Journal of Financial Economics 135 (2020) 271–292

ELSEVIER

Contents lists available at ScienceDirect

Journal of Financial Economics

journal homepage: www.elsevier.com/locate/jfec

Shrinking the cross-section[☆]

Serhiy Kozak[a,*], Stefan Nagel[b], Shrihari Santosh[c]

[a] University of Maryland, 7621 Mowatt Lane, College Park, MD 20742, United States
[b] University of Chicago, NBER, and CEPR, 5807 S Woodlawn Ave, Chicago, IL 60637, United States
[c] University of Colorado at Boulder, 995 Regent Dr, Boulder, CO 80309, United States

ARTICLE INFO

Article history:
Received 21 November 2017
Revised 23 July 2018
Accepted 14 August 2018
Available online 15 June 2019

JEL classification:
C11
G11
G12

Keyword:
Factor models
SDF
Cross section
Shrinkage
Machine learning,

ABSTRACT

We construct a robust stochastic discount factor (SDF) summarizing 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 contributions of low-variance principal components of the candidate characteristics-based factors. We find that characteristics-sparse SDFs formed from a few such factors—e.g., the four- or five-factor models in the recent literature—cannot adequately summarize the cross-section of expected stock returns. However, an SDF formed from a small number of principal components performs well.

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 that is linear in only a few such factors. Unfortunately, it seems that as new cross-sectional predictors emerge, these factor models need to be modified and expanded to capture the new evidence: Fama and French (1993) proposed a three factor model, Hou et al. (2015) have moved on to four, Fama and French (2015) to five factors, and Barillas and Shanken (2018) argue for a six-factor model. Even so, research in this area has tested these factor models only on portfolios constructed from a relatively small subset of known cross-sectional return predictors. These papers do not tell us how well characteristics-sparse factor models would do if one confronted them with a much larger set of cross-sectional return predictors—and an examination of this question is statistically challenging due to the high-dimensional nature of the problem.[1]

[☆] We thank Svetlana Bryzgalova, Mikhail Chernov, John Cochrane, Kent Daniel, Gene Fama, Stefano Giglio, Amit Goyal, Lars Hansen, Raymond Kan, Bryan Kelly, Ralph Koijen, Lubos Pastor, Michael Weber, Goufu Zhou, the referee Jay Shanken, participants at BYU Red Rock Conference, Colorado Finance Summit, New Methods For the Cross-Section of Returns Conference, NBER Asset Pricing, UBC Winter Finance Conference, and seminars at ASU, City University of Hong Kong, Dartmouth, HEC Montreal/McGill joint seminar, HKUST, Lausanne, Michigan, Stanford, Stockholm School of Economics, UCLA, UCSD, Washington University in St. Louis, and Yale for helpful comments and suggestions.

[*] Corresponding author.
E-mail addresses: sekozak@umd.edu (S. Kozak), stefan.nagel@chicagobooth.edu (S. Nagel), shrihari.santosh@colorado.edu (S. Santosh).

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

https://doi.org/10.1016/j.jfineco.2019.06.008
0304-405X/© 2019 Elsevier B.V. All rights reserved.

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

The conventional approach would be to estimate SDF coefficients with a cross-sectional regression of average returns on covariances of returns and factors. Due to the large number of potential factors, this conventional approach would lead to spurious overfitting. To overcome this high-dimensionality challenge, we use a Bayesian approach with a novel specification of prior beliefs. Asset pricing models of various kinds generally imply that much of the variance of the SDF should be attributable to high-eigenvalue (i.e., high-variance) principal components (PCs) of the candidate factor returns. Put differently, first and second moments of returns should be related. Therefore, if a factor earns high expected returns, it must either itself be a major source of variance or load heavily on factors that are major sources of variance. This is true not only in rational expectations models in which pervasive macroeconomic risks are priced but also, under plausible restrictions, in models in which cross-sectional variation in expected returns arises from biased investor beliefs (Kozak et al., 2018).

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

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

To allow for factor selection, we augment the estimation criterion with an additional penalty on the sum of absolute SDF coefficients (L^1 norm), which is typically used in lasso regression (Tibshirani, 1996) and naturally leads to sparse solutions. Our combined specification employs both L^1 and L^2 penalties, similar to the elastic net technique in machine learning. This combined specification achieves our two primary goals: (i) regularization based on an economically motivated prior, and (ii) it allows for sparsity by setting some SDF coefficients to zero. We pick the strength of penalization to maximize the (cross-validated) cross-sectional OOS R^2.

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

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

Interestingly, the results on sparsity are very different if we first transform the characteristics-portfolio returns into their PCs before applying our dual-penalty method. A sparse SDF that includes a few of the high-variance PCs delivers a good and robust out-of-sample fit of the cross-section of expected returns. Little is lost, in terms of explanatory power, by setting the SDF coefficients of low-variance PCs to zero. This finding is robust across our three primary sets of portfolios and the two extremely high-dimensional data sets 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. (2018) 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 three, four, or five 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) six-factor model in the out-of-sample period 2005–2017, allowing us to reject the hypothesis that the six-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 Brandt et al. (2009) and DeMiguel et al. (2009). Moreover, as DeMiguel et al. (2009) show, and as can be readily seen from the analytic expression of our estimator, portfolio optimization under L^2-norm constraints on weights shares similarities with portfolio optimization with a covariance matrix shrunk toward the identity matrix, as in Ledoit and Wolf (2004). However, despite some similarity of the solutions, there are important differences. First, our L^2-penalty results in level shrinkage of all SDF coefficients toward zero. This would not be the case with a shrunk covariance matrix. Second, in covariance matrix shrinkage approaches, the optimal amount of shrinkage would depend on the size of the parameter uncertainty in covariance estimation. Higher uncertainty about the covariance matrix parameters would call for stronger shrinkage. In contrast, our estimator is derived under the assumption that the covariance matrix is known (we use daily returns to estimate covariances precisely) and means are unknown. Shrinkage in our case is due to this uncertainty about means and our economically motivated 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 covariance uncertainty has negligible impact on coefficient estimates once uncertainty in means is accounted for.

Our paper contributes to an emerging literature that applies machine learning techniques in asset pricing to deal with the high-dimensionality challenge. Rapach et al. (2013) applies lasso to select a few predictors from a large set of candidates to forecast global stock markets. Stambaugh and Yuan (2016) use covariance cluster analysis to identify two groups of "related" anomalies and then construct factors based on stocks' average within-cluster characteristic rank. Kelly et al. (2018) show how to perform dimensionality reduction of the characteristics space. They extend projected-PCA (Fan et al., 2016) to allow for time-varying factor loadings and apply it to extract common latent factors from the cross-section of individual stock returns. Their method explicitly maps these latent factors to principal components of characteristic-managed portfolios (under certain conditions). Kelly et al. (2018) and Kozak et al. (2018) further show that an SDF constructed using few such dominant principal components prices the cross-section of expected returns reasonably well. The selection of a few dominant sources of covariance as pricing factors in these papers is an ad-hoc imposition of the asset pricing restriction that links factor mean returns and variances. Rather than imposing a PC-sparse SDF representation ex-ante, our methodology automatically recovers such sparsity if it improves out-of-sample performance.

DeMiguel et al. (2017), Freyberger et al. (2017) and Feng et al. (2017) focus on characteristics-based factor selection in lasso-style estimation with L^1-norm penalties. Their findings are suggestive of a relatively high degree of redundancy among cross-sectional stock return predictors. Yet, as our results show, for the purposes of SDF estimation with characteristics-based factors, a focus purely on factor selection with L^1-norm penalties is inferior to an approach with L^2-norm penalties that shrinks SDF coefficients toward zero to varying degrees but does not impose sparsity on the SDF coefficient vector. This is in line with results from the statistics literature where researchers have noted that lasso does not perform well when regressors are correlated, and that ridge regression (with L^2-norm penalty) or elastic net (with a combination of L^1- and L^2-norm penalties) delivers better prediction performance than lasso in these cases (Tibshirani, 1996; Zou and Hastie, 2005). Since many of the candidate characteristics-based factors in our application have substantial correlation, it is to be expected that an L^1-norm penalty alone will lead to inferior prediction performance. For example, instead of asking the estimation procedure to choose between the value factor and the correlated long-run-reversals factor for the sake of sparsity in terms of characteristics, it appears to be beneficial, 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. Most papers focus on estimating risk premia, i.e., the extent to which a stock characteristic is associated with variation in expected returns.[2] 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 2005–2017, a period that starts after or overlaps very little with the sample period used in studies that uncovered the anomalies (McLean and Pontiff, 2016).

2. Asset pricing with characteristics-based factors

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

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

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

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

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

2.1. Characteristics-based factor SDF

Characteristics-based asset pricing models parametrize the SDF loadings as

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

where Z_{t-1} is an $N \times H$ matrix of asset characteristics, and b is an $H \times 1$ vector of time-invariant coefficients. Without further restrictions, this representation is without loss of generality.[3] To obtain models with empirical content, researchers search for a few measurable asset attributes that approximately span b_{t-1}. For example, Fama and French (1993) use two characteristics: market capitalization and the book-to-market equity ratio. Our goal is to develop a statistical methodology that allows us to entertain a large number of candidate characteristics and estimate their coefficients b in such a high-dimensional setting.

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

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

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

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

[2] See, for example, Freyberger et al. (2017), Moritz and Zimmermann (2016), Huerta et al. (2013), and Tsai et al. (2011). An exception is Feng et al. (2017).

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

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

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

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

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

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

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

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

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

2.2. Sparsity in characteristics-based factor returns

Much of the existing characteristics-based factor model literature has sidestepped this high-dimensionality problem by focusing on models that include only a small number of factors. We will refer to such models as characteristics-sparse models. Whether such a characteristics-sparse model can adequately describe the SDF in a cross-section with a large number of stock characteristics is a key empirical question that we aim to answer in this paper.

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

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

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

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

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

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

In practice, however, neither expected profitability nor (planned) investment are observable. The usual approach is to use proxies, such as lagged profitability and lagged investment as potential predictors of unobserved quantities. Yet many additional characteristics are likely relevant for capturing expected profitability and planned investment and, therefore, expected returns. Moreover, considering that the model above is a vast simplification of reality to begin with, many more factors are likely to be required to approximate an SDF of a more realistic and complex model. The bottom line is that, in practice, q-theory does not necessarily provide much economic reason to expect sparse SDFs in the space of observable characteristics.

