# ZOOKEEPING

## A COMPARISON OF FACTOR PRUNING METHODS

**AXEL RILEY**

**HUGO RIPPE**

Bachelor Thesis

Stockholm School of Economics

2026

**Zookeeping: A Comparison of Factor Pruning Methods**

Abstract:

The proliferation of proposed risk factors in empirical asset pricing, often referred to as the factor zoo, complicates the identification of a true stochastic discount factor (SDF). To address this, we assess in- and out-of-sample cross-sectional performance for Fama-French models, Principal Component Analysis (PCA), Risk-Premium PCA (RP-PCA), Bayesian Model Averaging (BMA), and a nonlinear Deep Latent Factor Model (DLFM) across varying regularization regimes. We find that the performance of dense and latent factor models depends strongly on the choice of regularization method and strength, and that the optimal choice of least-squares method itself depends on how factors are constructed, while a regularization-agnostic BMA model provides robust pricing of the cross-section. Finally, we find a positive relationship between model dimensionality and cross-sectional pricing ability; however, performance improvements diminish rapidly for latent factor models, suggesting that the underlying SDF is likely to be dense in observable factors, yet adequately represented by a relatively small set of latent factors.

Keywords:

Asset pricing, Stochastic discount factor, Machine learning, Regularization

Authors:

Axel Riley (26130)
Hugo Rippe (25992)

Tutor:

Ye Zhang, Associate Professor, Department of Finance

Examiner:

Ramin Baghai, Professor, Department of Finance

Bachelor Thesis
Bachelor Program in Business & Economics
Stockholm School of Economics

# I. Introduction

The central goal of asset pricing is the identification of a set of risk factors that can be used to model and predict a cross-section of expected returns. Taken together, these risk factors make up the parameters of the stochastic discount factor (SDF), which acts as the underlying source of asset price movements. Over time, the number of "identified" candidate factors for the SDF has greatly increased, increasing the complexity of this modeling task. The pursuit of accurate and interpretable models in the face of a growing set of parameters is often referred to as the *factor zoo* problem. While many approaches to the issue have been proposed, it remains difficult to close in on what approach (if any) may be considered optimal, due to the variety of datasets available and metrics one may optimize for. In addition, the different proposed methods for thinning the factor zoo tend to build on greatly varying theoretical frameworks.

This work aims to address this gap by jointly evaluating model performance on a common dataset, enabling a direct comparison of their strengths and weaknesses. The models we investigate are the Fama-French three- and five-factor models, Principal Component Analysis (PCA), Risk-Premium PCA (RP-PCA), Bayesian Model Averaging (BMA), and a nonlinear Deep Latent Factor Model (DLFM). By applying these methods to a common dataset, it is possible to draw meaningful insights regarding each model's properties, such as which regularization methods work best, whether sparse or dense models price the cross-section more accurately, and whether high- or low-dimensional models tend to outperform one another. The distinction between sparsity and low dimensionality is subtle yet important. In this paper, sparsity is defined as the restriction of including a low number of observable factors as input to a model. This is distinct from a model being low-dimensional, as some asset pricing methods employ a low-rank set of *latent* factors, which are nonetheless calculated using a large set of observable factors. In total, our research question can be summarized as follows: Given a large set of candidate factors, which approaches to factor reduction best capture the cross-sectional pricing information out-of-sample, and what does the answer reveal about the stochastic discount factor? Our results suggest that optimal regularization is model-dependent, that dense models perform better than sparse ones, and that the relationship between model dimensionality and performance is positive. However, the returns from greater dimensionality rapidly diminish, and our findings show that a small set of latent factors can price the cross-section effectively.

The proliferation of proposed risk factors, referred to as the factor zoo (Cochrane, 2011), remains a central challenge in empirical asset pricing. This paper addresses this challenge through three primary contributions.

First, we contribute to the factor selection literature by providing a systematic, head-to-head comparison between sparse and dense SDF representations. While existing research justifies the need for thinning the factor zoo due to the high rate of false discoveries (Harvey, Liu, and Zhu, 2016; Feng, Giglio, and Xiu, 2020), the literature is divided on the nature of the true SDF. Many workhorse models assume sparsity (Sharpe, 1964; Fama and French, 1993; Carhart, 1997), a view supported by recent Bayesian and Lasso-based selection methods (Freyberger, Neuhierl, and Weber,

2020; Bryzgalova, Huang, and Julliard, 2023). Conversely, Kozak, Nagel, and Santosh (2020) argue that the SDF is fundamentally dense, with many small loadings aggregating to explain the cross-section. We bridge this gap by evaluating these competing philosophies (sparse versus dense) under a uniform testing framework using shared datasets and temporal splitting. We find that while sparse models offer interpretability, dense models consistently achieve higher out-of-sample $R^2$ performance at optimal configurations. The best dense specifications tested achieve $R^2$ values approximately 0.15 higher than the Fama-French five-factor model out-of-sample, albeit with wider confidence intervals. As such, our experiments suggest that the factor zoo contains meaningful information that sparse selection methods discard.

Second, we contribute to the dimensionality reduction literature by evaluating whether low-dimensional latent factor models can resolve the tension between SDF density and model tractability. Prior work has sought to compress the factor zoo using PCA-based methods (Onatski, 2012; Lettau and Pelger, 2020) or nonlinear autoencoders (Gu, Kelly, and Xiu, 2021, 2020). However, recent evidence suggests that high-dimensional models with significantly more parameters than time periods may actually outperform low-rank alternatives (Didisheim, Ke, Kelly, and Malamud, 2024; Kelly and Malamud, 2021). We analyze these conflicting conclusions by testing the limits of dimensionality reduction. The results show that Risk-Premium PCA using $k = 8$ latent factors achieves a cross-sectional $R^2$ out-of-sample of 0.50, comparable to a regularized 51-factor regression with over 6 times as many dimensions. Our nonlinear deep latent factor model also performs similarly; with the same number of latent factors it exhibits a mean $R^2$ value of 0.44. Mean absolute pricing errors are also comparable across these models. Taken together, these findings highlight the promise of low-dimensional, yet dense, latent factor models.

Third, we contribute to the econometric estimation literature by synthesizing traditional estimators with high-dimensional regularization techniques. Building on the foundational ordinary least squares (OLS) and generalized least squares (GLS) frameworks of Hansen (1982) and Jagannathan and Wang (1996), and the shrinkage techniques of Tibshirani (1996) and Kozak et al. (2020), we provide a comparison framework of estimator performance. We analyze how the interaction between weighting matrices and regularization types (Lasso vs. Ridge) affects the stability of risk prices. Our findings reveal that the optimal choice of weighting matrix and regularization method is highly model-dependent, and as such this design choice is especially critical when constructing an asset pricing model. Furthermore, we find that while RP-PCA collapses under the GLS regime, our deep latent factor model benefits substantially from the same estimator. These results indicate that the choice of least squares method is dependent on how factors are constructed. This view is largely missing from the standard Fama-MacBeth literature, where factors are typically taken as observables. Bayesian model averaging exhibits an out-of-sample cross-sectional $R^2$ of 0.29 (95% CI $[0.25, 0.34]$) without any tuning, while latent factor models and the 51-factor regression are heavily dependent on well-tuned regularization. Whereas the former shows that competitive cross-sectional pricing is achievable without hyperparameter tuning, the latter, running at their optimal configurations, exhibit $R^2$ values approximately 0.2 higher than BMA, but with confidence intervals two

to three times wider. The trade-off between model robustness and peak performance suggests that gains from richer factor compositions also create exposure to sample-specific noise; this trade-off should be weighed explicitly when selecting and benchmarking factor pricing models.

# II. Model framework

This section introduces the models compared throughout the paper. The set is chosen to span the axes along which factor models substantially differ: sparse versus dense, observable versus latent, and linear versus nonlinear. We first present the general linear factor model that underpins estimation and the regularization schemes applied at the cross-sectional stage. We then describe the Fama-French three- and five-factor models, principal component analysis, its risk-premium variant, Bayesian model averaging, and finally our proposed deep latent factor model. For all models except Bayesian model averaging, premia are estimated for every model through standard Fama-MacBeth regression.

## *A. Linear models*

Linear models form the foundation of empirical asset pricing. Typically optimized by least-squares methods, they are both computationally efficient and interpretable. As such, they constitute the common framework through which most asset pricing models are conceived. Fundamentally, linear models assume that the co-movement of assets is explained by a set of factors, and that any remaining movement is specific to the asset in question. Given a dataset of $N$ assets and $K$ factors over $T$ observation periods, excess returns may be modeled as a matrix $R \in \mathbb{R}^{N\times T}$ while factor values are denoted by $F \in \mathbb{R}^{K\times T}$. Formally, a general linear model is described by

$$R = \boldsymbol{a}\mathbf{1}^T + \beta F + \mathcal{E} \tag{1}$$

In Eq. 1, $\boldsymbol{a} \in \mathbb{R}^N$ denotes a vector of asset intercepts, while $\beta \in \mathbb{R}^{N\times K}$ denotes factor loadings such that $\beta_{ik}$ describes the sensitivity of asset $i$ to changes in factor $k$. Furthermore, $\mathcal{E}$ is modeled as a random variable corresponding to idiosyncratic asset returns. Linear models assume that the idiosyncratic return satisfies $\mathbb{E}[\mathcal{E}_{ti}] = 0$ for all $t, i$ and that $\mathrm{Cov}(\mathcal{E}, F) = 0$. Importantly, if $F$ is normalized such that $\mathbb{E}[F] = 0$, then the intercepts $\boldsymbol{a}$ will correspond to the expected average returns of each asset over time.

