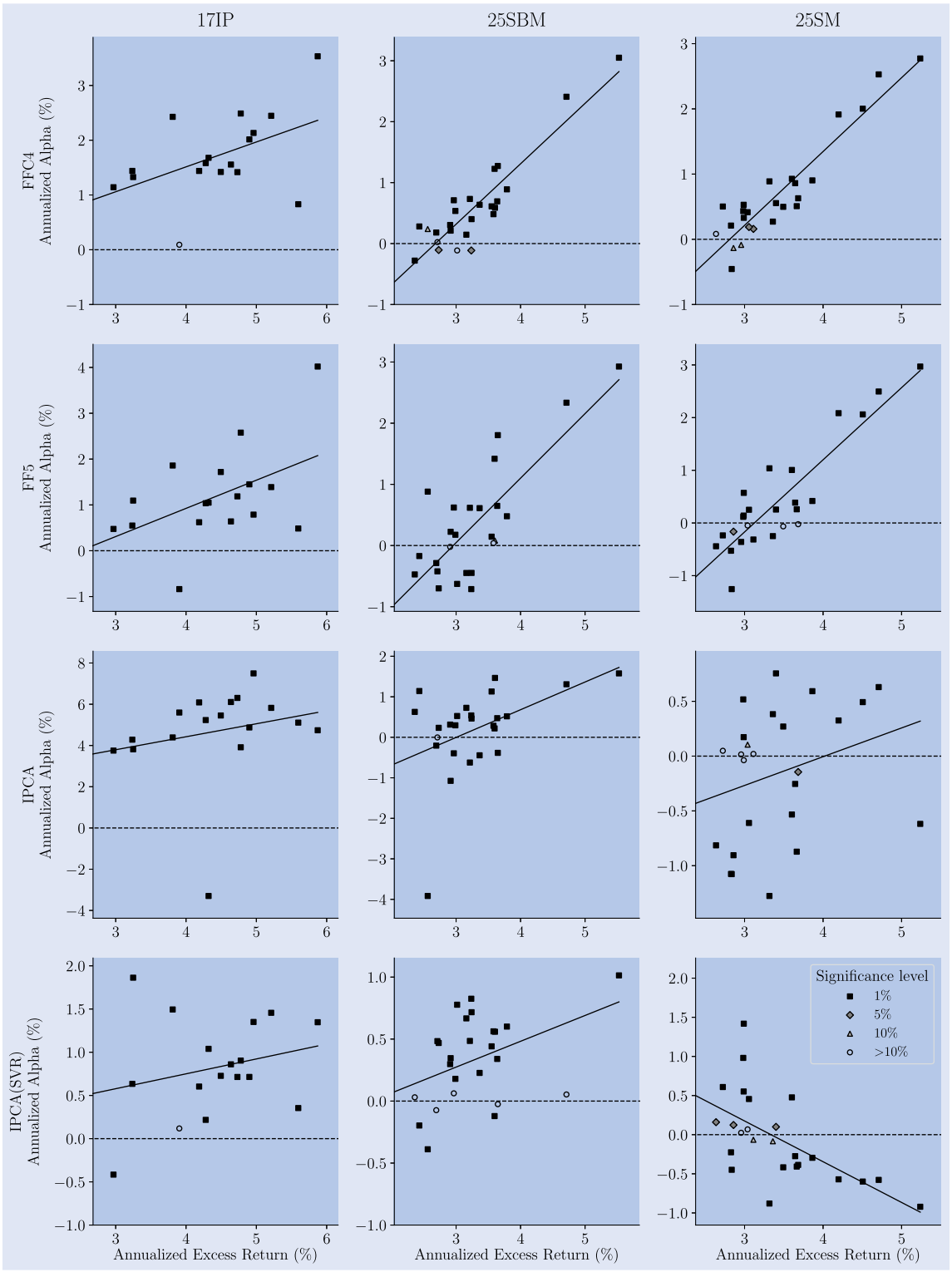


Figure 4. Alphas of priced portfolios.
Notes: The figure illustrates full sample unconditional annualized alphas using traditional models and IPCA for 25 Size and Book-to-Market (25SBM), 25 Size and Momentum (25SM), and 17 Industry Portfolios (17IP). The black line represents the linear relationship between annualied alphas and excess returns. Black squares, grey diamonds, light grey triangles, and white circles represent values that are statistically significant at 1%, 5%, 10% and less than 10% significance level, respectively.

From the perspective of explaining the variability of returns of anomaly portfolios, figure 4 depicts the generated time-series alphas for 25SBM, 25SM and 17IP portfolios, where, all methods produce mostly significant (at 1% level) alphas. Contrary to Kelly *et al.* (2019) this suggests that IPCA estimation produces factors that are not necessarily mean-variance efficient. For 17IP, 25SBM, and 25SM portfolios, IPCA(SVR) tends to produce smaller annualized alphas compared to other models, with some being insignificant, however, they are inversely associated with the size of the excess

returns for the 25SM portfolios. Based on 25SBM and 25SM portfolios, traditional models produce larger alphas compared to ML methods (with 25SBM being an exception for the classic IPCA), with a fraction yielding weak and non-significant values. Overall, the results indicate that IPCA is not able to fully explain the fluctuations of anomaly portfolios.

5. Conclusion

This study shows that machine-learning enhancements to the IPCA specification proposed by Kelly *et al.* (2019) can improve the modelling of the cross-section of equity returns. Using a set of 46 firm characteristics and 9689 equities, we find that while the original IPCA specification exhibits strong in-sample performance, incorporating support vector regression as a form of regularization generally produces more stable latent factors and superior out-of-sample results, suggesting that regularization helps mitigate overfitting and better identifies persistent return-characteristic relationships. An additional contribution of this study is the examination of the sensitivity of the results to the choice of SVR hyperparameters. Using a Grid Search procedure combined with walk-forward cross-validation, we find that the resulting estimates, pricing-performance measures, and out-of-sample results remain qualitatively similar to those obtained under the baseline calibration procedure. This evidence suggests that the main conclusions of the paper are not driven by a particular choice of hyperparameters and reinforces the view that the benefits of regularization within the IPCA framework are robust across alternative calibration approaches.