For this reason, we pursue an approach that does not impose that the SDF is necessarily characteristics-sparse. Moreover, it leads us to seek a method that can accommodate an SDF that involves a potentially very large number of characteristics-based factors, but at the same time, still ensures good out-of-sample performance and robustness against in-sample overfitting. At the same time, we would also like our method to be able to handle cases in which some of the candidate factors are not contributing to the

SDF at all. This situation may be particularly likely to arise if the analysis includes characteristics that are not known, from prior literature, to predict returns in the cross-section. It could also arise if there is truly some redundancy among the cross-sectional return predictors documented in the literature. To accommodate these cases, we want our approach to allow for the possibility of sparsity but without necessarily requiring sparsity to perform well out of sample. This will then allow us to assess the degree of sparsity empirically.

2.3. Sparsity in principal components of characteristics-based factor returns

While there are not strong economic reasons to expect characteristics-sparsity of the SDF, one may be able to find rotations of the characteristics factor data that admit, at least approximately, a sparse SDF representation. Motivated by the analysis in Kozak et al. (2018), we consider sparse SDF representations in the space of PCs of characteristic-based factor returns.

Based on the eigendecomposition of the factor covariance matrix,

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

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

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

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

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

In Kozak et al. (2018), we argue that absence of near-arbitrage (extremely high Sharpe ratios) implies that factors earning substantial risk premia must be major sources of co-movement. This conclusion obtains under very mild assumptions and applies equally to “rational” and “behavioral” models. Furthermore, for typical sets of test assets, returns have a strong factor structure dominated by a small number of PCs with the highest variance (or eigenvalues d_j). Under these two conditions, an SDF with a small number of these high-variance PCs as factors should explain most of the cross-sectional variation in expected returns. Motivated by this theoretical result, we explore empirically whether an SDF sparse in PCs can be sufficient to describe the cross-section of expected returns, and we compare it, in terms of their pricing performance, with SDFs that are sparse in characteristics.

3. Methodology

Consider a sample with size T. We denote

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

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

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

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

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

The resulting estimator is the sample version of Eq. (6),[4]

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

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

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

3.1. Shrinkage estimator

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

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

where $\tau = \text{tr}[\Sigma]$, and κ is a constant controlling the “scale” of μ that may depend on τ and H. As we will discuss, this family encompasses priors that have appeared in earlier asset pricing studies, albeit not in a high-dimensional setting. At this general level, this family of priors can broadly capture the notion—consistent with

[4] When $T < H$, we use Moore–Penrose pseudoinverse of the covariance matrix.

a wide class of asset pricing theories—that first moments of factor returns have some connection to their second moments. Parameter η controls the "shape" of the prior. It is the key parameter for the economic interpretation of the prior because it determines how exactly the relationship between first and second moments of factor returns is believed to look like under the prior.

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

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

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

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

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

This fact about the distribution of eigenvalues immediately makes clear that the assumption of $\eta = 0$ (as, e.g., in Harvey et al., 2008) is economically implausible. In this case, the mean Sharpe ratio of a PC factor in Eq. (20) is inversely related to the PC's eigenvalue. Therefore, the prior implies that the expected Sharpe ratios of low-eigenvalue PCs explode toward infinity. In other words, $\eta = 0$ would imply existence of near-arbitrage opportunities. As Kozak et al. (2018) discuss, existence of near-arbitrage opportunities is not only implausible in rational expectations models, but also in models in which investors have biased beliefs, as long as some arbitrageurs are present in the market.

Pástor (2000) and Pástor and Stambaugh (2000) work with $\eta = 1$. This assumption is more plausible in the sense that it is consistent with absence of near-arbitrage opportunities. However, as Eq. (20) makes clear, $\eta = 1$ implies that Sharpe ratios of low-eigenvalue PCs are expected to be of the same magnitude as Sharpe ratios of high-eigenvalue PCs. We do not view this as economically plausible. For instance, in rational expectations models in which cross-sectional differences in expected returns arise from exposure to macroeconomic risk factors, risk premia are typically concentrated in one or a few common factors. This means that Sharpe ratios of low-eigenvalue PCs should be smaller than those of the high-eigenvalue PCs that are the major source of risk premia. Kozak et al. (2018) show that a similar prediction also arises in plausible behavioral models in which investors have biased beliefs. They argue that to be economically plausible, such a model should include arbitrageurs in the investor population, and it should have realistic position size limits (e.g., leverage constraints or limits on short selling) for the biased-belief investors (who are likely to be less sophisticated). As a consequence, biased beliefs can only have substantial pricing effects in the cross-section if these biased beliefs align with high-eigenvalue PCs; otherwise, arbitrageurs would find it too attractive to aggressively lean against the demand from biased investors, leaving very little price impact. To the extent it exists, mispricing then appears in the SDF mainly through the risk prices of high-eigenvalue PCs. Thus, within both classes of asset pricing models, we would expect Sharpe ratios to be increasing in the eigenvalue, which is inconsistent with $\eta \le 1$.

Moreover, the portfolio that an unconstrained rational investor holds in equilibrium should have finite portfolio weights. Indeed, realistic position size limits for the biased-belief investors in Kozak et al. (2018) discussed above translate into finite equilibrium arbitrageur holdings and therefore finite SDF coefficients. Our prior should be consistent with this prediction. Since the optimal portfolio weights of a rational investor and SDF coefficients are equivalent, we want a prior that ensures $b'b$ remains bounded. A minimal requirement for this to be true is that $\mathbb{E}[b'b]$ remains bounded. With $b = \Sigma^{-1}\mu$, the decomposition $\Sigma = QDQ'$, and the prior (18), we can show

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

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

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

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

where $\gamma = \frac{\tau}{\kappa^2 T}$.[5] The posterior variance of b is given by

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

which we use in Section 4 to construct confidence intervals.

[5] We obtain this formula by first computing the posterior mean of μ based on the standard formula for the conjugate multivariate normal prior with a known covariance matrix. That is, letting the prior parameters $\mu_0 = 0$ and $\Sigma_0 = \frac{\kappa^2}{\tau}\Sigma^{\eta}$, we get the posterior means $\hat{\mu} = (\Sigma_0^{-1} + T\Sigma^{-1})^{-1}(\Sigma_0^{-1}\mu_0 + T\Sigma^{-1}\bar{\mu}) = (\Sigma + \gamma\Sigma^{(2-\eta)})^{-1}\Sigma\bar{\mu}$. Next, we use the fact that $\hat{b} = \Sigma^{-1}\hat{\mu}$ and $\eta = 2$ to obtain Eq. (22).

3.1.1. Economic interpretation

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

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

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

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

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

3.1.2. Representation as a penalized estimator

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

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

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

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

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

Eqs. (26) and (27) resemble ridge regression, a popular technique in machine learning (e.g., see Hastie et al., 2011) but with some important differences. A standard ridge regression objective function would impose a penalty on the L^2-norm of coefficients, $b'b$ in Eq. (26), or, equivalently, weight the pricing errors with the identity matrix instead of Σ^{-1} in Eq. (27). One can show that this standard ridge regression would correspond to a prior with $\eta = 3$, which would imply even more shrinkage of low-eigenvalue PCs than with our prior of $\eta = 2$. However, the estimator one obtains from a standard ridge approach is not invariant to how the estimation problem is formulated. For example, if one estimates factor risk premia λ in a beta-pricing formulation of the model, minimizing $(\bar{\mu} - I\lambda)'(\bar{\mu} - I\lambda)$ subject to a standard ridge penalty on $\lambda'\lambda$, the resulting estimator corresponds to a prior with $\eta = 1$, that, as we have argued, is not economically plausible. In contrast, in our approach the estimator is pinned down by the asset pricing Eq. (5) combined with the economically motivated prior (18).

3.2. Sparsity

The method that we have presented so far deals with the high-dimensionality challenge by shrinking SDF coefficients toward zero, but none of the coefficients are set to exactly zero. In other words, the solution we obtain is not sparse. As we have argued in Section 2, the economic case for extreme sparsity with characteristics-based factors is weak. However, it may be useful to allow for the possibility that some factors are truly redundant in terms of their contribution to the SDF. Moreover, there are economic reasons to expect that a representation of the SDF that is sparse in terms of PCs could provide a good approximation.

For these reasons, we now introduce an additional L^1 penalty $\gamma_1 \sum_{j=1}^{H} |b_j|$ in the penalized regression problem given by Eq. (27). The approach is motivated by lasso regression and elastic net (Zou and Hastie, 2005), which combines lasso and ridge penalties. Due to the geometry of the L^1 norm, it leads to some elements of $\hat{b}$ being set to zero, that is, it accomplishes sparsity and automatic factor selection.[6] The degree of sparsity is controlled by the strength of the penalty. Combining both L^1 and L^2

[6] L^2 regularization penalizes the square of SDF weights, while L^1 regularization penalizes their absolute value. Relative to L^1 regularization, L^2 regularization therefore focuses on pushing big weights down substantially more than tiny ones. More precisely, the first derivative of the L^2 penalty in a small neighborhood around a zero SDF weight is approximately zero. Therefore, changing the weight from zero to a small number has virtually no effect on the penalty. As a consequence, the estimator retains an SDF factor that does not contribute much explanatory power with a small, but nonzero, weight. In contrast, for the L^1, the first derivative is far from zero even in a close neighborhood around zero SDF weight. Changing a weight from a small value to exactly zero can therefore have a substantial effect on the penalty, which makes it "costly" to retain SDF factors with small explanatory power at a nonzero weight.

penalties, our estimator solves the problem[7]:

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

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

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

While we will ultimately let the data speak about the optimal values of the penalties γ_1 and γ_2, there is reason to believe that completely switching off the L^2 penalty and focusing purely on lasso-style estimation would not work well in this asset pricing setting. Lasso is known to suffer from relatively poor performance compared with ridge and elastic net when regressors are highly correlated (Tibshirani, 1996; Zou and Hastie, 2005). An L^2 penalty leads the estimator to shrink coefficients of correlated predictors toward each other, allowing them to borrow strength from each other (Hastie et al., 2011). In the extreme case of k identical predictors, they each get identical coefficients with $1/k$th the size that any single one would get if fit alone. The L^1 penalty, on the other hand, ignores correlations and will tend to pick one variable and disregard the rest. This hurts performance because if correlated regressors each contain a common signal and uncorrelated noise, a linear combination of the regressors formed based on an L^2 penalty will typically do better in isolating the signal than a single regressor alone. For instance, rather than picking book-to-market as the only characteristic to represent the value effect in an SDF, it may be advantageous to consider a weighted average of multiple measures of value, such as book-to-market, price-dividend, and cashflow-to-price (CF/P) ratios. This reasoning also suggests that an L^1-only penalty may work better when we first transform the characteristics-based factors into their PCs before estimation. We examine this question in our empirical work below.