These mean returns can also be explained through the risk premium model formulation, which relates the exposure matrix $\hat{\beta}$ obtained by regressing Eq. 1 to the factor-specific risk prices, denoted as $\boldsymbol{\lambda} \in \mathbb{R}^K$. Risk prices, or risk premia, can be interpreted as the return an investor expects to receive from each "unit" of exposure to a given factor, which will in turn correspond to how much that investor should be willing to pay for that exposure. The risk premia for a given set of factors can be estimated through the regression

$$\bar{\mathbf{r}} = \lambda_{\mathbf{c}}\mathbf{1} + \hat{\beta}\boldsymbol{\lambda} + \boldsymbol{\alpha} \tag{2}$$

In Eq. 2, $\bar{\mathbf{r}} \in \mathbb{R}^N$ represents the arithmetic mean of time-series asset returns, $\boldsymbol{\lambda}_c$ is a scalar average mispricing that should be zero under proper model specification, and $\boldsymbol{\alpha} \in \mathbb{R}^N$ represents the vector of asset-specific pricing errors. Risk premia play an important role in asset pricing, as they contribute to explaining the long-term mean return of assets, rather than explaining their short-term variance. This paper focuses on different approaches towards predicting these risk premia, rather than focusing on how well different methods can predict short-term asset price variance. In this work, we apply Lasso and Ridge regularization, also known as $\ell_1$ and $\ell_2$ regularization, to the risk premium regression in Eq. 2 to assess how regularization impacts performance across the models tested.

### *B. Fama-French models*

Fama and French (1993) find a set of empirically determined variables, which they first proposed as a three-factor model. They determine three explanatory factors for the cross-section of average returns. These are:

1. Market excess return (denoted $MKT$)
2. Small companies' outperformance of large companies (denoted $SMB$)
3. Value stocks' outperformance of growth stocks (denoted $HML$)

A value stock is a stock with a high book-to-market ratio, while a growth stock is a stock with a low book-to-market ratio .

In Fama and French (2015), the same authors extend the framework to a new five-factor model which includes two new variables: $RMW$, the difference in returns between companies with robust and weak profitability, and $CMA$, the difference in returns between companies with low (Conservative) and high (Aggressive) investment. In total, the Fama-French five-factor model takes the form in Eq. 1 with the factor set $F = \{MKT, SMB, HML, RMW, CMA\}$.

### *C. Principal component analysis*

Principal component analysis (PCA) is a linear dimensionality reduction method that estimates new variables, known as principal components, as linear combinations of the input variables. Given a demeaned matrix of values $\tilde{X} = X - \bar{X}, \quad \tilde{X} \in \mathbb{R}^{N \times T}$, PCA finds the directions $\mathbf{w}_i,\ i \in \{1, 2, \ldots, N\}$ which explain as much variance in $\tilde{X}$ as possible. Mathematically, the objective amounts to finding the eigenvectors $\mathbf{w}_i$ and associated eigenvalues $d_i$ of the covariance matrix $\Sigma = \frac{1}{T}\tilde{X}\tilde{X}^T$ as

$$\Sigma \mathbf{w}_i = d_i \mathbf{w}_i, \quad \|\mathbf{w}_i\| = 1 \tag{3}$$

Sorting the resulting eigenvectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N$ by the magnitudes of their associated eigenvalues $d_1, d_2, \ldots, d_N$ yields an ordered set of weight vectors. Taking each weight vector and projecting

$\tilde{X}$ onto it gives principal components as $\boldsymbol{PC}_i = \mathbf{w}_i^T \tilde{X}$ such that different principal components have no covariance over the dataset and such that the variance explained by each subsequent component is decreasing. To achieve a reduction in dataset dimensionality, a subset of the principal components must be chosen. Since each principal component explains more variance than the next, this is done in ordered fashion: for each integer $0 < k \leq N$, the set $\boldsymbol{PC}_1, \boldsymbol{PC}_2, \ldots, \boldsymbol{PC}_k$ is the set of $k$ components which maximally explain the variance in $\tilde{X}$.

### *D. Risk-premium PCA*

In the context of asset pricing models, PCA finds the latent variables that optimally explain the variance in the returns data. However, this objective is misaligned with that of explaining the cross-section of returns: since high-variance factors aren't necessarily associated with risk premia, PCA may select factors that have no relevance to asset pricing. To address this issue, Lettau and Pelger (2020) introduce Risk-Premium PCA, or RP-PCA, which applies PCA to a modified second-moment matrix $\Sigma_\gamma = \frac{1}{T} X X^T + \gamma \mu \mu^T$, where $\mu \in \mathbb{R}^N$ denotes the mean across observations. Note that $X$ is the non-demeaned matrix. The parameter $\gamma$ controls the degree of mean overweighting, which biases the resulting principal components towards explaining the cross-section of returns, rather than purely maximizing explained variance. The result is a new set of principal components which strike a balance between explaining variance and the cross-section.

### *E. Bayesian model averaging*

One method of mitigating the effects of model uncertainty inherent to traditional risk premium estimation methods is through the usage of Bayesian statistics. One such method is described by Bryzgalova et al. (2023) who, rather than relying on frequentist point estimates, treat the SDF parameters $\boldsymbol{\lambda}$ as random variables. This recovers the full posterior distribution of risk prices and enables the implementation of BMA, which integrates over the model space to provide robust inference even in the presence of weak factors. We define a linear SDF of the form $M_t = 1 - \mathbf{b}^T(\mathbf{F}_t - \mathbb{E}[\mathbf{F}_t])$, where the pricing kernels are identified via the moment conditions $\mathbb{E}[R_{it} M_t] = 0$. Under the assumption that excess returns follow a multivariate normal distribution, the likelihood of the data is combined with a "spike-and-slab" prior to perform factor selection. This prior is a mixture of a point mass at zero and a diffuse distribution:

$$p(\lambda_j | \gamma_j, \sigma^2) = \gamma_j \mathcal{N}(0, \omega^2 \sigma^2) + (1 - \gamma_j)\delta_0 \tag{4}$$

Where $\gamma_j \in \{0, 1\}$ is a binary indicator for factor inclusion, $\delta_0$ represents a Dirac delta function (the "spike"), and $\omega^2$ governs the variance of the "slab." By setting the spike at exactly zero, the model effectively regularizes the parameter space, shrinking the posterior inclusion probability (PIP) of redundant or weak factors toward zero. Estimation is then carried out using a Gibbs sampler to iterate through the conditional posterior distributions:

1. **Sample risk prices:** Draw $\boldsymbol{\lambda}$ from its conditional posterior distribution $p(\boldsymbol{\lambda}|\boldsymbol{\gamma}, \sigma^2)$, which, under a conjugate normal prior, results in a multivariate normal distribution.

2. **Update inclusion indicators:** Draw each $\gamma_j$ from a Bernoulli distribution with probability $\pi_j$, representing the conditional posterior probability that factor $j$ contributes to the SDF given the current values of $\boldsymbol{\lambda}$ and the data.

3. **Sample hyperparameters:** Update the residual variance $\sigma^2$ and any associated hyperparameters from their respective inverse-Gamma posteriors.

4. **Repeat** for $N$ iterations, discarding the initial burn-in period. The final PIP for each factor is calculated as the sample average of $\gamma_j$ across the remaining draws.

### *F. Deep latent factor model*

PCA-based models introduce a linear inductive bias in the construction of latent factors from observable ones. Although this approach often shows strong results, there is no reason to assume the existence of a linear relationship between observable data and the true factors governing the SDF. To address the potential misspecification which arises from making this inflexible assumption, we propose a deep learning model which parametrizes latent factors as a nonlinear function of the inputs. The conditional autoencoder by Gu et al. (2021) is a model which, given asset returns, uses firm-level characteristics to predict factor loadings $\hat{\beta}$. In contrast, the proposed model does not rely on firm-specific data, and its factor loadings and risk premia can easily be estimated using Fama-MacBeth regression. Furthermore, the proposed architecture allows for non-tradable factors as inputs. The model consists of an encoder-decoder pair $(E_\theta, D_\theta)$. Given a set of factors $\mathbf{f}_t$, the model attempts to reconstruct the returns $\mathbf{r}_t$ for the same time period. Note that the model tries to learn a map from factors to returns rather than reconstruct the input as is the case with autoencoders. Let $\hat{\mathbf{r}}_t = M(\mathbf{f}_t) = D_\theta(E_\theta(\mathbf{f}_t))$ and $T$ denote the maximum time step of the training data. The model is optimized by empirical risk minimization as

$$\underset{\theta}{\text{argmin}} \quad \frac{1}{T}\sum_{t=1}^{T}\|\mathbf{r}_t - M_\theta(\mathbf{f}_t)\|_2^2 \tag{5}$$

In practice, this objective is optimized by gradient descent methods. Once optimized, one may discard the decoder and use the encoder to produce latent factors as $\mathbf{z}_t = E_\theta(\mathbf{f}_t)$. Importantly, although we choose our encoder to be a nonlinear function, the decoder is a linear projection given the latent factors: $D_\theta(\mathbf{z}_t) = \beta_D \mathbf{z}_t + \mathbf{a}_D \; : \; \beta_D \in \mathbb{R}^{N\times k}, \mathbf{a}_D \in \mathbb{R}^N$ . By construction, this architecture causes the model to produce latent factors that are amenable to a linear model as in Eq. 1. Therefore, the deep latent factor model asserts a factor model of the form

$$R = \mathbf{a}\mathbf{1}^T + \beta E_\theta(F) + \mathcal{E} \tag{6}$$

The simple form in Eq. 6 allows the model's factor loadings $\hat{\beta}$ and risk premia $\boldsymbol{\lambda}$ to be estimated

through Fama-MacBeth regression as in all other models tested in this work. This choice of model architecture also allows latent factors to include nontradable factors as input, in contrast to the conditional autoencoder and PCA-based methods. If the loadings and risk premia are estimated by Fama-MacBeth regression, the decoder may be discarded. Although not explored in this work, the decoder loadings and estimated intercepts $(\beta_D, \mathbf{a}_D)$ can be used in place of the Fama-MacBeth estimates.

