UNIL | Université de Lausanne
Département de finance

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

FINANCE RESEARCH SEMINAR SUPPORTED BY UNIGESTION

"Shrinking the Cross Section"

Prof. Stefan NAGEL

University of Chicago, Booth School of Business

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.

Friday, September 8, 2017, 10:30-12:00

Room 126, Extranef building at the University of Lausanne

Shrinking the Cross Section*

Serhiy Kozak[†], **Stefan Nagel**[‡], **Shrihari Santosh**[§]

August 28, 2017

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.

*We thank Svetlana Bryzgalova, Mike Chernov, Stefano Giglio, Lars Hansen, Bryan Kelly, Ralph Koijen, participants at the NBER AP workshop, the Brazilian Finance Association Meetings, and seminar participants at City University of Hong Kong, HKUST, Michigan, and UCLA for helpful comments and suggestions.

[†]Stephen M. Ross School of Business, University of Michigan, 701 Tappan St., Ann Arbor, MI 48109, e-mail: sekozak@umich.edu.

[‡]Booth School of Business, University of Chicago, 5807 S. Woodlawn Ave., Chicago, IL 60637, e-mail: stefan.nagel@chicagobooth.edu.

[§]R.H. Smith School of Business, University of Maryland. Email: shrihari@umd.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 (2016) 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.

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 potential 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 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) principal components (PCs) of the candidate factor returns. Put differently, first and second

moments of returns should be related. If a factor earns high expected returns, then 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, as Kozak et al. (2017) show, under plausible conditions that allow mispricing but no near-arbitrage opportunities, in models in which cross-sectional variation in expected returns arises from biased investor beliefs.

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 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 price of risk of zero. 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 speci-

fication employs both L^1 and L^2 penalties, similarly 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 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 is in turn equivalent to portfolio optimization with a covariance matrix shrunk towards the identity matrix as in Ledoit and Wolf (2004). However, despite the similarity of the solutions, there is an important difference. 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, and 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 assumption 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 the latter type of uncertainty is quantitatively negligible once the 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. 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 doesn't 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 Methodology

For any point in time t, let R_t denote an $N \times 1$ vector of excess returns, and Z_{t-1} an $N \times H$ matrix of asset characteristics (with H possibly quite large, potentially thousands of characteristics). Let Z_{t-1} be centered and standardized cross-sectionally at each t.

2.1 SDF

Consider a projection of the true SDF onto the space of excess returns,

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

where b_t is an $N \times 1$ vector of SDF coefficients. We parametrize the coefficients b_t as a linear function of characteristics,

$$b_{t-1} = Z_{t-1} b, \tag{2}$$

where b is an $H \times 1$ vector of time-invariant coefficients. Therefore, rather than estimating SDF coefficients for each stock at each point in time, we estimate them as a single function of characteristics that applies to all stocks over time. The idea behind this approach is similar to Brandt et al. (2009) and DeMiguel et al. (2017). Plugging Eq. 2 into Eq. 1 delivers an SDF that is in the linear span of the H (basis) trading strategy returns $F_t = Z'_{t-1} R_t$ that can be created based on stock characteristics, i.e.,

$$M_t = 1 - b' \left(F_t - \mathbb{E} F_t \right). \tag{3}$$

2.1.1 Rotation

The transformation in Eq. 3 defines a rotation of the space of individual stock returns $R_t \in \mathbb{R}^N$ into the space of "managed portfolios" $F_t \in \mathbb{R}^H$. Z_{t-1} defines a transformation $\mathbb{R}^N \to \mathbb{R}^H$, i.e., maps the space of N individual stock returns into a space of H trading strategies (managed portfolios) as follows:

$$F_t = Z'_{t-1} R_t. \tag{4}$$

This rotation is motivated by an implicit assumption that the characteristics fully capture all aspects of the joint distribution of returns that are relevant for the purpose of constructing an SDF, i.e., that expected returns, variances, and covariances are stable functions of characteristics such as size and book-to-market ratio, and not security names (Cochrane, 2011). This (implicit) assumption was the driving force for using portfolio sorts in cross-sectional

asset pricing in the first place. Managed portfolios allow us to generalize this idea and be more flexible.

Note that even though we assume that all coefficients b are constant, it is without loss of generality, because we can always re-state a model with time-varying b_{t-1} as a model with constant b and an extended set of factors F_t. For instance, suppose we can capture time variation in b_{t-1} by some set of time-series instruments z_{t-1}. Then we can simply rewrite the SDF as $M_t = 1 - b'\tilde{F}_t$, where $\tilde{F}_t = z_{t-1} \otimes F_t$ and $\otimes$ denotes the Kronecker product of two vectors (Brandt et al., 2009, Cochrane, 2005, Ch. 8).

2.1.2 The MVE portfolio

Given the asset pricing equation,

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

in population we could solve for

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

the (SDF) coefficients in a (cross-sectional) projection with H "explanatory variables" and H "dependent variables"; where $\Sigma \equiv \text{cov}\left(F_t\right) = \mathbb{E}\left[\left(F_t - \mathbb{E}F_t\right)\left(F_t - \mathbb{E}F_t\right)'\right]$. The SDF coefficients are also the weights of the mean-variance-efficient (MVE) portfolio. In what follows, consider F_t as residuals from a factor model, so the returns have zero covariance with "known" factors, but potentially have non-zero pricing errors, α (the factor model may be incomplete). Hence, we are trying to estimate how best to extend the baseline model. In our empirical work we treat the CAPM is the "null model". We therefore normalize all managed portfolio to have zero unconditional covariance with the CRSP value-weighted market portfolio.

2.1.3 Sample estimators

Consider a sample with size T, where $T > H$, but possibly $T < N$. We denote

$$\bar{\mu} = \frac{1}{T}\sum_{t=1}^{T} F_t \tag{7}$$

$$\Sigma_T = \frac{1}{T}\sum_{t=1}^{T}\left(F_t - \mu_t\right)\left(F_t - \mu_t\right)' \tag{8}$$

the maximum likelihood estimates (MLE) of means and covariances, respectively.[1] A natural, but naïve, "plug-in" estimator of b is

$$\hat{b} = \left(\frac{T-N-2}{T}\right) \Sigma_T^{-1} \bar{\mu}, \tag{9}$$

where Σ_T^{-1} is the Moore-Penrose pseudo-inverse of Σ_T and $\left(\frac{T-N-2}{T}\right)$ is a bias adjustment due to estimation uncertainty in Σ_T. This estimator is unbiased (under joint normality of returns), but is imprecise.[2]

To see this, 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 $M_t = 1 - b_P' \left(P_t - \mathbb{E}P_t\right)$ we have

$$\hat{b}_P = \left(\frac{T-N-2}{T}\right) D_T^{-1} \bar{\mu}_P. \tag{10}$$

Consider the analytically simple case when D is known and replace $\left(\frac{T-N-2}{T}\right) D_T^{-1}$ with D^{-1}.[3] Then we have

$$\sqrt{T}\left(\hat{b}_P - b_P\right) \sim \mathcal{N}\left(0,\, D^{-1}\right), \tag{11}$$

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.[4] Then we can compute the theoretical variance of these estimates is simply,

$$\text{var}\left(\hat{b}\right) = \frac{k}{T} D^{-1}, \tag{12}$$

which is larger than in Eq. 11 by a factor of k due to the shorter samples. Figure 1 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 3. The solid line plots the relationship derived in Eq. 12. The good fit confirms that the theoretical relationship given in Eq. 11 is valid even with non-normally distributed actual return data.[5] Notice that the ratio of largest to smallest eigenvalue is of

[1] These estimators are MLE under joint normality and time-series independence of returns.

[2] Under joint-normality, $\bar{\mu}$ and Σ_T are independent, and are unbiased estimators of μ and Σ, respectively.

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

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

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

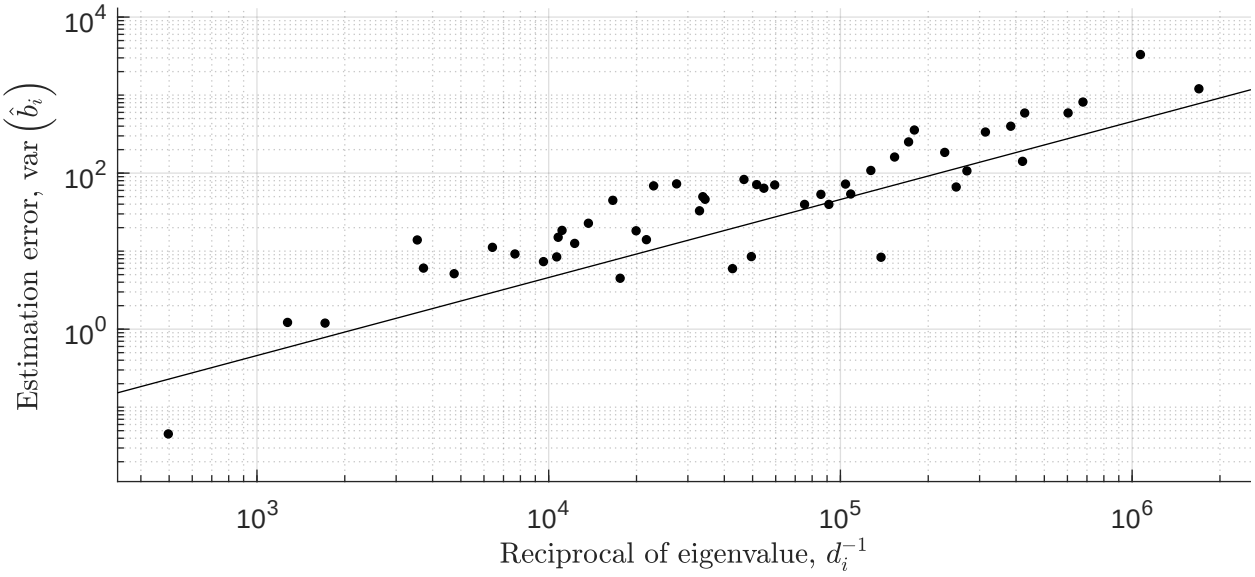


Figure 1: Sampling variance of b. The figure shows sample values of var $\left(\hat{b}_i\right)$ vs reciprocal eigenvalue d_i^{-1} (on a log-log scale) for the PCs of the 50 anomaly portfolios we use in Section 3. The solid line plots the theoretical relationship derived in Eq. 12.

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 Section B.2 we discuss covariance estimation uncertainty.

2.2 Regularization

The ill-conditioned problem of estimating b cries out for regularization. A natural way to address the above unreliability of MLE is Bayesian estimation, which combines additional prior information with the data to generate a posterior distribution for the unknown parameters. If the prior information is well-motivated (truly informative), the resulting bias is more than offset by a reduction in variance, reducing the mean squared error of the estimator. Since $b \equiv \Sigma^{-1}\mu$ and we treat Σ as known, a prior distribution for b maps one-to-one to a prior for μ.[6]

[6]This obtains as long as Σ is not rank deficient (no assets are redundant).

Consider the family of priors,

$$\mu \sim \mathcal{N}\left(0,\, \frac{\kappa^2}{\tau}\Sigma^{\eta}\right), \tag{13}$$

where $\tau = \text{trace}\,[\Sigma]$ and κ is a constant controlling the "scale" of μ and may depend on τ and N. Parameter η controls the "shape" of the prior and is key to controlling the essential characteristic of the posterior estimate (as shown below). This family nests commonly used priors: $\eta = 0$ gives the (conjugate) diffuse prior in Harvey et al. (2008) (and others); $\eta = 1$ gives the "asset pricing" prior in Pástor (2000) and Pástor and Stambaugh (2000); $\eta = 2$ implies the prior we use in this paper (and justify below),

$$\mu \sim \mathcal{N}\left(0,\, \frac{\kappa^2}{\tau}\Sigma^{2}\right), \tag{14}$$

which is novel to our knowledge.

Importantly for us, the family of priors (13) is quite flexible while delivering analytic posterior densities. Finally, the family maps to Tikhonov regularization, a commonly used procedure well suited to the current problem. We discuss this link further in Section 2.2.2. Recall from above, we can form principal component portfolios $P_t = Q'F_t$ with $\Sigma = QDQ'$. Expressing the family of priors (13) in terms of PC portfolios we have:

$$\mu_P \sim \mathcal{N}\left(0,\, \frac{\kappa^2}{\tau}D^{\eta}\right) \tag{15}$$

$$b_P \sim \mathcal{N}\left(0,\, \frac{\kappa^2}{\tau}D^{\eta-2}\right). \tag{16}$$

Considering, specifically, the lowest variance (smallest eigenvalue) PC portfolio (indexed by $j = H$), we obtain:

$$\text{var}\,(\,b_{P,H}|\,\text{prior}) = \frac{\kappa^2}{\tau}d_H^{\eta-2}. \tag{17}$$

To narrow down the potential choices of η, we investigate the limit of Eq. 17 in the case of redundant assets,

$$\lim_{d_H\to 0}\text{var}\,(\,b_{P,H}|\,\text{prior}) = \lim_{d_H\to 0}\frac{\kappa^2}{\tau}d_H^{\eta-2}, \tag{18}$$

and observe that it is finite for $\eta \geq 2$, but divergent for $\eta < 2$.[7] We now argue that "unbounded" SDF coefficients are implausible, even with nearly redundant assets.

[7]Throughout we assume $\kappa > 0$ so that the prior distribution is not degenerate.

Kozak et al. (2017) develop a model in which deviations from the CAPM are due to some (sentiment) investors' distorted demands, induced by distorted beliefs or hedging motives. The remaining investors (arbitrageurs) take the opposite position, mitigating, but not eliminating the equilibrium price impact of the distortion. Importantly, the magnitude of the demand distortion is bounded, motivated by leverage constraints and short-selling costs, which are particularly binding on the "retail" sentiment investors relative to the "institutional" arbitrageurs. In equilibrium, this bound leads to finite arbitrageur holdings, and therefore, finite SDF coefficients. Consider nearly perfectly correlated assets, i and j. If sentiment investors believe i has a higher expected payoff, they would like to buy a nearly infinite quantity of i, funded by short-selling a nearly infinite quantity of j. Alas, the leverage and short-selling constraints prevent this, leading to a finite sized bet. The arbitrageurs elastically accommodate this demand with little compensation, since their resulting long-short portfolio is nearly risk-free (due to the near perfect correlation). Since they require little compensation, equilibrium prices exhibit little distortion, which prevents SDF coefficients from being "too high" in the first place.

The least restrictive prior consistent with finite variance of SDF coefficients is $\eta = 2$, which we showed in Eq. 14. We therefore proceed with that parameter choice, which implies an i.i.d. prior on SDF coefficients, $b \sim \mathcal{N}\left(0, \frac{\kappa^2}{\tau} I\right)$. In this case, the posterior mean of b is

$$\hat{b} = \left[(\Sigma + \gamma I)^{-1} \Sigma\right] \left(\Sigma^{-1} \bar{\mu}\right) \tag{19}$$

$$= (\Sigma + \gamma I)^{-1} \bar{\mu}, \tag{20}$$

where $\gamma = \frac{\tau}{\kappa^2 T}$. In our empirical work we use the second representation since it is numerically more stable with large dimensional Σ. The posterior variance of b is given by

$$\text{var}(b) = \frac{1}{T} (\Sigma + \gamma I)^{-1}, \tag{21}$$

which we use in Section 3 to construct confidence intervals. We further discuss the economics of Eq. 19 below.

