Shrinking the cross section

Muaz Chowdhury, Matias Data, Lorenzo Latini

December 27, 2023

Contents

1 Introduction

The main contribution of Kozak et al. (2020) is the introduction of a robust methodology to estimate a stochastic discount factor (SDF) that performs well in a high dimensional setting with a large number of stock characteristics that act as return predictors. Past papers in the asset pricing literature typically focused on linear factor models with a small number of characteristics-based factors such as the well known Fama-French three-factor model (Fama and French 1993). This paper shows that this sparsity requirement in the characteristic space comes at a high cost of predictive performance, and non-sparse models with large number of factors have much better out-of-sample performance when they are used with appropriate (Bayesian) regularisation methods. However, the authors show that a sparse representations of the SDF with high explanatory value is found in the principal component (PCs) factor space.

This report is organised as follows: Section 2 gives a self-contained introduction to the subject of asset pricing, where in particular the concept of an SDF is introduced as well its equivalence with beta-representations and mean-variance efficiency. Moreover we introduce factor models and the role of conditioning information; section 3 gives an overview of the Bayesian methodology introduced in Kozak et al. (2020), where we include proofs for most of the statements in the paper; finally section 4 contains the empirical results where we replicate the results obtained in the paper.

2 Preliminaries

2.1 Regression and linear prediction

Given random variables $X_1, \dots, X_K$ and Y all with finite second moments (that is, belonging to $L^2(\Omega, \Sigma, \mathbb{P})$) we consider the problem of finding the closest random variable in the linear

span of $X_1, \dots, X_K$, i.e. $S = \{\beta^\top X \mid \beta \in \mathbb{R}^K\}$ where $X = [X_1, \dots, X_K]^\top$. This is no other than the orthogonal projection of Y into this subspace, as in any Hilbert space. It can be characterised as follows:

$$\text{proj}_S(Y) \in S, \quad \langle Y - \text{proj}_S(Y), X_j \rangle = 0, \quad j = 1, \dots, K.$$

Using that the inner product is the expectation, and writing this in a vector form we get that:

$$\mathbb{E}[YX^\top] = \mathbb{E}[\text{proj}_S(Y)X^\top] = \mathbb{E}[\beta^\top XX^\top] = \beta^\top \mathbb{E}[XX^\top]$$

Hence assuming $\mathbb{E}[XX^\top]$ is invertible (no redundant variables, i.e. the variables are linearly independent) there exists a unique β given by $\beta = \mathbb{E}[XX^\top]^{-1}\mathbb{E}[XY^\top]$. The same formula holds if we want to project N random variables $Y = [Y_1, \dots, Y_N]^\top$, and that is the reason why we have kept the transpose over Y.

Notice that we can restate this as $Y = \beta^\top X + \varepsilon$, where ε is such that $\mathbb{E}[\varepsilon X^\top] = 0$. Also, observe that it does not necessarily hold that $\mathbb{E}[\varepsilon] = 0$. However, it does hold if we add a constant and consider the subspace generated by $\mathbb{1}, X_1, \dots, X_K$.

Proposition 2.1. Suppose we add a constant and project Y onto the space generated by $\mathbb{1}, X_1, \dots, X_K$. Given $\beta = [\beta_0, \beta_1, \dots, \beta_K]^\top \in \mathbb{R}^{K+1}$ the regression coefficients, denote β_0 the coefficient corresponding to the constant $\mathbb{1}$ and $\beta_1 = [\beta_1, \dots, \beta_K]^\top$ correspond to $X_1, \dots, X_K$ respectively. Assume $\mathbb{1}, X_1, \dots, X_K$ are linearly independent. Then $\beta_0 = \mathbb{E}[Y] - \beta_1^\top \mathbb{E}[X]$, and $\beta_1 = \mathbb{V}[X]^{-1}\text{Cov}[X, Y]$.

Proof. Since $\mathbb{1}$ is in the space where we are projecting, then

$$0 = \mathbb{E}[(Y - \text{proj}_S(Y)1] = \mathbb{E}[Y - \beta_0 - \beta_1^\top X] = \mathbb{E}[Y] - \beta_0 - \beta_1^\top \mathbb{E}[X].$$

Now going back to the normal equations for X we get

$$0 = \mathbb{E}[(Y - \beta_0 - \beta_1^\top X)X^T] = \mathbb{E}[(Y - \mathbb{E}[Y] + \beta_1^\top \mathbb{E}[X] - \beta_1^\top X)X^T]$$

Thus,

$$0 = \mathbb{E}[(Y - \mathbb{E}[Y])X^\top] - \beta_1^\top \mathbb{E}[(X - \mathbb{E}[X])X^T] = \text{Cov}[Y, X] - \beta_1^\top \mathbb{V}[X]$$

Given that $\mathbb{1}, X_1, \dots, X_K$ are linearly independent, then $X_1 - \mathbb{E}[X_1], \dots, X_K - \mathbb{E}[X_K]$ are linearly independent. Hence $\mathbb{V}[X]$ is invertible and thus $\beta_1 = \mathbb{V}[X]^{-1}\text{Cov}[X, Y]$. □

Proposition 2.2. Given variables $X_1, \dots, X_K$, consider a regression of Y onto X without constant $Y = \beta^\top X + \varepsilon$ (i.e. $\mathbb{E}[\varepsilon X^\top] = 0$). Assume $\mathbb{1}, X_1, \dots, X_K$ are linearly independent. Then the following are equivalent:

1. $\mathbb{E}[\varepsilon] = 0$.

2. $\beta_0 = 0$ if we run a regression with a constant.

If either holds then $\beta = \mathbb{V}[X]^{-1}\mathrm{Cov}[X, Y]$.

Proof. 1. $\Rightarrow$ 2. If $\mathbb{E}[\varepsilon] = 0$ then ε is orthogonal to the constant, thus $\beta^\top X$ is the orthogonal projection onto the space generated by $\mathbb{1}, X_1, \dots, X_K$.

2. $\Rightarrow$ 1. We run a regression of Y onto X plus a constant $Y = \beta_0 + \beta_1^\top X + \nu$, with $\mathbb{E}[\nu X^\top] = 0$ and $\mathbb{E}[\nu] = 0$. If $\beta_0 = 0$, then $\beta_0 + \beta_1^\top X = \beta_1^\top X$ is in the linear space generated by $X_1, \dots, X_K$ and is orthogonal to all X_j's. Since $\mathbb{1}, X_1, \dots, X_K$ are linearly independent then β_1 is uniquely determined. Whence it follows that $\beta_1 = \beta$, $\varepsilon = \nu$, and then $\mathbb{E}[\varepsilon] = 0$. □

2.2 The Law of one price and stochastic discount factors

We follow here Cochrane (2009) to give a short introduction on the main themes of Asset Pricing. We think of a *payoff* as a random variable $X : \Omega \to \mathbb{R}$ where $X(\omega)$ represents the payoff of an asset (or portfolio of assets) when the state of nature is ω for $(\Omega, \Sigma, \mathbb{P})$, a probability space. The *payoff space*, denoted $\underline{X}$, is the set of all payoffs available in the market which we assume is a vector space. That is, if $X, Y \in \underline{X}$ then $aX + bY \in \underline{X}$. Moreover, we assume that the sample space Ω is finite, i.e. $\mathbb{P}(\omega_i) > 0$ for $\omega_1, \dots, \omega_N$ that add up to one. This space comes naturally equipped with the inner product given by

$$\langle X, Y \rangle = \mathbb{E}[XY] = \sum_{i=1}^{N} X(\omega_i) Y(\omega_i) \mathbb{P}(\omega_i).$$

Hence, being finite dimensional (it can be identified with a subspace of $\mathbb{R}^N$, where $N = \#\Omega$), it is a Hilbert space. The theory we develop here is in a one-period market, but it applies to a sequential market in each period, conditioning on the information of past observations.

We notice that the assumption that Ω is finite can be dropped: this is just for mathematical simplicity, but results can be generalised provided we work in an appropriate Hilbert

space (e.g. $L^2(\Omega, \Sigma, \mathbb{P})$). Moreover, in practice one can usually work with finite dimensional spaces (the span of N basis assets, e.g. stocks, bonds, etc.) in an appropriate Hilbert space.

We assume that every payoff $X \in \underline{X}$ has an associated price $p(X) \in \mathbb{R}$, hence we have a function $p : \underline{X} \to \mathbb{R}$.

Definition 2.3. The *Law of One Price (LOOP)* holds if $p(aX + bY) = ap(X) + bp(Y)$ for all $X, Y \in \underline{X}$, $a, b \in \mathbb{R}$.

The LOOP simply states that the price is a linear function, i.e. the price of a portfolio is the sum of the weights multiplied by the prices of the basic assets.

Definition 2.4. A *stochastic discount factor* (SDF) $M : \Omega \to \mathbb{R}$ is a random variable that represents the prices of all the payoffs, i.e. M is such that:

$$p(X) = \mathbb{E}[MX], \tag{1}$$

for all $X \in \underline{X}$.

Theorem 2.5. The Law of One Price holds in $(\underline{X}, p)$ if and only if there exists M an SDF. Moreover, if any of these conditions hold, there is a unique SDF in the payoff space $X^* \in \underline{X}$.