The findings also highlight the distinction between predictive performance and asset-pricing validity. Although IPCA(SVR) generates economically meaningful factors with strong risk-adjusted performance, significant pricing errors remain, indicating that improved prediction does not necessarily imply a complete description of expected returns.

From an economic perspective, the latent factors are primarily associated with value, profitability, and trading-friction characteristics. Rather than uncovering entirely new sources of risk, the machine-learning specifications appear to combine established firm characteristics into more efficient representations of the underlying determinants of returns. The results further suggest that only a relatively small subset of characteristics contributes materially to model performance, emphasizing the importance of regularization in high-dimensional asset-pricing applications.

Overall, the findings support the use of machine-learning techniques as a complement rather than a substitute for economic asset-pricing theory. Machine learning appears particularly valuable in identifying persistent characteristic-return relationships and constructing more efficient latent factors, but the resulting models remain closely linked to traditional economic concepts such as value, profitability, and market risk. Future research could investigate whether more sophisticated regularization procedures, alternative machine-learning algorithms, more computationally efficient hyperparameter-selection methods, or efficient characteristic portfolios (e.g. Daniel *et al.* 2020) further improve the balance between predictive accuracy, economic interpretability, and pricing performance.

Acknowledgments

There are no conflict of interests to declare. No funding was received. We thank for the opportunity to run the computational analysis on the High Performance Computing Cluster supported by the Research and Specialist Computing Support service at the University of East Anglia (UEA).

Author contributions

CRediT: **Konstantinos Baltas:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing; **Gustavo Fruet Dias:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing; **Artiom Rumiancev:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Konstantinos Baltas http://orcid.org/0000-0002-5665-2871
Gustavo Fruet Dias http://orcid.org/0000-0001-6428-5074

References

Banz, R.W., The relationship between return and market value of common stocks. *J. Financ. Econ.*, 1981, **9**(1), 3–18.

BARRA, *United states equity*, 1998 (Barra Inc.,: Berkeley). Available online at: http://www.alacra.com/alacra/help/barra_handbook_US.pdf (24 July 2022).

Basu, S., Investment performance of common stocks in relation to their price-earnings ratios. *J. Finance*, 1977, **32**(3), 663–682.

Boser, B., Guyon, I. and Vapnik, V., A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, pp. 144–152, 1992 (Association for Computing Machinery: Pittsburgh, New York).

Bray, M., *The arbitrage pricing theory is not robust 1*, 1994 (London School of Economics: London). LSE Financial Markets Group Discussion Paper. Available online at: http://pi.lib.uchicago.edu/1001/cat/bib/1624086 (25 July 2022).

Büchner, M. and Bybee, L., Instrumented principal components analysis (IPCA), 2021 [Python Package (WIP)]. Available online at: https://github.com/bkelly-lab/ipca (24 September 2022).

