Erasmus University Rotterdam

Erasmus School of Economics

Bachelor Thesis BSc^2 Econometrics and Economics

Nonlinearities in the cross-section of asset returns

Final version
July 4th 2021

Author
Jody Lesterhuis (470411)

Supervisor
dr. M. Grith

Second assessor
dr. M.D. Zaharieva

Abstract

This paper investigates if and to what extent relaxing the linearity assumption of the method of Kozak et al. (2020) that the Stochastic Discount Factor is a linear combination of factors improves the explanation of the cross-section of asset returns. This is investigated by setting the Generalized Linear Model of Gu et al. (2020) in a risk pricing framework. Namely, the basis functions from this model are used to approximate nonlinear factors. The paper finds little empirical evidence that a flexible functional form improves the out-of-sample performance of the model. However, the model recovers that raw portfolio returns do not allow for a sparse Stochastic Discount Factor (SDF) representation, but transforming the portfolios into their Principal Components first allows for a sparse SDF representation.

Contents

1 Introduction

Explaining the cross-section of asset returns by only a few predictors has been a widely discussed topic in asset pricing literature. Since the number of stock return predictors has grown steadily over the past decades, selecting only a few predictors to explain the cross-sectional variation of stock returns has become obsolete. Also, it is not necessarily desirable anymore because the recent introduction of machine learning techniques in asset pricing has made the use of high-dimensional datasets containing many characteristics more convenient. Machine learning can avoid overfitting when including a high number of predictors, a problem that occurs using other, more classical econometric techniques.

Researchers that attempt to account for the cross-sectional variation in stock returns by using a large set of characteristics are for example Light et al. (2017), Feng et al. (2020) and Kozak et al. (2020). Kozak et al. (2020) disagree with the vision that the cross-section of asset returns can be explained by only a few factors, an idea that is widely born in previous research. In contrast, they propose a new machine learning based estimation procedure that allows for the consideration of many characteristics but can also return a good characteristic-sparse representation if that is desirable.

Next to Kozak et al. (2020), Gu et al. (2020) offered research that lays out and compares several machine learning methods that can be used to price assets. Whereas the objective of Gu et al. (2020) lies in the estimation of risk premia, Kozak et al. (2020) focus on the estimation of risk prices. The risk premium measures expected return above the risk-free rate, that is, how much a stock characteristic is exposed to variation in excess returns. The risk price is the extent to which a characteristic helps in pricing assets by contributing to the variation in the SDF (Kozak et al., 2020).

Albeit Kozak et al. (2020) establish that their proposed model performs significantly better out-of-sample than characteristic-sparse models, one of the limitations of the paper is that a linear functional form is assumed to hold. Freyberger et al. (2020), who suggest a nonparametric method that focuses on assessing which characteristics add additional information to the cross-section of asset returns, note that assuming linearity is a rather strong assumption and is sensitive to outliers. Furthermore, empirical research has shown that the characteristics and returns appear to be nonlinearly related (Fama and French, 2008). These findings point towards the need for more investigation into nonlinear relationships between stock returns and the SDF. Therefore, my extension lies in investigating whether relaxing the linearity assumption in the methods proposed by Kozak et al. (2020) still results in good out-of-sample performance. Kozak et al. (2020) essentially make two linear assumptions. The first one is that the risk prices are a linear combination of characteristics. The second one is that the SDF is linear in factors, where the factors are portfolios of stocks constructed based on characteristics. For the scope of this research, I investigate the relaxation of the second assumption. Hence, I assume that the SDF is a nonlinear function of the factors. The nonlinear functional form is approximated using the basis functions proposed in the Generalized Linear Method from Gu et al. (2020). Thus, the investigation answers the question if and to what extent relaxing the linearity assumption of the method of Kozak et al. (2020) that the SDF is a linear combination of factors improves the explanation of the cross-section of asset returns. To answer this question, the following subquestions are answered throughout the research:

1. How can the risk prices be estimated when it is assumed that the SDF is nonlinear in factors?
2. How can the out-of-sample R^2 be calculated for the nonlinear functional form?
3. How does the out-of-sample performance of the nonlinear model compare to the linear model?

I hypothesise that the method including a nonlinear functional form has better out-of-sample performance because of the earlier mentioned arguments by Freyberger et al. (2020) and Fama and French (2008). However, to safely draw a conclusion, all methods should be compared on an equal basis. This is done by comparing the out-of-sample performance of both models via the cross-sectional R^2 using K-fold cross-validation while imposing L^1 and L^2 penalties to allow for sparsity and impose shrinkage, respectively. For this purpose, two datasets are used. The first one considered is the Fama-French 25 Size and Book-to-Market sorted portfolios from 1926 until 2017 (French, 2020), and the second is the 50 anomaly zero-investment long-short portfolios from 1973 until 2017 (Kozak et al., 2020).

The paper finds little empirical evidence that a flexible functional form improves the out-of-sample performance compared to the linear model. However, the paper does find that the raw portfolio returns do not allow for a sparse SDF representation, but transforming the portfolios into their Principal Components first does allow for a sparse SDF representation.

The contribution to the literature of this paper is two-fold. First, the paper contributes to the recently started empirical investigation in using machine learning models in asset pricing. More specifically, the paper extends the research in applying machine learning to estimate risk prices by using the Generalized Linear Model of Gu et al. (2020) in a risk pricing framework. This is also an extension to the paper of Kozak et al. (2020), which assume a linear functional form. Secondly, this paper empirically applies and evaluates a recently introduced machine learning method, a shrinkage technique based on an economically motivated prior. This allows me to assess the performance and robustness of this model critically.

The research is economically relevant because increased robustness of cross-sectional return predictions and the opportunity to base the return predictions on many characteristics available in the finance literature simultaneously can help asset managers to make better decisions regarding the construction of their portfolios. This could lead to increased Sharpe ratios on the managed portfolios, leading to increased returns for investors.

The remainder of this paper is divided into the following sections. Section 2 lays out related existing literature. Section 3 shows the methods used in this research. Section 4 presents the data used for the empirical application, of which the results are discussed in section 5. The paper concludes in section 6.

2 Literature

In this section, the existing literature in the field of asset pricing is touched upon. In the first part, a general overview of the existing literature is presented. In the second part, I dive into research applying machine learning techniques to asset pricing, and I especially focus on nonlinear models.

The Capital Asset Pricing Model (CAPM) proposed by Sharpe (1964), with contributions from Lintner (1965) and Mossin (1965), forms the cornerstone of the asset pricing literature. The CAPM models a linear relationship between the expected return of a stock and systematic risk, measured by the market risk premium (excess returns beyond the risk-free rate). This model is later extended by Fama and French (1993), who added A Small-minus-Big (SMB) and High-minus-Low (HML) factor to the asset pricing model.[1] This resulted in a wide appliance of the three-factor model but also gave rise to the discussion and evaluation of many other factors that could possibly summarize the cross-section of asset returns. To only name a few examples, Hou et al. (2015) proposed a four-factor model that next to the three Fama-French factors includes a factor of average returns on low profitability stocks. Fama and French (2015) extended their own three-factor model with factors representing the profitability and investment patterns. Barillas and Shanken (2018) argue that a model including a momentum factor on top of the five-factor model beats all previously mentioned models. These examples are only a small fraction of the potential factors considered in the asset pricing literature. Thus, the question which factors are relevant to include in the asset pricing model to explain the cross-section of asset returns accurately has been an ongoing debate in the literature.

Although explaining the cross-section of asset returns by only a few factors has attractive benefits such as clear interpretability or lacking overfitting, considering a large number of factors simultaneously has become a new strand of literature in recent years. Lewellen (2015) forms a starting point in this strand of literature by considering 15 characteristics to explain the cross-section of expected stock returns using Fama-MacBeth regressions.[2] Furthermore, Light et al. (2017) consider an even larger set of 26 characteristics and use partial least squares to estimate the cross-section of stock returns.

Next to using (variations of) ordinary least squares regressions, there have been proposed methods using Principal Component Analysis (PCA) to explain the cross-section of stock returns. Fan et al. (2016) propose a Projected PCA method, in wich PCA is applied to the projected data matrix onto a linear span of stock predictors. This helps to reduce noise when the dimensionality of the dataset becomes large. Moreover, Kelly et al. (2019) combine PCA with observable characteristics that are used as instruments for unobservable dynamic loadings. This results in an Instrumented PCA method that allows for both latent factors and time-varying loadings because the loadings now depend linearly on the characteristics.

While in the above-named methods the objective is to find out which stock characteristics best explain the cross-section of returns, Kozak et al. (2020) take a rather different perspective. First of all, they question whether it is economically rational to explain the cross-section of asset returns by only a few factors. They argue that it is simply impossible to state beforehand that the cross-section of expected returns can be explained by only a few factors, as many characteristics contribute individually towards the expected return. However, they find that applying a Principal Component (PC) transformation to this set of characteristics allows for a PC characteristic-sparse representation of SDF. This result is driven by a finding from Kozak et al. (2018), which state that if near arbitrage opportunities do not exist,

[1] The SMB factor represents the average return on three small-capitalization portfolios minus the average return on three big capitalization portfolios and the HML factor represents the average return on two value portfolios minus the average return on two growth portfolios (Fama and French, 1993).

[2] Fama and MacBeth (1973)

factors that capture substantial risk premia, which is the case for the first few PCs, explain a large part of the co-movement of the returns. To take into account a large set of stock return predictors, Kozak et al. (2020) construct a novel type of SDF. They consider a Bayesian approach that allows setting an economically motivated prior which, in combination with regularization procedures, avoids overfitting and leads to good out-of-sample performance.

Next to Kozak et al. (2020), other papers aim to find a solution to deal with a complex and large set of characteristics. Feng et al. (2020) propose a model selection method, a bootstrap procedure, to determine which factors contribute significant additional information above a high-dimensional set of existing factors and therefore should be included. They note that "A fundamental task facing the asset pricing field today is to bring more discipline to the proliferation of factors" (Feng et al., 2020, p. 1327). Similar to Kozak et al. (2020) and in contrast to the majority of papers in the field of asset pricing, they focus on the estimation of risk prices rather than risk premia.

Kozak et al. (2020) and Feng et al. (2020) are examples of papers in which machine learning techniques are used to explain the cross-section of stock returns while possibly accounting for a large number of observed characteristics. Machine learning is attractive because it can deal with high-dimensional datasets while implementing regularization methods to avoid overfitting (Gu et al., 2020). The paper of Gu et al. (2020) gives an overview of and comparison between most of the existing machine learning methods in asset pricing to recover which method is best in forecasting asset risk premia. Next to using machine learning in the form of regularization methods to deal with the high dimensionality problem, Gu et al. (2020) also propose machine learning methods that can capture nonlinearities. One of the models they consider is a Generalized Linear Model regularized by the group LASSO, which allows to model expected returns as a nonlinear function of characteristics. Freyberger et al. (2020) use a similar approach. They use a nonparametric estimation procedure, an adaptive group LASSO, to select characteristics that best explain the cross-section and estimate how they affect conditional expected returns while allowing for high-dimensional datasets and not imposing a strong functional form. The following papers also use machine learning to capture nonlinearities. Moritz and Zimmermann (2016) Introduce tree-based conditional portfolio sorts to evaluate which variables, out of a large set of stock return predictors and their interactions, contribute independent information to the prediction of stock returns. Bryzgalova et al. (2020) also use a tree-based method. They introduce a tree-based asset pricing model to estimate the SDF while extending the proposed prior of Kozak et al. (2020). They impose not only LASSO and Ridge shrinkage but also shrink the estimated mean towards the average return. Gu et al. (2021) propose a latent factor conditional asset pricing model in which they use an autoencoder neural network to price assets. The neural network shrinks the returns into a low-dimensional set of factors while allowing the covariates to map nonlinear into factor loadings.

3 Methodology

This section describes the methods used in this paper. First, the methodology for the replication part of this paper is laid out. Secondly, the methodology for the extension is presented.

3.1 Asset pricing model

Many models in the finance literature find their origin in the asset pricing definition, in the spirit of Cochrane (2005). To model the SDF, the price of an asset is defined as

$$P_{t-1} = \mathbb{E}_t \left[M_t X_t \right], \tag{1}$$

where P_{t-1} is $N \times 1$ vector of the prices of N assets at any point in time $t-1$. M_t is the $N \times 1$ SDF, X_t is the $N \times 1$ vector of pay-offs of the assets at time t, and the subscript t of the expectation indicates that the expectation is conditional on all information known to the investor at time $t-1$. This equation states that the price of an asset at time $t-1$ equals the expected payoff of that asset at time t discounted by the SDF M_t, given all information available until time t. The vector of gross returns of the assets equals $R_t = X_t / P_{t-1}$. Thus, dividing Equation (1) by P_{t-1} allows us to express the asset pricing equation in terms of gross returns

$$1 = \mathbb{E}_t \left[M_t R_t \right]. \tag{2}$$

Using the fundamental asset pricing assumption that there do not exist arbitrage opportunities, the above equation can be rewritten in terms of excess returns, $R_t^e = R_t - R_t^f$, which equals the return of the assets over the risk-free rate:

$$0 = \mathbb{E}_t \left[M_t R_t^e \right]. \tag{3}$$

Excess returns rather than gross returns are most often used in the finance literature because the effect a characteristic has on a particular stock is isolated from the effect that characteristic has on all stocks in the market. For the replication part of the paper, it is assumed, similar to Kozak et al. (2020) and many other papers in the finance literature, that the SDF is a linear combination of excess returns,

$$M_t = 1 - b_{t-1}' (R_t^e - \mathbb{E} \left[R_t^e \right]), \tag{4}$$

where M_t is the SDF, R_t^e is the $N \times 1$ vector of excess returns at time t, and b_{t-1} is the vector of SDF loadings, also the vector of risk prices.

3.2 Characteristic-based factor SDF

Expressing the SDF as a characteristic-based factor model allows the use of financial ratios and characteristics as factors in the SDF to price assets. This enables us to determine which characteristics are the most important in explaining the cross-sectional variation of asset returns. The characteristic-based factor SDF used in this paper follows the definitions in line with Kozak et al. (2020). The risk prices in a characteristic-based SDF are defined as

$$b_{t-1} = Z_{t-1} b, \tag{5}$$

where Z_{t-1} is an $N \times H$ matrix of asset characteristics at time $t-1$ and b is a $H \times 1$ vector of time-invariant coefficients. Expressing the loadings in this manner allows to contribute the time variation in the loadings

to the characteristics. Substituting these loadings in Equation (4) and defining the characteristic-based factor returns as $F_t = Z'_{t-1}R^e_t$ gives

$$M_t = 1 - b'(F_t - \mathbb{E}[F_t]), \tag{6}$$

in which the SDF is a linear function of F_t. I focus on the unconditional asset pricing equation $\mathbb{E}[M_t F_t]$ instead of the conditional one because it is more convenient for estimation as noted by, amongst others, Hansen and Jagannathan (1991). Inserting Equation (6) in the unconditional asset pricing equation and solving for b gives[3]

$$b = \Sigma^{-1}\mu, \tag{7}$$

where $\Sigma^{-1} = \mathbb{E}[(F_t - \mathbb{E}[F_t])(F_t - \mathbb{E}[F_t])']$ is the $N \times N$ covariance matrix of F_t and $\mu = \mathbb{E}[F_t]$ is the $H \times 1$ mean vector of F_t. Thus, b can be expressed in terms of the mean and (co)variance of the factors.

