NBER WORKING PAPER SERIES

SHRINKING THE CROSS SECTION

Serhiy Kozak
Stefan Nagel
Shrihari Santosh

Working Paper 24070
http://www.nber.org/papers/w24070

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2017

We thank Svetlana Bryzgalova, Mikhail Chernov, Gene Fama, Stefano Giglio, Amit Goyal, Lars Hansen, Bryan Kelly, Ralph Koijen, Lubos Pastor, Michael Weber, Goufu Zhou, and seminar participants at ASU, City University of Hong Kong, HKUST, Lausanne, Michigan, UCLA, UCSD, Washington University in St. Louis, and Yale for helpful comments and suggestions. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Shrinking the Cross Section
Serhiy Kozak, Stefan Nagel, and Shrihari Santosh
NBER Working Paper No. 24070
November 2017
JEL No. C38,G12

ABSTRACT

We construct a robust stochastic discount factor (SDF) that summarizes the joint explanatory power of a large number of cross-sectional stock return predictors. Our method achieves robust out-of-sample performance in this high-dimensional setting by imposing an economically motivated prior on SDF coefficients that shrinks the contributions of low-variance principal components of the candidate factors. While empirical asset pricing research has focused on SDFs with a small number of characteristics-based factors—e.g., the four- or five-factor models discussed in the recent literature—we find that such a characteristics-sparse SDF cannot adequately summarize the cross-section of expected stock returns. However, a relatively small number of principal components of the universe of potential characteristics-based factors can approximate the SDF quite well.

Serhiy Kozak
University of Michigan
701 Tappan Ave, R5350
Ann Arbor, MI 48104
sekozak@umich.edu

Shrihari Santosh
Van Munching Hall, #4416
7621 Mowatt Lane
College Park, MD 20742
shrisantosh@gmail.com

Stefan Nagel
University of Chicago
Booth School of Business
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
stefan.nagel@chicagobooth.edu

1 Introduction

The empirical asset pricing literature has found a large number of stock characteristics that help predict cross-sectional variation in expected stock returns. Researchers have tried to summarize this variation with factor models that include a small number of characteristics-based factors. That is, they seek to find a *characteristics-sparse* stochastic discount factor (SDF) representation which is linear in only a few such factors. Unfortunately, it seems that as new cross-sectional predictors emerge, these factor models need to be modified and expanded to capture the new evidence: Fama and French (1993) proposed a three factor model, Hou et al. (2015) have moved on to four, Fama and French (2015) to five factors, and Barillas and Shanken (2017) argue for a six-factor model. Even so, research in this area has tested these factor models only on portfolios constructed from a relatively small subset of known cross-sectional return predictors. These papers do not tell us how well characteristics-sparse factor models would do if one confronted them with a much larger set of cross-sectional return predictors—and an examination of this question is statistically challenging due to the high-dimensional nature of the problem.[1]

In this paper, we tackle this challenge. We start by questioning the economic rationale for a characteristics-sparse SDF. If it were possible to characterize the cross-section in terms of a few characteristics, this would imply extreme redundancy among the many dozens of known anomalies. However, upon closer examination, models based on present-value identities or q-theory that researchers have used to interpret the relationship between characteristics and expected returns do not really support the idea that only a few stock characteristics should matter. For example, a present-value identity can motivate why the book-to-market ratio and expected profitability could jointly explain expected returns. *Expected* profitability is not directly observable, though. A large number of observable stock characteristics could potentially be useful for predicting cross-sectional variation in future profitability—and, therefore, also for predicting returns. For these reasons, we seek a method that allows us to estimate the SDF's loadings on potentially dozens or hundreds of characteristics-based factors without imposing that the SDF is necessarily characteristics-sparse.

The conventional approach would be to estimate SDF coefficients with a cross-sectional regression of average returns on covariances of returns and factors. Due to the large number of potential factors, this conventional approach would lead to spurious overfitting. To overcome this high-dimensionality challenge, we use a Bayesian approach with a novel specification of prior beliefs. Asset pricing models of various kinds generally imply that much of the variance of the SDF should be attributable to high-eigenvalue (i.e., high-variance)

[1] Cochrane (2011) refers to this issue as "the multidimensional challenge."

principal components (PCs) of the candidate factor returns. Put differently, first and second moments of returns should be related. Therefore, if a factor earns high expected returns, it must either itself be a major source of variance or load heavily on factors that are major sources of variance. This is true not only in rational expectations models in which pervasive macroeconomic risks are priced but also, under plausible restrictions, in models in which cross-sectional variation in expected returns arises from biased investor beliefs (Kozak et al., 2017).

We construct a prior distribution that reflects these economic considerations. Compared to the naïve OLS estimator, the Bayesian posterior shrinks the SDF coefficients towards zero. Our prior specification shares similarities with the prior in Pástor (2000) and Pástor and Stambaugh (2000). Crucially, however, the degree of shrinkage in our case is *not* equal for all assets. Instead, the posterior applies significantly more shrinkage to SDF coefficients associated with low-eigenvalue PCs. This heterogeneity in shrinkage is consistent with our economic motivation for the prior and it is empirically important as it leads to better out-of-sample (OOS) performance. Our Bayesian estimator is similar to ridge regression—a popular technique in machine learning—but with important differences. The ridge version of the regression of average returns on factor covariances would add a penalty on the sum of squared SDF coefficients (L^2 norm) to the least-squares objective. In contrast, our estimator imposes a penalty based on the maximum squared Sharpe Ratio implied by the SDF—in line with our economic motivation that near-arbitrage opportunities are implausible and likely spurious. This estimator is in turn equivalent to one that minimizes the Hansen and Jagannathan (1997) distance and imposes a penalty on the sum of squared SDF coefficients (L^2 norm).

Our baseline Bayesian approach results in shrinkage of many SDF coefficients to nearly, but not exactly zero. Thus, while the resulting SDF may put low weight on the contribution of many characteristics-based factors, it will not be sparse in terms of characteristics. However, we also want to entertain the possibility that the weight of some of these candidate factors could truly be zero. First, a substantial existing literature focuses on SDFs with just a few characteristics-based factors. While we have argued above that the economic case for this extreme degree of characteristics-sparsity is weak, we still want to entertain it as an empirical hypothesis. Second, we may want to include among the set of candidate factors ones that have not been previously analyzed in empirical studies and which may therefore be more likely to have a zero risk price. For these reasons, we extend our Bayesian method to allow for automatic factor selection, that is, finding a good sparse SDF approximation.

To allow for factor selection, we augment the estimation criterion with an additional penalty on the sum of absolute SDF coefficients (L^1 norm), which is typically used in Lasso

regression (Tibshirani, 1996) and naturally leads to sparse solutions. Our combined specification employs both L^1 and L^2 penalties, similar to the elastic net technique in machine learning. This combined specification achieves our two primary goals: (i) regularization based on an economically motivated prior, and (ii) it allows for sparsity by setting some SDF coefficients to zero. We pick the strength of penalization to maximize the (cross-validated) cross-sectional OOS R^2.

In our empirical application of these methods, we first look at a familiar setting in which we know the answer that the method should deliver. We focus on the well known 25 ME/BM sorted portfolios from Fama and French (1993). We show that our method automatically recovers an SDF that is similar to the one based on the SMB and HML factors constructed intuitively by Fama and French (1993).

We then move on to a more challenging application in which we examine 50 well known anomaly portfolios, portfolios based on 80 lagged returns and financial ratios provided by Wharton Research Data Services (WRDS), as well as more than a thousand powers and interactions of these characteristics. We find that: (i) the L^2-penalty-only based method (our Bayesian approach) finds robust non-sparse SDF representations that perform well OOS; therefore, if sparsity is not required, our Bayesian method provides a natural starting point for most applications; (ii) L^1-penalty-only based methods often struggle in delivering good OOS performance in high-dimensional spaces of base characteristics; and (iii) sparsity in the space of characteristics is limited in general, even with our dual-penalty method, suggesting little redundancy among the anomalies represented in our data set. Thus, in summary, achieving robustness requires shrinkage of SDF coefficients, but restricting the SDF to just a few characteristics-based factors does not adequately capture the cross-section of expected returns.

Interestingly, the results on sparsity are very different if we first transform the characteristics-portfolio returns into their PCs before applying our dual-penalty method. A sparse SDF that includes a few of the high-variance PCs delivers a good and robust out-of-sample fit of the cross-section of expected returns. Little is lost, in terms of explanatory power, by setting the SDF coefficients of low-variance PCs to zero. This finding is robust across our three primary sets of portfolios and the two extremely high-dimensional datasets that include the power and interactions of characteristics. No similarly sparse SDF based on the primitive characteristics-based factors can compete in terms of OOS explanatory power with a sparse PC-based SDF.

That there is much greater evidence for sparsity in the space of principal component portfolios returns than in the original space of characteristics-based portfolio returns is economically sensible. As we argued earlier, there are no compelling reasons why one should be

able to summarize the cross-section of expected returns with just a few stock characteristics. In contrast, a wide range of asset pricing models implies that a relatively small number of high-variance PCs should be sufficient to explain most of the cross-sectional variation in expected returns. As Kozak et al. (2017) discuss, absence of near-arbitrage opportunities implies that factors earning substantial risk premia must be a major source of co-movement—in models with rational investors as well as ones that allow for investors with biased beliefs. Since typical sets of equity portfolio returns have a strong factor structure dominated by a small number of high-variance PCs, a sparse SDF that includes some of the high-variance PCs should then be sufficient to capture these risk premia.

In summary, our results suggest that the empirical asset-pricing literature's multi-decade quest for a sparse characteristics-based factor model (e.g., with 3, 4, or 5 characteristics-based factors) is ultimately futile. There is just not enough redundancy among the large number of cross-sectional return predictors for such a characteristics-sparse model to adequately summarize pricing in the cross-section. As a final test, we confirm the statistical significance of this finding in an out-of-sample test. We estimate the SDF coefficients, and hence the weights of the mean-variance efficient (MVE) portfolio, based on data until the end of 2004. We then show that this MVE portfolio earns an economically large and statistically highly significant abnormal return relative to the Fama and French (2016) 5-factor model in the out-of-sample period 2005–2016, allowing us to reject the hypothesis that the 5-factor model describes the SDF.

Conceptually, our estimation approach is related to research on mean-variance portfolio optimization in the presence of parameter uncertainty. SDF coefficients of factors are proportional to their weights in the MVE portfolio. Accordingly, our L^2-penalty estimator of SDF coefficients maps into L^2-norm constrained MVE portfolio weights obtained by DeMiguel et al. (2009). Moreover, as DeMiguel et al. (2009) show, and as can be readily seen from the analytic expression of our estimator, portfolio optimization under L^2-norm constraints on weights shares similarities with portfolio optimization with a covariance matrix shrunk towards the identity matrix as in Ledoit and Wolf (2004a). However, despite some similarity of the solutions, there are important differences. First, our L^2-penalty results in level shrinkage of all SDF coefficients towards zero. This would not be the case with a shrunk covariance matrix. Second, in covariance matrix shrinkage approaches, the optimal amount of shrinkage would depend on the size of the parameter uncertainty in covariance estimation. Higher uncertainty about the covariance matrix parameters would call for stronger shrinkage. In contrast, our estimator is derived under the assumption that the covariance matrix is *known* (we use daily returns to estimate covariances precisely) and means are unknown. Shrinkage in our case is due to this uncertainty about means and our economically motivated assump-

tion that ties means to covariances in a particular way. Notably, the amount of shrinkage required in our case of uncertain means is significantly higher than in the case of uncertain covariances. In fact, when we allow for uncertainty in both means and covariances, we find that covariance uncertainty has negligible impact on coefficient estimates once uncertainty in means is accounted for.

Our paper contributes to an emerging literature that applies machine learning techniques in asset pricing to deal with the high-dimensionality challenge. Kelly et al. (2017) show how to perform dimensionality reduction of the characteristics space. They extend PCA and Projected-PCA (Fan et al., 2016) to allow for time-varying factor loadings and apply it to extract common latent factors from the cross-section of individual stock returns. Their method explicitly maps these latent factors to principal components of characteristic-managed portfolios (under certain conditions). Kelly et al. (2017) and Kozak et al. (2017) further show that an SDF constructed using few such dominant principal components prices the cross-section of expected returns well. While Kelly et al. (2017) focuses purely on a factor variance criterion in selecting the factors, we exploit the asset pricing link between expected returns and covariances and use information from both moments in constructing an SDF.

DeMiguel et al. (2017), Freyberger et al. (2017) and Feng et al. (2017) focus on characteristics-based factor selection in Lasso-style estimation with L^1-norm penalties. Their findings are suggestive of a relatively high degree of redundancy among cross-sectional stock return predictors. Yet, as our results show, for the purposes of SDF estimation with characteristics-based factors, a focus purely on factor selection with L^1-norm penalties is inferior to an approach with L^2-norm penalties that shrinks SDF coefficients towards zero to varying degrees, but does not impose sparsity on the SDF coefficient vector. This is in line with results from the statistics literature where researchers have noted that Lasso does not perform well when regressors are correlated and that ridge regression (with L^2-norm penalty) or elastic net (with a combination of L^1- and L^2-norm penalties) delivers better prediction performance than Lasso in these cases (Tibshirani, 1996; Zou and Hastie, 2005). Since many of the candidate characteristics-based factors in our application have substantial correlation, it is to be expected that an L^1-norm penalty alone will lead to inferior prediction performance. For example, instead of asking the estimation procedure to choose between the value factor and the correlated long-run-reversals factor for the sake of sparsity in terms of characteristics, there appears to be value, in terms of explaining the cross-section of expected returns, in extracting the predictive information common to both.

Another important difference between our approach and much of this recent machine learning literature in asset pricing lies in the objective. Many papers (e.g., Freyberger et al.

(2017); Huerta et al. (2013); Moritz and Zimmermann (2016); Tsai et al. (2011), with the exception of Feng et al. (2017)) focus on estimating risk *premia*, i.e., the extent to which a stock characteristic is associated with variation in expected returns. In contrast, we focus on estimation of risk *prices*, i.e., the extent to which the factor associated with a characteristic helps price assets by contributing to variation in the SDF. The two perspectives are not the same because a factor can earn a substantial risk premium simply by being correlated with the pricing factors in the SDF, without being one of those pricing factors. Our objective is to characterize the SDF, hence our focus on risk prices. This difference in objective from much of the existing literature also explains why we pursue a different path in terms of methodology. While papers focusing on risk premia can directly apply standard machine learning methods to the cross-sectional regressions or portfolio sorts used for risk premia estimation, a key contribution of our paper is to adapt the objective function of standard ridge and Lasso estimators to be suitable for SDF estimation and consistent with our economically motivated prior.

Finally, our analysis is also related to papers that consider the statistical problems arising from researchers' data mining of cross-sectional return predictors. The focus of this literature is on assessing the statistical significance of individual characteristics-based factors when researchers may have tried many other factors as well. Green et al. (2017) and Harvey et al. (2015) adjust significance thresholds to account for such data mining. In contrast, rather than examining individual factors in isolation, we focus on assessing the *joint* pricing role of a large number of factors and the potential redundancy among the candidate factors. While our tests do not directly adjust for data mining, our approach implicitly includes some safeguards against data-mined factors. First, for data-mined factors there is no reason for the (spurious in-sample) mean return to be tied to covariances with major sources of return variance. Therefore, by imposing a prior that ties together means and covariances, we effectively downweight data-mined factors. Second, our final test using the SDF-implied MVE portfolio is based on data from 2005–2016, a period that starts after or overlaps very little with the sample period used in studies that uncovered the anomalies (McLean and Pontiff, 2016).

2 Asset Pricing with Characteristics-Based Factors

We start by laying out the basic asset pricing framework that underlies characteristics-based factor models. We first describe this framework in terms of population moments, leaving aside estimation issues for now. Building on this, we can then proceed to describe the estimation problem and our proposed approach for dealing with the high-dimensionality of

this problem.

For any point in time t, let R_t denote an $N \times 1$ vector of excess returns for N stocks. Typical reduced-form factor models express the SDF as a linear function of excess returns on stock portfolios. Along the lines of Hansen and Jagannathan (1991), one can find an SDF in the linear span of excess returns,

$$M_t = 1 - b'_{t-1} \left(R_t - \mathbb{E} R_t \right), \tag{1}$$

by solving for the $N \times 1$ vector of SDF loadings b_{t-1} that satisfies the conditional pricing equation

$$\mathbb{E}_{t-1}[M_t R_t] = 0. \tag{2}$$

2.1 Characteristics-based factor SDF

Characteristics-based asset pricing models parametrize the SDF loadings as

$$b_{t-1} = Z_{t-1} b, \tag{3}$$

where Z_{t-1} is an $N \times H$ matrix of asset characteristics and b is an $H \times 1$ vector of time-invariant coefficients. Without further restrictions, this representation is without loss of generality.[2] To obtain models with empirical content, researchers search for a few measurable asset attributes that approximately span b_{t-1}. For example, Fama and French (1993) use two characteristics: market capitalization and the book-to-market equity ratio. Our goal is to develop a statistical methodology that allows us to entertain a large number of candidate characteristics and estimate their coefficients b in such a high-dimensional setting.

Plugging eq. (3) into eq. (1) delivers an SDF that is in the linear span of the H characteristics-based factor returns, $F_t = Z'_{t-1} R_t$, that can be created based on stock characteristics, i.e.,

$$M_t = 1 - b' \left(F_t - \mathbb{E} F_t \right). \tag{4}$$

In line with much of the characteristics-based factor model literature, we focus on the unconditional asset pricing equation,

$$\mathbb{E}\left[M_t F_t \right] = 0, \tag{5}$$

[2]For example, at this general level, the SDF coefficient of an asset could serve as the "characteristic," $Z_{t-1} = b_{t-1}$, with $b = 1$. That we have specified the relationship between b_{t-1} and characteristics as linear is generally not restrictive as Z_{t-1} could also include nonlinear functions of some stock characteristics. Similarly, by working with cross-sectionally centered and standardized characteristics, we focus on cross-sectional variation, but it would be straightforward to generalize to Z_t that includes variables with time-series dynamics that could capture time-variation in conditional moments.

where the factors F_t serve simultaneously as the assets whose returns we are trying to explain as well as the candidate factors that can potentially enter as priced factors into the SDF.

In our empirical work we cross-sectionally demean each column of Z so that the factors in F_t are returns on zero-investment long-short portfolios. Typical characteristics-based factor models in the literature add a market factor to capture the level of the equity risk premium, while the long-short characteristics factors explain cross-sectional variation. In our specification, we focus on understanding the factors that help explain these cross-sectional differences and we do not explicitly include a market factor, but we orthogonalize the characteristics-based factors with respect to the market factor. This is equivalent, in terms of the effect on pricing errors, to including a market factor in the SDF. It is therefore useful here to think of the elements of F as factors that have been orthogonalized. In our empirical analysis, we also work with factors orthogonalized with respect to the market return.

With knowledge of population moments, we could now solve eq. (4) and eq. (5) for the SDF coefficients

$$b = \Sigma^{-1}\mathbb{E}\left(F_t\right), \tag{6}$$

where $\Sigma \equiv \mathbb{E}\left[\left(F_t - \mathbb{E}F_t\right)\left(F_t - \mathbb{E}F_t\right)'\right]$. Rewriting this expression as

$$b = \left(\Sigma\Sigma\right)^{-1}\Sigma\mathbb{E}\left(F_t\right) \tag{7}$$

shows that the SDF coefficients can be interpreted as the coefficients in a cross-sectional regression of the expected asset returns to be explained by the SDF, which in this case are the H elements of $\mathbb{E}\left(F_t\right)$, on the H columns of covariances of each factor with the other factors and with itself.