Proof. If M is an SDF then p is linear by the linearity of expectations. Conversely, if $p : \underline{X} \to \mathbb{R}$ is linear there is a unique payoff $X^* \in \underline{X}$ that represents this functional, i.e. such that $p(X) = \mathbb{E}[X^* X]$. If there is X^* that represents p and X is of price zero (i.e. $X \in K = \mathrm{Ker}(p)$), then $\mathbb{E}[X^* X] = p(X) = 0$, thus $X^* \in K^\perp$. Choose any non-zero vector X in the line $K^\perp$. For a vector X^* in this line (i.e. $X^* = \lambda X$) to be an SDF it needs to price itself correctly, i.e. we need $p(X^*) = \mathbb{E}[X^* X^*]$. Then we need to solve

$$\lambda p(X) = p(\lambda X) = p(X^*) = \mathbb{E}[X^* X^*] = \mathbb{E}[(\lambda X)^2] = \lambda^2 \mathbb{E}[X^2],$$

hence choose $\lambda = p(X)/\mathbb{E}[X^2]$. Then $X^* = \frac{p(X)}{\mathbb{E}[X^2]} X$ price itself correctly. It also prices every payoff $Y \in \underline{X}$ correctly, since $Y = aX^* + Z$ for unique $a \in \mathbb{R}$ and $Z \in K$, thus

$$\mathbb{E}[X^* Y] = \mathbb{E}[X^*(aX^* + Z)] = a\mathbb{E}[X^*] + 0 = ap(X^*) + p(Z) = p(Y).$$

Second Proof. Assume first that the market is complete, in the sense that $\underline{X} = \mathbb{R}^\Omega$. Then there is a unique SDF M^* that represents $p : \mathbb{R}^\Omega \to \mathbb{R}$ since

$$M^*(\omega_i)\mathbb{P}(\omega_i) = \mathbb{E}[M^* \mathbb{1}_{\{\omega_i\}}] = p(\mathbb{1}_{\{\omega_i\}}).$$

That is, $M^*(\omega_i)\mathbb{P}(\omega_i)$ is the price of the asset that pays exactly one unit in state ω_i and zero otherwise (these are called *Arrow-Debreu Assets*). If $\underline{X} \subset \mathbb{R}^\Omega$ then we can always extend the functional $p : \underline{X} \to \mathbb{R}$ to the whole space $\mathbb{R}^\Omega$, take an orthogonal basis of $\underline{X}$ and extend it with an orthogonal basis of $\underline{X}^\perp$, and define its extension $\overline{p} : \mathbb{R}^\Omega \to \mathbb{R}$ in any way on the orthogonal complement. Then $\overline{p}(X) = \mathbb{E}[MX]$ for a unique random variable M for all $X \in \mathbb{R}^\Omega$ as before.

Third Proof. Take a basis of N asset payoffs (e.g. N stocks) that span the payoff space. Denote by $X = [X_1, \ldots, X_N]^\top$ a $N \times 1$ random vector. The payoff space is then given by $\underline{X} = \{w^\top X \mid w \in \mathbb{R}^N\}$, i.e. all portfolios are in the span of the basis assets. Denote by $p \in \mathbb{R}^N$ the vector of prices of the basis assets, i.e. $p_i = p(X_i)$. We want an SDF $X^* \in \underline{X}$, hence $X^* = w^\top X$. But then if X^* prices correctly the basis assets

$$p = \mathbb{E}[X^* X] = \mathbb{E}[X X^*] = \mathbb{E}[X X^\top w] = \mathbb{E}[X X^\top] w$$

Then $w = \mathbb{E}[X X^\top]^{-1} p$. Notice that the matrix $\mathbb{E}[X X^\top]$ is invertible since it is the matrix of the expectation inner product $\langle , \rangle$ in the basis $X_1, \ldots, X_N$. Thus,

$$X^* = w^\top X = p\,\mathbb{E}[X X^\top]^{-1} X.$$

□

In infinite dimensional spaces the same follows from the Riesz representation theorem provided p is continuous.

Notice that in general, there could be an infinite number of SDFs, just pick any random variable in the orthogonal complement of the payoff space, i.e. $\varepsilon \in \underline{X}^\perp$ and then $M = X^* + \varepsilon$ is an SDF since:

$$\mathbb{E}[MX] = \mathbb{E}[X^* X] + \mathbb{E}[\varepsilon X] = \mathbb{E}[X^* X] = p(X),$$

for every payoff $X \in \underline{X}$. The converse also holds, if M is an SDF, then $M = X^* + \varepsilon$ with ε in the orthogonal complement of the payoff space.

It is more common in finance to talk about returns rather than prices and payoffs. To go from payoffs to returns we need the payoff of the asset (or portfolio) X to have a price p distinct from zero. In that case, we can define its return as $R = X/p$. This defines a payoff whose price is one. Moreover, we can define the space of returns as the hyperplane of assets with price one. The fundamental pricing equation for a return R is then:

$$1 = \mathbb{E}[MR].$$

The only problem with returns is that they are *not* a subspace, but they generate the payoff space. Thus, we can still restrict our work to price returns correctly.

A particularly important asset is the one whose payoff is equal to 1 in all states of nature ($\mathbb{1}(\omega) = 1$ for all $\omega \in \Omega$). This is what we call a *risk-free discount bond*. We will assume that this asset is traded in our markets of interest. Define the *risk-free rate* R_f as the return of this investment, hence if M is an SDF:

$$R_f = 1/\mathbb{E}[M].$$

Another important type of payoff is that of an *excess return*, which can be thought of as buying one unit (e.g. dollar) of one asset (or portfolio of assets) and selling short one unit in a second asset (i.e. buying -1 units of this asset). If we call a and b these assets/portfolios with returns R^a and R^b respectively, then the payoff of this strategy is then:

$$R^e = R^a - R^b$$

These are also called *zero-cost portfolios* since they lie in the hyperplane of zero-price payoffs:

$$\mathbb{E}[MR^e] = \mathbb{E}[MR^a] - \mathbb{E}[MR^b] = 0.$$

Moreover, the excess returns are the subspace of zero-cost portfolios. Of particular importance are the excess returns over the risk-free rate, i.e. when we borrow at the risk-free rate selling short one dollar of the risk-free discount bond.

Remark 2.6. Notice that from the arguments given on Theorem 2.5, it follows that the excess returns characterise the SDF in the payoff space up to a scalar.

Let us show an alternative formula for an SDF in terms of excess returns which is particularly useful in practice.

Proposition 2.7. *(Hansen-Jagannathan Formula)* Given a basis $R^e_1, \dots, R^e_N$ of the excess returns space, and assume the risk-free asset (with payoff $\mathbb{1}$) is traded in the market. Then,

$$X^* = \frac{1}{R^f} - \frac{1}{R^f}\mathbb{E}[R^e]^\top \Sigma^{-1}(R^e - \mathbb{E}[R^e]), \tag{2}$$

is an SDF, where $\Sigma = \mathbb{V}[R^e]$.

Proof. Observe that if $R_1^e, \dots, R_N^e$ is a basis of the space of excess returns then

$$\mathbb{1}, R_1^e - \mathbb{E}[R_1^e], \dots, R_N^e - \mathbb{E}[R_N^e]$$

is a basis of the payoff space since $\mathbb{1}$ is not an excess return unless the price of all payoffs is zero. Notice that $\mathbb{1}$ orthogonal to the other elements of this basis. Take the vector $X = [\mathbb{1}, R_1^e - \mathbb{E}[R_1^e], \dots, R_N^e - \mathbb{E}[R_N^e]]^\top$. Then, using the construction of the third proof of Theorem 2.5 we get an SDF X^* in the payoff space given by:

$$X^* = p\mathbb{E}[XX^\top]^{-1}X.$$

Denote by $\Sigma = \mathbb{V}[R^e] = \mathbb{E}[(R^e - \mathbb{E}[R^e])(R^e - \mathbb{E}[R^e])^\top]$, i.e. the covariance matrix of the excess returns. Then we compute and get:

$$p(X) = \left[\frac{1}{R^f}, -\frac{\mathbb{E}[R_1^e]}{R^f}, \dots, -\frac{\mathbb{E}[R_N^e]}{R^f}\right]^\top, \quad \mathbb{E}[XX^\top] = \begin{bmatrix} 1 & 0 \\ 0 & \Sigma \end{bmatrix}.$$

Hence, it follows that the formula holds:

$$X^* = p\mathbb{E}[XX^\top]^{-1}X = \frac{1}{R^f} - \frac{1}{R^f}\mathbb{E}[R^e]^\top\Sigma^{-1}(R^e - \mathbb{E}[R^e]).$$

Second Proof. First start by looking for an M given by $M = 1 - b^\top(R^e - \mathbb{E}[R^e])$ that prices correctly all excess returns, i.e. $\mathbb{E}[MR^e] = 0$. Then,

$$0 = \mathbb{E}[MR^e] = \mathrm{Cov}[R^e, M] + \mathbb{E}[R^e]\mathbb{E}[M] = \mathbb{V}[R^e](-b) + \mathbb{E}[R^e]$$