3.3 Bayesian estimator

According to Kozak et al. (2020), using an ordinary cross-sectional regression of mean returns on the covariances of returns and factors to estimate Equation (7) results in a very imprecise estimate of b and bad out-of-sample performance. This result is driven by the fact that the factor means are regressed on the covariances of the factors to get an estimator of b, and the estimates of these factor means are often noisy. Hence, regressing a large number of variables in the cross-section will falsely overfit this noise. To prevent this, an economically motivated prior belief about the relationship between the factor means and variances is set to shrink the SDF coefficients towards zero (Kozak et al., 2020). Furthermore, to isolate the uncertainty about factor means, the covariance matrix of the factors, Σ, is assumed to be known. The following family of priors, which appeared in previous asset pricing research and allows to account for the fact that the factor means and variances are related, is considered:

$$\mu \sim \mathcal{N}(0, \frac{\kappa^2}{\tau}\Sigma^\eta). \tag{8}$$

Here, $\tau = \text{tr}[\Sigma]$, κ is a constant that controls the "scale" of μ, and η is a parameter that controls the "shape" of of the prior. This is the most important parameter, as it shows the relationship between the factor mean and variance we believe to hold under the prior. In Kozak et al. (2020), it is shown that the novel parameter specification of $\eta = 2$ is both economically motivated and deviates least from prior assumptions in previous literature, such as Pastor and Stambaugh (2000). This results in a prior belief

$$b \sim \mathcal{N}(0, \frac{\kappa^2}{\tau}\mathcal{I}_N) \tag{9}$$

about the SDF coefficients. Combining this prior belief with the sample mean $\bar{\mu}$, assuming a multivariate normal distribution, results in the posterior estimator of b[4]:

$$\hat{b} = (\Sigma + \gamma\mathcal{I}_N)^{-1}\bar{\mu}, \tag{10}$$

[3]The derivation can be found in appendix A.

[4]The algorithm of Kozak et al. (2020) is used to calculate $\hat{b}$.

where $\gamma = \frac{\tau}{\kappa^2 T}$ and $\mathcal{I}_N$ is an $N \times N$ identity matrix. The variance of b equals $\text{Var}(b) = \frac{1}{T}(\Sigma + \gamma \mathcal{I}_N)^{-1}$. To obtain Equation (10), the posterior mean of μ is needed to calculate $\hat{b}$. The posterior mean is based on the formula of the conjugate multivariate normal prior. That is, $\hat{\mu} = \left(\Sigma_0^{-1} + \ T\Sigma^{-1}\right)^{-1} \left(\Sigma_0^{-1}\mu_0 + T\Sigma^{-1}\bar{\mu}\right) = \left(\Sigma + \gamma\Sigma^{(2-\eta)}\right)^{-1} \Sigma\bar{\mu}$, where $\mu_0 = 0$ and $\Sigma_0 = \frac{\kappa^2}{\tau}\Sigma^{\eta}$. Plugging $\hat{\mu}$ in $\hat{b} = \Sigma^{-1}\hat{\mu}$ and using that $\eta = 2$ gives the desired formula of $\hat{b}$. The choice of $\eta = 2$ results in a root expected maximum squared Sharpe ratio of $\mathbb{E}\left[\mu\Sigma^{-1}\mu\right]^{1/2} = \kappa$. Thus, γ has an implicit economic interpretation.

It can be shown that the Bayesian estimator can be written as a common machine learning estimator. A machine learning estimator usually solves an optimization problem constrained with some penalty function. Hence, we want to write the estimator from Equation (10) in such a way as well. This is possible because when maximizing the models' cross-sectional R^2 extended by a penalty on the maximum squared Sharpe ratio,

$$\hat{b} = \arg\max_b \{(\bar{\mu} - \Sigma b)'(\bar{\mu} - \Sigma b) + \gamma b'\Sigma b\}, \tag{11}$$

the estimator found from in this maximization problem is equivalent to the estimator found in Equation (10). Also, the estimator resulting from minimizing the HJ-distance (Hansen and Jagannathan, 1991) extended by an L^2 norm penalty,

$$\hat{b} = \arg\min_b \{(\bar{\mu} - \Sigma b)'\Sigma^{-1}(\bar{\mu} - \Sigma b) + \gamma b'b\}, \tag{12}$$

is equivalent to the estimator found in Equation (10). It now becomes apparent that the above two equations are similar to a ridge regression (Hastie et al., 2011).

Although an L^2 penalty shrinks the SDF coefficients substantially towards zero, it does not recover a sparse SDF representation. To recover a sparse SDF, Equation (12) can be extended with an L^1 penalty, $\gamma_1 \sum_{i=1}^{H} |b_i|$. This causes some coefficients to be set exactly to zero. Thus, combining both the L^1 and L^2 penalty results in the minimization problem

$$\hat{b} = \arg\min_b \{(\bar{\mu} - \Sigma b)'\Sigma^{-1}(\bar{\mu} - \Sigma b) + \gamma_1 \sum_{i=1}^{H} |b_i| + \gamma_2 b'b\}, \tag{13}$$

which is similar to the elastic net technique[5] (Zou and Hastie, 2005). Imposing both an L^1 and L^2 penalty has the benefit of using prior economic beliefs to avoid overfitting as well as to deliver a sparse SDF when this is desired. Thus, the degree of shrinkage increases when strengthening L^2 penalty, whereas the degree of sparsity is increased when strengthening the L^1 penalty.

3.4 Penalty choice and performance evaluation

To choose the penalties γ or γ_1 and γ_2 such that the out-of-sample (OOS) R^2 is maximized, K-fold cross-validation (CV) is performed. This allows selecting the penalties based on the data instead of determining them beforehand based on assumptions. The K-fold CV works as follows. First, the historical data is divided into K equally sized subsamples. Secondly, in $K-1$ of those subsamples, for each penalty (γ or γ_1 and γ_2), $\hat{b}$ is estimated. Third, the OOS fit is evaluated based upon the one subsample which is not

[5] The LARS-EN algorithm from Zou and Hastie (2005) is used to calculate $\hat{b}$.

used in the estimation of $\hat{b}$,

$$R^2_{OOS} = 1 - \frac{(\bar{\mu}_2 - \bar{\Sigma}_2\hat{b})'(\bar{\mu}_2 - \bar{\Sigma}_2\hat{b})}{\bar{\mu}_2'\bar{\mu}_2}, \tag{14}$$

where subscript 2 indicates that the mean or variance are estimated OOS in the withheld subsample. The previous two steps are repeated K times, each time withholding a different subsample. Then, the cross-validated out-of-sample R^2 is calculated by averaging the K individual R^2_{OOS}. Finally, the penalties that maximize the cross-validated OOS R^2 are chosen. An important thing to note here is that the cross-validated OOS R^2 is an upward biased estimator of the true OOS R^2 because all historical data is used to calculate the estimator (Tibshirani and Tibshirani, 2009).

3.5 Nonlinear factors in a characteristic-based factor SDF

As mentioned in the Introduction Section, previous literature has shown that the cross-section of asset returns exhibit nonlinearities. Therefore, as an extension to the paper of Kozak et al. (2020), I relax the assumption that the SDF is linear in factors[6]. For this purpose, the factors are portfolios evaluated at some basis functions rather than the portfolios themselves. Using a basis expansion to approximate nonlinearities in the cross-section has the benefit that the factors are evaluated at one point such that b can still be estimated utilizing a linear regression. To arrive at an expression of the SDF, I assume that the SDF is linear in the nonlinear factors, that is

$$M_t = 1 - b'(f(F_t) - \mathbb{E}[f(F_t)]), \text{ with } f(F_t) = \begin{pmatrix} f(F_{t,1}) \\ f(F_{t,2}) \\ \vdots \\ f(F_{t,N}) \end{pmatrix} = \begin{pmatrix} c_1'p_1 \\ c_2'p_2 \\ \vdots \\ c_N'p_N \end{pmatrix}, \tag{15}$$

where M_t is the SDF, b is an $N \times 1$ vector of risk prices and $f(F_t)$ is an $N \times 1$ vector of linear combinations of F_t. Compared to the linear case in Equation (6), $f(F_t)$ replaces F_t. However, the assumption that b is linear in characteristics still holds, which allows b to be time-invariant in this setup. Each element of $f(F_t)$, $f(F_{i,t}) = c_i'p_i$ for all $i = 1, ..., N$. p_i is a $K \times 1$ vector of basis functions evaluated for the factor $F_{i,t}$. That is, each p_i is a column of the matrix P_F defined as

$$P_F = \{p_{j,i} = p_j(F_{i,t})\} = \begin{pmatrix} p_1(F_{1,t}) & p_1(F_{2,t}) & \dots & p_1(F_{N,t}) \\ p_2(F_{1,t}) & p_2(F_{2,t}) & \dots & p_2(F_{N,t}) \\ \vdots & \vdots & & \vdots \\ p_K(F_{1,t}) & p_K(F_{2,t}) & \cdots & p_K(F_{N,t}) \end{pmatrix}, \quad p_i = \begin{pmatrix} p_{1,i} \\ c_{2,i} \\ \vdots \\ c_{K,i} \end{pmatrix}. \tag{16}$$

c_i is the $K \times 1$ vector of corresponding coefficients. As basis functions, similar to Gu et al. (2020), a spline series of order two is used: $(1, F_t, (F_t - c_1)^2, (F_t - c_2)^2, ..., (F_t - c_{K-2})^2)$, where $c_1, c_2, ..., c_{K-2}$ are

[6] I want to thank Grith (2021a) for her contributions to the methodology of pricing nonlinear factors.

knots[7]. Thus, each c_i is a column of the matrix C defined as

$$C = \{c_{j,i}\} = \begin{pmatrix} c_{1,1} & c_{1,2} & \dots & c_{1,N} \\ c_{2,1} & c_{2,2} & \dots & c_{2,N} \\ \vdots & \vdots & & \vdots \\ c_{K,1} & c_{K,2} & \dots & c_{K,N} \end{pmatrix}, \quad c_i = \begin{pmatrix} c_{1,i} \\ c_{2,i} \\ \vdots \\ c_{K,i} \end{pmatrix}. \tag{17}$$

3.5.1 Derivation coefficients

For the aim of the extension, nonlinear instead of linear factors are priced. This allows us still to obtain a true variance-covariance matrix of the nonlinear factors. If you would price linear factors, the variance-covariance matrix is not square, making the implementation to obtain an estimator very difficult. Thus, from here onwards, I focus on pricing nonlinear factors $f(F_t)$. To start the derivation of the coefficients, I use the definition of expected returns resulting from the fundamental asset pricing equation. It holds that the expected value of $f(F_t)$ is related to the covariance between $f(F_t)$ and M_t as

$$\mathbb{E}[f(F_t)] = -\text{Cov}[f(F_t), M_t]. \tag{18}$$

Writing the covariance between $f(F_t)$ and M_t in terms of expected values, the above equation equals

$$\mathbb{E}[f(F_t)] = -\,\mathbb{E}\,[f(F_t) - \mathbb{E}\,f(F_t)][M_t - \mathbb{E}\,(M_t)]. \tag{19}$$

Plugging in the definition of the SDF, $M_t = 1 - b'(f(F_t) - \mathbb{E}[f(F_t)])$, into this equation results in

$$\mathbb{E}[f(F_t)] = -\,\mathbb{E}\,[f(F_t) - \mathbb{E}\,f(F_t)][1 - b'(f(F_t) - \mathbb{E}\,f(F_t)) - \mathbb{E}\,(M_t)]. \tag{20}$$

Rewriting the equation such that b is outside the expected value and using that $\mathbb{E}\,(M_t) = 1$ gives

$$\mathbb{E}[f(F_t)] = \mathbb{E}[f(F_t) - \mathbb{E}\,f(F_t)][(f(F_t) - \mathbb{E}\,f(F_t)]'b, \tag{21}$$

in which the expected value is by definition the variance-covariance matrix of $f(F_t)$. This results in

$$\mathbb{E}[f(F_t)] = \text{Var}[f(F_t)]b. \tag{22}$$

Plugging in the definition of $f(F_{i,t}) = c_i' p_i$ into the variance-covariance matrix of $f(F_t)$ results in the following matrix of variances and covariances of the vectors of basis functions:

$$\text{Var}[f(F_t)] = \begin{pmatrix} c_1' R_{11} c_1 & c_1' R_{12} c_2 & \dots & c_1' R_{1N} c_N \\ c_2' R_{21} c_2 & c_2' R_{22} c_2 & \dots & c_2' R_{2N} c_N \\ \vdots & \vdots & & \vdots \\ c_N' R_{N1} c_1 & c_N' R_{N2} c_2 & \cdots & c_N' R_{NN} c_N \end{pmatrix}, R_{ij} = \begin{cases} \text{Cov}\,(p_i, p_j) & \text{if } i \neq j \\ \text{Var}(p_i) & \text{if } i = j \end{cases} \tag{23}$$

Here, the off-diagonal elements R_{ij} represent the covariance between two columns of P_F, p_i and p_j, and

[7] The basis functions from the algorithm of Gu et al. (2020 & 2021) are used to calculate P_F.

the diagonal elements represent the variance of one column of P_F. This representation of the variance-covariance matrix allows to write the matrices R and $\mathbb{C}$ as

$$R = \begin{pmatrix} R_{11} & R_{12} & \dots & R_{1N} \\ R_{21} & R_{21} & \dots & R_{2N} \\ \vdots & \vdots & & \vdots \\ R_{N1} & P_{N2} & \cdots & R_{NN} \end{pmatrix}, \quad \mathbb{C} = \begin{pmatrix} c_1 & 0 & \dots & 0 \\ 0 & c_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & c_N \end{pmatrix}, \tag{24}$$

where R is an $(N \times K) \times (N \times K)$ variance-covariance matrix of P_F and $\mathbb{C}$ is an $(N \times K) \times N$ diagonal matrix of coefficients. Each c_i is $K \times 1$ and equivalent to c_i from Equation (17), and each 0 is a $K \times 1$ vector of zeros. These definitions result in the following expression of the variance of $f(F_t)$:

$$\text{Var}[f(F_t)] = \mathbb{C}' R \mathbb{C}. \tag{25}$$

Next to that, the definition of $f(F_{i,t}) = c_i' p_i$ can also be plugged in the the expected value of $f(F_t)$, which results in the following representation of the expected value of $f(F_t)$:

$$\mathbb{E}\left[f(F_t)\right] = \begin{pmatrix} c_1' \, \mathbb{E}\left[p_1\right] \\ c_2' \, \mathbb{E}\left[p_2\right] \\ \vdots \\ c_N' \, \mathbb{E}\left[p_N\right] \end{pmatrix} = \mathbb{C}' \, \mathbb{E}[P_v], \text{ where } \mathbb{E}[P_v] = \begin{pmatrix} \mathbb{E}[p_1] \\ \mathbb{E}[p_2] \\ \vdots \\ \mathbb{E}[p_N] \end{pmatrix} \tag{26}$$

is the $(N \times K) \times 1$ vectorization of the expected values of each p_i and $\mathbb{C}'$ is the $N \times (N \times K)$ matrix of the transpose of each element in $\mathbb{C}$ from Equation (24). Plugging the above derived definitions of the expected value and the variance-covariance matrix into Equation (22) results in

$$\mathbb{C}' \, \mathbb{E}[P_v] = \mathbb{C}' R \mathbb{C} b. \tag{27}$$

It is assumed that $\mathbb{C}\mathbb{C}' = \mathcal{I}$. I elaborate on this assumption in Section 3.5.3. Then, the coefficients $\mathbb{C}$ and b can be isolated and written in terms of the first and second moments of the basis functions evaluated at each factor as

$$\mathbb{C} b = R^{-1} \, \mathbb{E}[P_v]. \tag{28}$$

However, $\mathbb{C}$ and b cannot be separated, which does not allow to estimate the risk price corresponding to the nonlinear factors separate from the coefficients corresponding to the basis functions. Hence, the estimates resulting from this method do not represent the risk prices of the factors but rather the risk price weighted by the coefficients corresponding to the basis function the factor is evaluated at.

3.5.2 Estimator

Although $\mathbb{C}$ and b cannot be separated, it is possible to estimate the joint coefficients while applying L^2 shrinkage to the variance-covariance matrix of P_F, similar to Kozak et al. (2020). This results in the

following estimator[8]:

$$(\hat{\mathbb{C}b}) = \tilde{R}^{-1}\bar{P}_v, \qquad \tilde{R} = \begin{cases} \tilde{R}_{ii} = R_{ii}(1+\gamma\mathcal{I}_K) & \text{if } i = j \\ \tilde{R}_{ij} = R_{ij} & \text{if } i \neq j \end{cases} \tag{29}$$

Here, $\bar{P}_v$ is the sample version of $\mathbb{E}[P_v]$, R is the second moment of P_F as defined in Equation (24), and $\mathcal{I}_K$ is a $(K \times K)$ identity matrix. Hence, the diagonal elements of the variance-covariance matrix R are extended with a penalty $\gamma = \frac{\tau}{\kappa^2 T}$. This penalty comes from an economically motivated prior belief about the relationship between factor means and (co)variances. This prior belief is set such that the coefficients are shrunk towards zero. For a thorough explanation of how the penalty results from an economically motivated prior belief, I refer to Section (3.3).