In practice, without knowledge of population moments, estimating the SDF coefficients by running such a cross-sectional regression in sample would result in overfitting of noise, with the consequence of poor out-of-sample performance, unless H is small. Since SDF coefficients are also weights of the mean-variance-efficient (MVE) portfolio, the difficulty of estimating SDF coefficients with big H is closely related to the well-known problem of estimating the weights of the MVE portfolio when the number of assets is large. The approach we propose in Section 3 is designed to address this problem.

2.2 Sparsity in characteristics-based factor returns

Much of the existing characteristics-based factor model literature has sidestepped this high-dimensionality problem by focusing on models that include only a small number of factors. We will refer to such models as *characteristics-sparse* models. Whether such a *characteristics-*

sparse model can adequately describe the SDF in a cross-section with a large number of stock characteristics is a key empirical question that we aim to answer in this paper.

Before going into the empirical methods and analysis to tackle these questions, it is useful to first briefly discuss what we might expect regarding characteristics-sparsity of the SDF based on some basic economic arguments. While the literature's focus on characteristics-sparse factor models has been largely ad-hoc, there have been some attempts to motivate the focus on a few specific characteristics.

One such approach is based on the q-theory of firm investment. Similar predictions also result from present-value identity relationships like those discussed in Fama and French (2015) or Vuolteenaho (2002). To provide a concrete example, we briefly discuss the two-period q-theory model in Lin and Zhang (2013). The key idea of the model is that an optimizing firm should choose investment policies such that it aligns expected returns (cost of capital) and profitability (investment payoff). In the model, firms take the SDF as given when making real investment decisions. A firm has a one-period investment opportunity. For an investment I_0 the firm will make profit ΠI_0. The firm faces quadratic adjustment costs with marginal cost cI_0 and the investment fully depreciates after one period. Every period, the firm has the objective

$$\max_{I_0} \mathbb{E}[M\Pi I_0] - I_0 - \frac{c}{2} I_0^2. \tag{8}$$

Taking this SDF as given and using the firm's first-order condition, $I_0 = \frac{1}{c}\left(\mathbb{E}[M\Pi] - 1\right)$, we can compute a one-period expected return,

$$\mathbb{E}\left[R\right] = \mathbb{E}\left(\frac{\Pi}{\mathbb{E}\left[M\Pi\right]}\right) = \frac{\mathbb{E}\left[\Pi\right]}{1 + cI_0}. \tag{9}$$

For example, a firm with high expected return, and hence high cost of capital, must either have high profitability or low investment, or a combination thereof. By the same token, expected profitability and investment jointly reveal whether the firm has high or low loadings on the SDF. For this reason, factors for which stocks' weights are based on expected profitability and investment help capture the factors driving the SDF. The model therefore implies a sparse characteristic-based factor model with two factors: expected profitability $\mathbb{E}\left[\Pi\right]$ and investment I_0, which seems to provide a partial motivation for the models in Hou et al. (2015) and Fama and French (2015).

In practice, however, neither expected profitability nor (planned) investment are observable. The usual approach is to use proxies, such as lagged profitability and lagged investment as potential predictors of unobserved quantities. Yet many additional characteristics are

likely relevant for capturing expected profitability and planned investment and, therefore, expected returns. Moreover, considering that the model above is a vast simplification of reality to begin with, many more factors are likely to be required to approximate an SDF of a more realistic and complex model. The bottom line is that, in practice, q-theory does not necessarily provide much economic reason to expect sparse SDFs in the space of *observable* characteristics.

For this reason, we pursue an approach that does not impose that the SDF is necessarily characteristics-sparse. Moreover, it leads us to seek a method that can accommodate an SDF that involves a potentially very large number of characteristics-based factors, but at the same time, still ensures good out-of-sample performance and robustness against in-sample overfitting. At the same time, we would also like our method to be able to handle cases in which some of the candidate factors are not contributing to the SDF at all. This situation may be particularly likely to arise if the analysis includes characteristics that are not known, from prior literature, to predict returns in the cross-section. It could also arise if there is truly some redundancy among the cross-sectional return predictors documented in the literature. To accommodate these cases, we want our approach to allow for the possibility of sparsity, but without necessarily requiring sparsity to perform well out of sample. This will then allow us to assess the degree of sparsity empirically.

2.3 Sparsity in principal components of characteristics-based factor returns

While there are not strong economic reasons to expect characteristics-sparsity of the SDF, one may be able to find rotations of the characteristics factor data that admit, at least approximately, a sparse SDF representation. Motivated by the analysis in Kozak et al. (2017), we consider sparse SDF representations in the space of principal components (PCs) of characteristic-based factor returns.

Based on the eigendecomposition of the factor covariance matrix,

$$\Sigma = QDQ' \qquad \text{with} \quad D = \text{diag}(d_1, d_2, ..., d_H), \tag{10}$$

where Q is the matrix of eigenvectors of Σ and D is the diagonal matrix of eigenvalues ordered in decreasing magnitude, we can construct PC factors

$$P_t = Q'F_t. \tag{11}$$

Using all PCs, and with knowledge of population moments, we could express the SDF as

$$M_t = 1 - b_P' \left(P_t - \mathbb{E} P_t \right), \qquad \text{with} \quad b_P = D^{-1} \mathbb{E}[P_t]. \tag{12}$$

In Kozak et al. (2017) we argue that absence of near-arbitrage (extremely high Sharpe Ratios) implies that factors earning substantial risk premium must be a major source of co-movement. This conclusion obtains under very mild assumptions and applies equally to "rational" and "behavioral" models. Furthermore, for typical sets of test assets, returns have a strong factor structure dominated by a small number of PCs with the highest variance (or eigenvalues d_j). Under these two conditions, an SDF with a small number of these high-variance PCs as factors should explain most of the cross-sectional variation in expected returns. Motivated by this theoretical result, we explore empirically whether an SDF sparse in PCs can be sufficient to describe the cross-section of expected returns and we compare it, in terms of their pricing performance, with SDFs that are sparse in characteristics.

3 Methodology

Consider a sample with size T. We denote

$$\bar{\mu} = \frac{1}{T} \sum_{t=1}^{T} F_t, \tag{13}$$

$$\overline{\Sigma} = \frac{1}{T} \sum_{t=1}^{T} \left(F_t - \bar{\mu} \right) \left(F_t - \bar{\mu} \right)'. \tag{14}$$

A natural, but naïve, GMM estimator of the coefficients b of the SDF in eq. (4), could be constructed based on the sample moment conditions

$$\mu - \frac{1}{T} \sum_{t=1}^{T} F_t = 0, \tag{15}$$

$$\frac{1}{T} \sum_{t=1}^{T} M_t F_t = 0. \tag{16}$$

The resulting estimator is the sample version of eq. (6),[3]

$$\hat{b} = \overline{\Sigma}^{-1} \bar{\mu}. \tag{17}$$

[3]When $T < H$ we use Moore-Penrose pseudoinverse of the covariance matrix.

However, unless H is very small relative to T, this naïve estimator yields very imprecise estimates of b. The main source of imprecision is the uncertainty about μ. Along the same lines as for the population SDF coefficients in Section 2.1, the estimator $\hat{b}$ effectively results from regressing factor means on the covariances of these factors with each other. As is generally the case in expected return estimation, the factor mean estimates are imprecise even with fairly long samples of returns. In a high-dimensional setting with large H, the cross-sectional regression effectively has a large number of explanatory variables. As a consequence, the regression will end up spuriously overfitting the noise in the factor means, resulting in a very imprecise $\hat{b}$ estimate and bad out-of-sample performance. Estimation uncertainty in the covariance matrix can further exacerbate the problem, but as we discuss in greater detail in Appendices A and B, the main source of fragility in our setting are the factor means, not the covariances.

To avoid spurious overfitting, we bring in economically motivated prior beliefs about the factors' expected returns. If the prior beliefs are well-motivated and truly informative, this will help reduce the (posterior) uncertainty about the SDF coefficients. In other words, bringing in prior information then *regularizes* the estimation problem sufficiently to produce robust estimates that perform well in out-of-sample prediction. We first start with prior beliefs that shrink the SDF coefficients away from the naïve estimator in eq. (17), but without imposing sparsity. We then expand the framework to allow for some degree of sparsity as well.

3.1 Shrinkage estimator

To focus on uncertainty about factor means, the most important source of fragility in the estimation, we proceed under the assumption that Σ is known. Consider the family of priors,

$$\mu \sim \mathcal{N}\left(0,\, \frac{\kappa^2}{\tau}\Sigma^{\eta}\right), \tag{18}$$

where $\tau = \text{tr}\left[\Sigma\right]$ and κ is a constant controlling the "scale" of μ that may depend on τ and H. As we will discuss, this family encompasses priors that have appeared in earlier asset pricing studies, albeit not in a high-dimensional setting. At this general level, this family of priors can broadly capture the notion—consistent with a wide class of asset pricing theories—that first moments of factor returns have some connection to their second moments. Parameter η controls the "shape" of the prior. It is the key parameter for the economic interpretation of the prior because it determines how exactly the relationship between first and second moments of factor returns is believed to look like under the prior.

To understand the economic implications of particular values of η, it is useful to consider the PC portfolios $P_t = Q'F_t$ with $\Sigma = QDQ'$ that we introduced in Section 2.3. Expressing the family of priors (18) in terms of PC portfolios we get

$$\mu_P \sim \mathcal{N}\left(0, \frac{\kappa^2}{\tau} D^{\eta}\right). \tag{19}$$

For the distribution of Sharpe Ratios of the PCs, we obtain

$$D^{-\frac{1}{2}}\mu_P \sim \mathcal{N}\left(0, \frac{\kappa^2}{\tau} D^{\eta-1}\right). \tag{20}$$

We can evaluate the plausibility of assumptions about η by considering the implied prior beliefs about Sharpe Ratios of small-eigenvalue PCs. For typical sets of asset returns, the distribution of eigenvalues is highly skewed: a few high-eigenvalue PCs account for most of the return variance, many PCs have much smaller eigenvalues, and the smallest eigenvalues of high-order PCs are tiny.

This fact about the distribution of eigenvalues immediately makes clear that the assumption of $\eta = 0$ (as, e.g., in Harvey et al. (2008)) is economically implausible. In this case, the mean Sharpe Ratio of a PC factor in eq. (20) is inversely related to the PC's eigenvalue. Therefore, the prior implies that the expected Sharpe Ratios of low-eigenvalue PCs explode towards infinity. In other words, $\eta = 0$ would imply existence of *near-arbitrage* opportunities. As Kozak et al. (2017) discuss, existence of near-arbitrage opportunities is not only implausible in rational expectations models, but also in models in which investors have biased beliefs, as long as some arbitrageurs are present in the market.

Pástor (2000) and Pástor and Stambaugh (2000) work with $\eta = 1$. This assumption is more plausible in the sense that it is consistent with absence of near-arbitrage opportunities. However, as eq. (20) makes clear, $\eta = 1$ implies that Sharpe Ratios of low-eigenvalue PCs are expected to be of the same magnitude as Sharpe Ratios of high-eigenvalue PCs. We do not view this as economically plausible. For instance, in rational expectations models in which cross-sectional differences in expected returns arise from exposure to macroeconomic risk factors, risk premia are typically concentrated in one or a few common factors. This means that Sharpe Ratios of low-eigenvalue PCs should be smaller than those of the high-eigenvalue PCs that are the major source of risk premia. Kozak et al. (2017) show that a similar prediction also arises in plausible "behavioral" models in which investors have biased beliefs. Kozak et al. argue that to be economically plausible, such a model should include arbitrageurs in the investor population and it should have realistic position size limits (e.g., leverage constraints or limits on short selling) for the biased-belief investors (who are likely

to be less sophisticated). As a consequence, biased beliefs can only have substantial pricing effects in the cross-section if these biased beliefs align with high-eigenvalue PCs; otherwise arbitrageurs would find it too attractive to aggressively lean against the demand from biased investors, leaving very little price impact. To the extent it exists, mispricing then appears in the SDF mainly through the risk prices of high-eigenvalue PCs. Thus, within both classes of asset pricing models, we would expect Sharpe Ratios to be increasing in the eigenvalue, which is inconsistent with $\eta \leq 1$.

Moreover, the portfolio that an unconstrained rational investor holds in equilibrium should have finite portfolio weights. Indeed, realistic position size limits for the biased-belief investors in Kozak et al. (2017) discussed above translate into finite equilibrium arbitrageur holdings, and therefore, finite SDF coefficients. Our prior should be consistent with this prediction. Since the optimal portfolio weights of a rational investor and SDF coefficients are equivalent, we want a prior which ensures that $b'b$ remains bounded. A minimal requirement for this to be true is that $\mathbb{E}[b'b]$ remains bounded. With $b = \Sigma^{-1}\mu$, the decomposition $\Sigma = QDQ'$, and the prior (18), we can show

$$\mathbb{E}[b'b] = \frac{\kappa^2}{\tau} \sum_{i=1}^{H} d_i^{\eta-2}, \tag{21}$$

where d_i are the eigenvalues on the diagonal of D. Since the lowest eigenvalue, d_H, in a typical asset return data set is extremely close to zero, the corresponding summation term $d_i^{\eta-2}$ is extremely big if $\eta < 2$. In other words, with $\eta < 2$ the prior would imply that the optimal portfolio of a rational investor is likely to place huge bets on the lowest-eigenvalue PCs. Setting $\eta \geq 2$ avoids such unrealistic portfolio weights. To ensure the prior is plausible, but at the same is also the least restrictive ("flattest") Bayesian prior which deviates as little as possible from more conventional prior assumptions like those in Pástor and Stambaugh's work, we set $\eta = 2$.

To the best of our knowledge, this prior specification is novel in the literature, but, as we have argued, there are sound economic reasons for this choice. Based on this assumption, we get an i.i.d. prior on SDF coefficients, $b \sim \mathcal{N}\left(0, \frac{\kappa^2}{\tau} I\right)$. Combining these prior beliefs with information about sample means $\bar{\mu}$ from a sample with size T, assuming a multivariate-normal likelihood, we obtain the posterior mean of b

$$\hat{b} = \left(\Sigma + \gamma I\right)^{-1} \bar{\mu}, \tag{22}$$

where $\gamma = \frac{\tau}{\kappa^2 T}$. The posterior variance of b is given by

$$\text{var}(b) = \frac{1}{T}(\Sigma + \gamma I)^{-1}, \tag{23}$$

which we use in Section 4 to construct confidence intervals.

3.1.1 Economic interpretation

To provide an economic interpretation of what this estimator does, it is convenient to consider a rotation of the original space of returns into the space of principal components. Expressing the SDF based on the estimator (22) in terms of PC portfolio returns, $P_t = Q'F_t$, with coefficients $\hat{b}_P = Q'\hat{b}$, we obtain a vector with elements

$$\hat{b}_{P,j} = \left(\frac{d_j}{d_j + \gamma}\right)\frac{\bar{\mu}_{P,j}}{d_j}, \tag{24}$$

Compared with the naïve exactly identified GMM estimator from eq. (17), which would yield SDF coefficients for the PCs of

$$\hat{b}^{\text{ols}}_{P,j} = \frac{\bar{\mu}_{P,j}}{d_j}, \tag{25}$$

our Bayesian estimator (with $\gamma > 0$) shrinks the SDF coefficients towards zero with the shrinkage factor $d_j/(d_j+\gamma) < 1$. Most importantly, the shrinkage is stronger the smaller the eigenvalue d_j associated with the PC. The economic interpretation is that we judge as implausible that a PC with low eigenvalue could contribute substantially to the volatility of the SDF and hence to the overall maximum squared Sharpe Ratio. For this reason, the estimator shrinks the SDF coefficients of these low-eigenvalue PCs particularly strongly. In contrast, with $\eta = 1$ in the prior—which we have argued earlier is economically implausible—the estimator would shrink the SDF coefficients of all PCs equally.

3.1.2 Representation as a penalized estimator

We now show that our Bayesian estimator maps into a penalized estimator that resembles estimators common in the machine learning literature. If we maximize the model cross-sectional R^2 subject to a penalty on the model-implied maximum squared Sharpe ratio $\gamma b'\Sigma b$,

$$\hat{b} = \arg\min_b \left\{(\bar{\mu} - \Sigma b)'(\bar{\mu} - \Sigma b) + \gamma b'\Sigma b\right\}, \tag{26}$$

the problem leads to exactly the same solution as in eq. (22). Equivalently, minimizing the model HJ-distance (Hansen and Jagannathan, 1991) subject to an L^2 norm penalty $\gamma b'b$,

$$\hat{b} = \arg\min_{b} \left\{ (\bar{\mu} - \Sigma b)' \Sigma^{-1} (\bar{\mu} - \Sigma b) + \gamma b'b \right\}, \tag{27}$$

leads again to the same solution as in eq. (22). Looking at this objective again in terms of factor returns that are transformed into their principal components, one can see intuitively how the penalty in this case induces shrinkage effects concentrated on low-eigenvalue PCs in the same way as the prior beliefs do in the case of the Bayesian estimator above. Suppose the estimation would shrink towards zero the coefficient $\hat{b}_{P,j}$ on a low-eigenvalue PC. This would bring a benefit in terms of the penalty, but little cost because for a given magnitude of the SDF coefficient, a low eigenvalue PC contributes only very little to SDF volatility and so shrinking its contribution has little effect on the HJ distance. In contrast, shrinking the coefficient on a high-eigenvalue PC by the same magnitude would bring a similar penalty benefit, but at a much larger cost because it would remove a major source of SDF volatility from the SDF. As a consequence, the estimation tilts towards shrinking SDF coefficients of low-eigenvalue PCs.

Equations (26) and (27) resemble ridge regression, a popular technique in machine learning (e.g., see Hastie et al., 2011), but with some important differences. A standard ridge regression objective function would impose a penalty on the L^2-norm of coefficients, $b'b$ in eq. (26), or, equivalently, weight the pricing errors with the identity matrix instead of Σ^{-1} in eq. (27). One can show that this standard ridge regression would correspond to a prior with $\eta = 3$, which would imply even more shrinkage of low-eigenvalue PCs than with our prior of $\eta = 2$. However, the estimator one obtains from a standard ridge approach is not invariant to how the estimation problem is formulated. For example, if one estimates factor risk premia λ in a beta-pricing formulation of the model, minimizing $(\bar{\mu} - I\lambda)' (\bar{\mu} - I\lambda)$ subject to a standard ridge penalty on $\lambda'\lambda$, the resulting estimator corresponds to a prior with $\eta = 1$, that, as we have argued, is not economically plausible. In contrast, in our approach the estimator is pinned down by the asset pricing equation (refeq:AP-eq) combined with the economically motivated prior (18).

3.2 Sparsity

The method that we have presented so far deals with the high-dimensionality challenge by shrinking SDF coefficients towards zero, but none of the coefficients are set to exactly zero. In other words, the solution we obtain is not sparse. As we have argued in Section 2, the economic case for extreme sparsity with characteristics-based factors is weak. However, it

may be useful to allow for the possibility that some factors are truly redundant in terms of their contribution to the SDF. Moreover, there are economic reasons to expect that a representation of the SDF that is sparse in terms of PCs could provide a good approximation.