# III. Method

## *A. Data*

The data used in this work was chosen to fulfill three criteria: Firstly, it needed to include a wide range of factors and returns, in order for the factor zoo to be adequately represented, and for the implemented dense models to have a wide range of factors to draw from. Secondly, the data should cover a significant time period, and contain enough observations to allow for proper data splitting. Thirdly, noise should be kept to a minimum. With these criteria in mind, a dataset was chosen that contained monthly data between October 1973 and December 2016, containing 51 factors (17 non-tradable, 34 tradable) and 60 long-short portfolios which capture well-documented cross-sectional anomalies. 34 of these portfolios correspond to the tradable factors. The dataset was constructed from a range of sources, in accordance with Bryzgalova et al. (2023).

In order to preserve temporal coherence, the data was always split such that model training used "past" data before a given cutoff date, while evaluation was only performed on the subsequent data points. Due to the risk of results depending on the choice of cutoff date, most results shown make use of expanding data windows. At minimum, all models are optimized on $T_{\text{train}} = 96$ months of data and tested on the remaining observations. The number of training months is increased by $t = 12$ months in each increment, up to a minimum testing limit of $T_{\text{test}} = 60$ months. Given the structure of the dataset, this corresponded to the minimum training cutoff date being October 1981, and the latest being December 2011, resulting in a total of $W = 32$ windows. Using this method, we obtained mean in- and out-of-sample cross-sectional $R^2$ values for each model. The expanding window tests whether pricing relationships are robust across different sample lengths and out-of-sample periods, while allowing the model to benefit from an increasing amount of training data. Tables I and II show summary statistics of the test portfolios and factors respectively. All returns shown are excess returns. To calculate $t$-statistics, a Newey-West estimator Newey and West (1987) with a lag of $m = 5$ was used to account for heteroskedasticity and autocorrelation. More information on the data is available in Bryzgalova et al. (2023).

## *B. Implementation*

The models chosen for investigation in this work can be divided into three categories: sparse low-dimensional models (FF3, FF5, BMA), dense low-dimensional latent factor models (low-dimensional PCA and DLFM), and dense high-dimensional models (51-factor regressions, high-dimensional PCA

**Table I**

## Cross-sectional summary statistics for the 60 test portfolios

This table reports cross-sectional summary statistics across the $N = 60$ test portfolios, based on monthly observations from October 1973 to December 2016 ($T = 519$). For each portfolio, the time-series mean, standard deviation, annualized Sharpe ratio, computed as the monthly Sharpe ratio $\frac{\mathbb{E}[R_{i,t}]}{\sigma_i}$ where $R_{i,t}$ denotes the monthly excess return of portfolio $i$, and multiplying by $\sqrt{12}$. $t$-statistics computed using Newey and West (1987) with $m = 5$ lags. The table summarizes the resulting cross-sectional distribution of each statistic. The 60 test portfolios consist of long-short anomaly portfolios constructed from firm-level characteristics, of which 34 also appear in the factor set used for pricing. $P_{10}$ and $P_{90}$ denote the 10th and 90th percentiles.

| | Mean | SD | $P_{10}$ | Median | $P_{90}$ | Min | Max |
|---|---|---|---|---|---|---|---|
| Monthly mean return (%) | 0.09 | 0.20 | -0.00 | 0.00 | 0.41 | -0.00 | 0.76 |
| Standard deviation (%) | 0.62 | 1.24 | 0.03 | 0.04 | 2.64 | 0.01 | 4.58 |
| Annualized Sharpe ratio | 0.33 | 0.30 | -0.03 | 0.34 | 0.69 | -0.30 | 1.14 |
| $t$-statistic of mean | 2.02 | 1.84 | -0.17 | 2.00 | 4.21 | -1.74 | 7.47 |
| Fraction of portfolios with $\|t\| > 2$: 0.48 | | | | | | | |

and DLFM). All models, as listed in the introduction, were implemented in Python, primarily using the NumPy and PyTorch packages. Each model, when trained on a given set of data, should be able to output two results: the exposure matrix $\beta$ that describes the relationship between portfolio returns and the model's chosen factors, as well as the risk premia $\boldsymbol{\lambda}$ that can be used to predict each portfolio's mean return. For all non-Bayesian methods, $\beta$ is calculated first. Thereafter, the risk premia $\boldsymbol{\lambda}$ are estimated from the cross-sectional regression (see Eq. 2). If the models do not specify a way to estimate $\beta$, and instead perform factor selection directly, then $\beta$ is estimated through the regression in Eq. 1, with $F$ taken as the model's selected factors. BMA returns $\boldsymbol{\lambda}$ directly and uses standard regression to estimate $\beta$. In all experiments, BMA uses spike-and-slab prior variances $\sigma^2_{\text{slab}} = 20, \sigma^2_{\text{spike}} = 10^{-5}$, a prior inclusion probability of $\gamma_{\text{inc}} = 0.5$, and risk price variance prior $\sigma^2_\lambda = 1.0$. Following Lettau and Pelger (2020) we set the cross-sectional weighting parameter to $\gamma = 10$ for RP-PCA. All generalized least squares optimization uses linear shrinkage of the covariance matrix as proposed in Ledoit and Wolf (2004); this enables more stable estimation of the covariance matrix in high-dimensional settings. For the DLFM, we use two hidden layers in the encoder with dimensionalities 64 and 32 respectively, with layer normalization and ReLU activation functions in between the hidden layers. Additionally, we use dropout as proposed in Srivastava, Hinton, Krizhevsky, Sutskever, and Salakhutdinov (2014) with a probability $p = 0.2$ to prevent overfitting. All models are trained for $n = 100$ epochs using the Adam optimizer from Kingma and Ba (2015) with a learning rate of $\eta = 10^{-3}$.

### *C. Experimental design*

Our evaluation proceeds in two stages. In the first stage, we compute mean cross-sectional pricing performance out-of-sample across a grid of regularization strengths $\alpha_{\text{reg}}$ for both Lasso and Ridge by means of our expanding window regime. This produces the sensitivity plots presented in

## Table II
## Summary statistics for the factor set

This table reports summary statistics for the 51 factors used in this study, based on monthly observations from October 1973 to December 2016 ($T = 519$). Means and standard deviations are reported in percent for tradable factors. Sharpe ratios are computed as the ratio of mean to standard deviation, annualized by $\sqrt{12}$. $t$-statistics are computed using Newey-West Newey and West (1987) standard errors with 5 lags. Panel A reports statistics for the five Fama-French factors individually. Panel B reports a cross-sectional summary across the remaining 29 tradable factors, where 'Mean of monthly means' is the average across factors of each factor's time-series mean. Panel C lists the 17 non-tradable factors individually; these are reported in their native units, so means and standard deviations across this panel are not directly comparable, and $t$-statistics and Sharpe ratios are not reported. For more details, see the data construction process outlined in Bryzgalova et al. (2023).

| | Mean | SD | $t$-stat | Sharpe |
|---|---|---|---|---|
| *Panel A: Fama-French factors* | | | | |
| MKT | 0.56 | 4.58 | 2.67 | 0.43 |
| SMB | 0.26 | 3.01 | 2.00 | 0.30 |
| HML | 0.38 | 2.94 | 2.49 | 0.44 |
| RMW | 0.29 | 2.33 | 2.51 | 0.43 |
| CMA | 0.35 | 1.98 | 3.56 | 0.62 |
| *Panel B: Other tradable factors (cross-sectional summary)* | | | | |
| Number of factors | | | | 29 |
| Mean of monthly means | | | | 0.13 |
| Range of means | | | | [0.00, 0.76] |
| Mean SD | | | | 0.73 |
| Mean annualized Sharpe | | | | 0.55 |
| Fraction with $\|t\| > 2$ | | | | 0.83 |
| *Panel C: Non-tradable factors* | | | | |
| | Mean | SD | Min | Max |
| Liquidity factor | -0.001 | 0.057 | -0.39 | 0.28 |
| Innovations to intermediary capital ratio | 0.001 | 0.068 | -0.28 | 0.40 |
| Financial uncertainty | 0.950 | 0.137 | 0.73 | 1.43 |
| Real economic uncertainty | 0.751 | 0.061 | 0.66 | 1.00 |
| Macroeconomic uncertainty | 0.808 | 0.104 | 0.68 | 1.21 |
| Investor sentiment, Baker and Wurgler (2006) | -0.095 | 0.920 | -1.19 | 2.88 |
| Investor sentiment, Huang, Jiang, Tu, and Zhou (2015) | -0.006 | 0.889 | -2.27 | 2.84 |
| Term spread | 1.240 | 1.759 | -6.51 | 3.85 |
| Default rates | 1.107 | 0.463 | 0.55 | 3.38 |
| Dividend yield | 2.924 | 1.281 | 1.11 | 6.24 |
| Unemployment rate | 6.428 | 1.565 | 3.80 | 10.80 |
| Price-earnings ratio | 18.854 | 9.637 | 6.57 | 86.84 |
| Nondurable consumption growth | 0.095 | 0.672 | -3.95 | 3.35 |
| Growth rate service expenditure | 0.067 | 0.129 | -0.44 | 0.51 |
| IP growth rate | 0.152 | 0.726 | -4.43 | 2.05 |
| Monthly growth rate of the price of petroleum | 0.391 | 9.046 | -36.74 | 48.50 |
| Change in difference between 10-year/3-month bond yields | 0.011 | 0.546 | -2.89 | 5.34 |

Section IV.A, which reveal how each model's pricing ability depends on the choice and strength of regularization. The purpose of this stage is to investigate each model's dependence on regularization. Note that BMA (Bryzgalova et al. (2023)) doesn't use any regularization terms, as its risk premia are estimated by iteratively updating a posterior rather than through Fama-MacBeth regression.