Hence we get that $b = \mathbb{V}[R^e]^{-1}\mathbb{E}[R^e]$. Thus,

$$M = 1 - b^\top(R^e - \mathbb{E}[R^e]) = 1 - \mathbb{E}[R^e]^\top\mathbb{V}[R^e]^{-1}(R^e - \mathbb{E}[R^e])$$

prices correctly all excess returns. However, it does not price the risk-free asset correctly, since $\mathbb{E}[M] = 1 \neq 1/R^f = p(\mathbb{1})$. But then M/R^f still price correctly all excess returns and the risk-free asset, hence prices all payoffs, thus

$$X^* = M/R^f = \frac{1}{R^f} - \frac{1}{R^f}\mathbb{E}[R^e]^\top\Sigma^{-1}(R^e - \mathbb{E}[R^e]),$$

is an SDF. □

2.3 Stochastic discount factors and mean-variance efficiency

The Hansen-Jagannathan formula is reminiscent of mean-variance efficient portfolios. Let us now clarify this connection. Consider an agent that can invest one dollar in the risk-free asset and the risky assets $R_1, \dots, R_N$ (assume as before, that their excess returns are a basis of the excess return space). Suppose he invests $w \in \mathbb{R}^N$ in the risky assets and he borrows at the risk-free rate for each of these investments, then the return on this portfolio is given by

$$R^f + w^\top (R - R^f) = R^f + w^\top R^e,$$

where $R = [R_1, \dots, R_N]^\top$. Now assume he has the objective of mean-variance optimization on this return, i.e. he wants to maximize the expected return of the portfolio subject to a bound on the variance. This can be formulated with a lagrangian as the following problem:

$$\max_w R^f + \mu^\top w - \frac{\gamma}{2} w^\top \Sigma w,$$

where $\mu = \mathbb{E}[R^e]$, $\Sigma = \mathbb{V}[R^e]$ and γ is a "risk aversion" coefficient. Differentiating with respect to w we get that at the optimum:

$$0 = \mu^\top - \gamma (w^*)^\top \Sigma$$

Thus, the optimal portfolio weights are given by

$$w^* = \frac{1}{\gamma} \Sigma^{-1} \mu = \frac{1}{\gamma} \mathbb{V}[R^e]^{-1} \mathbb{E}[R^e]. \tag{3}$$

The *mean-variance efficient portfolio* for γ is has following return:

$$R^{mv} = R^f + (w^*)^\top R^e = R^f + \frac{1}{\gamma} \mathbb{E}[R^e]^\top \mathbb{V}[R^e]^{-1} R^e \tag{4}$$

By changing the risk aversion γ we get the mean-variance efficient frontier, which is the frontier cone in the "mean - standard deviation" space (i.e. the space of (μ, σ) for all returns in the market).

We have found that the coefficients of b in the Hansen-Jagannathan formula (Equation 2) are the weights w^* of a mean-variance efficient portfolio (Equation 3) for some parameter γ. Thus, we get the following result.

Theorem 2.8. Assume a risk-free asset is traded. Given M the SDF on the payoff space, then $M = a + bR^{mv}$ where R^{mv} is a mean-variance efficient portfolio for some scalars $a, b \in \mathbb{R}$.

Conversely if R^{mv} then there are scalars $c, d \in \mathbb{R}$ such that $R^{mv} = c + dM$ is the SDF on the payoff space.

Proof. With what we have already done there's not much to do. Simply observe that from equations 2 and 4 we get:

$$\gamma(R^{mv} - R^f) = \mathbb{E}[R^e]^\top \mathbb{V}[R^e]^{-1} R^e = (1 + \mathbb{E}[R^e]^\top \mathbb{V}[R^e]^{-1} \mathbb{E}[R^e]) - R_f M.$$

From this we can solve R^{mv} in terms of M and viceversa. □

2.4 Stochastic discount factors and beta representations

Given M an SDF, we get a *single-factor* beta representation as follows. Given R_i a return of an asset, then

$$1 = \mathbb{E}[MR_i] = \mathrm{Cov}[R_i, M] + \mathbb{E}[M]\mathbb{E}[R_i]$$

Thus, dividing by $\mathbb{E}[M]$ we get

$$\mathbb{E}[R_i] = \gamma - \frac{\mathrm{Cov}[R_i, M]}{\mathbb{E}[M]} = \gamma + \left(\frac{\mathrm{Cov}[R_i, M]}{\mathbb{V}[M]}\right)\left(-\frac{\mathbb{V}[M]}{\mathbb{E}[M]}\right) = \gamma + \beta_i \lambda \tag{5}$$

where $\gamma = 1/\mathbb{E}[M]$ (which is R^f provided a risk-free asset is traded). The coefficient β_i is the slope of the regression of R_i onto M and an intercept (i.e. the linear projection onto the space generated by M plus a constant, also known as the best linear predictor). Notice that γ and λ are asset independent. What this says is that the expected return of every asset can be explained perfectly by how it covaries with the SDF. Expected returns should all lie on the line that starts at γ (the risk-free return) and slope λ, and β_i determines where on this line the return R_i should be. By Theorem 2.8 we know that the SDF and any mean-variance efficient portfolio are perfectly correlated, hence we can run regressions against any R^{mv} instead of the SDF and get a single-factor beta representation as well

$$\mathbb{E}[R_i] = \gamma + \beta_{i,mv} \lambda_{mv},$$

where $\beta_{i,mv}$ is the regression coefficient of R_i onto R^{mv}. Since R^{mv} is also a return, we can regress it onto itself and its beta must be one, thus we obtain that $\lambda_{mv} = \mathbb{E}[R^{mv}] - R^f$. Hence we get the formula

$$\mathbb{E}[R_i] - R^f = \beta_{i,mv}(\mathbb{E}[R^{mv}] - R^f).$$

2.5 Factor models

Usually in asset pricing the models are formulated as a beta-representation with multiple factors that explain the expected returns of assets:

$$\mathbb{E}[R_i] = \gamma + \beta_{i,1}\lambda_1 + \ldots \beta_{i,K}\lambda_K = \gamma + \beta_i^\top \lambda$$

where β_i are the regression coefficients of R_i onto the linear subspace generated by the factors $F_1, \ldots, F_K$ plus a constant. For example, the CAPM is a one factor model:

$$\mathbb{E}[R_i] = R^f + \beta_i(\mathbb{E}[R^m] - R^f).$$

where R^m is the return on the "market" portfolio. Another example is the Fama-French three-factor model (Fama and French 1993), which is given by:

$$\mathbb{E}[R_i] = R^f + \beta_{i,m}(\mathbb{E}[R^m] - R^f) + \beta_{i,smb}\mathbb{E}[F^{SMB}] + \beta_{i,hml}\mathbb{E}[F^{HML}],$$

where R^m is the "market" return, F^{SMB} is the "Small-Minus-Big" factor (constructed as a portfolio sort of small companies minus big companies; it is an excess return by construction), and F^{HML} is the "High-Minus-Low" factor (a portfolio sort of companies according to the book-to-market value ratio).

We now show that models of this type are equivalent to models to a restriction on the SDF to be on the span of the factors.

Theorem 2.9. Given the model

$$M = 1 - b^\top (F - \mathbb{E}[F]), \quad \mathbb{E}[MR^e] = 0, \tag{6}$$

then there exists $\lambda \in \mathbb{R}^K$ such that

$$\mathbb{E}[R_i^e] = \beta_i^\top \lambda, \tag{7}$$

where β_i are the regression coefficients of the excess returns onto F (without a constant) for every excess return R_i.

Conversely, if equation 7 holds for all excess returns then there is a vector $b \in \mathbb{R}^K$ that satisfies equation 6.

Proof. Take $F = [F_1, \ldots, F_K]^\top$ and $R^e = [R_1^e, \ldots, R_N^e]^\top$. From $\mathbb{E}[MR^e] = 0$, the bilinearity of the covariance and $\mathbb{E}[M] = 1$ it follows that

$$0 = \mathbb{E}[MR^e] = \mathrm{Cov}[R^e, M] + \mathbb{E}[R^e]\mathbb{E}[M] = \mathrm{Cov}[R, F](-b) + \mathbb{E}[R^e].$$

Thus,

$$\mathbb{E}[R^e] = \mathrm{Cov}[R^e, F]b = \mathrm{Cov}[R^e, F]\mathbb{V}[F]^{-1}(\mathbb{V}[F]b) = \beta^\top \lambda$$

where $\beta = \mathbb{V}[F]^{-1}\mathrm{Cov}[F, R^e] \in \mathbb{R}^{K\times N}$ are the regression betas of the excess returns of the assets onto the factors and $\lambda = \mathbb{V}[F]b \in \mathbb{R}^K$. The content of the theory is that the intercept is zero for all assets, hence the formula for the β holds. Conversely given the regression betas $\beta = \mathbb{V}[F]^{-1}\mathrm{Cov}[F, R^e]$ and $\lambda \in \mathbb{R}^K$, define $b = \mathbb{V}[F]^{-1}\lambda$, and going backwards through the computation we get that $M = 1 - b^\top(F - \mathbb{E}[F])$ satisfies $\mathbb{E}[MR^e] = 0$. □

Theorem 2.10. Given the model

$$M = a - b^\top F, \quad 1 = \mathbb{E}[MR], \tag{8}$$

then there exists $\gamma \in \mathbb{R}$ and $\lambda \in \mathbb{R}^K$ such that

$$\mathbb{E}[R_i] = \gamma + \beta_i^\top \lambda, \tag{9}$$

where β_i are the multiple regression coefficients of R_i onto F with a constant. Conversely, given γ and λ in a factor model such as 9, one can find a, b such that 8 holds.