For these reasons, we now introduce an additional L^1 penalty $\gamma_1 \sum_{j=1}^{H} |b_j|$ in the penalized regression problem given by eq. (27). The approach is motivated by Lasso regression and *elastic net* (Zou and Hastie, 2005), which combines Lasso and ridge penalties. Due to the geometry of the L^1 norm, it leads to some elements of $\hat{b}$ being set to zero, that is, it accomplishes sparsity and automatic factor selection. The degree of sparsity is controlled by the strength of the penalty. Combining both L^1 and L^2 penalties, our estimator solves the problem:[4]

$$\hat{b} = \arg\min_{b} \left(\bar{\mu} - \Sigma b\right)' \Sigma^{-1} \left(\bar{\mu} - \Sigma b\right) + \gamma_2 b'b + \gamma_1 \sum_{i=1}^{H} |b_i| \,. \tag{28}$$

This dual-penalty method enjoys much of the economic motivation behind the L^2-penalty-only method with an added benefit of potentially delivering sparse SDF representations. We can control the degree of sparsity by varying the strength of the L^1 penalty and the degree of economic shrinkage by varying the L^2 penalty.

Despite the visual similarities, there are important, economically motivated differences between our method and a standard elastic net estimator. First, we minimize the HJ-distance instead of minimizing (unweighted) pricing errors. Second, unlike in typical elastic net applications, we do not normalize or center variables: the economic structure of our setup imposes strict restrictions between means and covariances and leaves no room for intercepts or arbitrary normalizations.

While we will ultimately let the data speak about the optimal values of the penalties γ_1 and γ_2, there is reason to believe that completely switching off the L^2 penalty and focusing purely on Lasso-style estimation would not work well in this asset-pricing setting. Lasso is known to suffer from relatively poor performance compared with ridge and elastic net when regressors are highly correlated (Tibshirani, 1996; Zou and Hastie, 2005). An L^2 penalty leads the estimator to shrink coefficients of correlated predictors towards each other, allowing them to borrow strength from each other (Hastie et al., 2011). In the extreme case of k identical predictors, they each get identical coefficients with $1/k$-th the size that any single one would get if fit alone. The L^1 penalty, on the other hand, ignores correlations and will tend to pick one variable and disregard the rest. This hurts performance because if correlated regressors each contain a common signal and uncorrelated noise, a linear combination of the regressors formed based on an L^2 penalty will typically do better in isolating the signal than a single regressor alone. For instance, rather than picking book-to-market as the only characteristic

[4]To solve the optimization problem in eq. (28) we use the LARS-EN algorithm in Zou and Hastie (2005).

to represent the value effect in an SDF, it may be advantageous to consider a weighted average of multiple measures of value, such as book-to-market, price-dividend, and cashflow-to-price ratios. This reasoning also suggests that an L^1-only penalty may work better when we first transform the characteristics-based factors into their PCs before estimation. We examine this question in our empirical work below.

3.3 Data-driven penalty choice

To implement the estimators (22) and (28), we need to set the values of the penalty parameters γ and γ_1, γ_2, respectively. In the L^2-only penalty specification, the penalty parameter $\gamma = \frac{\tau}{\kappa T}$ following from the prior (18) has an economic interpretation. With our choice of $\eta = 2$, the root expected maximum squared Sharpe Ratio under the prior is

$$\mathbb{E}[\mu\Sigma^{-1}\mu]^{1/2} = \kappa, \tag{29}$$

and hence γ implicitly represents views about the expected squared Sharpe Ratio. For example, an expectation that the maximum Sharpe Ratio cannot be very high, i.e., low κ, would imply high γ and hence a high degree of shrinkage imposed on the estimation. Some researchers pick a prior belief based on intuitive reasoning about the likely relationship between the maximum squared Sharpe Ratio and the historical squared Sharpe Ratio of a market index.[5] However, these are intuitive guesses. It would be difficult to go further and ground beliefs about κ in deeper economic analyses of plausible degrees of risk aversion, risk-bearing capacity of arbitrageurs, and degree of mispricing. For this reason, we prefer a data-driven approach. But we will make use of eq. (29) to express the magnitude of the L^2-penalty that we apply in estimation in terms of an economically interpretable root expected maximum squared Sharpe Ratio.

The data-driven approach involves estimation of γ via K-fold cross validation. We divide the historic data into K equal sub-samples. Then, for each possible γ (or each possible pair of γ_1, γ_2 in the dual penalty specification), we compute $\hat{b}$ by applying eq. (22) to $K-1$ of these sub-samples. We evaluate the "out-of-sample" (OOS) fit of the resulting model on the single withheld subsample. Consistent with the penalized objective, eq. (26), we compute the OOS R-squared as

$$R^2_{\text{oos}} = 1 - \frac{\left(\bar{\mu}_2 - \overline{\Sigma}_2\hat{b}\right)'\left(\bar{\mu}_2 - \overline{\Sigma}_2\hat{b}\right)}{\bar{\mu}_2'\bar{\mu}_2}, \tag{30}$$

where the subscript 2 indicates an OOS sample moment from the withheld sample. We

[5] Barillas and Shanken (2017) is a recent example. See, also, MacKinlay (1995) and Ross (1976) for similar arguments.

repeat this procedure K times, each time treating a different sub-sample as the OOS data. We then average the R^2 across these K estimates, yielding the cross-validated R^2_{oos}. Finally, we choose γ (or γ_1, γ_2) that generates the highest R^2_{oos}.

We chose $K = 3$ as a compromise between estimation uncertainty in $\hat{b}$ and estimation uncertainty in the OOS covariance matrix $\overline{\Sigma}_2$. The latter rises as K increases due to difficulties of estimating the OOS covariance matrix precisely. With high K, the withheld sample becomes too short for $\overline{\Sigma}_2$ to be well-behaved, which distorts the fitted factor mean returns $\overline{\Sigma}_2\hat{b}$. However, our results are robust to using moderately higher K.

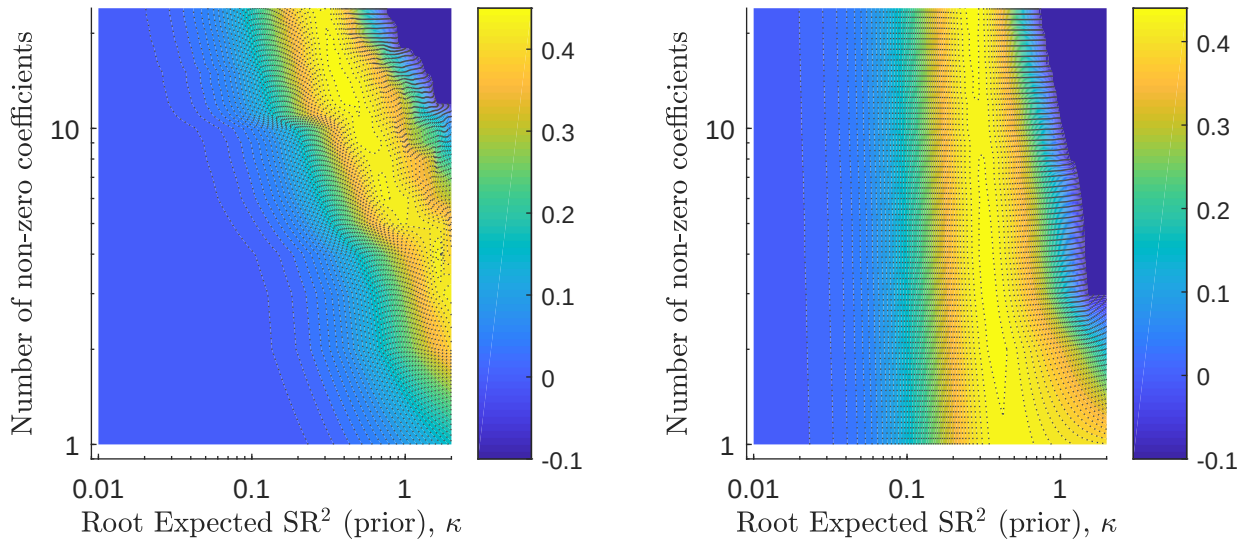
4 Empirical Analysis

4.1 Preliminary analysis: Fama-French SZ/BM portfolios

We start with an application of our proposed method to daily returns on the 25 Fama-French ME/BM-sorted (FF25) portfolios from 1926 to 2016, which we orthogonalize with respect to the CRSP value-weighted index return using βs estimated in the full sample.[6] In this analysis, we treat the 25 portfolio membership indicators as stock characteristics and we estimate the SDF's loadings on these 25 portfolios. These portfolios are not the challenging high-dimensional setting for which our method is designed, but this initial step is useful to verify that our method produces reasonable results before we apply it to more interesting and statistically challenging high-dimensional sets of asset returns where classic techniques are infeasible.

For the FF25 portfolios, we know quite well from earlier research what to expect and we can check whether our method produces these expected results. From Lewellen et al. (2010), we know that the FF25 portfolio returns have such a strong factor structure that the 25 portfolio returns (orthogonalized w.r.t. to the market index return) are close to being linear combinations of the SMB and HML factors. As a consequence, essentially any selection of four portfolios out of the 25 with somewhat different loadings on the SMB and HML factors should suffice to span the SDF. Thus, treating the portfolio membership indicators as characteristics, we should find a substantial degree of sparsity. From Kozak et al. (2017), we know that the SMB and HML factors essentially match the first and the second PCs of the FF25 (market-neutral) portfolio returns. Therefore, when we run the analysis using the PCs of the FF25 portfolio returns as the basis assets, we should find even more sparsity: two PCs at most should be sufficient to describe the SDF well.

[6]The resulting abnormal returns are $F_{i,t} = \tilde{F}_{i,t} - \beta_i R_{m,t}$ where $\tilde{F}_{i,t}$ is the raw portfolio return. We thank Ken French for providing FF25 portfolio return data on his website.



(a) Raw Fama-French 25 portfolios

(b) PCs of Fama-French 25 portfolios

Figure 1: OOS R^2 from dual-penalty specification (Fama-French 25 ME/BM portfolios). OOS cross-sectional R^2 for families of models that employ both L^1 and L^2 penalties simultaneously using 25 Fama-French ME/BM sorted portfolios (Panel a) and 25 PCs based on Fama and French portfolios (Panel b). We quantify the strength of the L^2 penalty by prior root expected SR^2 (κ) on the x-axis. We show the number of retained variables in the SDF, which quantifies the strength of the L^1 penalty, on the y-axis. Warmer (yellow) colors depict higher values of OOS R^2. Both axes are plotted on logarithmic scale.

Figure 1 presents results for our dual-penalty estimator in eq. (28). The results using the raw FF25 portfolio returns are shown in the left-hand side in Figure 1a; those using PCs of these returns are shown in the right-hand side plot Figure 1b. Every point on the plane in these plots represents a particular combination of the two penalties γ_1 and γ_2 that control sparsity and L^2-shrinkage, respectively. We vary the degree of L^2-shrinkage on the horizontal axis, going from extreme shrinkage on the left to no shrinkage at all at the right border of the plot. To facilitate interpretation, we express the degree of shrinkage in terms of κ. In the L^2-only penalty case, κ has a natural economic interpretation: it is the square root of the expected maximum squared Sharpe ratio under the prior in eq. (18) and it is inversely related to the shrinkage penalty $\gamma = \frac{\tau}{\kappa^2 T}$. Variation along the vertical axis represents different degrees of sparsity. We express the degree of sparsity in terms of how many factors remain in the SDF with non-zero coefficients. Thus, there is no sparsity at the top end of the plot and extreme sparsity at the bottom. Both axes are depicted on logarithmic scale.

The contour maps show the OOS R^2 calculated as in eq. (30) for each of these penalty

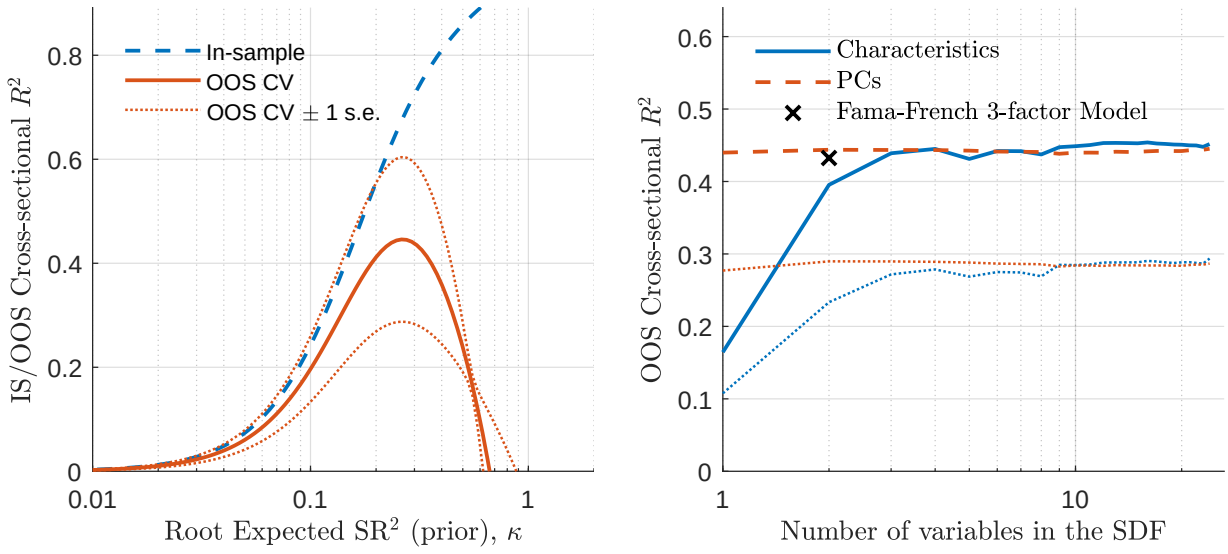
combinations. Our data-driven penalty choice selects the combination with the highest OOS R^2, but in this figure we show the OOS R^2 for a wide range of penalties to illustrate how L^2-shrinkage and sparsity (L^1 penalty) influences the OOS R^2. Warmer (yellow) colors indicate higher OOS R^2. To interpret the magnitudes it is useful to keep in mind that with our choice of $K = 3$, we evaluate the OOS R^2 in withheld samples of about 23 years in length, i.e., the OOS R^2 show how well the SDF explains returns averaged over a 23-year period.

Focusing first on the raw FF25 portfolio returns in Figure 1a, we can see that for this set of portfolios, sparsity and L^2-shrinkage are substitutes in terms of ensuring good OOS performance: the contour plot features a diagonal ridge with high OOS R^2 extending from the top edge of the plot (substantial L^2-shrinkage, no sparsity) to the right edge (substantial sparsity, no shrinkage). As we outlined above, this is what we would expect for this set of asset returns: a selection of 3-4 portfolios from these 25 should be sufficient to span the SDF that prices all 25 well, and adding more portfolio returns to the SDF hurts OOS performance unless more L^2-shrinkage is imposed to avoid overfitting. Unregularized models that include all 25 factors (top-right corner) perform extremely poorly in the OOS evaluation.[7]

Figure 1b, which is based on the PCs of the FF25 portfolio returns, also shows the expected result: even one PC is already sufficient to get close to the maximum OOS R^2 and two PCs are sufficient to attain the maximum. Adding more PCs to the SDF doesn't hurt the OOS performance as along as some L^2-shrinkage is applied. However, with PCs, the ridge of close-to-maximum OOS R^2 is almost vertical and hence very little additional L^2-shrinkage is needed when sparsity is relaxed. The reason is that our estimator based on the L^2 penalty in eq. (27) already downweights low-variance PCs by pushing their SDF coefficients close to zero. As a consequence, it makes little difference whether one leaves these coefficients close to zero (without the L^1 penalty at the top edge of the plot) or forces them to exactly zero (with substantial L^1 penalty towards the bottom edge of the plot).

In Figure 2, we further illustrate the role of L^2-shrinkage and sparsity by taking some cuts of the contour plots in Figure 1. Figure 2a focuses on L^2-shrinkage by taking a cut along the top edge of the contour plot for the raw FF25 portfolio returns in Figure 1a where we only shrink using the L^2-penalty, but do not impose sparsity. The OOS R^2 is shown by the solid red line. In line with Figure 1a, this plot shows that the OOS R^2 is maximized for $\kappa \approx 0.23$. For comparison, we also show the in-sample cross-sectional R^2 (dashed blue). The contrast with the OOS R^2 vividly illustrates how the in-sample R^2 can be grossly misleading about the ability of an SDF to explain expected returns OOS—and especially so without substantial shrinkage.

[7] We impose a floor on negative R^2 at -0.1 in these plots. In reality unregularized models deliver R^2 significantly below this number.



(a) L^2 model selection **(b)** Sparsity

Figure 2: L^2 Model Selection and Sparsity (Fama-French 25 ME/BM portfolios). Panel (a) plots the in-sample cross-sectional R^2 (dashed) and OOS cross-sectional R^2 based on cross validation (solid). Dotted lines depict ± 1 s.e. bounds of the CV estimator. In Panel (b) we show the maximum OOS cross-sectional R^2 attained by a model with n factors (on the x-axis) across all possible values of L^2 shrinkage, for models based on original characteristics portfolios (solid) and PCs (dashed). Dotted lines depict -1 s.e. bounds of the CV estimator. The "X" mark indicates OOS performance of the Fama-French model that uses only SMB and HML factors.

Figure 2b presents the OOS R^2 for various degrees of sparsity, choosing the optimal (i.e., OOS R^2 maximizing) amount of L^2-shrinkage for each level of sparsity. In other words, we are following the ridge of the highest values in the contour plots from the bottom edge of the plot to the top. The solid blue line is based on the raw FF25 portfolio returns and the dashed red line based on the PCs. Dotted lines on the plot show approximate -1 standard error bounds for the CV estimator.[8] This plot makes even more transparent our earlier point that a sparse SDF with just a few of the FF25 portfolio is sufficient to get maximal OOS performance—comparable to the an SDF with SMB and HML shown by the black "X"[9]—and that in PC-space even one PC is enough. The PC that is eliminated last as we raise the degree of sparsity is PC1 (i.e., with the one with the highest variance). PC1 is

[8]We estimate these by computing variance of the CV estimator under the assumption that $K = 3$ CV estimates are IID. In that case, $\text{var}\left(R^2_{\text{CV estimator}}\right) = \text{var}\left(\frac{1}{K}\sum_{j=1}^{K}\hat{R}_j^{\,2}\right) \approx \frac{1}{K}\text{var}\left(\hat{R}_j^{\,2}\right)$, where $\hat{R}_j^{\,2}$ is an estimate of the OOS R^2 in the j-th fold of the data. Standard errors of the CV estimator can thus be computed as $\frac{1}{\sqrt{K}}\text{sd}\left(\hat{R}_1^{\,2}, ..., \hat{R_K}^{\,2}\right)$.

[9]To put both approaches on equal footing, we shrink Fama-French coefficients towards zero based on the amount of "level" shrinkage implied by our method. This modification significantly improves OOS performance of the FF factors. Since SMB and HML are long-short factors, one could also view them as representing four portfolio returns rather than the two that we assumed here.

highly correlated with the HML factor (and somewhat with SMB); the SDF based on PC1 is therefore effectively the same as Fama-French's and performs similarly.

To summarize, these results confirm that our method can recover the SDF that Fama and French (1993) constructed intuitively for this set of portfolios. The method also can detect sparsity where it should (few portfolios and very few PCs are sufficient to represent the SDF) for this well-known set of portfolios. The true strength of our method, however, comes in dealing with multidimensional settings characterized by a vast abundance of characteristics and unknown factors where classic techniques are inadequate. We turn to these more challenging settings next.