In the second stage, we use a range of configurations chosen a priori for each model and evaluate them using an expanding window procedure. For this stage, all models are initialized over a set of 10 seeds to account for variations in network initialization. This yields a sequence of out-of-sample cross-sectional $R^2$ estimates from which we compute bootstrap confidence intervals with $n = 1000$ samples. For each resample, $W$ values are selected from the list of $R^2$ values with replacement and averaged, yielding $R_i^2$. The 95% confidence interval is then obtained by sorting the list $R^2 = (R_1^2, R_2^2, \ldots, R_n^2)$ and finding the values at the 2.5th and 97.5th percentiles. Additionally, the mean absolute percentage error (MAPE) of each model configuration was recorded as an additional way to gauge the performance of each method.

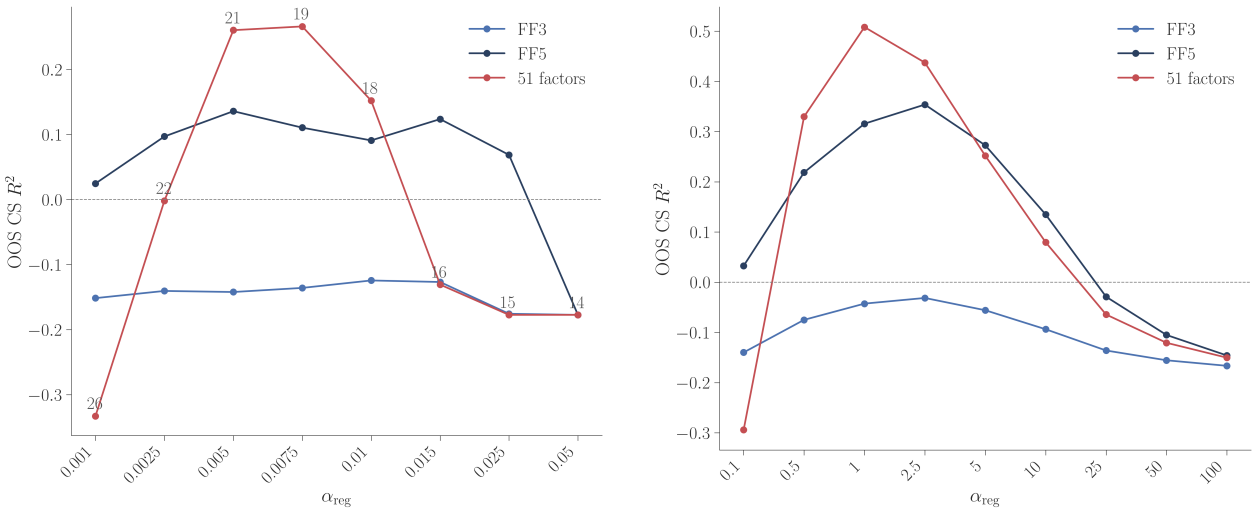
# IV. Results

This section presents the empirical results. We first characterize each model's sensitivity to the regularization regime, since the regularizer interacts with factor extraction and risk pricing. Furthermore, a fair cross-model comparison requires identifying each model's best operating conditions. We then report mean cross-sectional $R^2$ in- and out-of-sample at optimal configurations, separately under OLS and GLS objectives, and conclude by examining how pricing performance varies across the expanding evaluation windows. Reported intervals are 95% bootstrap confidence intervals, and all $R^2$ values are averaged across 10 random seeds to ensure robustness to variance from initialization.

## *A. Regularization sensitivity*

Firstly, the performance of the observable linear models was evaluated as a function of regularization strength. It is well-known that an ordinary least-squares regression becomes ill-conditioned in high-dimensional settings, leading to poor performance out-of-sample. Figure 1 shows these models' performance as a function of regularization strength:

The plot shows clear trends for the five-factor model and 51-factor regression. The Fama-French three-factor model exhibits very stable, although unimpressive, performance out-of-sample. The five factor model yields stable results across most of the tested regularization strengths, peaking at approximately $\alpha_{\text{reg}} = 2.5$ for Ridge regularization. The full regression on all variables showcases extremely strong dependence on regularization strength, with performance peaking around a Lasso strength of $\alpha_{\text{reg}} \in (0.005, 0.0075)$ or a Ridge strength of $\alpha_{\text{reg}} = 1$. At high regularization strengths, all models tend to collapse towards the same predictions and thus show very similar mean $R^2$ scores. The five-factor and 51-factor models strongly favor Ridge over Lasso regularization.




**(a)** Lasso **(b)** Ridge

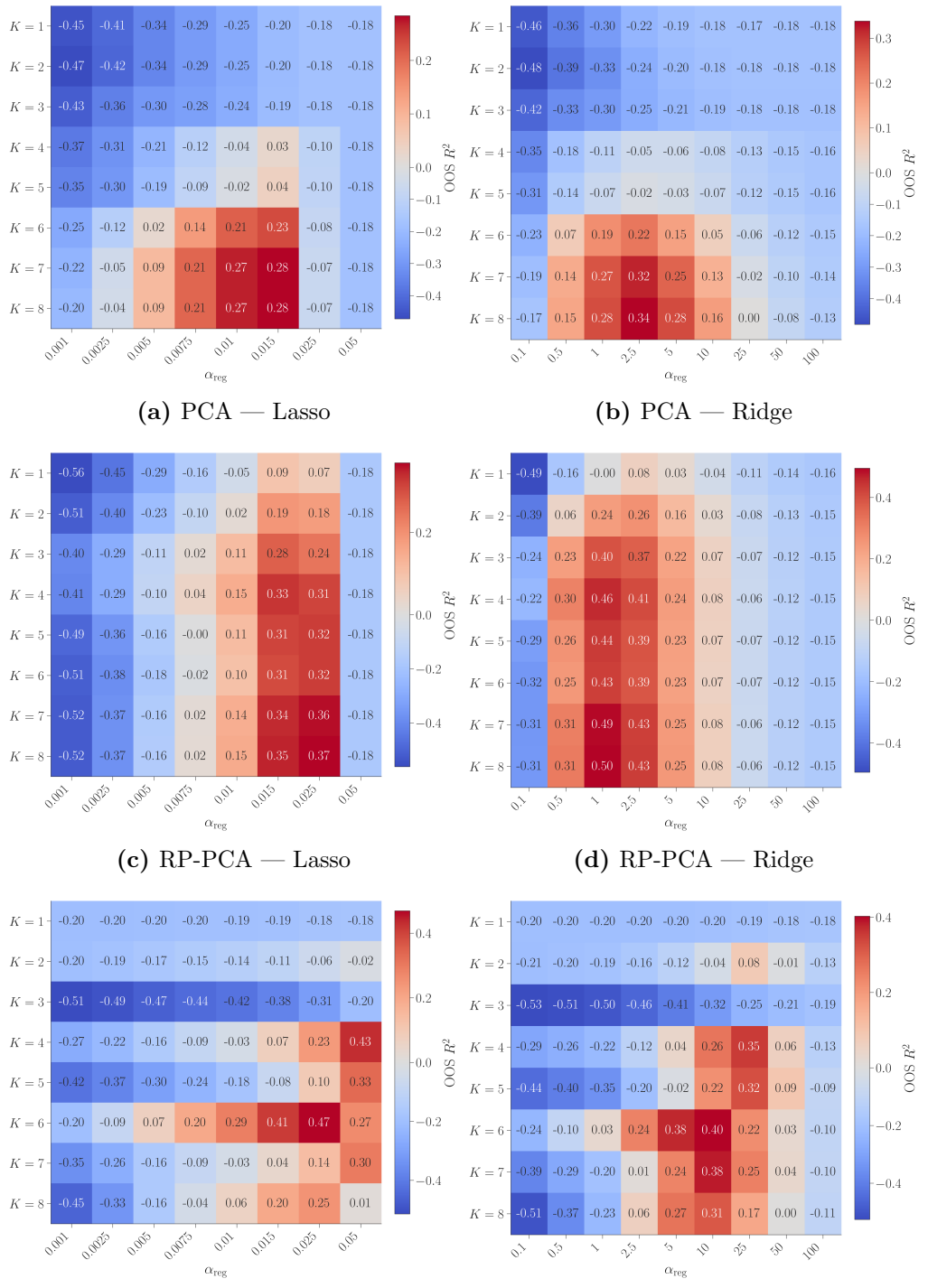
**Figure 1.** Cross-sectional out-of-sample $R^2$ for Fama-French models and 51-factor regression across regularization strengths. For Lasso, the numbers above the datapoints denote the number of factors associated with non-zero risk premia. Cross-sectional performance is evaluated using expanding windows and averaged across windows.

To analyze the effect of regularization strength on the latent factor models, we produce heatmaps of the out-of-sample cross-sectional $R^2$ as a function of $\alpha_{\text{reg}}$ and the number of latent factors $K$ for both Lasso and Ridge. Each cell in Figure 2 corresponds to a single $(K, \alpha_{\text{reg}})$ configuration, with red indicating positive pricing ability and blue indicating that the model underperforms the cross-sectional mean on average.

The heatmaps show that, similarly to the benchmark linear models tested in Figure 1, there are "sweet spots" with regards to regularization strength. The clearest trend occurs for the RP-PCA model, which showcases high performance across almost all numbers of latent factors tested in a specific range of $\alpha_{\text{reg}} \in [0.01, 0.025]$ for Lasso and $\alpha_{\text{reg}} \in [0.5, 10]$ for Ridge regularization. The PCA model consistently underperforms the cross-sectional mean at low values of $K$, which is indicative of risk premia often not being associated with the variables that have highest variance. Overall, the DLFM favors higher regularization strengths, especially for Ridge regularization. In contrast to the observable linear models, none of the latent factor models show clearly better performance when using one regularization method over the other.