3.3. Data-driven penalty choice

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

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

and hence γ implicitly represents views about the expected squared Sharpe ratio. For example, an expectation that the maximum Sharpe ratio cannot be very high, i.e., low κ, would imply high γ and hence a high degree of shrinkage imposed on the estimation. Some researchers pick a prior belief based on intuitive reasoning about the likely relationship between the maximum squared Sharpe Ratio and the historical squared Sharpe ratio of a market index.[8] However, these are intuitive guesses. It would be difficult to go further and ground beliefs about κ in deeper economic analyses of plausible degrees of risk aversion, risk-bearing capacity of arbitrageurs, and degree of mispricing. For this reason, we prefer a data-driven approach. But we will make use of Eq. (29) to express the magnitude of the L^2-penalty that we apply in estimation in terms of an economically interpretable root expected maximum squared Sharpe ratio.

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

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

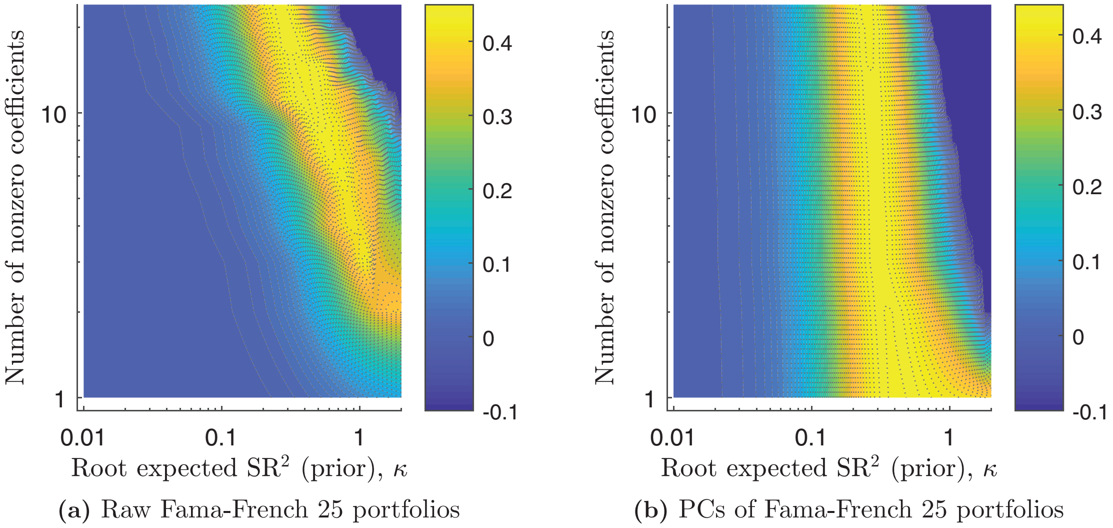
where the subscript 2 indicates an OOS sample moment from the withheld sample. We repeat this procedure K times, each time treating a different subsample as the OOS data. We then average the R^2 across these K estimates, yielding the cross-validated R^2_{oos}. Finally, we choose γ (or γ_1, γ_2) that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

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

This penalty choice procedure uses information from the whole sample to find the penalty parameters that minimize the R^2 based on Eq. (30). The cross-validated OOS R^2 at the optimal values of the penalty parameters is therefore typically an upward-biased estimate of the true OOS R^2 that one would obtain in a new data set that has not been used for penalty parameter estimation (Tibshirani and Tibshirani, 2009; Varma and Simon, 2006). Our interest centers on the optimal degree of regularization and

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

[8] Barillas and Shanken (2018) is a recent example. See also MacKinlay (1995) and Ross (1976) for similar arguments.



(a) Raw Fama-French 25 portfolios

(b) PCs of Fama-French 25 portfolios

Fig. 1. OOS R^2 from dual-penalty specification (Fama-French 25 ME/BM portfolios). OOS cross-sectional R^2 for families of models that employ both L^1 and L^2 penalties simultaneously using 25 Fama-French ME/BM-sorted portfolios (Panel a) and 25 PCs based on Fama and French portfolios (Panel b). We quantify the strength of the L^2 penalty by prior root expected SR^2 (κ) on the x-axis. We show the number of retained variables in the SDF, which quantifies the strength of the L^1 penalty, on the y-axis. Warmer (yellow) colors depict higher values of OOS R^2. Both axes are plotted on logarithmic scale. The sample is daily from July 1926 to December 2017. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

we therefore are only concerned about the relative performance of models at various degrees of regularization, not the level of the OOS R^2. For this purpose, the cross-validated OOS R^2 is a well-motivated target just like in the OLS case in which it is optimal to estimate parameters by maximizing the sample R^2 even though the sample R^2 is an upward-biased estimate of the OOS R^2. In a subsequent step, in Section 5, we perform a true OOS exercise where we evaluate the estimated SDF on a part of the sample that has not been used to estimate the penalty parameters.

4. Empirical analysis

4.1. Preliminary analysis: Fama-French ME/BM portfolios

We start with an application of our proposed method to daily returns on the 25 Fama-French ME/BM-sorted (FF25) portfolios from July 1926 to December 2017 to December 2017, which we orthogonalize with respect to the Center for Research in Security Prices (CRSP) value-weighted index return using βs estimated in the full sample.[9] Portfolio returns are further rescaled to have standard deviations equal to the in-sample standard deviation of the excess return on the aggregate market index. In this analysis, we treat the 25 portfolio membership indicators as stock characteristics, and we estimate the SDF's loadings on these 25 portfolios. These portfolios are not the challenging high-dimensional setting for which our method is designed, but this initial step is useful to verify that our method produces reasonable results before we apply it to more interesting and statistically challenging high-dimensional sets of asset returns where classic techniques are infeasible.

For the FF25 portfolios, we know quite well from earlier research what to expect, and we can check whether our method produces these expected results. From Lewellen et al. (2010), we know that the FF25 portfolio returns have such a strong factor structure that the 25 portfolio returns (orthogonalized with respect to the market index return) are close to being linear combinations of the SMB and HML factors. As a consequence, essentially any selection of a couple of portfolios out of the 25 with somewhat different loadings on the SMB and HML factors should suffice to span the SDF. Thus, treating the portfolio membership indicators as characteristics, we should find a substantial degree of sparsity. From Kozak et al. (2018), we know that the SMB and HML factors essentially match the first and the second PCs of the FF25 (market-neutral) portfolio returns. Therefore, when we run the analysis using the PCs of the FF25 portfolio returns as the basis assets, we should find even more sparsity: two PCs at most should be sufficient to describe the SDF well.

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

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

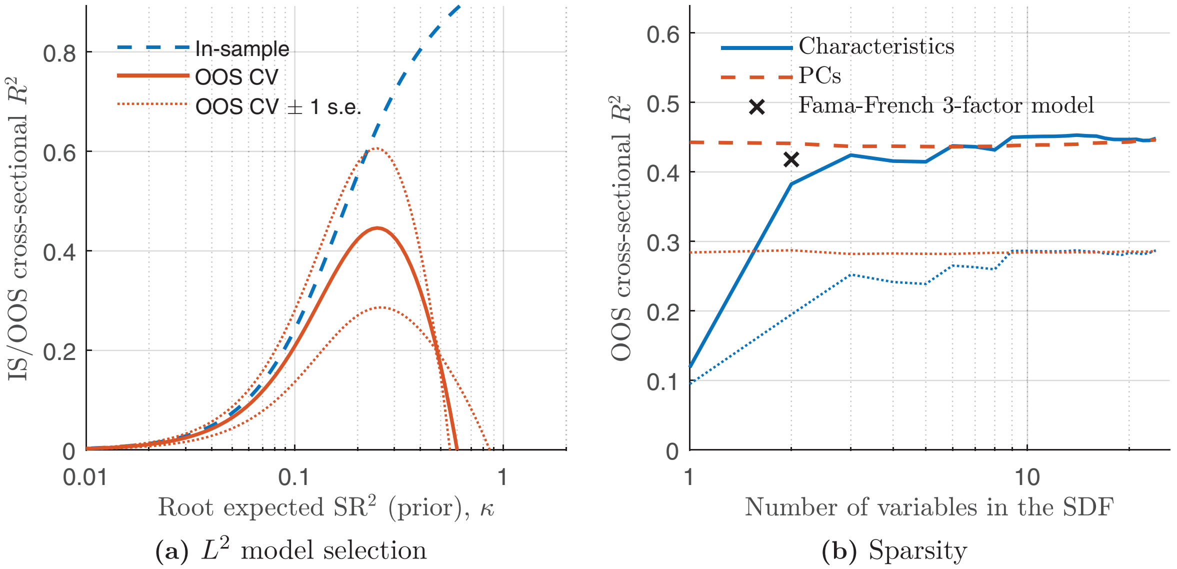


Fig. 2. L^2 model selection and sparsity (Fama-French 25 ME/BM portfolios). Panel (a) plots the in-sample cross-sectional R^2 (dashed) and OOS cross-sectional R^2 based on cross-validation (solid), with no sparsity imposed. Dotted lines depict ±1 standard error (s.e.) bounds of the CV estimator. In Panel (b), we show the maximum OOS cross-sectional R^2 attained by a model with n factors (on the x-axis) across all possible values of L^2 shrinkage, for models based on original characteristics portfolios (solid) and PCs (dashed). Dotted lines depict −1 s.e. bounds of the CV estimator. The X mark indicates OOS performance of the Fama-French model that uses only SMB and HML factors. The sample is daily from July 1926 to December 2017 to December 2017. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

remain in the SDF with nonzero coefficients. Thus, there is no sparsity at the top end of the plot and extreme sparsity at the bottom. Both axes are depicted on logarithmic scale.