2.2.1 Economic interpretation

To dissect Eq. 19, it is convenient to consider a rotation of the original space of returns into the space of principal components as in Section 2.1.3. Fitted means and SDF coefficients of

PC portfolios are given by:

$$\hat{\mu}_{P,j} = \left(\frac{d_j}{d_j + \gamma}\right) \bar{\mu}_{P,j} \tag{22}$$

$$\hat{b}_{P,j} = \left(\frac{d_j}{d_j + \gamma}\right) \frac{\bar{\mu}_{P,j}}{d_j}, \tag{23}$$

i.e., for the j-th PC of returns with a small eigenvalue d_j, its fitted mean $\bar{\mu}_{P,j}$ is forced to be close to zero. Compare this to the "just identified" least squares solution without penalty,

$$\hat{\mu}^{\text{ols}}_{P,j} = \bar{\mu}_{P,j} \tag{24}$$

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

which involves no shrinkage of sample means. Since under our prior ($\eta = 2$) it is unlikely that a low-variance portfolio has high mean, the estimator applies relatively stronger shrinkage to the means of small-eigenvalue PCs. Notice that as $d_j \to 0$, the ratio of our regularized estimator to the least squares estimator goes to zero:

$$\lim_{d_j \to 0} \frac{\hat{b}_{P,j}}{\hat{b}^{\text{ols}}_{P,j}} = 0. \tag{26}$$

Hence, when some assets are nearly redundant, our estimator does not allow explosively large SDF coefficients.

Under the prior, the expected maximum squared Sharpe ratio (variance of the SDF) is given by:

$$\mathbb{E}\left[\mu \Sigma^{-1} \mu\right] = \mathbb{E}\left[\sum_{j=1}^{H} \frac{\mu^2_{P,j}}{d_j}\right] = \sum_{j=1}^{H} \frac{\kappa^2}{\tau} d_j^{\eta - 1}. \tag{27}$$

For $\eta = 1$, this becomes $\sum_{j=1}^{H} \frac{\kappa^2}{\tau} = \frac{H\kappa^2}{\tau}$, showing that each PC portfolio has the same expected contribution. With $\eta = 2$, the expression becomes:

$$\mathbb{E}\left[\mu \Sigma^{-1} \mu\right] = \sum_{j=1}^{H} \frac{\kappa^2}{\tau} d_j = \kappa^2, \tag{28}$$

that is, the expected contribution from each PC is proportional to its eigenvalue, d_i. This aspect of our prior is consistent with the result in Kozak et al. (2017) that "it is not possible to generate much cross-sectional variation in expected returns without having the first [few] principal components ... explain almost all the cross-sectional variation in expected returns."

That is, unless most of total SDF variance comes from large PC portfolios, prices will only be minimally distorted and SDF variance will be low. This "belief" embedded in our prior leads to shrinkage of the contributions of small PCs to the (model-implied) maximum squared Sharpe ratio under the posterior,

$$\text{var}\left(M_t\right) = \max \text{SR}^2 = \hat{b}_P' D \hat{b}_P = \bar{\mu}_P' \left(D + \gamma I\right)^{-1} D \left(D + \gamma I\right)^{-1} \bar{\mu}_P \tag{29}$$

$$= \bar{\mu}_P' D \left(D + \gamma I\right)^{-2} \bar{\mu}_P = \sum_{j=1}^{H} \frac{\left(\bar{\mu}_{P,j}\right)^2}{d_j} \left(\frac{d_j}{d_j + \gamma}\right)^2, \tag{30}$$

where $\hat{b}$ is given in Eq. 20, d_j are diagonal elements of D and $\frac{\left(\bar{\mu}_{P,j}\right)^2}{d_j}$ is the contribution of the j-th PC to the maximum in-sample squared Sharpe ratio. Note that since $\gamma \geq 0$, we have $0 < \frac{d_j}{d_j+\gamma} \leq 1$. The multiplication with $\left(\frac{d_j}{d_j+\gamma}\right)^2$ therefore causes contributions to the max squared SR from low-eigenvalue PCs to be shrunk more than contributions of high-eigenvalue PCs. The economic interpretation of such shrinkage is that we judge as economically implausible the case that a principal component of the candidate factors has high mean return (or high contribution to the total squared Sharpe ratio), but a small eigenvalue.

2.2.2 Penalized Estimation

We now show how our Bayesian prior maps to penalized estimation, more commonly used in machine learning. If we maximize the model cross-sectional R^2 subject to a penalty on the model-implied maximum squared Sharpe ratio $\gamma b' \Sigma b$,

$$\hat{b} = \arg\min_b \left\{ \left(\mu_T - \Sigma b\right)' \left(\mu_T - \Sigma b\right) + \gamma b' \Sigma b \right\}, \tag{31}$$

the problem leads to the solution given by Eq. 20, showing the tight link to our Bayesian approach with $\eta = 2$. Equivalently, we can represent the model in terms of risk premia, $\mathbb{E}\left[F\right]$, instead of risk prices (b's),

$$\hat{\lambda} = \arg\min_\lambda \left\{ \left(\mu_T - I\lambda\right)' \left(\mu_T - I\lambda\right) + \gamma \lambda' \Sigma^{-1} \lambda \right\}, \tag{32}$$

which leads to the solution

$$\hat{\lambda} = \left[\left(\Sigma + \gamma I\right)^{-1} \Sigma\right] \left(\bar{\mu}\right). \tag{33}$$

The shrinkage term, $\left[\left(\Sigma + \gamma I\right)^{-1} \Sigma\right]$, is identical to the previous solution given by Eq. 19.

Yet another alternative specification is to minimize the model HJ-distance (Hansen and

Jagannathan, 1997) subject to an L^2 norm penalty $\gamma b'b$,

$$\hat{b} = \arg\min_b \left\{ (\bar{\mu} - \Sigma b)' \Sigma^{-1} (\bar{\mu} - \Sigma b) + \gamma b'b \right\}, \tag{34}$$

which leads to the same solution as the problem in Eq. 31. This representation is useful when including an additional L^1 penalty on b, which we pursue in Section 2.4.

Eq. 31 and Eq. 32 resemble ridge regression (e.g., see Hastie et al., 2011), which imposes a slightly different constraint on the L^2-norm of coefficients, $b'b \leq B$ or $\lambda'\lambda \leq C$, or uses different objective compared to Eq. 34. Our representations and ridge regression are not equivalent, however. The differences stem from economic restrictions embedded in our prior (14). Indeed, if instead of considering the Bayesian interpretation of an estimator, we naïvely apply classic ridge regression, the estimators are quite different. When parametrizing the model in terms of λ, ridge regression is given by:

$$\hat{\lambda} = \arg\min_b \left\{ (\bar{\mu} - I\lambda)' (\bar{\mu} - I\lambda) + \gamma \lambda'\lambda \right\}, \tag{35}$$

which yields the solution $\hat{\lambda} = \frac{1}{1+\gamma}\mu_T$, equivalent to the implausible prior $\eta = 1$.

The appealing property of the L^2 penalty mentioned above (that it shrinks the contributions of small PCs to the total squared Sharpe ratio) is specific to our setup and the representation we are working with. Because we started with the asset pricing equation Eq. 5, our explanatory variables are covariances with candidate factors, which are test asset returns themselves (so that the number of test assets equals the number of candidate factors). The PC-based interpretation we discussed obtains because of this latter fact. On the contrary, in the Fama-Macbeth regression setting (e.g., Freyberger et al. 2017) where one predicts expected returns with stock-level characteristics, an L^2 penalty does not have such clear interpretation.

2.3 Model Selection

The estimator Eq. 20 depends on the prior hyper-parameter, κ, which controls the degree of regularization and is, thus far, undetermined.[8] To fix κ one could simple rely on prior data, theory, or intuition, as in Barillas and Shanken (2017). They set $\kappa \approx 0.5$ annualized.[9] Alternatively, one could have a "hyper-prior" distribution over κ, which would then introduce additional hyper-parameters. We instead choose to estimate κ empirically, via 5-fold cross

[8] Recall that $\gamma = \frac{\tau}{\kappa T}$.

[9] Barillas and Shanken effectively use the prior $\eta = 1$ and set the expected max squared Sharpe ratio of zero-β assets equal to $1.25 \times \text{SR}^2_{\text{mkt}}$.

validation.[10] We divide the historic data into five equal sub-samples. Then, for each possible κ, we compute $\hat{b}$ by applying Eq. 20 to "in-sample" constructed by removing one of the sub-samples from the data, and evaluate the "out-of-sample" (OOS) fit of the resulting model on that withheld data. Consistent with the penalized objective, Eq. 31, we compute the OOS R^2 as

$$R^2 = 1 - \frac{\left(\mu_2 - \Sigma_2\hat{b}\right)'\left(\mu_1 - \Sigma_2\hat{b}\right)}{\mu_2'\mu_2}, \tag{36}$$

where the subscript 2 indicates an OOS sample moment and $\hat{b}$ is the estimator using the in-sample data. We repeat this procedure five times, each time treating a different sub-sample as the OOS data. We then average the R^2 across these five estimates, yielding the cross-validated R^2_{oos}. Finally, we choose κ which generates the highest R^2_{oos}. Our definition of R^2 implicitly treats the "risk-neutral" model, $\mu = b = 0$, as the benchmark. This is reasonable since our assets all have zero market β; under the "null model" (CAPM), all such assets should have zero mean. A negative value of R^2_{oos} indicates that a given model performs worse than the risk-neutral model out-of-sample.

2.4 Sparsity

Popular asset pricing models often predict sparsity of the weights on zero-investment portfolios in the SDF. For instance, the CAPM predicts a single factor representation; the 5-factor model of Fama and French (2016) and investment-based asset pricing models represent an SDF in terms of size, book-to-market and/or investment, and profitability characteristics. Yet the techniques we introduced so far do not automatically deliver such sparse SDF representations. In this section we explore economic reasons behind sparsity and enhance our toolkit to look for sparse SDF representations in the data.

2.4.1 Sparsity in characteristics-based factors

As a concrete example, consider a simple two-period "q-theory" model of firm investment similar to the one 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 and

[10] Our findings are robust with respect to the number of folds used.

the investment fully depreciates after one period. Every period, the firm has the objective

$$\max_{I_0} \mathbb{E}[M\Pi I_0] - I_0 - \frac{c}{2} I_0^2. \tag{37}$$

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,

$$\mathbb{E}\left[R\right] = \mathbb{E}\left(\frac{\Pi}{\mathbb{E}\left[M\Pi\right]}\right) = \frac{\mathbb{E}\left[\Pi\right]}{1 + cI_0}. \tag{38}$$

The model therefore implies a sparse characteristic-based factor model with two factors: expected profitability $\mathbb{E}\left[\Pi\right]$ and investment I_0.

Now consider a test of this model in the data. Neither expected profitability nor (planned) investment are observable. The usual approach is to use proxies, such as realized profitability and realized investment as potential predictors of unobserved quantities. Yet many additional characteristics are likely relevant for predicting 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 q-theory does not necessarily provide much economic reason to expect sparse SDFs in the space of observable characteristics. Most characteristics we use are inevitably noisy measures of true theoretic (unobservable) quantities. Models are gross simplifications of reality. Both of these facts lead us to believe that searching for sparse models in the space of observable characteristics is not well-motivated by theory and ultimately impractical. Indeed, we find little empirical support for this kind of sparsity (Section 3).

2.4.2 Sparsity in principal components of characteristics-based factor returns

The concept of sparsity is somewhat ill-defined in the first place. Because sparsity is not invariant to rotations of the data, one might potentially find significantly sparser representations in transformed rather than original data. By far the most common approach is to look for sparsity in the space of characteristics as discussed above. While practically appealing, we showed that such definition of sparsity might be problematic.

We propose an alternative approach motivated by the analysis in Kozak et al. (2017). We consider sparse SDF representations in the space of principal components of characteristic-based factor returns. The rationale behind such an approach is the following. 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 strong factor structure dominated by a small number of largest principal components. Under these two observations, an SDF with a small number of the largest PCs as factors should explain most of the cross-sectional variation in expected returns. An SDF should therefore be relatively sparse in the space of PCs and only large PCs should generally enter an SDF as factors. We exploit this theoretical result in our empirical section by exploring sparsity in the space of PCs and comparing it to evidence of sparsity in the space of characteristics. Consistent with our intuition, in Section 3 we do indeed find much greater evidence of the sparsity of the former kind.

2.4.3 The dual-penalty specification

Recall that our Bayesian formulation is equivalent to solving a penalized regression problem given by Eq. 34. The L^2 penalty, $\gamma_2 b'b$, results in shrinkage of elements of $\hat{b}$, but none of the coefficients is set to identically zero. To accomplish sparsity, we additionally impose an L^1 penalty , $\gamma_1 \sum_{j=1}^{H} |b_j|$. The approach is motivated by lasso regression and *elastic net* (Zou and Hastie, 2005). Due to 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[11]:

$$\hat{b} = \arg\min_{b} \left(\bar{\mu} - \Sigma b\right)' \Sigma^{-1} \left(\bar{\mu} - \Sigma b\right) + \gamma_2 b'b + \gamma_1 \sum_{i=1}^{H} |b_i| \,. \tag{39}$$

The L^2 penalty (and ridge regression) is known 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. In our setting this penalty will tend to use information in multiple related characteristics. For instance, rather than picking book-to-market as the only characteristic to represent the value effect in an SDF, it will instead average out multiple measures of value, such as book-to-market, price-dividend, and cashflow-to-price ratios.

The L^1 penalty, on the other hand, ignores correlations, and will tend to pick one characteristic and disregard the rest. In the extreme case of k identical predictors above, the

[11] To solve the optimization problem in Eq. 39 we use the LARS-EN algorithm in Zou and Hastie (2005), with few small modifications that impose economic restrictions specific to our setup.

lasso problem breaks down.

Our dual-penalty method enjoys much of economic motivation behind the L^2-penalty-only method with an added benefit of delivering sparse SDF representations. We can control the degree and cost of imposing sparsity by varying the strength of the L^1 penalty. By switching off the L^2 penalty we can naturally compare the performance of our method to pure Lasso (with GLS objective). We find that Lasso often performs poorly, indicating the vital importance of the L^2 penalty in emphasizing large PCs, as motivated by theory.

Despite of the visual similarities, there are important differences between our method and the elastic net. First, our method is economically motivated. Second, 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. Third, our objective function differs: we maximize the squared Sharpe ratio (minimize the distance to the mean-variance frontier) instead of minimizing (unweighted) pricing errors (our objective includes the weighting matrix Σ^{-1}).

2.5 Covariance Estimation Uncertainty

In the prior analyses, we have treated covariances (Σ and D) as known. Many papers, such as Ledoit and Wolf (2003, 2004), 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 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. Detailed derivations are provided in Section B.2.