Proof. Take $F = [F_1, \ldots, F_K]^\top$ and $R = [R_1, \ldots, R_N]^\top$. We can assume without loss of generality that $\mathbb{E}[F] = 0$ by changing a. Then,

$$1 = \mathbb{E}[MR] = \mathrm{Cov}[R, M] + \mathbb{E}[M]\mathbb{E}[R].$$

Thus, given that $\mathbb{E}[F] = 0$,

$$\mathbb{E}[R] = \frac{1}{\mathbb{E}[M]} + \mathrm{Cov}[R, F]\mathbb{V}[F]^{-1}\left(\frac{\mathbb{V}[F]b}{\mathbb{E}[M]}\right) = \frac{1}{\mathbb{E}[M]} + \left(\mathbb{E}[RF^\top]\mathbb{E}[FF^\top]^{-1}\right)\left(\frac{\mathbb{E}[FF^\top]b}{\mathbb{E}[M]}\right).$$

Hence it follows that

$$\mathbb{E}[R] = \gamma + \beta^\top \lambda, \tag{10}$$

where $\gamma = 1/\mathbb{E}[M] = 1/a$, $\beta = \mathbb{E}[FF^\top]^{-1}\mathbb{E}[FR^\top] \in \mathbb{R}^{K\times N}$ are the betas of a regression of the returns R against the factors F with a constant, and $\lambda = \mathbb{V}[F]\frac{b}{a}$. Conversely, given γ, λ we can define $a = 1/\gamma$ and $b = \frac{1}{\gamma}\mathbb{V}[F]^{-1}\lambda$, and check that $M = a - b^\top F$ is such that $\mathbb{E}[MR] = 1$. □

Remark 2.11. When the factors in 7 or 9 are excess returns of assets, then they can be regressed onto themselves and we get that $\lambda = \mathbb{E}[F]$.

We might ask ourselves why does the theory need to be based on a number of factors. In principle we can use all returns as factors and we know that there is an SDF given by the formula 2, or equivalently we have the formula 4 for the mean-variance efficient portfolios. The problem with these formulas is that they depend on population moments which are known to be very hard to estimate: means and variances of assets do change over time and we cannot simply assume they are independent and identically distributed. Moreover, if we have a large number of assets then the covariance matrix $\mathbb{V}[R^e]$ will typically be ill conditioned since we will have assets that are highly correlated to a portfolio of assets which closely "replicates" it. In other words, the principal components of this matrix typically has many eigenvalues which are close to zero. Then its inverse will have very large eigenvalues and small errors in the estimation of $\hat{\mathbb{V}}[R^e]$ will produce large errors in its inverse.

If instead we restrict ourselves to work with a small number of factors which are not highly correlated, then we don't have these problems and we can perhaps better estimate b and the regression β's.

2.6 Conditioning information

Let us now briefly discuss the role of conditioning information. Suppose we have now a filtered probability space $(\Omega, \Sigma, \mathbb{P}, (I_t)_t)$ where I_t represents the information available to markets participants when trading at time t for $t = 0, 1, \ldots$ (i.e. we still assume discrete time). At each time there are payoffs $X_t \in \underline{X}_t \subset mI_t$ in a payoff space (i.e. adapted and also assume with finite second moments) with prices $p_{t-1}(X_t)$. Moreover, prices are known ahead of time as their name suggests it, $p_{t-1}(X_t) \in mI_{t-1}$. An SDF $(M_t)_t$ then must satisfy first that it is adapted ($M_t \in mI_t$) and also that

$$p_{t-1} = \mathbb{E}[M_t X_t \mid I_{t-1}], \tag{11}$$

for all payoffs and times t. We can restate this condition in terms of returns or excess returns, for example assuming the risk free asset is tradable then equation 11 is equivalent to:

$$0 = \mathbb{E}[M_t R^e_t \mid I_{t-1}], \tag{12}$$

for all excess returns R_t^e, plus $1/R_t^f = \mathbb{E}[M_t \mid I_{t-1}]$. The condition on the risk free rate follows since $R_t^f \in mI_{t-1}$, i.e. it is known at time $t-1$.

By the law of iterated conditioning it follows that it must also satisfy the unconditional asset pricing equation:

$$0 = \mathbb{E}[\mathbb{E}[M_t R_t^e \mid I_{t-1}]] = \mathbb{E}[M_t R_t^e],$$

but this is not enough to guarantee 12. By definition of conditional expectation, equation 12 holds if and only if for all random variables Z_{t-1} which are I_{t-1}-measurable:

$$0 = \mathbb{E}[\mathbb{E}[M_t R_t^e \mid I_{t-1}] Z_{t-1}] = \mathbb{E}[M_t R_t^e Z_{t-1}]$$

This means that in order to get equation 12 we should, instead of just looking at the basic assets traded in the market, also look for dynamic or *managed portfolios*, and then the asset pricing equation is again in terms of unconditional moments. Let's say that there are $R_t^e = [R_{1,t}, \dots R_{N,t}]^\top$ excess returns at each time t, and $(Z_t)_t$ is an adapted process ($Z_t \in mI_t$) with values in $\mathbb{R}^N$. Then we can construct the managed portfolio $F_t = Z_{t-1}^\top R_t^e$ which is an excess return and from the above we must have:

$$0 = \mathbb{E}[M_t (R_t^e)^\top Z_{t-1}] = \mathbb{E}[M_t F_t] \tag{13}$$

If equation 13 holds for all managed portfolios then the conditional asset pricing equation 12 holds. Clearly, conditioning on all market information (i.e. all managed portfolios) is too much to ask, as this gives an infinite number of moment equations. Our hope is that by conditioning with a good set of managed portfolios, we can get a tractable and well performing model.

Now we can state more precisely what is an unconditional factor model as in equation 6. Suppose we have K factors $(F_t)_t$ which are excess returns of managed portfolios, i.e. $F_t = Z_{t-1}^\top R_t^e$ for an adapted process $(Z_t)_t$ with values in $\mathbb{R}^{N\times K}$ for all t.

Definition 2.12. An *unconditional (or fixed weight) asset pricing model* for excess returns is given by

$$M_t = 1 - b^\top (F_t - \mathbb{E}[F_t]), \quad 0 = \mathbb{E}[M_t R_t^e \tilde{Z}_{t-1}].$$

for all $\tilde{Z}_{t-1} \in mI_{t-1}$.

Importantly, the vector of weights b is not time dependant, however the factors are dynamic portfolios so M_t is time dependent and the weights of each basic asset change in time as $b_t = Z_{t-1}b$. As the second condition is intractable in practice, we typically replace it with a simpler set of moment conditions. For example $0 = \mathbb{E}[M_t F_t]$, i.e. we use the factors themselves as test assets, or possibly a larger set of managed portfolios.

3 Asset pricing with characteristics-based factors

3.1 Characteristics-based factor SDF

The setting of Kozak et al. (2020) starts with an $N \times 1$ vector R_t of excess returns for N stocks at time t. The model is given by H characteristics-based factors F_t, defined by an $N \times H$ matrix Z_{t-1} of assets characteristics, and then the factors are given by

$$F_t = Z_{t-1}^\top R_t.$$

We notice that the characteristics Z_{t-1} are observable at time $t-1$ as their name suggests, hence the factors are tradeable (or investable) portfolios. The j-th factor F_t^j is given by the inner product of the j-th column of Z_{t-1} and R_t, and thus we can interpret the entries of Z_{t-1}^{*j} as the *weights* of the factor portfolio.

The matrix Z_t in turn is defined as follows. We have H observable stock characteristics, which give a real number $c_{i,t}^j$ for each stock $i = 1, \ldots, N$. We rank them cross-sectionally, that is we sort the stocks according to each characteristic from 1 to N. Then we normalise all ranks by dividing by $N+1$, i.e.

$$rc_{i,t}^j = \frac{\text{rank}(c_{i,t}^j)}{N+1}.$$

Finally, we center these and divide by the sum of absolute deviations from the mean:

$$Z_{i,t}^j = \frac{rc_{i,t}^j - \bar{rc}_t^j}{\sum_{i=1}^N |rc_{i,t}^j - \bar{rc}_t^j|}$$

where $\bar{rc}_t^j = \frac{1}{N}\sum_{i=1}^N rc_{i,t}^j$. The resulting portfolios based on these transformed characteristics are zero-cost since $\sum_{i=1}^N Z_{i,t}^j = 0$, are insensitive to outliers in the original characteristics, and they have a fixed leverage as the absolute exposure is one, i.e. $\sum_{i=1}^N |Z_{i,t}^j| = 1$. The approach

differs from the standard practice of sorting stocks into deciles and then creating a portfolio long the 1^{st} decile and short the 10^{th} decile.

First Kozak et al. (2020) consider $H = 50$ "anomaly" characteristics known to have some predictive power on returns (at least in-sample). The second set of characteristics they consider are given by 68 financial ratios from the WRDS Industry Financial Ratios, supplemented by 12 past monthly returns, adding up to $H = 80$ managed portfolios. They also consider in each case adding the interactions between each pair of (basic) characteristics which for example for the $n = 50$ anomalies results in $H = \frac{1}{2}n(n-1) + 2n = 1325$ factors.