4.2 Large sets of characteristics portfolios

We start with the universe of U.S. firms in CRSP. We construct two independent sets of characteristics. The first set relies on characteristics underlying common "anomalies" in the literature. We follow standard anomaly definitions in Novy-Marx and Velikov (2016), McLean and Pontiff (2016), and Kogan and Tian (2015) and compile our own set of 50 such characteristics. The second set of characteristic is based on 70 financial ratios as defined by WRDS: "WRDS Industry Financial Ratios" (WFR) is a collection of most commonly used financial ratios by academic researchers (often for purposes other than return prediction). There are in total over 70 financial ratios grouped into the following seven categories: Capitalization, Efficiency, Financial Soundness/Solvency, Liquidity, Profitability, Valuation, and Others" (Table 6 in the Appendix lists the ratios). We supplement this dataset with 12 portfolios sorted on past monthly returns in months $t-1$ through $t-12$. The combined dataset contains 80 managed portfolios (we drop two variables due to their short time series and end up with 68 WRDS ratios in the final dataset).

In order to focus exclusively on the cross-sectional aspect of return predictability, remove the influence of outliers, and keep constant leverage across all portfolios, we perform certain normalizations of characteristics that define our characteristics-based factors. First, similarly to Asness et al. (2014) and Freyberger et al. (2017), we perform a simple rank transformation for each characteristic. For each characteristic i of a stock s at a given time t, denoted as $c^i_{s,t}$, we sort all stocks based on the values of their respective characteristics $c^i_{s,t}$ and rank them cross-sectionally (across all s) from 1 to n_t, where n_t is the number of stocks at t for which this characteristic is available.[10] We then normalize all ranks by dividing by $n_t + 1$ to

[10]If two stocks are "tied", we assign the average rank to both. For example, if two firms have the lowest value of c, they are both assigned a rank of 1.5 (the average of 1 and 2). This preserves any symmetry in the underlying characteristic.

obtain the value of the rank transform:

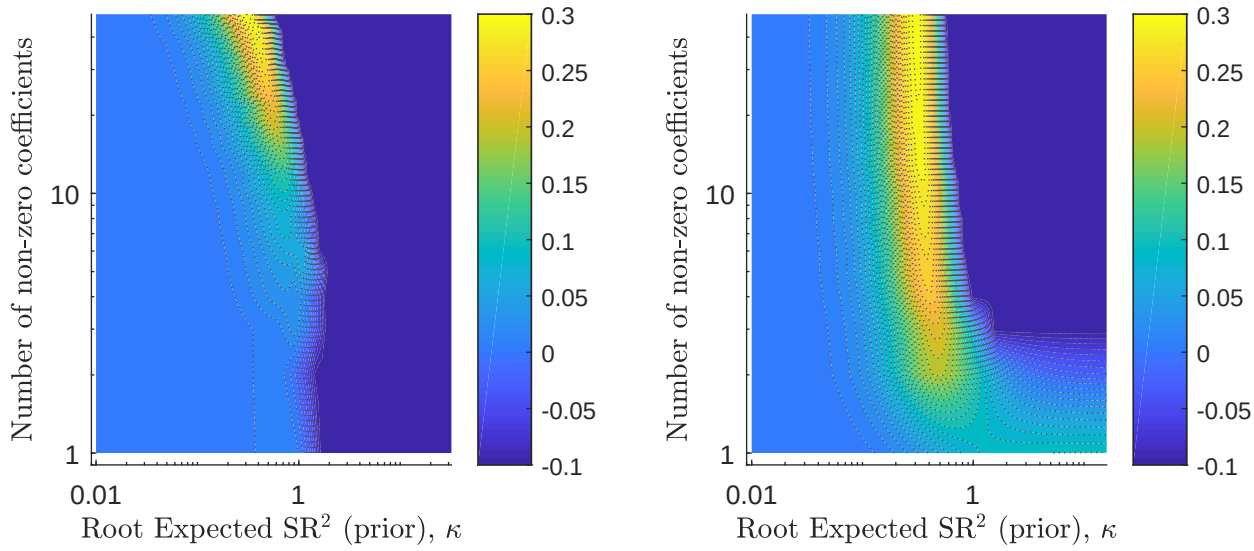
$$rc^i_{s,t} = \frac{\text{rank}\left(c^i_{s,t}\right)}{n_t + 1}. \tag{31}$$

Next, we normalize each rank-transformed characteristic $rc^i_{s,t}$ by first centering it cross-sectionally and then dividing by sum of absolute deviations from the mean of all stocks:

$$z^i_{s,t} = \frac{\left(rc^i_{s,t} - \bar{rc}^i_t\right)}{\sum_{s=1}^{n_t}\left|rc^i_{s,t} - \bar{rc}^i_t\right|}, \tag{32}$$

where $\bar{rc}^i_t = \frac{1}{n_t}\sum_{s=1}^{n_t} rc^i_{s,t}$. The resulting zero-investment long-short portfolios of transformed characteristics $z^i_{s,t}$ are insensitive to outliers and allow us to keep the absolute amount of long and short positions invested in the characteristic-based strategy (i.e., leverage) fixed. For instance, doubling the number of stocks at any time t has no effect on the strategy's gross exposure.[11] Finally, we combine all transformed characteristics $z^i_{s,t}$ for all stocks into a matrix of instruments Z_t.[12] Interaction with returns, $F_t = Z'_{t-1}R_t$, then yields one factor for each characteristic.

To ensure that the results are not driven by very small illiquid stocks, we exclude small-cap stocks with market caps below 0.01% of aggregate stock market capitalization at each point in time.[13] In all of our analysis we use *daily* returns from CRSP for each individual stock. Using daily data allows us to estimate second moments much more precisely than with monthly data and focus on uncertainty in means while largely ignoring negligibly small uncertainty in covariance estimates (with exceptions as noted below). We adjust daily portfolio weights on individual stocks within each month to correspond to a monthly-rebalanced buy-and-hold strategy during that month. All portfolios' returns are further rescaled to have standard deviations equal to the in-sample standard deviation of the excess return on the aggregate market index. Table 5 in the Appendix shows the annualized mean returns for the anomaly portfolios. Mean returns for the WFR managed portfolios are reported in Appendix Table 6. Finally, as in the previous section, we orthogonalize all portfolio returns with respect to the CRSP value-weighted index return using βs estimated in the full sample.



(a) Raw 50 anomaly portfolios

(b) PCs of 50 anomaly portfolios

Figure 3: OOS R^2 from dual-penalty specification (50 anomaly portfolios). OOS cross-sectional R^2 for families of models that employ both L^1 and L^2 penalties simultaneously using 50 anomaly portfolios (Panel a) and 50 PCs based on anomaly portfolios (Panel b). We quantify the strength of the L^2 penalty by prior root expected SR^2 (κ) on the x-axis. We show the number of retained variables in the SDF, which quantifies the strength of the L^1 penalty, on the y-axis. Warmer (yellow) colors depict higher values of OOS R^2. Both axes are plotted on logarithmic scale.

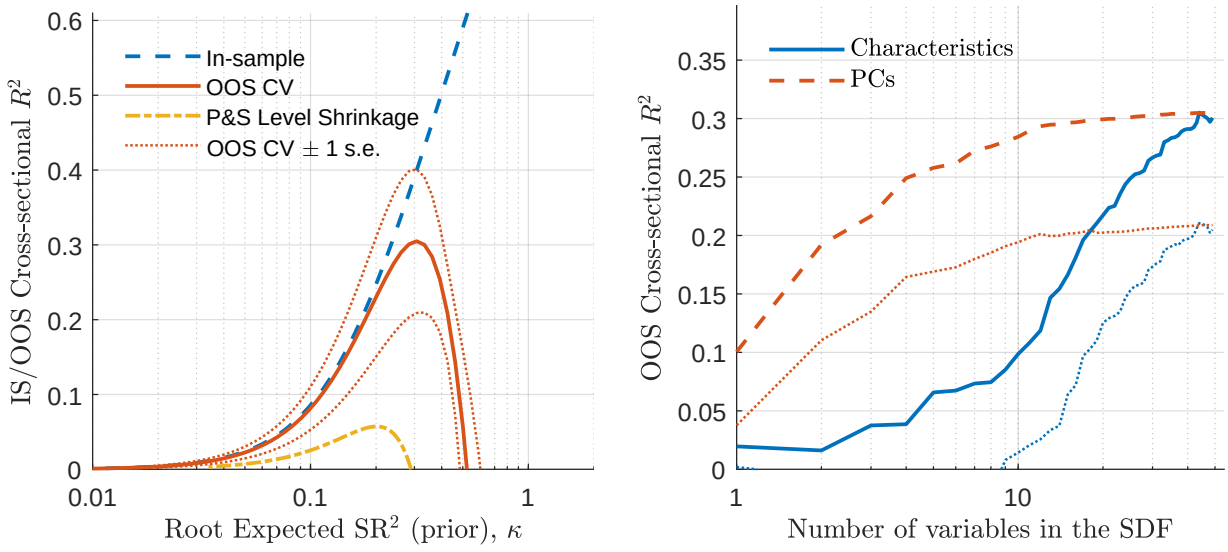
4.2.1 50 anomaly characteristics

We now turn to our primary dataset of 50 portfolios based on anomaly characteristics. Figure 3 presents the OOS R^2 from our dual-penalty specification as a function of κ (on the x-axis) and the number of non-zero SDF coefficients (on the y-axis). A comparison with our earlier Figure 1 for the FF25 portfolios shows some similarities, but also features that are drastically different. Focusing on the left-hand Figure 3a based on raw returns of the 50 anomaly portfolios, one similarity is that unregularized models (top-right corner) demonstrate extremely poor performance with OOS R^2 substantially below 0. Hence, substantial regularization is needed to get good OOS performance. However, unlike for the FF25 portfolios, there isn't much substitutability between L^1 and L^2-regularization here. To attain the maximum OOS R^2, the data calls for substantial L^2-shrinkage, but essentially no sparsity. Imposing sparsity (i.e., moving down in the plot) leads to a major deterioration in OOS R^2. This indicates that there is almost no redundancy among the 50 anomalies. The FF25 port-

[11]Since the portfolio is long-short the net exposure is always zero.

[12]If $z^i_{s,t}$ is missing we replace it with the mean value, zero.

[13]For example, for an aggregate stock market capitalization of \$20tn, we keep only stocks with market caps above \$2bn.



(a) L^2 model selection **(b)** Sparsity

Figure 4: L^2 Model Selection and Sparsity (50 anomaly portfolios). Panel (a) plots the in-sample cross-sectional R^2 (dashed), OOS cross-sectional R^2 based on cross validation (solid), and OOS cross-sectional R^2 based on the proportional shrinkage (dash-dot) from Pástor and Stambaugh (2000). In Panel (b) we show the maximum OOS cross-sectional R^2 attained by a model with n factors (on the x-axis) across all possible values of L^2 shrinkage, for models based on original characteristics portfolios (solid) and PCs (dashed). Dotted lines in Panel (b) depict -1 s.e. bounds of the CV estimator.

folios have so much redundancy that a small subset of these portfolios is sufficient to span the SDF. In contrast, to adequately capture the pricing information in the 50 anomalies one needs to include basically all of these 50 factors in the SDF. Shrinking their SDF coefficients is important to obtain good performance, but forcing any of them to zero to get a sparse solution hurts the OOS R^2. In other words, a *characteristics*-sparse SDF with good pricing performance does not exist. Hence, many anomalies do in fact make substantial marginal contributions to OOS explanatory power of the SDF.

If we take the PCs of the anomaly portfolio returns as basis assets, as shown in Figure 3b, the situation is quite different. A relatively sparse SDF with only four PCs, for example, does quite well in terms of OOS R^2 and with 10 PCs we get close to the maximum OOS R^2. Thus, a *PC*-sparse SDF prices the anomaly portfolios quite well.

Figure 4 provides a more precise picture of the key properties of OOS R^2 by taking cuts of the contour plots. The solid red line in Figure 4a represents a cut along the top edge of Figure 3 with varying degrees of L^2-shrinkage, but no sparsity. As the figure shows, the OOS R^2 is maximized for $\kappa \approx 0.30$. The standard error bounds indicate that OOS R^2 around this value of κ is not only economically, but also statistically quite far above zero. Table 1a lists

Table 1: Largest SDF factors (50 anomaly portfolios)

Coefficient estimates and absolute t-statistics at the optimal value of the prior root expected SR^2 (based on cross-validation). Panel (a) focuses on the original 50 anomaly portfolios. Panel (b) pre-rotates returns into PC space and shows coefficient estimates corresponding to these PCs. Coefficients are sorted descending on their absolute t-statistic values.

(a) Raw 50 anomaly portfolios

	b	t-stat
Industry Rel. Rev. (L.V.)	-0.92	3.67
Ind. Mom-Reversals	0.50	1.98
Industry Rel. Reversals	-0.44	1.75
Seasonality	0.36	1.44
Earnings Surprises	0.35	1.39
Return on Market Equity	0.32	1.28
Value-Profitablity	0.31	1.22
Composite Issuance	-0.26	1.03
Return on Equity	0.24	0.94
Investment/Assets	-0.23	0.92
Momentum (12m)	0.23	0.91

(b) PCs of 50 anomaly portfolios

	b	t-stat
PC 5	-0.91	3.82
PC 1	-0.57	3.41
PC 2	-0.54	2.58
PC 4	0.49	2.08
PC 11	-0.48	1.92
PC 15	0.42	1.66
PC 10	-0.36	1.41
PC 6	-0.30	1.25
PC 19	-0.26	1.00
PC 14	0.24	0.94
PC 9	0.21	0.84

the anomaly factors with the largest absolute t-statistics, where standard errors are based on eq. (23). The largest coefficients and t-statistics are associated with industry relative reversals (low vol.), industry momentum-reversals, industry relative-reversals, seasonality, earnings surprises, ROE, value-profitability, momentum, etc. Not surprisingly, these are the anomalies that have been found to be among the most robust in the literature. Our method uncovers them naturally. The t-statistics are quite low, but it is important to keep in mind that what matters for the SDF is the joint significance of linear combinations of 50 of these factors. Table 1b shows t-statistics for particular linear combinations: the PCs of the 50 portfolio returns. As the table shows, the loadings on PC1, PC2, PC4, and PC5 are all significantly different from zero at conventional significance levels.[14] Our earlier analysis in Figure 4b showed that the SDF already achieves a high OOS R^2 with only these four PCs. It is also consistent with our economic arguments in the beginning of the paper that the PCs with the biggest absolute coefficients are PCs with the highest variance.

In Section 3.1 we argued on economic grounds that our prior specification with $\eta = 2$ is reasonable. However, it would be useful to check whether this economic motivation is

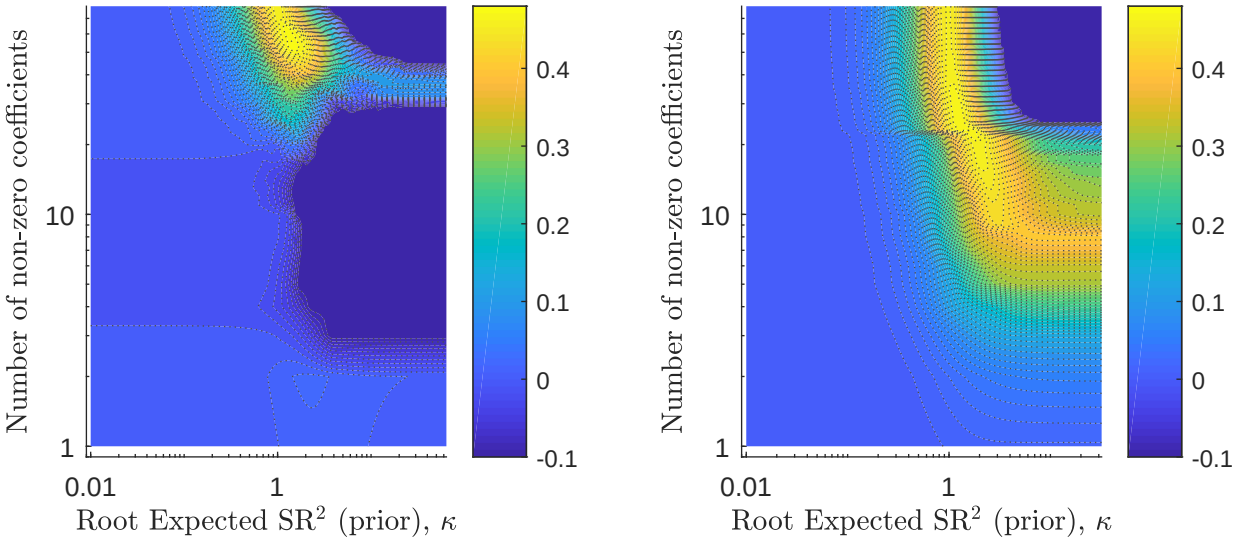
[14]Since L^2 regularization is rotation invariant, we obtain the same solution (in terms of the weight that an individual anomaly factor obtains in the SDF) whether we first estimate the model on the original assets and then rotate into PC space or directly estimate in PC space. Thus, the coefficients Table 1b are linear combinations of those in Table 1a.

also accompanied by better performance in the data. To do this, the yellow dash-dot line in Figure 4a plots the OOS R^2 we would get with the more commonly used prior of Pástor and Stambaugh (2000) with $\eta = 1$.[15] Recall that our method performs both level shrinkage of all coefficients, as well as relative shrinkage (twist) which down-weights the influence of small PCs. The method in Pástor and Stambaugh (2000) employs only level shrinkage. We can see that optimally-chosen level shrinkage alone achieves OOS R^2 lower than 5% (an improvement over the OLS solution), but falls substantially short of the 30% R^2 delivered by our method. Relative shrinkage, which is the key element of our method, therefore, contributes a major fraction of the total out-of-sample performance.

Figure 4b takes a cut in the contour plots along the ridge of maximal OOS R^2 from bottom to top where we vary sparsity and choose the optimal L^2-shrinkage for each level of sparsity. The solid blue line shows very clearly how characteristics-sparse SDFs perform poorly. The OOS R^2 only starts rising substantially at the lowest sparsity levels towards the very right of the plot. In PC space, on the contrary, very sparse models perform exceedingly well: a model with a single PC-based factor captures roughly a third of the total OOS cross-sectional R^2, while adding a second factor raises the R^2 to about 65% of the maximal one. A model with 10 PC factors achieves nearly maximal R^2, while a model with ten factors in the space of characteristics-based factors achieves less a third of the maximum—as much as a model with only one PC does. Many of the PC factors that our dual-penalty approach picks in PC-sparse SDF representations are the same as the PCs with highest t-statistics in Table 1. For instance, the first selected factor is PC1, followed by PC5, PC4, and PC2. (see Figure 11 in the Appendix for more details).

To summarize, there is little redundancy among the 50 anomalies. As a consequence, it is not possible to find a sparse SDF with just a few characteristics-based factors that delivers good OOS performance. For this reason, it is also important to deal with the high-dimensional nature of the estimation problem through an L^2-shrinkage rather than just an L^1-penalty and sparsity. L^2-shrinkage delivers much higher OOS R^2 than a pure L^1-penalty Lasso-style approach and the dual-penalty approach with data-driven penalty choice essentially turns off the L^1 penalty for this set of portfolios. However, if these portfolio returns are transformed into their PCs, a sparse representation of the SDF emerges. These findings are consistent with the point we made in Section 2 that the economic arguments for a characteristics-sparse SDF are rather weak, while there are good reasons to expect sparsity in terms of PCs.