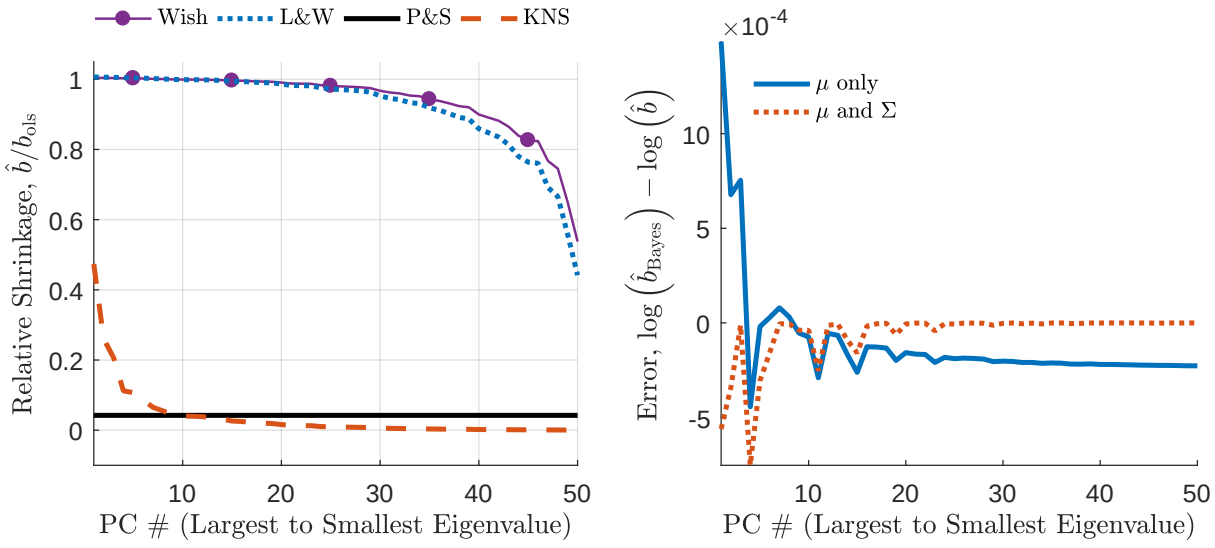
Ledoit and Wolf (2003) develop a covariance estimator which (asymptotically) optimally shrinks the sample covariance towards a target, typically a scaled identity matrix. They "concentrate on the covariance matrix alone without worrying about expected returns." Hence, their estimation of means is simply $\hat{\mu} = \bar{\mu}$.[12]

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

$$\Sigma^{-1} \sim \mathcal{W}\left(H, \frac{1}{H}\Sigma_0^{-1}\right), \tag{40}$$

where $\Sigma_0 = \frac{1}{H}\text{tr}(\Sigma_T) I$, 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

[12] In a series of papers (2002, 2003, 2004) L&W propose various shrinkage estimators. The estimator in L&W 2004 is most appropriate in our empirical setting of zero-β anomaly portfolios.



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

Figure 2: 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.

relatively "diffuse". 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.

Figure 2a 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. 19. 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. That plot shows that this priors shrinks all coefficients uniformly towards zero.[13] The L&W and Wishart methods deliver very similar estimators. Importantly, they are characteristically different from our method (KNS). 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

[13]The degree of shrinkage is determined by cross-validation, as described in Section 2.3

to $1,700$.

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

$$\mu | \, \Sigma \sim \mathcal{N}\left(0, \, \frac{\kappa^2}{\tau}\Sigma^2\right), \tag{41}$$

where $\tau = \mathrm{tr}\,(\Sigma_T)$. For Σ, we use a similar prior to Eq. 40, with a slight modification for numerical tractability since the posterior is not fully analytic.[14] We also consider an approximation given by the following procedure: first regularize the covariance matrix according to the Wishart prior, Eq. 40. Then, treating the covariance matrix as known, apply Eq. 19. This method is fully analytic and closely approximates the fully Bayesian solution. Figure 2b shows the ratio of the full Bayes estimate to the approximate Bayes estimate, 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 3.4.

[14] Details are given in Section B.2

3 Empirics

3.1 Data

We start with the universe of U.S. firms in CRSP. We use Compustat to compute a set of characteristics associated with each firm. For robustness, we exclude small-cap stocks[15], center and standardize all characteristics as explained in Section 3.2. We then interact characteristics and stock excess returns in the cross-section at each time t. This leads to a rotation of the space of N individual stocks returns into a space of H "managed portfolios" based on such H characteristics.

We construct three independent sets of characteristics. The first set uses (implicitly defined) binary characteristics that lead to a rotation which exactly matches Fama-French 25 portfolios sorted on ME and BE/ME. It is straightforward to define such characteristics: for any given portfolio, a corresponding characteristic equals 1 for a stock that enters that portfolio and 0 for all other stocks (e.g., Feng et al. 2017, Freyberger et al. 2017).

The second 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. We interact characteristics with underlying stock excess returns linearly and use all stocks in doing so. This approach differs slightly from the common practice of sorting stocks into deciles and forming a long-short portfolio of deciles 10-1. While the outcomes of both approaches are qualitatively similar, our method does not throw away information (portfolios 2-9) and produces better diversified portfolios.

The anomalies and excess returns on managed portfolios (which are linear in characteristics) are listed in Table 6. The table shows mean excess returns in three subperiods: full sample, pre-2005, and the most recent post-2005 sample. All managed portfolios' excess returns are rescaled to have standard deviations equal to the in-sample standard deviation of excess return on the aggregate market index.

The third set of characteristic is based on 70 financial ratios as defined in WRDS: "WRDS Industry Financial Ratios" (WFR hereafter) is a collection of most commonly used financial ratios by academic researchers. There are in total over 70 financial ratios grouped into the following seven categories: Capitalization, Efficiency, Financial Soundness/Solvency, Liquidity, Profitability, Valuation and Others." We list mean returns on WFR managed portfolios in Table 7. We supplement this dataset with 12 portfolios sorted on past monthly returns

[15]We drop all stocks with market caps below 0.01% of aggregate stock market capitalization at each point in time. For example, for an aggregate stock market capitalization of \$20trln, we keep only stocks with market caps above \$2bln.

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).

Finally, for our analysis in Section 3.4, we supplements the sets of 50 anomaly and WFR raw characteristics with characteristics based on second and third powers and linear first-order interactions of characteristics, which we construct using the approach of Section 3.4. 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.

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.

3.2 Normalizations and Rescaling of Characteristics

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 managed portfolios in Eq. 4. 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.[16] We then normalize all ranks by dividing by n_t+1 to obtain the value of the rank transform:

$$rc^i_{s,t} = \frac{\text{rank}\left(c^i_{s,t}\right)}{n_t+1}. \tag{42}$$

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:

$$z^i_{s,t} = \frac{\left(rc^i_{s,t} - \bar{rc}^i_t\right)}{\sum_{s=1}^{n_t}\left|rc^i_{s,t} - \bar{rc}^i_t\right|}, \tag{43}$$

[16] 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.

where $\bar{rc}_t^i = \frac{1}{n_t}\sum_{s=1}^{n_t} rc_{s,t}^i$. The resulting portfolios of transformed characteristics $z_{s,t}^i$ are insensitive to outliers and allow us to keep the total exposure of a portfolio to a characteristic-based strategy (leverage) fixed. For instance, doubling the number of stocks at any time t has no effect on the overall exposure of a strategy. Finally, we combine all transformed characteristics $z_{s,t}^i$ for all stocks into a matrix of instruments Z_t, which we use in our analysis in Eq. 4.

3.3 Results

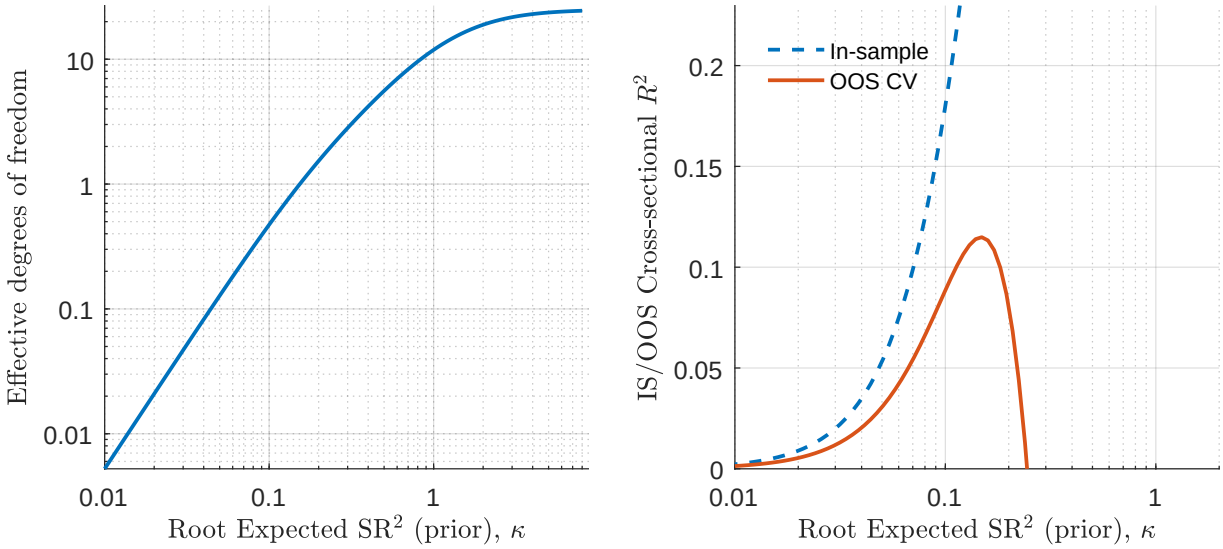
We start with the basic and well-known set of test assets (Fama-French 25 ME and BE/ME sorted portfolios) and explain how our method generalizes and improves upon the technique by Fama and French (1993). We then proceed to our main datasets of 50 anomalies and 80 WFR (including 12 return lags) to illustrate the power of the method in highly multidimensional environments where classic techniques are infeasible.

3.3.1 Fama-French 25 ME/BM -sorted portfolios

We consider 25 Fama-French ME/BM sorted portfolios which we orthogonalize with respect to the aggregate market in the full sample. These portfolios exhibit very strong factor structure: most of the variance can be explained by only two factors (HML and SMB). Fama and French (1993) construct these factors manually. Kozak et al. (2017) show that HML and SMB factors essentially match the first and the second PCs of the 25 (market-neutral) portfolio returns. Extracting and using only two such factors to explain the cross-section of expected returns, therefore, is a form of regularization, known as *principal component regression* (PCR), but done implicitly. In our method, the economic reasoning and Bayesian priors described in Section 2 lead to a modified *ridge regression*, which can be thought of as a continuous version of PCR. Whereas PCR ignores small PCs completely, ridge regression strongly down-weights them instead.

In the case of only L^2 regularization, how should one choose the amount of shrinkage? We first note that there are several equivalent ways to quantify this. The most natural and our preferred method is to focus on κ in our Bayesian prior for mean returns. We showed that κ has a natural economic interpretation: it is the square root of the expected maximum squared Sharpe ratio under the prior. Alternatively, one can focus on the strength of penalty $\gamma = \frac{\tau}{\kappa^2 T}$ in Eq. 34. In the ridge regression setting it is common to map γ into the *effective degrees of freedom*[17] which also quantifies the strength of L^2 regularization:

[17] A similar definition is often used in a ridge regression setup. See Hastie et al. (2011).



(a) Effective degrees of freedom **(b)** L^2 model selection

Figure 3: L^2 Model Selection (Fama-French 25 ME/BM portfolios). In Panel (a) we show the map between effective degrees of freedom and the annualized Root Expected SR^2 (κ) under the prior. Panel (b) plots the in-sample cross-sectional R^2 (dashed) and OOS cross-sectional R^2 based on cross validation (solid).

$$\mathrm{df}(\gamma) = \mathrm{tr}\left[\Sigma\left(\Sigma+\gamma I\right)^{-1}\right] = \sum_{j=1}^{H}\frac{d_j}{d_j+\gamma}, \tag{44}$$

where d_j is the j-th eigenvalue of Σ and γ is the parameter which governs the strength of L^2 penalty as defined in Section 2.2. Note that $\gamma = 0$, which corresponds to no shrinkage, gives $\mathrm{df}(\gamma) = H$, the number of free parameters. There is one-to-one mapping between κ and $\mathrm{df}(\gamma)$, which we illustrate in Figure 3a. The Fama-French 25 market-neutral portfolios are sorted on two dimensions (ME and BE/ME) and therefore possess roughly two degrees of freedom, that is, most of the variance of their returns is explained by two dominant PCs (which roughly match HML and SMB factors). We see from the plot that $\mathrm{df}(\gamma) = 2$ gives the $\sqrt{\mathbb{E}\left(\mathrm{SR}^2\right)} \approx 0.2$ under the prior. Without such prior knowledge, one would have to resort to data-driven methods, such as cross validation (see Section 2.3), to determine κ.

We now turn to model selection. Figure 3b focuses on picking the optimal κ using cross-validation. It plots the cross-sectional R^2 on the y-axis as a function of $\sqrt{\mathbb{E}\left(\mathrm{SR}^2\right)}$ on the x axis. We show two paths: the in-sample cross-sectional R^2 (dashed blue) and cross-validated out-of-sample cross-sectional R^2 (solid red). The out-of-sample R^2 are computed using the 5-fold cross-validation method previously explained in Section 2.3. The cross-validation criteria suggest a level of regularization $\kappa \approx 0.15$ — which roughly coincides with the value implied

Table 1: Coefficient estimates and t-statistics (Fama-French 25 ME/BM 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 raw Fama-French 25 BE/BM sorted portfolios. Panel (b) pre-rotates FF25 returns into PC space and shows coefficient estimates corresponding to these PCs. 10 portfolios with largest t-statistics are shown. Standard errors are calculated using Eq. 21 and do not account for uncertainty in κ.