## *B. Cross-sectional performance*

To evaluate the methods outlined in the literature under different optimization regimes, we performed training and evaluation of all chosen model types using different weighting and regularization configurations. Results are shown for two of the tested combinations, along with a summary of each model's best cross-sectional OOS $R^2$ performance across all regularization choices. Additionally, optimal model configurations were again calculated for both the GLS and OLS weighting




**(a)** PCA — Lasso **(b)** PCA — Ridge

**(c)** RP-PCA — Lasso **(d)** RP-PCA — Ridge

**(e)** DLFM — Lasso **(f)** DLFM — Ridge

**Figure 2.** Cross-sectional out-of-sample $R^2$ across regularization strengths and number of latent factors $K$ for PCA (top), RP-PCA (middle), and DLFM (bottom). Left column: Lasso. Right column: Ridge. Note that color bars are centered per-plot. All model optimization is done using OLS. Regularization strengths chosen a priori.

methods separately, so that the best regime for each method may be determined.

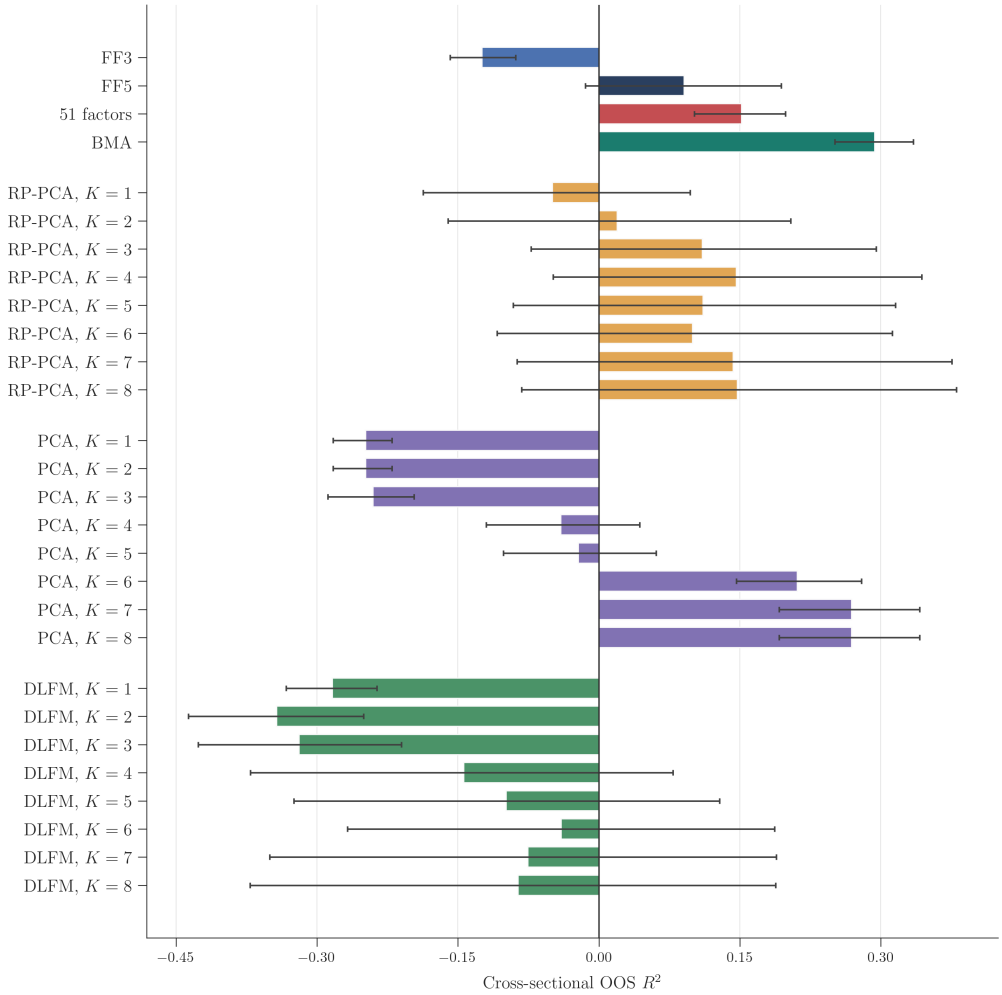



**Figure 3.** Cross-sectional $R^2$ out-of-sample for the different models. All non-Bayesian models used OLS and Lasso regularization with $\alpha_{\text{reg}} = 0.01$. RP-PCA, PCA and DLFM models were tested for different latent dimensionality levels, up to $K = 8$. Results are averaged over the the range of all data split configurations. Black line bars show the standard deviation of each model across the range of data splits.

Figure 3 shows that OLS regression with a weak Lasso regularization term results in a high $R^2$ for the high-dimensional PCA methods, but not when used with the DLFM approach. The RP-PCA results may seem promising, but have significant uncertainty compared to the other methods. BMA, which does not depend on regularization methods, has the most significant result, achieving an OOS $R^2$ of 0.297 with little uncertainty.

Figure 4 demonstrates that when the strength of the Lasso regularization increases, the DLFM performs significantly better, while all other regularization-dependent models collapse into the same poor estimation. These findings corroborate the findings from Section IV.A which showed that the DLFM favored higher levels of regularization.

Table III shows that for most models, using ordinary least squares with appropriate regular-

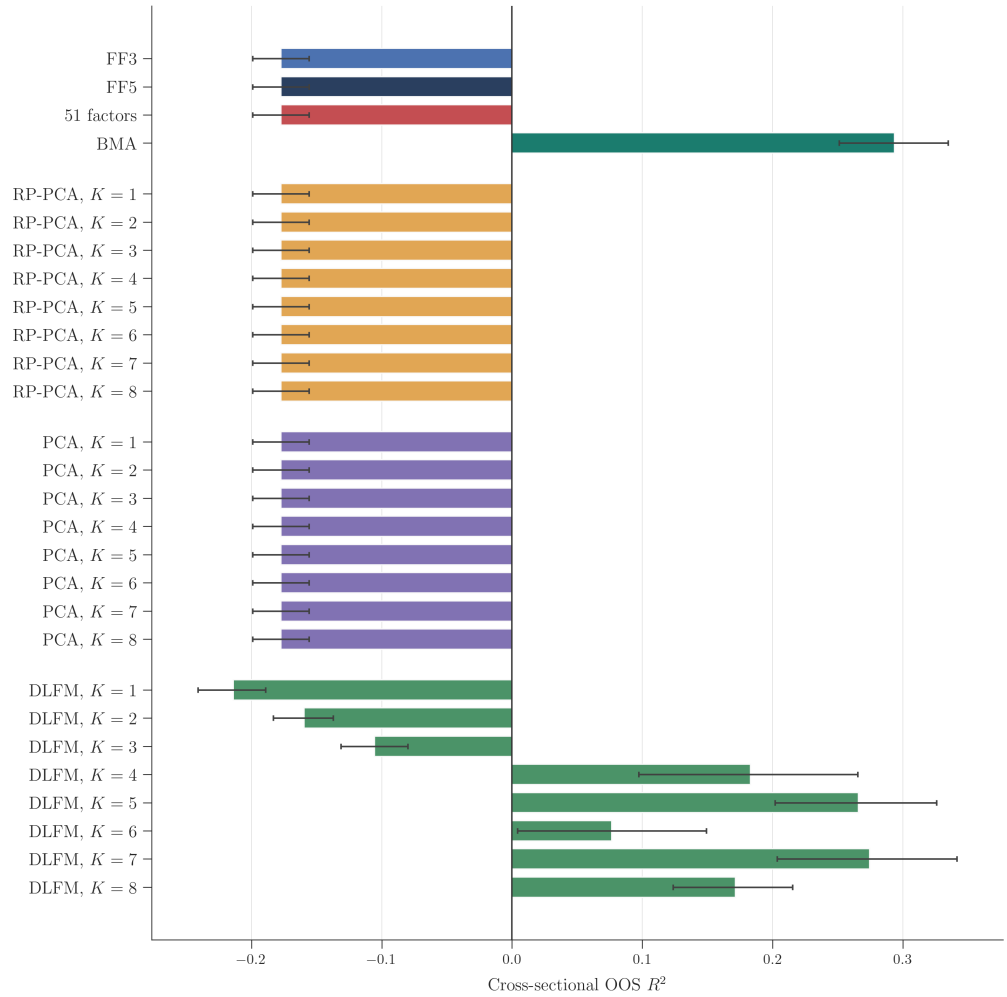



**Figure 4.** Cross-sectional $R^2$ out-of-sample for the different models. All non-Bayesian models used OLS and Lasso regularization with $\alpha_{\text{reg}} = 0.05$. RP-PCA, PCA and DLFM models were tested for different latent dimensionality levels, up to $K = 8$. Results are averaged over the the range of all data split configurations. Black line bars show the standard deviation of each model across the range of data splits.

ization weights produces the highest $R^2$ values out-of-sample. The optimization processes of the RP-PCA and 51-factor models yield strongly negative $R^2$ values when using GLS, which can occur because of noisy finite-sample estimates of the residual covariance matrix, making it ill-conditioned for least-squares objectives. For the 51-factor regression, this is likely due to the number of factors being large relative to the number of assets. For shorter training periods especially, this may cause the objective to be ill-conditioned. To understand why RP-PCA faces issues when using GLS, it is important to note that GLS upweights the directions in which idiosyncratic variance is small. RP-PCA constructs factors that are aligned with the directions corresponding to risk premia and thus extracts the signal from the directions with low idiosyncratic variance. Since GLS upweights precisely those directions, this is likely to lead to numerical instability. Regular PCA also suffers losses when using GLS, although the difference is smaller. This is contrasted by the DLFM, which

**Table III**

## Optimal model performance under GLS and OLS

This table shows all models' best cross-sectional performance out-of-sample when using GLS and OLS objectives on the risk premium regression. The amount of latent dimensions used by the latent factor models is indicated by $k$. Optimal configuration are specified with respect to the regularization method, as well as the regularization strength $\alpha_{\text{reg}}$. The displayed 95% confidence interval is taken with respect to the out-of-sample $R^2$ value. All confidence intervals are computed using bootstrapping with $n = 1000$ samples. Model results averaged over 10 seeds.

