Shrinking the Cross Section*

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

February 2, 2017

PRELIMINARY AND INCOMPLETE

[PRINT IN COLOR]

Abstract

We propose a new method of tackling the "multi-dimensionality challenge" in the cross section of equity returns. Our approach relies on exploiting economically-driven regularization to construct a robust stochastic discount factor (SDF) using individual stock returns and a vast array of characteristics. We impose penalties on estimated SDF coefficients in L_2 and L_1 norms, similar to the elastic nets technique in machine learning. The penalties are motivated by the need to down-weight contributions of small principal components to the total squared Sharpe ratio and sparsity of SDFs implied by most economic models, respectively. Our economically-motivated estimator delivers robust, sparse SDF representations that perform well out of sample.

*We thank Mike Chernov and seminar participants at Michigan for helpful comments and suggestions.

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

[‡]Stephen M. Ross School of Business and Department of Economics, University of Michigan, 701 Tappan St., Ann Arbor, MI 48109, e-mail: stenagel@umich.edu.

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

1 Introduction

Studies on cross-sectional stock return predictability have found relationships between expected returns and hundreds of stock characteristics. An economic interpretation of the collection of these findings is difficult for a number of reasons. First, how many of these characteristics capture pricing information that is of marginal value over and above all the other characteristics? Empirical studies typically check whether a characteristics-factor offers alpha relative to a few popular small-scale factor models, but this does not address the concern that many of the discovered predictability relationships could be redundant. Second, do the characteristics interact in their relationship to expected returns? Existing studies have explored a very small number of interactions – often with size or value – but a very large number of potential interactions remains unexplored. Third, could additional predictors, beyond the many that have already been explored, contain useful information about the cross-section of expected returns?

Due to the high-dimensional nature of the problem, these questions cannot be addressed with conventional statistical methods. We use tools from machine learning to do so. However, rather relying on a purely statistical motivation for the regularization that we impose on the problem, we develop the regularization from economic restrictions.

Our starting point is that we seek to summarize the investment opportunities offered by a large cross-section of stocks in a stochastic discount factor (SDF). By looking for an SDF representation, we are seeking factors that help price the cross-section. In this way, the SDF approach automatically leads towards exclusion of factors that are redundant for pricing because their expected return premium fully derives from covariance with other factors in the SDF. We then recognize that first and second moments of returns should be linked. As Kozak et al. (2015) argue, absence of near-arbitrage opportunities is plausible in "behavioral" as well as "rational" models of asset prices. As a consequence, it should not be possible to earn large expected return premia without commensurate common factor risk exposures. Moreover, much of the limited Sharpe Ratio that can be earned in the cross-section should come from dominant volatile factors rather than obscure small-variance factors.

The restrictions we impose amount to penalizing both the L_1 and L_2 norms of SDF coefficients, similar to the elastic nets technique in machine learning.[1] The L_2 penalty alone is similar to ridge regression. It is motivated by our prior work (Kozak et al., 2015) as a restriction on the total size of "sentiment" belief distortions in the context of a model in that paper. Alternatively, and more generally, one can interpret this penalty as a restriction of the total size of arbitrageurs' cross-sectional deviations from the market portfolio weights. The

[1] See Hastie et al. (2011) for a textbook treatment of ridge regression, lasso regression, and elastic nets.

model in Kozak et al. (2015) implies that much of the variance of the SDF (or maximum squared Sharpe ratio) should come from high eigenvalue principal components (PCs) of returns. Based on this economic reasoning, we focus on penalizing high contributions from low eigenvalue PCs. We show that our L_2 penalty amounts exactly to such a down-weighting of contributions from small PCs. We further argue that the L_1 penalty, which is similar to lasso regression, can also be motivated by "sentiment" belief distortions or by many other economic models which predict a sparse SDF representation (e.g., CAPM). It is then natural to impose a penalty in a norm that leads to a sparse solution.

One key difference of our approach from others that penalize Fama-MacBeth regression coefficients (prices of risk in a "beta" representation) is that there is an economic reason for low eigenvalue PCs to have negligible contribution to SDF variance (and to not help price assets). In contrast, regression coefficients in Fama-MacBeth regressions are factor mimicking portfolio mean returns, which can be non-zero even if the factors don't help price assets. Thus, it is much more natural to impose a penalty on SDF coefficients (risk prices) rather Fama-MacBeth regression coefficients (risk premia).

Contrary to classical approaches in the literature, our method does not require the intermediate steps of modeling expected returns, construction of ad hoc factors (long-short strategies of univariate portfolio sorts), or verifying that expected returns line us with factor covariances in the cross-section. All of the economic content behind these steps is directly embedded into the method and automatically imposed during estimation. The essence of the method lies in using economic theory to effectively combine all *individual* stock returns into an out-of-sample mean-variance-efficient (OOS-MVE) portfolio that satisfies the aforementioned economic constraints.

We rely on a vast cross-section of stock characteristics in our analysis, yet we do not require sorting stocks into a small number of portfolios in the classical sense. Rather, we use a cross-section of characteristics to rotate the space of individual stocks into a space of "managed portfolios" which allow for non-linearities and interactions.[2]. Our method is able to efficiently incorporate many *thousands* of such derived characteristics to produce an SDF which prices the cross-section well out of sample (does not overfit) and delivers high Sharpe ratios based on robust predictors. The final SDF representation allows us to examine which factors, interactions, and non-linearities are most important and most robust out-of-sample.

Characteristics that our method uses may, but need not be directly associated with typical characteristics that underlie well-known asset pricing "anomalies". Even if some characteristic is not priced unconditionally in the cross-section, it may show up in the SDF due to its interactions with other cross-sectional characteristics or time-varying instruments.

[2]See Cochrane (1991).

For instance, in one of our tests, we include Fama and French industry classification dummies in the set of characteristics. We find significant interactions of known anomalies with industry dummies.

Unlike classical methods that require us to be particularly mindful about the number and identity of characteristics used to test a model (to avoid over-fitting), our method tends to be more powerful as we expand the set of potential predictors. This eliminates the need to pre-screen and "fish" for factors and lets the data "speak". In fact, our "out-of-sample" tests are unbiased only if a set of characteristics that we use was not data-mined in the validation part of our sample ("anomalies" potentially are). To address this issue, we use post-2005 period in our out-of-sample tests. We include only anomaly characteristics that have been discovered before that date.

Our out-of-sample analysis demonstrates that the method successfully avoids in-sample over-fitting and generates high Sharpe ratios which are robust out of sample. These Sharpe ratios are delivered by efficiently combining known anomalies, as well as uncovering some interesting and unknown interactions of anomalies with other anomalies or other individually unpriced characteristics.[3] For example, we find that more aggressive "cubic" strategies are important for many anomalies, and some interesting interactions such as value-momentum, size-momentum, momentum-beta-arbitrage naturally appear in our analysis.

Our approach is closely related to several powerful regularization techniques used in machine learning: ridge regression, lasso regression, and elastic nets. It differs in several important ways. First, our penalties are economically motivated and derived from economic principles: we penalize contributions of the smallest PCs to the total maximum squared Sharpe ratio the most. Second, unlike in ridge/lasso regressions, we need not normalize and center all variables: our predictions are much sharper – the pricing equation must hold without an intercept; expected returns should be proportional to covariances. Third, our objective is to maximize the squared Sharpe ratio (minimize the distance to the mean-variance frontier) rather than the prediction error as in the purely statistical techniques used in machine learning.

[3] As Harvey et al. (2015) note "it is possible that a particular factor is very important in certain economic environments and not important in other environments. The unconditional test might conclude the factor is marginal."

2 Methodology

For any point in time t, let R_t denote an $N \times 1$ vector of excess returns, and Z_{t-1} an $N \times H$ matrix of asset characteristics (with H possibly quite large, potentially thousands of characteristics). Let Z_{t-1} be centered and standardized cross-sectionally at each t.[4]

2.1 Rotation

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

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

This rotation is motivated by an implicit assumption that underlies everything we do: expected returns, variances, and covariances are stable functions of characteristics such as size and book-to-market ratio, and not security names (Cochrane, 2011). This (implicit) assumption was the driving force for using portfolio sorts in cross-sectional asset pricing in the first place. Managed portfolios allow us to generalize this idea and be more flexible.

2.2 SDF

We are looking for an SDF that's in the linear span of the H (basis) trading strategy returns F_t that can be created based on stock characteristics, i.e.,

$$M_t = 1 - b' \left(F_t - \mathbb{E} F_t \right), \tag{2}$$

and let $\Sigma \equiv \operatorname{cov}(F_t) = \mathbb{E}\left[\left(F_t - \mathbb{E} F_t \right) \left(F_t - \mathbb{E} F_t \right)' \right]$. In the equation (2) above, we assumed that all SDF coefficients b are constant. It is without loss of generality, because we can always re-state a model with time-varying b_{t-1} as a model with constant b and an extended set of factors F_t. For instance, suppose we can capture time variation in b_t by some set of time-series instruments z_{t-1}, i.e., $b_{t-1} = z_{t-1} b$ (Cochrane, 2005, Ch. 8). Then we can simply rewrite the SDF as $M_t = 1 - b' \left(\tilde{F}_t - \mathbb{E} \tilde{F}_t \right)$, where $\tilde{F}_t = z'_{t-1} \left(F_t - \mathbb{E} F_t \right)$.

Given the asset pricing equation,

$$\mathbb{E}\left[M_t F_t \right] = 0, \tag{3}$$

[4] This means each column of Z_{t-1} is mean zero and L^2 norm equals N.

and ignoring the SDF variance constraint (for now), in population we could solve for

$$b = \Sigma^{-1}\mathbb{E}\left(F_t\right), \tag{4}$$

a projection coefficient in a (cross-sectional) projection with H "explanatory variables" and H "dependent variables."

2.2.1 Sample estimators

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

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

be the sample estimates of means and covariances, respectively.

2.3 Approximate SDF: L_2 penalty

In sample, just replacing population moments with sample moments in Eq. 4 would not work well as we would not get invertibility (or at least, non-robust results). Instead, we focus on finding an approximate SDF with good out-of-sample properties. Our aim therefore is to find an out-of sample mean-variance efficient ("OOS-MVE") portfolio that is "close" to the in-sample mean-variance frontier and is robust out of sample. We achieve this goal by minimizing the HJ-distance (Hansen and Jagannathan, 1997) between the in-sample ex-post SDF that prices all assets and our target "robust" SDF.