Next to an L^2 penalty, we can also impose an L^1 penalty to allow for the possibility to obtain a sparse SDF representation. Namely, an L^2 penalty only shrinks the coefficients towards zero, but does not set them exactly equal to zero. This is where the L^1 penalty comes in place. The stronger the L^1 penalty, the more coefficients are set to zero and the sparse the SDF representation. Maximizing the cross-sectional R^2 subject to the L^2 penalty of the maximum squared Sharpe ratio as well as an L^1 penalty defined as $\gamma_1 \sum_{j=1}^{K}\sum_{i=1}^{N}|c_{j,i}b_i|$ results in the following estimator of $(\mathbb{C}b)$:

$$(\hat{\mathbb{C}b}) = \arg\max_{\mathbb{C}b}\{(\bar{P}_v - R(\mathbb{C}b))'(\bar{P}_v - R(\mathbb{C}b)) + \gamma_2(\mathbb{C}b)'(\mathbb{C}b) + \gamma_1 \sum_{j=1}^{K}\sum_{i=1}^{N}|c_{j,i}b_i|\} \tag{30}$$

3.5.3 Penalty choice and performance evaluation

To choose the penalties such that the OOS R^2 is optimized, K-fold cross-validation (CV) is performed as well in the nonlinear setup. For a thorough explanation of cross-validation, I refer to Section (3.4). The R^2_{OOS} of the withheld subsample in the case of pricing nonlinear factors equals[9]

$$R^2_{OOS} = 1 - \frac{(\bar{P}_{v,2} - \bar{R}_2(\hat{\mathbb{C}b}))'\hat{\mathbb{C}}\hat{\mathbb{C}}'(\bar{P}_{v,2} - \bar{R}_2(\hat{\mathbb{C}b}))}{\bar{P}'_{v,2}\hat{\mathbb{C}}\hat{\mathbb{C}}'\bar{P}_{v,2}}, \tag{31}$$

where subscript 2 indicates that the mean or variance are estimated OOS in the withheld subsample. $\bar{P}_{v,2}$ is the OOS mean of P_v and $\bar{R}_2$ is the OOS variance-covariance matrix of P_F. Assuming that $\hat{\mathbb{C}}\hat{\mathbb{C}}' = \mathcal{I}$ results in

$$R^2_{OOS} = 1 - \frac{(\bar{P}_{v,2} - \bar{R}_2(\hat{\mathbb{C}b})'(\bar{P}_{v,2} - \bar{R}_2(\hat{\mathbb{C}b}))}{\bar{P}'_{v,2}\bar{P}_{v,2}}. \tag{32}$$

The assumption that $\hat{\mathbb{C}}\hat{\mathbb{C}}' = \mathcal{I}$ is made because it is not possible to estimate b and $\mathbb{C}$ separately. Hence, we do not know the isolated estimator of $\mathbb{C}$. Therefore, to calculate the OOS R^2, a simplifying assumption should be made to cancel $\hat{\mathbb{C}}\hat{\mathbb{C}}'$. Because the original factors $F_{i,t}$ are zero-investment long-short portfolios[10], which means that the factors are orthogonal, this assumption is plausible.

[8] For a more detailed derivation of this estimator I refer to Appendix B.

[9] A more detailed derivation of the out-of-sample R^2 can be found in Appendix C.

[10] I elaborate on the construction of these portfolios in Section 4.

4 Data

For the empirical applications of the methodology in this paper, two datasets from Kozak et al. (2020) are used. The first dataset consists of the Fama-French 25 sorted portfolios, which are retrieved from French (2020). Daily returns on the 25 Fama-French Size and Book-to-Market sorted portfolios are used. This dataset consists of stocks listed on the NYSE, AMEX and NASDAQ for which there exists market- and book equity data. The period considered ranges from 1926/07/01 until 2017/12/29. The stocks are divided into 5×5 portfolios of stocks in a given Size and Book-to-Market quintile. This means that the stock returns are ordered from high to low based on their size or book-to-market equity ratio and divided into five equally sized quantiles. Firms that are in the intersection of a certain Size and Book-to-Market quintile form a portfolio. For example, one portfolio could be all firms in the second quintile of Size and the fifth quintile of Book-to-Market. The Center for Research in Security Prices (CRSP) value-weighted index return is subtracted from these portfolio returns, and they are rescaled to have the same standard deviations as the in-sample standard deviation of the aggregate market index excess return.

The second dataset consists of 50 anomaly portfolios that are constructed by Kozak et al. (2020), based on frequently used anomalies in the finance literature. Daily returns of the US firms retrieved from CRSP from a period of 01/11/1973 until 20/12/2017 are considered. At each time t, stocks with a market capitalization smaller than 0.01% are removed to prevent the results from being driven by high-volatile penny stocks. To arrive at anomaly portfolios, a set of 50 characteristics is manipulated as follows. First, for each characteristic, a rank transformation is performed. Let $c^i_{s,t}$ be a characteristic i of stock s at time t. All stocks are sorted based on the value of the respective characteristic i and are ranked from 1 until n_t, where n_t is the number of stocks at time t. Then the rank-transformed characteristic is calculated as $rc^i_{t,s} = \text{rank}(c^i_{s,t})/(1 + n_t)$. Secondly, the rank transformed characteristic is further normalized as

$$z^i_{t,s} = \frac{(rc^i_{t,s} - \bar{rc}^i_{t,s})}{\sum_{s=1}^{n_t} |rc^i_{t,s} - \bar{rc}^i_{t,s}|}. \tag{33}$$

This results in zero-investment long-short portfolios per characteristic. All characteristics are combined in one instrument matrix Z_t. Multiplying this with the stock returns gives the desired matrix of anomaly portfolios.

5 Results

The empirical results using the methods discussed in this paper and the datasets mentioned in the previous section are reported in this section. First, the results of the replication, and second of the extension are presented and discussed.

5.1 Replication

The key objective of Kozak et al. (2020) is to show that their proposed method with the use of an economically motivated prior can conquer the *multidimensional challenge* (Cochrane, 2011). They want to show that their method has good out-of-sample performance while considering a large set of stock

characteristics. They do this by first applying the method to a set of well-known portfolios: the 25 Fama-French Size and Book-to-Market (FF25) portfolios. Thereafter, they consider two sets of stock anomalies: a self-constructed set of 50 anomaly portfolios based on popular characteristics used in the finance literature and the WRDS financial ratio set. Lastly, to raise the dimensional challenge, they supplement each of these sets with second and third powers and linear first-order interactions. In this section, I focus on replicating the results from Kozak et al. (2020) for the FF25 portfolios and the self-constructed 50 anomaly portfolios[11].

5.1.1 25 Fama-French Size and Book-to-Market portfolios

Before the method from Kozak et al. (2020) is applied to a high-dimensional dataset, its performance is evaluated for a well-known and commonly used dataset: the FF25 portfolios. For a complete description of the dataset, I refer to Section 4. These portfolios are considered because their returns are known to have a clear factor structure. This means that the returns are linear combinations of the SMB and HML factors[12] (Lewellen et al., 2010). Furthermore, the first and second PCs of the FF25 are known to correspond with the SMB and HML factors. Therefore, it can be assessed whether the method recovers the expected sparsity. This gives a good first indication of the performance of the model before moving to the high-dimensional setting.

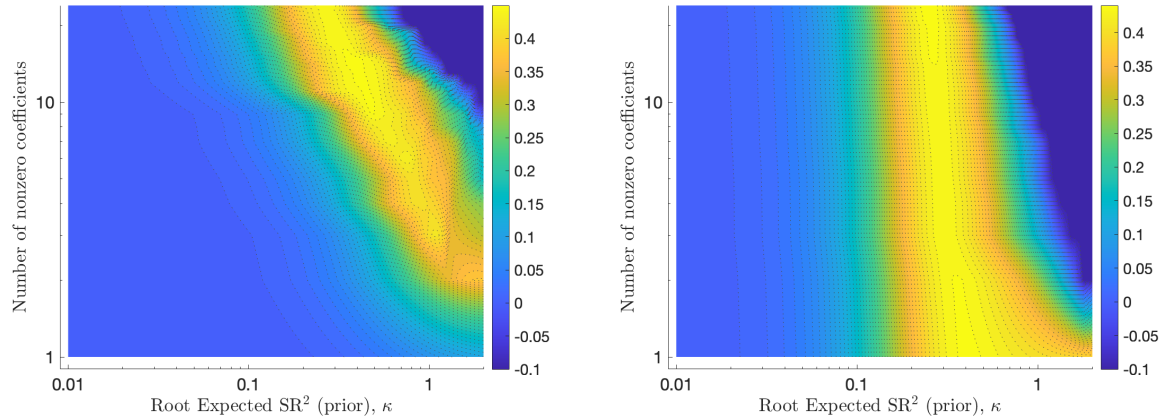
In Figure 1, contour maps of the FF25 portfolios (a) and the PCs of the FF25 portfolios (b) are presented. The contour map plots the degree of sparsity, expressed in the number of nonzero SDF coefficients, against the amount of shrinkage, expressed by κ. The amount of shrinkage is chosen to be expressed by κ because it has an economic interpretation: it equals the square root of the expected maximum squared Sharpe ratio. The degree of sparsity corresponds to γ_1 and the amount of shrinkage to γ_2 in Equation 13. The values for γ_1 and γ_2 are depicted on a logarithmic scale. Viewing the map from along the y-axis, the top shows little sparsity (many factors), whereas the bottom shows extreme sparsity (only one factor). From along the x-axis, the most left side depicts extreme shrinkage, and the most right side no shrinkage at all. The third dimension of the map displays the level of the out-of-sample (OOS) cross-sectional R^2. The different levels are expressed in colours, and which colour corresponds to which level is given by the bar right next to the figures. The darker blue, the smaller the R^2 and the more yellow, the higher the R^2. The R^2 is bounded by a minimum of -0.1. In summary, for each pair of (γ_1, γ_2), the contour map shows the corresponding OOS R^2. Thus, this contour map visualizes the levels of out-of-sample performance of the model expressed by R^2 for different degrees of sparsity and shrinkage.

Zooming in on Figure 1a, the somewhat diagonal shape of the yellow grid shows that less sparsity should be compensated with more shrinkage to arrive at the same level of out-of-sample performance. Also, a good OOS R^2 can be attained for a small number of nonzero coefficients, that is, when only a few factors are included. This confirms the expectations of the FF25 portfolios, from which it is known that they have a clear factor structure.

[11] I want to thank Kozak et al. (2020) for making available most of their code.

[12] For further explanation of these factors, I refer to the paper of Fama and French (1993) and the website of French (2020).

For the PCs of the FF25 portfolios, displayed in Figure 1b, the factor structure becomes even more clear. The almost vertical yellow grid shows that no additional shrinkage needs to be imposed to compensate for the use of more factors. This implies that the method can capture many factors without resulting in overfitting and in turn, worse out-of-sample performance. However, some amount of shrinkage is needed to arrive at a good level of OOS R^2 because the dark blue plane to the right from the yellow vertical plane indicates that applying no shrinkage at all leads to terrible results. Furthermore, including many factors is possible but unnecessary, as the optimal OOS R^2 is already nearly attained when including only one factor. This is also expected from a statistical point of view because the first few PC transformed factors capture the highest proportion of the variance explained in the cross-section of asset returns. When a set of portfolios already has a clear factor structure, the PCs transformed portfolios should have an even more sparse structure.



(a) Fama-French 25 portfolios

(b) PCs of Fama-French 25 portfolios

Figure 1: Cross-sectional OOS R^2 for possible combinations of L^1 and L^2 penalties considering the linear pricing model. The left figure (a) displays the OOS R^2 for 25 Fama-French Size and Book-to-Market sorted portfolios. The right figure (b) for their PC transformation. The x-axis in both figures shows the amount of shrinkage (L^2 penalty), with no shrinkage at the right and extreme shrinkage at the left edge of the figure. The y-axis shows the degree of sparsity (L^1 penalty), which is expressed as the number of nonzero coefficients. A small number indicates high sparsity, and a large number indicates almost no sparsity. The coloured contour map depicts the OOS R^2 for combinations of the L^1 and L^2 penalties, where the darker blue means the lower the R^2 and the lighter yellow the higher the R^2. The R^2 bounded at a minimum of -0.1.

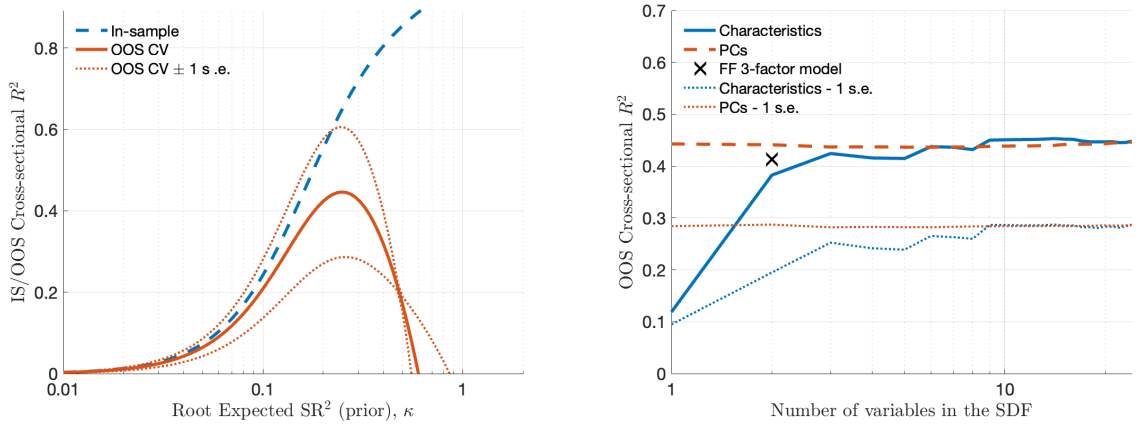
The aforementioned results are very similar to those of Kozak et al. (2020), which is completely logical because I use their code and dataset. However, there are made some assumptions in their paper that I found rather arbitrary. First, the choice of $K = 3$ in the K-fold CV. Kozak et al. (2020) base this choice on the fact that they have to compromise between estimation uncertainty in $\hat{b}$ and $\bar{\Sigma}_2$. The estimation precision of the OOS covariance matrix lowers when the withheld sample becomes smaller. However, they note that their output is robust for values slightly higher than $K = 3$. It is not clear to me why they chose $K = 3$ over for example $K = 5$. The second decision that I found arbitrary is the choice of a withheld sample of 31 years. Kozak et al. (2020) do not provide any explanation on why this sample size is chosen and whether their results are robust to changing this sample size.

Figure 2 shows some cuts along the axes of the contour maps in Figure 1. For the left panel

(a), a cut is taken along the top of Figure 1a. This means that the cross-sectional R^2 of the raw FF25 portfolios is displayed for various levels of shrinkage without imposing any sparsity. Figure 2a shows that the maximum R^2 of approximately 0.45 is attained at a value of $\kappa \approx 0.25$. The plot also shows that evaluating the in-sample R^2 can lead to very different conclusions regarding the amount of shrinkage necessary for a model to perform well out-of-sample. This indicates that in-sample evaluation is not the appropriate method to evaluate out-of-sample performance.

The right panel (b) of Figure 2 displays the maximum cross-sectional R^2 obtained at various degrees of sparsity, choosing the amount of shrinkage that maximizes the R^2 at a certain level of sparsity. This is done for both the raw FF25 portfolios and their PC transformations. Thus, in figure 2b, a cut is taken from the bottom to the top of the maps in Figure 1 along all R^2 maximizing amounts of shrinkage. Figure 2b shows that for the FF25 portfolios, a sparse SDF representation, similar to the Fama-French 3-factor model (indicated by the cross), can be obtained. The PC transformation of the FF portfolios allows for an even more sparse representation, as an R^2 close to the maximum can already be obtained by including only 1 factor in the model.