The contour maps show the OOS R^2 calculated as in Eq. (30) for each of these penalty combinations. Our data-driven penalty choice selects the combination with the highest OOS R^2, but in this figure, we show the OOS R^2 for a wide range of penalties to illustrate how L^2-shrinkage and sparsity (L^1 penalty) influences the OOS R^2. Warmer (yellow) colors indicate higher OOS R^2. To interpret the magnitudes, it is useful to keep in mind that with our choice of $K = 3$, we evaluate the OOS R^2 in withheld samples of about 31 years in length, i.e., the OOS R^2 show how well the SDF explains returns averaged over a 31-year period.

Focusing first on the raw FF25 portfolio returns in Fig. 1a, we can see that for this set of portfolios, sparsity and L^2-shrinkage are substitutes in terms of ensuring good OOS performance: the contour plot features a diagonal ridge with high OOS R^2 extending from the top edge of the plot (substantial L^2-shrinkage, no sparsity) to the right-edge (substantial sparsity, no shrinkage). As we outlined above, this is what we would expect for this set of asset returns: a selection of two to three portfolios from these 25 should be sufficient to span the SDF that prices all 25 well, and adding more portfolio returns to the SDF hurts OOS performance unless more L^2-shrinkage is imposed to avoid overfitting. Unregularized models that include all 25 factors (top-right corner) perform extremely poorly in the OOS evaluation.[10]

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

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

Fig. 2b presents the OOS R^2 for various degrees of sparsity, choosing the optimal (i.e., OOS R^2 maximizing) amount of L^2-shrinkage for each level of sparsity. In other words, we are following the ridge of the highest values in the contour plots from the bottom edge of the plot to the

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

top. The solid blue line is based on the raw FF25 portfolio returns and the dashed red line based on the PCs. Dotted lines on the plot show approximate -1 standard error bounds for the CV estimator.[11] This plot makes even more transparent our earlier point that a sparse SDF with just a few of the FF25 portfolio is sufficient to get maximal OOS performance—comparable to the an SDF with SMB and HML shown by the black X[12]—and that in PC-space even one PC is enough. The PC that is eliminated last as we raise the degree of sparsity is PC1 (i.e., with the one with the highest variance). PC1 is highly correlated with the HML factor (and somewhat with SMB); the SDF based on PC1 is therefore effectively the same as Fama-French's and performs similarly.

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

4.2. Large sets of characteristics portfolios

We start with the universe of US firms in CRSP. We construct two independent sets of characteristics. The first set relies on characteristics underlying common “anomalies” in the literature. We follow standard anomaly definitions in Novy-Marx and Velikov (2016), McLean and Pontiff (2016), Kogan and Tian (2015), and Hou et al. (2015) and compile our own set of 50 such characteristics. The second set of characteristics is based on 70 financial ratios as defined by WRDS: “WRDS Industry Financial Ratios (WFR) is a collection of most commonly used financial ratios by academic researchers (often for purposes other than return prediction). There are in total over 70 financial ratios grouped into the following seven categories: capitalization, efficiency, financial soundness/solvency, liquidity, profitability, valuation, and others.” We supplement this data set with 12 portfolios sorted on past monthly returns in months $t-1$ through $t-12$. The combined data set contains 80 managed portfolios (we drop two variables due to their short time series and end up with 68 WRDS ratios in the final data set). We provide definitions of all variables in both data sets in Internet Appendix D.[13]

To focus exclusively on the cross-sectional aspect of return predictability, remove the influence of outliers, and keep constant leverage across all portfolios, we perform certain normalizations of characteristics that define our characteristics-based factors. First, similarly to Asness et al. (2014) and Freyberger et al. (2017), we perform a simple rank transformation for each characteristic. For each characteristic i of a stock s at a given time t, denoted as $c^i_{s,t}$, we sort all stocks based on the values of their respective characteristics $c^i_{s,t}$ and rank them cross-sectionally (across all s) from 1 to n_t, where n_t is the number of stocks at t for which this characteristic is available.[14] We then normalize all ranks by dividing by n_t+1 to obtain the value of the rank transform:

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

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

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

where $\bar{rc}^i_t=\frac{1}{n_t}\sum_{s=1}^{n_t}rc^i_{s,t}$. The resulting zero-investment long-short portfolios of transformed characteristics $z^i_{s,t}$ are insensitive to outliers and allow us to keep the absolute amount of long and short positions invested in the characteristic-based strategy (i.e., leverage) fixed. For instance, doubling the number of stocks at any time t has no effect on the strategy's gross exposure.[15] Finally, we combine all transformed characteristics $z^i_{s,t}$ for all stocks into a matrix of instruments Z_t.[16] Interaction with returns, $F_t=Z'_{t-1}R_t$, then yields one factor for each characteristic.

To ensure that the results are not driven by very small illiquid stocks, we exclude small-cap stocks with market caps below 0.01% of aggregate stock market capitalization at each point in time.[17] In all of our analysis, we use daily returns from CRSP for each individual stock. Using daily data allows us to estimate second moments much more precisely than with monthly data and focus on uncertainty in means while largely ignoring negligibly small uncertainty in covariance estimates (with exceptions as noted below). We adjust daily portfolio weights on individual stocks within each month to correspond to a monthly rebalanced buy-and-hold strategy during that month. Table 1 in the Internet Appendix shows the annualized mean returns for the anomaly portfolios. Mean returns for the WFR managed portfolios are reported in the Internet Appendix, Table 2. Finally, as in the previous section, we orthogonalize all portfolio returns with respect to the

[11] We estimate these by computing variance of the CV estimator under the assumption that $K=3$ CV estimates are i.i.d. In that case, $\mathrm{var}(R^2_{\text{CV estimator}})=\mathrm{var}(\frac{1}{K}\sum_{j=1}^{K}\hat{R_j}^2)\approx\frac{1}{K}\mathrm{var}(\hat{R_j}^2)$, where $\hat{R_j}^2$ is an estimate of the OOS R^2 in the jth fold of the data. Standard errors of the CV estimator can thus be computed as $\frac{1}{\sqrt{K}}\mathrm{sd}(\hat{R_1}^2,\ldots,\hat{R_K}^2)$.

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

[13] We make the data available at: https://www.serhiykozak.com/data.

[14] 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.

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

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

[17] For example, for an aggregate stock market capitalization of $20 trillion, we keep only stocks with market caps above $2 billion.

Table 1
Largest SDF factors (50 anomaly portfolios).
Coefficient estimates and absolute t-statistics at the optimal value of the prior root expected SR^2 (based on cross-validation). Panel (a) focuses on the original 50 anomaly portfolios. Panel (b) pre-rotates returns into PC space and shows coefficient estimates corresponding to these PCs. Coefficients are sorted descending on their absolute t-statistic values. The sample is daily from November 1973to December 2017.

(a) Raw 50 anomaly portfolios			(b) PCs of 50 anomaly portfolios		
	b	t-stat		b	t-stat
Industry rel. rev. (L.V.)	−0.88	3.53	PC 4	1.01	4.25
Ind. mom-reversals	0.48	1.94	PC 1	−0.54	3.08
Industry rel. reversals	−0.43	1.70	PC 2	−0.56	2.65
Seasonality	0.32	1.29	PC 9	−0.63	2.51
Earnings surprises	0.32	1.29	PC 15	0.32	1.27
Value-profitablity	0.30	1.18	PC 17	−0.30	1.18
Return on market equity	0.30	1.18	PC 6	−0.29	1.18
Investment/Assets	−0.24	0.95	PC 11	−0.19	0.74
Return on equity	0.24	0.95	PC 13	−0.17	0.65
Composite issuance	−0.24	0.95	PC 23	0.15	0.56
Momentum (12m)	0.23	0.91	PC 7	0.14	0.56

Table 2
Largest SDF factors (WFR portfolios).
Coefficient estimates and t-statistics at the optimal value of the prior root expected SR^2 (based on cross-validation). Panel (a) focuses on the original WFR portfolios. Panel (b) pre-rotates returns into PC space and shows coefficient estimates corresponding to these PCs. Coefficients are sorted descending on their absolute t-statistic values. The sample is daily from September 1964to December 2017.

(a) Raw WFR portfolios			(b) PCs of WFR portfolios		
	b	t-stat		b	t-stat
Free cash flow/Operating cash flow	3.64	5.49	PC 7	−3.39	6.64
Accruals/Average assets	2.85	4.13	PC 19	−3.69	6.00
P/E (diluted, incl. EI)	−2.46	3.51	PC 6	2.48	5.06
Month $t-9$	1.83	3.03	PC 20	−2.83	4.59
Month $t-11$	1.64	2.71	PC 26	2.78	4.20
Operating CF/Current liabilities	1.89	2.65	PC 10	1.61	2.85
Cash flow/Total debt	1.80	2.48	PC 2	−0.59	2.66
Trailing P/E to growth (PEG) ratio	−1.63	2.47	PC 8	1.38	2.56
P/E (diluted, excl. EI)	−1.72	2.43	PC 5	0.98	2.44
Month $t-1$	−1.40	2.31	PC 36	1.47	2.05
Enterprise value multiple	−1.43	2.11	PC 25	1.31	2.00

CRSP value-weighted index return using βs estimated in the full sample.

4.2.1. Fifty anomaly characteristics

We now turn to our primary data set of 50 portfolios based on anomaly characteristics. The sample is daily from November 1973to December 2017. Fig. 3 presents the OOS R^2 from our dual-penalty specification as a function of κ (on the x-axis) and the number of nonzero SDF coefficients (on the y-axis). A comparison with our earlier Fig. 1 for the FF25 portfolios shows some similarities but also features that are drastically different. Focusing on the left-hand Fig. 3a based on raw returns of the 50 anomaly portfolios, one similarity is that unregularized models (top-right corner) demonstrate extremely poor performance with OOS R^2 substantially below zero. Hence, substantial regularization is needed to get good OOS performance. However, unlike for the FF25 portfolios, there is not much substitutability between L^1 and L^2-regularization here. To attain the maximum OOS R^2, the data calls for substantial L^2-shrinkage but essentially no sparsity. Imposing sparsity (i.e., moving down in the plot) leads to a major deterioration in OOS R^2. This indicates that there is almost no redundancy among the 50 anomalies. The FF25 portfolios have so much redundancy that a small subset of these portfolios is sufficient to span the SDF. In contrast, to adequately capture the pricing information in the 50 anomalies, one needs to include basically all of these 50 factors in the SDF. Shrinking their SDF coefficients is important to obtain good performance, but forcing any of them to zero to get a sparse solution hurts the OOS R^2. In other words, a characteristics-sparse SDF with good pricing performance does not exist. Hence, many anomalies do in fact make substantial marginal contributions to OOS explanatory power of the SDF.