More specifically, let $\tilde{M}_t = 1 - \mu_T'\Sigma_T^{-1}F_t$ be the ex-post SDF that prices all managed portfolios and let $M_t = 1 - b'F_t$ be the SDF we are trying to find. We use the measure of model of misspecification from Hansen and Jagannathan (1997):

$$d\left(\tilde{M}_t, M_t\right) = \left[\mathbb{E}\left(\tilde{M}_t - M_t\right)^2\right]^{\frac{1}{2}}.$$

Next, we minimize the in-sample HJ-distance between two SDFs by using the sample estimates of required moments,

$$\min_b \hat{d}\left(\tilde{M}_t, M_t\right) = \min_b \left(F_t(\Sigma_T^{-1}\mu_T - b)\right)'\left(F_t(\Sigma_T^{-1}\mu_T - b)\right).$$

We further re-scale all excess returns F_t to have the same standard deviations and impose

a penalty on the size of b into the objective function, $b'b$, which we will argue amounts to down-weighting contributions of small PCs to the total squared Sharpe ratio. This penalty is motivated by our prior work (Kozak et al., 2015) as a restriction on the total size of "sentiment" belief distortions. More generally, one can interpret this penalty as a restriction of the total size of arbitrageurs' cross-sectional deviations from the market portfolio weights. The model in Kozak et al. (2015) implies that much of the maximum squared SR should come from high eigenvalue PCs, not low eigenvalue PCs. Based on this economic reasoning, what we focus on penalizing high contributions to the maximum SR from low eigenvalue PCs.

Combing the objective and the penalty leads to the following problem:

$$\hat{b} = \arg\min_{b} \left\{ (\mu_T - \Sigma_T b)' \Sigma_T^{-1} (\mu_T - \Sigma_T b) + \lambda b'b \right\}, \tag{5}$$

where λ is the Lagrange penalty parameter.

The FOC yields the solution:

$$\hat{b} = (\Sigma_T + \lambda I)^{-1} \mu_T. \tag{6}$$

2.3.1 Interpretation

We argued that the penalty on $b'b$ leads to the shrinkage of contributions of small PCs to the total squared Sharpe ratio. We can see this if we let Q be the matrix of eigenvectors of Σ_T and D the diagonal matrix of eigenvalues,

$$\begin{aligned} \text{var}(M_t) = \max \text{SR}^2 = \hat{b}' \Sigma_T \hat{b} &= \mu_T' (\Sigma_T + \lambda I)^{-1} \Sigma_T (\Sigma_T + \lambda I)^{-1} \mu_T \\ &= \mu_T' Q D (D + \lambda I)^{-2} Q' \mu_T = \sum_{j=1}^{H} \frac{\left(q_j' \mu_T\right)^2}{d_j} \left(\frac{d_j}{d_j + \lambda} \right)^2, \end{aligned}$$

where d_j are diagonal elements of D, q_j are columns of Q, and $\frac{\left(q_j' \mu_T\right)^2}{d_j}$ is the contribution of the j-th PC to the maximum in-sample squared Sharpe Ratio. Note that since $\lambda \geq 0$, we have $\frac{d_j}{d_j + \lambda} \leq 1$. The multiplication with $\left(\frac{d_j}{d_j + \lambda}\right)^2$ therefore causes contributions to the max squared SR from low-eigenvalue PCs to get penalized more than contributions of high-eigenvalue PCs.

Fitted means are shrunk in a similar fashion,

$$\begin{aligned}\hat{\mu}_T &= \Sigma_T \left(\Sigma_T + \lambda I\right)^{-1} \mu_T = QD \left(D + \lambda I\right)^{-1} Q' \mu_T \\ &= \sum_{j=1}^{H} q_j \left(\frac{d_j}{d_j + \lambda}\right) q_j' \mu_T,\end{aligned}$$

with stronger shrinkage applied to smaller d_j. Small d_j correspond to directions in the space of factors having small variance.[5] To clearly see this, consider the rotation of original space of returns into the space of principal components:

$$\hat{\mu}_T^{\mathrm{PC}_j} \equiv q_j' \hat{\mu}_T = \frac{d_j}{d_j + \lambda} \mu_T^{\mathrm{PC}_j},$$

i.e., for PC factors with small eigenvalues d_j, fitted means $\hat{\mu}_T^{\mathrm{PC}_j}$ are forced to be close to zero. Compare this to OLS,

$$\hat{\mu}_T^{\mathrm{PC}_j,\mathrm{LS}} = \mu_T^{\mathrm{PC}_j}.$$

Our procedure therefore jointly tilts covariance matrix (PC rotation) and expected return estimates by limiting contributions of small PCs to the total squared Sharpe ratio.

The economic interpretation of such shrinkage is that we judge as economically implausible the case that a principal component of the candidate factors has high mean return (or high contribution to the total squared Sharpe ratio), but a small eigenvalue. Kozak et al. (2015), for instance, show that much of the squared Sharpe ratio must come from large PCs; otherwise the variance of an SDF becomes too large. They also argue (in the context of a "behavioral" asset pricing model) that the only way to generate large cross-sectional variance of expected returns is to have sentiment investors'/arbitrageurs' demands line up with few large PCs.

The procedure is closely related to ridge regression – a popular technique in machine learning used to regularize regression-based problems with many RHS variables. Our method differs in several important ways. First, our penalty is economically motivated (see Kozak et al., 2015); the resulting penalty is derived from economic principles and has a very intuitive interpretation: we penalize contributions of the smallest PCs to the total maximum squared Sharpe ratio the most. Second, unlike in ridge regression, we need not to normalize and center all variables: our objective must hold without intercepts. Third, our objective is to maximize the squared Sharpe ratio (minimize the distance to the mean-variance frontier) rather than the prediction error as in ridge regression.

[5] Similar calculations for OLS estimates give $\hat{\mu}_T^{\mathrm{LS}} = \Sigma_T \hat{b}^{\mathrm{LS}} = \Sigma_T \left(\Sigma_T' \Sigma_T\right)^{-1} \Sigma_T' \mu_T = QQ' \mu_T = \mu_T$ in the case when Σ_T is square and symmetric. The intuition is straightforward: we regress N variables on N predictors, so we perfectly fit all means.

2.3.2 Effective degrees of freedom

We define the *effective degrees of freedom*[6] as follows:

$$\mathrm{df}(\lambda) = \mathrm{tr}\left[\Sigma_T \left(\Sigma_T + \lambda I\right)^{-1}\right] = \sum_{j=1}^{H} \frac{d_j}{d_j + \lambda}. \tag{7}$$

Note that $\lambda = 0$, which corresponds to no shrinkage, gives $\mathrm{df}(\lambda) = H$, the number of free parameters.

Finally, since the estimator is linear, it is straightforward to calculate the variance-covariance matrix,

$$\mathrm{var}\left(\hat{b}\right) = \sigma^2 \left(\Sigma_T + \lambda I\right)^{-2}.$$

2.3.3 Alternative Formulation

The formulation in Eq. 5 minimizes the HJ-distance to the in-sample mean-variance frontier subject to the L_2 penalty on b. Intuitively, we are maximizing the squared Sharpe ratio while penalizing contributions of small PCs.

Interestingly, an alternative formulation which minimizes cross-sectional pricing errors subject to the constraint on the maximum squared Sharpe ratio (variance of the SDF) leads to the same solution. To see this, start with the asset pricing equation in Eq. 3 and consider Eq. 2 subject to an SDF second moment constraint:

$$b'\Sigma b \leq B, \tag{8}$$

i.e., we are looking for an SDF with bounded variance that does not allow too high Sharpe ratios.

We can then replace the pricing moments implied by Eq. 3 by their in-sample counterparts and arrive at the following formulation:

$$\hat{b} = \arg\min_b \left(\mu_T - \Sigma_T b\right)' \left(\mu_T - \Sigma_T b\right) + \lambda b'\Sigma_T b$$

The solution to this problem coincides with Eq. 6.

2.4 Approximate SDF: L_1 penalty

Based on the ridge regression, elements of $\hat{b}$ are shrunk, but none of them are set to zero. In this section we argue that many economic theories predict sparse SDF representations.

[6] Similar definition is often used in the ridge regression setup. See Hastie et al. (2011).

It is natural then to impose an L_1-analog of the penalty on $b'b$. Such an approach leads to a version of lasso regression. As a result, we can achieve sparsity – get some elements of $\hat{b}$ set to zero. This amounts to automatic factor selection, which may be an attractive feature because it allows us to express the SDF based on factors constructed from a relatively small set of stock characteristics.

We thus impose an L_1-analog of the penalty on $b'b$ by penalizing the sum of absolute values of SDF coefficients, $\sum_{j=1}^{H} |b_j|$. One key difference of our approach from others that work off Fama-MacBeth (returns on betas) regression coefficients (and penalize those) is that there is an economic reason for low eigenvalue PCs to make low SR contributions (and to not help price assets in an SDF). In contrast, regression coefficients in Fama-MacBeth regressions are factor mimicking portfolio mean returns and the factors can have non-zero expected returns even if they don't help price assets. Thus, it's not clear that it makes sense to impose a penalty based on the sum of absolute or sum of squares of these cross-sectional regression coefficients.

2.4.1 Interpretation

Economic models. Economic theories predict sparsity of the weights on zero-investment portfolios in the SDF.

Rational models: Existing equilibrium theories imply sparse SDFs. Most extreme case: CAPM. Roughly speaking, the CAPM would imply an SDF of the form $M_t = a + b \times R_t^M$. Other examples: investment-based asset pricing models that work with, say, two-factor reduced SDFs where these models imply that a few firm characteristics span exposures to these few factors.

Behavioral models: One can argue that limited attention implies that only a relatively small number of common factors are subject to sentiment. Kozak et al. (2015) restrict the size of behavioral demand of sentiment investors in L_2-norm, but similar considerations can be used to motivate L_1 constraints.