(a) Raw Fama-French 25 portfolios

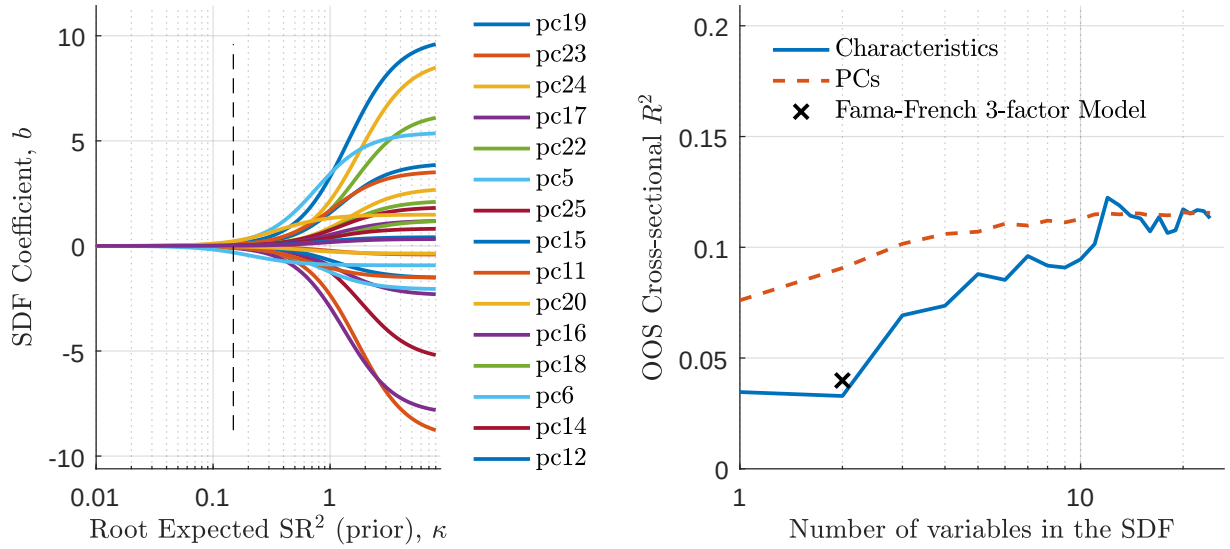
	b	t-stat
SMALLHiBM	0.18	0.98
ME3BM5	0.16	0.87
ME1BM4	0.16	0.84
ME3BM4	0.14	0.74
ME4BM4	0.14	0.73
ME4BM1	0.00	0.02
ME5BM3	-0.00	0.00
ME3BM1	-0.06	0.33
ME2BM1	-0.08	0.45
SMALLLoBM	-0.16	0.84

(b) PCs of Fama-French 25 portfolios

	b	t-stat
PC 1	-0.31	2.03
PC 2	0.22	1.28
PC 5	0.20	1.08
PC 19	0.11	0.57
PC 17	-0.10	0.53
PC 4	-0.07	0.39
PC 11	0.07	0.37
PC 23	-0.07	0.37
PC 6	-0.07	0.36
PC 24	0.06	0.33

by two degrees of freedom. Moreover, Figure 3b vividly illustrates that we require a very significant amount of shrinkage to obtain a well-performing SDF. Since we do not generally have a strong prior regarding the appropriate degrees of freedom, we use cross-validation throughout this paper to determine the strength of regularization.

Table 1 lists coefficient estimates at this optimal level of regularization, $\kappa \approx 0.15$. Table 1a seeks to represent an SDF in terms of original Fama-French (market-neutral) 25 portfolio returns as factors, while Table 1b pre-rotates assets into PC space and looks for an SDF representation in PC space, as in Figure 3b. Since L^2 regularization is rotation invariant, we obtain the same solution whether we first estimate the model on the original assets and then rotate into PC space or directly estimate in PC space. We show the 5 largest (positive) and 5 smallest (negative) coefficient estimates in Table 1a and 10 largest absolute coefficient estimates in Table 1b. We can see from Table 1b that, at this optimal level of regularization, the largest and most significant coefficients our method picks up are associated with PC1 and PC2, consistent with our economic intuition. This is in stark contrast to the unregularized OLS solution (shown below in Figure 4a), which assigns the highest SDF loadings to some of the smallest PCs. In terms of the original portfolios, Table 1a shows that the optimal SDF assigns positive weights to small and value portfolios and shorts growth and large portfolios, exactly as done by the approach of Fama and French.



(a) L^2 Coefficient Paths **(b)** Sparsity

Figure 4: L^2 Coefficient Paths and Sparsity (Fama-French 25 ME/BM portfolios). Panel (a) plots paths of coefficients as a function of the prior Root Expected SR^2 (κ) for 25 Fama-French ME/BM sorted portfolios. Labels are ordered according to absolute values of coefficients (descending) at the right edge of the plot, which corresponds to the OLS solution. 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). The "x" mark indicates OOS performance of the Fama-French model that uses only SMB and HML factors.

In Figure 4a plot paths of coefficients associated with each PC in the final SDF as a function of strength of regularization, as measured by κ ($\sqrt{\mathbb{E}\left(\text{SR}^2\right)}$ under the prior). The right edge of the plot corresponds to the unregularized OLS solution. Labels are ordered according to absolute values of OLS coefficients (descending). As we increase the strength of the penalty (move to the left), coefficients shrink toward zero and the model-SDF approaches the risk-neutral (constant) SDF (left edge of the plot). We can see that shrinkage is stronger for small PCs and weaker for large PCs, which reflects our prior belief that small PCs are unlikely to be important in the SDF . Recall, the prior $\eta = 1$ (Pástor and Stambaugh, 2000) uniformly shrinks all coefficients toward zero since under that prior all PCs have equal importance.

Figure 4a clearly demonstrates the need for regularization: because the 25 Fama and French ME/BM sorted portfolios are highly correlated, estimating the MVE portfolio (SDF coefficients) with no regularization leads to very high SDF coefficients (± 10) and extreme SDF variance (not shown). Most importantly, such an SDF loads heavily on small PCs (PCs 19, 23, 24, 17, 22 have the highest coefficients on the plot), quite different from the

optimally regularized coefficients given in Table 1. As we show later, the OLS solution performs extremely poorly out of sample.

For the family of models with two penalties we need to cross-validate two parameters: κ and the number of variables in the SDF (or, equivalently, the L^1 penalty parameter). Before doing that, we first consider fixing the number of non-zero SDF coefficients (L^1 penalty), then determining κ by cross-validation. Figure 4b plots the maximum (cross-validated) cross-sectional R^2 achievable by a sparse model with n factors (on the x-axis). Following our discussion in Section 2.4, we explore sparsity both in the space of characteristics and PCs. The plot shows two important results. First, we are able to obtain much sparser SDF representations in PC space than in original space of characteristics. Note that the first two PCs optimally chosen are the first and second PCs (not shown). An SDF representation with these PC factors performs as well as a ten factor model using the primitive characteristic factors. Second, including more than two PCs barely improves the model performance, suggesting that sparse models formed from a few large PCs can approximate the SDF quite well.

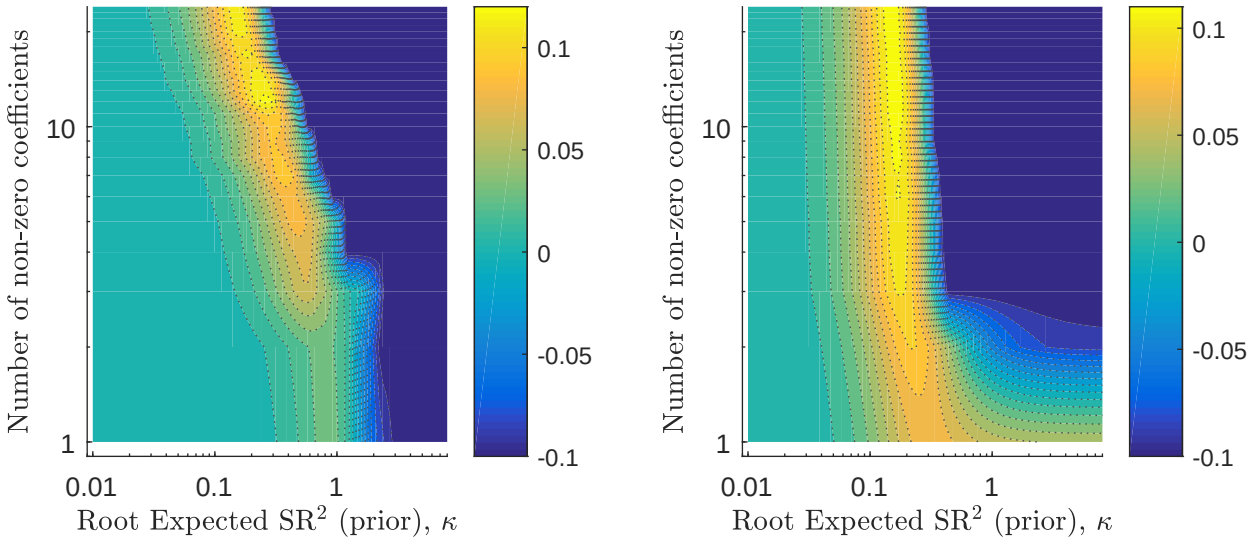
The black "X" depicts the performance of the Fama-French model based on HML and SMB.[18] Because PC1 and PC2 essentially recover HML and SMB, respectively, our SDF is effectively the same as Fama-French's. It, however, performs slightly better for two reasons: (i) PC1 and PC2 as factors are slightly more "efficient" because they combine all portfolios in their construction; and (ii) our method performs substantial relative shrinkage (twist), shrinking PC2 (SMB) more than PC1 (HML).

We now perform full cross-validation on both penalty parameters. Figure 5 shows a contour map depicting cross-validated 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). Warmer (yellow) colors reflect higher OOS R^2. Like with the L^2 penalty alone, cross validation favors very aggressive regularization, emphasizing the importance of shrinkage in the case when assets have a strong factor structure. Note that unregularized models that include all 25 factors (top-right corner) demonstrate extremely poor performance with OOS R^2 significantly below 0.[19] The bottom-left point reflects the SDF with maximum combined penalties which result in essentially flat (risk-neutral) SDF and deliver the R^2 of 0. Optimal regularization (yellow regions) significantly outperform such an SDF.

Figure 5 shows that L^2-penalty-only based models (top edge of a plot) perform very

[18]To put both approaches on equal footing, we shrink FF coefficients towards zero based by the amount of "level" shrinkage implied by our method. This modification significantly improves OOS performance of FF method.

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



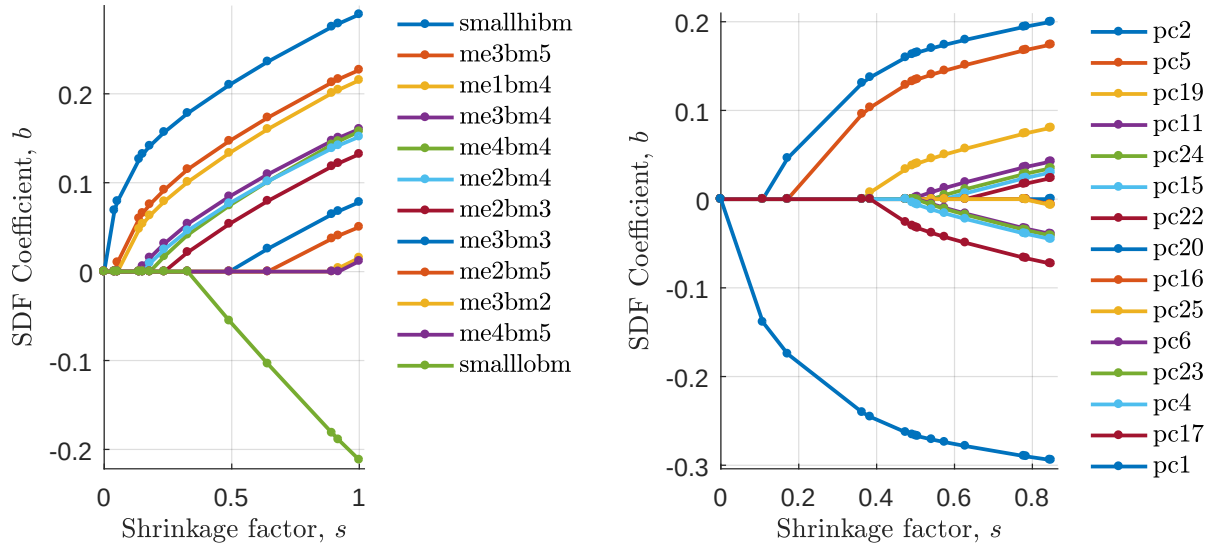
(a) Raw Fama-French 25 portfolios

(b) PCs of Fama-French 25 portfolios

Figure 5: Sparse Model Selection (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 to 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.

well and attain nearly maximal R^2, while L^1-only "Lasso" based models (right edge of the plot) completely fail in the space of characteristics (Figure 5a) and are still substantially sub-optimal in PC space (Figure 5b). When sparsity is a goal, our message is that Lasso alone is inadequate; optimal sparse models should combine both penalties. Figure 5b shows that such models perform well and significantly greater sparsity can be achieved in PC space compared to the space of characteristics. Additionally, the vertical shape of the optimal (yellow) region in Figure 5b suggests that there is essentially no cost of dropping most of the (small) PCs from the SDF – their respective coefficients were already set to almost (but not exactly) zero by the L^2 penalty. As a result, the L^1 penalty simply forces those coefficients to zero but has little impact on coefficients associated with largest PCs. This is clearly not the case in the space of characteristics (Figure 5a), where both penalties are somewhat substitutable in that both are shrinking most of the coefficients towards zero. As a result the optimal region extends diagonally.

Lastly, we explore the identity of factors in the optimal SDF approximations. Figure 6 plots paths of SDF coefficients as a function of L^1 penalty when fixing the L^2 penalty at



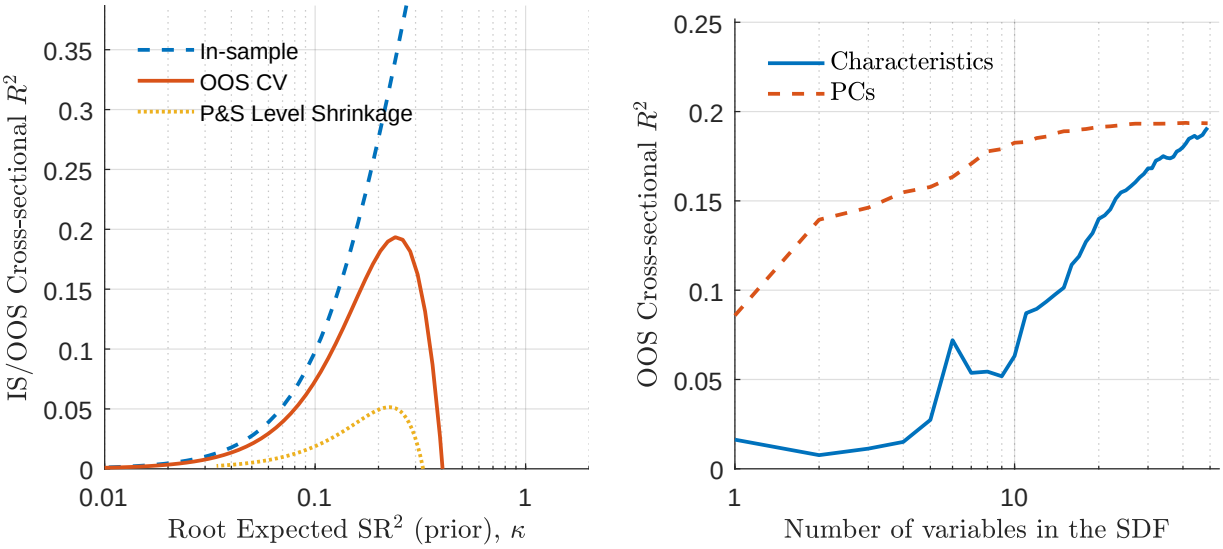
(a) Raw Fama-French 25 portfolios **(b)** PCs of Fama-French 25 portfolios

Figure 6: L^1 coefficient paths for the optimal model (Fama-French 25 ME/BM portfolios). Paths of coefficients as a function of shrinkage factor s based on the optimal (dual-penalty) sparse model that uses 25 for Fama-French ME/BM sorted portfolios (Panel a) and 25 PCs based on Fama and French 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.

the cross-validated optimum. The right-edge of each panel gives the coefficient estimates of the dual-penalty optimal model. Moving to the left, we show the estimates when the L^2 parameter is kept at the optimum, but the L^1 penalty is increased. For ease of presentation the x-axis is a monotone transform of the L^1 penalty, γ_1,

$$s = \frac{\sum_{i=1}^{H} |b_i|}{\sum_{i=1}^{H} \left|\tilde{b}_i\right|},$$

where $\tilde{b}_i$ are coefficients corresponding to a solution with optimal levels of both penalties. Figure 6a shows that the first 5 variables picked by the dual-penalty method coincide with the top 5 variables in Table 1a. The resulting SDF is long small and value stocks, similarly to our L^2-penalty-only SDF and in the spirit of Fama and French (1993). Similarly, the first 2 PCs selected in PC space in Figure 6b are PC1 and PC2 – exactly the same variables as those with highest t-statistics in Table 1b. The resulting SDF contains two dominant principal components and is nearly identical to the SDF constructed by Fama and French (1993). The first few variables selected by L^2-only and dual-penalty methods are nearly always identical for all sets of test assets we consider, further suggesting that optimal models that employ



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

Figure 7: 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 (dotted) in 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).

only an L^2 penalty typically perform on par with dual-penalty sparse models.