Notice that if we formulate our factor model for the SDF as

$$M_t = 1 - b^\top (F_t - \mathbb{E}[F_t]), \quad \mathbb{E}[M_t R_t] = 0, \tag{14}$$

as in equation 6, we find that our coefficients $b \in \mathbb{R}^H$ are constant in time, but the coefficients of each asset will depend in time and will be given by $b_{t-1} = Z_{t-1} b$. To see this just observe that

$$M_t = 1 - b^\top (F_t - \mathbb{E}[F_t]) = 1 - b^\top (Z_{t-1}^\top R_t - Z_{t-1}^\top \mathbb{E}[R_t]) = 1 - b_{t-1}^\top (R_t - \mathbb{E}[R_t]).$$

This is important because by means of these "managed portfolios" (since their composition changes every period, e.g. monthly, yearly) we're able to add a time dependency in response to conditioning information. Adding managed portfolios can in theory incorporate all the extra information in conditioning, and we can focus in estimating unconditional moments.

In equation 14 we are using all excess returns as test assets, however Kozak et al. (2020) consider only the factors themselves as tests assets arriving at the equation

$$\mathbb{E}[M_t F_t] = 0 \tag{15}$$

We can solve this equation as follows

$$0 = \mathbb{E}[M_t F_t] = \mathrm{Cov}[F_t, M_t] + \mathbb{E}[F_t]\mathbb{E}[M_t] = \mathbb{V}[F_t](-b) + \mathbb{E}[F_t],$$

recalling our definition of M_t in equation 14. Thus,

$$b = \mathbb{V}[F_t]^{-1}\mathbb{E}[F_t]. \tag{16}$$

3.2 Shrinkage estimator

According to Kozak et al. (2020), the main weakness of estimating b with the formula 16 using plug-in sample moments as:

$$b = \bar{\Sigma}^{-1}\bar{\mu},$$

where $\bar{\mu} = \frac{1}{T}\sum_{t=1}^{T} F_t$ and $\bar{\Sigma} = \frac{1}{T}\sum_{t=1}^{T}(F_t - \bar{\mu})(F_t - \bar{\mu})^\top$, comes from the uncertainty in the sample means, which is high even with long samples of returns. If the number of factors H is large, this estimator which is essentially running a regression of $\bar{\mu}$ onto the covariances of the factors will end up overfitting the sample with poor performance out of sample.

To avoid this overfitting, the authors introduce a Bayesian prior on the mean returns of factors which will shrink their means to zero by adding a regularisation term that produces a more robust estimator. Assume that the covariance matrix of the factors Σ is known. The family of priors introduced is given by:

$$\mu \sim \mathcal{N}(0, \frac{\kappa^2}{\tau}\Sigma^\eta), \tag{17}$$

where $\tau = \mathrm{tr}[\Sigma]$ and κ is a parameter that controls the "scale" of μ. To get an intuition on how this family of priors works, we diagonalise the covariance matrix $\Sigma = Q\Lambda Q^\top$, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_H)$ and Q orthogonal due to the spectral theorem. If we consider the principal component factors $P_t = Q^\top F_t$, then the prior on these is given by

$$\mu_P \sim \mathcal{N}(0, \frac{\kappa^2}{\tau}\Lambda^\eta).$$

Notice that the covariance matrix of the principal component factors is Λ, hence the Sharpe ratio distribution of the PCs is

$$\Lambda^{-\frac{1}{2}}\mu_P \sim \mathcal{N}(0, \frac{\kappa^2}{\tau}\Lambda^{\eta-1}). \tag{18}$$

This allows us to discard the idea of setting $\eta = 0$, as then the Sharpe ratios of PCs portfolios would be inversely proportional to their volatility. This would imply the existence of near arbitrage oportunities, as the Sharpe ratio of the higher-order PCs which typically are very small would be then extremely high. Also, $\eta = 1$ does not look like a plausible assumption as then Sharpe ratios of small-eigenvalue PCs would be of the same magnitude as the high-eigenvalue PCs.

Proposition 3.1. Under the prior in equation 17, the expected squared maximum Sharpe ratio is given by:

$$\mathbb{E}[\mu^\top \Sigma^{-1} \mu] = \sum_{j=1}^{H} \frac{\kappa^2}{\tau} \lambda_j^{\eta-1}.$$

Proof. As we observed in Theorem 2.8, $b = \Sigma^{-1}\mu$ are the weights of a mean variance efficient portfolio, thus we can compute the Sharpe ratio of $b^\top F_t$. It's easy to check $\mathbb{E}[b^\top F_t] = \mu^\top \Sigma^{-1} \mu$ and $\mathbb{V}[b^\top F_t] = \mu^\top \Sigma^{-1} \mu$, hence the Sharpe ratio is $\sqrt{\mu^\top \Sigma^{-1} \mu}$. Thus under the prior, the expected maximum Sharpe ratio is

$$\mathbb{E}[\mu^\top \Sigma^{-1} \mu] = \mathbb{E}[\mu_P^\top \Lambda^{-1} \mu_P] = \sum_{j=1}^{H} \frac{\kappa^2}{\tau} \lambda_j^{\eta-1}.$$

where the last equation follows from the fact that from equation 18, $\lambda_j^{-\frac{1}{2}} \mu_{P,j}$ are independent normals with mean zero and variance $\frac{\kappa^2}{\tau} \lambda_j^{\eta-1}$. □

Remark 3.2. Notice that if $\eta = 2$ then under the prior the square root expected maximum Sharpe ratio squared is $\sqrt{SR^2} = \sqrt{\mathbb{E}[\mu^\top \Sigma^{-1} \mu]} = \kappa$.

Proposition 3.3. Suppose μ has prior given by equation 17, then $\mathbb{E}[b^\top b] = \frac{\kappa^2}{\tau} \sum_{j=1}^{H} \lambda_j^{\eta-2}$.

Proof. Since $b = \Sigma^{-1}\mu$, then $b \sim \mathcal{N}(0, \frac{\kappa^2}{\tau} \Sigma^{-1} \Sigma^\eta \Sigma^{-1}) = \mathcal{N}(0, \frac{\kappa^2}{\tau} \Sigma^{\eta-2})$. Now observe that

$$b^\top b = \|b\|^2 = \|Q^\top b\|^2 = \|b_P\|^2,$$

since Q is orthogonal. Thus it suffices to compute $\mathbb{E}[\|b_P\|^2]$. Since $b_P \sim \mathcal{N}(0, \frac{\kappa^2}{\tau} \Lambda^{\eta-2})$, then the random variables $(b_P)_j$ for $j = 1, \ldots, H$ are independent normal. Moreover, $\mathbb{E}[(b_P)_j^2] = \frac{\kappa^2}{\tau} \lambda_j$ as the standardised $(b_P)_j^2$ are χ_1^2, which have expectation one. □

Now typically for a large number of factors, the smallest principal components will have eigenvalues which are very small. If $\eta < 2$ then then $\lambda_j^{\eta-2}$ will be very big for the smaller eigenvalues resulting in a very large expected 2-norm of b. In equilibrium, portfolio weights of optimal portfolios should be bounded. Thus setting $\eta \geq 2$ avoids the unrealistically high portfolio weights.

Proposition 3.4. Consider a linear regression model:

$$y = Xg + \varepsilon$$

where $\varepsilon \sim \mathcal{N}(0,\Sigma)$ with known Σ and assume a prior $g \sim \mathcal{N}(0,\Sigma_g)$. Then the posterior distribution of g given the data y is normal with mean given by

$$g_p = (X^\top \Sigma^{-1} X + \Sigma_g^{-1})^{-1} X^\top \Sigma^{-1} y \tag{19}$$

and posterior variance $\Sigma_p = (X^\top \Sigma^{-1} X + \Sigma_g^{-1})^{-1}$.

Proof. It is well known that the normal is self-conjugate, i.e. that the posterior distribution is again normal, so to the mean of the distribution is also the mode. Thus, we take minus the logarithm of the posterior distribution and find the minimum. By dropping constants, the problem becomes:

$$\arg\min_g \frac{1}{2}(y - Xg)^\top \Sigma^{-1}(y - Xg) + \frac{1}{2} g^\top \Sigma_g^{-1} g$$

Then by differentiating, we get:

$$0 = (y - Xg)^\top \Sigma^{-1}(-X) + g^\top \Sigma_g^{-1} = -y\Sigma^{-1}X + g^\top (X^\top \Sigma^{-1} X + \Sigma_g^{-1})$$

whence the formula follows. To get the posterior variance we need to compute the posterior distribution. Once we know the mean, this can be computed as follows:

$$\begin{aligned} f(g \mid y) \propto f(y \mid g) f(g) &= \exp\left((y - Xg)^\top \Sigma^{-1}(y - Xg) + \frac{1}{2} g^\top \Sigma_g^{-1} g\right) \\ \propto \exp\left(g^\top (X^\top \Sigma^{-1} X + \Sigma_g^{-1}) g + g^\top X^\top \Sigma^{-1} y\right) &= \exp\left(g^\top \Sigma_p^{-1} g + g^\top \Sigma_p^{-1} (\Sigma_p X^\top \Sigma^{-1} y)\right) \\ = \exp\left(g^\top \Sigma_p^{-1} g + g^\top \Sigma_p^{-1} g_p\right) &\propto \exp\left((g - g_p)^\top \Sigma_p^{-1} (g - g_p)\right) \end{aligned}$$

whence $g \mid y \sim \mathcal{N}(g_p, \Sigma_p)$ where $\Sigma_p = (X^\top \Sigma^{-1} X + \Sigma_g^{-1})^{-1}$ and $g_p = \Sigma_p X^\top \Sigma^{-1} y$. □

Proposition 3.5. Suppose μ has prior given by equation 17 for $\eta = 2$, and assume that $F_t - \mu \sim \mathcal{N}(0,\Sigma)$, then the posterior mean of b given a sample is:

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

where $\gamma = \frac{\tau}{\kappa^2 T}$. Moreover, the posterior variance of b is $\mathbb{V}[b] = \frac{1}{T}(\Sigma + \gamma I)^{-1}$.