Given this, it makes economic sense to think of the prior distribution of b_i associated with characteristics that we, as econometricians, might try to use. Laplace priors[7] naturally lead to the lasso-type penalty we employ (see Section 2.5.3).

[7]Laplace distribution is also sometimes called the double exponential distribution, because it can be thought of as two exponential distributions spliced together back-to-back. Its density is $f(b|\tau) = \frac{1}{2\tau} \exp\left(-\frac{|b|}{\tau}\right)$.

2.5 Specification

We are therefore looking for an SDF that maximizes the OOS Sharpe ratio, i.e., is not "too far" from the in-sample MVE portfolio, but is robust out of sample. In our main specification we will impose both L_1 and L_2 penalties simultaneously. We believe both constraints are important and economically motivated as discussed above. We therefore seek to solve the following problem:

$$\hat{b} = \arg\min_b \left(\mu_T - \Sigma_T b\right)' \Sigma_T^{-1} \left(\mu_T - \Sigma_T b\right) + \lambda b'b + \gamma \sum_{i=1}^{H} |b_i| \,.$$

The method is similar to elastic nets. There are important differences, however. First, both L_1 and L_2 penalties are economically motivated and derived from first principles. Second, we need not to normalize and center all variables: our objective is the HJ-distance and no regression intercepts are needed. Third, we maximize the squared Sharpe ratio (minimize the distance to the mean-variance frontier) instead of minimizing (unweighted) pricing errors.

Finally, because we are looking for an SDF in the span of excess returns, we do not necessarily care about its scale. As in regularized regression-based methods, our penalties introduce bias, a lot of which is coming from the "level" bias. We are effectively doing two things: (i) shrinking the total variance of an SDF by re-scaling all b_i in the direction of zero; and (ii) re-scaling b_i cross-sectionally in a way that penalizes the smallest PCs the most. In our application we later undo the "level" shrinkage and focus only on the cross-sectional aspect (this has no effect on Sharpe ratios).

2.5.1 L_1 and L_2 penalties: Ridge vs. Lasso

Ridge regression is known to shrink the coefficients of correlated predictors towards each other, allowing them to borrow strength from each other. 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. From a Bayesian point of view, the ridge penalty is ideal if there are many predictors, and all have non-zero coefficients (drawn from a Gaussian distribution).

Lasso, on the other hand, is somewhat indifferent to very correlated predictors, and will tend to pick one and ignore the rest. In the extreme case above, the lasso problem breaks down. The Lasso penalty corresponds to a Laplace prior, which expects many coefficients to be close to zero, and a small subset to be larger and nonzero.

The elastic net performs much like the lasso, but removes any degeneracies and wild behavior caused by extreme correlations. Combining both penalties creates a useful com-

promise between ridge and lasso. As we vary relative strength of two types of penalties, the sparsity of the solution (i.e. the number of coefficients equal to zero) increases monotonically from 0 to the sparsity of the lasso solution (Friedman et al., 2010).

2.5.2 Solution method

Least Angle Regression (LAR) algorithm. We use a modified version of the LAR algorithm to solve the problem. Hastie et al. (2011) argue it is extremely efficient and computes the entire lasso path at a cost comparable to OLS. Moreover, predictors are added sequentially, so we can easily stop at any number of predictors (usually small).[8] This means that with a pre-computed Gram matrix, it is $O(kn^2)$ in our application (where k is the number of predictors at which to stop), i.e., even faster than OLS at $O(n^3)$.

The LAR algorithm starts with the empty set of active variables. At the first step it identifies the variable most correlated with the response. LAR then moves the coefficient of this variable continuously towards its least squares value. Walking along this direction, the angles between the variables and the residual vector are measured. Along this walk, the angles will change; in particular, the correlation between the residual vector and the active variable will shrink linearly towards 0. At some stage before this point, another variable will obtain the same correlation with respect to the residual vector as the active variable. The walk stops and the new variable is added to the active set. The new direction of the walk is towards the least squares solution of the two active variables, and so on. After p steps, the full least squares solution will be reached. Hastie et al. (2011) further show that a single modification to the LAR algorithm gives the entire lasso path. Namely, if a non-zero coefficient hits zero, we need to drop the variable from the active set and recompute the current joint least squares solution.

Finally, in the case when both L_1 and L_2 penalties are present (elastic nets), it is straightforward to show that one can turn this into a lasso problem using an augmented version of X and y. In the classic elastic nets setup, such augmentation is equivalent to replacing the Gram matrix $X'X$ used in computations of OLS coefficients at each step with its ridge counterpart $X'X + \delta I$. In our setting a similar simplification obtains.

Our implementation exploits these basic principles and adapts them to our setting. First, because our variables are not (and should not be) centered and standardized, we replace a measure of direction with inner product (instead of correlation). Inner product naturally measures both the angle and the size of the move – both are important in our setting. Second, since our objective is non-standard, we construct altered version of X and y as

[8] This is similar to early stopping regularization in boosting methods.

follows: $\tilde{X} = X^{\frac{1}{2}}$ and $\tilde{y} = X^{-\frac{1}{2}}y$, and the Gram matrix $(X + \delta I)$. One can easily verify that in the absence of L_1 penalty, such modifications lead to the solution discussed in Section 2.5.

QP problem. We can also formulate our final specification as a quadratic programming (QP) problem with linear constraints. For N predictors, we need $2N+1$ constraints and $2N$ variables. We use this method to validate the LAR algorithm above; it produces an identical solution. Since QP is a very general method that does not take advantage of specific structure of the problem, however, it is computationally quite inefficient.

2.5.3 Bayesian Interpretation

General Bayesian interpretation of ridge and lasso. The ridge penalty corresponds to an i.i.d. Normal prior. From a Bayesian point of view, the ridge penalty is ideal if there are many predictors, and all have non-zero coefficients (drawn from a Gaussian distribution).

The Lasso penalty corresponds to a Laplace prior, which expects many coefficients to be close to zero, and a small subset to be larger and nonzero.

Bayesian interpretation of our L_2 penalty. We can equivalently solve the following problem:

$$\hat{\lambda} = \arg\min_{\lambda} (\mu_T - \lambda)' (\mu_T - \lambda) \quad \text{s.t.} \quad \lambda' \Sigma_T^{-1} \lambda \leq B \tag{9}$$

where we substituted $\beta_T = I$, since we include all LHS variables as factors on the RHS. Note that this is simply a second-stage cross-sectional regression with an additional penalty on the maximum squared Sharpe ratio (SDF variance). This problem maps directly into the alternative formulation in Section 2.3.3 by a simple change of variables: $\lambda = \Sigma_T b$.

We can express the problem in PC space as:

$$\hat{\lambda}^{\text{PC}} = \arg\min_{\lambda^{\text{PC}}} \left(\mu_T^{\text{PC}} - \lambda^{\text{PC}}\right)' \left(\mu_T^{\text{PC}} - \lambda^{\text{PC}}\right) \quad \text{s.t.} \quad \lambda^{\text{PC}\prime} D^{-1} \lambda^{\text{PC}} \leq B, \tag{10}$$

where D is the matrix of eigenvalues of Σ. The FOC yields:

$$-\mu_T^{\text{PC}} + \hat{\lambda}^{\text{PC}} + \gamma D^{-1} \hat{\lambda}^{\text{PC}} = 0$$

where γ is the Lagrange multiplier. Then

$$\hat{\lambda}^{\text{PC},j} = \frac{d_j}{d_j + \gamma} \mu_T^{\text{PC},j},$$

which is exactly the solution (for $\hat{\mu}_T^{\text{PC}_j}$) in the previous section.

Consider the Bayesian interpretation of Eq. 10: our prior is that all standardized risk prices of PCs, λ^{PC}, are i.i.d Normal (and i.i.d Normal realized returns). This again underlines the difference between our approach and other approaches that penalize Fama-MacBeth (returns on betas) regression coefficients. These coefficients are factor mimicking portfolios means. Unless factors are orthogonal (e.g., PCs), some can have non-zero expected returns even if they don't help price assets.

Table 1: List of anomalies

The table lists all raw "anomaly" characteristics used in our analysis. We include powers, interactions of these raw characteristics, as well as other "non-anomaly" characteristics in many of our tests.

Size	Investment	Long-term Reversals
Value	Inv/Cap	Value (M)
Profitability	Investment Growth	Net Issuance (M)
Value-Profitability	Sales Growth	SUE
F-score	Leverage	Return on Book Equity
Debt Issuance	Return on Assets (A)	Return on Market Equity
Share Repurchases	Return on Equity (A)	Return on Assets
Net Issuance (A)	Sales/Price	Short-term Reversals
Accruals	Growth in LTNOA	Idiosyncratic Volatility
Asset Growth	Momentum (6m)	Beta Arbitrage
Asset Turnover	Industry Momentum	Seasonality
Gross Margins	Value-Momentum	Industry Rel. Reversals
D/P	Value-Prof-Momentum	Industry. Rel. Rev. (LV)
E/P	Short Interest	Industry Momentum-Rev
CF/P	Momentum (12m)	Composite Issuance
Net Operating Assets	Momentum-Reversals	Stock Price

3 Empirics

3.1 Data

We start with the universe of U.S. firms in CRSP. For each stock, using Compustat data we compute 50 "anomaly" characteristics commonly studied in the literature (listed in Table 1). For robustness, we exclude small-cap stocks[9], trim (drop top 5% and bottom 5%), center, and standardize all characteristics. Apart from that, we follow the anomaly definitions in Novy-Marx and Velikov (2016) and Kogan and Tian (2015).

In many of our tests we use powers and interactions of those 50 raw characteristics. Interactions expand the set of possible predictors exponentially. For instance, with only first-order interactions of 50 raw characteristics, we obtain $\frac{1}{2}n\left(n+1\right)+n=1,325$ predictors. In addition to using anomaly characteristics, we add indicators for Fama and French industry classification (49 industry dummies). With this expanded set of characteristics and all their

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

cross products, the number of predictors exceeds $5,000$.

We use *daily* returns from CRSP for each stock. Using daily data allows us to estimate second moments much more precisely than with monthly data. It does, however, make the problem somewhat computationally challenging, however. With thousands of stocks, trading days, and characteristics, we nearly one hundred of million daily stock-level return observations, each having thousands of predictors. The problem therefore requires performing trillions of computations and sufficient amount of memory to store intermediate steps.