Büchner, M. and Kelly, B., A factor model for option returns. *J. Financ. Econ.*, 2022, **143**(3), 1140–1161.
Bybee, L., Kelly, B. and Su, Y., Narrative asset pricing: Interpretable systematic risk factors from news text. *Rev. Financ. Stud.*, 2023, **36**(12), 4759–4787.
Carhart, M.M., On persistence in mutual fund performance. *J. Finance*, 1997, **52**(1), 57–82.
Chamberlain, G. and Rothschild, M., Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 1983, **51**(5), 1281–1304.
Chapelle, O., Vapnik, V., Bousquet, O. and Mukherjee, S., Choosing multiple parameters for support vector machines. *Mach. Learn.*, 2001, **46**(1), 131–159.
Chen, L., Pelger, M. and Zhu, J., Deep learning in asset pricing. *Manage. Sci.*, 2024, **70**(2), 714–750.
Chen, N., Roll, R. and Ross, S., Economic forces and the stock market. *J. Bus.*, 1986, **59**(1), 383–403.
Cherkassky, V. and Ma, Y., Practical selection of SVM parameters and noise estimation for SVM regression. *Neural. Netw.*, 2004, **17**(1), 113–126.
Cochrane, J.H., *Asset Pricing*, 2009 (Princeton University Press: New Jersey).
Cochrane, J.H., Presidential address: Discount rates. *J. Finance*, 2011, **66**(4), 1047–1108.
Connor, G. and Korajczyk, R.A., Performance measurement with the arbitrage pricing theory. *J. Financ. Econ.*, 1986, **15**(3), 373–394.
Connor, G. and Korajczyk, R.A., Risk and return in an equilibrium apt. *J. Financ. Econ.*, 1988, **21**(2), 255–289.
Cortes, C. and Vapnik, V., Support-vector networks. *Mach. Learn.*, 1995, **20**(3), 273–297.
Daniel, K. and Titman, S., Evidence on the characteristics of cross sectional variation in stock returns. *J. Finance.*, 1997, **52**(1), 1–33.
Daniel, K. and Titman, S., Testing factor-model explanations of market anomalies. *Critical Finance Rev.*, 2012, **1**(1), 103–139.
Daniel, K.D., Mota, L., Rottke, S. and Santos, T., The cross-section of risk and return. *Rev. Financ. Stud.*, 2020, **33**(5), 1927–1979.
Daniel, K.D., Discussion of 'Characteristics are covariances: A unified model of risk and return' by Kelly, B., Pruitt, S. and Su, Y., (2019). In *28th Utah Winter Finance Conference*, 2018 (Columbia University: Snowbird, New York).
Fama, E.F. and French, K.R., Common risk factors in the returns on stocks and bonds. *J. Financ. Econ.*, 1993, **33**(1), 3–56.
Fama, E.F. and French, K.R., A five-factor asset pricing model. *J. Financ. Econ.*, 2015, **116**(1), 1–22.
Fama, E.F. and MacBeth, J.D., Risk, return, and equilibrium: Empirical tests. *J. Polit. Economy*, 1973, **81**(3), 607–636.
Fan, J., Furger, A. and Xiu, D., Incorporating global industrial classification standard into portfolio allocation. *J. Bus. Econ. Stat.*, 2016, **34**(4), 489–503.
Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R. and Lin, C.J., Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 2008, **9**(9), 1871–1874.
Frazzini, A. and Pedersen, L.H., Betting against beta. *J. Financ. Econ.*, 2014, **111**(1), 1–25.
French, K.R., *Sorted portfolios*, 2021 (Dartmouth College: Hanover). Available online at: https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html (31 July 2022).
Freyberger, J., Neuhierl, A. and Weber, M., Dissecting characteristics nonparametrically. *Rev. Financ. Stud.*, 2020, **33**(5), 2326–2377.
Gibbons, M., Ross, S. and Shanken, J., A test of effciency of a given portfolio. *Econometrica*, 1989, **57**(5), 1121–1152.
Gonçalves, S. and Kilian, L., Bootstrapping autoregressions with conditional heteroskedasticity of unknown form. *J. Econom.*, 2004, **123**, 89–120.
Gu, S., Kelly, B.T. and Xiu, D., Autoencoder asset pricing models. *J. Econom.*, 2021, **222**(1), 429–450.
Hansen, L.P. and Richard, S.F., The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing models. *Econometrica: J. Econom. Soc.*, 1987, **55**(3), 587–613.
Hsieh, C.J., Chang, K.W., Lin, C.J., Keerthi, S. and Sellamanickam, S., A dual coordinate descent method for large-scale linear svm. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 408–415, 2008 (Association for Computing Machinery: Helsinki, New York).
Hsu, C.W., Chang, C.C. and Lin, C.J., A practical guide to support vector classification. Department of Computer Science Technical Report, National Taiwan University, 2003.
Jagannathan, R. and Wang, Z., The conditional CAPM and the cross-section of expected returns. *J. Finance.*, 1996, **51**(1), 3–53.
Jarque, C.M. and Bera, A.K., A test for normality of observations and regression residuals. *Int. Stat. Rev./ Revue Internationale de Statistique*, 1987, **55**(2), 163–172.
Jensen, T., Kelly, B.T. and Pedersen, L.H., Is there a replication crisis in finance? *J. Finance.*, 2023, **78**(5), 2465–2518.
Keerthi, S., Sindhwani, V. and Chapelle, O., An efficient method for gradient-based adaptation of hyperparameters in svm models. In *Advances in Neural Information Processing Systems*, Vancouver and Whistler, BC, Canada, edited by B. Schölkopf, J. Platt and T. Hoffman, Vol. 19, 2006 (MIT Press).
Kelly, B.T., Pruitt, S. and Su, Y., Characteristics are covariances. *J. Financ. Econ.*, 2019, **134**(3), 501–524.
Kelly, B.T., Pruitt, S. and Su, Y., Instrumented principal component analysis. SSRN, 2020. Available online at: http://dx.doi.org/10.2139/ssrn.2983919 (24 July 2022).
Kou, S., Peng, X. and Zhong, H., Asset pricing with spatial interaction. *Manage. Sci.*, 2018, **64**(5), 2083–2101.
Kozak, S., Nagel, S. and Santosh, S., Shrinking the cross-section. *J. Financ. Econ.*, 2020, **135**(2), 271–292.
Lewellen, J., Nagel, S. and Shanken, J., A skeptical appraisal of asset pricing tests. *J. Financ. Econ.*, 2010, **96**(2), 175–194.
Lintner, J., Security prices, risk, and maximal gains from diversification. *J. Finance*, 1965a, **20**(4), 587–615.
Lintner, J., The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Rev. Econ. Stat.*, 1965b, **47**(1), 13–37.
Lu, C.J., Lee, T.S. and Chiu, C.C., Financial time series forecasting using independent component analysis and support vector regression. *Decis. Support. Syst.*, 2009, **47**(2), 115–125.
Markowitz, H., Portfolio selection. *J. Finance*, 1952, **7**(1), 77–91.
Mossin, J., Equilibrium in a capital asset market. *Econometrica: J. Econom. Soc.*, 1966, **34**(4), 768–783.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E., Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 2011, **12**(1), 2825–2830.
Ross, S.A., The arbitrage theory of capital asset pricing. *J. Econ. Theory.*, 1976, **13**(3), 341–360.
Shanken, J., On the estimation of beta-pricing models. *Rev. Financ. Stud.*, 1992, **5**(1), 1–33.
Sharpe, W.F., Capital asset prices. *J. Finance*, 1964, **19**(3), 425–442.
Stattman, D., Book values and stock returns. *The Chicago MBA: A J. Sel. Papers*, 1980, **4**(1), 25–45.
Talbi, E.G., *Metaheuristics*, 2009 (Wiley Publishing: New Jersey).
Treynor, J.L., Market value, time, and risk, 1961. Unpublished. Available online at: http://dx.doi.org/10.2139/ssrn.2600356 (24 July 2022).
Treynor, J.L., Towards a theory of market value of risky assets', 1962. Unpublished. Available online at: http://dx.doi.org/10.2139/ssrn.628187 (24 July 2022).
Welch, B.L., The generalization of 'STUDENT'S' problem when several different population variances are involved. *Biometrica*, 1947, **34**(1-2), 28–35.
Wharton Research Data Services, CRSP monthly update monthly stock file and compustat—capital IQ North America fundamentals quarterly, 2021 (Wharton School of the University of Pennsylvania: Philadelphia). Available online at: https://wrds-www.wharton.upenn.edu/pages/get-data/ (31 July 2022).
Zhou, Y., Wilkinson, D., Schreiber, R. and Pan, R., Large-scale parallel collaborative filtering for the netflix prize. In *Algorithmic Aspects in Information and Management*, edited by R. Fleischer and J. Xu, pp. 337–348, 2008 (Springer: Berlin, Heidelberg).