Proof. If $\mu \sim \mathcal{N}(0, \frac{\kappa^2}{\tau}\Sigma^2)$ then the prior distribution of b is given by

$$b = \Sigma^{-1}\mu \sim \mathcal{N}(0, \frac{\kappa^2}{\tau}\Sigma^{-1}\Sigma^2\Sigma^{-1}) = \mathcal{N}(0, \frac{\kappa^2}{\tau} I).$$

Also, the likelihood of $\varepsilon = \bar{\mu} - \Sigma b = \bar{\mu} - \mu$ given b is

$$\varepsilon = \bar{\mu} - \mu \sim N(0, \frac{1}{T}\Sigma),$$

since $\bar{\mu} = \frac{1}{T}\sum_{t=1}^{T} F_t$. Thus, using Proposition 3.4 we get that,

$$\hat{b} = (\Sigma T \Sigma^{-1} \Sigma + \frac{\tau}{\kappa^2} I)^{-1} \Sigma T \Sigma^{-1} \bar{\mu} = \left(\Sigma + \gamma I\right)^{-1} \bar{\mu}.$$

and that the variance is $\left(\Sigma + \gamma I\right)^{-1}$. □

3.3 Representation as a penalised estimator

We now show that the Bayesian estimator can be obtained as a penalised estimator. Importantly, the penalty can be interpreted as one on the maximum model-implied Sharpe ratio $\gamma b^\top \Sigma b$.

Proposition 3.6. Let $\hat{b}$ be the estimator given by formula 20, then it is the solution of the following optimization problem:

$$\hat{b} = \arg\min_b (\bar{\mu} - \Sigma b)^\top (\bar{\mu} - \Sigma b) + \gamma b^\top \Sigma b.$$

Proof. By differentiating and setting the derivative to zero, we get that at the optimum:

$$0 = 2\Sigma(-\bar{\mu}^\top + b^\top(\Sigma + \gamma I))$$

Given that Σ is positive definite we get that the second term must be zero, then we get the formula for the Bayesian estimator $\hat{b} = (\Sigma + \gamma I)^{-1}\bar{\mu}$. □

Alternatively, the estimator can be obtained by minimizing the Hansen-Jagannathan distance subject to an L^2 penalty $\gamma b^\top b$:

Proposition 3.7. Let $\hat{b}$ be the estimator given by formula 20, then it is the solution of the following optimization problem:

$$\hat{b} = \arg\min_b (\bar{\mu} - \Sigma b)^\top \Sigma^{-1} (\bar{\mu} - \Sigma b) + \gamma b^\top b.$$

Proof. Same as before, if we differentiate we get

$$0 = -\bar{\mu}^\top + b^\top(\Sigma + \gamma I),$$

which directly gives the estimator. □

Moreover, the estimator is invariant under change of basis, in particular we can express it in the basis of principal components.

Proposition 3.8. Let $\hat{b}$ be the estimator given by formula 20, and consider the basis of principal component factors $P_t = Q^\top F_t$, where $\Sigma = Q\Lambda Q^\top$. Then,

$$\hat{b}_P = (\Lambda + \gamma I)^{-1}\bar{\mu}_P,$$

where $\hat{b}_P = Q^\top b$.

Proof. Recall that $\hat{b}$ is the solution of the optimization problem defined in proposition 3.6. Apply that change of variables $y = Q^\top x$ to this problem. Then, we get that

$$\hat{b}_P = \underset{b_P}{\arg\min}\ (\bar{\mu}_P - \Lambda b_P)^\top(\bar{\mu}_P - \Lambda b_P) + \gamma b_P^\top \Lambda b_P.$$

where $\hat{b}_P = Q^\top b$. As before, differentiating we get that $\hat{b}_P = (\Lambda + \gamma I)^{-1}\bar{\mu}_P$.

Second proof. Alternatively, we can show directly from the definition,

$$\hat{b}_P = Q^\top b = Q^\top(\Sigma + \gamma I)^{-1}\bar{\mu} = Q^\top(\Sigma + \gamma I)^{-1}Q\bar{\mu}_P = (\Lambda + \gamma I)^{-1}\bar{\mu}_P$$

since $(\Sigma + \gamma I)^{-1} = Q(\Lambda + \gamma I)^{-1}Q^\top$. □

3.4 Sparsity

While it is not the case that we can get a good SDF which is sparse in the factors, it might be possible to still get a good SDF which is sparse in PCs factors. As the authors argued economically, low-eigenvalue PCs should have small Sharpe ratios, which is why we choose the estimator 17 for $\eta = 2$ which shrinks low eigenvalue PCs coefficients to zero faster.

If we want to get a solution b which is sparse, then we can add an L^1 penalty to the problem in Proposition 3.7. Thus the sparse estimator is given by:

$$\hat{b} = \underset{b}{\arg\min}\ (\bar{\mu} - \Sigma b)^\top \Sigma^{-1}(\bar{\mu} - \Sigma b) + \gamma_2 b^\top b + \gamma_1 \|b\|_1, \tag{21}$$

where $\|b\|_1 = \sum_{j=1}^{H} |b_j|$. This estimator does depends on the basis chosen, so if we first change the basis to the PCs basis and then add the L^1 penalty to the coefficients in that basis, we get a different solution.

3.5 Covariance estimation uncertainty

Previously we assumed that the covariance matrix Σ is known, whereas in practice we have to replace it with its sample estimator $\bar{\Sigma}$. When the number of factors (H) is small compared to the number of observations (T), this is not a problem. However, in a high dimensional setting when H is of the same order as T, the sample covariance matrix behaves poorly in practice.

A well known method introduced in Ledoit and Wolf (2004) consists of shrinking the covariance matrix towards a target matrix, typically a scaled identity. Alternatively, one can use a Bayesian approach specifying a prior Wishart distribution for Σ^{-1}, however then the posterior mean of b has no analytic solution as in 20. The approach used by Kozak et al. (2020), in which they take the mean of the posterior distribution of Σ^{-1} given a Wishart prior $\Sigma^{-1} \sim \mathcal{W}(H, \frac{1}{H}\Sigma_0^{-1})$, gives a shrinkage:

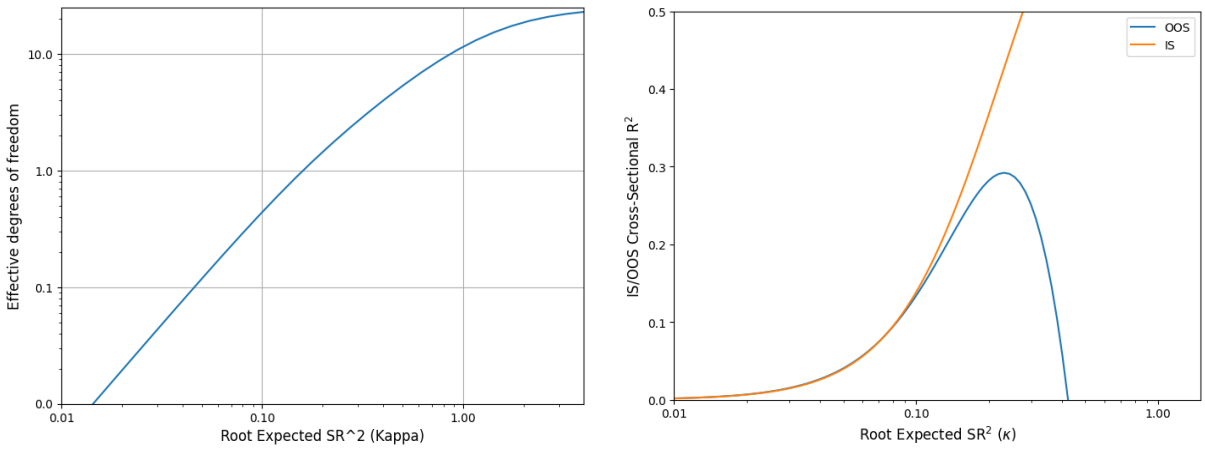
$$\hat{\Sigma} = a\Sigma_0 + (1-a)\bar{\Sigma}$$

where $\Sigma_0 = \frac{\text{tr}(\bar{\Sigma})}{H} I_H$, $\bar{\Sigma}$ is the sample covariance, and $a = \frac{H}{H+T}$ (thus $1-a = \frac{T}{H+T}$). Then b is obtained as a plug-in estimator where we replace Σ by $\hat{\Sigma}$ in equations 20 and 21.

4 Empirical results

4.1 Fama-French ME/BM portfolios

As a first test of the methodology introduced, Kozak et al. (2020) consider a low dimensional setting where we have 25 ME/BM-sorted portfolios. ME stands for the size of a company, i.e. the number of shares times the stock price, while BM is the book-to-market ratio, i.e. the ratio of the book value (BE) of a firm to its market value ($BM = BE/ME$). Fama and French (1993) consider bivariate portfolio sorts of these two characteristics which we use as a basis of factors. We expect that the excess returns of these factors when orthogonalised against the market component should be explained well by two factors similar to the SMB and HML factors of Fama and French. If the methodology is sound, we should be able to recover a sparse SDF that prices all excess returns appropriately.