3.2 Results

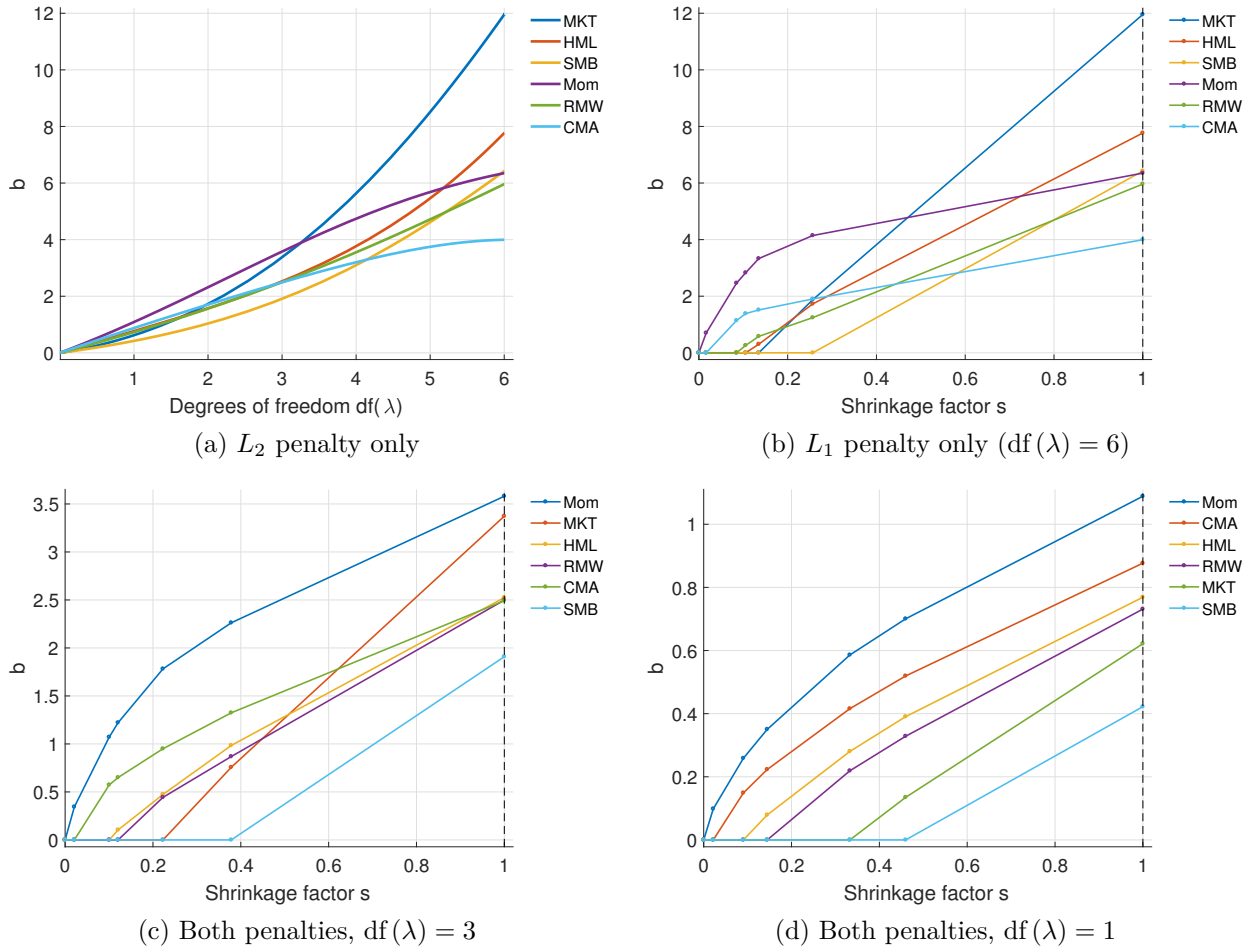
We start from the most basic example involving only a few test assets and characteristics, and proceed progressively towards our final specification that utilizes a broad range of characteristics, their powers and first-order interactions. The basic examples are revealing in terms of grasping intuition and comparing performance to classic techniques used in finance. The latter examples are infeasible for classic techniques and should be judged purely on their out-of-sample performance and new insights they uncover.

3.2.1 5 Fama-French Factors

In our first exercise, we use 5 Fama-French factors from Ken French's website to compare our "managed portfolios" approach to the simplest MVE portfolio of 5 Fama-French anomaly strategies. First, let's consider a case when our instrumented portfolios (test assets and candidate factors) coincide with the 5 original Fama-French factors: SMB, HML, MOM, RMW, CMA. These factors capture different anomalies and are nearly uncorrelated. We thus do not expect our regularization methods to have a lot of bite in this case.

Panel (a) of Figure 1 shows paths of estimated coefficients when only the L_2 penalty is imposed. The rightmost points on the plot correspond to unrestricted coefficients of the in-sample MVE portfolio. As we increase the strength of the penalty (move left on the plot), coefficients shrink toward zero and SDF variance drops (not shown). We can see that shrinkage is clearly non-linear and its strength varies substantially across different factors.

In Panels (b)-(d), for a fixed level of L_2 penalty, we additionally impose the L_1 penalty. The x-axis shows the shrinkage factor, $s = \frac{\sum_{i=1}^{H}|b_i|}{\sum_{i=1}^{H}|\tilde{b}_i|}$, where $\tilde{b}_i$ are coefficients corresponding to no L_1 penalty. Panel (b) corresponds to no L_2 penalty (6 degrees of freedom). As we tighten the L_1 constraint, coefficients shrink and eventually some of them get set to zero. The variables that are set to zero the earliest (we again move from right to left on the plot) are the least important. We therefore obtain sparse representations naturally, depending on how strong the L_1 penalty is.



(a) L_2 penalty only

(b) L_1 penalty only ($\mathrm{df}(\lambda) = 6$)

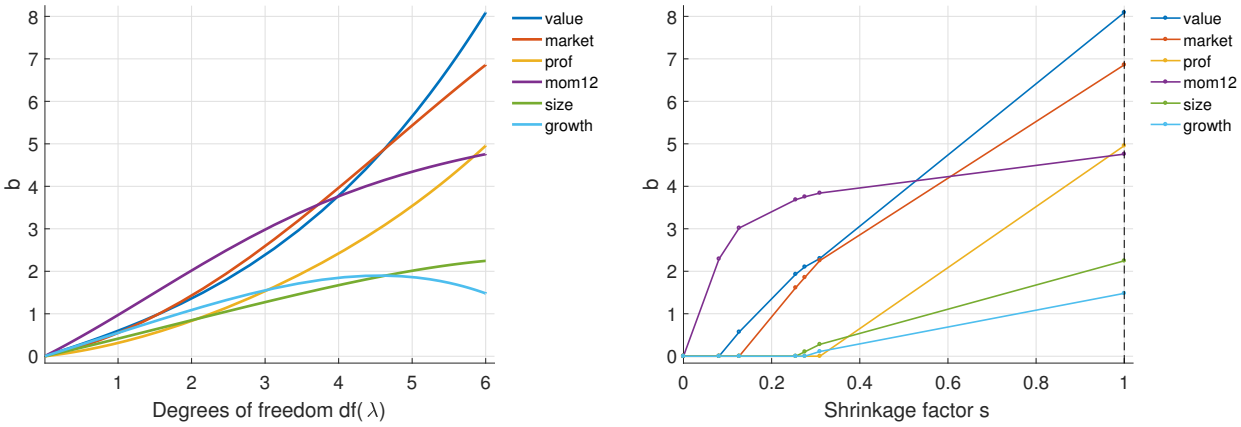
(c) Both penalties, $\mathrm{df}(\lambda) = 3$

(d) Both penalties, $\mathrm{df}(\lambda) = 1$

Figure 1: The figure shows paths of coefficients as a function of penalty strengths for original 5 Fama-French factors. Factor returns data is from Ken French's website. We impose only an L_2-norm penalty in Panel (a). We impose only an L_1-norm penalty in Panel (b). In Panels (c) and (d) we vary the strength of the L_1 penalty for a fixed L_2 penalty that corresponds to 3 and 1 effective degrees of freedom, respectively.

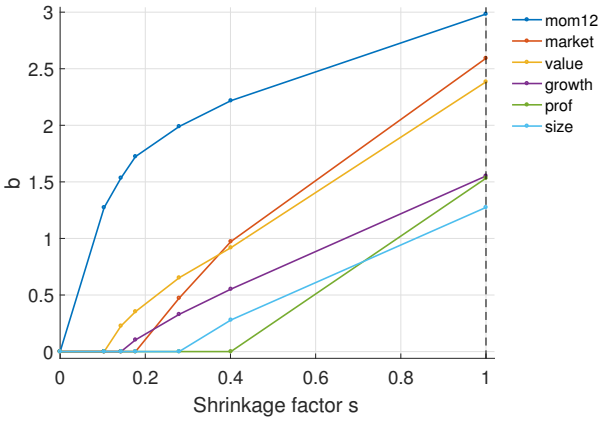
In Panel (c) we fixed the L_2 penalty at the level that corresponds to 3 effective degrees of freedom (Eq. 7). Panel (d) sets the L_2 penalty at 1 degree of freedom, As we tighten the L_1 constraint, coefficients shrink towards zero in both panels. We can see that the path of L_1 (lasso) coefficients becomes more monotone as more shrinkage is applied at the first stage (ridge). This happens because high level of L_2 shrinkage essentially eliminates small PCs and resulting non-monotonicities in the lasso coefficients paths they produce.