Appendix

A.1. Characteristics

Table A1. Characteristic variables.

No.	Acronym in dataset	Name | Category
	Intangibles	
2.	oacc[1]	Opperating accruals to book equity
25.	oaccruals_at	Opearting accruals to assets
26.	opex_at	Operating leverage
28.	pcm	Price to cost margin
	Investment	
13.	ppeinv_gr1a	Change in property, plant and equipment to lagged assets
17.	at_gr1	One year asset growth
24.	noa_gr1a	Net operating asset one year growth
	Past returns	
20.	ret_60_12	Long-term reversal
32.	ret_3_1[2]	Lagged two month return
33.	ret_12_1	Momentum
34.	ret_12_7	Intermediate momentum
35.	ret_36_12	Long-term momentum
44.	ret_1_0	Short-term reversal
	Profitability	
4.	sale_me	Sales to market equity
10.	at_turnover	Asset turnover
11.	d2a	Depreciation and amortization to assets
15.	fc2y	Fix costs to sales
27.	ope_be	Operating profits to book equity
29.	ebit_sale	Profit margin
30.	cop_atl1[3]	Gross profitability to lagged book assets
38.	ebit_bev	Return on net operating assets
39.	niq_at	Return on assets
40.	ni_be	Return on book equity
42.	sga2s	Selling, general and administrative expenses to sales
	Trading frictions	
3.	assets	Total assets
6.	betabab_1260d[4]	CAPM beta
16.	ivol_ff3_21d[5]	Idiosyncartic volatility
19.	lme	Market equity
21.	turnover_126d[6]	Share turnover
22.	beta_60m	5-year market beta
23.	chcsho_12m	Net stock issues
36.	prc_highprc_252d	Lagged price to highest price in 12 months
37.	resid_var[7]	Residual variance
43.	bidaskhl_21d	The average bid-ask spread in the previous month
45.	dolvol_var_126d[8]	Coefficient of variation for dollar trading volume
46.	rvol_21d[9]	Realized volatility
	Value	
1.	at_me	Assets to market equity
5.	be_me	Book to market equity
7.	cash_at	Cash to assets
8.	cf	Free cash flow to book equity
9.	ocf_me	Cash flow to market quity
12.	div12m_me	Dividend yield to market equity
14.	ni_me	Earinings to market equity
18.	lev	Debt to debt and equity
31.	Q	Tobin's Q
41.	s2p	Sales to price

Notes: The table below presents 46 firm characteristics aligned with the variables used by Chen *et al.* (2024) and are located in French (2021) library or applied int the 2017 version of Freyberger *et al.* (2020). The majority of characteristics are obtained or constructed from data generated by running the SAS code provided in Jensen *et al.* (2023). For further details on their computation and detailed economic interpretation, please refer to these sources and the 'Internet Appendix' (IA) of Chen *et al.* (2024) as well as the code documentation of Jensen *et al.* (2023). Some characteristics are replaced with analogues due missing actual, or lack of primary variables needed for their computation. No. is the serial number of an identical or replaced characteristic in the IA of Chen *et al.* (2024). Acronym is the index of the characteristic used in the computer dataset. Name — Category are the characteristic names and the category they are allocated by Chen *et al.* (2024). Footnote numbers represent the replaced characteristics. Their revised computation method is as follows:[1]Changes in non-cash working capital minus depreciation scaled by book equity—$\frac{\texttt{oaccurals_atxassets}}{\texttt{book_equity}}$.[2]Return over the one-month lagged previous two months.[3]Gross profit divided by the last available book equity.[4]Frazzini-Pedersen market beta (Frazzini and Pedersen 2014).[5]21 days idiosyncratic volatility of the Fama and French (1993) 3 factor model.[6]Share turnover over the last 6 months.[7]Residual variance of the CAPM for the past 21 days.[8]Dollar volume volatility over the past 6 months.[9]Realized return volatility for the past 21 days in place of variance.



Figure A1. Characteristic correlation.
Notes: The figure above depicts the correlation matrix of 46 characteristics used in the analysis. Colours range from ice blue to red and represent the direction of the correlation coefficient. An absolute correlation of above 80% is present in 12 pairs of characteristics. Some characteristics are correlated due to their computational structure (e.g. `oaccruals_at` and `oacc` share the same numerator), while others have similar origin (`resid_var` and `ivol_ff3_21d` are residuals estimated using CAPM and FF3, respectively). Hence, the high correlations of certain characteristics, without regularization, may lead to multicollinearity issues in least squares optimization.