| | GLS | | | | OLS | | | |
|---|---|---|---|---|---|---|---|---|
| Model | Reg. | $\alpha_{\text{reg}}$ | OOS $R^2$ | 95% CI | Reg. | $\alpha_{\text{reg}}$ | OOS $R^2$ | 95% CI |
| FF3 | Ridge | 10.0 | -0.016 | [-0.111, 0.071] | Ridge | 2.5 | -0.031 | [-0.071, 0.011] |
| FF5 | Ridge | 10.0 | 0.308 | [0.214, 0.396] | Ridge | 2.5 | 0.354 | [0.299, 0.409] |
| 51 factors | Ridge | 10.0 | -0.624 | [-1.062, -0.194] | Ridge | 1.0 | 0.508 | [0.386, 0.625] |
| BMA | None | | 0.294 | [0.251, 0.335] | None | | 0.294 | [0.251, 0.335] |
| PCA, $k = 1$ | Ridge | 10.0 | -0.172 | [-0.192, -0.153] | Ridge | 10.0 | -0.177 | [-0.194, -0.159] |
| PCA, $k = 2$ | Lasso | 0.025 | -0.177 | [-0.199, -0.156] | Lasso | 0.025 | -0.177 | [-0.199, -0.156] |
| PCA, $k = 3$ | Lasso | 0.025 | -0.177 | [-0.199, -0.156] | Lasso | 0.025 | -0.177 | [-0.199, -0.156] |
| PCA, $k = 4$ | Lasso | 0.015 | -0.002 | [-0.056, 0.052] | Lasso | 0.015 | 0.025 | [-0.028, 0.080] |
| PCA, $k = 5$ | Lasso | 0.015 | 0.005 | [-0.046, 0.059] | Lasso | 0.015 | 0.042 | [-0.008, 0.095] |
| PCA, $k = 6$ | Ridge | 5.0 | -0.004 | [-0.055, 0.053] | Lasso | 0.015 | 0.233 | [0.187, 0.285] |
| PCA, $k = 7$ | Ridge | 5.0 | 0.099 | [0.028, 0.170] | Ridge | 2.5 | 0.319 | [0.254, 0.383] |
| PCA, $k = 8$ | Ridge | 5.0 | 0.115 | [0.044, 0.187] | Ridge | 2.5 | 0.338 | [0.272, 0.398] |
| RP-PCA, $k = 1$ | Ridge | 10.0 | -0.267 | [-0.564, 0.037] | Lasso | 0.025 | 0.186 | [0.125, 0.249] |
| RP-PCA, $k = 2$ | Ridge | 10.0 | 0.074 | [-0.117, 0.249] | Ridge | 2.5 | 0.334 | [0.287, 0.378] |
| RP-PCA, $k = 3$ | Ridge | 10.0 | -0.236 | [-0.414, -0.051] | Ridge | 1.0 | 0.410 | [0.319, 0.498] |
| RP-PCA, $k = 4$ | Ridge | 10.0 | -0.445 | [-0.787, -0.109] | Ridge | 1.0 | 0.462 | [0.362, 0.557] |
| RP-PCA, $k = 5$ | Ridge | 10.0 | -0.629 | [-1.043, -0.227] | Ridge | 1.0 | 0.436 | [0.323, 0.543] |
| RP-PCA, $k = 6$ | Ridge | 10.0 | -0.764 | [-1.224, -0.311] | Ridge | 1.0 | 0.429 | [0.315, 0.542] |
| RP-PCA, $k = 7$ | Ridge | 10.0 | -0.561 | [-0.979, -0.155] | Ridge | 1.0 | 0.491 | [0.372, 0.607] |
| RP-PCA, $k = 8$ | Ridge | 10.0 | -0.592 | [-1.017, -0.173] | Ridge | 1.0 | 0.496 | [0.374, 0.613] |
| DLFM, $k = 1$ | Lasso | 0.05 | -0.302 | [-0.330, -0.273] | Lasso | 0.05 | -0.214 | [-0.241, -0.189] |
| DLFM, $k = 2$ | Lasso | 0.05 | -0.126 | [-0.149, -0.106] | Lasso | 0.05 | -0.160 | [-0.183, -0.137] |
| DLFM, $k = 3$ | Ridge | 10.0 | -0.096 | [-0.113, -0.077] | Lasso | 0.05 | -0.106 | [-0.131, -0.080] |
| DLFM, $k = 4$ | Ridge | 10.0 | 0.229 | [0.168, 0.287] | Lasso | 0.05 | 0.183 | [0.097, 0.265] |
| DLFM, $k = 5$ | Ridge | 10.0 | 0.338 | [0.304, 0.367] | Lasso | 0.05 | 0.266 | [0.202, 0.326] |
| DLFM, $k = 6$ | Ridge | 10.0 | 0.261 | [0.220, 0.301] | Ridge | 10.0 | 0.231 | [0.106, 0.358] |
| DLFM, $k = 7$ | Ridge | 10.0 | 0.350 | [0.301, 0.396] | Ridge | 10.0 | 0.290 | [0.148, 0.428] |
| DLFM, $k = 8$ | Ridge | 5.0 | 0.435 | [0.350, 0.515] | Ridge | 10.0 | 0.331 | [0.204, 0.457] |

performs significantly better under the GLS regime. While traditional linear dense models and RP-PCA exhibit significant instability or performance degradation when moving from OLS to GLS, the DLFM achieves its peak $R^2$ (0.435) specifically when optimized via GLS.

Figure 5 and Table IV summarize the best OOS $R^2$ results obtained for each method across the space of regularization parameter choices. The 51-factor regression obtained the highest mean $R^2$ (0.508), but with a relatively high degree of uncertainty; the confidence interval shows that RP-PCA performs similarly well. In general, there is a positive relationship between performance

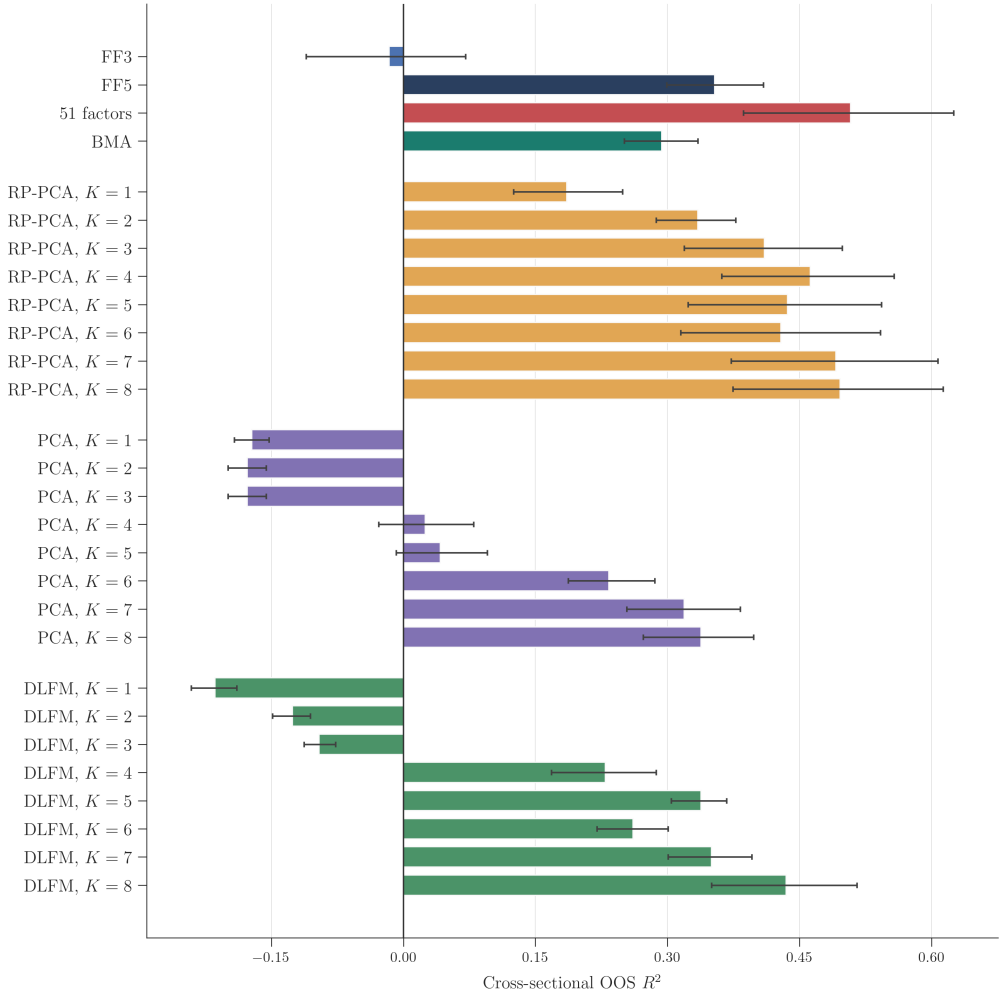



**Figure 5.** Cross-sectional $R^2$ out-of-sample for the different models. RP-PCA, PCA and DLFM models were tested for different latent dimensionality levels, up to $k = 8$. All model results are shown for their highest-performing configurations. Results are averaged over the the range of all data split configurations. Black line bars show the standard deviation of each model across the range of data splits.

and dimensionality when parameters are chosen optimally. By construction, BMA doesn't require tuning of regularization hyperparameters, and is the most "robust" well-performing model in terms of its uncertainty, while the volatile RP-PCA and 51-factor models show greater peak performances. Additionally, the results show that FF3, FF5, BMA, and regular PCA perform similarly well or better out-of-sample than in-sample, which shows that they are less prone to overfitting. Although RP-PCA shows significant overfitting, the model performs well for all but the lowest values of $k$.