Figure 2 uses a different method of constructing instrumented portfolios. Instead of relying on Fama-French factors, we use the characteristics which underlie those factors (B/M, ME, past-12m returns, profitability, growth) to construct instruments Z_t. We standardize and center each of these instruments, so that an instrumented portfolio (cross-product of

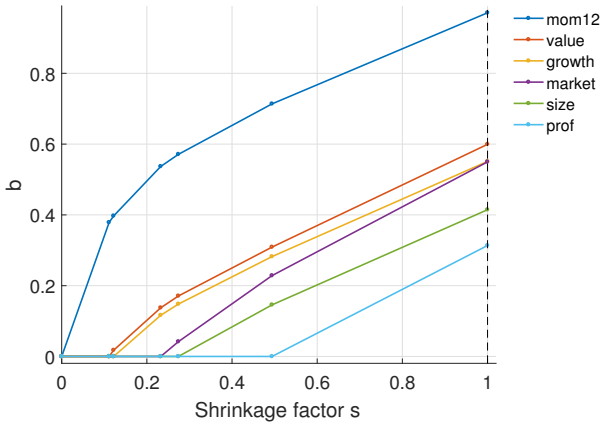


(a) No-near-arbitrage constraint only

(b) L_1 penalty only ($\mathrm{df}(\lambda) = 6$)



(c) Both penalties, $\mathrm{df}(\lambda) = 3$



(d) Both penalties, $\mathrm{df}(\lambda) = 1$

Figure 2: The figure shows paths of coefficients as a function of penalty strength for 5 linear instruments based on characteristics that underlie 5 Fama-French factors. We impose only L_2-norm penalty in Panel (a). We impose only L_1-norm penalty in Panel (b). In Panels (c) and (d) we vary the strength of the L_1 penalty for a fixed L_2 penalty that corresponds to 3 and 1 effective degrees of freedom, respectively.

an instrument and returns) has a natural interpretation of a long-short portfolio formed as a linear function of a characteristic.[10] We further use these instrumented portfolios as test assets and candidate factors in our estimation method. There are two primary differences between our simple linear instruments and FF factors: (a) we exclude small stocks; and (b) Fama and French use a somewhat more sophisticated approach to construct their factors.

Similarly to Figure 1, Panel (a) of Figure 2 shows paths of coefficients with only the L_2 penalty imposed, while Panel (b) imposes only L_1 penalty. Panels (c)-(d) impose two types of penalties simultaneously and correspond to different fixed (at 3 and 1 degrees of freedom, respectively) levels of L_2 penalty. We flipped signs on size and growth anomalies to match those of Fama-French factors for ease of interpretation. Comparing results in Figure 1 and Figure 2, we see that the "managed portfolios" method produces very similar results to the method that uses FF factors. Apart from profitability anomaly, for which our definition differs slightly from Fama and French (we follow Novy-Marx and Velikov, 2016), the remaining factors show very similar coefficient paths using either of two methods: momentum and value are the strongest anomalies; size is the weakest.

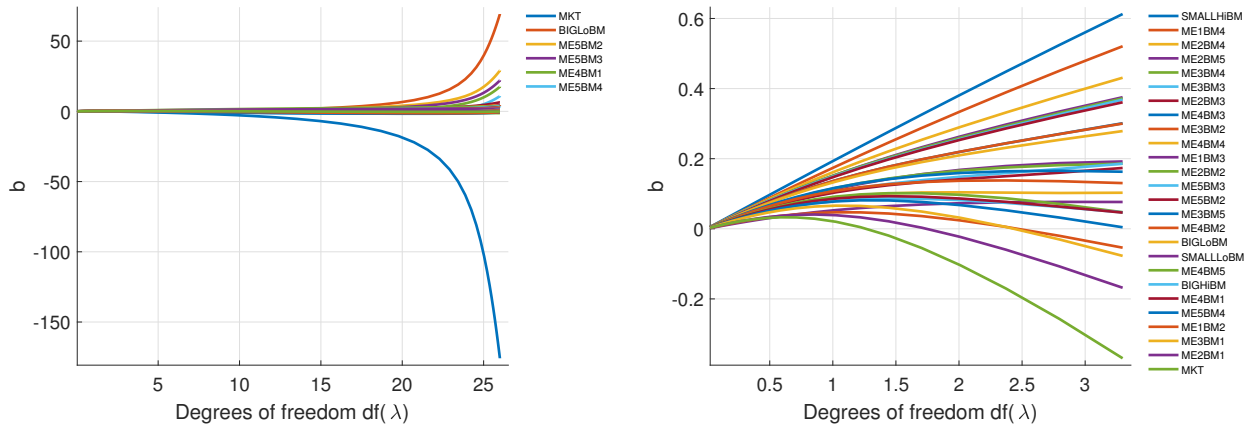
Overall, the evidence above suggests that our method nests standard approaches in cross-sectional asset pricing. It's important to remember, though, that those approaches are viable only in cases when factors are known and there are only few of them or with unknown factors when the universe of test assets is small (e.g., sorted portfolios) and the assets do not exhibit strong factor structure. In the previous example the universe of test assets we considered included only 5 strategy portfolios and had weak factor structure. In the next section we consider a simple example of strong factor structure in test asset returns.

3.2.2 Fama-French 25 ME/BM -sorted portfolios

In this example we consider 25 Fama-French ME/BM sorted portfolios that exhibit a very strong factor structure. Most of the variance can be explained by only three factors (market, SMB, HML). Fama and French (1992) construct these factors manually. Kozak et al. (2015) show that SMB and HML factors essentially match the second and the third PCs of 25 portfolio returns. Extracting and using only three such factors, therefore, is a form of regularization, known as *principal component regression* (PCR) in machine learning, but done implicitly. In our method, the no-near-arbitrage constraint leads to a variant of *ridge regression*, which is effectively a continuous version of PCR. Whereas PCR ignores small PCs completely, ridge regression strongly down-weights them instead.

Panel (a) of Figure 3 clearly show the need for regularization in this case: because FF25 portfolios are highly correlated, estimating the MVE portfolio with no constraints leads to

[10] Centering and standardizing implies that weights sum to zero and squared weights sum to N.



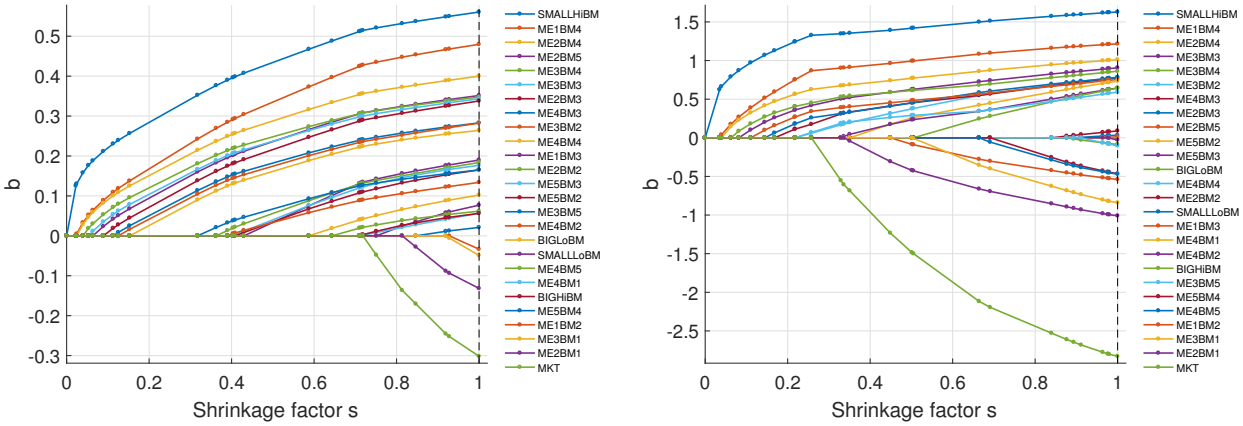
(a) Full path

(b) Regularized (zoomed in)

Figure 3: The figure shows paths of coefficients as a function of penalty strength for 25 Fama-French ME/BM sorted portfolios. We impose the L_2-norm penalty and vary its strength. Panel (a) shows the full path of coefficients. Panel (b) zooms in on the regularized range (between 0 and 3.5 degrees of freedom). Labels in panel (b) are ordered according to the vertical ordering of estimates at the rightmost point.

highly exaggerated SDF coefficients and high SDF variance. Panel (b) plots the profile of coefficients for a regularized problem, when coefficients are shrunk in a way that results in less than 3.5 effective degrees of freedom (in Panel (b) we simply zoom in on the portion of the plot in Panel (a)). Coefficient estimates are much more reasonable in this case and they have an intuitive pattern: highly positive coefficients correspond mostly to small and value portfolios, while negative coefficients primarily to growth or large portfolios (ordering of labels in Panel (b) coincides with the vertical ordering of coefficient estimates at the rightmost point). Effectively then, our regularized SDF (OOS-MVE portfolio) is long small and value stocks and short big and growth stocks — the same as the SDF implied by Fama and French (1992).

Figure 4 shows the path of SDF coefficients when both penalties are imposed. We fix the corresponding L_2 penalty to 3 or 10 degrees of freedom and vary the L_1 penalty. Labels are ordered according to the vertical ordering of estimates at the rightmost point (at the dashed black vertical line). Most of coefficients along the path are positive. This does not always have to be the case, but is likely, because with long-only test assets and positive market equity premium, our trading-costs constraints act similarly to short-selling constraints (Brodie et al., 2009). Most positive coefficients at high level of shrinkage (low s) can be attributed to high BM portfolios and small size portfolios. Additionally, because the resulting portfolio is long only, it has strong exposure to aggregate market. We can also see that a high level of L_2 (ridge) shrinkage leads to more averaging and potentially more robust SDF that goes long in



(a) Both constraints, $\mathrm{df}(\lambda) = 3$

(d) Both constraints, $\mathrm{df}(\lambda) = 10$

Figure 4: The figures show paths of coefficients as a function of L_1 penalty strength for 25 Fama-French ME/BM sorted portfolios. We impose the L_2-norm penalty that leads to 3 (Panel a) or 10 (Panel b) effective degrees of freedom. We further add the L_1 penalty. Labels are ordered according to the vertical ordering of estimates at the rightmost point (at the dashed black vertical line).