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

Fig. 4 provides a more precise picture of the key properties of OOS R^2 by taking cuts of the contour plots. The solid red line in Fig. 4a represents a cut along the top edge of Fig. 3 with varying degrees of L^2-shrinkage but no sparsity. As the figure shows, the OOS R^2 is maximized for $\kappa \approx 0.30$.

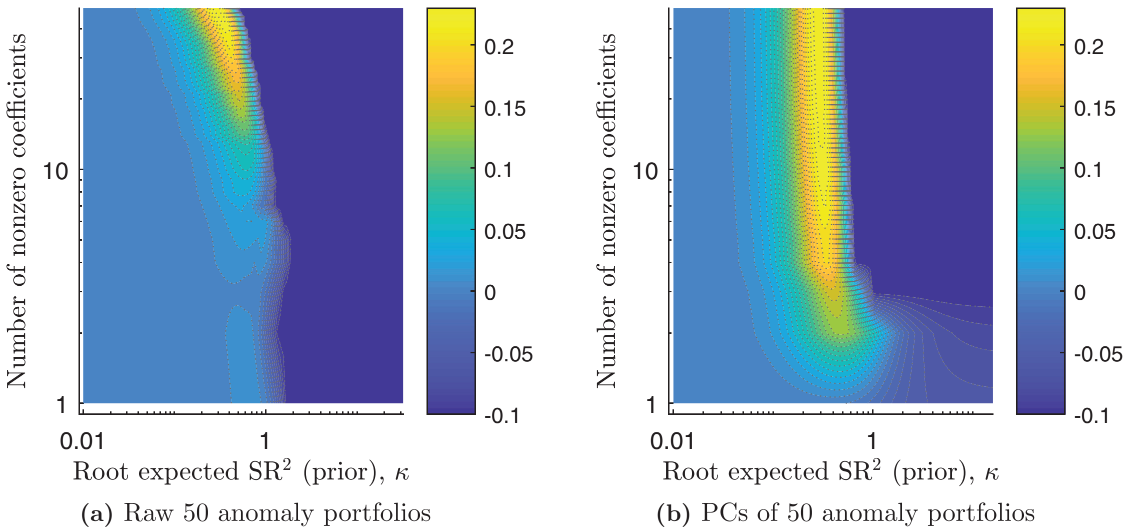


Fig. 3. OOS R^2 from dual-penalty specification (50 anomaly portfolios). OOS cross-sectional R^2 for families of models that employ both L^1 and L^2 penalties simultaneously using 50 anomaly portfolios (Panel a) and 50 PCs based on anomaly portfolios (Panel b). We quantify the strength of the L^2 penalty by prior root expected SR^2 (κ) on the x-axis. We show the number of retained variables in the SDF, which quantifies the strength of the L^1 penalty, on the y-axis. Warmer (yellow) colors depict higher values of OOS R^2. Both axes are plotted on logarithmic scale. The sample is daily from November 1973to December 2017. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

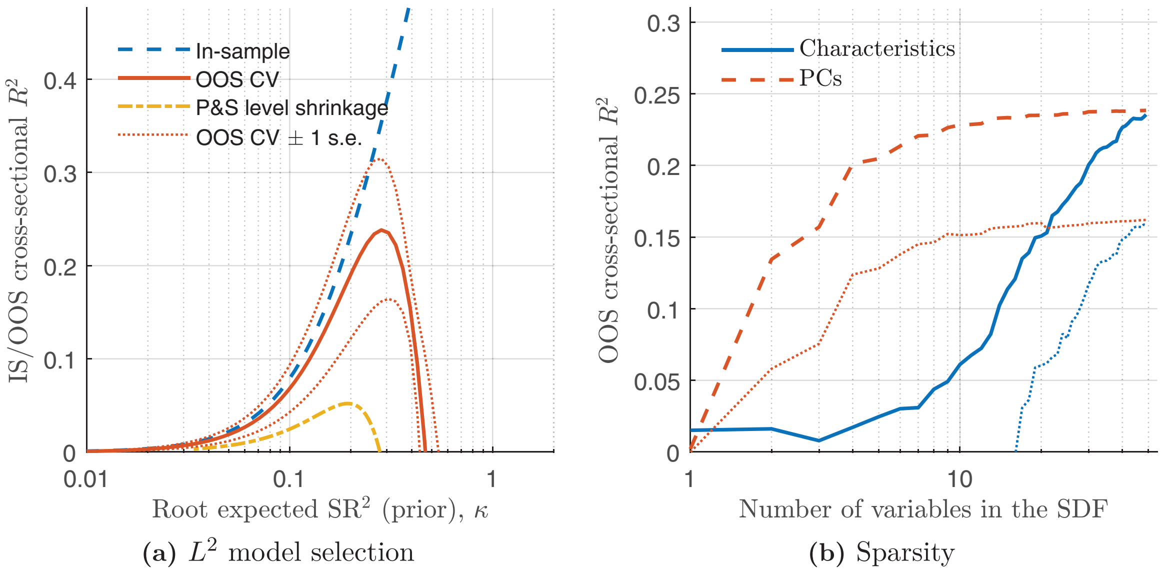


Fig. 4. L^2 model selection and sparsity (50 anomaly portfolios). Panel (a) plots the in-sample cross-sectional R^2 (dashed), OOS cross-sectional R^2 based on cross-validation (solid), and OOS cross-sectional R^2 based on the proportional shrinkage (dash-dot) from Pástor and Stambaugh (2000). In Panel (b), we show the maximum OOS cross-sectional R^2 attained by a model with n factors (on the x-axis) across all possible values of L^2 shrinkage, for models based on original characteristics portfolios (solid) and PCs (dashed). Dotted lines in Panel (b) depict −1 s.e. bounds of the CV estimator. The sample is daily from November 1973to December 2017.

The standard error bounds indicate that OOS R^2 around this value of κ is not only economically but also statistically quite far above zero. Table 1a lists the anomaly factors with the largest absolute t-statistics, where standard errors are based on Eq. (23). The largest coefficients and t-statistics are associated with industry relative-reversals (low vol.), industry momentum-reversals, industry relative-reversals, seasonality, earnings surprises, return on equity (ROE), value-profitability, momentum, etc. Not surprisingly, these are the anomalies that have been found to be among the most robust in the literature. Our method uncovers them naturally. The t-statistics are quite low, but it is important to keep in mind that what matters for the SDF is the joint significance of linear combinations of 50 of these factors. Table 1b shows t-statistics for particular linear combinations: the PCs of the 50 portfolio returns. As the table shows, the loadings on PC1, PC2, PC4, and PC9 are all significantly different from zero at conventional significance levels.[18] Our earlier analysis in Fig. 4b showed that the SDF already achieves a high OOS R^2 with only these four PCs. It is also consistent with our economic arguments in the beginning of the paper that the PCs with the biggest absolute coefficients are PCs with the highest variance.

In Section 3.1, we argued on economic grounds that our prior specification with $\eta = 2$ is reasonable. However, it would be useful to check whether this economic motivation is also accompanied by better performance in the data. To do this, the yellow dash-dot line in Fig. 4a plots the OOS R^2 we would get with the more commonly used prior of Pástor and Stambaugh (2000) with $\eta = 1$.[19] Recall that our method performs both level shrinkage of all coefficients, as well as relative shrinkage (twist) that downweights the influence of small PCs. The method in Pástor and Stambaugh (2000) employs only level shrinkage. We can see that optimally chosen level shrinkage alone achieves OOS R^2 lower than 5% (an improvement over the OLS solution) but falls substantially short of the 30% R^2 delivered by our method. Relative shrinkage, which is the key element of our method, therefore contributes a major fraction of the total out-of-sample performance.

Fig. 4b takes a cut in the contour plots along the ridge of maximal OOS R^2 from bottom to top where we vary sparsity and choose the optimal L^2-shrinkage for each level of sparsity. The solid blue line shows very clearly how characteristics-sparse SDFs perform poorly. The OOS R^2 only starts rising substantially at the lowest sparsity levels toward the very right of the plot. In PC space, on the contrary, very sparse models perform exceedingly well: a model with only two PC-based factors captures roughly two-thirds of the total OOS cross-sectional R^2. A model with ten PC factors achieves nearly maximal R^2, while a model with ten factors in the space of characteristics-based factors achieves less than a third of the maximum. Many of the PC factors that our dual-penalty approach picks in PC-sparse SDF representations are the same as the PCs with highest t-statistics in Table 1. For instance, the first selected factor is PC1, followed by PC4, PC2, and PC9. (see Fig. 3 in Internet Appendix E for more details).

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

4.2.2. WRDS financial ratios (WFR)

The data set of 50 anomalies is special in the sense that all of these characteristics are known, from the past literature, to be related to average returns. Our method is useful to check for redundancy among these anomalies, but this set of asset returns did not expose the method to the challenge of identifying entirely new pricing factors from a high-dimensional data set. For this reason, we now look at 80 characteristics-based factors formed based on the WFR data set. We supplement the data set with 12 portfolios sorted on past monthly returns in months $t-1$ through $t-12$. The sample is daily from September 1964to December 2017. Some of the characteristics in the WFR data set are known to be related to expected returns (e.g., several versions of the P/E ratio), but many others are not. It is therefore possible that a substantial number of these 80 factors are irrelevant for pricing. It will be interesting to see whether our method can: (i) properly de-emphasize these pricing-irrelevant factors and avoid overfitting them; (ii) pick out pricing factors that are similar to those that our analysis of 50 anomalies found relevant; and (iii) potentially find new pricing factors.