To assess how the different models tested perform during the different windows, we plot the out-of-sample cross-sectional $R^2$ for each window.

Figure 6 shows several interesting results. Firstly, when testing between 1980 and 2000, the models perform well, with the RP-PCA and 51 factors models dominating. During this period, all models save for the FF3 model exhibit similar variance. Around the 2008 financial crisis, the Fama-

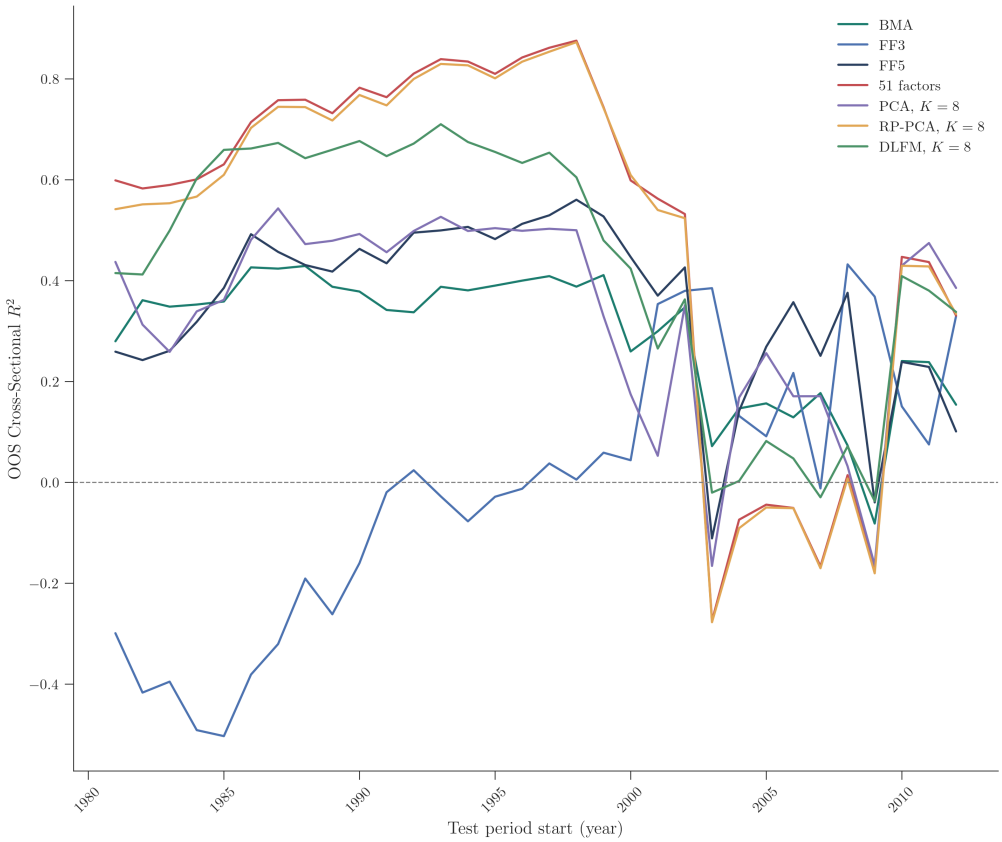



**Figure 6.** Cross-sectional $R^2$ out-of-sample during different windows, meaning that the displayed OOS $R^2$ is obtained when the model is trained on the data before the test period start date, and evaluated on the data after the test period start date. All model results are shown for their highest-performing regularization configurations, including only the best dimensionality choice for all of the latent factor models. All cross-sectional $R^2$ values are averaged over 10 seeds.

French models, BMA, and PCA generally have positive $R^2$ values while the others exhibit significant issues with pricing. However, the models in this work are designed to predict the expected returns rather than the returns at a given point in time. Furthermore, RP-PCA and the regression based on all observable factors achieve remarkably similar performance across time, further indicating that a large number of observable factors can be compressed into a low-dimensional latent space.

**Table IV**

## Overall model performance at optimal configurations

This table shows each optimally configured model's performance with regards to in-sample $R^2$, out-of-sample $R^2$, and MAPE. Which configuration that is considered optimal is chosen based on OOS $R^2$ performance, in accordance with Table III. The displayed 95% confidence interval is taken with respect to the out-of-sample $R^2$ value. All confidence intervals are computed using bootstrapping with $n = 1000$ samples. Model results averaged over 10 seeds.

| Model | IS $R^2$ | OOS $R^2$ | 95% CI | MAPE | Best config |
|---|---|---|---|---|---|
| FF3 | -0.044 | -0.016 | [-0.111, 0.071] | 0.0600 | GLS, Ridge, $\alpha_{\text{reg}} = 10.0$ |
| FF5 | 0.300 | 0.354 | [0.299, 0.409] | 0.0458 | OLS, Ridge, $\alpha_{\text{reg}} = 2.5$ |
| 51 factors | 0.696 | 0.508 | [0.386, 0.625] | 0.0349 | OLS, Ridge, $\alpha_{\text{reg}} = 1.0$ |
| BMA | 0.276 | 0.294 | [0.251, 0.335] | 0.0487 | N/A |
| PCA, $k = 1$ | -0.193 | -0.172 | [-0.192, -0.153] | 0.0649 | GLS, Ridge, $\alpha_{\text{reg}} = 10.0$ |
| PCA, $k = 2$ | -0.215 | -0.177 | [-0.199, -0.156] | 0.0660 | OLS, Lasso, $\alpha_{\text{reg}} = 0.025$ |
| PCA, $k = 3$ | -0.215 | -0.177 | [-0.199, -0.156] | 0.0660 | GLS, Lasso, $\alpha_{\text{reg}} = 0.025$ |
| PCA, $k = 4$ | 0.034 | 0.025 | [-0.028, 0.080] | 0.0557 | OLS, Lasso, $\alpha_{\text{reg}} = 0.015$ |
| PCA, $k = 5$ | 0.049 | 0.042 | [-0.008, 0.095] | 0.0552 | OLS, Lasso, $\alpha_{\text{reg}} = 0.015$ |
| PCA, $k = 6$ | 0.202 | 0.233 | [0.187, 0.285] | 0.0517 | OLS, Lasso, $\alpha_{\text{reg}} = 0.015$ |
| PCA, $k = 7$ | 0.332 | 0.319 | [0.254, 0.383] | 0.0465 | OLS, Ridge, $\alpha_{\text{reg}} = 2.5$ |
| PCA, $k = 8$ | 0.346 | 0.338 | [0.272, 0.398] | 0.0461 | OLS, Ridge, $\alpha_{\text{reg}} = 2.5$ |
| RP-PCA, $k = 1$ | 0.262 | 0.091 | [-0.010, 0.198] | 0.0489 | OLS, Lasso, $\alpha_{\text{reg}} = 0.015$ |
| RP-PCA, $k = 2$ | 0.255 | 0.265 | [0.213, 0.314] | 0.0483 | OLS, Ridge, $\alpha_{\text{reg}} = 2.5$ |
| RP-PCA, $k = 3$ | 0.527 | 0.396 | [0.312, 0.479] | 0.0421 | OLS, Ridge, $\alpha_{\text{reg}} = 1.0$ |
| RP-PCA, $k = 4$ | 0.604 | 0.460 | [0.365, 0.549] | 0.0381 | OLS, Ridge, $\alpha_{\text{reg}} = 1.0$ |
| RP-PCA, $k = 5$ | 0.630 | 0.436 | [0.328, 0.537] | 0.0386 | OLS, Ridge, $\alpha_{\text{reg}} = 1.0$ |
| RP-PCA, $k = 6$ | 0.637 | 0.428 | [0.318, 0.536] | 0.0388 | OLS, Ridge, $\alpha_{\text{reg}} = 1.0$ |
| RP-PCA, $k = 7$ | 0.685 | 0.489 | [0.370, 0.604] | 0.0361 | OLS, Ridge, $\alpha_{\text{reg}} = 1.0$ |
| RP-PCA, $k = 8$ | 0.692 | 0.495 | [0.374, 0.612] | 0.0355 | OLS, Ridge, $\alpha_{\text{reg}} = 1.0$ |
| DLFM, $k = 1$ | -0.137 | -0.214 | [-0.241, -0.189] | 0.0645 | OLS, Lasso, $\alpha_{\text{reg}} = 0.05$ |
| DLFM, $k = 2$ | -0.120 | -0.126 | [-0.149, -0.106] | 0.0619 | GLS, Lasso, $\alpha_{\text{reg}} = 0.05$ |
| DLFM, $k = 3$ | -0.025 | -0.096 | [-0.113, -0.077] | 0.0600 | GLS, Ridge, $\alpha_{\text{reg}} = 10.0$ |
| DLFM, $k = 4$ | 0.271 | 0.229 | [0.168, 0.287] | 0.0481 | GLS, Ridge, $\alpha_{\text{reg}} = 10.0$ |
| DLFM, $k = 5$ | 0.255 | 0.338 | [0.304, 0.367] | 0.0465 | GLS, Ridge, $\alpha_{\text{reg}} = 10.0$ |
| DLFM, $k = 6$ | 0.233 | 0.261 | [0.220, 0.301] | 0.0483 | GLS, Ridge, $\alpha_{\text{reg}} = 10.0$ |
| DLFM, $k = 7$ | 0.289 | 0.350 | [0.301, 0.396] | 0.0453 | GLS, Ridge, $\alpha_{\text{reg}} = 10.0$ |
| DLFM, $k = 8$ | 0.467 | 0.435 | [0.350, 0.515] | 0.0399 | GLS, Ridge, $\alpha_{\text{reg}} = 5.0$ |

# V. Discussion

The findings from the results presented are several. Firstly, the data indicates that sparse factor models, comprising the Fama-French models and BMA, show more consistent performance across regularization parameters. Our findings suggest that the Fama-French three-factor model is inadequate at pricing the cross section, while the five-factor model performs competitively. This is likely due to the nature of the portfolio data set used. As the data set consists of anomaly portfolios, many of the portfolios were constructed specifically to produce large pricing errors in models such as the Fama-French three-factor model and CAPM. As BMA is independent of regularization, it consistently shows out-of-sample $R^2$ values around $0.25 - 0.30$ and shows extremely similar performance in- and out-of-sample.