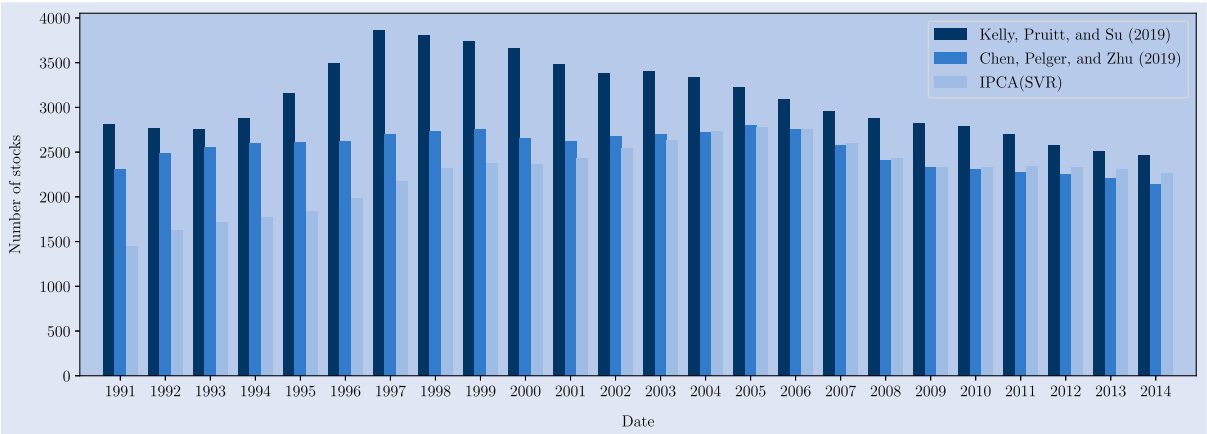


Figure A2. Stock count.
Notes: The figure presents the sample size comparison between Kelly *et al.* (2019), and Chen *et al.* (2024) and the current analysis dataset, on the overlapping dates. As evident, Kelly *et al.* (2019) utilize a more diverse set of stocks for testing, however, the number of characteristics in their analysis is only 36, compared to Chen *et al.* (2024) and the current study's 46. Furthermore, Kelly *et al.* (2019) replace missing characteristics with cross sectional medians, hence, the current dataset is best compared to Chen *et al.* (2024), who claim to possess the most complete sample available. As such, the current dataset lacks observations for the first 10 years in comparison to Chen *et al.* (2024), although after 5 years the gap shrinks by half with future yearly stock count being above 2000 units. Overall, the current sample is diverse for robust in-sample estimation as well as out-of-sample pricing.

Table A2. Correlation of created portfolios with original portfolios.

17IP		25SBM		25SM	
Portfolio	ρ	Portfolio	ρ	Portfolio	ρ
Food	0.982	SMLL	0.943	SMLL	0.953
Mines	0.855	SMLM1	0.927	SMLM1	0.937
Oil	0.993	SMLM2	0.941	SMLM2	0.922
Clths	0.973	SMLM3	0.953	SMLM3	0.92
Durbl	0.916	SMLH	0.958	SMLH	0.922
Chems	0.971	MEDM1	0.951	MEDM1	0.935
Cnsum	0.984	MEDM2	0.952	MEDM2	0.945
Cnstr	0.986	MEDM3	0.965	MEDM3	0.948
Steel	0.898	MEDM4	0.954	MEDM4	0.939
FabPr	0.933	MEDM5	0.946	MEDM5	0.892
Machn	0.983	MEDM6	0.952	MEDM6	0.894
Cars	0.963	MEDM7	0.964	MEDM7	0.949
Trans	0.978	MEDM8	0.961	MEDM8	0.937
Utils	0.989	MEDM9	0.942	MEDM9	0.928
Rtail	0.982	MEDM10	0.929	MEDM10	0.879
Finan	0.93	MEDM11	0.967	MEDM11	0.9
Other	0.969	MEDM12	0.974	MEDM12	0.935
		MEDM13	0.948	MEDM13	0.955
		MEDM14	0.951	MEDM14	0.944
		MEDM15	0.914	MEDM15	0.845
		BGL	0.985	BGL	0.837
		BGM1	0.972	BGM1	0.925
		BGM2	0.93	BGM2	0.951
		BGM3	0.907	BGM3	0.945
		BGH	0.827	BGH	0.833
Max	0.993		0.985		0.955
Min	0.855		0.827		0.833
μ	0.958		0.945		0.919

Notes: The table below presents correlations (ρ) of created portfolios with originals from French (2021) library. These are 17 Fama-French industry portfolios, 25 Size and Book-to-Market and 25 Size and Momentum portfolios. Overall, considering the smaller dataset used in the portfolio formation, there is a considerably high correlation between these portfolios and the originals, generally within the 90% + range and rarely falling below 85%.

A.2. Data

Table A3. Descriptive statistics of common risk factors.

	RMRF	SMB(1)	SMB(2)	HML	RMW	CMA	WML
μ	0.008***	0.002	0.002	0.001	0.003**	0.002	0.004*
	(3.368)	(1.024)	(1.160)	(0.535)	(2.258)	(1.579)	(1.770)
Min	− 0.172	− 0.168	− 0.149	− 0.140	− 0.185	− 0.069	− 0.344
Max	0.136	0.212	0.181	0.126	0.134	0.096	0.184
σ	0.043	0.032	0.031	0.031	0.026	0.020	0.048
$\tilde{\mu}_3$	− 0.615	0.619	0.375	0.096	− 0.412	0.639	− 1.430
$\tilde{\mu}_4$	1.440	7.182	4.264	2.749	10.881	2.475	10.450
$JB - st.$	59.200	285.340	32.414	1.505	941.804	28.614	955.295
$p(JB)$	0.000	0.000	0.000	0.471	0.000	0.000	0.000
T				360			
RMRF	1.000						
SMB(1)	0.244	1.000					
SMB(2)	0.241	0.977	1.000				
HML	− 0.101	− 0.224	− 0.042	1.000			
RMW	− 0.381	− 0.543	− 0.452	0.366	1.000		
CMA	− 0.337	− 0.145	− 0.042	0.638	0.250	1.000	
WML	− 0.282	0.014	− 0.040	− 0.221	0.066	0.011	1.000

Notes: This table illustrates factor descriptive statistics. RMRF is the excess value-weighted market return. SMB(1) and HML are Fama and French (1993) size and growth proxies, while SMB(2) is the revised size proxy (Fama and French 2015). RMW, and CMA are Fama and French (2015) profitability and investment factors. WML is the Carhart (1997) momentum factor. μ is the average proxy return, while its t-statistic is in parenthesis. Min and Max are the lowest and highest observation in the sample, respectively. σ, $\tilde{\mu}_3$ and $\tilde{\mu}_4$ represent standard deviation, skewness, and kurtosis of the data. $JB - st.$ is the Jarque and Bera (1987) normality distribution test wile $p(JB)$ is its p-value. *, **, *** represent values that are statistically significant at 10%, 5%, and 1% significance level, respectively.