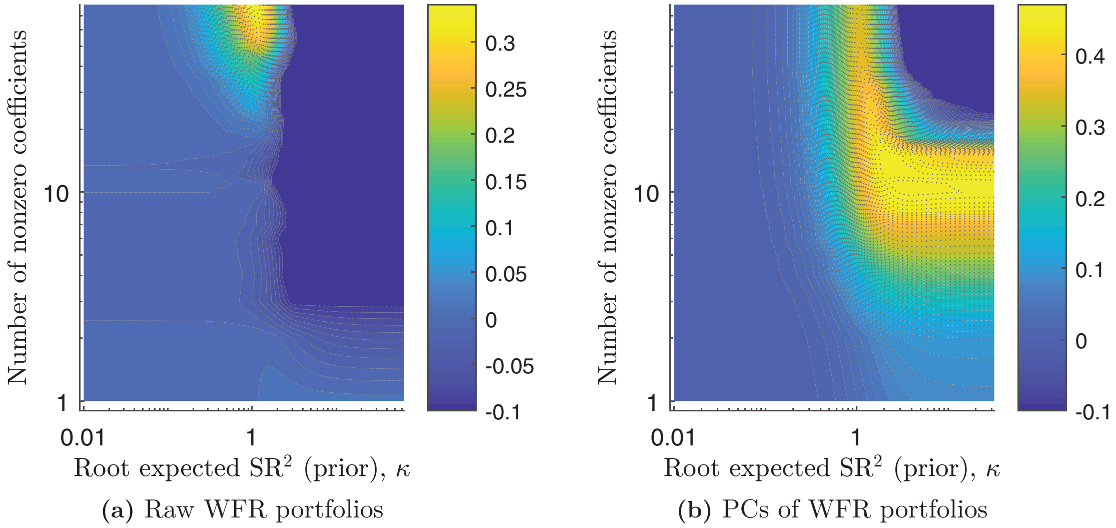
The contour map of OOS R^2 in Fig. 5 looks quite similar to the earlier one for the 50 anomaly portfolios in Fig. 3. Unregularized models (top-right corner) again perform extremely poorly with OOS R^2 significantly below zero. L^2-penalty-only based models (top edge of a plot) perform well for both raw portfolio returns and PCs. L^1-penalty-only "lasso" based models (right-edge of the plot) work poorly for raw portfolio returns in Fig. 5a.

However, there are some differences as well. As can be seen toward the right-edge of Fig. 5b, a PC-sparse SDF not only does quite well in terms of OOS R^2, but it does so even without much L^2-shrinkage. A potential explanation of this finding is that the data mining and publication bias toward in-sample significant factors may play a bigger role in the anomalies data set, which is based on published anomalies, than in the WFR data set. As a consequence, stronger shrinkage of SDF coefficients toward zero may be needed in the anomalies data set to prevent these biases from impairing OOS performance, while there is less need for shrinkage in the WFR data set because in- and out-of-sample returns are not so different.

This explanation is further consistent with the fact that the OOS R^2-maximizing $\kappa \approx 1$, which is much higher than in the anomalies data set. Fig. 6a illustrates this even more transparently by taking a cut along the top edge of Fig. 5a. The solid red line shows the OOS R^2. Its peak is much farther to the right than in the analogous figure for the anomalies data set (Fig. 4a), consistent with our intuition that WFR are less likely to have been datamined in an early part of the sample compared to the published anomalies and therefore do not require as much shrinkage. Standard errors are smaller, too, due to more stable performance of WFR portfolios across time periods relative to anomalies, which experienced significant deterioration

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

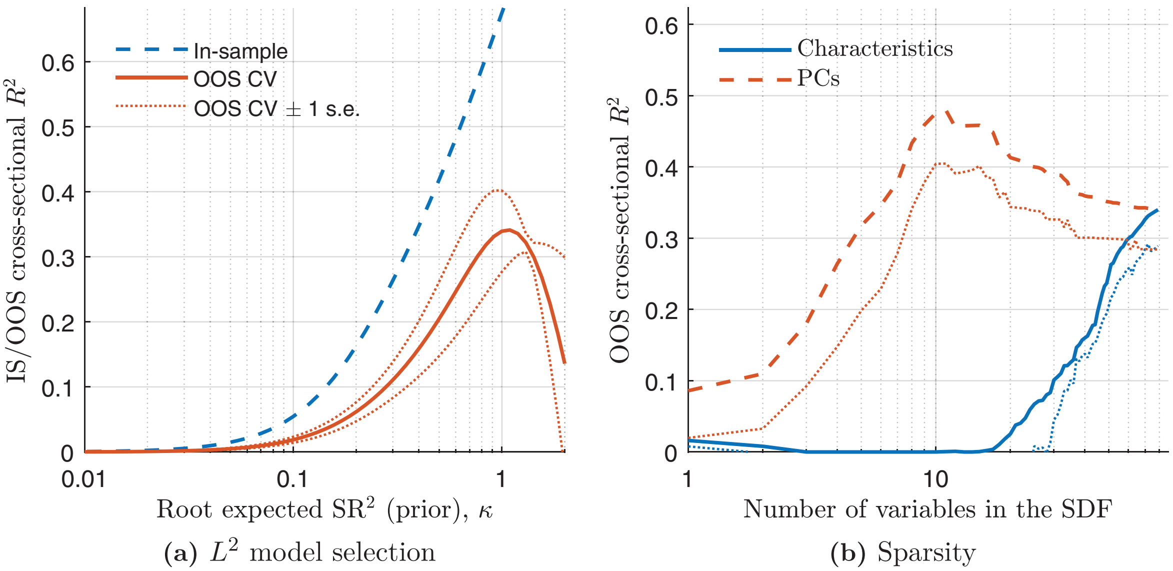
[19] For the Pástor and Stambaugh (2000) level shrinkage estimator, we show $\mathbb{E}(\mathrm{SR}^2)$ under the prior on the x-axis, but it no longer coincides with the κ parameter in this case.



(a) Raw WFR portfolios

(b) PCs of WFR portfolios

Fig. 5. OOS R^2 from dual-penalty specification (WFR portfolios). OOS cross-sectional R^2 for families of models that employ both L^1 and L^2 penalties simultaneously using 80 WFR portfolios (Panel a) and 80 PCs based on WFR portfolios (Panel b). We quantify the strength of the L^2 penalty by prior root expected SR^2 (κ) on the x-axis. We show the number of retained variables in the SDF, which quantifies the strength of the L^1 penalty, on the y-axis. Warmer (yellow) colors depict higher values of OOS R^2. Both axes are plotted on logarithmic scale. The sample is daily from September 1964to December 2017. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) L^2 model selection

(b) Sparsity

Fig. 6. L^2 model selection and sparsity (WFR portfolios). Panel (a) plots the in-sample cross-sectional R^2 (dashed) and OOS cross-sectional R^2 based on cross-validation (solid). In Panel (b), we show the maximum OOS cross-sectional R^2 attained by a model with n factors (on the x-axis) across all possible values of the prior root expected SR^2 (κ) for models based on original characteristics portfolios (solid) and PCs (dashed). Dotted lines in Panel (b) depict -1 s.e. bounds of the CV estimator. The sample is daily from September 1964to December 2017.

in the latest (not datamined) part of the sample (McLean and Pontiff, 2016).

Table 2 lists coefficient estimates at this optimal level of L^2-only penalty. Coefficients are sorted descending on their absolute t-statistic values. Table 2a focuses on original WFR portfolio returns. It shows that our method tends to estimate high weights on factors based on characteristics known to be associated with expected returns. Among the picks there are few measures of valuation ratios (price/earnings (PE), PE/G (PEG)), investment (free CF/operating CF, which equals 1 - capital expenditure/operating CF), accruals (accruals/average assets), financial soundness (operating CF/current liabilities, CF/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 measures (averaging them by the means of the L^2 penalty) and delivers a robust SDF that performs well out of sample. Combining several measures of each signal (e.g., valuation measures) performs much better out of sample than using any single ratio.

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

Fig. 6b takes a cut in the contour plots along the ridge of maximal OOS R^2 from bottom to top where we vary sparsity and choose the optimal shrinkage for each level of sparsity. This figure illustrates that like in the case of the 50 anomalies, there is little sparsity in the space of characteristics. Even so, sparsity is again much stronger in PC space. A model with six factors delivers nearly maximum OOS R^2.

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

4.3. Interactions

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

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

We construct the nonlinear weights and interactions as follows. For any two given rank-transformed characteristics $z^i_{s,t}$ 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 renormalized using Eq. (32) as follows:

$$z^{ij}_{s,t} = \frac{\left(z^i_{s,t} z^j_{s,t} - \frac{1}{n_t}\sum_{s=1}^{n_t} z^i_{s,t} z^j_{s,t}\right)}{\sum_{s=1}^{n_t}\left| z^i_{s,t} z^j_{s,t} - \frac{1}{n_t}\sum_{s=1}^{n_t} z^i_{s,t} z^j_{s,t}\right|}. \tag{33}$$

We include all first-order interactions in our empirical tests. In addition to interactions, we also include second and third powers of each characteristic, which are defined analogously based on interaction of the characteristic with itself. Note that although we renormalize all characteristics after interacting or raising to powers, we do not rerank 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 but has the same gross exposure (leverage). We illustrate how this approach maps into more conventional two-way portfolio sorts portfolios in Internet Appendix C.

Due to the extremely high number of characteristics-based factors in this case, our three-fold cross-validation method runs into numerical instability issues in covariance matrix inversion, even with daily data. For this reason, we switch to two-fold cross-validation. This gives us a somewhat longer sample to estimate the covariance matrix, and this sample extension is sufficient to obtain stable behavior.[21]

Fig. 7 shows contour maps of the OOS cross-sectional R^2 as a function of κ (on the x-axis) and the number of nonzero SDF coefficients (on the y-axis). Plots for the raw portfolio returns are shown in the top row, and plots for the PCs are in the bottom row. Focusing first on the results for the raw portfolio returns, it is apparent that a substantial degree of sparsity is now possible for both the anomalies and the WFR portfolios without deterioration in the OOS R^2. Strengthening the L^1-penalty to the point that only around 100 of the characteristics and their powers and interactions remain in the SDF (out of 1375 and 3400, respectively) does not reduce the OOS R^2 as long as one picks the L^2-penalty optimal for this level of sparsity. As before, an L^1-penalty-only approach leads to poor OOS performance.

The plots in the bottom row show contour maps for PCs. These results are drastically different from the ones in the top row in terms of how much sparsity can be im-

[20] Table 3 in Internet Appendix E shows the variances of the PCs.

[21] Because some interactions are missing in the earlier part of the sample, our sample periods shorten to February 1974– December 2017and September 1968– December 2017for anomaly and WFR characteristics, respectively.