This example illustrates that regularization is particularly important when factor structure is strong or some candidate factors are highly correlated. Overall, our method tends to recover an SDF that is closely related to the SDF implied by Fama and French (1993). Our method therefore simply generalizes their approach and slightly improves its out-of-sample performance. The true strength of our method however comes in dealing with multidimensional settings characterized by vast abundance of characteristics and unknown factors, where classic techniques are inadequate. Importantly, we find that analytically tractable estimation with only an L^2 penalty achieves nearly optimal results. Again, this becomes important with dealing with extremely large cross-sections.

3.3.2 50 anomaly characteristics

We now turn to our primary dataset – 50 portfolios based on anomaly characteristics listed in Table 6. Since characteristics are centered and are linear in the cross-sectional centered rank of a characteristic, managed portfolios have an intuitive long-short strategy interpretation. Further, because the managed portfolios underlying the anomalies in Table 6 are not as highly correlated as the 25 Fama and French ME/BM sorted portfolios, we expect the overall factor

Table 2: Coefficient estimates and t-statistics (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 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) 50 anomaly portfolios

	b	t-stat
Industry Rel. Rev. (L.V.)	-0.58	2.92
Ind. Mom-Reversals	0.32	1.61
Industry Rel. Reversals	-0.28	1.43
Earnings Surprises	0.22	1.11
Seasonality	0.22	1.10
Return on Market Equity	0.21	1.05
Value-Profitablity	0.20	1.02
Composite Issuance	-0.17	0.85
Investment/Assets	-0.16	0.79
Return on Equity	0.16	0.78
Momentum (12m)	0.15	0.75

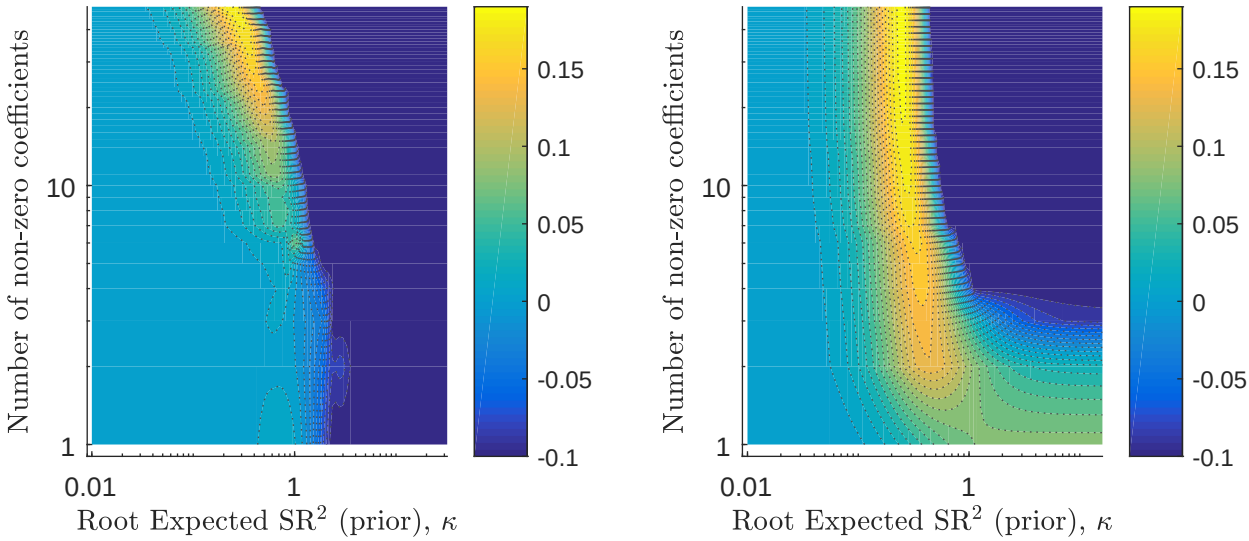
(b) PCs of 50 anomaly portfolios

	b	t-stat
PC 4	0.58	2.94
PC 1	-0.43	2.89
PC 2	-0.36	2.02
PC 5	-0.37	1.86
PC 11	-0.35	1.70
PC 15	-0.28	1.34
PC 6	-0.26	1.32
PC 14	0.21	1.01
PC 19	0.17	0.83
PC 23	0.14	0.69
PC 10	0.13	0.64

structure to be weaker. Some of the individual anomalies, however, are highly correlated. For instance, one can correctly conjecture that many price ratios, such as D/P, E/P, B/M, CF/P, have non-trivial correlations. We therefore expect that some regularization is needed to handle those correlated anomalies.

Figure 7a plots the in-sample cross-sectional R^2 (dashed) and OOS cross-sectional R^2 based on cross validation (solid). Similarly to the case with 25 ME/BM portfolios, cross validation favors aggressive regularization with $\kappa \equiv \sqrt{\mathbb{E}\left(\mathrm{SR}^2\right)} \approx 0.22$. Table 2 lists coefficient estimates at this optimal level of regularization. The original anomalies are shown in Table 2a and PCs are listed in Table 2b. The largest coefficients and t-statistics are associated with industry relative reversals (low vol.), industry momentum-reversals, industry relative-reversals, earnings surprises, seasonality, 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. Focusing on the SDF represented in the space of PCs, we see that at the optimal level of regularization the largest and significant coefficients our method picks up are associated with PC1, PC2, and PC4, consistent with our economic intuition.

To compare the effect of our shrinkage method to that of Pástor and Stambaugh (2000), we additionally plot the OOS performance of their method as a dotted yellow line in Fig-



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

Figure 8: Sparse Model Selection (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 to 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.

ure 7a.[20] Recall that our method performs both "level" (proportional) 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 5% OOS R^2 (an improvement over the OLS solution), but falls substantially short of the 20% 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 7b explores the performance of sparse models which employ both L^1 and L^2 penalties. There is little evidence of sparsity in the space of characteristics (solid line): more than twenty factors are needed in the SDF to capture a significant fraction of total R^2. In PC space, on the contrary, very sparse models perform exceedingly well: a model with a single PC-based factor captures roughly half of the total OOS cross-sectional R^2, while adding a second factor raises the R^2 to about 70% of the maximal one. A model with ten factors captures achieves maximal R^2, while a model with ten factors in the space of characteristics

[20]For the P&S level shrinkage estimator we show $\mathbb{E}\left(SR^2\right)$ under the prior on the x-axis, but it no longer coincides with the κ parameter in that model.

achieves less than a third of maximum. Therefore, once again we witness strong evidence in favor of sparsity in PC space, but not in the space of original characteristics.

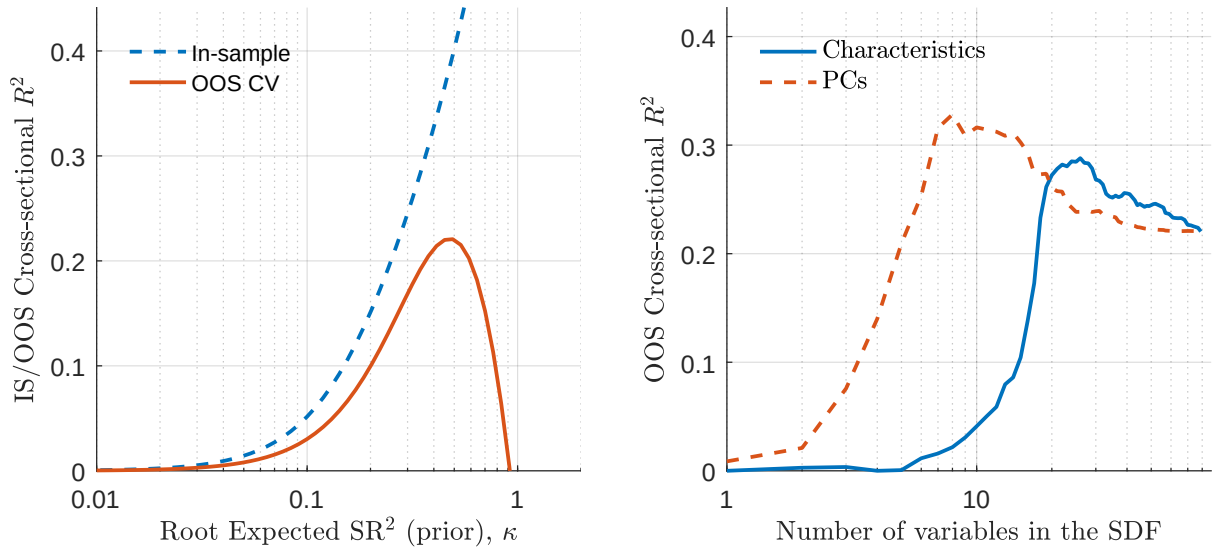
In Figure 8 we show a contour map depicting cross-validated OOS cross-sectional R^2 as a function of κ (on the x-axis) and the number of non-zero SDF coefficients (on the y-axis). Warmer (yellow) colors reflect higher OOS R^2. Cross validation again favors very aggressive regularization. Unregularized models (top-right corner) demonstrate extremely poor performance with OOS R^2 significantly below 0. L^2-penalty-only based models (top edge of a plot) perform very well and attain nearly maximal R^2, while L^1-only "Lasso" based models (right edge of the plot) completely fail in the space of characteristics (Figure 8a). In PC space (Figure 8b), a Lasso model based on a single PC (PC1; see Figure 13 in the Appendix) delivers roughly half of the total R^2, but performance deteriorates quickly if more factors are added.

The factors that our dual-penalty approach picks in sparse SDF representations significantly overlap with factors with highest t-statistics in Table 2. For instance, in PC space, the first selected factor is PC1, followed by PC4, PC2, and PC5. We omit this plot from the main text for brevity (see Figure 13 in the Appendix).

Our main findings in this section are threefold: (i) the L^2-penalty-only based method finds robust SDF representations that perform exceptionally well out of sample; (ii) the dual-penalty method delivers robust and sparse SDF representations in PC space; and (iii) sparsity in the space of characteristics is limited, suggesting little redundancy among the anomalies we consider.

3.3.3 Robustness: WRDS financial ratios (WFR)

One potential problem with the anomaly dataset in the previous section is that many of existing anomalies have been heavily data-mined in sample and might not generalize well outside the period in which they were discovered (Harvey et al., 2015, McLean and Pontiff, 2016). To address this concern we construct a new dataset which is much less prone to this issue. We focus on various financial ratios as defined by WRDS, which is simply is "a collection of the most commonly used financial ratios by academic researchers" with no direct link to the cross-section of expected returns. There are in total 68 financial ratios grouped into the following seven categories: Capitalization, Efficiency, Financial Soundness/Solvency, Liquidity, Profitability, Valuation and Others (Table 7 in the Appendix lists the ratios). We supplement this dataset with twelve portfolios sorted on past monthly returns in months $t-1$ through $t-12$ to allow for any potential predictability stemming from return autocorrelations. The final dataset therefore contains eighty managed portfolios (WFR hereafter), none of which was formed specifically to generate dispersion in expected returns. We now



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

Figure 9: 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).

use our method to construct an SDF based on these new data.

Figure 9a plots the in-sample cross-sectional R^2 (dashed) and OOS cross-sectional R^2 based on cross validation (solid). Cross validation favors less aggressive regularization than with 50 anomaly portfolios, with root expected SR^2 of $\kappa \approx 0.45$, consistent with our intuition that WFR are much less likely to have been datamined in an early part of the sample compared to anomalies and therefore do not require as much shrinkage as the anomaly-based dataset. Table 3 lists coefficient estimates at this optimal level of regularization. Table 3a focuses on original WFR portfolio returns, while Table 3b pre-rotates assets into PC space. Coefficients are sorted descending on their absolute t-statistic values. Table 3b shows that most of the variables that enter the SDF are indeed relatively large PC, as predicted by our theory.

Table 3a shows that our method tends to pick up 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), accruals (Accruals/Average Assets), financial soundness (Free Cash Flow/Operating Cash Flow, Operating CF/Current Liabilities, Cash Flow/Total Debt), momentum (months $t-9$, $t-11$) and short-term reversals (month $t-1$). 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

Table 3: Coefficient estimates and t-statistics (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) WFR portfolios

	b	t-stat
Free Cash Flow/Operating Cash Flow	0.96	2.95
P/E (Diluted, Incl. EI)	-0.75	2.28
Accruals/Average Assets	0.75	2.28
Operating CF/Current Liabilities	0.63	1.90
P/E (Diluted, Excl. EI)	-0.58	1.76
Enterprise Value Multiple	-0.56	1.70
Cash Flow/Total Debt	0.55	1.65
Trailing P/E to Growth (PEG) ratio	-0.52	1.60
Month $t-9$	0.51	1.59
Month $t-11$	0.47	1.44
Month $t-1$	-0.42	1.31

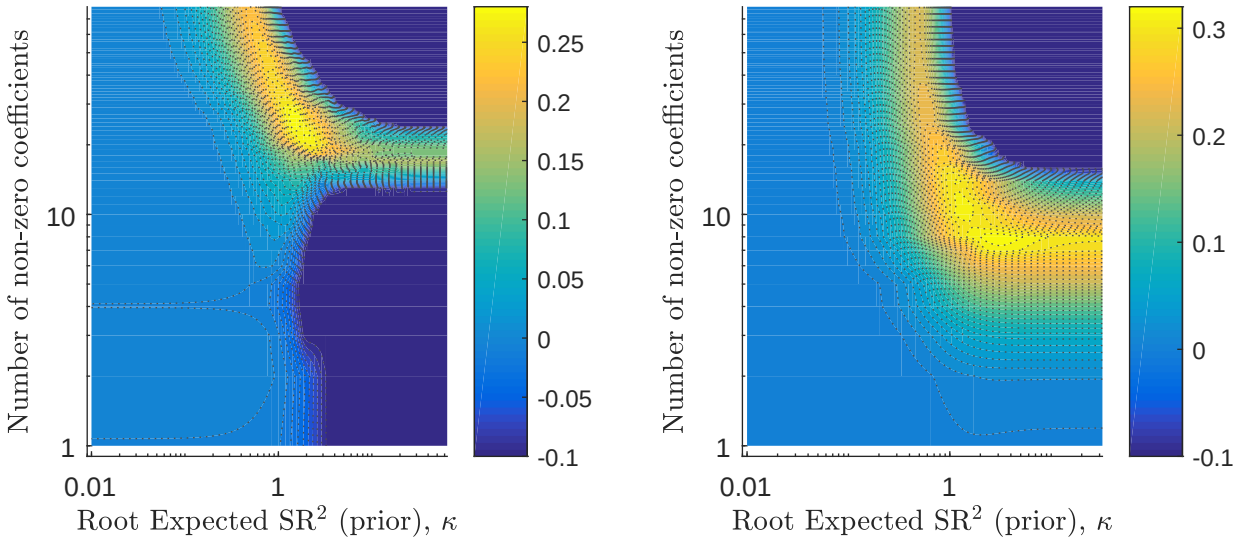
(b) PCs of WFR portfolios

	b	t-stat
PC 7	-1.11	3.81
PC 19	-1.08	3.49
PC 6	0.86	2.99
PC 26	0.77	2.46
PC 2	-0.34	1.83
PC 9	0.52	1.72
PC 5	-0.36	1.38
PC 1	0.21	1.37
PC 15	-0.34	1.11
PC 20	-0.33	1.07
PC 36	-0.34	1.05

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. This can shed lights on findings of Hou et al. (2015) relative to Fama and French (2016), for instance, but also emphasizes that using any single measure in constructing low-dimensional factor models is often sub-optimal when it comes to out-of-sample performance. Our approach offers a powerful alternative to standard methods in those papers.