In summary, the raw FF25 portfolio returns trade-off shrinkage and sparsity to attain good out-of-sample performance, whereas the PC transformation of the portfolio returns allows for extreme sparsity without applying additional shrinkage. However, some shrinkage is needed to ensure good out-of-sample performance.

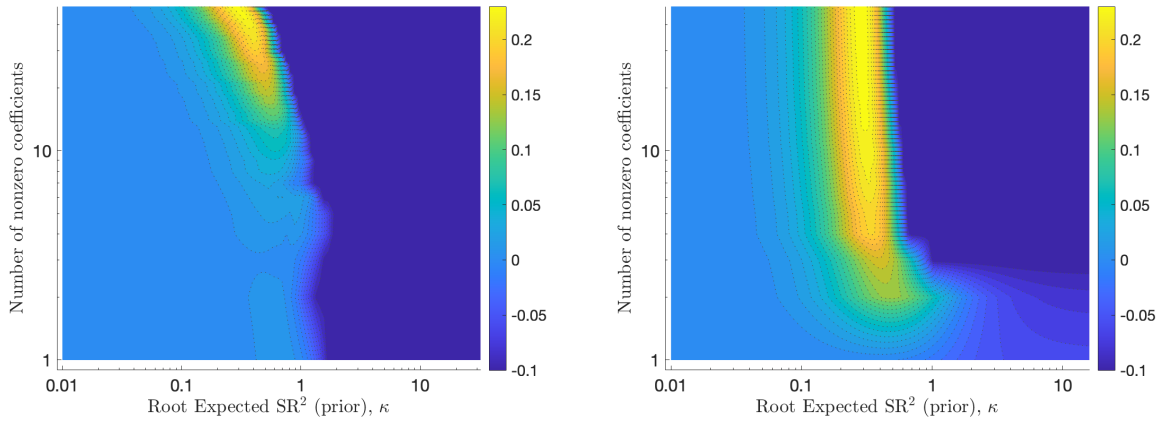


(a) Shrinkage (b) Sparsity

Figure 2: Cross-sectional OOS R^2 for various levels of shrinkage (a) and sparsity (b), based on the 25 Fama-French Size and Book-to-Market sorted portfolios. The left panel (a) plots the in-sample and out-of-sample R^2 against the amount of shrinkage (L^2 penalty) without imposing sparsity. The dashed blue line displays the in-sample R^2, the solid red line the OOS R^2 and the dotted red lines the $\pm$ 1 standard error (s.e.) bounds of the OOS R^2. The right panel (b) plots the maximum OOS R^2 attained for optimal levels of shrinkage against the degree of sparsity. The solid blue line depicts the R^2 for the raw 25 Fama-French portfolios, with the dotted blue line its -1 s.e. The dashed red line depicts the PCs of the 25 Fama-French portfolios, with the dotted red line its -1 s.e. The cross ('X') indicates the R^2 attained for the Fama-French three-factor model, which includes the risk-free market rate, SMB, and HML factor.

5.1.2 50 anomaly portfolios

Applying the model to the FF25 portfolios forms a good starting point for evaluating the models' performance. Still, to assess whether the model can tackle the multidimensional challenge, a higher dimensional dataset should be considered as well. In this section, the results of applying the model to the second dataset considered in this paper, the 50 anomaly portfolios, are discussed. In Figure 3a and 3b, the cross-sectional OOS R^2 is plotted for various combinations of the degree of sparsity and shrinkage. In the left panel (a) for the raw 50 anomaly portfolios, and the right panel (b) for their PCs. From Figure 3a, it visible that for the 50 anomaly portfolios, essentially no sparsity can be attained while at the same time ensuring a good OOS R^2. Moreover, the maximum OOS R^2 is attained for an SDF representation including 48 out of 50 factors. This is clear evidence for the observation of Kozak et al. (2020) that the cross-section of asset returns cannot be summarized by only a few factors. Many characteristics contribute small parts to the cross-sectional variation; therefore, all characteristics should be taken into account. Although almost no sparsity is allowed, a substantial degree of shrinkage should be imposed to attain good OOS performance. This can be seen from the large dark blue plain on the right-hand side of the contour map. The OOS R^2 is bounded by a minimum value of -0.1, and lifting this bound would show an even worse OOS R^2 when almost no shrinkage is applied. Lastly, for the 50 anomaly portfolios, in contrast to the FF25 portfolios, sparsity and shrinkage are no substitutes anymore in ensuring good OOS performance.



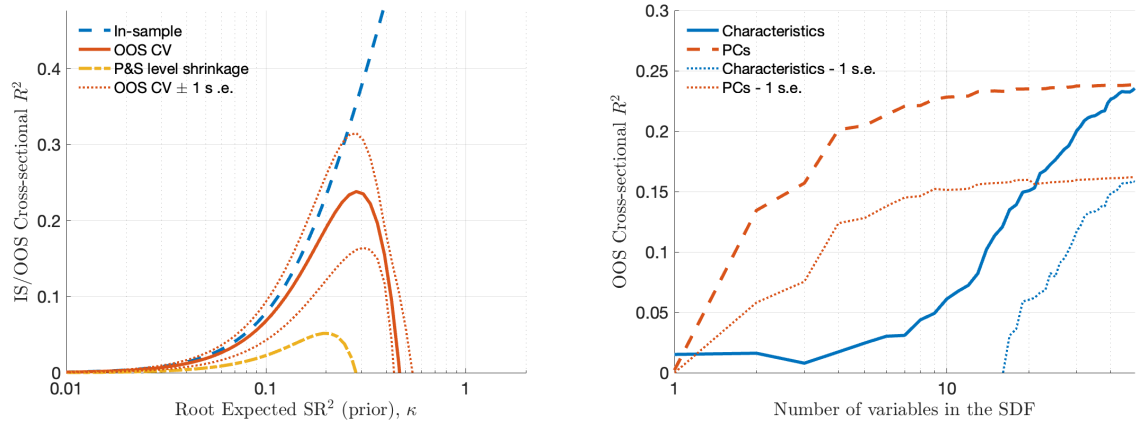
(a) 50 anomaly portfolios

(b) PCs of 50 anomaly portfolios

Figure 3: Cross-sectional OOS R^2 for possible combinations of L^1 and L^2 penalties considering the linear pricing model. The left figure (a) displays the OOS R^2 for 50 anomaly portfolios constructed by Kozak et al. (2020). The right figure (b) for their PC transformation. The x-axis in both figures shows the amount of shrinkage (L^2 penalty), with no shrinkage at the right and extreme shrinkage at the left edge of the figure. The y-axis shows the degree of sparsity (L^1 penalty), expressed as the number of nonzero coefficients. A small number indicates high sparsity, and a large number indicates almost no sparsity. The coloured contour map depicts the OOS R^2 for combinations of the L^1 and L^2 penalties, where the darker blue means the lower the R^2 and the lighter yellow the higher the R^2. The R^2 bounded at a minimum of -0.1.

The right subfigure, Figure 3b, shows a rather different conclusion. When the 50 anomaly portfolios are transformed into their PCs, the model attains a high OOS R^2 even when allowing for substantial sparsity. This conclusion is seen from the almost vertical yellow plain. The plain ends at a degree of

around 3/4 factors. Hence, the PCs of 50 anomaly portfolios can be explained well by around 4 factors. Similar to the raw 50 anomaly portfolios, still substantial shrinkage should be applied to ensure good OOS performance.



(a) Shrinkage (b) Sparsity

Figure 4: Cross-sectional OOS R^2 for various levels of shrinkage (a) and sparsity (b), based on the 50 anomaly portfolios constructed by Kozak et al. (2020). The left panel (a) plots the in-sample and OOS R^2 against the amount of shrinkage (L^2 penalty) without imposing sparsity. The dashed blue line displays the in-sample R^2, the solid red line the OOS R^2 and the dotted red lines the $\pm$ 1 standard error (s.e.) bounds of the OOS R^2. The yellow dash-dotted line shows the OOS R^2 based on the proportional shrinkage method from Pastor and Stambaugh (2000). The right panel (b) plots the maximum OOS R^2 attained for possible levels of shrinkage against the degree of sparsity. The solid blue line depicts the R^2 for the raw 50 anomaly portfolios, with the dotted blue line its -1 s.e. bound, and the dashed red line depicts the PCs of the 50 anomaly portfolios, with the dotted red line its -1 s.e. bound.

Figure 4 shows some cuts along the axes of the contour maps in Figure 3. For the left panel (a), a cut is taken along the top of Figure 3a. This means that the cross-sectional R^2 of the raw 50 anomaly portfolios is displayed for various levels of shrinkage without imposing any sparsity. For comparison purposes, Figure 4a also includes a plot of the OOS R^2 attained by the proportional shrinkage method from Pastor and Stambaugh (2000). This is equivalent to choosing a prior with $\eta = 1$. Figure 4a shows that the maximum R^2 of approximately 0.28 is attained at a value of $\kappa \approx 0.25$. Moreover, it can be seen that the R^2 obtained by the method of Kozak et al. (2020) is substantially higher than the one obtained by the proportional shrinkage method from Pastor and Stambaugh (2000). However, to obtain this line, I had to apply various manipulations such as mirroring the line and moving the line to the left. Next to that, the estimator of b, according to the proportional shrinkage method should be $\hat{b} = (\Sigma + \gamma\Sigma)^{-1}\bar{\mu}$. But the estimator of b that derives this line in the code of Kozak et al. (2020) equals $\hat{b} = \gamma(\Sigma + 0 \times I_n)^{-1}\bar{\mu}$. I have tried my best to implement the former estimator, but the code keeps giving an error unless the L^2 penalty is changed drastically. Hence, I am doubtful about the values the code returns in this perspective. Next to the OOS R^2, the plot shows the in-sample R^2. It again becomes apparent that evaluating the in-sample performance of a method can lead to very different conclusions regarding the amount of shrinkage necessary for a model to perform well out-of-sample.

The right panel (b) of Figure 4 shows the maximum cross-sectional R^2 obtained at various degrees

of sparsity, choosing the amount of shrinkage that maximizes the R^2 at a certain level of sparsity. This is done for both the raw 50 anomaly portfolios and their PC transformations. Thus, in figure 4b, a cut is taken from the bottom to the top of the maps in Figure 3 along all R^2 maximizing amounts of shrinkage.

Figure 4b shows clearly that applying a PC transformation to the 50 anomaly portfolios leads to a concave function of the R^2 in the number of variables in the SDF, as opposed to a convex function for the raw portfolios. This means that the PC transformation of the 50 anomaly portfolios can produce a good out-of-sample R^2 with a sparse SDF representation, whereas the raw portfolios cannot and a large number of variables need to be included in the SDF to produce good out-of-sample results.

In summary, for the 50 anomaly portfolio returns, almost all characteristics need to be considered to obtain good out-of-sample performance. The PC transformation of the returns allows for more sparsity. Significant shrinkage should be applied to both return series to ensure good out-of-sample performance.

Table 1: The SDF coefficients and absolute t-statistics of the ten factors with the highest explanatory power at the optimal value of the prior root expected square Sharpe ratio, ordered in a descending manner on their t-statistic value. The raw 50 anomaly portfolios are considered on the left (a) and their PCs on the right (b).

(a) 50 anomaly portfolios			(b) PCs of 50 anomaly portfolios		
	b	t-stat		b	t-stat
Industry rel. rev. (L.V.)	-0.879	3.527	PC 4	1.014	4.249
Ind. mom-reversals	0.484	1.945	PC 1	-0.537	3.081
Industry rel. reversals	-0.425	1.705	PC 2	-0.556	2.653
Seasonality	0.322	1.292	PC 9	0.635	2.514
Earnings surprises	0.323	1.291	PC 15	-0.324	1.265
Value-profitability	0.297	1.184	PC 17	0.303	1.182
Return on market equity	0.299	1.183	PC 6	-0.287	1.176
Investment/Assets	-0.238	0.948	PC 11	0.189	0.744
Return on equity	0.238	0.947	PC 13	0.166	0.654
Composite issuance	-0.240	0.947	PC 23	0.146	0.564
Momentum (12m)	0.227	0.906	PC 7	-0.140	0.561

Next to evaluating the out-of-sample performance of the SDF representation proposed by Kozak et al. (2020), it is interesting to see which variables included in the 50 anomaly portfolios have the highest explanatory power. Therefore, the ten characteristics[13] and ten PCs of the 50 anomaly portfolios with the highest absolute t-statistics, ordered from highest to lowest, are presented in Table 1. The full table including all 50 characteristics and PCs is shown in appendix B. The left panel (a) of the table shows the characteristics that contribute most to the cross-sectional variation in asset returns. The list contains characteristics that are found among the most robust in asset pricing literature. Hence, the method of Kozak et al. (2020) can retrieve these as well. The right panel (b) of the table shows the PCs

[13]For a full description of the characteristics used, I refer to the paper of Kozak et al. (2020).

that contribute most to the cross-sectional variation in asset returns. Among the ten with the highest t-statistic, six are out of the first ten PCs. This is in line with the finding that the first couple of PCs represent linear combinations of characteristics such that the combined variance is highest.

5.2 Extension

In this section, I report the results from my extension of the methods of Kozak et al. (2020). As an extension, I relax the assumption that the SDF is linear in factors by allowing the relationship between the SDF and the factors to be nonlinear. For this purpose, a number of 5 basis functions are used. Furthermore, instead of pricing linear factors, I price the nonlinear factors. To compare whether relaxing the linearity assumption improves the estimation, I use the same two datasets that I used for the replication section. First, I consider the FF25 portfolios. Secondly, I use the 50 anomaly portfolios.

5.2.1 Fama-French Size and Book-to-Market portfolios

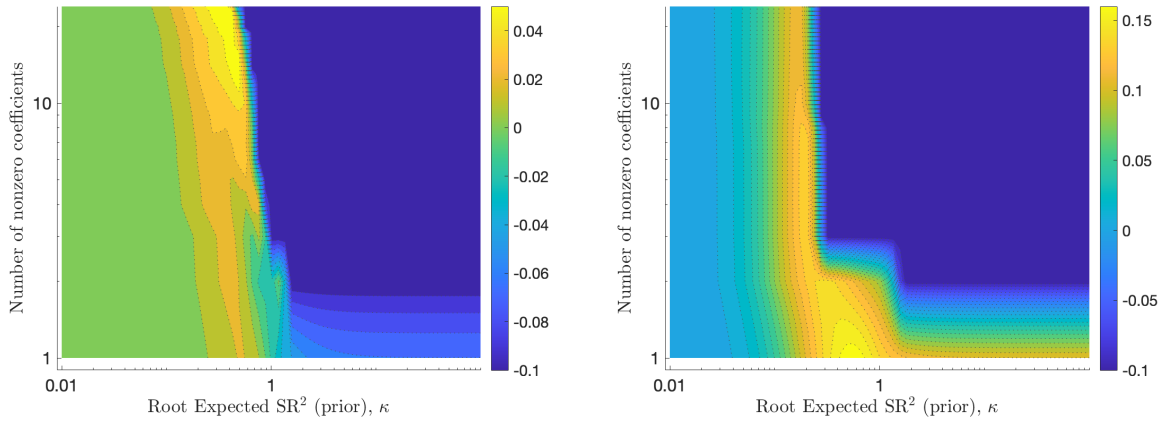
Similar to Section 5.1.1, first the methods are applied to a well-known dataset: the FF25 portfolios. Because the model is known to have a clear factor structure, it is interesting to see how the performance of the model changes when nonlinearities are accounted for.

In Figure 5, contour maps of the FF25 portfolios (a) and the PCs of the FF25 portfolios (b) are shown. The contour map plots the degree of sparsity, expressed as the number of nonzero coefficients, against the amount of shrinkage, expressed by κ. In the results of the extension, the coefficients are not the single risk prices, but the joint coefficients of the basis functions with the risk price of each factor, that is $c_i' b_i$. The amount of shrinkage is chosen to be expressed by κ because it has an economic interpretation: it equals the square root of the expected maximum squared Sharpe ratio. The degree of sparsity corresponds to γ_1 and the amount of shrinkage to γ_2 in Equation (30). The values for γ_1 and γ_2 are depicted on a logarithmic scale. Viewing the map from along the y-axis, the top shows little sparsity (a lot of factors), whereas the bottom shows extreme sparsity (only one factor). From along the x-axis, the most left side depicts extreme shrinkage, and the most right side no shrinkage at all. The third dimension of the map displays the level of the out-of-sample R^2. The different levels are expressed in colours and which colour corresponds to which level is given by the bar right next to the figures. The darker blue, the smaller the R^2 and the more yellow, the higher the R^2. The R^2 is bounded by a minimum of -0.1. In summary, for each pair of (γ_1, γ_2), the contour map shows the corresponding out-of-sample R^2. Thus, this contour map visualizes the levels of out-of-sample performance of the model expressed by R^2 for different degrees of sparsity and shrinkage.