[15] For the Pástor and Stambaugh (2000) level shrinkage estimator we show $\mathbb{E}\left(\mathrm{SR}^2\right)$ under the prior on the x-axis, but it no longer coincides with the κ parameter in this case.



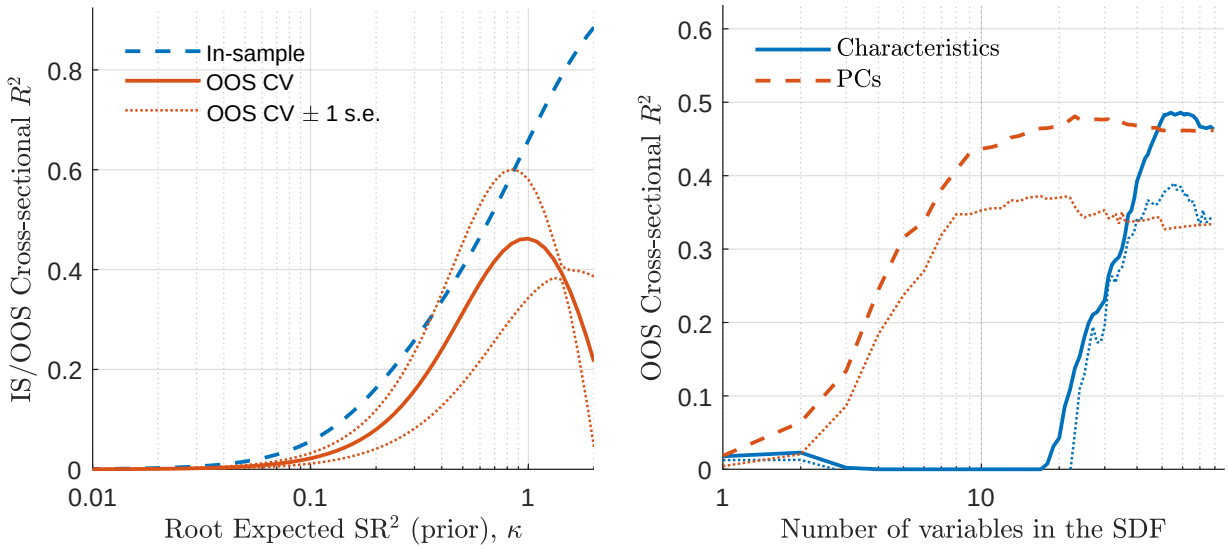
(a) Raw WFR portfolios **(b)** PCs of WFR portfolios

Figure 5: OOS R^2 from dual-penalty specification (WFR portfolios). OOS cross-sectional R^2 for families of models that employ both L^1 and L^2 penalties simultaneously using 80 WFR portfolios (Panel a) and 80 PCs based on WFR portfolios (Panel b). We quantify the strength of the L^2 penalty by prior root expected SR^2 (κ) on the x-axis. We show the number of retained variables in the SDF, which quantifies the strength of the L^1 penalty, on the y-axis. Warmer (yellow) colors depict higher values of OOS R^2. Both axes are plotted on logarithmic scale.

4.2.2 WRDS financial ratios (WFR)

The dataset of 50 anomalies is special in the sense that all of these characteristics are known, from the past literature, to be related to expected returns. Our method is useful to check for redundancy among these anomalies, but this set of asset returns did not expose the method to the challenge of identifying entirely new pricing factors from a high-dimensional data set. For this reason, we now look at 80 characteristics-based factors formed based on the WFR data set, including 12 portfolios sorted on past monthly returns in months $t-1$ through $t-12$. Some of the characteristics in the WFR data set are known to be related to expected returns (e.g., several versions of the P/E ratio), but many others are not. It is therefore possible that a substantial number of these 80 factors are irrelevant for pricing. It will be interesting to see whether our method can: (i) properly de-emphasize these pricing-irrelevant factors and avoid overfitting them; (ii) pick out pricing factors that are similar to those that our analysis of 50 anomalies found relevant; (iii) potentially find new pricing factors.

The contour map of OOS R^2 in Figure 5 looks quite similar to the earlier one for the 50 anomaly portfolios in Figure 3. Unregularized models (top-right corner) again perform extremely poorly with OOS R^2 significantly below 0. L^2-penalty-only based models (top



(a) L^2 model selection **(b)** Sparsity

Figure 6: L^2 Model Selection and Sparsity (WFR portfolios). Panel (a) plots the in-sample cross-sectional R^2 (dashed) and OOS cross-sectional R^2 based on cross validation (solid). In Panel (b) we show the maximum OOS cross-sectional R^2 attained by a model with n factors (on the x-axis) across all possible values of the prior root expected SR^2 (κ) for models based on original characteristics portfolios (solid) and PCs (dashed). Dotted lines in Panel (b) depict -1 s.e. bounds of the CV estimator.

edge of a plot) perform well for both raw portfolio returns and PCs. L^1-penalty-only "Lasso" based models (right edge of the plot) work poorly for raw portfolio returns in Figure 5a.

However, there are some differences as well. First, in Figure 5a, the area of very high OOS R^2 extends down from the top more than it does in the case of the 50 anomalies (due to the log scale on the y-axis this feature is relatively subtle). This indicates that imposing a small degree of sparsity is not harmful to OOS performance. Apparently, some of the WFR characteristics are not important for pricing and can be left out. Second, as can be seen towards the right-edge of Figure 5b, a PC-sparse SDF not only does quite well in terms of OOS R^2, but it does so even without much L^2-shrinkage. A potential explanation of this finding is that the data mining and publication bias towards in-sample significant factors may play a bigger role in the anomalies data set, which is based on published anomalies, than in the WFR data set. As a consequence, stronger shrinkage of SDF coefficients towards zero may be needed in the anomalies data set to prevent these biases from impairing OOS performance, while there is less need for shrinkage in the WFR data set because in- and out-of-sample returns are not so different.

This explanation is further consistent with the fact that the OOS R^2-maximizing $\kappa \approx 1$, which is much higher than in the anomalies data set. Figure 6a illustrates this even more

Table 2: Largest SDF factors (WFR portfolios)

Coefficient estimates and t-statistics at the optimal value of the prior root expected SR^2 (based on cross-validation). Panel (a) focuses on the original WFR portfolios. Panel (b) pre-rotates returns into PC space and shows coefficient estimates corresponding to these PCs. Coefficients are sorted descending on their absolute t-statistic values.

(a) Raw WFR portfolios

	b	t-stat
Free Cash Flow/Operating Cash Flow	3.03	5.06
Accruals/Average Assets	2.41	3.89
P/E (Diluted, Incl. EI)	-2.09	3.33
Month $t-9$	1.59	2.86
Operating CF/Current Liabilities	1.72	2.73
Trailing P/E to Growth (PEG) ratio	-1.46	2.53
Month $t-11$	1.39	2.49
Cash Flow/Total Debt	1.53	2.37
P/E (Diluted, Excl. EI)	-1.47	2.33
Month $t-12$	1.21	2.17
Enterprise Value Multiple	-1.31	2.16

(b) PCs of WFR portfolios

	b	t-stat
PC 7	-3.19	6.75
PC 19	-3.57	6.33
PC 26	2.58	4.32
PC 6	-1.90	4.22
PC 2	-0.51	2.37
PC 17	-1.31	2.35
PC 9	-1.20	2.34
PC 10	1.17	2.25
PC 18	1.22	2.19
PC 5	0.82	2.16
PC 25	1.14	1.93

transparently by taking a cut along the top edge of Figure 5a. The solid red line shows the OOS R^2. Its peak is much farther to the right than in the analogous figure for the anomalies data set (Figure 4a), consistent with our intuition that WFR are less likely to have been data-mined in an early part of the sample compared to the published anomalies and therefore do not require as much shrinkage. Standard errors are smaller, too, due to more stable performance of WFR portfolios across time periods relative to anomalies, which experienced significant deterioration in the latest (not data-mined) part of the sample (McLean and Pontiff, 2016).

Table 2 lists coefficient estimates at this optimal level of L^2-only penalty. Coefficients are sorted descending on their absolute t-statistic values. Table 2a focuses on original WFR portfolio returns. It shows that our method tends to estimate high weights on factors based on characteristics known to be associated with expected returns. Among the picks there are few measures of valuation ratios (P/E, PEG, Enterprise Value Multiple), investment (Free CF/Operating C, which equals 1 - Capital Expenditure/Operating CF), accruals (Accruals/Average Assets), financial soundness (Operating CF/Current Liabilities, Operating CF/Total Debt), and momentum (months $t-9$, $t-11$, $t-12$). None of these variables on their own, however, are likely to be optimal measures of the "true" underlying signal (factor). Our method combines information in many such imperfect measures (averaging

them by the means of the L^2 penalty) and delivers a robust SDF that performs well out of sample. Combining several measures of each signal (e.g., valuation measures) performs much better out of sample than using any single ratio.

Table 2b pre-rotates assets into PC space. Most of the entries in this table belong to the top 20 high-variance PCs. However, compared with the anomaly portfolio PCs in Table 1b, there are a few more of the lower variance PCs on this list as well. If we also impose some sparsity through an L^1 penalty in a dual-penalty specification, these lower variance PCs drop out. For example, the best sparse model with 5 factors, which achieves about three-quarters of the maximal OOS R^2, includes PC 1, PC 2, PC 6, PC 7, PC 19. This is broadly consistent with our economic arguments that important pricing factors are most likely to be found among high-variance PCs, although, of course, not every high-variance PC is necessarily an important factor in the SDF.

Figure 6b takes a cut in the contour plots along the ridge of maximal OOS R^2 from bottom to top where we vary sparsity and choose the optimal shrinkage for each level of sparsity. This figure illustrates that—unlike in the case of the 50 anomalies—some degree of sparsity does not hurt the OOS R^2. As the solid blue line shows, the OOS R^2 reaches its maximum at around 45 characteristics, which means that half of the WFR characteristics-based factors can be omitted from the SDF. Even so, sparsity is again much stronger in PC space. A model with five factors captures a large fraction of the total OOS cross-sectional R^2, while a model with nine factors delivers nearly maximum OOS R^2.

In summary, the analysis of the WFR data set shows that our method can handle well a data set that mixes factors that are relevant for pricing with others that are not. Sensibly, the characteristics-based factors that our method finds to be the ones most relevant with the highest weight in the SDF are closely related to those that help price the 50 anomaly portfolios. If sparsity is desired, a moderate level of L^1-penalty can be used to omit the pricing-irrelevant factors, but a L^2-penalty-only method works just as well in terms of OOS R^2.

4.3 Interactions

To raise the statistical challenge, we now consider extremely high-dimensional data sets. We supplement the sets of 50 anomaly and 80 WFR raw characteristics with characteristics based on second and third powers and linear first-order interactions of characteristics. This exercise is interesting not only in terms of the statistical challenge, but also because it allows us to relax the rather arbitrary assumption of linearity of factor portfolio weights in the characteristics when we construct the characteristics-based factors.

In fact, for some anomalies like the idiosyncratic volatility anomaly, it is known that the expected return effect is concentrated among stocks with extreme values of the characteristic. Fama and French (2008); Freyberger et al. (2017) provide evidence of nonlinear effects for other anomalies, but in terms of portfolio sorts and cross-sectional return prediction rather than SDF estimation. Furthermore, while there is existing evidence of interaction effects for a few anomalies (Asness et al., 2013; Fama and French, 2008), interactions have not been explored in the literature for more than these few—presumably a consequence of the extreme high-dimensionality of the problem. Interactions expand the set of possible predictors exponentially. For instance, with only first-order interactions of 50 raw characteristics and their powers, we obtain $\frac{1}{2}n\left(n+1\right)+2n=1,375$ candidate factors and test asset returns. For 80 WFR characteristics, we obtain a set of $3,400$ portfolios.

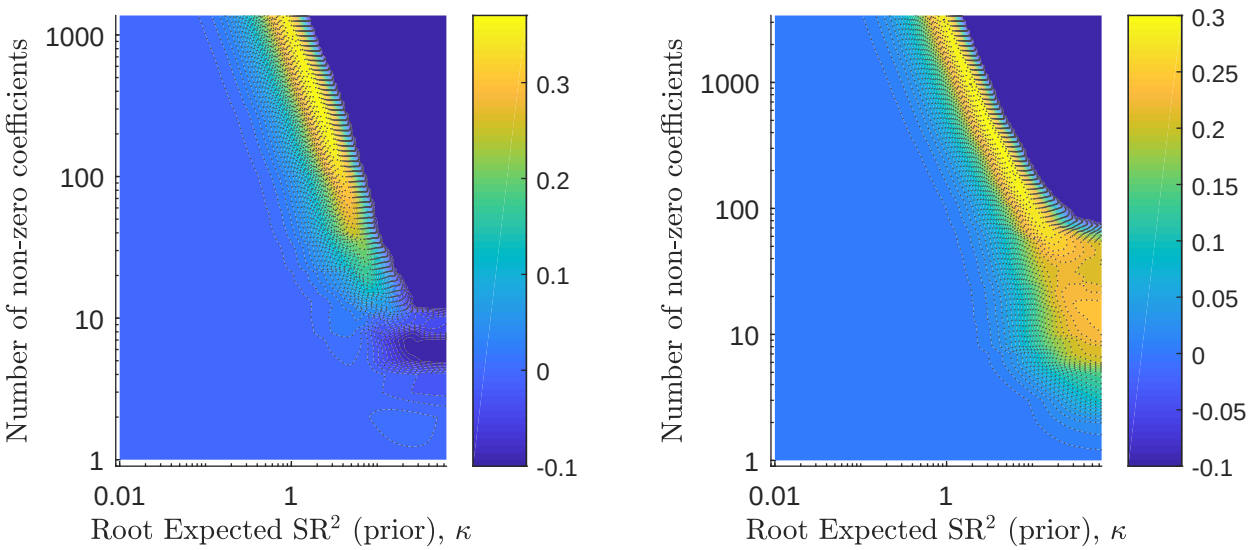
We construct the nonlinear weights and interactions as follows. For any two given rank-transformed characteristics $z_{s,t}^{i}$ and $z_{s,t}^{j}$ of a stock s at time t, we define the first-order interaction characteristic $z_{s,t}^{ij}$ as the product of two original characteristics that is further re-normalized using eq. (32) as follows:

$$z_{s,t}^{ij}=\frac{\left(z_{s,t}^{i}z_{s,t}^{j}-\frac{1}{n_t}\sum_{s=1}^{n_t}z_{s,t}^{i}z_{s,t}^{j}\right)}{\sum_{s=1}^{n_t}\left|z_{s,t}^{i}z_{s,t}^{j}-\frac{1}{n_t}\sum_{s=1}^{n_t}z_{s,t}^{i}z_{s,t}^{j}\right|}. \tag{33}$$

We include all first-order interactions in our empirical tests. In addition to interactions, we also include second and third powers of each characteristic, which are defined analogously based on interaction of the characteristic with itself. Note that although we re-normalize all characteristics after interacting or raising to powers, we do not re-rank them. For example, the cube of any given characteristic then is a new different characteristic that has stronger exposures to stocks with extreme realization of the original characteristic. We illustrate how this approach maps into more conventional two-way portfolio sorts portfolios in Appendix C.

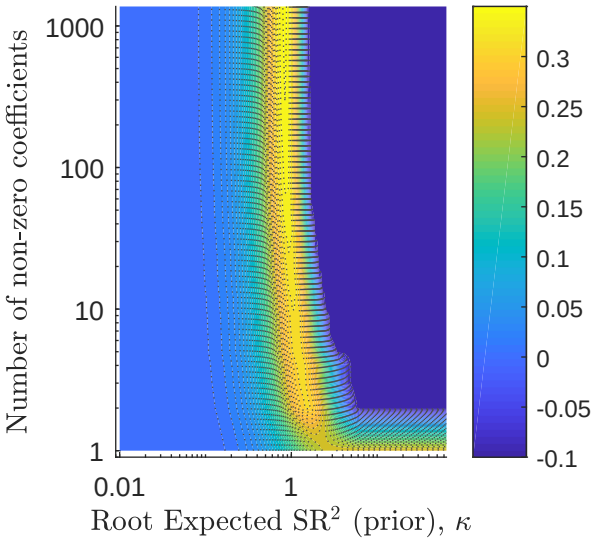
Due to the extremely high number of characteristics-based factors in this case, our 3-fold cross-validation method runs into numerical instability issues in covariance matrix inversion, even with daily data. For this reason, we switch to 2-fold cross-validation. This gives us a somewhat longer sample to estimate the covariance matrix and this sample extension is sufficient to obtain stable behavior.

Figure 7 shows contour maps of the OOS cross-sectional R^2 as a function of κ (on the x-axis) and the number of non-zero SDF coefficients (on the y-axis). Plots for the raw portfolio returns are shown in the top row and plots for the PCs are in the bottom row. Focusing first on the results for the raw portfolio returns, it is apparent that a substantial

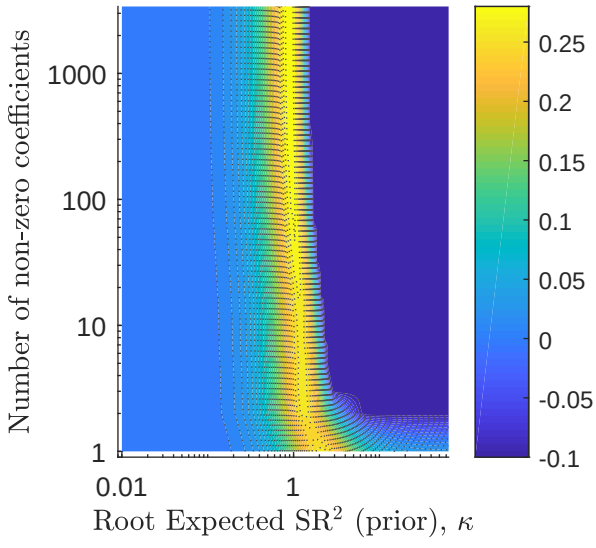


(a) 50 anomalies

(b) WFR portfolios



(c) PCs of 50 anomalies



(d) PCs of WFR portfolios

Figure 7: OOS R^2 from dual-penalty specification for models with interactions. OOS cross-sectional R^2 for families of models that employ both L^1 and L^2 penalties simultaneously using portfolio returns based on interactions of 50 anomaly (Panel a) and 80 WFR (Panel b) characteristics, and PCs of these portfolio returns (Panels c and d). We quantify the strength of the L^2 penalty by prior root expected SR^2 (κ) on the x-axis. We show the number of retained variables in the SDF, which quantifies the strength of the L^1 penalty, on the y-axis. Warmer (yellow) colors depict higher values of OOS R^2. Both axes are plotted on logarithmic scale.

Table 3: Largest SDF factors (models with interactions)

Coefficient estimates and t-statistics at the optimal value of the prior root expected SR^2 (based on cross-validation). Panel (a) focuses on the SDF constructed from PCs portfolio returns based on interactions of 50 anomaly characteristics. Panel (b) shows coefficient estimates corresponding to PCs of portfolio returns based on interactions of WFR. Coefficients are sorted descending on their absolute t-statistic values.