Table A4. Descriptive statistics of 25 size and book-to-market portfolios.

	μ		Min	Max	σ	$\tilde{\mu}_3$	$\tilde{\mu}_4$	$JB-st.$	$p(JB)$
SMLL	0.011***	(2.619)	−0.250	0.428	0.080	0.494	3.157	15.025	0.001
SMLM1	0.014***	(3.905)	−0.219	0.327	0.067	0.308	2.749	6.653	0.036
SMLM2	0.014***	(4.214)	−0.230	0.251	0.062	−0.153	2.096	13.647	0.001
SMLM3	0.016***	(5.211)	−0.234	0.292	0.059	−0.116	2.873	1.045	0.593
SMLH	0.019***	(5.245)	−0.288	0.323	0.068	0.261	3.688	11.184	0.004
MEDM1	0.011***	(3.108)	−0.248	0.300	0.068	−0.203	2.054	15.895	0.000
MEDM2	0.013***	(4.123)	−0.240	0.193	0.059	−0.613	2.129	33.953	0.000
MEDM3	0.013***	(4.149)	−0.224	0.245	0.057	−0.420	2.119	22.246	0.000
MEDM4	0.013***	(4.225)	−0.224	0.221	0.056	−0.333	2.299	14.017	0.001
MEDM5	0.012***	(2.932)	−0.357	0.436	0.077	0.301	6.020	142.263	0.000
MEDM6	0.012***	(3.720)	−0.248	0.244	0.063	−0.346	1.810	28.431	0.000
MEDM7	0.013***	(4.688)	−0.219	0.229	0.054	−0.448	2.262	20.187	0.000
MEDM8	0.011***	(3.821)	−0.222	0.248	0.056	−0.395	2.515	12.891	0.002
MEDM9	0.011***	(3.664)	−0.283	0.247	0.056	−0.336	3.642	12.947	0.002
MEDM10	0.012***	(3.568)	−0.312	0.296	0.062	−0.204	4.843	53.448	0.000
MEDM11	0.012***	(3.996)	−0.219	0.190	0.056	−0.342	1.766	29.836	0.000
MEDM12	0.010***	(3.732)	−0.220	0.208	0.051	−0.518	3.038	16.129	0.000
MEDM13	0.012***	(4.280)	−0.258	0.232	0.054	−0.627	4.307	49.200	0.000
MEDM14	0.011***	(3.593)	−0.337	0.287	0.056	−0.724	7.451	328.685	0.000
MEDM15	0.011***	(3.205)	−0.418	0.306	0.064	−0.422	8.086	398.688	0.000
BGL	0.011***	(4.397)	−0.205	0.146	0.046	−0.482	1.889	32.452	0.000
BGM1	0.010***	(4.187)	−0.202	0.134	0.046	−0.587	2.085	33.209	0.000
BGM2	0.011***	(4.362)	−0.217	0.161	0.048	−0.765	3.315	36.632	0.000
BGM3	0.009***	(3.595)	−0.233	0.139	0.047	−0.638	2.927	24.536	0.000
BGH	0.010***	(3.193)	−0.293	0.325	0.057	−0.199	6.912	231.971	0.000
T	360								

Notes: The sample descriptive statistics illustrate the statistical properties of each of the 25 portfolios. SML(small), MED(medium) and BG(large) refer to the portfolio size group while L(low), M(medium), and H(high) represent book-to-market equity, with numbers from 1–15 indicating the lower to upper tier of the middle group. Consequently, MEDM7 is a medium size, medium book-to-market portfolio of tier 7. μ is the average raw portfolio return for the period, while its t-statistic is in brackets. Min and Max are the lowest and highest observation in the sample, respectively. σ, $\tilde{\mu}_3$, and $\tilde{\mu}_4$ are the second, third, and fourth central moments of the sample, with represent standard deviation, skewness, and kurtosis of the data. $JB-st.$ is the Jarque and Bera (1987) normality distribution test with $p(JB)$ representing its p-value. *, **, *** represent values that are statistically significant at 10%, 5%, and 1% significance level, respectively. Overall, only SMLM3 portfolio returns are normally distributed under the JB test at 99%, 95% and 90% confidence intervals, while SMLM1 returns are normally distributed only at 95% confidence. All portfolios have a statistically significant mean at a 1% significance level with the majority also demonstrating negative skewness, indicating at fatter left tails. Kurtosis varies noticeably between portfolios, with leptokurtic, mesokurtic and platykurtic examples, however, it seems that small portfolios tend to be more platykurtic-mesokurtic, while big portfolios go from being platykurtic to leptokurtic as book-to-market value increases.

Table A5. Descriptive statistics of 25 size and momentum portfolios.