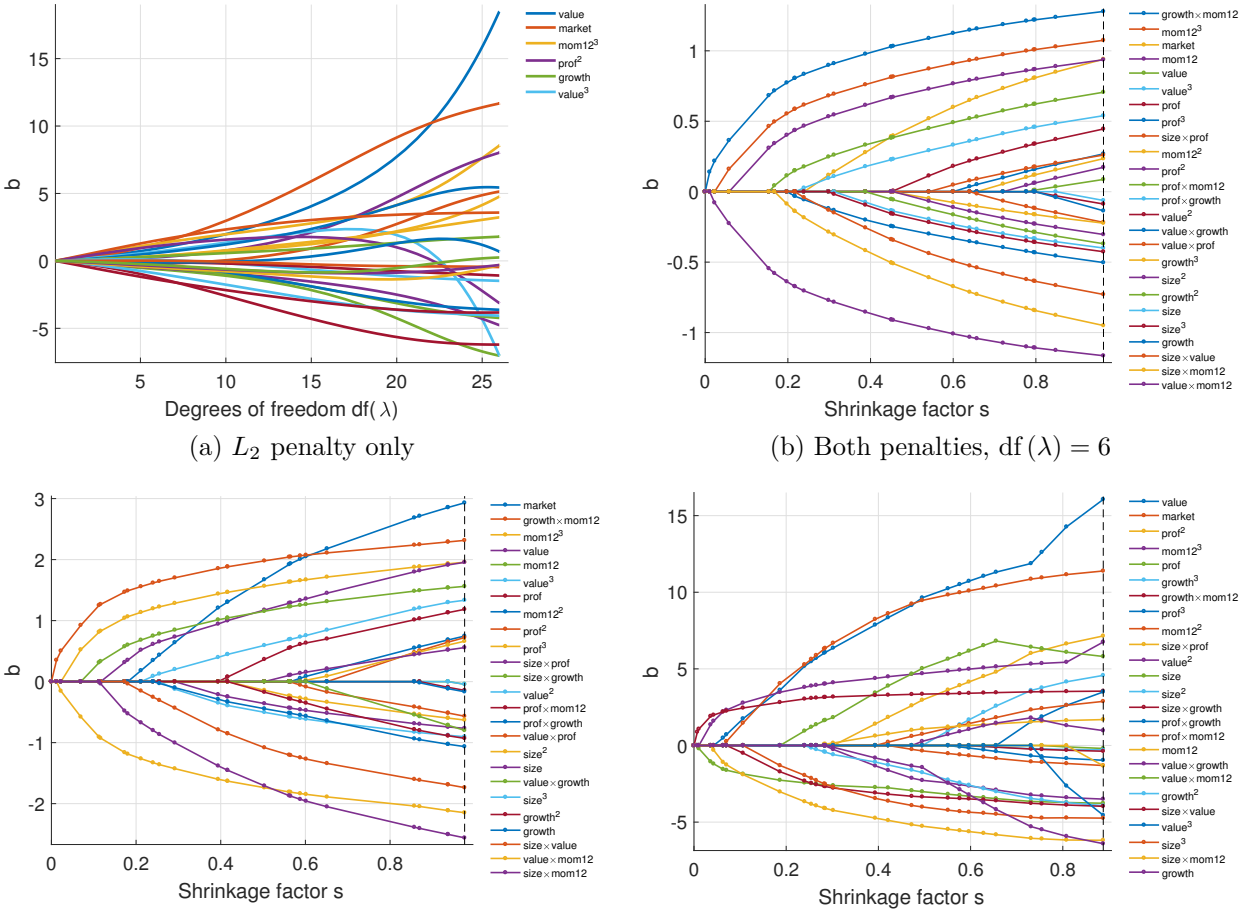
small value stocks, shorts the market and growth firms, and allocates small positive weights to the rest. The path of coefficients is much more monotone (as a result of small PCs being essentially eliminated) relative to the case of weak L_2 shrinkage in Panel (b).

Overall, our method tends to recover an SDF that is related to the SDF implied by Fama and French (1992). Further regularization tends to slightly favor the value strategy relative to size. This example illustrates that regularization is particularly important when factor structure is strong or some candidate factors are highly correlated. In the simple context of portfolio sorts, the intuitive need for such regularization was realized and accomplished implicitly by Fama and French (1992). The real strength of our method however comes when dealing with vast abundance of characteristics and unknown factors.

In what follows we work only with managed long-short portfolios and iteratively increase the size and complexity of models we consider.

3.2.3 Interactions and powers of 5 characteristics

We now go beyond using 5 FF factors or only 5 managed portfolios that are linear in implied characteristics. Namely, we extend the list of characteristics to include their second and third moments, as well as all first-order interactions, for a total of 26 instruments (market $+ \ 2 \times 5 + \frac{1}{2}5 \times (5+1)$). All derived characteristics, as always, are re-standardized and re-centered.



(a) L_2 penalty only

(b) Both penalties, $\mathrm{df}(\lambda) = 6$

(c) Both penalties, $\mathrm{df}(\lambda) = 10$

(d) Both penalties, $\mathrm{df}(\lambda) = 26$

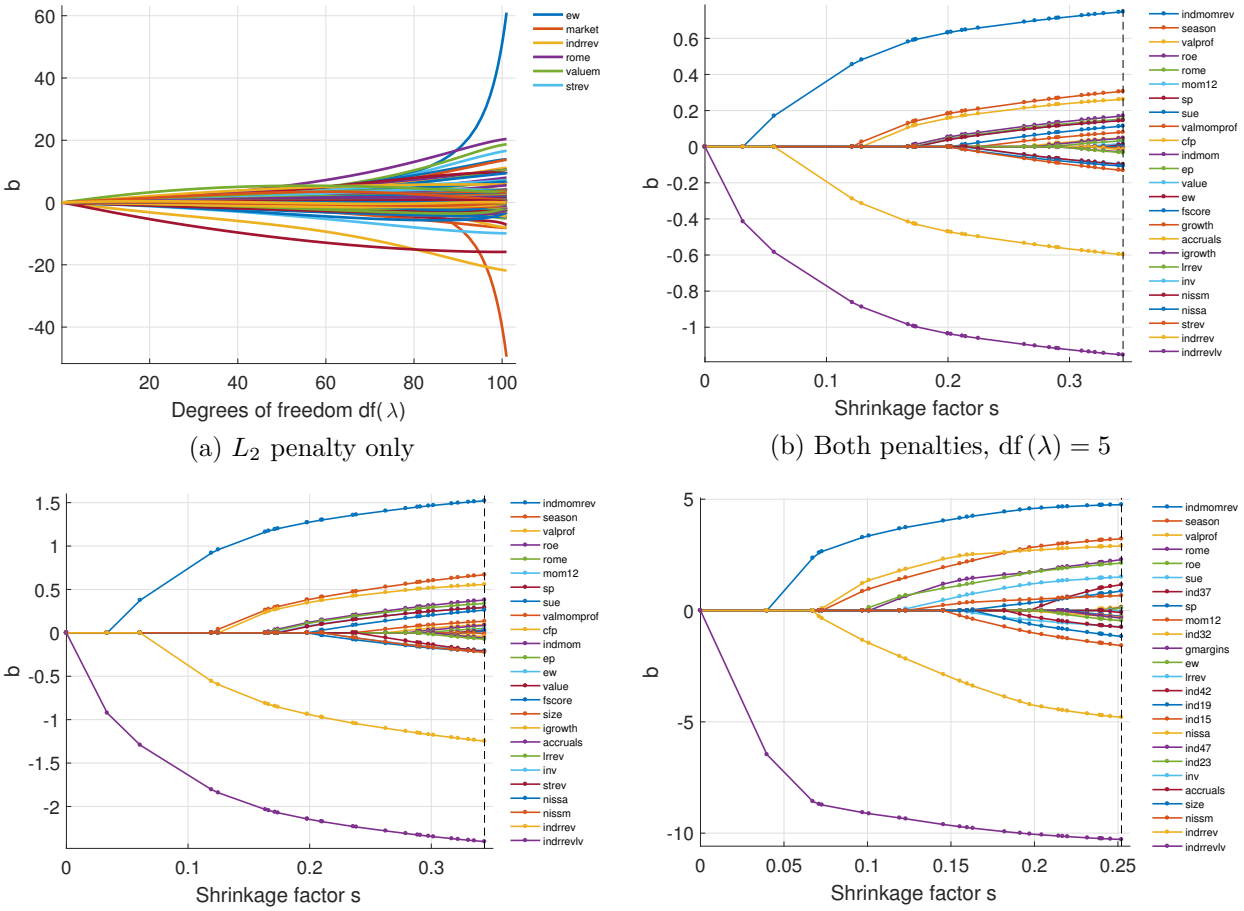
Figure 5: The figure shows paths of coefficients as a function of penalty strength for all characteristics underlying 5 Fama-French factors, their second and third moments, and all cross-products. We impose only the L_2-norm penalty in Panel (a). In Panels (b)-(d) we vary the strength of L_1 penalty for a fixed L_2 penalty that corresponds to 6, 10, and 26 effective degrees of freedom, respectively. Labels are ordered according to the vertical ordering of estimates at the rightmost point (at the dashed black vertical line).

Figure 5 shows the results. Similarly to the case with 25 ME/BM portfolios, without regularization some of the coefficients explode and result in high SDF variance. Some form of regularization is therefore needed. We combine both the L_2 penalty in Panel (a), and L_1 penalty that further shrinks coefficients in Panels (b)-(d) for a fixed level of L_2 penalty that corresponds to 6, 10, and 26 effective degrees of freedom, respectively. Panel (b) also shows that interactions and powers are important: the resulting SDF contains the cubes of momentum and value strategies (these are more aggressive versions of respective linear strategies with more weight put on tails) and interactions, namely interactions of value and momentum, and interactions of these two strategies with size. Our methodology, therefore, automatically recovers known patterns and interactions without requiring researcher "intuition".

3.2.4 50 anomaly characteristics and 49 FF industries

We now use the 50 anomaly characteristics listed in Table 1 and 49 FF industry dummies. Since characteristics are always re-centered, managed portfolios corresponding to industry dummies are effectively long that industry stocks and short the rest of the equity market. Because anomalies in Table 1 are not very correlated overall, we expect the overall factor structure to be weaker compared to 25 FF portfolios, where 3 PCs contributed most of the variance. However, some of the individual anomalies, and certainly many industry portfolios, are highly correlated. We therefore expect that some regularization is needed to handle those correlated anomalies.