(a) PCs of interactions of anomaly portfolios

	b	t-stat
PC 1	-0.24	3.97
PC 2	0.28	3.31
PC 17	0.28	2.28
PC 40	-0.30	2.28
PC 60	0.27	2.03
PC 19	0.23	1.89
PC 67	0.24	1.86
PC 30	-0.21	1.62
PC 63	-0.21	1.60
PC 10	-0.19	1.59
PC 21	-0.18	1.49

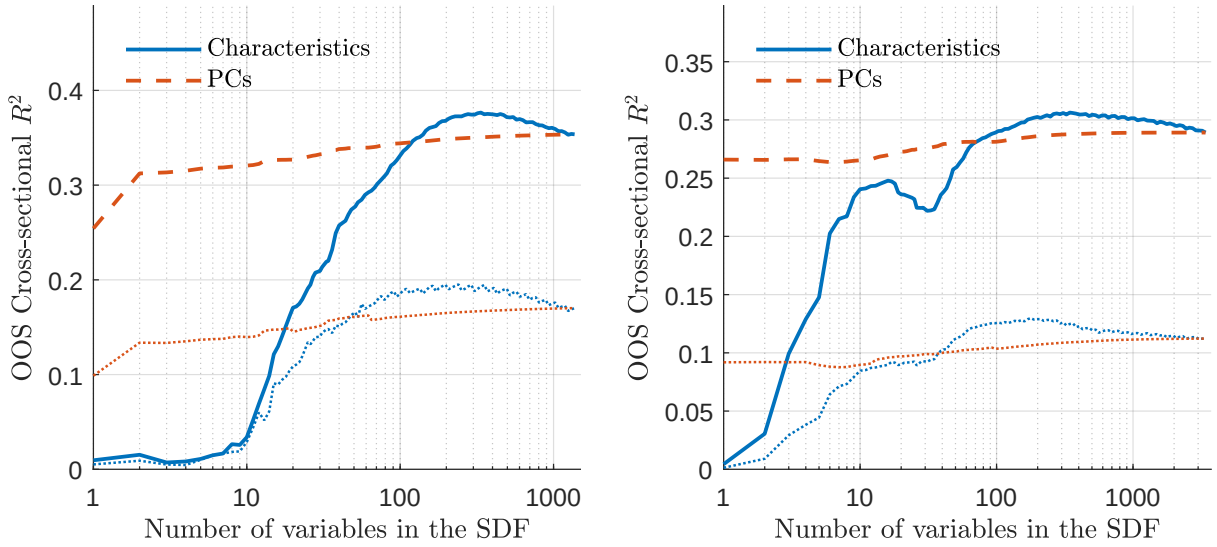
(b) PC of interactions of WFR portfolios

	b	t-stat
PC 1	-0.11	3.02
PC 5	-0.13	1.78
PC 2	-0.08	1.50
PC 50	0.13	1.45
PC 7	-0.10	1.23
PC 4	-0.08	1.17
PC101	-0.11	1.16
PC 20	0.10	1.15
PC 83	0.11	1.12
PC112	-0.11	1.11
PC 30	-0.09	1.04

degree of sparsity is now possible for both the anomalies and the WFR portfolios without deterioration in the OOS R^2. Strengthening the L^1-penalty to the point that only around 200 of the characteristics and their powers and interactions remain in the SDF (out of $1,375$ and $3,400$, respectively) does not reduce the OOS R^2 as long as one picks the L^2-penalty optimal for this level of sparsity. As before, an L^1-penalty-only approach leads to poor OOS performance.

The plots in the bottom row show contour maps for PCs. These results are drastically different in terms of how much sparsity can be imposed without hurting OOS performance. Very few PCs—or even just one—suffice to obtain substantial OOS explanatory power. But here, too, the combination of sparsity with an optimally chosen L^2 penalty is very important. Adding more PCs does not hurt as long as substantial L^2 shrinkage is imposed, but it does not improve OOS performance much either.

Table 3 lists coefficient estimates at the optimal level of L^2 regularization (i.e., the maximum along the top edge of the contour plots). Table 3a focuses on the SDF constructed from PCs of portfolio returns based on interactions of 50 anomaly characteristics. Table 3b shows coefficient estimates corresponding to PCs of portfolio returns based on interactions of WRDS financial ratios (WFR). PC1 has the highest t-statistic for both sets of portfolios.



(a) 50 anomalies **(b)** WFR portfolios

Figure 8: L^1 Sparsity of models with interactions. We show the maximum OOS cross-sectional R^2 attained by a model with n factors (on the x-axis) across all possible values of the prior root expected SR^2 (κ) for models based on interactions of original characteristics portfolios (solid) and PCs (dashed). Panel (a) focuses on the SDF constructed from PCs of interactions of 50 anomaly portfolios. Panel (b) shows coefficient estimates corresponding to PCs based on interactions of WFR portfolios. Dotted lines depict -1 s.e. bounds of the CV estimator.

PC1 is also the last survivor if one imposes enough sparsity that only one PC remains. The estimated SDF coefficients are quite similar for many of the other PCs in this table that are ranked lower than PC1 in terms of their t-statistic. However, since these other PCs have lower variance, their contribution to SDF variance, and hence the overall squared Sharpe Ratio captured by the SDF, is lower.

The two plots in Figure 8 take a cut in the contour plots along the ridge of maximal OOS R^2 from bottom to top where we vary sparsity and choose the L^2 optimal shrinkage for each level of sparsity. These plots reinforce the point we noted from the contour plots that many of the powers and interactions of the characteristics are not adding pricing-relevant information to the SDF and can be omitted. The SDF which attains the highest OOS R^2 is relatively sparse with about 200 factors for both the anomalies in Figure 8a and the WFR portfolios in Figure 8b. However, as the wide standard error bands show, statistical precision is quite low. The very large number of portfolios in this case pushes the method to its statistical limits.

Overall, these results show that many of the powers and interactions of characteristics seem to be redundant in terms of their pricing implications. A majority of them can be excluded from the SDF without adverse impact on OOS pricing performance. But, as before,

L^2-shrinkage is crucial for obtaining good OOS performance.

5 Asset Pricing Tests: Performance Compared with Sparse Models

Our cross-validation method evaluates a model's performance on the part of a sample not used in the estimation of the SDF coefficients; it is, therefore, by construction an OOS metric. Yet our choice of the strength of regularization (L^1 and L^2 penalties) is based on the entire sample. It is possible that the penalty that is optimal within one sample does not generalize well on new or fully withheld data. To address this potential issue we now conduct a pure OOS test. Using our L^2-penalty method, we conduct the entire estimation, including the choice of penalty, based on data until the end of 2004. Post-2004 data is completely left out of the estimation. We evaluate performance of this SDF in the 2005–2016 OOS period. This analysis also allows us to assess the statistical significance of our earlier claim that characteristics-sparse SDFs cannot adequately describe the cross-section of stock returns.

This OOS exercise further helps to gain robustness against the effects of data mining in prior published research. Especially for the data set of 50 known anomalies, there is a concern that the full-sample average returns may not be representative of the ex-ante expected returns of these largely ex-post selected portfolios. Implicitly, our analysis so far has already employed some safeguards against data mining bias. For data-mined spurious anomalies, there is no economic reason why their average returns should be related to exposures to high-variance PCs—and if they are not, our L^2 and dual-penalty specifications strongly shrink their contribution to the SDF. Even so, an OOS test on a fully withheld sample of post-2004 data provides additional assurance that the results are not unduly driven by data-mined anomalies.

We proceed as follows. We first orthogonalize all managed portfolio returns with respect to the market using βs estimated in the pre-2005 sample.[16] Given the estimate $\hat{b}$ based on our L^2-penalty Bayesian method in this sample, we construct the time-series of the implied mean-variance efficient (MVE) portfolio $P_t = \hat{b}'F_t$ in the 2005–2016 OOS period. We focus on three sets of portfolios in constructing an SDF: the 50 anomaly portfolios, the 80 WFR portfolios, and the interactions and powers of 50 anomaly characteristics.[17] As in our earlier estimation, we choose penalties by 3-fold cross-validation (2-fold if interactions are included),

[16]The resulting abnormal returns are $F_{i,t} = \tilde{F}_{i,t} - \beta_i R_{m,t}$ where $\tilde{F}_{i,t}$ is the raw portfolio return. In our previous analysis, we used the full data to estimate β_i.

[17]We do not report results for interactions of WFR portfolios due to issues in estimating covariances in an even shorter sample with an extremely high number of characteristics-based factors in this case.

Table 4: MVE portfolio's annualized OOS α in the withheld sample (2005-2016), %

The table shows annualized alphas (in %) computed from the time-series regression of the SDF-implied OOS-MVE portfolio's returns (based on L^2 shrinkage only) relative to four restricted benchmarks: CAPM, Fama-French 5-factor model, optimal sparse model with 5 factors, and optimal PC-sparse model with at most 5 PC-based factors. MVE portfolio returns are normalized to have the same standard deviation as the aggregate market. Standard errors in parentheses.

SDF factors \ Benchmark	CAPM	FF 5-factor	Char.-sparse	PC-sparse
50 anomaly portfolios	14.34	10.58	11.58	3.21
	(5.67)	(5.32)	(4.42)	(2.16)
80 WFR portfolios	20.23	20.19	17.26	2.88
	(5.67)	(5.70)	(5.42)	(2.92)
1,375 interactions of anomalies	27.38	24.97	24.76	13.54
	(5.67)	(5.56)	(5.47)	(3.31)

but with shorter blocks because we only use the pre-2005 sample here.[18]

We then estimate abnormal returns of this OOS-MVE portfolio with respect to three characteristics-based benchmarks: CAPM; the 5-factor model of Fama and French (2016); and our dual-penalty model where we have set the L^1 penalty such that the SDF contains only 5 characteristics-based factors. To compare the models on equal footing, we construct the MVE portfolio implied by these benchmarks. Since we work with candidate factor returns orthogonalized to the market return, the benchmark in the CAPM case is simply a mean return of zero. For Fama-French 5-factor model, we estimate the unregularized MVE portfolio weights, $\hat{w} = \hat{\Sigma}^{-1}\hat{\mu}$, from the 5 non-market factors in the pre-2005 period.[19] We then apply these weights to the 5 factor returns in the OOS period to construct a single benchmark return. Finally, for the dual-penalty sparse model with 5 factors, we estimate $\hat{b}$ in the pre-2005 period and then apply these optimal portfolio weights to returns in the OOS period. If our earlier claim is correct that the SDF cannot be summarized by a small number of characteristics-based factors, then our OOS-MVE portfolio constructed from the full set of candidate factors should generate abnormal returns relative to the MVE portfolio constructed from these sparse benchmarks.

Table 4 confirms that the MVE portfolio implied by our SDF performs well in the withheld data. The table presents the intercepts (alphas) from time-series regressions of the OOS-MVE portfolio returns on the benchmark portfolio return in %, annualized, with standard

[18] We plot the time-series of returns of the MVE portfolios in Figure 13 in the Appendix.

[19] As before, we orthogonalize these factors (SMB, HML, UMD, RMW, CMA) with respect to the market using βs estimated in the pre-2005 sample.

errors in parentheses. To facilitate interpretation of magnitudes, we scale MVE portfolio returns so that they have the same standard deviation as the market index return in the OOS period. The first column shows that the OOS-MVE portfolio offers a large abnormal return relative to the CAPM for all three sets of candidate factor returns. For example, for the OOS-MVE portfolio based on the 50 anomalies, we estimate an abnormal return of 14.34% which is more than two standard errors from zero, despite the short length of the evaluation sample. The abnormal returns are even larger for the other two sets of portfolios. As the second column shows, the abnormal returns are very similar in magnitude for the FF 5-factor model and we can reject the hypothesis of zero abnormal returns at a 5% level or less for all three sets of candidate factor portfolios. The third column shows that the results for the sparse 5-factor model based on our dual-penalty method is almost identical to the FF 5-factor model. Overall, the evidence in this table confirms our claim that characteristics-sparse models do not adequately describe the cross-section of expected stock returns.

In our earlier analysis, we also found that sparse models based on PCs do much better than sparse characteristics-based models. This result also holds up in this OOS analysis. The last column shows that the PC-sparse MVE portfolio, which includes only 5 optimally-selected PC-based factors using our dual-penalty method, performs uniformly better than characteristics-sparse models. Abnormal returns are much smaller and in two cases (50 anomaly portfolios and 80 WFR portfolios), not statistically significantly different from zero.

6 Conclusion

Our results suggest that the multi-decade quest to summarize the cross-section of stock returns with sparse characteristics-based factor models containing only a few (e.g., 3, 4, or 5) characteristics-based factors is ultimately futile. There is simply not enough redundancy among the large number of cross-sectional return predictors that have appeared in the literature for such a characteristics-sparse model to adequately price the cross-section. To perform well, the SDF needs to load on a large number of characteristics-based factors. Sparsity is generally elusive.

In this high-dimensional setting, shrinkage of estimated SDF coefficients towards zero is critical for finding an SDF representation that performs well out-of-sample. L^2-penalty (ridge) based methods that shrink, but do not set to zero, the contributions of candidate factors to the SDF work very well. In contrast, purely L^1-penalty (lasso) based techniques perform poorly because they tend to impose sparsity even where there is none. For some data sets—e.g., one where we include an extremely large number of interactions and powers

of stock characteristics—inclusion of the L^1-penalty in combination with an L^2-penalty can help eliminate some useless factors, but the L^2-penalty is still most important for out-of-sample performance and the number of required factors in the SDF is still very large.

In addition to being empirically successful, the L^2-penalty approach also has an economic motivation. We derive our particular L^2-penalty specification from an economically plausible prior that existence of near-arbitrage opportunities is implausible and major sources of return co-movement are the most likely sources of expected return premia. Lasso-style L^1-penalty approaches, on the other hand, lack such an economic justification.

In line with this economic motivation, a sparse SDF approximation is achievable if one seeks it in the space of principal components of characteristics-based portfolio returns, rather than raw characteristics-sorted portfolio returns. A relatively small number of high-variance principal components in the SDF typically suffices to achieve good out-of-sample performance. This approach inherently still uses all characteristics (factors) in constructing an optimal SDF, but distilling their SDF contributions in a few principal components factors can be fruitful for future research on the economic interpretation of the SDF. Researchers can focus their efforts on linking these few factors to sources of economic risk or investor sentiment.

The mean-variance efficient portfolio implied by our estimated SDF can also serve as a useful test asset to evaluate any potential model of the cross-section of equity returns. This portfolio summarizes the pricing information contained in a large number of characteristics-based factors, and a candidate factor model can be tested in a single time-series regression. In an application of this sort, we have shown that the 5-factor model of Fama and French (2016) leaves much of the cross-section of equity returns unexplained.

References

Asness, C. S., A. Frazzini, and L. H. Pedersen (2014). Quality minus junk. Technical report, Copenhagen Business School.

Asness, C. S., T. J. Moskowitz, and L. H. Pedersen (2013). Value and momentum everywhere. *Journal of Finance*, 929–985.

Barillas, F. and J. Shanken (2017). Comparing asset pricing models. *Journal of Finance*, forthcoming.

Cochrane, J. H. (2011). Presidential address: Discount rates. *Journal of Finance 66*(4), 1047–1108.

DeMiguel, V., L. Garlappi, F. J. Nogales, and R. Uppal (2009). A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science 55*(5), 798–812.

DeMiguel, V., A. Martin-Utrera, F. J. Nogales, and R. Uppal (2017). A portfolio perspective on the multitude of firm characteristics.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*, 23–49.

Fama, E. F. and K. R. French (2008). Dissecting anomalies. *The Journal of Finance 63*(4), 1653–1678.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of Financial Economics 116*(1), 1–22.

Fama, E. F. and K. R. French (2016). Dissecting anomalies with a five-factor model. *The Review of Financial Studies 29*(1), 69–103.

Fan, J., Y. Liao, and W. Wang (2016). Projected principal component analysis in factor models. *Annals of statistics 44*(1), 219.

Feng, G., S. Giglio, and D. Xiu (2017). Taming the factor zoo. Technical report, University of Chicago.

Freyberger, J., A. Neuhierl, and M. Weber (2017). Dissecting characteristics nonparametrically. Technical report, University of Chicago.

Green, J., J. R. Hand, and X. F. Zhang (2017). The characteristics that provide independent information about average us monthly stock returns. *Review of Financial Studies*, hhx019.

Hansen, L. P. and R. Jagannathan (1991). Implications of security market data for models of dynamic economies. *Journal of Political Economy 99*, 225–262.

Hansen, L. P. and R. Jagannathan (1997). Assessing specification errors in stochastic discount factor models. *Journal of Finance 52*, 557–590.

Harvey, C. R., J. C. Liechty, and M. W. Liechty (2008). Bayes vs. resampling: a rematch. *Journal of Investment Management 6 No. 1*, 29–45.

Harvey, C. R., Y. Liu, and H. Zhu (2015). ... and the cross-section of expected returns. *Review of Financial Studies 29*(1), 5–68.

Hastie, T. J., R. J. Tibshirani, and J. H. Friedman (2011). *The elements of statistical learning: data mining, inference, and prediction*. Springer.

Hou, K., C. Xue, and L. Zhang (2015). Digesting anomalies: An investment approach. *The Review of Financial Studies 28*(3), 650–705.

Huerta, R., F. Corbacho, and C. Elkan (2013). Nonlinear support vector machines can systematically identify stocks with high and low future returns. *Algorithmic Finance 2*(1), 45–58.

Kelly, B. T., S. Pruitt, and Y. Su (2017). Some characteristics are risk exposures, and the rest are irrelevant.

Kogan, L. and M. Tian (2015). Firm characteristics and empirical factor models: a model-mining experiment. Technical report, MIT.

Kozak, S., S. Nagel, and S. Santosh (2017). Interpreting factor models. *Journal of Finance*.

Ledoit, O. and M. Wolf (2004a). A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices. *Journal of Multivariate Analysis 88*(2), 365–411.

Ledoit, O. and M. Wolf (2004b). Honey, I Shrunk the Sample Covariance Matrix. *The Journal of Portfolio Management 30*(4), 110–119.

Lewellen, J., S. Nagel, and J. Shanken (2010). A skeptical appraisal of asset-pricing tests. *Journal of Financial Economics 96*, 175–194.

Lin, X. and L. Zhang (2013). The investment manifesto. *Journal of Monetary Economics 60*, 351–66.

MacKinlay, A. C. (1995). Multifactor models do not explain deviations from the capm. *Journal of Financial Economics 38*(1), 3–28.

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

Moritz, B. and T. Zimmermann (2016). Tree-based conditional portfolio sorts: The relation between past and future stock returns. Technical report, Federal Reserve Board.

Novy-Marx, R. and M. Velikov (2016). A taxonomy of anomalies and their trading costs. *Review of Financial Studies 29*(1), 104–147.

Pástor, L. (2000). Portfolio selection and asset pricing models. *Journal of Finance 55*(1), 179–223.

Pástor, L. and R. F. Stambaugh (2000). Comparing asset pricing models: an investment perspective. *Journal of Financial Economics 56*(3), 335–381.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tsai, C.-F., Y.-C. Lin, D. C. Yen, and Y.-M. Chen (2011). Predicting stock returns by classifier ensembles. *Applied Soft Computing 11*(2), 2452–2459.

Vuolteenaho, T. (2002). What drives firm-level stock returns? *Journal of Finance 52*, 233–264.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 301–320.

Appendix

A Properties of the Naive SDF Coefficient Estimator

Consider an orthogonal rotation $P_t = Q'F_t$ with $\Sigma_T = QD_TQ'$, Q is the matrix of eigenvectors of Σ_T and D_T is the sample diagonal matrix of eigenvalues, d_j, ordered in decreasing magnitude. If we express the SDF as $M_t = 1 - b_P' (P_t - \mathbb{E}P_t)$ we have

$$\hat{b}_P = \left(\frac{T-N-2}{T}\right) D_T^{-1} \bar{\mu}_P. \tag{34}$$

Consider the analytically simple case when D is known and replace $\left(\frac{T-N-2}{T}\right) D_T^{-1}$ with D^{-1}.[20] Then we have

$$\sqrt{T}\left(\hat{b}_P - b_P\right) \sim \mathcal{N}\left(0,\, D^{-1}\right), \tag{35}$$

which shows that estimated SDF coefficients on small-eigenvalue PCs (small d_i) have explosive uncertainty.