(a) (b)

Figure 1: a) Shows the effective degrees of freedom as a function of the annualized Root Expected SR2 (κ) under the prior. b) displays the in-sample cross-sectional R^2 (orange) and OOS cross-sectional R^2 based on cross validation (blue).

We first consider the Bayesian estimator of b of Proposition 3.5. We have one free parameter, namely κ, which we have interpreted as the square root expected maximum Sharpe ratio squared in Proposition 3.1. We choose κ optimally with 3-fold cross-validation according to the out of sample cross-sectional R^2. In Figure 1 (b) we observe that the optimum is obtained at $\kappa \approx 0.231$. We also show the in sample Sharpe ratio which is increasing in κ (as $\kappa \to +\infty$, the regularisation goes to zero and we get the OLS solution). However the out of sample performance of these more flexible models is bad as the R^2_{OOS} is negative for large κ. Alternatively we can look at panel (a) where we plot the *effective degrees of freedom* as a function of κ which measures the model complexity defined as the trace of the hat matrix:

$$\mathrm{df}(\gamma) = \mathrm{tr}(\Sigma(\Sigma+\gamma I)^{-1}) = \sum_{j=1}^{H} \frac{\lambda_j}{\lambda_j+\gamma}$$

When there is no regularisation this is just H, the number of regressors. As γ increases, the effective degrees of freedom decrease and go eventually to zero. We can observe in Figure 1 (a) that $\mathrm{df}(\gamma) = 2$ corresponds roughly to $\kappa \approx 0.25$, so given our prior knowledge of the Fama-French factors we should expect such κ to have a good OOS performance.

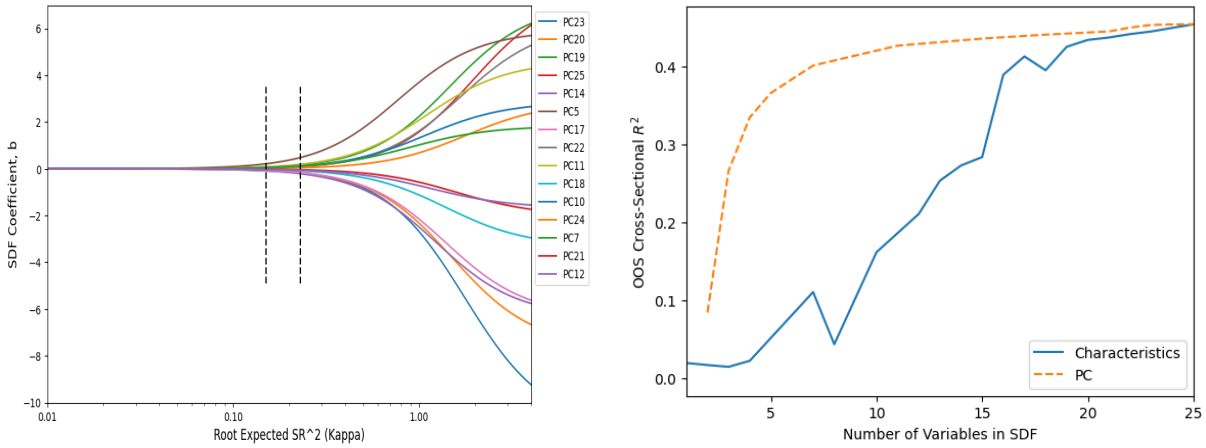
	b ($\kappa = 0.231$)	t-stat ($\kappa = 0.231$)	b ($\kappa = 0.15$)	t-stat ($\kappa = 0.15$)
SMALL HiBM	0.37	1.32	0.18	1.00
ME1 BM4	0.31	1.12	0.16	0.86
ME3 BM5	0.30	1.06	0.15	0.83
ME4 BM4	0.30	1.06	0.15	0.82
ME3 BM4	0.29	1.04	0.15	0.82
ME4 BM1	0.01	0.03	0.00	0.00
ME5 BM4	-0.05	0.18	-0.01	0.06
ME3 BM1	-0.16	0.57	-0.06	0.35
ME2 BM1	-0.22	0.78	-0.08	0.45
SMALL LoBM	-0.37	1.31	-0.14	0.78

Table 1: Left: Coefficient estimates and absolute t-statistics at the optimal value of the prior Root Expected SR^2 (κ, based on cross-validation) for the raw Fama-French 25 portfolios. Right: Coefficient estimates and absolute t-statistics at the optimum found on the paper ($\kappa \approx 0.15$). 10 portfolios with largest t-statistics are shown. Standard errors are calculated as in Proposition 3.5 and do not account for uncertainty in κ.

	b ($\kappa = 0.231$)	t-stat ($\kappa = 0.231$)	b ($\kappa = 0.15$)	t-stat ($\kappa = 0.15$)
PC1	-0.52	2.61	-0.32	2.04
PC5	0.48	1.74	0.21	1.16
PC2	0.42	1.72	0.21	1.22
PC14	-0.21	0.74	-0.09	0.49
PC11	0.19	0.69	0.08	0.45
PC23	-0.19	0.65	-0.08	0.42
PC20	-0.17	0.61	-0.07	0.40
PC19	0.17	0.60	0.07	0.39
PC17	-0.17	0.59	-0.07	0.39
PC10	0.12	0.44	0.05	0.29

Table 2: Left: Coefficient estimates and absolute t-statistics at the optimal value of the prior Root Expected SR^2 (κ, based on cross-validation) for the raw PCs portfolios. Right: Coefficient estimates and absolute t-statistics at the optimum found on the paper ($\kappa \approx 0.15$). 10 portfolios with largest t-statistics are shown. Standard errors are calculated as in Proposition 3.5 and do not account for uncertainty in κ.

Table 1 shows the coefficient estimates for the optimal level of regularisation ($\kappa = 0.231$) as well as the one obtained in Kozak et al. (2020). We display the 10 portfolios with largest absolute coefficients on the SDF as well as their t-statistics. The table shows that the optimal SDF assigns positive weights to small and value portfolios and shorts growth and large portfolios, as done by Fama and French. Table 2 shows the same but for the principal component portfolios. Since the Bayesian estimator with L^2 regularisation is rotation invariant, the solution we obtain is the same SDF in a different basis. We find that PC1, PC5 and PC2 have the largest and most significant coefficients at the optimal level, similar to the findings of the authors.

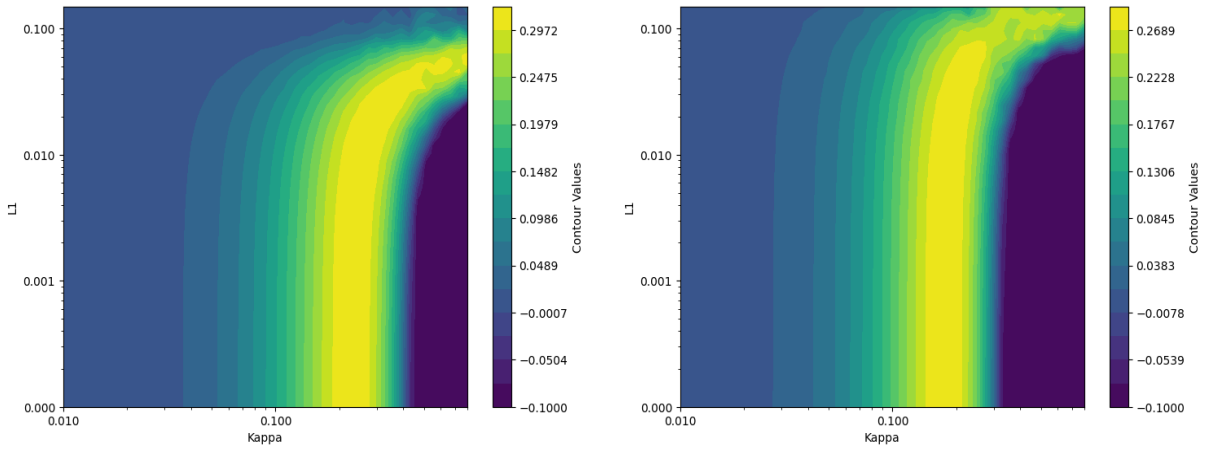


(a) (b)

Figure 2: L^2 **coefficient paths and sparsity (Fama-French 25 ME/BM portfolios)**. Panel (a) plots paths of coefficients as a function of the prior Root Expected SR^2 (κ) for the 25 Fama-French PCs portfolios. Labels are ordered according to absolute values of coefficients (descending) at the right edge of the plot, which corresponds to the OLS solution. A vertical line shows the shrinked coefficients for the optimal value κ found by us and the authors respectively. 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).

Figure 2 (a) displays the coefficients of for the SDF of the PCs portfolios as a function of κ. As $\kappa \to 0$ the coefficients go to zero as well, whereas for $\kappa \to +\infty$ the solution is the OLS estimator. Notice that the the OLS solution without regularisation has very large coefficients

for many small eigenvalue portfolios while for the optimal values only high eigenvalue PCs have large coefficients. Figure 2 (b) plots the maximum cross-sectional R^2_{OOS} that can be achieved by a sparse model with n factors. The blue line displays the R^2_{OOS} for sparse solutions in the characteristic portfolios, whereas the dashed orange line is in the PCs portfolios. We observe that in the PCs space we can obtain a high R^2_{OOS} with a small number of factors (e.g. less than 5) whereas the performance in the characteristic portfolios space deteriorates very rapidly for SDF with less than 15 factors.