Figure 9b explores the performance of sparse models which employ both penalties. There is little evidence of sparsity in the space of characteristics (solid line): an SDF with ten characteristic-based factors captures negligible fraction of total R^2. In the PC space, on the contrary, relatively sparse models perform well; a model with five factors captures a large fraction of the total OOS cross-sectional R^2, while a model with eight factors delivers maximum OOS R^2. Therefore, once again we witness strong evidence in favor of sparsity in PC space, but not in the space of original characteristics.

In Figure 10 we show a contour map depicting cross-validated 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). Warmer (yellow) colors reflect higher OOS R^2. Cross validation again favors very aggressive regularization. Unregularized models (top-right corner) demonstrate extremely



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

Figure 10: Sparse Model Selection (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 to 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.

poor performance with OOS R^2 significantly below 0. L^2-penalty-only based models (top edge of a plot) perform well and attain high OOS R^2. L^1-only "Lasso" based models (right edge of the plot) work somewhat better than before, especially in PC space, where a model with 7-8 factors captures most of the R^2.

The factors that our dual-penalty approach picks in sparse SDF representations significantly overlap with factors with highest t-statistics in Table 3. For instance, in PC space PC7, PC19, PC2, PC6, PC5, PC1 are the first ones to be added to the SDF. We omit this plot from the main text for brevity (see Figure 14 in the Appendix).

From the analysis in this section we conclude that our main messages remain intact: (i) the L^2-penalty-only based method finds robust SDF representations that perform well out of sample; (ii) the dual-penalty method delivers robust and sparse SDF representations in PC space; and (iii) sparsity in the space of characteristics is not guaranteed and is often unreliable. We also find that our approach is relatively immune to the issue of datamined anomalies, mostly due to two reasons: (i) our method relies on no-near-arbitrage assumption which requires that expected returns line up with large sources of co-movements in the data (big PCs) and therefore it crushes SDF coefficients associated with "fished" anomaly factors,

which instead line up with small PCs; and (ii) our sample extends to 2016 while most of the anomalies had been discovered prior to 2005; these "extra" ten years of data provide powerful incentive for the cross-validation method to pick high levels of regularization.

3.4 Interactions

Up to this point, the cross-section of test assets was based on the cross-product of returns with a small set of pre-specified characteristics (in each cross-section). This methodology imposes linearity and independence (no interactions) of characteristics. In reality, there is little reason to limit ourselves to such strict functional form assumptions. There is evidence of both non-linear effects (Fama and French, 2008, Freyberger et al., 2017) and interaction (Asness et al., 2013, Fama and French, 2008). We now extend our analysis to allow for more flexible specifications.

Specifically, for any two given rank-transformed characteristics $z^i_{s,t}$ and $z^j_{s,t}$ of a stock s at time t, we define the first-order interaction characteristic $z^{ij}_{s,t}$ as the product of two original characteristics that is further re-normalized using Eq. 43 as follows:

$$z^{ij}_{s,t} = \frac{\left(z^i_{s,t} z^j_{s,t} - \frac{1}{n_t}\sum_{s=1}^{n_t} z^i_{s,t} z^j_{s,t}\right)}{\sum_{s=1}^{n_t}\left|z^i_{s,t} z^j_{s,t} - \frac{1}{n_t}\sum_{s=1}^{n_t} z^i_{s,t} z^j_{s,t}\right|}. \tag{45}$$

We include all first-order interactions in our empirical tests in Section 3. In addition to interactions, we also include second and third powers of each characteristic, which are defined analogously. Note that although we re-normalize all characteristics post 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 (tails). We provide an economic interpretation of interactions portfolios in Appendix A.

3.4.1 Results: Interactions of anomaly and WFR portfolios

We start with the 50 characteristics from Section 3.3.2 and 80 WFR from Section 3.3.3, compute their second and third powers and all first-order interactions as explained in Section 3.4. We obtain the total of 1,375 and 3,400 derived characteristics and candidate factors for anomaly and WFR portfolios, respectively.

Table 4 lists coefficient estimates at the optimal level of L^2 regularization. Table 4a focuses on the SDF constructed from PCs of portfolio returns based on interactions of 50 anomaly characteristics. Table 4b shows coefficient estimates corresponding to PCs of portfolio returns based on interactions of WRDS financial ratios (WFR). Coefficients are sorted

Table 4: Coefficient estimates and t-statistics (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	-0.18	3.42
PC 2	0.19	2.75
PC 17	-0.16	1.78
PC 60	-0.13	1.41
PC 40	0.12	1.33
PC 30	-0.11	1.19
PC 10	-0.10	1.15
PC 6	-0.09	1.15
PC 20	-0.10	1.12
PC 67	-0.10	1.11
PC 12	-0.09	1.04

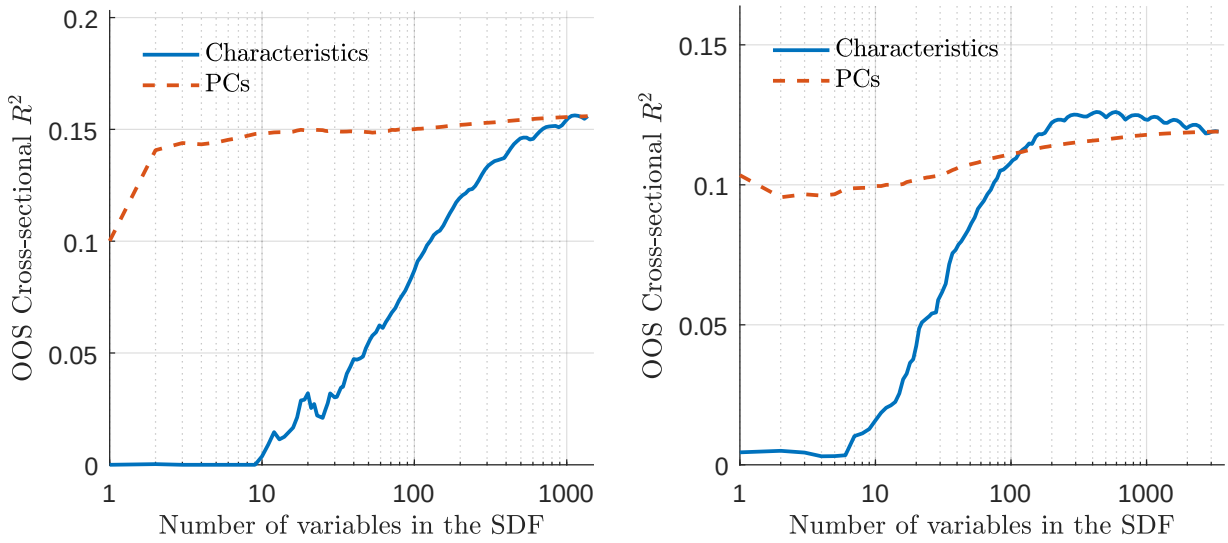
(b) PC of interactions of WFR portfolios

	b	t-stat
PC 1	-0.10	2.77
PC 5	0.09	1.43
PC 2	-0.07	1.43
PC 20	-0.08	1.15
PC 7	0.07	1.06
PC 50	-0.07	0.90
PC114	-0.06	0.80
PC 48	-0.06	0.78
PC 39	0.05	0.75
PC 28	-0.05	0.72
PC 52	0.05	0.71

descending on their absolute t-statistic values. We find that PC1 and PC2 are the only two significant variables and are the first ones included into an SDF that prices all portfolios based on interactions of 50 anomaly characteristics (Table 4a). Table 4b shows that only one principal component, PC1, is statistically significant for pricing managed portfolios based on interactions of WFR.

Figure 11 explores the performance of sparse models which employ both penalties simultaneously. There is little evidence of sparsity in the space of characteristics (solid lines): an SDF with 10 characteristic-based factors captures negligible fraction of total R^2 for both anomaly-based portfolios (Figure 11a) and WFR-based portfolios (Figure 11b). In the PC space (dashed lines), on the contrary, very sparse models perform well: a model with only two factors captures a large fraction of the total OOS cross-sectional R^2 for anomaly-based portfolios, while a model with a single PC delivers high OOS R^2 for WFR-based portfolios. Therefore, once again we witness strong evidence in favor of sparsity in PC space, but not in the space of original characteristics.

Figure 12 show contour maps depicting cross-validated 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). Warmer (yellow) colors reflect higher OOS R^2. Cross validation favors very aggressive regularization.

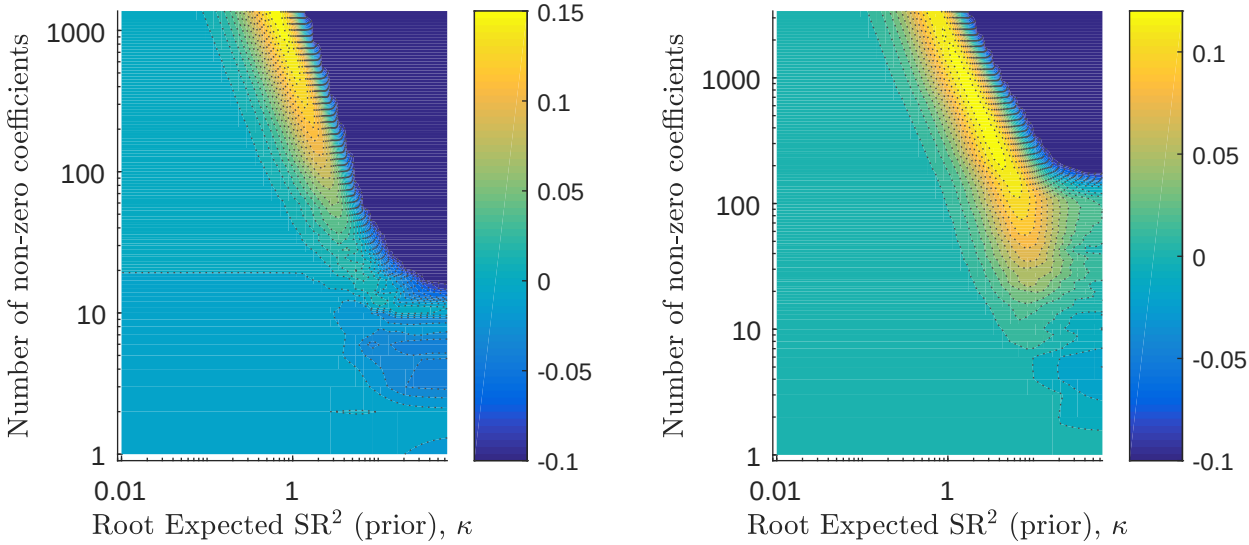


(a) 50 anomalies **(b)** WFR portfolios

Figure 11: 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.

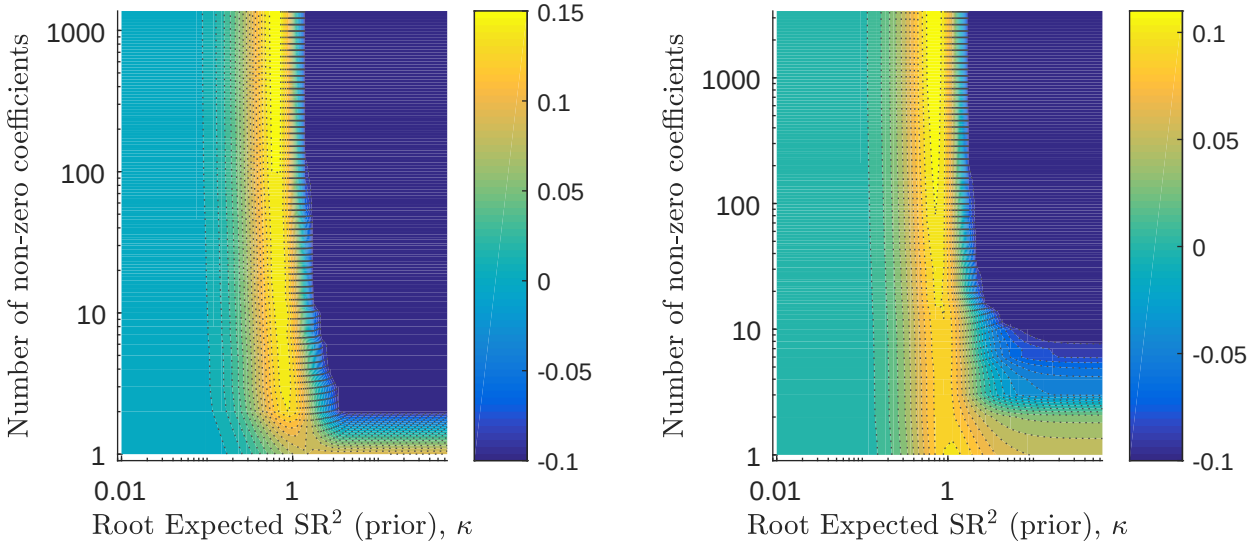
As usual, unregularized models (top-right corner) demonstrate extremely poor performance with OOS R^2 significantly below 0. L^2-penalty-only based models (top edge of a plot) perform very well and attain nearly maximal R^2, while L^1-only "Lasso" based models (right edge of the plot) completely fail in the space of characteristics (Panels a and b). In PC space (Panels c and d), a Lasso model based on a single PC delivers satisfactory results for both sets of test assets.