Secondly, the results strongly indicate a positive relationship between high dimensionality and ability to price the cross-section. As seen in Figure 5, the 51-factor model performs better than its FF5 counterpart, and significantly better than FF3. For the PCA approach, the mean OOS $R^2$ switches from being negative to positive around the inclusion of $k = 6$ model dimensions, and the RP-PCA model performs at its best for $k = 8$ dimensions and at its worst for $k = 1$. However, while the performance improves as dimensionality increases, the returns from larger amounts of latent factors seems to diminish very rapidly; The RP-PCA model obtains powerful results with as few as $k = 4$ latent dimensions, and both the PCA and DLFM methods seem to plateau at approximately $k = 7$. This lends credibility to the conclusion that the true SDF, while likely being dense, doesn't require a particularly high-dimensional model to represent it, reinforcing the conclusions of Kozak et al. (2020). The empirical findings in this article indicate that the factor zoo is redundant, not because the majority of observed factors are irrelevant, but because the pricing information they convey can be compressed into a small set of latent factors. As discussed in Kozak, Nagel, and Santosh (2017), absence of near-arbitrage implies that factors associated with large risk premia should be an important source of asset covariance. Our results provide empirical support for this view. However, in Kozak et al. (2020), the authors assert that typical test assets contain a factor structure such that a few high-variance principal components price the cross-section well. Although standard PCA models achieve performance competitive with the FF5 benchmark at high $k$, performance is substantially improved with RP-PCA and our DLFM model. The cross-sectional prior specified in Kozak et al. (2020) relies on high-eigenvalue principal components corresponding to higher Sharpe ratios than those with low eigenvalues; this prior is not fully supported by our empirical findings. As such, the results suggest that, rather than predicting variance, models more explicitly optimized for return prediction produce stronger results. Our findings contrast with those of Didisheim et al. (2024) and Kelly and Malamud (2021) in the context of pricing the cross-section, as out-of-sample performance plateaus quickly when using latent factor models in our experiments. This suggests that the high-dimensional structure used in these papers can be captured by both linear and nonlinear compressions. However, we acknowledge that the largest models studied in this work are of modest size compared to the regime studied in this literature.

Thirdly, all models tested, save for BMA, depend strongly on the level of regularization. For

both observable and latent factor models, differences in regularization strength or regime can mean the difference between highly inadequate and best-in-class results. This phenomenon is especially prevalent across the latent factor models. Table IV shows that the 51-factors and RP-PCA models perform very well in their best configurations. However, our findings indicate that a thorough analysis of regularization strength must be carried out in order for that excellent performance to be realized. Additionally, the RP-PCA and 51-factor models perform extremely poorly when using GLS, further establishing that these models must be optimized carefully. Even in their best configurations, these models show significant overfitting to noise in-sample, shown by the large discrepancies between in-sample and out-of-sample $R^2$. The observable linear models show significantly improved performance using Ridge regression, while latent factors show no such preference. The contrast between RP-PCA's collapse under GLS and the DLFM's substantial improvement under the same regime is intriguing. By construction, Risk-Premium PCA upweights the directions with low idiosyncratic variance. Once the signal from those directions is extracted, the residuals used in GLS optimization typically become small in magnitude, leading to numerical instability. Although we use Ledoit-Wolf shrinkage to mitigate collapse in the covariance matrix, this doesn't eliminate the issue in our experiments. In contrast, our proposed deep latent factor model heavily favors generalized least squares since its objective is not adversarial to the stability of the GLS optimization process. This implies that the choice between ordinary- and generalized least squares approaches is heavily dependent on how the factors are constructed. This consideration is largely missing from standard Fama-MacBeth regression, where factors are typically taken as observable.

The conclusions drawn from the results, as well as the absolute performance metrics and broader applicability of these findings, must be contextualized by limitations in the data and methods used in this work. The scope of the data is one of non-recency, with the latest data being from December 2016. Furthermore, it is smaller than the data sets used in some of the related literature. There have been several large changes to the financial landscape since then. As such, some of the reported results may not be representative of current model performance. Although several of the prescriptions in Lewellen, Nagel, and Shanken (2010) were followed, such as avoiding size-B/M portfolios, including both OLS and GLS $R^2$ results, and reporting confidence intervals, there was no enforcement that tradable factors should price themselves. As 34 out of the 60 test portfolios correspond to tradable factors, the absolute cross-sectional $R^2$ values may be inflated and should be interpreted with caution. However, the relative ranking of models is more robust, as all models face this bias on the same test assets.

Additionally, there is some indirect look-ahead bias in the process of selecting the best regularization regime for each model, since that decision is made on the out-of-sample performance, which is intentionally chosen to be future data points. As such, the performance estimates of each model are best-case estimates, except for the BMA model, which is not dependent on regularization choices and therefore more robust.

Lastly, the latent models that were evaluated were only tested up to a maximum of 8 latent dimensions. There may be additional conclusions to draw from increasing this ceiling further to

see if the plateauing pattern observed in the results continues as model dimensionality increases further.

# VI. Conclusion

This thesis provides a systematic approach to the "factor zoo" problem by evaluating the structural trade-offs between sparsity, density, and latent representations. By unifying these disparate modeling philosophies within a single empirical framework, we offer a clearer understanding of how to navigate the high-dimensional nature of empirical asset pricing.

Our analysis yields three high-level insights for the literature. First, we provide evidence that the SDF is fundamentally dense but informationally low-rank; while sparse models discard meaningful pricing signals, the relevant information in the broader zoo can be efficiently compressed into a parsimonious set of latent factors. Second, we demonstrate that the choice of estimator is as critical as the choice of factors. While linear models often struggle with the numerical instability of GLS in high dimensions, nonlinear architectures like the DLFM thrive under these regimes, successfully extracting pricing information from residuals that linear benchmarks treat as noise. This finding shows that the choice of least-squares regime is dependent on factor construction. Finally, we highlight BMA as a robust alternative for practitioners, offering the stability of shrinkage without the fragility of hyperparameter tuning.

Ultimately, these findings suggest that "taming the factor zoo" is not a simple task of factor selection, but a multidimensional challenge requiring the integration of latent factor compression and regularization-aware estimation. We provide a framework that allows future researchers to move beyond the search for the "best" individual factors toward a more robust, model-agnostic approach to cross-sectional pricing.

# REFERENCES

Baker, Malcolm and Jeffrey Wurgler (2006): "Investor Sentiment and the Cross-Section of Stock Returns," *The Journal of Finance*, 61 (4), 1645–1680.

Bryzgalova, Svetlana, Jiantao Huang, and Christian Julliard (2023): "Bayesian Solutions for the Factor Zoo: We Just Ran Two Quadrillion Models," *The Journal of Finance*, 78 (1), 487–557.

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

Cochrane, John H. (2011): "Presidential Address: Discount Rates," *The Journal of Finance*, 66 (4), 1047–1108.

Didisheim, Antoine, Shikun (Barry) Ke, Bryan Kelly, and Semyon Malamud (2024): *APT or "AIPT"? The Surprising Dominance of Large Factor Models*, National Bureau of Economic Research.

Fama, Eugene F and Kenneth R French (1993): "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, 33 (1), 3–56.

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

Feng, Guanhao, Stefano Giglio, and Dacheng Xiu (2020): "Taming the Factor Zoo: A Test of New Factors," *The Journal of Finance*, 75 (3), 1327–1370.

Freyberger, Joachim, Andreas Neuhierl, and Michael Weber (2020): "Dissecting Characteristics Nonparametrically," *The Review of Financial Studies*, 33 (5), 2326–2377.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu (2020): "Empirical Asset Pricing via Machine Learning," *The Review of Financial Studies*, 33 (5), 2223–2273.

——— (2021): "Autoencoder asset pricing models," *Journal of Econometrics*, 222 (1), 429–450.

Hansen, Lars Peter (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50 (4), 1029–1054.

Harvey, Campbell R., Yan Liu, and Heqing Zhu (2016): "and the Cross-Section of Expected Returns," *The Review of Financial Studies*, 29 (1), 5–68.

Huang, Dashan, Fuwei Jiang, Jun Tu, and Guofu Zhou (2015): "Investor Sentiment Aligned: A Powerful Predictor of Stock Returns," *The Review of Financial Studies*, 28 (3), 791–837.

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

Kelly, Bryan T. and Semyon Malamud (2021): "The Virtue of Complexity in Machine Learning Portfolios," *SSRN Electronic Journal*.

Kingma, Diederik P. and Jimmy Ba (2015): "Adam: A Method for Stochastic Optimization," in *International Conference on Learning Representations*.

Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh (2017): "Shrinking the Cross Section," *SSRN Electronic Journal*.

——— (2020): "Shrinking the cross-section," *Journal of Financial Economics*, 135 (2), 271–292.

Ledoit, Olivier and Michael Wolf (2004): "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, 88 (2), 365–411.

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

Lewellen, Jonathan, Stefan Nagel, and Jay Shanken (2010): "A skeptical appraisal of asset pricing tests," *Journal of Financial Economics*, 96 (2), 175–194.

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

Onatski, Alexei (2012): "Asymptotics of the principal components estimator of large factor models with weakly influential factors," *Journal of Econometrics*, 168 (2), 244–258.

Sharpe, William F. (1964): "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *The Journal of Finance*, 19 (3), 425–442.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014): "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, 15 (56), 1929–1958.

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