In Figure 5a, it can be seen, from the dark blue plane on the right side of the map, that a substantial degree of shrinkage is necessary for the model to perform well out-of-sample. However, contrary to the results of the linear model, no sparsity can be attained, as the model should include at least 12 factors to still produce a close to maximum OOS R^2. Furthermore, the optimal OOS R^2 is quite low, attaining a maximum level of approximately 0.05. Because the FF25 portfolios are known to have a clear factor structure, it is expected that the portfolios exhibit few nonlinearities. This causes that pricing a nonlinear function of those factors is less accurate than pricing linear factors.

Zooming in on Figure 5b, where the contour map using the PCs of the FF25 portfolios is presented, it is visible that again shrinkage is necessary to attain a high OOS R^2. However, in contrast to the linear method, even without shrinkage, a positive R^2 can be attained in an SDF representation of one factor. Moreover, extreme sparsity of only one factor leads to the best OOS R^2, and in contrast to the linear model, the R^2 reduces again when the number of factors included in the model increases. This is evidence for the fact that when raw portfolios are transformed into their PCs, a sparse SDF representation can lead to good OOS performance. Also, when the number of nonzero coefficients increases, more shrinkage is needed to produce a positive OOS R^2.

Comparing Figure 5a and 5b, transforming the FF25 portfolios into their PCs results in better performance of the nonlinear model, especially when the SDF representation is sparse.



(a) Fama-French 25 portfolios

(b) PCs of Fama-French 25 portfolios

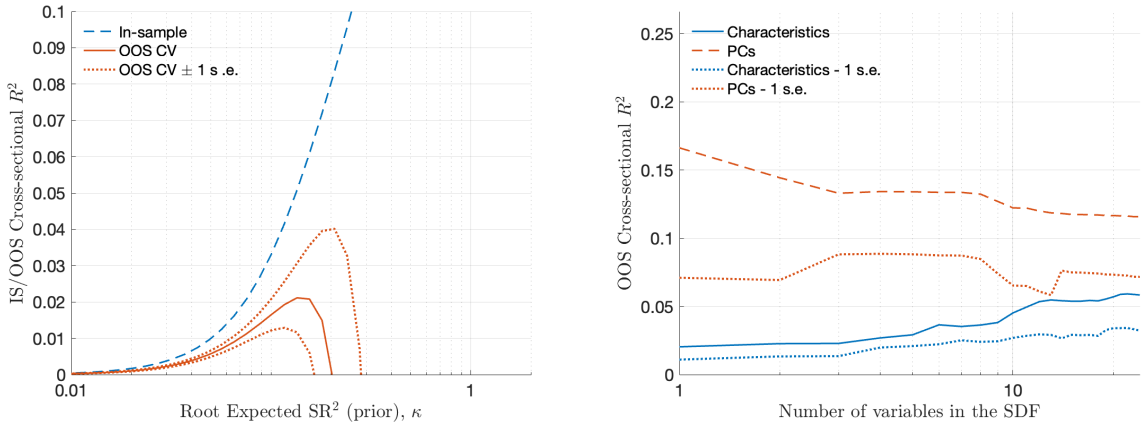
Figure 5: Cross-sectional OOS R^2 for possible combinations of L^1 and L^2 penalties considering the nonlinear pricing model. The left figure (a) displays the OOS R^2 for 25 Fama-French Size and Book-to-Market sorted portfolios. The right figure (b) for their PC transformation. The x-axis in both figures shows the amount of shrinkage (L^2 penalty), with no shrinkage at the right and extreme shrinkage at the left edge of the figure. The y-axis shows the degree of sparsity (L^1 penalty), which is expressed as the number of nonzero coefficients. A small number indicates high sparsity, and a large number indicates almost no sparsity. The coloured contour map depicts the OOS R^2 for combinations of the L^1 and L^2 penalties, where the darker blue means the lower the R^2 and the lighter yellow the higher the R^2. The R^2 bounded at a minimum of -0.1.

Figure 6 shows some cuts along the axes of the contour maps in Figure 5. For the left panel (a), a cut is taken along the top of Figure 5a. This means that the cross-sectional OOS R^2 of the raw FF25 portfolios is displayed for various levels of shrinkage without imposing any sparsity. Figure 6a shows that an OOS R^2 of just above 0.02 is attained when no sparsity is imposed. This level of R^2 is attained for a $\kappa \approx 0.14$. From this figure, it becomes again clear that pricing the nonlinear transformation of the FF25 portfolios result in worse OOS performance compared to pricing the linear portfolios. Furthermore, the figure shows that using in-sample performance evaluation leads to different conclusions about the amount of shrinkage needed for the model to perform well out-of-sample. This shows that in-sample performance evaluation is not a representative measure of out-of-sample performance.

The right panel (b) of Figure 6 displays the maximum cross-sectional R^2 obtained at various degrees

of sparsity, choosing the amount of shrinkage that maximizes the R^2 at a certain level of sparsity. This is done for both the raw FF25 portfolios and their PC transformations. Thus, in figure 6b, a cut is taken from the bottom to the top of the maps in Figure 5 along all R^2 maximizing amounts of shrinkage. From 6b, it becomes clear that for the raw FF25 portfolios, essentially more sparsity leads to worse out-of-sample performance, as the blue solid line shows a continuously increasing pattern. Moreover, transforming the FF25 portfolios into their PCs first always results in a higher OOS R^2. However, contrary to the linear case, the R^2 is decreasing in the number of variables included in the SDF.

In summary, the OOS performance of the nonlinear pricing model is worse than that of the linear pricing model, but transforming the portfolios into their PCs first achieves a higher out-of-sample performance compared to the raw portfolios and allows for a sparse SDF representation.

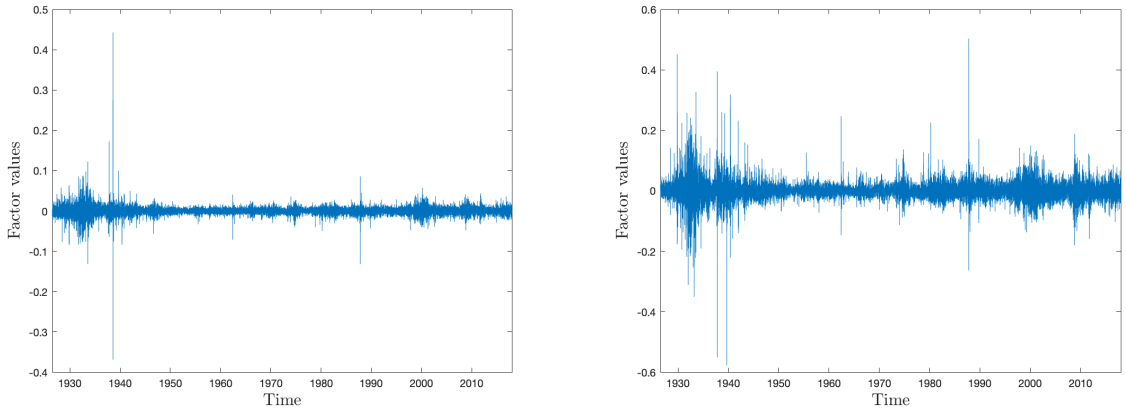


(a) Shrinkage (b) Sparsity

Figure 6: Cross-sectional OOS R^2 for various levels of shrinkage (a) and sparsity (b), based on the 25 Fama-French Size and Book-to-Market sorted portfolios. The left panel (a) plots the in-sample and OOS R^2 against the amount of shrinkage (L^2 penalty) without imposing sparsity. The dashed blue line displays the in-sample R^2, the solid red line the OOS R^2 and the dotted red lines the $\pm$ 1 standard error (s.e.) bounds of the OOS R^2. The right panel (b) plots the maximum OOS R^2 attained for optimal levels of shrinkage against the degree of sparsity. The solid blue line depicts the R^2 for the raw 25 Fama-French portfolios, with the dotted blue line its -1 s.e. The dashed red line depicts the PCs of the 25 Fama-French portfolios, with the dotted red line its -1 s.e.

Figure 7 shows the plots of the factor values with the most explanatory power according to the absolute t-statistic corresponding to the coefficient. The values are plotted for the raw FF25 portfolios in (a) and for their PC transformation in (b). The optimal factor of the raw portfolios is ME3BM3[14] and of the PC transformed FF25 portfolios PC1. From this figure, it becomes clear that the optimal PC factor contains way more variation than the raw optimal factor. Thus, this figure visualizes that the PC transformed factors can better capture the cross-section of asset returns with only a few factors than the raw factors.

[14]This is the Market Equity and Book-to-Market both of quintile 3 portfolio. For a thorough explanation of these portfolios, I refer to Section 4.



(a) Fama-French 25 portfolios

(b) PCs of Fama-French 25 portfolios

Figure 7: Factor values of the factor with the most explanatory power according to the absolute t-statistic for the FF25 portfolios (a) and their PC transformation (b) plotted over time. The optimal factor is ME3BM3 for the raw FF25 portfolios and PC1 for the PC transformations of the FF25 portfolios. The factor values equal $F_{t,i} = Z'_{t-1,i} R^e_{t,i}$ where i is the optimal factor. The time equivalent to the sample size of the FF25 portfolios data, which ranges from 1926 until 2017.

5.2.2 50 anomaly portfolios

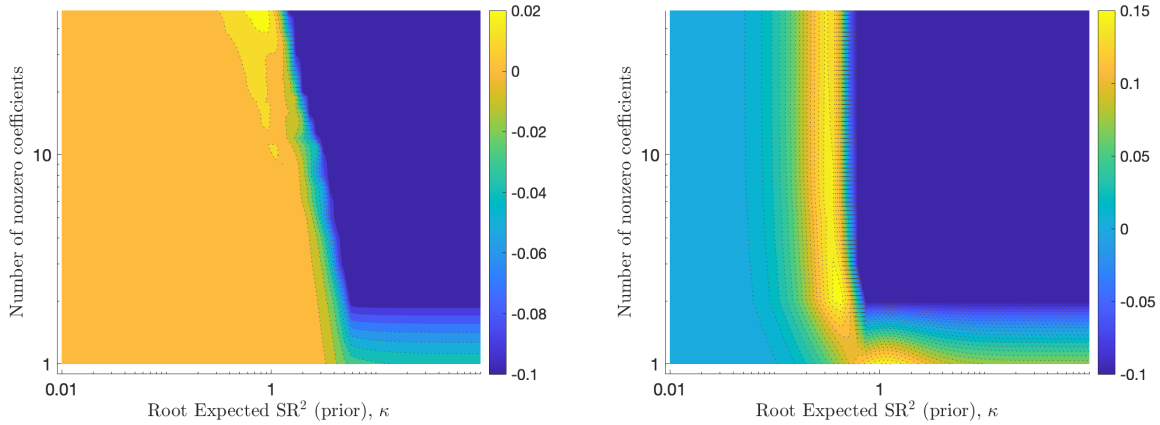
In this section, the results of applying the nonlinear model to the second dataset considered in this paper, the 50 anomaly portfolios, are discussed. In Figure 8, the cross-sectional OOS R^2 is plotted for various combinations of the degree of sparsity and shrinkage. In the left panel (a), the raw 50 anomaly portfolios are considered, and in the right panel (b) their PC transformation. Note that the coefficients are not the single risk prices, but the joint coefficients of the basis functions with the risk price of each factor, $c_i' b_i$.

From Figure 8a, it becomes clear that no sparsity can be attained while assuring a good OOS R^2. This supports the finding of Kozak et al. (2020) that the cross-section of asset returns cannot be summarized by only a few factors. Next to that, the dark blue plane on the right side of the map shows that a substantial degree of shrinkage should be imposed to attain a good OOS R^2. However, imposing too much shrinkage sets the R^2 equal to zero. This can be seen from the large orange plane on the left side of the contour map. Overall, the maximum R^2 of approximately 0.02 shows that pricing nonlinear factors has worse OOS performance than pricing linear factors, which was not to be expected from this dataset. Namely, previous research by for example Freyberger et al. (2020) has shown that asset-return data is subject to nonlinearities.

The above-mentioned results are rather different when transforming the 50 anomaly portfolios into their PCs first. Figure 8b shows that the model attains a high OOS R^2 even when allowing for substantial sparsity. This conclusion can be drawn from the almost vertical yellow plain. Moreover, when the SDF representation would contain only one factor, even less shrinkage needed to assure good OOS performance. The OOS R^2 is maximal at an SDF representation of two factors. Hence, the PCs of 50 anomaly portfolios can be explained best by two factors. Similar to the raw 50 anomaly portfolios, still substantial shrinkage should be applied to ensure good OOS performance.

Comparing Figure 8a and 8b, transforming the 50 anomaly portfolios into their PCs first results

in better performance of the nonlinear model, especially when the SDF representation is sparse.



(a) 50 anomaly portfolios

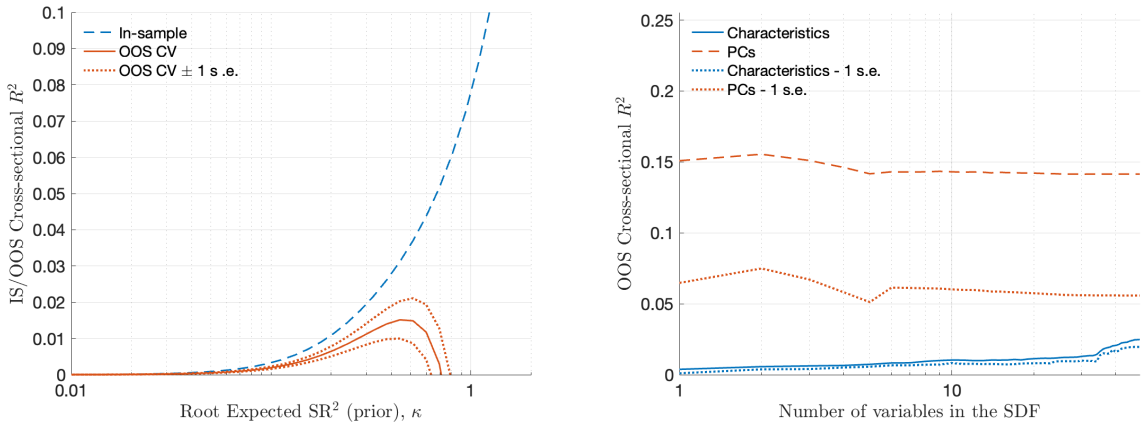
(b) PCs of 50 anomaly portfolios

Figure 8: Cross-sectional OOS R^2 for possible combinations of L^1 and L^2 penalties considering the nonlinear pricing model. The left figure (a) displays the OOS R^2 for 50 anomaly portfolios constructed by Kozak et al. (2020). The right figure (b) for their PC transformation. The x-axis in both figures shows the amount of shrinkage (L^2 penalty), with no shrinkage at the right and extreme shrinkage at the left edge of the figure. The y-axis shows the degree of sparsity (L^1 penalty), which is expressed as the number of nonzero coefficients. A small number indicates high sparsity, and a large number indicates almost no sparsity. The coloured contour map depicts the OOS R^2 for combinations of the L^1 and L^2 penalties, where the darker blue means the lower the R^2 and the lighter yellow the higher the R^2. The R^2 bounded at a minimum of -0.1.