Results of this section suggest that with more diverse cross-sections of test assets and candidate factors (e.g., if we include interactions) our main findings become even more robust: (i) sparsity is prevalent in the PC space, but not in the space of characteristics; (ii) models with just the few largest PCs price the cross-section very well; and (iii) L^2-penalty-only based models, which do not target sparsity, perform on par with optimal sparse models.



(a) 50 anomalies

(b) PCs of WFR portfolios



(c) PCs of 50 anomalies

(d) PCs of WFR portfolios

Figure 12: Sparse Model Selection 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 to 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.

3.5 Performance in a Fully-Withheld Sample

Our cross validation method always evaluates a model's performance on the part of a sample not used in the estimation; it is therefore by construction an out-of-sample metric. Yet our choice of the strength of regularization (needs to be identified from full sample. It is possible that this model selection step doesn't generalize well on new or fully withheld data. To address this potential issue we evaluate the out-of-sample performance of our method on the most recent fully withheld part of the sample.

Another potential concern is deterioration of anomalies' alphas over time (McLean and Pontiff, 2016) and anomaly "fishing". If the cross-section of test assets we use had been data-mined in the same sample in which we perform cross validation, our results might be misleading. We emphasize two points in addressing this potential issue. First, in addition to using 50 anomaly portfolio as our primary dataset, we confirmed our findings in the WFR dataset, which is significantly less prone to "fishing". Second, although data mining is certainly a concern when assessing Sharpe ratios (because means of certain strategies might be "too" high relative to their risk), it does not necessarily have a big impact on the final SDF and OOS cross-sectional fit as measured by OOS cross-sectional R^2. The reason for this is that our method tends to shrink those "fished" estimates of expected returns substantially. McLean and Pontiff (2016) argue that many anomalies essentially disappear following a research paper publication. For "unknown" anomalies in the earlier part of the sample– there is no clear reason why expected returns need to line up with covariances in the data, as required by our model. If they don't, and mostly line up with small PCs, our method will shrink them considerably. It is, therefore, only anomalies that have been data-mined and correspond to large systematic co-movements in the data that could be problematic for our procedure.

We truncate the estimation sample to end in 2005, then redo the cross validated estimation on this truncated sample. The data post-2005 are fully excluded from all elements of the estimation process. Given the estimate $\hat{b}$ in the pre-2005 sample, we construct the time-series of the SDF $\hat{M}_t = 1 - \hat{b}'(F_t - \mathbb{E}F_t)$ in the withheld sample, or, equivalently, the return on the robust MVE portfolio $P_t = \hat{b}'F_t$. We then use the time series of P_t to assess the quality of the SDF and the mean-variance efficient (MVE) portfolio implied by our SDF. In particular, we focus on alphas of the MVE portfolio with respect to two benchmarks: CAPM and Fama-French 5-factor model.

We present the results in Table 5. We construct out-of-sample SDFs based on two sets of test assets: (i) 50 basic anomalies portfolios from Table 6 and (ii) all interactions and powers of 50 basic anomalies. We then regress OOS-MVE portfolio's returns on the market (second column) and on five Fama-French factors (plus the market; third column), and

Table 5: MVE portfolio's annualized OOS α (%) in the withheld sample (2005-2016)

We construct the OOS SDFs in the withheld sample (2005-2016) based on two sets of test assets: (i) 50 basic anomalies portfolios from Table 6 and (ii) all interactions and powers of 50 basic anomalies. We then regress MVE portfolio's returns on the market (second column) and on five Fama-French factors (plus the market; third column), and report annualized alphas in %. MVE portfolio returns are normalized to have the same standard deviation as the aggregate market. Standard errors in parentheses.

	MVE portfolio's CAPM α	MVE portfolio's FF 5-factor α
50 anomaly portfolios	11.54	5.56
	(5.50)	(4.67)
1,375 interactions of anomaly portfolios	26.05	20.36
	(5.50)	(4.94)

report annualized alphas in %. The MVE portfolio returns are normalized to have the same standard deviation as the aggregate market. We use only post-2005 data in our tests – this portion of the sample was fully withheld from our method and was never used in the estimation of the SDF.

Table 5 confirms that the MVE portfolio implied by our SDF performs well in the withheld data. At the same level of volatility as the aggregate market, the market-neutral portfolio constructed using 50 basic anomalies delivers annualized alphas of 12% and 6% using CAPM and Fama-French 5-factor models as benchmarks, respectively. An MVE portfolio based on the set of test assets which adds interactions, delivers alphas of 26% and 20% in the 2005-2016 sample using the same benchmarks, respectively. These alphas are statistically different from zero. We plot the time-series of returns on the two MVE portfolios in the withheld sample in Panels (a) and (b) of Figure 15 in the Appendix. Panel (c) shows the time-series of returns on the MVE portfolio based on interactions in full sample.

Tests of asset pricing models. In addition to being useful for evaluating the performance of our method on the withheld data, the estimated MVE portfolio can be viewed as a useful asset to test any potential model of the cross-section of equity returns. For example, He et al. (2016) construct a 2-factor SDF which includes the excess market return and the change in aggregate financial intermediary leverage. They estimate risk prices (SDF coefficients) using a range of assets including FF25 equity portfolios, maturity sorted bond portfolios, commodities, sovereign debt, equity index options, corporate bonds, CDS contracts, and currencies. Since our equity MVE portfolio, P_t, is constructed from a vast cross-section of equities, most of which are not included in their estimation, it provides a

nearly ideal "out-of-sample" test asset for their model. Given the time-series of their estimated SDF, $\frac{\Delta\Lambda_{t+1}}{\Lambda_t}$, we can simply check whether or not the Euler equation holds for P_t; that is, does $\mathbb{E}\left[P_{t+1}\frac{\Delta\Lambda_{t+1}}{\Lambda_t}\right] = 0$ hold in the data.

Similarly, if another researcher proposes a new factor model, our SDF presents a straightforward way to assess its pricing performance in the wide cross-section of anomalies. The candidate model should be tested in terms of its ability to explain mean returns on our SDF-implied MVE portfolio. That is, one only needs to run a single time-series regression of returns on our MVE portfolio on the factors of the model at hand, and check whether the intercept (alpha) is significantly different from zero. Because our SDF was constructed using a broad cross-section of anomalies and their interactions, explaining it (reducing alpha to zero), however, is a very high threshold to achieve for most models. Table 5 therefore can be viewed as a test of the the 5-factor model of Fama and French (2016) on its ability to price our OOS-MVE portfolio. Given the size of alphas and their statistical significance in the table, the 5-factor model is strongly rejected in the data.

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*.

Brandt, M. W., P. Santa-Clara, and R. Valkanov (2009). Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns. *Review of Financial Studies 22*(9), 3411–3447.

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

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 (2016). Dissecting anomalies with a five-factor model. *The Review of Financial Studies 29*(1), 69–103.

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*.

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.

He, Z., B. Kelly, and A. Manela (2016). Intermediary asset pricing: New evidence from many asset classes. Technical report, National Bureau of Economic Research.

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.

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 (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis 88*(2), 365–411.

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

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.

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.

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 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 $\lambda_1 \equiv \mathbb{E}\left[z_s^1 R_s\right]$? It's simply $\frac{1}{2}\left(\mu_A + \mu_B - \mu_C - \mu_D\right)$. Similarly, $\lambda_2 \equiv \mathbb{E}\left[r_s z_s^2\right] = \frac{1}{2}\left(-\mu_A + \mu_B - \mu_C + \mu_D\right)$ and $\lambda_{12} \equiv \mathbb{E}\left[r_s\left(z_s^1 z_s^2\right)\right] = \frac{1}{2}\left(-\mu_A + \mu_B + \mu_C - \mu_D\right)$. Since we have $(\mu_A + \mu_B + \mu_C + \mu_D) = 0$, we can easily recover μ_i from knowledge of $\lambda_1, \lambda_2, \lambda_{12}$ by the identity

$$\lambda \equiv \begin{bmatrix} 0 \\ \lambda_1 \\ \lambda_2 \\ \lambda_{12} \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ -1 & 1 & -1 & 1 \\ -1 & 1 & 1 & -1 \end{bmatrix}\begin{bmatrix} \mu_A \\ \mu_B \\ \mu_C \\ \mu_D \end{bmatrix} = G\mu \tag{46}$$

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?[21] 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. 46. The mean of the desired strategy is $\mu_D - \mu_C$, which is also equal to

$$\mu_D - \mu_C = 2\iota_{DC}' G^{-1}\lambda$$

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.[22] Hence, combining the interaction with the base strategies allows for construction of any "mixed" strategies. Conceptually, what's required is that the managed portfolios form a "basis" of the potential strategies.

B Alternative Estimators

We consider a number of alternative shrinkage estimators.

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

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

B.1 Pástor and Stambaugh (2000) Prior ($\eta = 1$)

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 $\frac{1}{1+\gamma} \approx 4.3\%$.

B.2 Uncertainty in Σ

- uncertainty in estimated Σ potentially a problem
- particularly when H/T isn't "small"
- this isn't very important in our case. How to know?
- Ignore uncertainty in μ. focus on Σ

B.2.1 Bayesian

Recall, there are H number of assets and T time periods. Consider a Wishart prior

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

A common (and reasonable) choice for Σ_0 when working with long-short excess returns is $\Sigma_0 = \sigma^2 I$. σ^2 can be chosen by empirical Bayes as $tr\left(\Sigma_T\right)/H$. For a given Σ_0, setting degrees of freedom parameter to H gives the most "diffuse" prior. For any choice of Σ_0, the posterior is given by

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

with mean value

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

In our main estimation with WRDS ratios (and momentum lags), $\frac{H}{H+T} \approx 0.5\%$. This is augmented with a "flat" prior on μ so that $\hat{\mu} = \mu_T$.

B.2.2 Ledoit and Wolf

In a series of papers (2003, 2004), Ledoit and Wolf propose two estimators of Σ which trade off bias and variance. They are both shrinkage estimators with different targets, Σ_0:

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

One choice of Σ_0 is the diagonal matrix $\frac{tr(\Sigma_T)}{H}I$. The other preserves all 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 RMSE), given the choice of Σ_0. We implement their algorithm on the 50 anomaly portfolios and find $a \approx 0.7\%$ for both methods, quite close to the Wishart shrinkage parameter of 0.5%. Ledoit and Wolf "concentrate on the covariance matrix alone without worrying about expected returns." Hence, $\hat{\mu} = \mu_T$.

Figure 2a shows the amount of shrinkage for various estimators, where shrinkage is measured by $\frac{\hat{b}_i}{b_{i,ols}}$ and $b_{ols} = \Sigma_T^{-1}\mu_T$. Therefore, though our estimator is superficially similar to covariance

shrinkage estimators, it is practically quite different. For comparison, we include the pure "level" shrinkage estimator based on the prior in Pástor and Stambaugh (2000), $\mu \sim \mathcal{N}(0, \Sigma_T)$.

B.3 Both Uncertain

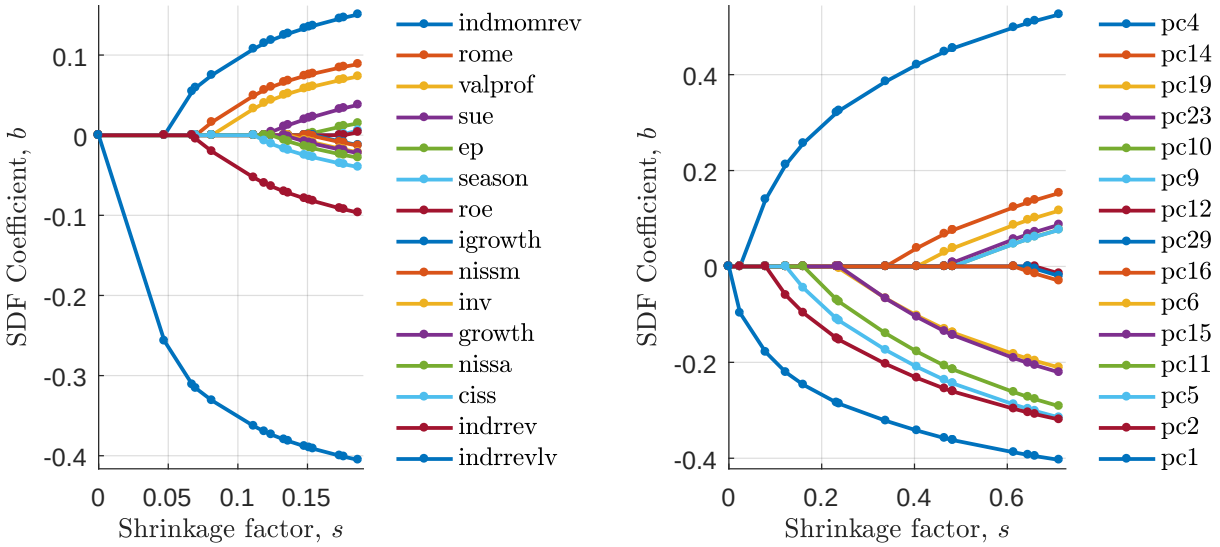
We allow for uncertainty in both μ and Σ. First, assume eigenvectors are known, so return covariance can be orthogonalized. Let D be the covariance of PC portfolios. For analytical tractability, we introduce independent scaled inverse-chi squared priors on each diagonal element of D^{-1}, with off-diagonals set to 0 always. 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. Then inference can be done PC-by-PC instead of jointly.

Let H be the number of assets, $\sigma^2 = tr(D_T)/H$, where D_T is the sample covariance matrix and T is the length of the data sample. Under the identity Wishart prior for D^{-1} (with known μ), we had $E_{prior}\left(d_i^{-1}\right) = \sigma^2$. 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:

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

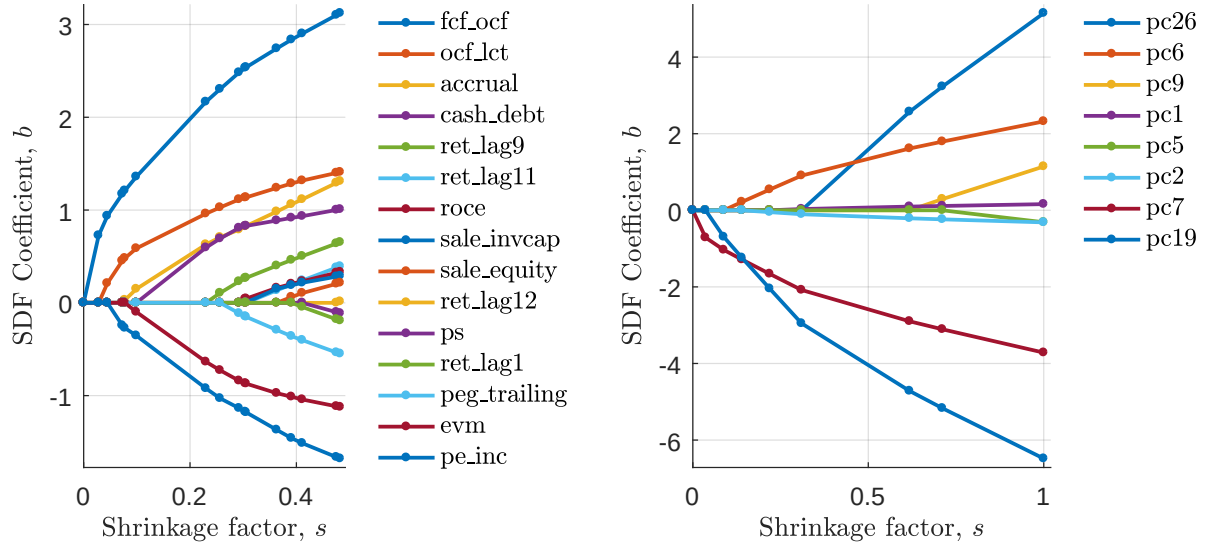
which preserves the uncertainty (degrees of freedom) in the prior.