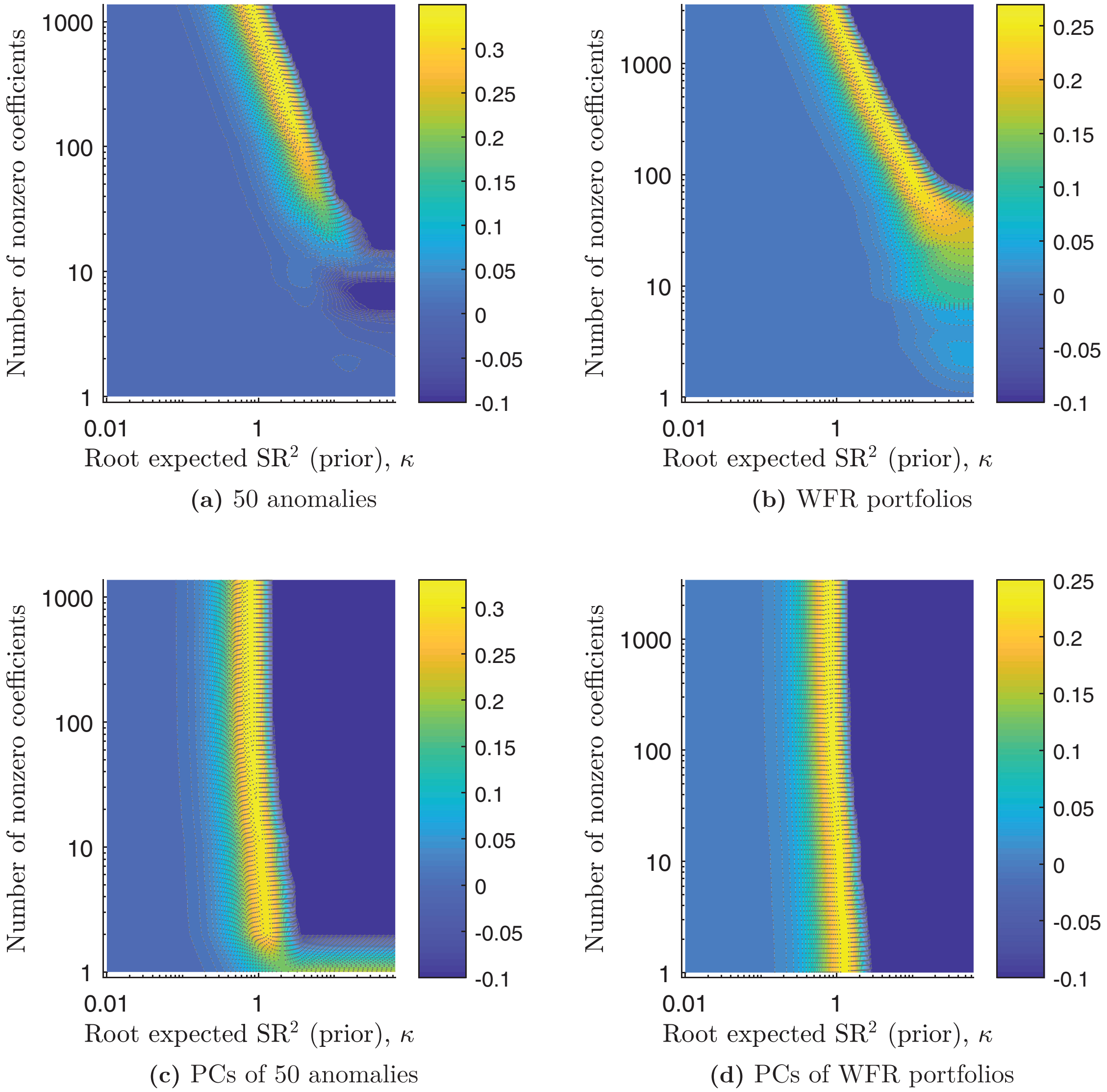


Fig. 7. OOS R^2 from dual-penalty specification for models with interactions. OOS cross-sectional R^2 for families of models that employ both L^1 and L^2 penalties simultaneously using portfolio returns based on interactions of 50 anomaly (Panel a) and 80 WFR (Panel b) characteristics and PCs of these portfolio returns (Panels c and d). We quantify the strength of the L^2 penalty by prior root expected SR^2 (κ) on the x-axis. We show the number of retained variables in the SDF, which quantifies the strength of the L^1 penalty, on the y-axis. Warmer (yellow) colors depict higher values of OOS R^2. Both axes are plotted on logarithmic scale. The sample is daily from February 1974to December 2017and September 1968to December 2017for anomaly and WFR characteristics, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

posed without hurting OOS performance. Very few PCs—or even just one—suffice to obtain substantial OOS explanatory power. But here, too, the combination of sparsity with an optimally chosen L^2 penalty is very important. Adding more PCs does not hurt as long as substantial L^2 shrinkage is imposed, but it does not improve OOS performance much either.

Table 3 lists coefficient estimates at the optimal level of L^2 regularization (i.e., the maximum along the top edge of the contour plots). Table 3a focuses on the SDF constructed from PCs of portfolio returns based on interactions of 50 anomaly characteristics. Table 3b shows coefficient estimates corresponding to PCs of portfolio returns based on interactions of WRDS financial ratios (WFR). PC1 has the highest t-statistic for both sets of portfolios. PC1 is also the last survivor if one imposes enough sparsity that only one PC remains. The estimated SDF coefficients are quite similar for many of the other PCs in this table that are ranked lower than PC1 in terms of their t-statistic. However, since these other PCs have lower variance, their contribution to SDF variance, and hence the overall squared Sharpe ratio captured by the SDF, is lower.

The two plots in Fig. 8 take a cut in the contour plots along the ridge of maximal OOS R^2 from bottom to top where we vary sparsity and choose the L^2 optimal shrinkage for each level of sparsity. These plots reinforce

Table 3
Largest SDF factors (models with interactions).
Coefficient estimates and t-statistics at the optimal value of the prior root expected SR^2 (based on cross-validation). Panel (a) focuses on the SDF constructed from PCs portfolio returns based on interactions of 50 anomaly characteristics. Panel (b) shows coefficient estimates corresponding to PCs of portfolio returns based on interactions of WFR. Coefficients are sorted descending on their absolute t-statistic values. The sample is daily from February 1974to December 2017and September 1968to December 2017for anomaly and WFR characteristics, respectively.

(a) PCs of interactions of anomaly portfolios			(b) PC of interactions of WFR portfolios		
	b	t-stat		b	t-stat
PC 1	−0.24	3.82	PC 1	−0.11	2.89
PC 2	0.27	3.22	PC 5	−0.14	1.97
PC 18	0.24	1.95	PC 2	−0.08	1.49
PC 17	0.24	1.95	PC 21	−0.13	1.46
PC 19	−0.23	1.87	PC 6	0.11	1.39
PC 41	0.23	1.79	PC 50	0.11	1.27
PC 34	0.22	1.78	PC 84	−0.09	1.02
PC 26	−0.20	1.59	PC 7	−0.07	0.97
PC 60	0.20	1.57	PC105	0.08	0.87
PC 10	−0.18	1.57	PC108	−0.08	0.86
PC 56	−0.20	1.55	PC114	−0.08	0.85

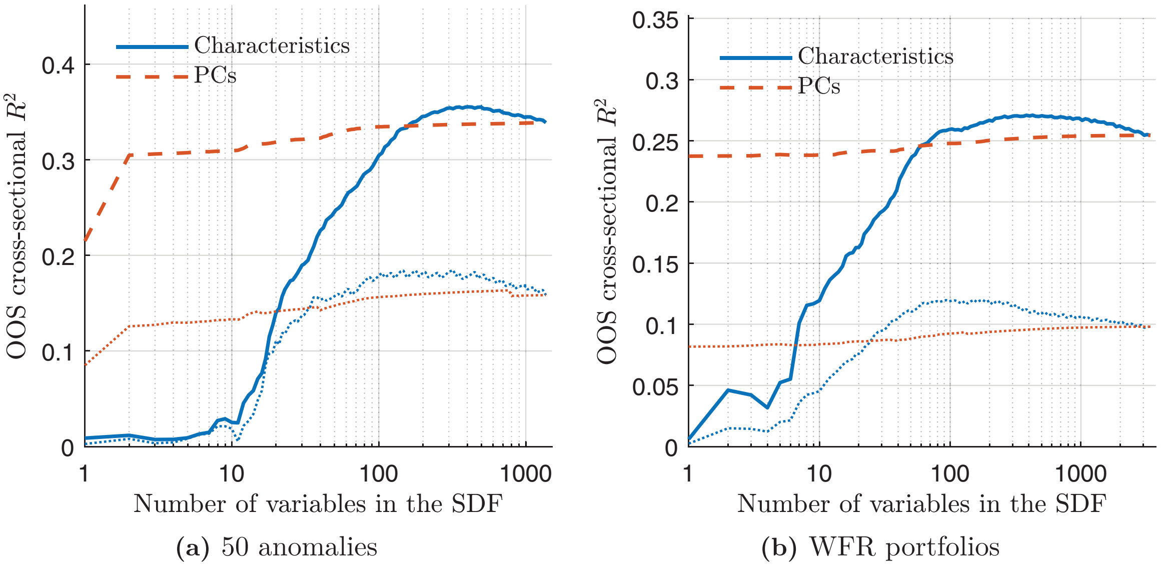


Fig. 8. L^1 sparsity of models with interactions. We show the maximum OOS cross-sectional R^2 attained by a model with n factors (on the x-axis) across all possible values of the prior root expected SR^2 (κ) for models based on interactions of original characteristics portfolios (solid) and PCs (dashed).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

the point we noted from the contour plots that many of the powers and interactions of the characteristics are not adding pricing-relevant information to the SDF and can be omitted. The SDF that attains the highest OOS R^2 is relatively sparse with about 100 factors for both the anomalies in Fig. 8a and the WFR portfolios in Fig. 8b. However, as the wide standard error bands show, statistical precision is quite low. The very large number of portfolios in this case pushes the method to its statistical limits.

Overall, these results show that many of the powers and interactions of characteristics seem to be redundant in terms of their pricing implications. A majority of them can be excluded from the SDF without adverse impact on OOS pricing performance. But as before, L^2-shrinkage is crucial for obtaining good OOS performance.