	μ		Min	Max	σ	$\tilde{\mu}_3$	$\tilde{\mu}_4$	$JB-st.$	$p(JB)$
SML	0.018***	(4.113)	−0.295	0.437	0.081	0.642	3.987	39.327	0.000
SMLM1	0.016***	(5.254)	−0.221	0.271	0.058	0.139	3.050	1.198	0.549
SMLM2	0.017***	(5.831)	−0.240	0.219	0.054	−0.411	3.125	10.363	0.006
SMLM3	0.015***	(5.330)	−0.189	0.178	0.053	−0.294	1.417	42.811	0.000
SMLH	0.013***	(3.805)	−0.257	0.323	0.066	0.045	2.653	1.926	0.382
MEDM1	0.013***	(2.817)	−0.335	0.455	0.085	0.730	6.318	197.169	0.000
MEDM2	0.012***	(3.924)	−0.267	0.342	0.059	−0.159	5.015	62.434	0.000
MEDM3	0.013***	(4.698)	−0.226	0.207	0.053	−0.505	2.267	23.378	0.000
MEDM4	0.012***	(4.583)	−0.216	0.171	0.051	−0.514	2.164	26.343	0.000
MEDM5	0.011***	(3.231)	−0.234	0.240	0.065	−0.307	1.508	39.026	0.000
MEDM6	0.011**	(2.489)	−0.352	0.409	0.081	0.304	5.097	71.549	0.000
MEDM7	0.012***	(4.048)	−0.275	0.306	0.057	−0.237	5.412	90.640	0.000
MEDM8	0.011***	(4.321)	−0.204	0.241	0.050	−0.319	2.686	7.599	0.022
MEDM9	0.011***	(4.274)	−0.194	0.189	0.051	−0.325	1.716	31.060	0.000
MEDM10	0.014***	(4.453)	−0.220	0.207	0.061	−0.411	1.312	52.844	0.000
MEDM11	0.011**	(2.497)	−0.418	0.392	0.083	−0.025	5.495	93.442	0.000
MEDM12	0.012***	(4.373)	−0.265	0.244	0.054	−0.535	5.179	88.417	0.000
MEDM13	0.011***	(4.292)	−0.219	0.220	0.047	−0.502	3.790	24.471	0.000
MEDM14	0.011***	(4.398)	−0.203	0.169	0.047	−0.448	2.460	16.400	0.000
MEDM15	0.011***	(3.727)	−0.237	0.218	0.058	−0.275	1.986	19.971	0.000
BGL	0.010**	(2.238)	−0.339	0.587	0.084	1.158	10.422	906.833	0.000
BGM1	0.010***	(3.685)	−0.263	0.262	0.051	−0.202	5.192	74.507	0.000
BGM2	0.011***	(4.853)	−0.186	0.162	0.042	−0.507	2.848	15.774	0.000
BGM3	0.011***	(4.912)	−0.169	0.137	0.041	−0.499	1.835	35.272	0.000
BGH	0.011***	(3.952)	−0.234	0.183	0.054	−0.503	2.232	24.029	0.000
T	360								

Notes: The sample descriptive statistics illustrate the statistical properties of each of the 25 portfolios. SML(small), MED(medium) and BG(large) refer to the portfolio size group while L(low), M(medium), and H(high) represent momentum, with numbers from 1–15 indicating the lower to upper tier of the middle group. Consequently, MEDM7 reads as medium size, medium momentum portfolio of tier 7. μ is the average raw portfolio return for the period, while its t-statistic is in brackets. Min and Max are the lowest and highest observation in the sample, respectively. σ, $\tilde{\mu}_3$, and $\tilde{\mu}_4$ are the second, third, and fourth central moments of the sample, with represent standard deviation, skewness, and kurtosis of the data. $JB-st.$ is the Jarque and Bera (1987) normality distribution test with $p(JB)$ representing its p-value. *, **, *** represent values that are statistically significant at 10%, 5%, and 1% significance level, respectively. Overall, two portfolios—SMLM1 and SMLH are normally distributed under the JB test at 99%, 95% and 90% confidence intervals, while MEDM11—under 95%. Most portfolios have a statistically significant mean at a 1% significance level, while MEDM6, MEDM11 and BGL are only significant at 5% level. Most portfolios demonstrate negative skewness, indicating at fatter left tails. Kurtosis varies noticeably between portfolios, with leptokurtic, mesokurtic and platykurtic examples, however, it appears to be leptokurtic for most portfolios.

Table A6. Descriptive statistics of 17 industry portfolios.

	μ		Min	Max	σ	$\tilde{\mu}_3$	$\tilde{\mu}_4$	$JB-st.$	$p(JB)$
Food	0.012***	(5.788)	−0.139	0.143	0.039	−0.252	1.592	33.541	0.000
Mines	0.014***	(3.285)	−0.343	0.426	0.083	0.299	2.217	14.562	0.001
Oil	0.010***	(3.076)	−0.322	0.320	0.064	0.139	4.465	33.365	0.000
Clths	0.017***	(5.338)	−0.203	0.239	0.060	−0.127	1.534	33.201	0.000
Durbl	0.015***	(5.057)	−0.220	0.295	0.057	−0.039	2.693	1.510	0.470
Chems	0.014***	(4.565)	−0.199	0.269	0.060	0.106	2.146	11.616	0.003
Cnsum	0.013***	(6.110)	−0.095	0.156	0.041	−0.121	0.216	117.154	0.000
Cnstr	0.016***	(5.273)	−0.223	0.203	0.059	−0.153	1.499	35.197	0.000
Steel	0.013***	(2.866)	−0.353	0.289	0.089	−0.138	1.908	19.017	0.000
FabPr	0.016***	(5.441)	−0.184	0.242	0.057	−0.111	1.428	37.792	0.000
Machn	0.020***	(5.347)	−0.267	0.237	0.071	−0.224	1.384	42.204	0.000
Cars	0.018***	(4.403)	−0.246	0.495	0.076	0.975	6.012	193.128	0.000
Trans	0.015***	(5.339)	−0.190	0.178	0.052	−0.538	1.737	41.322	0.000
Utils	0.011***	(5.441)	−0.119	0.128	0.039	−0.450	0.684	92.571	0.000
Rtail	0.016***	(6.243)	−0.126	0.192	0.048	0.068	0.902	66.274	0.000
Finan	0.016***	(5.617)	−0.207	0.206	0.055	−0.184	1.819	22.942	0.000
Other	0.016***	(6.469)	−0.155	0.154	0.046	−0.110	0.965	62.863	0.000
T	360								