C Supplementary Plots and Tables



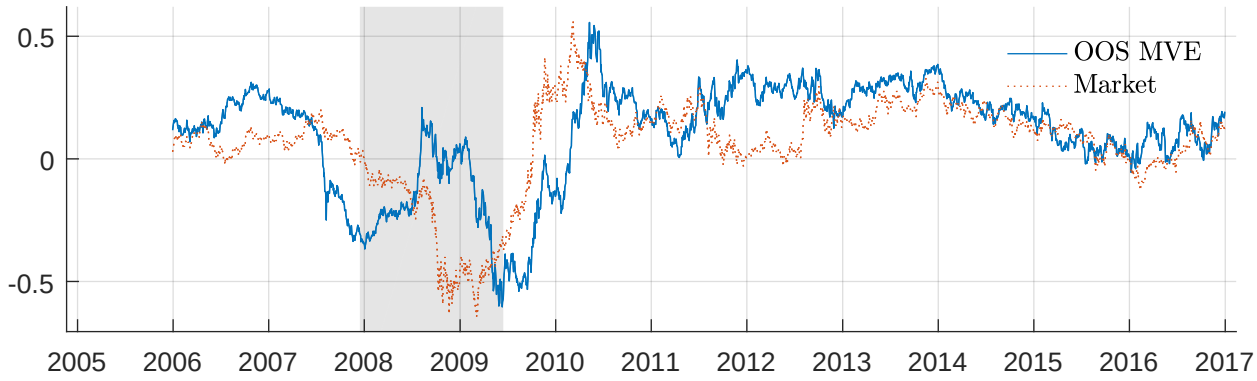
(a) Raw 50 anomaly portfolios **(b)** PCs of 50 anomaly portfolios

Figure 13: 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.

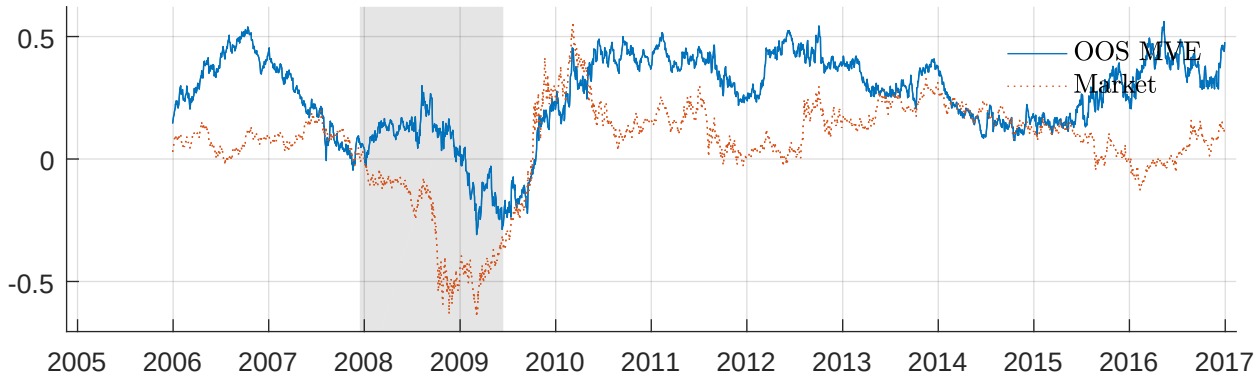


(a) Raw WFR portfolios **(b)** PCs of WFR portfolios

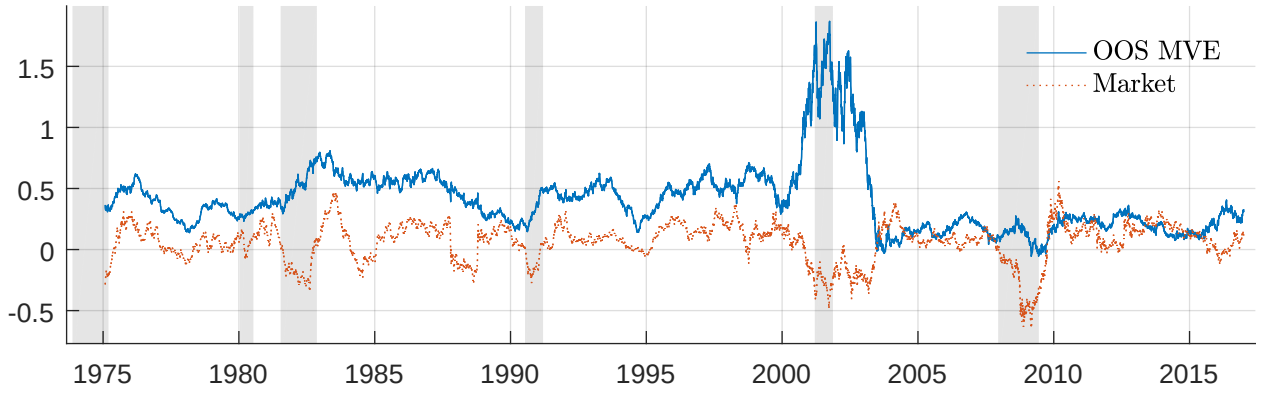
Figure 14: L^1 coefficient paths for the optimal model (WFR portfolios). Paths of coefficients based on the optimal (dual-penalty) sparse model that uses 80 WFR portfolios sorted portfolios (Panel a) and 80 PCs based on WFR portfolios (Panel b). Labels are ordered according to the vertical ordering of 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 15: 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 6: Part I: Mean annualized excess returns on anomaly portfolios, %

The table all basic "anomaly" characteristics used in our analysis and shows mean excess returns on managed portfolios which are linear in characteristics. Columns (1)-(3) show mean annualized returns (in %) for managed portfolios corresponding to all characteristics, net of risk-free rate, in the full sample, pre-2005 sample, and post-2005 sample, respectively. All managed portfolios' excess returns are 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	-2.9	-3.4	-1.5
2. Value (A)	6.5	8.4	1.4
3. Gross Profitability	3.4	3.2	3.9
4. Value-Profitablity	13.2	17.1	3.1
5. F-score	7.4	9.4	2.4
6. Debt Issuance	1.1	1.2	1.0
7. Share Repurchases	6.4	6.8	5.2
8. Net Issuance (A)	-8.6	-10.5	-3.4
9. Accruals	-5.7	-8.3	1.1
10. Asset Growth	-8.5	-10.4	-3.5
11. Asset Turnover	5.1	4.7	6.3
12. Gross Margins	-0.9	0.4	-4.2
13. Dividend/Price	3.4	4.9	-0.3
14. Earnings/Price	8.0	9.8	3.1
15. Cash Flows/Price	8.1	9.8	3.6
16. Net Operating Assets	1.7	2.7	-0.8
17. Investment/Assets	-9.5	-11.9	-3.2
18. Investment/Capital	-4.0	-4.5	-2.7
19. Investment Growth	-8.9	-10.4	-4.9
20. Sales Growth	-5.7	-5.5	-6.4
21. Leverage	5.2	6.9	0.8
22. Return on Assets (A)	1.9	1.0	4.1
23. Return on Book Equity (A)	4.0	4.5	2.7
24. Sales/Price	9.5	11.0	5.7
25. Growth in LTNOA	-2.3	-1.4	-4.5
26. Momentum (6m)	2.1	4.3	-3.5
27. Industry Momentum	5.5	7.8	-0.5
28. Value-Momentum	5.3	7.5	-0.6
29. Value-Momentum-Prof.	6.7	9.6	-0.9
30. Short Interest	-0.2	1.0	-3.5

continued on next page...

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

	(1)	(2)	(3)
31. Momentum (12m)	9.1	12.8	-0.7
32. Momentum-Reversals	-5.9	-7.6	-1.5
33. Long Run Reversals	-6.0	-7.8	-1.2
34. Value (M)	5.8	7.3	1.9
35. Net Issuance (M)	-7.8	-9.0	-4.7
36. Earnings Surprises	12.1	15.3	3.9
37. Return on Equity	9.6	11.7	4.3
38. Return on Market Equity	11.3	14.1	4.1
39. Return on Assets	6.7	7.5	4.5
40. Short-Term Reversals	-8.4	-12.0	0.9
41. Idiosyncratic Volatility	-1.6	-2.2	-0.2
42. Beta Arbitrage	-0.2	0.1	-0.8
43. Seasonality	12.1	18.8	-5.2
44. Industry Rel. Reversals	-18.1	-25.0	0.1
45. Industry Rel. Rev. (L.V.)	-35.1	-46.8	-4.6
46. Ind. Mom-Reversals	20.3	28.7	-1.4
47. Composite Issuance	-7.3	-9.1	-2.6
48. Price	-2.9	-2.7	-3.3
49. Age	3.1	4.1	0.6
50. Share Volume	-0.8	-1.0	-0.1

Table 7: Part I: Mean annualized excess returns on WFR portfolios, %

The table lists all basic WFR characteristics used in our analysis and shows mean excess returns on managed portfolios which are linear in characteristics. Columns (1)-(3) show mean annualized returns (in %) for managed portfolios corresponding to all characteristics, net of risk-free rate, in the full sample, pre-2005 sample, and post-2005 sample, respectively. All managed portfolios' excess returns are 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)	-10.3	-11.1	-7.6
2. P/E (Diluted, Incl. EI)	-13.2	-14.8	-8.0
3. Price/Sales	-7.8	-8.6	-5.2
4. Price/Cash flow	-4.4	-4.0	-5.8
5. Enterprise Value Multiple	-10.0	-11.2	-6.2
6. Book/Market	4.2	5.3	0.9
7. Shillers Cyclically Adjusted P/E Ratio	-5.7	-7.3	-0.2
8. Dividend Payout Ratio	-2.0	-2.4	-0.9
9. Net Profit Margin	1.5	2.5	-1.8
10. Operating Profit Margin Before Depreciation	2.2	3.4	-2.0
11. Operating Profit Margin After Depreciation	2.3	3.6	-1.7
12. Gross Profit Margin	1.2	2.5	-3.4
13. Pre-tax Profit Margin	2.2	3.3	-1.2
14. Cash Flow Margin	0.9	1.7	-1.8
15. Return on Assets	6.6	7.0	5.4
16. Return on Equity	6.5	7.2	3.8
17. Return on Capital Employed	8.1	8.3	7.7
18. After-tax Return on Average Common Equity	7.1	8.2	3.3
19. After-tax Return on Invested Capital	5.4	5.7	4.1
20. After-tax Return on Total Stockholders Equity	7.0	8.1	3.2
21. Pre-tax return on Net Operating Assets	6.6	7.7	2.9
22. Pre-tax Return on Total Earning Assets	6.3	7.5	2.4
23. Common Equity/Invested Capital	1.2	1.4	0.6
24. Long-term Debt/Invested Capital	-0.4	-0.5	-0.3
25. Total Debt/Invested Capital	-0.6	-0.5	-1.0
26. Interest/Average Long-term Debt	3.8	4.6	1.1
27. Interest/Average Total Debt	4.1	4.8	2.1
28. Cash Balance/Total Liabilities	1.0	1.5	-0.4
29. Inventory/Current Assets	-0.2	-1.1	2.8
30. Receivables/Current Assets	0.5	0.2	1.4

continued on next page...

Table 7: Part II: Mean annualized excess returns on WFR portfolios, %

	(1)	(2)	(3)
31. Total Debt/Total Assets	-2.6	-2.9	-1.6
32. Short-Term Debt/Total Debt	-0.6	0.2	-3.4
33. Current Liabilities/Total Liabilities	2.3	3.3	-0.7
34. Long-term Debt/Total Liabilities	-4.7	-5.9	-0.9
35. Free Cash Flow/Operating Cash Flow	16.6	20.5	3.6
36. Avertising Expenses/Sales	2.0	2.4	0.6
37. Profit Before Depreciation/Current Liabilities	3.3	4.2	0.3
38. Total Debt/EBITDA	-1.5	-1.5	-1.6
39. Operating CF/Current Liabilities	11.4	14.4	1.7
40. Total Liabilities/Total Tangible Assets	2.7	3.8	-0.8
41. Long-term Debt/Book Equity	-1.2	-1.4	-0.5
42. Total Debt/Total Assets	2.1	2.0	2.1
43. Total Debt/Capital	1.0	1.1	0.5
44. Total Debt/Equity	1.8	1.7	2.0
45. After-tax Interest Coverage	4.4	5.0	2.6
46. Cash Ratio	0.8	1.6	-1.8
47. Quick Ratio (Acid Test)	-1.1	-0.8	-2.2
48. Current Ratio	-1.4	-1.4	-1.6
49. Capitalization Ratio	-0.3	-0.4	0.1
50. Cash Flow/Total Debt	11.5	13.6	4.5
51. Inventory Turnover	2.6	3.5	-0.3
52. Asset Turnover	6.1	6.1	6.0
53. Receivables Turnover	3.0	3.0	3.0
54. Payables Turnover	-2.1	-4.2	5.2
55. Sales/Invested Capital	8.1	8.2	7.9
56. Sales/Stockholders Equity	7.3	7.6	6.2
57. Sales/Working Capital	2.6	3.1	0.7
58. Research and Development/Sales	3.0	3.5	1.4
59. Accruals/Average Assets	12.4	13.7	8.2
60. Gross Profit/Total Assets	6.4	6.7	5.2
61. Book Equity	0.4	0.5	0.2
62. Cash Conversion Cycle (Days)	-3.0	-3.6	-1.2
63. Effective Tax Rate	4.0	4.5	2.3
64. Interest Coverage Ratio	5.9	6.4	4.5
65. Labor Expenses/Sales	0.9	1.5	-0.9
66. Dividend Yield	3.4	4.3	0.4
67. Price/Book	-5.4	-6.4	-2.0
68. Trailing P/E to Growth (PEG) ratio	-10.7	-11.9	-6.9
69. Month $t-1$	-8.9	-11.8	0.7
70. Month $t-2$	-0.2	0.3	-1.9

continued on next page...

Table 7: Part III: Mean annualized excess returns on WFR portfolios, %

	(1)	(2)	(3)
71. Month $t-3$	3.0	4.3	-1.5
72. Month $t-4$	3.0	3.1	2.7
73. Month $t-5$	2.5	3.4	-0.6
74. Month $t-6$	5.9	8.4	-2.3
75. Month $t-7$	3.4	4.8	-1.2
76. Month $t-8$	3.5	3.5	3.4
77. Month $t-9$	9.6	11.9	2.0
78. Month $t-10$	5.4	8.4	-4.7
79. Month $t-11$	8.8	8.2	10.8
80. Month $t-12$	7.5	9.7	-0.1