Figure 6 shows the results. Similarly to the case with 25 ME/BM portfolios, without regularization some of the coefficients explode and result in high SDF variance. Some form of regularization is therefore needed. We combine both the L_2 penalty in Panel (a), and L_1 penalty (for a fixed level of L_2 penalty that corresponds to 5, 10, and 50 effective degrees of freedom) that is limited to 25 non-zero predictors in Panels (b)-(d). We can see that the most important anomalies that survive two stages of regularization include: industry momentum-reversals, industry relative reversals, value-profitability, ROE, seasonality, net issuance, sales/price, momentum, long-run reversals, 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. Interestingly, our method does not get confused by the presence of industry portfolios: none of them are included as factors in the final SDF representation if a sufficient level of L_2 penalty is applied (Panels (b) and (c)). In the case of weak L_2 shrinkage in Panel (d), however, many industries appear as factors. This example underlines the importance of L_2 shrinkage discussed earlier in the paper (in that it heavily down-weights small PCs, like those corresponding to movements in individual industries).



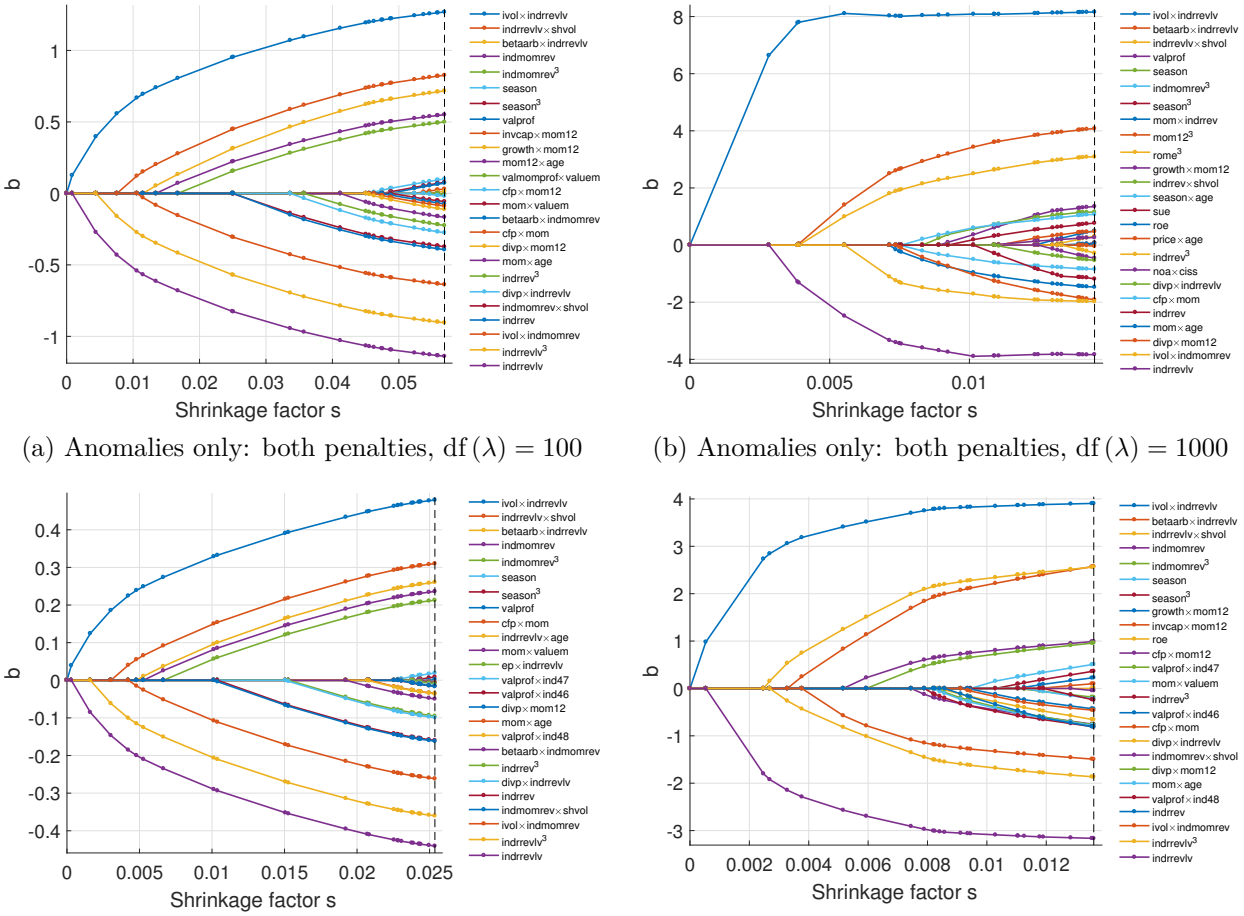
(a) L_2 penalty only

(b) Both penalties, $\mathrm{df}(\lambda) = 5$

(c) Both penalties, $\mathrm{df}(\lambda) = 10$

(d) Both penalties, $\mathrm{df}(\lambda) = 50$

Figure 6: The figure shows paths of coefficients as a function of penalty strength for 50 anomaly and 49 industry portfolios based on characteristics in Table 1 and FF 49 industry classification. We impose only L_1-norm penalty in Panel (a). In Panels (b)-(d) we vary the strength of L_1 penalty for a fixed level of L_2 penalty that corresponds to 5, 10, and 50 effective degrees of freedom, respectively. We truncate the path at 25 non-zero predictors. Labels are ordered according to the vertical ordering of estimates at the rightmost point (at the dashed black vertical line).



(a) Anomalies only: both penalties, $\mathrm{df}(\lambda) = 100$

(b) Anomalies only: both penalties, $\mathrm{df}(\lambda) = 1000$

(c) Anomalies and industries: both penalties, $\mathrm{df}(\lambda) = 100$

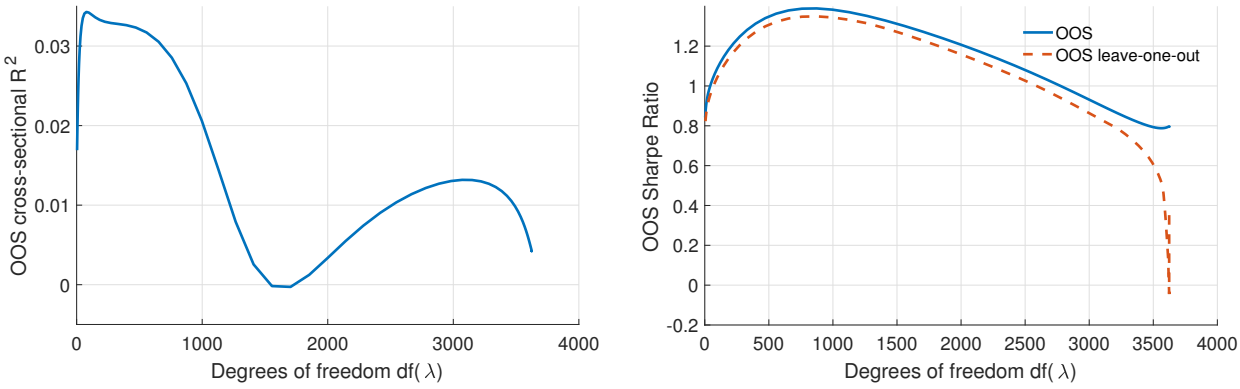
(d) Anomalies and industries: both penalties, $\mathrm{df}(\lambda) = 1000$

Figure 7: The figure shows paths of coefficients as a function of L_1 penalty strength for for 50 anomaly portfolios (Panels a and b) and 50 anomalies + 49 industry portfolios (Panels c and d) based on characteristics in Table 1 and FF 49 industry classification. We impose the L_2 penalty that corresponds to 100 or 1000 effective degrees of freedom. We then additionally impose L_1 penalty and show lasso coefficient paths. We truncate the number of factors in the SDF to 25. Labels are ordered according to the vertical ordering of estimates at the dashed black vertical line.

3.2.5 The ultimate model: 50 characteristics + 49 industries; their second and third moments and all cross-products

We now show the true power of our method by considering the case of vast abundance of characteristics and candidate factors. We start with 50 characteristics and 49 industry dummies, compute their second and third moments and all cross-products. We obtain more than 3,500 derived characteristics that we feed into our method (we drop interactions of two industry indicators).

The full path of L_2 regularization in the case of this many variables is explosive; we



(a) OOS cross-sectional R^2

(b) Out-of-sample Sharpe Ratios

Figure 8: The figure shows out-of-sample cross-sectional R^2 in Panel (a) and out-of-sample Sharpe ratios in Panel (b) as functions of the effective degrees of freedom (strength of L_2 penalty). We do not impose the L_1 penalty in these calculations. Panel (b) shows two lines: regular out-of-sample estimates of Sharpe ratios (blue solid line) and leave-one-out OOS Sharpe ratios which are obtained by excluding the single most important predictor. The out-of-sample period starts on 01/2005.

therefore do not show it. Instead, we apply two types of regularization and show only paths of coefficients implied by those. In doing so, we first impose the L_2 penalty corresponding to 100 or 1000 effective degrees of freedom. Next, we show the paths of coefficients implied by varying strength of L_1 penalty. Figure 7 shows the results. Panels (a)-(b) use only anomaly portfolios (and all interactions, second and third moments). Panels (c)-(d) extend the set of test assets to also include 49 FF industry portfolios. Industry momentum-reversals, industry relative reversals (low vol.) still seem to be important, but many interesting interactions now appear. Among the new ones are value-momentum and idiosyncratic volatility, asset growth and momentum, momentum and beta arbitrage. Some of interactions of anomalies and industries show up: valprof with real estate and financial trading industries.

3.3 Out-of-sample performance

We now evaluate the out-of-sample performance of our method. It is also useful if one is uncertain about how much regularization to use. Although we consider economic motivation to be the best guidance in picking regularization parameters (e.g., bound on maximum Sharpe ratio in Section 2.3.3), one might also want to explore the values of these parameters suggested by the data (validation or out-of-sample portion of the sample).

For our out-of-sample tests we use data prior to 2005 to estimate the model and compute all SDF coefficients. We then re-estimate means and covariances in the out-of-sample period (2005-2016) and use the pre-estimated SDF coefficients to construct measures of out-of-

sample cross-sectional R^2 and Sharpe ratio.