The above results give exact small sample distributions, assuming returns are jointly normal. As a simple robustness exercise, consider dividing the data into $k = 5$ sub-samples and estimating b_P separately in each.[21] Then we can compute the theoretical variance of these estimates is simply,

$$\text{var}\left(\hat{b}\right) = \frac{k}{T} D^{-1}, \tag{36}$$

which is larger than in eq. (35) by a factor of k due to the shorter samples. Figure 9 plots the sample values of var $\left(\hat{b}_i\right)$ vs d_i^{-1} (on a log-log scale) for the PCs of the 50 anomaly portfolios we use in Section 4. The solid line plots the relationship derived in eq. (36). The good fit confirms that the theoretical relationship given in eq. (35) is valid even with non-normally distributed actual return data.[22] Notice that the ratio of largest to smallest eigenvalue is of the order 10^3. This implies that the variance of the estimated b associated with the smallest eigenvalue portfolio has 3 orders of magnitude larger sampling variance as the b associated with the largest eigenvalue portfolio.

This problem is somewhat exacerbated when D^{-1} is unknown, and thus, estimated. It is well known that the sample eigenvalues of D (equivalently, Σ) are "over-dispersed" relative to true eigenvalues, especially when the number of characteristics, H, is comparable to the sample size, T. This implies that, on average, the smallest estimated eigenvalue is too small and hence the corresponding $\hat{b}_i$ has even greater variance than shown above. In Appendix B we discuss covariance estimation uncertainty.

B Covariance Estimation Uncertainty

In the prior analyses, we have treated covariances (Σ and D) as known. Many papers highlight the empirical difficulty in accurately estimating covariance matrices when the number of assets, H, is of the same order of magnitude as the number of time periods, T. In our main estimation with

[20] With high-frequency data (daily) and even hundreds of factors, D^{-1} is estimated quite well as measured by the loss function tr $\left(D_T^{-1}D - I\right)^2 / N^2$.

[21] Throughout, we assume D is known. For this exercise, we estimate D from the full sample.

[22] This is simply an example of the central limit theorem in full effect.

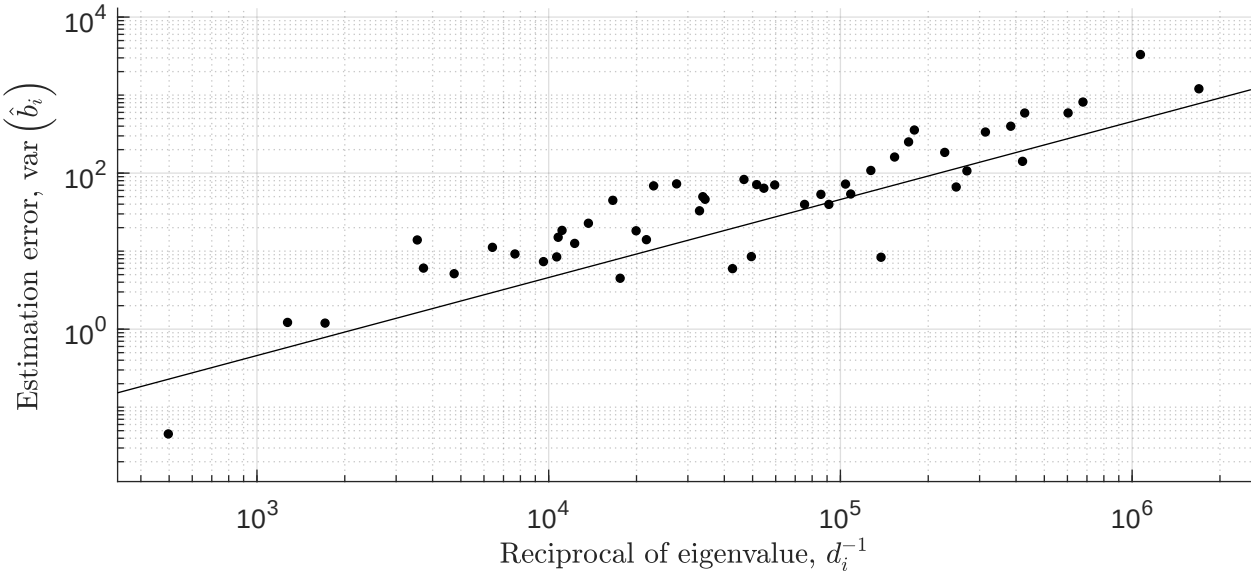


Figure 9: Sampling variance of b. The figure shows sample values of var $\left(\hat{b}_i\right)$ vs reciprocal eigenvalue d_i^{-1} (on a log-log scale) for the PCs of the 50 anomaly portfolios we use in Section 4. The solid line plots the theoretical relationship derived in eq. (36).

anomalies, this should not be of great concern, since $H = 50$ and $T \approx 11,000$. Still, we now analyze methods for dealing with covariance uncertainty in our empirical setting.

In a series of papers, Ledoit and Wolf (L&W) propose robust estimators of Σ which trade off small sample bias and variance by (asymptotically) optimally shrinking the sample covariance towards an a priori target.[23] They are conceptually similar but use different shrinkage targets, Σ_0:

$$\hat{\Sigma} = a\Sigma_0 + (1-a)\,\Sigma_T$$

One choice of Σ_0 is the diagonal matrix $\frac{\text{tr}(\Sigma_T)}{H} I$. The other preserves sample variances, but all correlations are set to $\bar{\rho}$, the average correlation coefficient extracted from Σ_T. The shrinkage parameter, a, is chosen to optimally balance bias and variance (to minimize estimated RMSE), given the choice of Σ_0. The scaled identity matrix proposed in Ledoit and Wolf (2004a) is most appropriate in our empirical setting of zero-β anomaly portfolios. We implement their algorithm on the 50 anomaly portfolios and find $a \approx 0.7\%$ for both methods. Ledoit and Wolf "concentrate on the covariance matrix alone without worrying about expected returns." Hence, they set $\hat{\mu} = \mu_T$. The final estimator of SDF coefficients is

$$\hat{b} = (a\Sigma_0 + (1-a)\,\Sigma_T)^{-1}\mu_T,$$

which appears similar to our estimator given in eq. (22).

A fully Bayesian approach (which delivers similar results) is to specify a Wishart prior for Σ^{-1}, with a "flat" prior on μ, $p(\mu|\,\Sigma) \propto 1$, with

$$\Sigma^{-1} \sim \mathcal{W}\left(H,\, \frac{1}{H}\Sigma_0^{-1}\right), \tag{37}$$

[23] See Ledoit and Wolf (2004a), and Ledoit and Wolf (2004b).

where $\Sigma_0 = \frac{1}{H}\text{tr}\left(\Sigma_T\right) I$, which ensures the total expected variation under the prior matches the data, as in the L&W method. Setting the degrees of freedom to H makes the prior relatively "diffuse." For any choice of Σ_0, the posterior is given by

$$\Sigma^{-1} \sim \mathcal{W}\left(H+T,\, [H\Sigma_0 + T\Sigma_T]^{-1}\right),$$

with expected value

$$\mathbb{E}\left(\Sigma^{-1}\right) = \left[\left(\frac{H}{H+T}\right)\Sigma_0 + \left(\frac{T}{H+T}\right)\Sigma_T\right]^{-1}.$$

For the 50 anomaly portfolios, $\frac{H}{H+T} \approx 0.5\%$, similar to the shrinkage coefficient of the L&W method. We augment this with a "flat" prior on μ so that $\hat{\mu} = \mu_T$. The final estimator of SDF coefficients is

$$\hat{b} = \left[\left(\frac{H}{H+T}\right)\Sigma_0 + \left(\frac{T}{H+T}\right)\Sigma_T\right]^{-1}\mu_T,$$

which is the same as the L&W estimator except that the shrinkage constant is now deterministic.

Both the L&W method and the Bayesian approach address the known phenomenon that eigenvalues of sample covariance matrices are "over-dispersed." That is, the largest estimated eigenvalue tends to be too large while the smallest is too small. Both methods end up shrinking all eigenvalues towards the average, $\bar{d} = \frac{1}{H}\text{tr}\left(\Sigma_T\right)$, while preserving the eigenvectors, Q. Since both use a flat prior for μ, they explicitly do not address uncertainty in estimating means.

Figure 10a shows the relative shrinkage applied to each PC portfolio of the anomalies (our main dataset) for the L&W, Wishart, and our mean-shrinkage method given by eq. (22). We define relative shrinkage as $\frac{\hat{b}_{P,j}}{\hat{b}^{\text{ols}}_{P,j}}$, with $\hat{b}^{\text{ols}}_P = Q'\Sigma_T^{-1}\bar{\mu}$. For comparison, we include the P&S "level" shrinkage of Pástor and Stambaugh (2000), which corresponds to our $\eta = 1$ prior.[24] That plot shows that this prior shrinks all coefficients uniformly towards zero.[25] The L&W and Wishart methods deliver very similar estimators. Importantly, these covariance shrinkage methods are characteristically different from our method (KNS) though they appear superficially similar. Whereas we shrink all coefficients, with greater shrinkage applied to smaller PCs, those methods actually slightly inflate the SDF coefficients associated with large PCs and apply much less shrinkage to small PCs. Indeed, for the smallest PC, the ratio of the L&W estimator to our estimator is approximately equal to $1,700$.

B.1 Σ and μ both uncertain

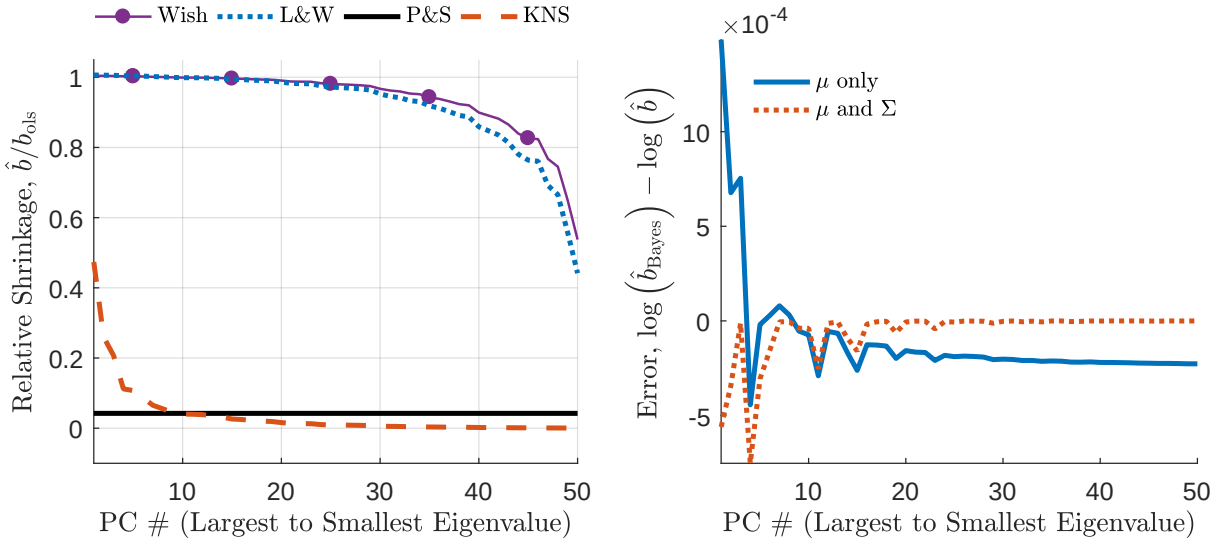
We now analyze the impact of recognizing uncertainty in both μ and Σ. As in our main estimation, we specify

$$\mu|\,\Sigma \sim \mathcal{N}\left(0,\, \frac{\kappa^2}{\tau}\Sigma^2\right), \tag{38}$$

where $\tau = \text{tr}\left(\Sigma_T\right)$. For Σ, we use a similar prior to eq. (37), with a slight modification for numerical tractability since the posterior is not fully analytic. First, we assume eigenvectors (but

[24] We repeat the cross-validation exercise using the prior $\eta = 1$, which induces the posterior estimate $\hat{\mu} = \frac{1}{1+\gamma}\mu_T$. For this shrinkage, the cross-validated optimum is attained at $\frac{1}{1+\gamma} \approx 4.3\%$.

[25] The degree of shrinkage is determined by cross-validation, as described in Section 3.3.



(a) Comparing Methods **(b)** Approximate Bayes

Figure 10: Relative Shrinkage by Method. Panel (a) plots the ratio of regularized estimates of PC SDF coefficients to OLS estimates for various methods. Panel (b) plots the relative difference between the fully Bayesian estimates taking into consideration uncertainty in both μ and Σ and two alternative estimators. The line "μ only" represents the estimator which treats the sample covariance matrix as the truth. The line "μ and Σ" represents the approximate Bayesian solution which first computes the posterior variance assuming sample means are the true means, then computes posterior means assuming the posterior variance is the true variance.

not eigenvalues) are known a priori, so the return covariance matrix can be orthogonalized. Let D be the covariance of PC portfolios. The marginal prior for each PC (each diagonal element of D^{-1}) is an independent scaled inverse-chi squared priors. Let $\sigma^2 = \text{tr}\left(D_T\right)/H$, where D_T is the sample covariance matrix of eigen-portfolios. Under the identity Wishart prior for D^{-1} (with known μ), we had $\mathbb{E}_{\text{prior}}\left(d_i^{-1}\right) = \sigma^2$. The independent priors can be constructed by letting each diagonal element of D^{-1} have a Wishart prior with the same parameters, except to collapse the distribution to one-dimensional:

$$d_i^{-1} \sim \mathcal{W}\left(H, \frac{1}{H}\frac{1}{\sigma}\right),$$

which preserves the level of uncertainty (degrees of freedom) relative to eq. (37). The assumption that eigenvectors are known implies that off-diagonals of D are set to identically 0 under the prior (and hence under the posterior). Along with conditional independence of $\mu|\, D$, this assumption implies that the prior, likelihood, and posterior can be factored into independent terms, one for each PC. Hence inference can be done PC-by-PC instead of jointly.[26]

We also consider an approximation given by the following procedure: first regularize the covariance matrix according to the Wishart prior, eq. (37). Then, we estimate $\hat{b}$ treating the covariance matrix as known. This method is fully analytic and closely approximates the fully Bayesian solution. Figure 10b shows the ratio of the full Bayes estimate to the approximate Bayes estimate,

[26]Since $\mu|\, D$ is multivariate normal with zero correlation across PCs, the elements of μ are conditionally independent.

and to the estimator which ignores covariance uncertainty, $\hat{b}_P = (D_T + \gamma I)^{-1} \bar{\mu}$ with $\gamma = \frac{\tau}{\kappa^2 T}$. As the figure shows, even the simple estimator which treats covariances as known provides a good approximation to the (numerically solved) Bayesian solution. The approximate solution is even better, delivering nearly identical estimates. Throughout our empirical work we use this approximate solution, since covariance uncertainty is potentially important when we consider thousands of portfolios in Section 4.3.

C Interpreting Interactions

What is the economic interpretation of interactions portfolios? For simplicity, consider two binary strategies with characteristic values that can be either high or low (± 1). Let z_s^1 and z_s^2 be the characteristic values for stock s. The pair $\{z_s^1\, z_s^2\}$ takes on four values, shown in the table below:

$z_s^1 \backslash z_s^2$	-1	$+1$
$+1$	A	B
-1	C	D

The letters A to D are names attached to each cell. Let μ_i, $i \in \{A, B, C, D\}$ by the mean returns of stocks in each cell. For simplicity, suppose the characteristics are uncorrelated so that each cell contains the same number of firms. Further, suppose returns are cross-sectionally demeaned (equivalent to including a time fixed-effect, or an equal-weight market portfolio factor). What is the expected return on the z_s^1 mimicking portfolio? That is, what is $\lambda_1 \equiv \mathbb{E}\left[z_s^1 R_s\right]$? Simply $\frac{1}{2}(\mu_A + \mu_B - \mu_C - \mu_D)$. Similarly, $\lambda_2 \equiv \mathbb{E}\left[z_s^2 R_s\right] = \frac{1}{2}(-\mu_A + \mu_B - \mu_C + \mu_D)$ and $\lambda_{12} \equiv \mathbb{E}\left[\left(z_s^1 z_s^2\right) R_s\right] = \frac{1}{2}(-\mu_A + \mu_B + \mu_C - \mu_D)$. Aggregate market clearing implies $(\mu_A + \mu_B + \mu_C + \mu_D) = 0$, so we can easily recover μ_i from knowledge of $\lambda_1, \lambda_2, \lambda_{12}$ by the identity

$$\lambda \equiv \begin{bmatrix} 0 \\ \lambda_1 \\ \lambda_2 \\ \lambda_{12} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ -1 & 1 & -1 & 1 \\ -1 & 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_A \\ \mu_B \\ \mu_C \\ \mu_D \end{bmatrix} = G\mu \tag{39}$$

since the matrix is invertible, where the first equation imposes market clearing (all our assets are market neutral, so the total risk premium on the portfolio of all stocks in the economy is zero).

Given the three managed portfolios, how would we construct something like the "small×value" strategy which buys small-value stocks and shorts small-growth stocks?[27] If z^1 measures market capitalization and z^2 measures BE/ME, the strategy is long D and short C. Let G be the square matrix in eq. (39). The mean of the desired strategy is $\mu_D - \mu_C$, which is also equal to

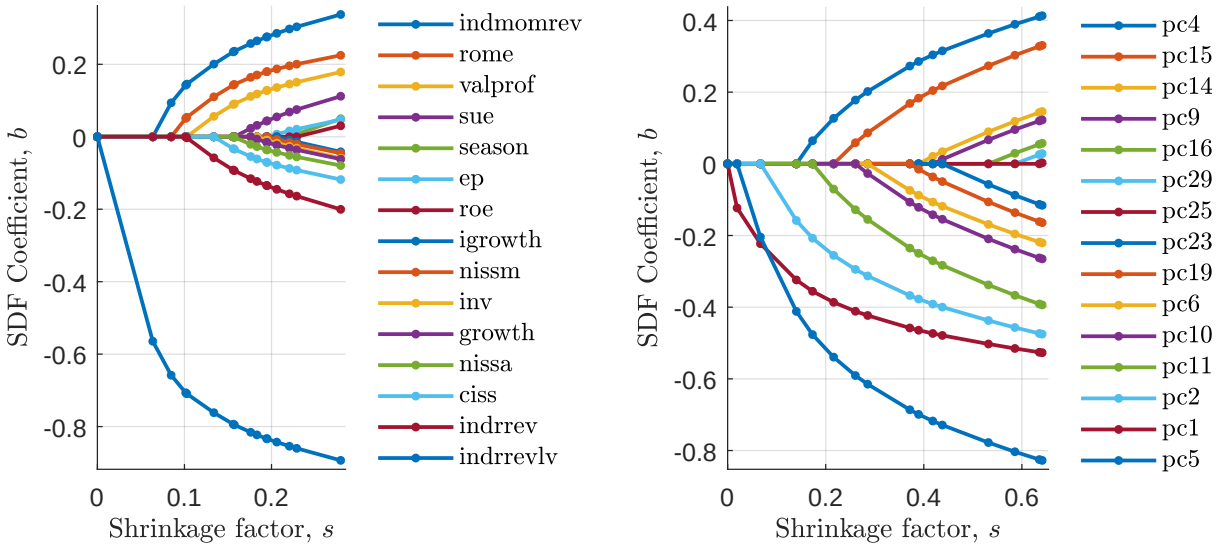
$$\mu_D - \mu_C = \iota_{DC}' G^{-1} \lambda$$

where $\iota_{DC} = \begin{bmatrix} 0 & 0 & -1 & 1 \end{bmatrix}'$, which shows the desired strategy of long D and short C can be constructed with weights equal to $\begin{bmatrix} 0 & 0 & 1 & -1 \end{bmatrix}$ on the four managed portfolio strategies.[28] Hence, combining the interaction with the base strategies allows for construction of any "mixed" strategies. Conceptually, what is required is that the managed portfolios form an approximate "basis" of the potential strategies.