Figure 9 shows some cuts along the axes of the contour maps in Figure 8. For the left panel (a), a cut is taken along the top of Figure 8a. This means that the cross-sectional R^2 of the raw 50 anomaly portfolios is displayed for various levels of shrinkage without imposing any sparsity. Figure 9a shows that a cross-sectional OOS R^2 of just below 0.02 is attained when no sparsity is imposed. This level of R^2 is attained for $\kappa \approx 0.44$. From this figure, it becomes again clear that pricing the nonlinear transformation of the 50 anomaly portfolios result in worse OOS performance compared to pricing the linear portfolios. Furthermore, the figure shows that using in-sample performance evaluation leads to different conclusions about the amount of shrinkage needed for the model to perform well out-of-sample. This shows that in-sample performance evaluation is not a representative measure of out-of-sample performance.

The right panel (b) of Figure 9 displays the maximum cross-sectional R^2 obtained at various degrees of sparsity, choosing the amount of shrinkage that maximizes the R^2 at a certain level of sparsity. This is done for both the raw 50 anomaly portfolios and their PC transformations. Thus, in figure 9b, a cut is taken from the bottom to the top of the maps in Figure 8 along all R^2 maximizing amounts of shrinkage. From 9b, it becomes clear that for the raw 50 anomaly portfolios, essentially more sparsity leads to worse out-of-sample performance, as the blue solid line shows a slowly increasing pattern over the number of variables included in the SDF. Moreover, transforming the 50 anomaly portfolios into their PCs first always results in a higher OOS R^2. However, contrary to the linear case, the R^2 first increases until two factors are included, then decreases again and then stays approximately equal in the number of variables included in the SDF.

In summary, the OOS performance of the nonlinear pricing model is worse than that of the linear pricing model, but transforming the portfolios into their PCs first achieves a higher out-of-sample performance compared to the raw portfolios and allows for a sparse SDF representation.



(a) Shrinkage (b) Sparsity

Figure 9: Cross-sectional OOS R^2 for various levels of shrinkage (a) and sparsity (b), based on the 50 anomaly portfolios constructed by Kozak et al. (2020). The left panel (a) plots the in-sample and OOS R^2 against the amount of shrinkage (L^2 penalty), without imposing sparsity. The dashed blue line displays the in-sample R^2, the solid red line the OOS R^2 and the dotted red lines the $\pm$ 1 standard error (s.e.) bounds of the OOS R^2. The right panel (b) plots the maximum OOS R^2 attained for optimal levels of shrinkage against the degree of sparsity. The solid blue line depicts the R^2 for the raw 50 anomaly portfolios, with the dotted blue line its -1 s.e. bound, and the dashed red line depicts the PCs of the 50 anomaly portfolios, with the dotted red line its -1 s.e. bound.

Next to evaluating the out-of-sample performance of the nonlinear SDF representation, it is interesting to see which variables included in the 50 anomaly portfolios have the highest explanatory power. Therefore, the ten characteristics and ten PCs of the 50 anomaly portfolios with the highest absolute t-statistics, ordered from highest to lowest based upon the absolute t-statistics corresponding to the first basis function, are presented in Table 2 and Table 3 respectively. The full table, including all 50 characteristics and PCs, are shown in Appendix F and G respectively.

The choice of ordering the characteristics and PCs of the 50 anomaly portfolios based upon the absolute t-statistics corresponding to the first basis function comes from the fact that the factor corresponding to the first basis function is equivalent to the linear factors. Therefore, choosing the t-statistics corresponding to the first basis function to order the anomalies gives the best comparison possible with the ten best anomalies resulting from the linear method. However, as mentioned earlier in the Methodology Section, the risk prices and coefficients corresponding to the basis functions cannot be distinguished. Thus, the values of the coefficients resulting from the nonlinear method cannot be compared directly with the linear method, but we can assess whether similar characteristics are present in the first ten anomalies with the most explanatory power.

Table 2 shows the ten characteristics that contribute most to the cross-sectional variation in asset returns, based upon their absolute t-statistics corresponding to the first basis function. Among these ten characteristics, the first five are momentum characteristics. Thus, momentum characteristics are

important in explaining the cross-section of nonlinear factors. Furthermore, some characteristics found among the most profound in the asset-pricing literature, such as *Price* and *Value-profitability* are present in Table 2 as well. However, some other characteristics that we would expect to be on this list are not, such as *Return on market equity* and *Investments/Assets*.

Table 2: The $\mathbb{C}b$ coefficients* and corresponding absolute t-statistics of the 10 factors with the highest explanatory power at the optimal level of L^2 shrinkage, ordered in a descending manner on their t-statistic value corresponding to the first basis function. In this table, the raw 50 anomaly portfolios are considered.

Variable name	Basis function 1		Basis function 2		Basis function 3		Basis function 4		Basis function 5	
	$C_l b$	t-stat	$C_l b$	t-stat	$C_l b$	t-stat	$C_l b$	t-stat	$C_l b$	t-stat
Value-momentum-prof.	-5.508	14.740	42.217	108.111	-0.046	0.118	-0.070	0.178	-0.070	0.178
Value-momentum	-5.171	13.808	51.366	131.631	-0.043	0.109	-0.065	0.166	-0.066	0.169
Momentum (6m)	-4.842	12.946	29.391	75.151	-0.030	0.075	-0.055	0.139	-0.058	0.148
Industry momentum	-4.698	12.640	36.786	94.180	-0.054	0.137	-0.075	0.190	-0.076	0.193
Momentum (12m)	-3.326	8.910	24.758	63.282	-0.005	0.012	-0.035	0.090	-0.042	0.108
Industry rel. reversals	-3.172	8.619	23.507	60.113	0.039	0.100	-0.007	0.017	-0.022	0.057
Price	-3.112	8.405	37.686	96.513	-0.001	0.001	-0.034	0.087	-0.044	0.113
Value-profitablity	-3.090	8.305	32.146	82.267	0.003	0.009	-0.041	0.103	-0.052	0.133
Industry rel. rev. (L.V.)	-3.036	8.300	9.802	25.021	0.136	0.347	0.055	0.141	0.021	0.054
Short-term reversals	-2.801	7.626	31.153	79.758	0.014	0.036	-0.022	0.055	-0.033	0.085

Note: *The coefficients $\mathbb{C}b$ are transformed into a matrix of which each column represents the coefficients in the table. This procedure is explained in Appendix E.

Table 3 shows the ten PCs that contribute most to the cross-sectional variation in asset returns, based upon their absolute t-statistic corresponding to the first basis function. Among the ten with the highest t-statistics, seven are out of the first ten PCs, and the first four are all among the first five PCs. This is in line with the finding that the first couple of PCs represent linear combinations of characteristics such that the combined variance is highest. Contrary to the linear method, the nonlinear model recovers the first PC as the one with the most explanatory power, which is expected from a statistical point of view. Overall, the PCs in Table 3 are of lower order than those in Table 1, which implies that the nonlinear method can recover the lower PCs better.

Supplementary to the tables, Figure 10 shows plots of the factor values with the most explanatory power according to the absolute t-statistic corresponding to the coefficients from the tables. This is done for the raw 50 anomaly portfolios in (a) and for their PC transformation in (b). The optimal factor of the raw portfolios, which can be seen from the tables as well, is *Value-momentum-prof* and of the PC transformed 50 anomaly portfolios PC1. In this higher-dimensional setting, compared to the FF25 portfolios, Figure 10 shows even better that the optimal PC factor captures way more variation than the raw optimal factor. This clearly visualizes that factors of the PC transformed portfolios can capture the cross-section of asset returns with only a few factors than better the raw factors.

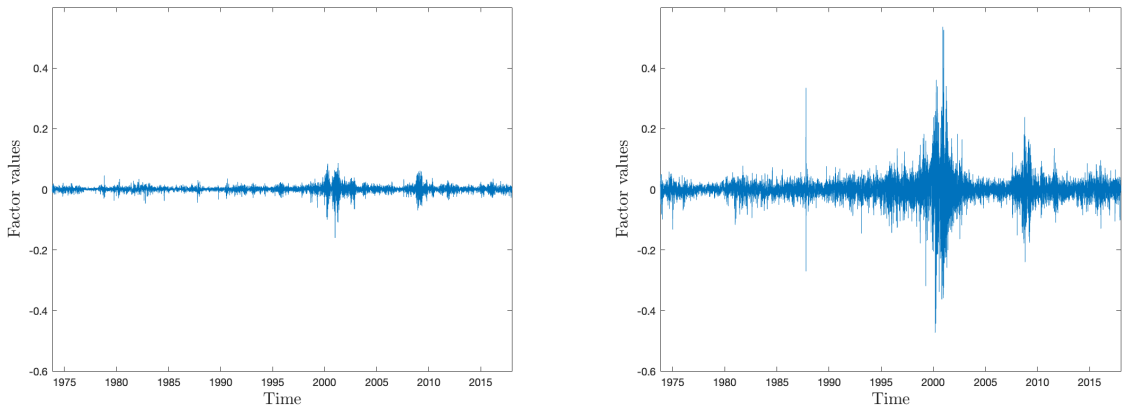
In summary, momentum characteristics are important in pricing nonlinear factors. Furthermore, some characteristics that are found among the most profound in the finance literature are included in

the list of characteristics with the highest explanatory power, but a few are missed. The PCs with the highest explanatory power are almost all low-order PCs, which is in line with the expectation from a statistical point of view.

Table 3: The $\mathbb{C}b$ coefficients* and corresponding absolute t-statistics of the 10 factors with the highest explanatory power at the optimal level of L^2 shrinkage, ordered in a descending manner on their t-statistic value corresponding to the first basis function. In this table, the PCs of the 50 anomaly portfolios are considered.

Variable name	Basis function 1		Basis function 2		Basis function 3		Basis function 4		Basis function 5	
	$C_l b$	t-stat	$C_l b$	t-stat	$C_l b$	t-stat	$C_l b$	t-stat	$C_l b$	t-stat
PC 1	-40.734	219.974	89.522	472.942	-5.367	23.675	-4.855	21.389	-4.336	19.084
PC 3	-10.390	51.147	70.319	325.416	-0.454	1.986	-0.428	1.872	-0.368	1.609
PC 2	-6.292	32.300	30.724	139.564	-0.768	3.364	-0.789	3.454	-0.759	3.325
PC 5	-0.573	2.666	9.395	41.252	0.063	0.276	0.008	0.035	-0.013	0.055
PC 17	0.539	2.386	2.328	10.182	0.014	0.063	0.004	0.019	0.000	0.002
PC 12	-0.488	2.179	1.510	6.603	0.035	0.155	0.018	0.078	0.010	0.042
PC 7	-0.431	1.955	8.338	36.536	0.059	0.258	0.029	0.125	0.014	0.063
PC 15	-0.375	1.662	1.673	7.318	0.018	0.079	0.004	0.017	-0.001	0.006
PC 9	0.300	1.346	3.561	15.581	0.034	0.149	0.014	0.059	0.005	0.022
PC 6	-0.244	1.120	4.869	21.328	0.072	0.314	0.027	0.119	0.008	0.034

Note: *The coefficients $\mathbb{C}b$ are transformed into a matrix of which each column represents the coefficients in the table. This procedure is explained in Appendix E.



(a) 50 anomaly portfolios

(b) PCs of 50 anomaly portfolios

Figure 10: Factor values of the factor with the most explanatory power according to the absolute t-statistic. for the 50 anomaly portfolios (a) and their PC transformation (b) plotted over time. The optimal factor is *Value-momentum-prof* for the raw 50 anomaly portfolios and PC1 for the PC transformations of the 50 anomaly portfolios. The factor values equal $F_{t,i} = Z'_{t-1,i} R^e_{t,i}$ where i is the optimal factor. The time equivalent to the sample size of the 50 anomaly portfolios data, which ranges from 1973 until 2017.

6 Discussion and conclusion

The main goal of this research was to identify whether and to what extent relaxing the linearity assumption of the method of Kozak et al. (2020) that the SDF is a linear combination of factors improves the explanation of the cross-section of asset returns. My interest in theoretically extending the methods from Kozak et al. (2020) was triggered by the paper of Gu et al. (2020), in which a lot of machine learning methods that can be used in asset-pricing are proposed and discussed. However, the methods by Gu et al. (2020) are proposed for a risk premia framework. Hence, I was especially interested in setting one of their models in a risk pricing framework. For this purpose, I have chosen to use the Generalized Linear Model, as this model conveniently provides a way to implement nonlinearities in the SDF representation. The method is convenient because the factors are evaluated for some nonlinear functions, which allows us to keep using a linear regression to estimate the coefficients.

To be able to compare the nonlinear model with the linear one, first the methods from the paper of Kozak et al. (2020) were implemented. From the empirical application of these methods using the Fama-French 25 portfolios, it can be concluded that shrinkage and sparsity are substitutes from one another. The more sparse the SDF representation, the less shrinkage is needed to achieve good out-of-sample performance and vice versa. Furthermore, transforming the factors into their Principal Components allows for a sparse SDF representation without additional shrinkage. An SDF representation of only one factor can explain the cross-section of asset returns well. These results were expected because the Fama-French portfolios possess a clear factor structure. Hence, the methods from Kozak et al. (2020) can recover this structure. Moving on to a higher dimensional dataset, the 50 anomaly portfolios are considered. From this application, it can be concluded that substantial shrinkage is needed to achieve good out-of-sample performance. Furthermore, transforming the factors into their Principal Components allows for a sparse SDF representation, whereas the raw factors do not allow for a sparse representation.

Secondly, the methods of the extension were implemented. The following conclusions can be drawn from the results of the empirical application of these methods. Using the Fama-French 25 portfolios, it can be concluded that substantial shrinkage is needed to achieve a positive out-of-sample R^2. Furthermore, contrary to the linear methods, shrinkage cannot be substituted with sparsity. The nonlinear method does not allow for a sparse SDF representation when considering the raw portfolios. However, transforming the factors into their PCs first does allow for a sparse SDF representation. In fact, an SDF representation of only one factor performs better than a representation including more factors. Moreover, the PC transformation generates a positive R^2 for an SDF representation of one factor even when no shrinkage is applied. Overall, the nonlinear method cannot outperform the linear method based on the out-of-sample R^2, particularly for the raw portfolios. This is a logical consequence when using the Fama-French 25 portfolios because those portfolios have a clear factor structure, which is better captured by the linear method.

Using the 50 anomaly portfolios, it can be concluded that again substantial shrinkage is needed to achieve a positive out-of-sample R^2. Moreover, no sparsity can be obtained to ensure good out-of-sample performance. These results are similar to the linear method. However, the R^2 is substantially lower than

in the linear case, which is not expected because from previous research of, for example, Freyberger et al. (2020) we know that asset returns possess nonlinearities. Transforming the portfolios into their PCs and applying substantial shrinkage results in good out-of-sample performance and allows for sparsity, even more than in the linear case.

In summary, the nonlinear method results in higher out-of-sample performance when transforming the portfolios to their PCs first compared to the raw portfolios. Still, overall, the nonlinear model performs worse than the linear model based on the cross-sectional out-of-sample R^2.

The main limitation of this paper is that the risk prices and the coefficients corresponding to the basis functions cannot be separated in the estimation of the coefficients. This causes the impossibility to retrieve the actual risk prices but the risk prices weighted by the coefficients corresponding to a basis function. Hence, the estimates cannot be interpreted directly as risk prices. Next to that, another limitation is the use of basis functions to approximate the nonlinearities in the portfolios. Using the Generalized Linear Method still restricts the number of possible nonlinear interactions. Using regression trees or neural networks would expand the number of possible nonlinear interactions substantially.

Although my results do not conclude that allowing for nonlinearities in the SDF representation leads to better out-of-sample performance, I think it is important to investigate this further because other papers have concluded that nonlinearities are important (Fama and French, 2008, and Bryzgalova et al., 2020). The way I allowed for nonlinearities in the SDF representation might not be the best. Thus, for further research, I propose incorporating another type of nonlinearity in the SDF representation. You could consider factors that are nonlinear in characteristics. This method looks promising, but I have not yet had the chance to implement it because it required retrieving the raw stock returns and characteristics, which was infeasible time-wise for this thesis. I have made a start on the theoretical setup of this approach in Appendix H.

References

Barillas, F., & Shanken, J. (2018). Comparing asset pricing models. *The Journal of Finance*, *73*(2), 715–754.

Bryzgalova, S., Pelger, M., & Zhu, J. (2020). Forest through the trees: Building cross-sections of stock returns. *Working paper, available at SSRN: https://dx.doi.org/10.2139/ssrn.3493458*.