Notes: The sample descriptive statistics illustrate the statistical properties of each of the 17 industry portfolios. μ is the average raw portfolio return for the period, while its t-statistic is in brackets. Min and Max are the lowest and highest observation in the sample, respectively. σ, $\tilde{\mu}_3$, and $\tilde{\mu}_4$ are the second, third, and fourth central moments of the sample, with represent standard deviation, skewness, and kurtosis of the data. $JB-st.$ is the Jarque and Bera (1987) normality distribution test with $p(JB)$ representing its p-value. *, **, *** represent values that are statistically significant at 10%, 5%, and 1% significance level, respectively. Overall, no portfolio is normally distributed under the JB test at any confidence intervals. All portfolios have a statistically significant mean at a 1% significance level and the majority has negative skewness. Kurtosis varies noticeably between portfolios, with leptokurtic, mesokurtic and platykurtic examples. 'Durables' are only normally distributed at 1% significance.

A.3. Algorithms

Algorithm 1: IPCA(SVR) model.

Input: $\boldsymbol{R}^c, \boldsymbol{Z}, \boldsymbol{R}^e, T, N, L, K, C, \varepsilon$
Result: $\bar{\boldsymbol{\Gamma}}, \bar{\boldsymbol{F}}$
// Set the maximum number of iterations
$p \leftarrow 1000$;
$L^n \Leftarrow \{n \in \mathbb{Z} : 0 < n \leq p+1\}$;
$\tilde{K} \leftarrow K+1$;
$\bar{\boldsymbol{\Gamma}}_n \leftarrow 0 \in \mathbb{R}^{L \times \tilde{K}}$;
$\bar{\boldsymbol{f}}_{t+1} \leftarrow 0 \in \mathbb{R}^{\tilde{K} \times 1}$ and $\bar{\boldsymbol{f}}_{t+1} \in \bar{\boldsymbol{F}}_n$;
// Apply unstandardized principal components analysis to characteristic weighted portfolios
$\frac{\boldsymbol{R}^c \boldsymbol{R}^{c'}}{T-1} = \boldsymbol{V}\boldsymbol{L}\boldsymbol{V}'$;
$\boldsymbol{\Gamma}_{n-1} \leftarrow \boldsymbol{V}$;
for $t \leftarrow 1$ **to** $T-1$ **do**
  $\boldsymbol{f}_{t+1} \leftarrow \left(\boldsymbol{\Gamma}'_{n-1}\boldsymbol{Z}'_t\boldsymbol{Z}_t\boldsymbol{\Gamma}_{n-1}\right)^{-1}\boldsymbol{\Gamma}'_{n-1}\boldsymbol{Z}'_t\boldsymbol{r}^e_{t+1}$;
end
// Fit $\bar{\boldsymbol{\Gamma}}$ and $\bar{\boldsymbol{F}}$ until convergence
while $\bar{\boldsymbol{\Gamma}}_n, \bar{\boldsymbol{F}}_n \xrightarrow{L^n} \bar{\boldsymbol{\Gamma}}, \bar{\boldsymbol{F}}$ **do**
  for $i \leftarrow 1$ **to** N **do**
    for $t \leftarrow 1$ **to** $T-1$ **do**
      if $\bar{\boldsymbol{\Gamma}}_\alpha = 0$ **then**
        $\bar{\boldsymbol{f}}_{t+1}\boldsymbol{z}_{i,t} \leftarrow \mathrm{vec}(([1, \boldsymbol{f}'_{t+1}]' \otimes \boldsymbol{z}_{i,t})')$; // restricted model
      else
        $\bar{\boldsymbol{f}}_{t+1}\boldsymbol{z}_{i,t} \leftarrow \mathrm{vec}((\bar{\boldsymbol{f}}_{t+1} \otimes \boldsymbol{z}_{i,t})')$; // unrestricted model
      end
      $\xi_{i,t+1} \leftarrow r^e_{i,t+1} - \mathrm{vec}(\bar{\boldsymbol{\Gamma}}_n)'\bar{\boldsymbol{f}}_{t+1}\boldsymbol{z}_{i,t}$;
      $\mathrm{vec}(\bar{\boldsymbol{\Gamma}}_n) \leftarrow \arg\min_{\bar{\boldsymbol{\Gamma}}} \frac{1}{2}\mathrm{vec}(\bar{\boldsymbol{\Gamma}})'\mathrm{vec}(\bar{\boldsymbol{\Gamma}}) + C\sum_{i=1}^{N}\sum_{t=1}^{T-1}\max(0, |\xi_{i,t+1}| - \varepsilon)$;
    end
  end
  $\bar{\boldsymbol{\Gamma}}_n \leftarrow \mathrm{vec}^{-1}_{L,\tilde{K}}(\mathrm{vec}(\bar{\boldsymbol{\Gamma}}_n))'$**and** enforce orthogonality **and** enforce sign convension;
  for $t \leftarrow 1$ **to** $T-1$ **do**
    $\boldsymbol{F}_{n,t+1} \leftarrow \left(\boldsymbol{\Gamma}'_\beta\boldsymbol{Z}'_t\boldsymbol{Z}_t\boldsymbol{\Gamma}_\beta\right)^{-1}\boldsymbol{\Gamma}'_\beta\boldsymbol{Z}'_t\left(\boldsymbol{r}^e_{t+1} - \boldsymbol{Z}_t\boldsymbol{\Gamma}_\alpha\right)$;
  end
  $\bar{\boldsymbol{F}}_n \leftarrow [\boldsymbol{1}, \boldsymbol{F}_n]$
end