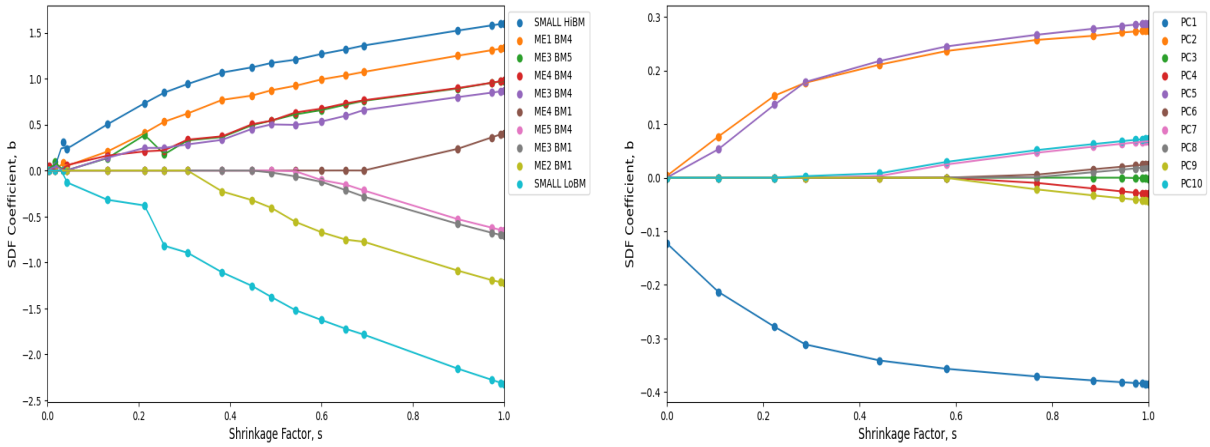


(a) (b)

Figure 3: **Sparse Model Selection (Fama-French 25 ME/BM portfolios)**. OOS cross-sectional R^2 for families of models that employ both L^1 and L^2 penalties simultaneously using 25 Fama-French ME/BM sorted portfolios (Panel a) and 25 PCs based on Fama-French portfolios (Panel b). We quantify the strength of the L^2 penalty by prior Root Expected SR2 (κ) and the strength of the L^1 penalty by their γ_1 coefficient in equation 21. Warmer (yellow) colors depict higher values of OOS R^2. Both axes are plotted on logarithmic scale.

Then, we show how the cross validation with L^1 regularisation on both parameters γ_1 and γ_2 works in figure 3 (a) with a contour map. Warmer colours represent higher R^2_{OOS} values. We see that L^2 regularisation is fundamental as models without small κ (i.e. large γ) perform very poorly (right edge of the plot). The model performs typically well for γ_1 small to moderate but when large performance is decreased. Figure 3 (b) shows that greater sparsity can be achieved in the PCs space without compromising the performance of the model. The top left region of high regularisation is essentially flat on zero and thus we converge to the

CAPM SDF. The vertical shape of Figure 3 (b) shows that there's a small cost to pay when dropping most of the small PCs from the SDF.



(a) (b)

Figure 4: L^1 **coefficient paths for the optimal model (Fama-French 25 ME/BM portfolios)**. Paths of coefficients as a function of shrinkage factor s based on the optimal (dual-penalty) sparse model that uses 25 for Fama-French ME/BM sorted portfolios (Panel a) and 25 PCs based on Fama and French portfolios (Panel b). Labels are ordered according to the vertical ordering of estimates at the right edge of the plot. In Panel b coefficient paths are truncated at the first 15 variables.

Lastly, Figure 4 shows how the coefficients are shrank as we increase the L^1 penalty, leaving the L^2 penalty fixed at its optimal value. The x-axis represents the ratio between the 1-norm of the coefficients over the optimal coefficients. As observed in Figure 4 (a) The resulting SDF is long small and value stocks, just like the Bayesian SDF. In Figure 4 (b) we observe that small eigenvalue PCs portfolios are shrunk much faster than high eigenvalue ones and results in a model with essentially three PCs factors. In conclusion, the method tends to recover an SDF that is closely related to the SDF implied by Fama and French (1993). However, the main advantage of this methodology comes when dealing with a large number of characteristics and unknown factors, where classic techniques are inadequate of ordinary least squares are inappropriate.

4.2 50 anomaly characteristics

We now consider 50 anomaly characteristics known to predict returns from the asset pricing literature and use the methodology introduced to construct an SDF. The complete list of anomalies can be found in the internet appendix to the paper Kozak et al. (2019). We found that the optimal shrinkage is at about $\kappa \approx 0.22$, same as the authors, see figure 6. However, we don't understand why we get a smaller R^2_{IS} than R^2_{OOS} for values of κ below the optimum.

In table 3 we observe the 10 largest coefficients of the SDF as well as their t-statistics. We found that the largest coefficients are associated to industry relative reversals (low volatility), industry momentum reversals, industry relative reversals, seasonality, earning surprises, etc. We also find in table 4 that PC5, PC1, PC4 and PC2 are the ones with largest coefficients consistent with the economic intuition.

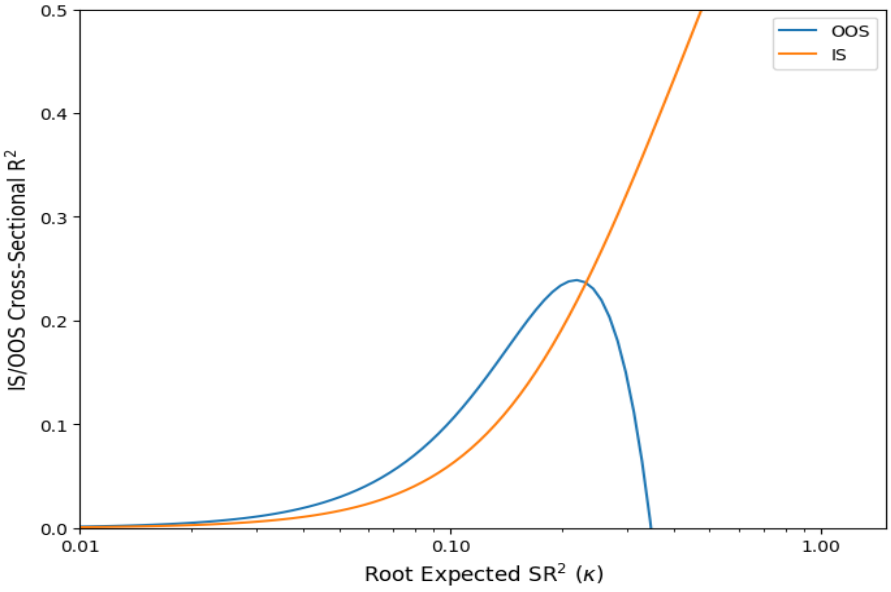


Figure 5

Figure 6: In-sample (orange) and out-of-sample (blue) cross-sectional R^2 for the 50 anomaly characteristics model.

	b ($\kappa = 0.219$)	t-stat ($\kappa = 0.219$)
Industry Rel. Rev. (L.V.)	-0.543693	2.781587
Ind. Mom-Reversals	0.316530	1.618897
Industry Rel. Reversals	-0.273105	1.396818
Seasonality	0.214621	1.097627
Earnings Surprises	0.184950	0.945860
Return on Market Equity	0.181511	0.925420
Value-Profitability	0.178857	0.913290
Composite Issuance	-0.150587	0.768426
Investment	-0.148822	0.759694
Momentum (12m)	0.135594	0.692750
Earnings/Price	0.134748	0.686838

Table 3: Coefficient estimates and absolute t-statistics at the optimal value of the prior Root Expected SR^2 ($\kappa = 0.219$, based on 3-fold cross-validation).

	b ($\kappa = 0.219$)	t-stat ($\kappa = 0.219$)
PC5	-0.591946	3.084821
PC1	0.435587	2.754067
PC4	0.261495	1.368606
PC2	0.240874	1.327756
PC14	0.246378	1.254655
PC9	-0.242013	1.237487
PC15	0.172033	0.875401
PC18	0.164441	0.834095
PC12	-0.147843	0.754960
PC6	0.132509	0.686164
PC10	-0.130348	0.666267

Table 4: Coefficient estimates and absolute t-statistics at the optimal value of the prior Root Expected SR^2 ($\kappa = 0.219$, based on 3-fold cross-validation) for the PCs portfolios.

References

[Coc09] John Cochrane. *Asset pricing: Revised edition*. Princeton university press, 2009.

[FF93] Eugene F Fama and Kenneth R French. "Common risk factors in the returns on stocks and bonds". In: *Journal of financial economics* 33.1 (1993), pp. 3–56.

[KNS19] Serhiy Kozak, Stefan Nagel, and Shrihari Santosh. "Internet Appendix to Shrinking the Cross-Section". In: (2019). URL: https://bpb-us-w2.wpmucdn.com/voices.uchicago.edu/dist/f/575/files/2020/07/SCSIA.pdf.

[KNS20] Serhiy Kozak, Stefan Nagel, and Shrihari Santosh. "Shrinking the cross-section". In: *Journal of Financial Economics* 135.2 (2020), pp. 271–292.

[LW04] Olivier Ledoit and Michael Wolf. "A well-conditioned estimator for large dimensional covariance matrices". In: *Journal of multivariate analysis* 88.2 (2004), pp. 365–411.