Cochrane, J. H. (2005). *Asset pricing*. Princeton University Press.

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

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*(1), 3–56.

Fama, E. F., & French, K. R. (2008). Dissecting anomalies. *Journal of Financial Economics*, *63*(4), 1653–1678.

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

Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, *81*(3), 607–636.

Fan, J., Liao, Y., & Wang, W. (2016). Projected principal component analysis in factor models. *Annals of Statistics*, *44*(1), 219–254.

Feng, G., Giglio, S., & Xiu, D. (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance*, *75*(3), 1327–1370.

French, K. R. (2020). https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Freyberger, J., Neuhierl, A., & Weber, M. (2020). Dissecting characteristics nonparametrically. *The Review of Financial Studies*, *33*(5), 2326–2377.

Grith, M. (2021a). *Notes on pricing nonlinear factors*.

Grith, M. (2021b). *Notes on shrinking the cross-section by Serhiy Kozak, Stefan Nagel and Shrihari Santosh*.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

Gu, S., Kelly, B., & Xiu, D. (2021). Autoencoder asset pricing models. *Journal of Econometrics*, *222*(1), 429–450.

Hansen, L. P., & Jagannathan, R. (1991). Implications of security market data for models of dynamic economies. *Journal of Political Economy*, *99*(2), 225–262.

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

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

Kelly, B. T., Pruitt, S., & Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, *134*(3), 501–524.

Kozak, S., Nagel, S., & Santosh, S. (2018). Interpreting factor models. *Journal of Finance*, *73*(3), 1183–1223.

Kozak, S., Nagel, S., & Santosh, S. (2020). Shrinking the cross-section. *Journal of Financial Economics*, *135*(2), 271–292.

Lewellen, J. (2015). The cross-section of expected stock returns. *Critical Finance Review*, *4*(1), 1–44.

Lewellen, J., Nagel, S., & Shanken, J. (2010). A skeptical appraisal of asset pricing tests. *Journal of Financial Economics*, *96*(2), 175–194.

Light, N., Maslov, D., & Rytchkov, O. (2017). Aggregation of information about the cross section of stock returns: A latent variable approach. *The Review of Financial Studies*, *30*(4), 1339–1381.

Lintner, J. (1965). Security prices, risk and maximal gains from diversification. *Journal of Finance*, *20*(4), 587–615.

Moritz, B., & Zimmermann, T. (2016). Tree-based conditional portfolio sorts: The relation between past and future stock returns. *Working Paper, Available at SSRN: https://dx.doi.org/10.2139/ssrn.2740751*.

Mossin, J. (1965). Equilibrium in a capital asset market. *Econometrica*, *34*(4), 768–783.

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

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, *19*(3), 425–442.

Tibshirani, R., & Tibshirani, R. (2009). A bias correction for the minimum error rate in cross-validation. *The Annals Of Applied Statistics*, *3*(2), 822–829.

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

A Derivation SDF coefficients

This appendix shows the derivation of the SDF coefficients based on the unconditional asset pricing equation. The derivation is based on the notes provided by Grith (2021b).

$$0 = \mathbb{E}\left[M_t F_t\right]$$
$$0 = \mathbb{E}\left[(1 - b'(F_t - \mathbb{E}\left[F_t\right]))F_t\right]$$
$$0 = \mathbb{E}\left[(1 - b'(F_t - \mathbb{E}\left[F_t\right]))(F_t - \mathbb{E}\left[F_t\right]\right] + \mathbb{E}\left[M_t\,\mathbb{E}\left[F_t\right]\right]$$
$$0 = \mathbb{E}\left[(1 - b'(F_t - \mathbb{E}\left[F_t\right]))(F_t - \mathbb{E}\left[F_t\right]\right] + \mathbb{E}\left[M_t\right]\mathbb{E}\left[F_t\right]\right]$$
$$0 = \mathbb{E}\left[(1 - b'(F_t - \mathbb{E}\left[F_t\right]))(F_t - \mathbb{E}\left[F_t\right]\right] + \mathbb{E}\left[F_t\right]\right]$$
$$0 = \mathbb{E}\left[(F_t - \mathbb{E}\left[F_t\right]) - b'(F_t - \mathbb{E}\left[F_t\right])(F_t - \mathbb{E}\left[F_t\right]\right] + \mathbb{E}\left[F_t\right]\right]$$
$$0 = \mathbb{E}\left[F_t - b'(F_t - \mathbb{E}\left[F_t\right])(F_t - \mathbb{E}\left[F_t\right]\right]$$
$$0 = \mathbb{E}\left[F_t\right] - b'\,\mathbb{E}\left[(F_t - \mathbb{E}\left[F_t\right])(F_t - \mathbb{E}\left[F_t\right]\right]$$
$$\mathbb{E}\left[F_t\right] = b'\,\mathbb{E}\left[(F_t - \mathbb{E}\left[F_t\right])(F_t - \mathbb{E}\left[F_t\right]\right]$$
$$\mu = b'\Sigma$$
$$b = \Sigma^{-1}\mu$$

B Derivation estimator $\mathbb{C}b$

This appendix shows the derivation of the estimator of $\mathbb{C}b$ by making use of an economically motivated prior. Using the prior belief about P_v that

$$P_v \sim \mathcal{N}(0, \frac{\kappa^2}{\tau} R), \tag{B.1}$$

results in the posterior mean of P_v, which is derived from the formula of the conjugate multivariate normal prior where $R_0 = \frac{\kappa^2}{\tau} R^\eta$ and $P_{v,0} = 0$,

$$\hat{P}_v = (R_0^{-1} + TR^{-1})^{-1}(R_0^{-1} P_{v,0} + TR^{-1}\bar{P}_v) = (R + \gamma R^{2-\eta})^{-1} R\bar{P}_v \tag{B.2}$$

Plugging this formula in $\hat{\mathbb{C}}b = R^{-1}\hat{P}_v$ equals

$$\hat{\mathbb{C}}b = R^{-1}(R + \gamma R^{2-\eta})^{-1} R\bar{P}_v. \tag{B.3}$$

Using that $\eta = 2$ gives

$$\hat{\mathbb{C}}b = (R + \gamma \mathcal{I}_{NK})^{-1}\bar{P}_v, \tag{B.4}$$

where $\gamma = \frac{\tau}{\kappa^2 T}$ and $\mathcal{I}_{NK}$ is an $(N \times K)$ identity matrix. Thus, to each diagonal element of R, a penalty equal to γ is added. This is equivalent to

$$(\hat{\mathbb{C}}b) = \tilde{R}^{-1}\bar{P}_v, \qquad \tilde{R} = \begin{cases} \tilde{R}_{ii} = R_{ii}(1 + \gamma \mathcal{I}_K) & \text{if } i = j \\ \tilde{R}_{ij} = R_{ij} & \text{if } i \neq j \end{cases} \tag{B.5}$$

because each R_{ij} is a $K \times K$ variance-covariance matrix of each factor evaluated at K basis functions.

C Derivation out-of-sample R^2 with a nonlinear SDF

In this appendix, the derivation of the out-of-sample R^2 in case of a nonlinear SDF is shown. Using the definition from Equation (27) that $\mathbb{C}'\,\mathbb{E}[P_v] = \mathbb{C}'R\mathbb{C}b$, the out-of-sample R^2 can be expressed as

$$R^2_{OOS} = 1 - \frac{[\hat{\mathbb{C}}'(\bar{P}_{v,2} - \bar{R}_2\hat{\mathbb{C}}b)]'[\hat{\mathbb{C}}'(\bar{P}_{v,2} - \bar{R}_2\hat{\mathbb{C}}b)]}{[\hat{\mathbb{C}}'\bar{P}_{v,2}]'[\hat{\mathbb{C}}'\bar{P}_{v,2}]} \tag{C.1}$$

which can be written as

$$R^2_{OOS} = 1 - \frac{(\bar{P}_{v,2} - \bar{R}_2\hat{\mathbb{C}}b)'\hat{\mathbb{C}}\hat{\mathbb{C}}'(\bar{P}_{v,2} - \bar{R}_2\hat{\mathbb{C}}b)}{\bar{P}'_{v,2}\hat{\mathbb{C}}\hat{\mathbb{C}}'\bar{P}_{v,2}}. \tag{C.2}$$

Assuming that $\hat{\mathbb{C}}\hat{\mathbb{C}}' = \mathcal{I}$ results in the final definition

$$R^2_{OOS} = 1 - \frac{(\bar{P}_{v,2} - \bar{R}_2\hat{\mathbb{C}}b)'(\bar{P}_{v,2} - \bar{R}_2\hat{\mathbb{C}}b)}{\bar{P}'_{v,2}\bar{P}_{v,2}} \tag{C.3}$$

D Full 50 anomalies coefficients table

Table D4: The SDF coefficients and corresponding absolute t-statistics of all 50 anomaly portfolios at the optimal value of the prior root expected square Sharpe ratio, ordered in a descending manner on their t-statistic value. The raw 50 anomaly portfolios are considered on the left (a) and their PCs on the right (b).

(a) 50 anomaly portfolios			(b) PCs of 50 anomaly portfolios		
	b	t-stat		b	t-stat
Industry rel. rev. (L.V.)	-0.879	3.527	PC 4	1.014	4.249
Ind. mom-reversals	0.484	1.945	PC 1	-0.537	3.081
Industry rel. reversals	-0.425	1.705	PC 2	-0.556	2.653
Seasonality	0.322	1.292	PC 9	0.635	2.514
Earnings surprises	0.323	1.291	PC 15	-0.324	1.265
Value-profitablity	0.297	1.184	PC 17	0.303	1.182
Return on market equity	0.299	1.183	PC 6	-0.287	1.176
Investment/Assets	-0.238	0.948	PC 11	0.189	0.744
Return on equity	0.238	0.947	PC 13	0.166	0.654
Composite issuance	-0.240	0.947	PC 23	0.146	0.564
Momentum (12m)	0.227	0.906	PC 7	-0.140	0.561
Net issuance (A)	-0.215	0.851	PC 25	-0.135	0.523
Investment growth	-0.210	0.838	PC 22	0.124	0.479
Net issuance (M)	-0.205	0.810	PC 3	-0.084	0.379
Asset growth	-0.195	0.772	PC 28	0.087	0.336
Earnings/Price	0.195	0.770	PC 21	-0.086	0.334
F-score	0.193	0.769	PC 33	0.085	0.327
Accruals	-0.172	0.690	PC 29	0.079	0.305
Cash Flows/Price	0.171	0.678	PC 24	0.069	0.266
Return on assets	0.170	0.676	PC 5	-0.061	0.253
Short-term reversals	-0.162	0.651	PC 31	0.065	0.250
Sales/Price	0.162	0.642	PC 19	-0.057	0.220
Dividend/Price	0.143	0.567	PC 36	-0.055	0.213
Debt issuance	0.141	0.561	PC 35	-0.047	0.183
Share repurchases	0.134	0.531	PC 45	-0.047	0.180
Value (M)	0.123	0.487	PC 38	0.035	0.136
Industry momentum	0.121	0.485	PC 34	-0.034	0.130
Value (A)	0.121	0.477	PC 30	-0.033	0.126
Sales growth	-0.118	0.467	PC 18	-0.031	0.122
Momentum-reversals	-0.116	0.464	PC 12	0.030	0.118
Value-momentum-prof.	0.114	0.456	PC 26	-0.030	0.114
Gross profitability	0.114	0.453	PC 8	-0.028	0.110
Size	-0.111	0.444	PC 32	-0.028	0.109
Asset turnover	0.101	0.406	PC 14	-0.027	0.105
Long-run reversals	-0.098	0.391	PC 48	0.025	0.095
Idiosyncratic volatility	-0.098	0.384	PC 49	-0.025	0.095
Beta arbitrage	-0.088	0.347	PC 37	-0.024	0.092
Value-momentum	0.085	0.341	PC 39	-0.023	0.088
Investment/Capital	-0.084	0.331	PC 41	-0.021	0.080
Return on book equity (A)	0.082	0.328	PC 40	-0.020	0.078
Growth in LTNOA	-0.068	0.269	PC 46	0.020	0.076
Leverage	0.068	0.269	PC 20	0.017	0.065
Return on assets (A)	0.066	0.266	PC 10	-0.015	0.059
Share volume	-0.054	0.213	PC 44	0.013	0.051
Short interest	-0.044	0.176	PC 47	-0.013	0.050
Gross margins	0.041	0.162	PC 27	-0.012	0.045
Net operating assets	-0.031	0.124	PC 42	-0.010	0.038
Age	0.022	0.086	PC 43	0.007	0.029
Momentum (6m)	0.018	0.070	PC 16	0.004	0.016
Price	-0.002	0.007	PC 50	-0.002	0.008

E Coefficients transformation for tables of nonlinear factors

For the purpose of the visualization of the coefficients in a table, I have chosen to transform the estimated vector of coefficients $\mathbb{C}b$ such that each coefficient corresponds to a characteristic. Instead of $\mathbb{C}b$, which is an $(N \times K) \times 1$ vector, one column of the table corresponds to one column $C_l b$ of the matrix

$$C_l b_N = \begin{pmatrix} c_{1,1}b_1 & c_{2,1}b_1 & \dots & c_{K,1}b_1 \\ c_{1,2}b_2 & c_{2,2}b_2 & \dots & c_{K,2}b_2 \\ \vdots & \vdots & & \vdots \\ c_{1,N}b_N & c_{2,N}b_N & \dots & c_{K,N}b_N \end{pmatrix}, \tag{E.1}$$

where $C_l b_N$ is an $N \times K$ matrix of coefficients per basis function corresponding to each characteristic. Hence, in the tables, each vector of coefficients corresponds to one column $C_l b$ of the above matrix.

F Full 50 anomalies coefficients table for nonlinear factors

Table F5: The $\mathbb{C}b$ coefficients* for and corresponding absolute t-statistics of all 50 anomaly factors and basis functions at the optimal level of L^2 shrinkage, ordered in a descending manner on the t-statistic value of the first basis function.