5. Asset pricing tests: performance compared with sparse models

Our cross-validation method evaluates a model's performance on the part of a sample not used in the estimation of the SDF coefficients; it is, therefore, by construction an OOS metric. Yet our choice of the strength of regularization (L^1 and L^2 penalties) is based on the entire sample. It is possible that the penalty that is optimal within one sample does not generalize well on new or fully withheld data. To address this potential issue, we now conduct a pure OOS test. Using our L^2-penalty method, we conduct the entire estimation, including the choice of penalty, based on data until the end of 2004. Post-2004 data is completely left out of the estimation. We evaluate

Table 4
MVE portfolio's annualized OOS α in the withheld sample (2005–2017), %.
The table shows annualized alphas (in %) computed from the time-series regression of the SDF-implied OOS-MVE portfolio's returns (based on L^2 shrinkage only) relative to four restricted benchmarks: CAPM, Fama-French six-factor model, optimal sparse model with five factors, and optimal PC-sparse model with at most five PC-based factors. MVE portfolio returns are normalized to have the same standard deviation as the aggregate market. Standard errors are in parentheses.

SDF factors/Benchmark	CAPM	FF 6-factor	Char.-sparse	PC-sparse
50 anomaly portfolios	12.35	8.71	9.55	4.60
	(5.26)	(4.94)	(3.95)	(2.22)
80 WFR portfolios	20.05	19.77	17.08	3.63
	(5.26)	(5.29)	(5.05)	(2.93)
1375 interactions of	25.00	22.79	21.68	12.41
anomalies	(5.26)	(5.18)	(5.03)	(3.26)

performance of this SDF in the 2005–2017 OOS period. This analysis also allows us to assess the statistical significance of our earlier claim that characteristics-sparse SDFs cannot adequately describe the cross-section of stock returns.

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

Our analysis is very much in the spirit of Barillas and Shanken (2018) in that we compare the Sharpe ratios of the MVE portfolios implied by competing factor models (rather than the alphas of some "test assets"), albeit with an OOS focus. We proceed as follows. We first orthogonalize all managed portfolio returns with respect to the market using βs estimated in the pre-2005 sample.[22] Given the estimate $\hat{b}$ based on our L^2-penalty Bayesian method in this sample, we construct the time series of the implied MVE portfolio $P_t = \hat{b}'F_t$ in the 2005–2017 OOS period. We focus on three sets of portfolios in constructing an SDF: the 50 anomaly portfolios, the 80 WFR portfolios, and the interactions and powers of 50 anomaly characteristics.[23] As in our earlier estimation, we choose penalties by three-fold cross-validation (two-fold if interactions are included) but with shorter blocks because we only use the pre-2005 sample here.[24]

We then estimate abnormal returns of this OOS-MVE portfolio with respect to three characteristics-based benchmarks: the capital asset pricing model (CAPM); the six-factor model of Fama and French (2016) (with five cross-sectional factors, including the momentum factor); and our dual-penalty model where we have set the L^1 penalty such that the SDF contains only five cross-sectional characteristics-based factors. To compare the models on equal footing, we construct the MVE portfolio implied by these benchmarks. Since we work with candidate factor returns orthogonalized to the market return, the benchmark in the CAPM case is simply a mean return of zero. For Fama-French six-factor model, we estimate the unregularized MVE portfolio weights, $\hat{w} = \hat{\Sigma}^{-1}\hat{\mu}$, from the five nonmarket factors in the pre-2005 period.[25] We then apply these weights to the five factor returns in the OOS period to construct a single benchmark return. Finally, for the dual-penalty sparse model with five factors, we estimate $\hat{b}$ in the pre-2005 period and then apply these optimal portfolio weights to returns in the OOS period. If our earlier claim is correct that the SDF cannot be summarized by a small number of characteristics-based factors, then our OOS-MVE portfolio constructed from the full set of candidate factors should generate abnormal returns relative to the MVE portfolio constructed from these sparse benchmarks.

Table 4 confirms that the MVE portfolio implied by our SDF performs well in the withheld data. The table presents the intercepts (alphas) from time-series regressions of the OOS-MVE portfolio returns on the benchmark portfolio return in %, annualized, with standard errors in parentheses. To facilitate interpretation of magnitudes, we scale MVE portfolio returns so that they have the same standard deviation as the market index return in the OOS period. The first column shows that the OOS-MVE portfolio offers a large abnormal return relative to the CAPM for all three sets of candidate factor returns. For example, for the OOS-MVE portfolio based on the 50 anomalies, we estimate an abnormal return of 12.35%, which is more than two standard errors from zero, despite the short length of the evaluation sample. The abnormal returns are even larger for the other two sets of portfolios. As the second column shows, the abnormal returns are very similar in magnitude for the FF six-factor model, and we

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

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

[24] We plot the time series of returns of the MVE portfolios in Fig. 5 in the Internet Appendix.

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

can reject the hypothesis of zero abnormal returns at a 5% level or less for two of the three sets of candidate factor portfolios. The third column shows that the results for the sparse six-factor model based on our dual-penalty method is almost identical to the FF six-factor model. Overall, the evidence in this table confirms our claim that characteristics-sparse models do not adequately describe the cross-section of expected stock returns.

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

6. Conclusion

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

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

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

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

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

References

Asness, C.S., Frazzini, A., Pedersen, L.H., 2014. Quality minus junk. AQR Capital Management and Copenhagen Business School. Working paper

Asness, C.S., Moskowitz, T.J., Pedersen, L.H., 2013. Value and momentum everywhere. J. Financ. 68 (3), 929–985.

Barillas, F., Shanken, J., 2018. Comparing asset pricing models. J. Financ. 73 (2), 715–754.

Brandt, M.W., Santa-Clara, P., Valkanov, R., 2009. Parametric portfolio policies: exploiting characteristics in the cross-section of equity returns. Rev. Financ. Stud. 22 (9), 3411–3447.

Cochrane, J.H., 2011. Presidential address: discount rates. J. Financ. 66 (4), 1047–1108.

DeMiguel, V., Garlappi, L., Nogales, F.J., Uppal, R., 2009. A generalized approach to portfolio optimization: improving performance by constraining portfolio norms. Manag. Sci. 55 (5), 798–812.

DeMiguel, V., Martin-Utrera, A., Nogales, F.J., Uppal, R., 2019. A transaction-cost perspective on the multitude of firm characteristics. Rev. Financ. Stud. ISSN 0893-9454 (In Press).

Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. J. Financ. Econ. 33, 23–49.

Fama, E.F., French, K.R., 2008. Dissecting anomalies. J. Financ. 63 (4), 1653–1678.

Fama, E.F., French, K.R., 2015. A five-factor asset pricing model. J. Financ. Econ. 116 (1), 1–22.

Fama, E.F., French, K.R., 2016. Dissecting anomalies with a five-factor model. Rev. Financ. Stud. 29 (1), 69–103.

Fan, J., Liao, Y., Wang, W., 2016. Projected principal component analysis in factor models. Ann. Stat. 44 (1), 219.

Feng, G., Giglio, S., Xiu, D., 2017. Taming the factor zoo. University of Chicago. Working paper

Freyberger, J., Neuhierl, A., Weber, M., 2017. Dissecting characteristics nonparametrically. Technical Report.

Green, J., Hand, J.R., Zhang, X.F., 2017. The characteristics that provide independent information about average US monthly stock returns. Rev. Financ. Stud. hhx019.

Hansen, L.P., Jagannathan, R., 1991. Implications of security market data for models of dynamic economies. J. Polit. Econ. 99, 225–262.

Hansen, L.P., Jagannathan, R., 1997. Assessing specification errors in stochastic discount factor models. J. Financ. 52, 557–590.

Harvey, C.R., Liechty, J.C., Liechty, M.W., 2008. Bayes vs. resampling: a rematch. J. Invest. Manag. 6, 29–45.

Harvey, C.R., Liu, Y., Zhu, H., 2015. ... and the cross-section of expected returns. Rev. Financ. Stud. 29 (1), 5–68.

Hastie, T.J., Tibshirani, R.J., Friedman, J.H., 2011. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York.

Hou, K., Xue, C., Zhang, L., 2015. Digesting anomalies: an investment approach. Rev. Financ. Stud. 28 (3), 650–705.

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

Kelly, B.T., Pruitt, S., Su, Y., 2018. Characteristics are covariances: a unified model of risk and return. National Bureau of Economic Research. Working paper

Kogan, L., Tian, M., 2015. Firm characteristics and empirical factor models: a model-mining experiment. MIT. Working paper

Kozak, S., Nagel, S., Santosh, S., 2018. Interpreting factor models. J. Financ. 73 (3), 1183–1223.

Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. . Multivar. Anal. 88 (2), 365–411.

Lewellen, J., Nagel, S., Shanken, J., 2010. A skeptical appraisal of asset-pricing tests. J. Financ. Econ. 96, 175–194.

Lin, X., Zhang, L., 2013. The investment manifesto. J. Monet. Econ. 60, 351–366.

MacKinlay, A.C., 1995. Multifactor models do not explain deviations from the CAPM. J. Financ. Econ. 38 (1), 3–28.

McLean, D.R., Pontiff, J., 2016. Does academic research destroy stock return predictability? J. Financ. 71 (1), 5–32.

Moritz, B., Zimmermann, T., 2016. Tree-based conditional portfolio sorts: the relation between past and future stock returns. Technical report.

Novy-Marx, R., Velikov, M., 2016. A taxonomy of anomalies and their trading costs. Rev. Financ. Stud. 29 (1), 104–147.

Pástor, L., 2000. Portfolio selection and asset pricing models. J. Financ. 55 (1), 179–223.

Pástor, L., Stambaugh, R.F., 2000. Comparing asset pricing models: an investment perspective. J. Financ. Econ. 56 (3), 335–381.

Rapach, D.E., Strauss, J.K., Zhou, G., 2013. International stock return predictability: what is the role of the United States? J. Financ. 68 (4), 1633–1662.

Ross, S.A., 1976. The arbitrage theory of capital asset pricing. J. Econ. Theory 13, 341–360.

Stambaugh, R.F., Yuan, Y., 2016. Mispricing factors. Rev. Financ. Stud. 30 (4), 1270–1315.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B (Methodol.) 58, 267–288.

Tibshirani, R.J., Tibshirani, R., 2009. A bias correction for the minimum error rate in cross-validation. Ann. Appl. Stat. 3, 822–829.

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

Varma, S., Simon, R., 2006. Bias in error estimation when using cross-validation for model selection. BMC Bioinformat. 7 (91).

Vuolteenaho, T., 2002. What drives firm-level stock returns? J. Financ. 52, 233–264.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B (Stat. Methodol.) 67 (2), 301–320.