[27] The value anomaly is larger for small stocks, which we would like our methodology to recover.

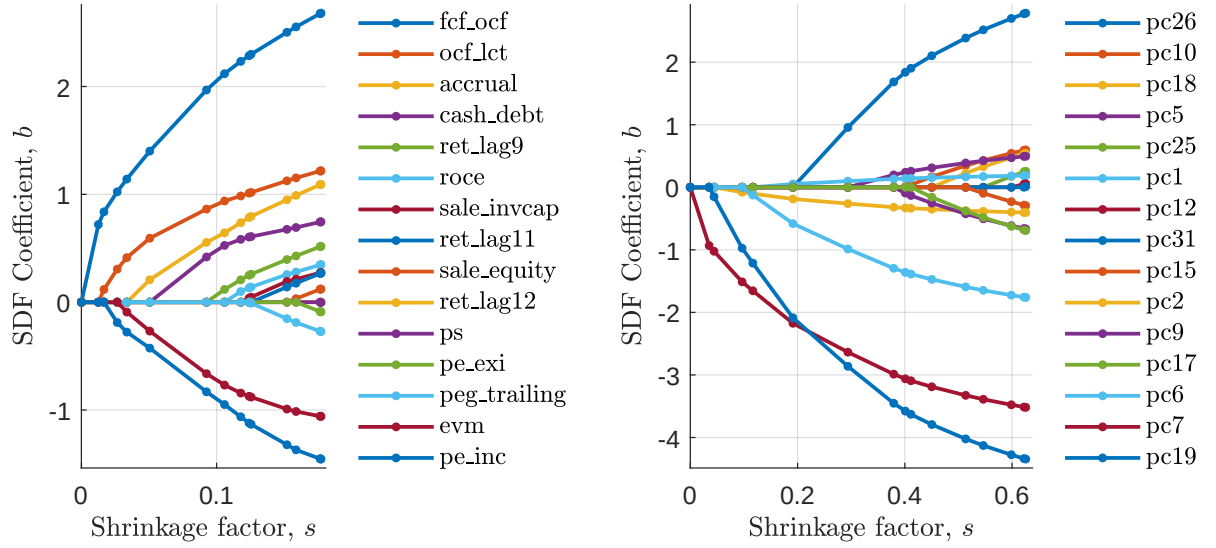
[28] We include the risk-free strategy (with zero excess) return for algebraic convenience.

D Supplementary Plots and Tables



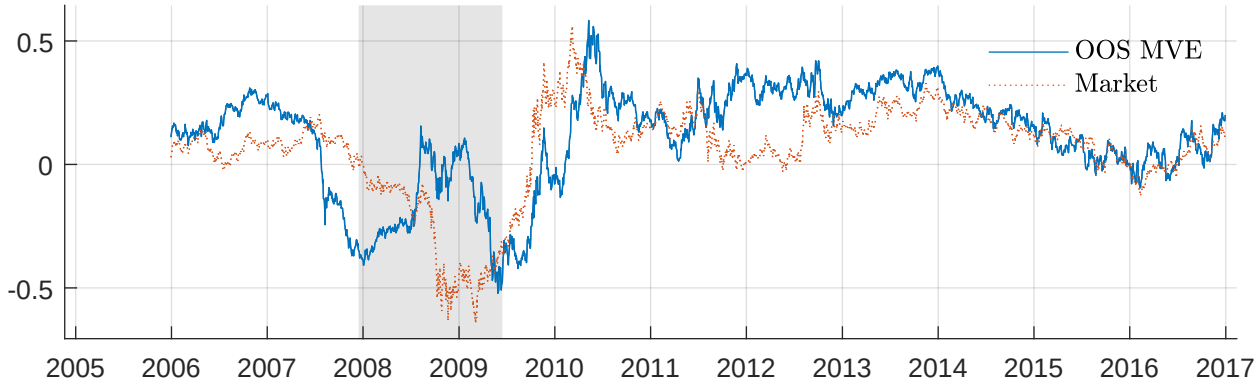
(a) Raw 50 anomaly portfolios **(b)** PCs of 50 anomaly portfolios

Figure 11: L^1 coefficient paths for the optimal model (50 anomaly portfolios). Paths of coefficients based on the optimal (dual-penalty) sparse model that uses 50 anomaly portfolios sorted portfolios (Panel a) and 50 PCs based on anomaly portfolios (Panel b). Labels are ordered according to the vertical ordering of estimates at the right edge of the plot. In Panel b coefficient paths are truncated at the first 15 variables.

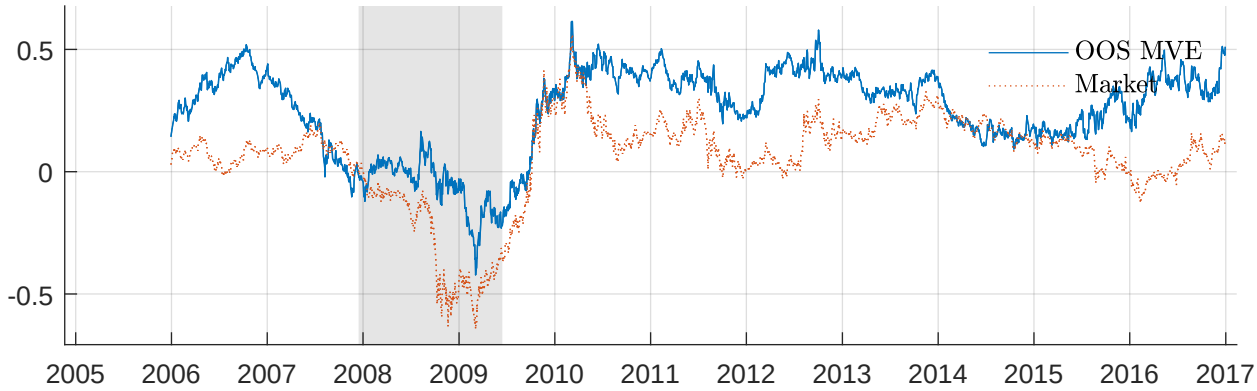


(a) Raw WFR portfolios **(b)** PCs of WFR portfolios

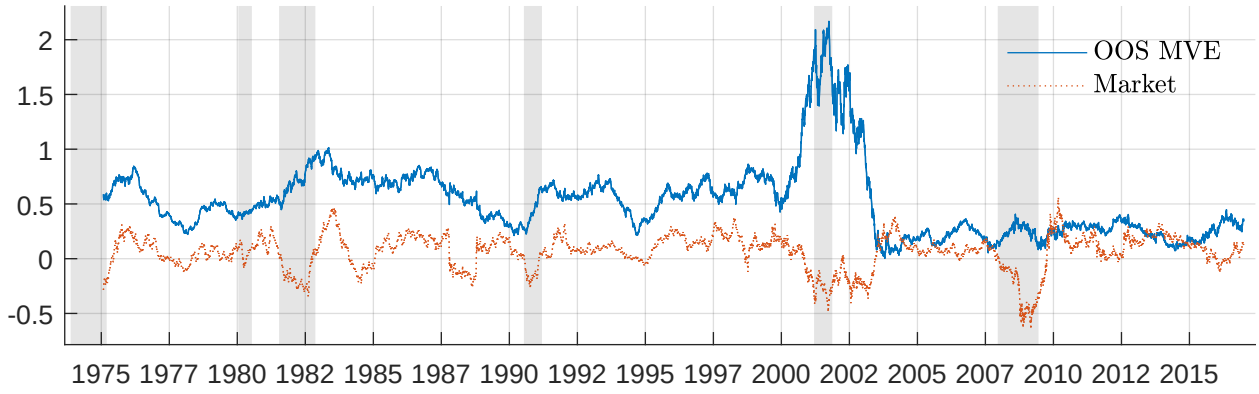
Figure 12: L^1 coefficient paths for the optimal model (WFR portfolios). Paths of coefficients based on the optimal (dual-penalty) sparse model that uses 80 WFR portfolios sorted portfolios (Panel a) and 80 PCs based on WFR portfolios (Panel b). Labels are ordered according to the vertical ordering of estimates at the right edge of the plot.



(a) Returns on MVE portfolio based on 50 anomalies in withheld sample



(b) Returns on MVE portfolio based on interactions of 50 anomalies in withheld sample



(c) Returns on MVE portfolio based on interactions of 50 anomalies in full sample

Figure 13: Time-series of returns on the MVE portfolio. The figure plots the time-series of one-year overlapping returns on the regularized market-neutral MVE portfolio implied by our SDF (blue solid line) and returns on the market (for comparison only; red dashed line). Panel (a) plots MVE portfolio returns in the withheld sample (2005-present) implied by the SDF that was constructed using 50 anomaly portfolios. Panel (b) plots MVE returns in the withheld sample using a model based on interactions of 50 anomalies. Panel (c) plots MVE returns in full sample implied by the model with interactions.

Table 5: Part I: Mean annualized returns on anomaly portfolios, %

The table lists all basic "anomaly" characteristics used in our analysis and shows annualized mean returns on managed portfolios which are linear in characteristics. Columns (1)-(3) show mean annualized returns (in %) for managed portfolios corresponding to all characteristics in the full sample, pre-2005 sample, and post-2005 sample, respectively. All managed portfolios' returns are based on a monthly-rebalanced buy-and-hold strategy and are further rescaled to have standard deviations equal to the in-sample standard deviation of excess returns on the aggregate market index. The sample is daily from May 1, 1974 till December 30, 2016.

	(1)	(2)	(3)
	Full Sample	Pre 2005	Post 2005
1. Size	-2.6	-2.8	-2.0
2. Value (A)	6.9	9.2	0.8
3. Gross Profitability	3.3	2.3	5.6
4. Value-Profitablity	14.0	17.8	4.0
5. F-score	8.1	10.0	3.3
6. Debt Issuance	1.3	1.0	2.2
7. Share Repurchases	7.0	7.6	5.5
8. Net Issuance (A)	-9.6	-11.5	-4.7
9. Accruals	-5.4	-7.8	0.9
10. Asset Growth	-9.2	-11.2	-4.1
11. Asset Turnover	5.1	3.9	8.4
12. Gross Margins	-1.4	0.1	-5.2
13. Dividend/Price	4.0	5.5	-0.0
14. Earnings/Price	8.7	10.7	3.6
15. Cash Flows/Price	8.6	10.4	3.6
16. Net Operating Assets	2.2	3.6	-1.5
17. Investment/Assets	-10.1	-12.6	-3.5
18. Investment/Capital	-4.5	-4.9	-3.2
19. Investment Growth	-9.5	-11.0	-5.8
20. Sales Growth	-6.2	-5.9	-7.0
21. Leverage	5.6	7.8	-0.1
22. Return on Assets (A)	2.0	0.4	6.2
23. Return on Book Equity (A)	4.5	4.7	4.0
24. Sales/Price	10.0	11.6	5.8
25. Growth in LTNOA	-2.4	-1.7	-4.1
26. Momentum (6m)	2.1	4.1	-3.1
27. Industry Momentum	5.8	8.0	0.1
28. Value-Momentum	5.5	7.8	-0.5
29. Value-Momentum-Prof.	6.8	9.6	-0.5
30. Short Interest	0.3	1.6	-2.9

continued on next page...

Table 5: Part II: Mean annualized returns on anomaly portfolios, %

	(1)	(2)	(3)
31. Momentum (12m)	9.2	12.7	-0.0
32. Momentum-Reversals	-6.1	-7.9	-1.6
33. Long Run Reversals	-6.1	-7.9	-1.3
34. Value (M)	6.0	8.0	1.0
35. Net Issuance (M)	-8.8	-9.9	-6.0
36. Earnings Surprises	12.4	15.2	5.1
37. Return on Equity	10.2	12.1	5.4
38. Return on Market Equity	12.4	15.3	5.0
39. Return on Assets	6.7	7.1	5.7
40. Short-Term Reversals	-8.1	-12.0	1.9
41. Idiosyncratic Volatility	-3.1	-3.7	-1.7
42. Beta Arbitrage	-0.8	-0.3	-2.0
43. Seasonality	12.3	18.7	-4.6
44. Industry Rel. Reversals	-18.1	-25.6	1.4
45. Industry Rel. Rev. (L.V.)	-35.7	-47.6	-4.5
46. Ind. Mom-Reversals	20.8	29.3	-1.2
47. Composite Issuance	-8.4	-10.2	-3.6
48. Price	-1.4	-1.1	-2.4
49. Age	3.8	4.7	1.3
50. Share Volume	-1.3	-1.3	-1.2

Table 6: Part I: Mean annualized returns on WFR portfolios, %

The table lists all basic WFR characteristics used in our analysis and shows annualized mean returns on managed portfolios which are linear in characteristics. Columns (1)-(3) show mean annualized returns (in %) for managed portfolios corresponding to all characteristics in the full sample, pre-2005 sample, and post-2005 sample, respectively. All managed portfolios' returns are based on a monthly-rebalanced buy-an-hold strategy and are further rescaled to have standard deviations equal to the in-sample standard deviation of excess returns on the aggregate market index. The sample is daily from May 1, 1974 till December 30, 2016.

	(1)	(2)	(3)
	Full Sample	Pre 2005	Post 2005
1. P/E (Diluted, Excl. EI)	-10.7	-11.8	-7.0
2. P/E (Diluted, Incl. EI)	-13.5	-15.4	-7.3
3. Price/Sales	-8.1	-9.0	-5.1
4. Price/Cash flow	-5.0	-5.0	-4.8
5. Enterprise Value Multiple	-10.5	-11.5	-7.1
6. Book/Market	4.5	5.7	0.4
7. Shillers Cyclically Adjusted P/E Ratio	-5.9	-7.8	0.4
8. Dividend Payout Ratio	-1.6	-1.9	-0.5
9. Net Profit Margin	1.9	2.9	-1.4
10. Operating Profit Margin Before Depreciation	2.3	3.9	-2.8
11. Operating Profit Margin After Depreciation	2.6	4.1	-2.3
12. Gross Profit Margin	1.0	2.5	-4.2
13. Pre-tax Profit Margin	2.7	3.7	-0.7
14. Cash Flow Margin	1.1	1.9	-1.8
15. Return on Assets	6.8	6.8	6.5
16. Return on Equity	7.0	7.6	4.9
17. Return on Capital Employed	8.7	8.6	9.0
18. After-tax Return on Average Common Equity	7.8	8.8	4.4
19. After-tax Return on Invested Capital	5.7	6.0	4.8
20. After-tax Return on Total Stockholders Equity	7.6	8.6	4.3
21. Pre-tax return on Net Operating Assets	7.0	8.1	3.2
22. Pre-tax Return on Total Earning Assets	6.6	7.8	2.7
23. Common Equity/Invested Capital	1.0	1.0	1.3
24. Long-term Debt/Invested Capital	-0.3	-0.1	-0.9
25. Total Debt/Invested Capital	-0.3	0.0	-1.6
26. Interest/Average Long-term Debt	3.5	4.4	0.6
27. Interest/Average Total Debt	3.7	4.3	1.5
28. Cash Balance/Total Liabilities	0.8	1.1	-0.3
29. Inventory/Current Assets	0.2	-0.8	3.8
30. Receivables/Current Assets	0.6	0.4	1.1

continued on next page...

Table 6: Part II: Mean annualized returns on WFR portfolios, %

	(1)	(2)	(3)
31. Total Debt/Total Assets	-2.5	-2.7	-2.0
32. Short-Term Debt/Total Debt	-0.2	0.8	-3.5
33. Current Liabilities/Total Liabilities	2.2	3.0	-0.4
34. Long-term Debt/Total Liabilities	-4.9	-6.0	-1.2
35. Free Cash Flow/Operating Cash Flow	17.0	21.1	3.4
36. Avertising Expenses/Sales	2.1	2.2	1.5
37. Profit Before Depreciation/Current Liabilities	3.3	4.3	-0.2
38. Total Debt/EBITDA	-1.4	-1.1	-2.5
39. Operating CF/Current Liabilities	11.6	14.8	1.0
40. Total Liabilities/Total Tangible Assets	2.9	4.2	-1.4
41. Long-term Debt/Book Equity	-1.1	-1.0	-1.4
42. Total Debt/Total Assets	2.5	2.8	1.5
43. Total Debt/Capital	1.3	1.8	-0.1
44. Total Debt/Equity	2.2	2.5	1.2
45. After-tax Interest Coverage	4.7	5.1	3.5
46. Cash Ratio	0.4	1.2	-2.0
47. Quick Ratio (Acid Test)	-1.6	-1.4	-2.5
48. Current Ratio	-1.9	-2.0	-1.6
49. Capitalization Ratio	-0.2	-0.1	-0.5
50. Cash Flow/Total Debt	11.3	13.0	5.6
51. Inventory Turnover	2.9	3.7	0.1
52. Asset Turnover	6.2	5.7	7.8
53. Receivables Turnover	3.4	3.0	4.9
54. Payables Turnover	-1.9	-4.5	6.6
55. Sales/Invested Capital	8.7	8.3	9.8
56. Sales/Stockholders Equity	7.8	7.9	7.4
57. Sales/Working Capital	3.0	3.6	1.0
58. Research and Development/Sales	3.2	3.6	1.9
59. Accruals/Average Assets	12.5	13.7	8.5
60. Gross Profit/Total Assets	6.4	6.3	6.6
61. Book Equity	1.0	1.4	-0.4
62. Cash Conversion Cycle (Days)	-3.7	-4.2	-2.1
63. Effective Tax Rate	3.8	4.0	3.1
64. Interest Coverage Ratio	6.3	6.6	5.5
65. Labor Expenses/Sales	0.9	1.8	-1.9
66. Dividend Yield	3.7	4.7	0.5
67. Price/Book	-5.5	-6.8	-1.1
68. Trailing P/E to Growth (PEG) ratio	-10.5	-11.9	-5.8
69. Month $t-1$	-8.7	-11.8	1.7
70. Month $t-2$	0.1	0.7	-1.6

continued on next page...

Table 6: Part III: Mean annualized returns on WFR portfolios, %

	(1)	(2)	(3)
71. Month $t-3$	3.2	4.5	-0.9
72. Month $t-4$	3.2	3.1	3.6
73. Month $t-5$	2.7	3.5	0.2
74. Month $t-6$	5.9	8.6	-2.9
75. Month $t-7$	3.2	4.6	-1.4
76. Month $t-8$	3.4	3.6	2.8
77. Month $t-9$	9.9	12.3	1.9
78. Month $t-10$	5.0	8.0	-5.0
79. Month $t-11$	9.0	8.4	10.9
80. Month $t-12$	7.9	10.0	0.8