Variable name	Basis function 1		Basis function 2		Basis function 3		Basis function 4		Basis function 5	
	$C_l b$	t-stat	$C_l b$	t-stat	$C_l b$	t-stat	$C_l b$	t-stat	$C_l b$	t-stat
Value-momentum-prof.	-5.508	14.740	42.217	108.111	-0.046	0.118	-0.070	0.178	-0.070	0.178
Value-momentum	-5.171	13.808	51.366	131.631	-0.043	0.109	-0.065	0.166	-0.066	0.169
Momentum (6m)	-4.842	12.946	29.391	75.151	-0.030	0.075	-0.055	0.139	-0.058	0.148
Industry momentum	-4.698	12.640	36.786	94.180	-0.054	0.137	-0.075	0.190	-0.076	0.193
Momentum (12m)	-3.326	8.910	24.758	63.282	-0.005	0.012	-0.035	0.090	-0.042	0.108
Industry rel. reversals	-3.172	8.619	23.507	60.113	0.039	0.100	-0.007	0.017	-0.022	0.057
Price	-3.112	8.405	37.686	96.513	-0.001	0.001	-0.034	0.087	-0.044	0.113
Value-profitablity	-3.090	8.305	32.146	82.267	0.003	0.009	-0.041	0.103	-0.052	0.133
Industry rel. rev. (L.V.)	-3.036	8.300	9.802	25.021	0.136	0.347	0.055	0.141	0.021	0.054
Short-term reversals	-2.801	7.626	31.153	79.758	0.014	0.036	-0.022	0.055	-0.033	0.085
Sales/Price	-2.318	6.112	35.340	90.338	-0.025	0.063	-0.059	0.151	-0.068	0.174
Return on equity	-1.905	5.098	9.062	23.123	0.044	0.112	-0.002	0.005	-0.018	0.046
Cash Flows/Price	-1.934	5.097	30.770	78.618	0.018	0.045	-0.025	0.063	-0.039	0.098
Ind. mom-reversals	-1.724	4.707	19.861	50.777	0.058	0.147	0.000	0.001	-0.021	0.054
Earnings/Price	-1.777	4.668	19.254	49.149	-0.007	0.018	-0.040	0.103	-0.050	0.129
Value (A)	-1.731	4.546	32.319	82.567	0.008	0.021	-0.031	0.080	-0.044	0.113
Return on book equity (A)	-1.663	4.484	18.726	47.855	0.055	0.140	-0.011	0.027	-0.035	0.089
Return on assets	-1.647	4.399	8.625	22.005	0.079	0.202	0.023	0.060	0.000	0.000
Asset turnover	-1.612	4.354	19.316	49.359	0.109	0.277	0.037	0.094	0.008	0.020
Leverage	-1.441	3.795	24.086	61.510	0.015	0.038	-0.026	0.065	-0.041	0.103
Age	-1.426	3.755	24.408	62.351	-0.019	0.048	-0.056	0.142	-0.069	0.176
Gross margins	1.378	3.704	7.632	19.472	0.140	0.357	0.057	0.147	0.020	0.050
Earnings surprises	-1.321	3.570	14.829	37.872	0.077	0.197	0.023	0.058	0.001	0.001
Return on market equity	-1.267	3.336	19.447	49.649	-0.020	0.051	-0.046	0.118	-0.054	0.138
Return on assets (A)	-1.233	3.314	14.025	35.809	0.127	0.324	0.044	0.113	0.010	0.025
F-score	-1.182	3.185	23.244	59.442	0.017	0.042	-0.030	0.077	-0.047	0.121
Accruals	-1.109	3.043	11.574	29.558	0.149	0.379	0.055	0.140	0.016	0.041
Size	1.094	2.975	17.525	44.801	0.109	0.278	0.026	0.065	-0.018	0.045
Investment/Capital	1.064	2.787	18.678	47.667	0.065	0.167	0.023	0.058	0.002	0.005
Value (M)	1.010	2.659	16.485	42.081	0.058	0.148	0.011	0.029	-0.012	0.029
Idiosyncratic volatility	0.983	2.566	19.096	48.723	0.058	0.148	0.021	0.052	0.004	0.010
Net issuance (A)	0.949	2.502	16.323	41.662	0.086	0.220	0.034	0.087	0.010	0.025
Net operating assets	-0.923	2.438	20.880	53.326	0.009	0.022	-0.039	0.100	-0.059	0.149
Long-run reversals	0.886	2.375	9.696	24.744	0.132	0.336	0.060	0.152	0.026	0.067
Short interest	-0.724	1.940	23.418	59.843	0.105	0.268	0.043	0.109	0.015	0.039
Asset growth	0.683	1.799	7.711	19.668	0.108	0.275	0.044	0.111	0.015	0.038
Seasonality	0.567	1.549	8.014	20.454	0.112	0.285	0.040	0.101	0.009	0.022
Beta arbitrage	0.565	1.470	8.699	22.183	0.047	0.120	0.021	0.053	0.008	0.021
Debt issuance	0.523	1.394	34.411	88.025	0.055	0.140	0.001	0.003	-0.021	0.053
Growth in LTNOA	-0.510	1.364	16.440	41.976	0.090	0.228	0.031	0.080	0.007	0.018
Gross profitability	-0.493	1.319	18.590	47.484	0.115	0.293	0.040	0.103	0.008	0.021
Investment growth	0.488	1.313	7.928	20.227	0.149	0.380	0.067	0.172	0.031	0.078
Share repurchases	-0.451	1.192	23.039	58.855	0.019	0.049	-0.020	0.051	-0.035	0.090
Sales growth	0.397	1.051	6.365	16.232	0.108	0.275	0.038	0.096	0.009	0.022
Share volume	0.388	1.013	15.794	40.291	0.054	0.138	0.019	0.048	0.003	0.008
Net issuance (M)	0.346	0.913	26.346	67.293	0.050	0.126	0.005	0.012	-0.014	0.035
Momentum-reversals	0.283	0.766	19.468	49.759	0.083	0.212	0.031	0.078	0.007	0.019
Investment/Assets	-0.250	0.671	9.167	23.390	0.136	0.347	0.055	0.140	0.020	0.052
Dividend/Price	-0.246	0.654	15.017	38.337	0.090	0.230	0.038	0.096	0.013	0.034
Composite issuance	0.199	0.525	14.160	36.132	0.092	0.235	0.034	0.086	0.008	0.020

Note: *The coefficients $\mathbb{C}b$ are transformed into a matrix of which each column represents the coefficients in the table. This procedure is explained in Appendix E.

G Full PCs coefficients table for nonlinear factors

Table G6: The $\mathbb{C}b$ coefficients* for and corresponding absolute t-statistics of all PC transformed 50 anomaly factors and basis functions at the optimal level of L^2 shrinkage, ordered in a descending manner on the t-statistic value of the first basis function.

Variable name	Basis function 1		Basis function 2		Basis function 3		Basis function 4		Basis function 5	
	$C_l b$	t-stat	$C_l b$	t-stat	$C_l b$	t-stat	$C_l b$	t-stat	$C_l b$	t-stat
PC 1	-40.734	219.974	89.522	472.942	-5.367	23.675	-4.855	21.389	-4.336	19.084
PC 3	-10.390	51.147	70.319	325.416	-0.454	1.986	-0.428	1.872	-0.368	1.609
PC 2	-6.292	32.300	30.724	139.564	-0.768	3.364	-0.789	3.454	-0.759	3.325
PC 5	-0.573	2.666	9.395	41.252	0.063	0.276	0.008	0.035	-0.013	0.055
PC 17	0.539	2.386	2.328	10.182	0.014	0.063	0.004	0.019	0.000	0.002
PC 12	-0.488	2.179	1.510	6.603	0.035	0.155	0.018	0.078	0.010	0.042
PC 7	-0.431	1.955	8.338	36.536	0.059	0.258	0.029	0.125	0.014	0.063
PC 15	-0.375	1.662	1.673	7.318	0.018	0.079	0.004	0.017	-0.001	0.006
PC 9	0.300	1.346	3.561	15.581	0.034	0.149	0.014	0.059	0.005	0.022
PC 6	-0.244	1.120	4.869	21.328	0.072	0.314	0.027	0.119	0.008	0.034
PC 8	0.238	1.075	6.931	30.355	0.060	0.262	0.027	0.119	0.012	0.053
PC 14	-0.225	1.002	2.686	11.749	0.011	0.049	-0.007	0.032	-0.013	0.057
PC 4	-0.179	0.838	18.335	80.870	-0.033	0.144	-0.072	0.316	-0.084	0.367
PC 29	0.169	0.743	0.583	2.550	0.008	0.035	0.003	0.014	0.001	0.005
PC 10	-0.150	0.673	3.589	15.705	-0.006	0.025	-0.025	0.110	-0.031	0.136
PC 25	-0.149	0.657	0.635	2.779	0.010	0.044	0.005	0.021	0.002	0.010
PC 28	0.116	0.508	0.394	1.721	0.007	0.028	0.002	0.007	0.000	0.002
PC 24	0.099	0.437	1.079	4.717	0.013	0.057	0.006	0.024	0.002	0.009
PC 31	0.094	0.413	0.322	1.408	0.006	0.027	0.003	0.013	0.001	0.006
PC 36	-0.087	0.380	0.201	0.877	0.003	0.013	0.001	0.005	0.000	0.001
PC 23	0.085	0.373	0.562	2.459	0.010	0.043	0.003	0.013	0.000	0.002
PC 27	0.074	0.327	0.815	3.563	0.006	0.028	0.002	0.009	0.000	0.002
PC 37	-0.074	0.324	0.270	1.180	0.002	0.008	0.000	0.001	0.000	0.001
PC 11	-0.059	0.263	2.688	11.758	0.029	0.125	0.006	0.024	-0.003	0.014
PC 35	-0.057	0.249	0.257	1.125	0.004	0.016	0.002	0.007	0.001	0.003
PC 41	-0.055	0.239	0.085	0.371	0.001	0.004	0.000	0.000	0.000	0.001
PC 32	-0.051	0.224	0.507	2.217	0.004	0.018	0.001	0.006	0.000	0.002
PC 13	0.050	0.222	3.056	13.371	0.036	0.159	0.018	0.077	0.009	0.040
PC 20	-0.047	0.207	1.022	4.469	0.016	0.068	0.006	0.026	0.002	0.008
PC 19	0.045	0.197	1.661	7.265	0.015	0.066	0.004	0.018	0.000	0.002
PC 40	0.038	0.165	0.155	0.677	0.002	0.009	0.001	0.004	0.000	0.001
PC 26	0.037	0.163	0.372	1.625	0.008	0.035	0.002	0.009	0.000	0.002
PC 34	-0.034	0.149	0.235	1.027	0.003	0.012	0.001	0.003	0.000	0.001
PC 48	0.033	0.146	0.041	0.179	0.001	0.003	0.000	0.001	0.000	0.000
PC 22	0.028	0.123	0.930	4.068	0.012	0.051	0.004	0.016	0.001	0.002
PC 30	-0.025	0.109	0.326	1.426	0.006	0.028	0.002	0.010	0.001	0.003
PC 44	0.020	0.088	0.080	0.349	0.001	0.006	0.001	0.002	0.000	0.001
PC 39	-0.018	0.078	0.285	1.247	0.002	0.010	0.001	0.004	0.000	0.001
PC 47	-0.017	0.076	0.024	0.105	0.001	0.004	0.000	0.002	0.000	0.001
PC 33	-0.017	0.074	0.412	1.803	0.004	0.017	0.002	0.008	0.001	0.004
PC 45	-0.015	0.068	0.138	0.605	0.001	0.004	0.000	0.001	0.000	0.000
PC 21	0.012	0.053	1.992	8.714	0.006	0.027	-0.002	0.009	-0.005	0.022
PC 43	0.012	0.052	0.098	0.430	0.001	0.005	0.000	0.002	0.000	0.000
PC 49	0.012	0.050	0.036	0.156	0.000	0.002	0.000	0.001	0.000	0.000
PC 42	0.008	0.036	0.138	0.604	0.001	0.005	0.000	0.001	0.000	0.000
PC 18	-0.007	0.033	3.277	14.334	0.014	0.063	0.003	0.013	-0.002	0.007
PC 50	0.003	0.014	0.018	0.081	0.000	0.001	0.000	0.000	0.000	0.000
PC 16	-0.003	0.013	3.096	13.542	0.011	0.049	-0.001	0.006	-0.006	0.028
PC 38	0.002	0.010	0.106	0.465	0.002	0.010	0.001	0.004	0.000	0.001
PC 46	-0.001	0.005	0.151	0.660	0.001	0.004	0.000	0.001	0.000	0.000

Note: *The coefficients $\mathbb{C}b$ are transformed into a matrix of which each column represents the coefficients in the table. This procedure is explained in Appendix E.

H Nonlinear risk prices

The extension of nonlinear risk prices lies in relaxing the assumption made in Kozak et al. (2020) that the risk prices are linear combinations of the characteristics, as is shown in Equation 5. Relaxing the assumption of linear risk prices leads to the following definition of the SDF,

$$M_t = 1 - b'_{t-1}(F_t - \mathbb{E}[F_t]), \tag{H.1}$$

where M_t is the SDF, $b_{t-1} = (b_{t-1,1}, b_{t-1,2}, ..., b_{t-1,N})'$ is an $N \times 1$ vector of factor loadings, each element corresponding to a factor $i = 1, ..., N$, and F_t is an $N \times 1$ vector of factors. When assuming that the factor loadings are nonlinear in characteristics, each element of b_{t-1} can be written as

$$b_{t-1,i} = f(Z_{t-1,i}) = \sum_{l=1}^{H} \sum_{j=1}^{K} a_{j,l} p_j(Z^l_{t-1,i}) = A'_v P_{Z_i,v}, \tag{H.2}$$

where Z_{t-1} is an $N \times H$ matrix of H characteristics observed for each of the N factors, A_v is the vectorization of the coefficient matrix A given below, resulting in A_v being a $(K \times H) \times 1$ vector of coefficients and $P_{Z_i,v}$ is a vectorization of the $K \times H$ matrix P_{Z_i} given below. Thus, $P_{Z_i,v}$ is a $(K \times H) \times 1$ vector. Each column of P_{Z_i} is a $K \times 1$ vector of basis functions $(p_1(\cdot), p_2(\cdot), ..., p_K(\cdot))'$. As basis functions, a spline series of order two, in the spirit of Gu et al. (2020), is chosen: $(1, Z_{t-1}, (Z_{t-1} - c_1)^2, (Z_{t-1} - c_2)^2, ..., (Z_{t-1} - c_{K-2})^2)$, where $c_1, c_2, ..., c_{K-2}$ are knots.

$$A = \{a_{j,l}\}_{j=1:K;l=1:H} = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,H} \\ a_{2,1} & a_{2,2} & \dots & a_{2,H} \\ \vdots & \vdots & & \vdots \\ a_{K,1} & a_{K,2} & \dots & a_{K,H} \end{pmatrix} \tag{H.3}$$

$$P_{Z_i} = \left\{p_{j,l} = p_j\left(Z^l_{t-1,i}\right)\right\}_{j=1:K;l=1:H} = \begin{pmatrix} p_1\left(Z^1_{t-1,i}\right) & p_1\left(Z^2_{t-1,i}\right) & \dots & p_1\left(Z^H_{t-1,i}\right) \\ p_2\left(Z^1_{t-1,i}\right) & p_2\left(Z^2_{t-1,i}\right) & \dots & p_2(Z^H_{t-1,i}) \\ \vdots & \vdots & & \vdots \\ p_K\left(Z^1_{t-1,i}\right) & p_K\left(Z^2_{t-1,i}\right) & \dots & p_K\left(Z^H_{t-1,i}\right) \end{pmatrix} \tag{H.4}$$

Thus, in order to estimate $b_{t-1,i}$, an expression of A_v is needed.

H.1 Derivation expression A_v

By the definition of expected returns, resulting from the fundamental asset pricing equation, it holds that for all $i = 1, ..., N$

$$E[F_{i,t}] = -\text{Cov}(F_{i,t}, M_t). \tag{H.5}$$

Writing the covariance matrix in terms of expected values and defining $\mu_i = E[F_{i,t}]$ gives

$$\mu_i = -\mathbb{E}\left[F_{i,t} - \mathbb{E}\left(F_{i,t}\right)\right][M_t - \mathbb{E}\left(M_t\right)]. \tag{H.6}$$

Plugging the definition of SDF, $M_t = 1 - b'_{t-1,i}(F_{t,i} - \mathbb{E}\, F_{t,i})$, into this equation results in

$$\mu_i = -\mathbb{E}\left[F_{i,t} - \mathbb{E}\, F_{i,t}\right][1 - b_{t-1,i}(F_{i,t} - \mathbb{E}\, F_{i,t}) - \mathbb{E}\left(M_t\right)] \tag{H.7}$$

Using that $\mathbb{E}(M_t) = 1$ and rewriting the equation such that $b_{t-1,i}$ is outside the expected value gives

$$\mu_i = b_{t-1,i}\,\mathbb{E}\,[F_{i,t} - \mathbb{E}\,F_{i,t}][F_{i,t} - \mathbb{E}\,F_{i,t}], \tag{H.8}$$

from which it becomes clear that the expectation can be written in terms of the variance of $F_{i,t}$ as

$$\mu_i = b_{t-1,i}\mathrm{Var}(F_{i,t}) \tag{H.9}$$

Isolating $b_{t-1,i}$ and defining the variance as $\sigma_i^2 = \mathrm{Var}(F_{i,t})$ results in

$$b_{t-1,i} = \frac{\mu_i}{\sigma_i^2}. \tag{H.10}$$

Now plugging in Equation (H.2), the definition of nonlinear risk prices, gives

$$f(Z_{t-1,i}) = A_v' P_{Z_i,v} = \frac{\mu_i}{\sigma_i^2}. \tag{H.11}$$

Isolating the coefficient matrix A_v' in this equation results in a final definition of A_v':

$$A_v' = \frac{\mu_i}{\sigma_i^2} P_{Z_i,v}(P_{Z_i,v}P_{Z_i,v})'. \tag{H.12}$$

Thus, to estimate the SDF coefficients when it is assumed that the risk prices are nonlinear in characteristics, the coefficient matrix A_v needs to be estimated and in turn can be used to retrieve the SDF coefficients when plugging A_v into Equation (H.2).