Figure 8 plots the out-of-sample cross-sectional R^2 in Panel (a) and out-of-sample Sharpe ratios in Panel (b) as functions of the effective degrees of freedom (strength of L_2 penalty). We do not impose L_1 constraints in these calculations. It is very difficult to achieve sparsity in pricing such a huge cross-section in the original space of managed portfolios (model fit and OOS R^2 keeps increasing with the number of predictors). Kozak et al. (2015) show that it is often possible to find an SDF in terms of small number of PCs (i.e., if we seek sparsity in the space of PCs — such an approach would require imposing L_1 bound on $Q'b$). When we relax the L_1 constraint sufficiently to allow for hundreds of predictors (and for sufficiently lax L_2 constraint), we obtain similar OOS plots when two types of penalties are present (not shown). Panel (b) shows two lines: regular out-of-sample estimates of Sharpe ratios (blue solid line) and leave-one-out OOS Sharpe ratios which are obtained by excluding the single most important predictor (more robust).

The figures show that the highest cross-sectional R^2 and highest OOS Sharpe ratios are achieved when regularization is aggressive: we need to reduce the effective number of degrees of freedom to around 500 (starting from 3500+). The latter SDF performs well out-of-sample: Sharpe ratios in the post-2005 period (which includes the financial crisis) are above 1.3. Sharpe ratios this high translate into $\approx 26\%$ annual risk premium for the same level of volatility as that of the aggregate market. Note however that resulting SDF is far from sparse: it requires hundreds of factors to achieve this level of OOS performance. Also notice that cross-sectional R^2 are very low (around 3.5%). This is not necessarily an issue. We are used to high cross-sectional R^2 because we typically work with univariate portfolio sorts and the set of test assets has only few meaningful sources of variation. This is not the case here.

Finally, we perform a 5-fold cross-validation exercise to assess the level of cross-sectional fit which is potentially less sample dependent than our sample-split exercise. We start by splitting the sample into five random non-contiguous parts. We then use four parts for estimation and the fifth part for assessing the OOS cross-sectional fit. We thus obtain 5 OOS estimates of a cross-sectional R^2 (for each of the 5 parts used as a validation sample), which we average out to compute a single estimate of expected cross-sectional OOS R^2. We repeat this entire exercise 10 times and average out the estimates to reduce dependency on the initial split.

One potential issue with this exercise is that many anomaly variables have been data-mined in the earlier part of the sample. Our estimates of expected returns are therefore biased upwards. However, this does not necessarily need to have big impact on the final SDF and OOS cross-sectional fit, since our method tends to shrink those estimates of expected returns

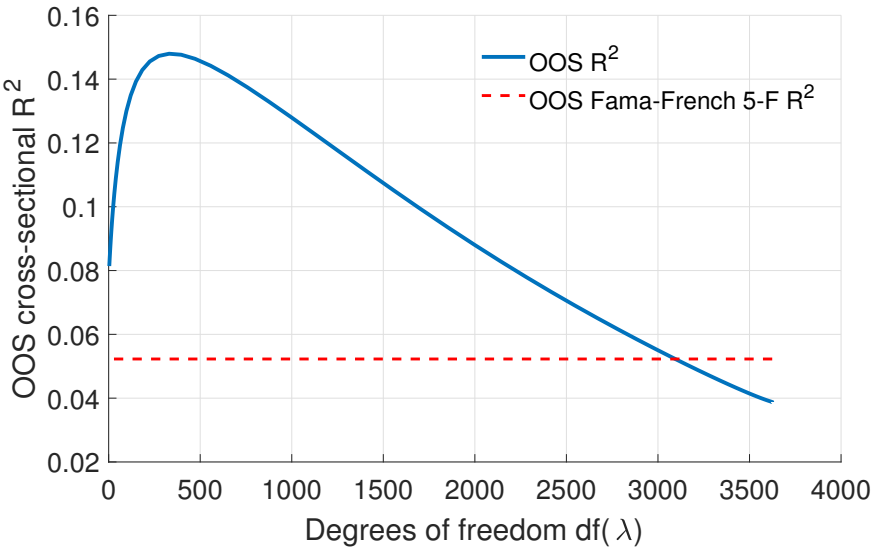


Figure 9: The figure shows out-of-sample cross-sectional R^2 obtained by performing a 5-fold cross-validation exercise. We start by splitting the sample into five random non-contiguous parts. We then use four parts for estimation and the fifth part for assessing the OOS cross-sectional fit. We thus obtain 5 OOS estimates of a cross-sectional R^2 (for each of the 5 parts used as a validation sample), which we average out to compute a single estimate of expected cross-sectional OOS R^2. We do not impose the L_1 penalty in these calculations. The dashed red line shows the OOS cross-sectional R^2 implied by the (unregularized) model that uses only 5 Fama-French factors to construct the SDF that explains the cross-section of all managed portfolios.

substantially. McLean and Pontiff (2016) argue that many anomalies essentially disappear following a research paper publication. For "unknown" anomalies in the earlier part of the sample then there is no clear reason why expected returns need to line up with covariances in the data (our model imposes this). If they don't, and mostly line up with small PCs, our algorithm will effectively shrink them. It is therefore only anomalies that have been data-mined and correspond to large systematic co-movements in the data that could be problematic for our procedure.

We show the results in Figure 9. We can see that the optimal strength of regularization is similar to the estimates which we obtained using the split-sample exercise. We still require strong regularization that shrinks the effective degrees of freedom to about 400. The fact that this estimate is so close to estimates obtained using the split-sample analysis suggests that the cross-validation approach might be relatively immune to anomaly "fishing" for the purposes of estimating the optimal degree of shrinkage.

For comparison, we also plot the OOS cross-sectional R^2 implied by the (unregularized) model that uses only 5 Fama-French factors to construct the SDF that explains the cross-section of all managed portfolios (the dashed red line). We use the pre-2005 sample to

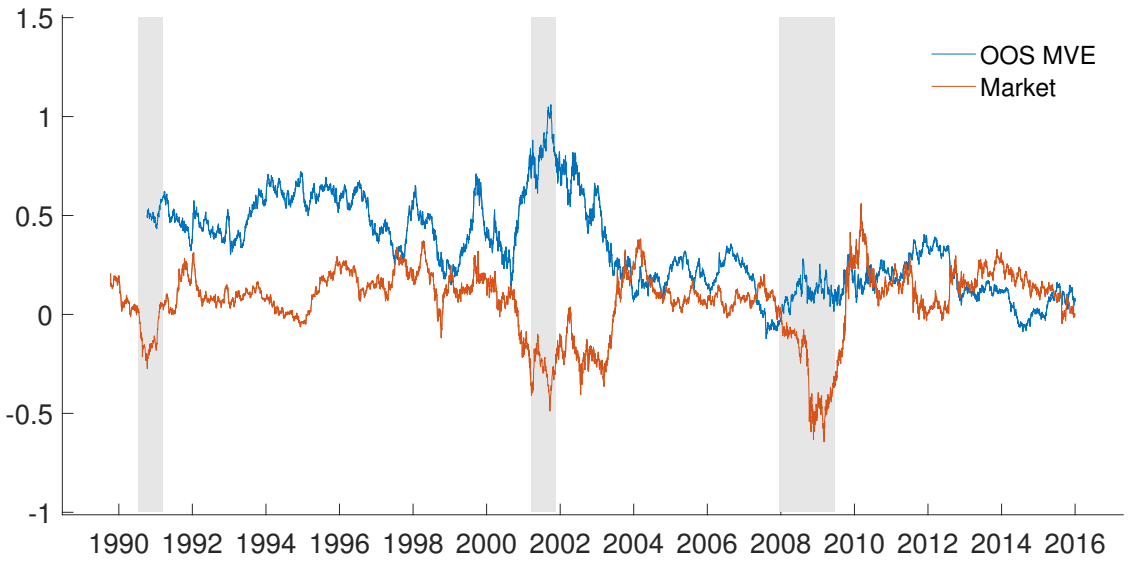


Figure 10: The figure plots the time-series of one-year overlapping returns on the regularized MVE portfolio implied by our SDF (blue solid line) and returns on the market (red dashed line). We use the first half of the sample (before 1990) to estimate SDF coefficients (we regularize the problem by imposing L_2 penalty that corresponds to 400 effective degrees of freedom) and then fix those coefficients to construct OOS returns on the MVE portfolio.

estimate SDF weights (on 5 FF factors and the market) and use those weights to restrict the MVE portfolio in the post-2005 sample. Our method that efficiently combines all managed portfolios in an SDF demonstrates substantially better performance than the model based only on the restricted linear combination of 5 FF factors (and the market).

Finally, we plot the time-series of one-year overlapping returns on the regularized OOS-MVE portfolio implied by our SDF in Figure 10 (blue solid line). We use the first half of the sample (before 1990) to estimate SDF coefficients (we regularize the problem by imposing L_2 penalty that corresponds to 400 effective degrees of freedom) and then fix those coefficients to construct OOS returns on the MVE portfolio in the second half of the sample. The red dashed line shows returns on the market index for comparison. Note that average returns on the MVE portfolio are much higher in the pre-2005 period, resulting in extremely high Sharpe ratios. This likely happens due to the fact that many anomalies have been data-mined in the earlier part of the sample. In the post-2005 period mean returns on the MVE portfolio deteriorate significantly; however, they are still substantially higher than mean returns on the stock market index. Note that our MVE portfolio performs significantly better than the market in recessions, often moving in the opposite direction to the market and generating very high returns (especially in 2001-2002).

References

Brodie, J., I. Daubechies, C. De Mol, D. Giannone, and I. Loris (2009). Sparse and stable markowitz portfolios. *Proceedings of the National Academy of Sciences 106*(30), 12267–12272.

Cochrane, J. H. (1991). Production-based asset pricing and the link between stock returns and economic fluctuations. *Journal of Finance 46*, 209–237.

Cochrane, J. H. (2005). The risk and return of venture capital. *Journal of Financial Economics 75*(1), 3–52.

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

Fama, E. F. and K. R. French (1992). The cross-section of expected stock returns. *Journal of Finance 47*, 427–465.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software 33*(1), 1.

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

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

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

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

Kozak, S., S. Nagel, and S. Santosh (2015). Interpreting factor models. Technical report, University of Michigan.

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

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