Kernel Principal Component Analysis for a Characteristics-Based Stochastic Discount Factor

Alexander Raaphorst (477556)
ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics
BSc^2 Econometrics / Economics
Supervisor: Dr. Maria Grith
Second assessor: Dr. Martina Zaharieva
Final version:

August 8, 2021

Abstract

In this paper, we replicate the results obtained by Kozak et al. (2020) and extend their methods. We evaluate the performance two different dimensionality reduction techniques, namely principal component analysis (PCA) and kernel principal component analysis (KPCA), for the estimation of the factor coefficients of a characteristics-based stochastic discount factor (SDF). For this, we use many different factor returns and compare the cross-sectional out of sample performance in terms of explanatory power. We impose different levels of shrinkage and sparsity on the factor coefficients, where the shrinkage is based on prior economic beliefs. We find that characteristics-sparse SDFs for which we use PCA and KPCA for the SDF coefficients, perform well for higher levels of sparsity. This is not the case when we do not use one of these methods. We find that KPCA performs worse than PCA when there is much redundancy between the different factor returns and shows a similar performance to PCA when there is almost no redundancy. A higher dimensionality might improve the performance of KPCA relative to PCA.

Contents

1 Introduction

In previous years, the availability of data has largely increased. For many applications, there is a lot of data, which means it can be used to improve previous outcomes of research. This also holds for asset pricing. There is however one large problem that occurs with the growing amount of available data, namely dimensionality. Models using a large amount of data need very high computational power, which is not always possible. Therefore, dimensionality reduction techniques exist to counter this problem. A well-known asset pricing model is the three factor model introduced by Fama and French (1993). However, as mentioned, there are many more factors available today to possibly improve asset pricing models. As mentioned by Kozak et al. (2020), there are many more cross-sectional characteristics to improve the predictions of such models. Evaluating the explanatory power of characteristics sparse models is the main focus of their research, specifically with the use of a Stochastic Discount Factor (SDF) they construct. They do this by using a large number of characteristics to then examine the amount of sparsity and shrinkage of the characteristics and the corresponding SDF loadings for which the model is still sufficiently explanatory, meaning the explanatory power does not differ too much compared to without sparsity and shrinkage. Furthermore, they apply a dimensionality reduction technique, namely Principal Component Analysis (PCA), to compare this with the characteristics based model. Their findings indicate that PCA is a useful method to reduce the dimensionality of the model, while still considering a large number of characteristics.

This paper will largely follow the research by Kozak et al. (2020) with an additional focus on dimensionality reduction techniques. There exist multiple dimensionality reduction techniques, of which Van Der Maaten et al. (2009) give a good overview. In previous research by Ince and Trafalis (2007), one of these different techniques, specifically Kernel PCA (KPCA) is used for stock price prediction, however not for the approach of an SDF. Besides this, there is not a lot of literature on different dimensionality reduction techniques used specifically for an SDF approach. Kozak (2019) does have a working paper on this subject, however there the kernel is applied to the raw characteristics, while in this paper we apply the kernel directly to the factor returns, which will be explained in more detail in this paper. Therefore, it would add to the existing literature to evaluate different techniques to attempt to improve the current SDF approach. Dimensionality reduction has the potential of becoming more important in the future, as more data becomes available. The main idea of this research will be how the cross-sectional performance of an SDF in a case of somewhat

high dimensionality can be improved with the use of different dimensionality reduction techniques. In addition to this, a comparison between these techniques will be made and the model will be compared with different data sets. The dimensionality reduction technique that will be considered is KPCA with different kernels.

This research could provide beneficial results for both scientific as practical applications. If these dimensionality reduction techniques work well for this specific application, future research could focus on applying these techniques to different dimensionality problems within the financial sector. Furthermore, not only is high dimensionality a problem for research, but also for companies that perform large analyses. Should these dimensionality reduction techniques prove to be useful, they could be applied in practical financial applications, or even outside the financial sector.

Previous research by Kan and Zhou (1999) gives some critique to the SDF method. They find that the risk premium estimate of the SDF performs worse in terms of standard error compared to traditional asset pricing methods, as the standard errors corresponding to the SDF are far higher. Besides this, they also find that in specification tests, the traditional methods, which in this case are linear factor models, perform better than the SDF method. Overall, they have a higher power, specifically for the rejection of incorrectly specified models.

In contrast to this, Jagannathan and Wang (2002) show us that the SDF method is a useful method to be considered in asset-pricing models. They compare this method to the well-known beta method used for the estimation of risk premiums. Their findings suggest that for this purpose, the SDF method is equally efficient and as powerful, based on specification tests. The research by Kozak et al. (2020) however, focuses on the estimation of risk prices rather than risk premia. The reason for this is that they want to focus on characterizing the SDF.

In a more recent and slightly different study by Lettau and Pelger (2020), an improved adaptation of PCA is created for the application of asset pricing. As normal PCA is very good in finding factors that have a large variance and contribute strongly through this variance, there might be some factors that have a small variance, yet contribute strongly to the asset pricing. The estimator created by Lettau and Pelger (2020) is able to identify these factors and is for this specific application and improvement on PCA. Even though this method is not used for characteristics based SDF and the goal of sparsity and shrinkage, it is a relevant finding. The problem that normal PCA has, namely that the empirical estimate of the covariance function might be poor, is of importance for this research. The use of different versions of PCA might overcome this problem in this case.

The rest of this paper is structured as follows. In Section 2, the methodology will be explained,

both the theoretical framework as the estimation for the replication of the research by Kozak et al. (2020), and the extension. After this, in Section 3, data will be briefly touched upon. Following this, in Section 4, the results will be given and discussed. Finally, the conclusion and a discussion will be provided in Section 5.

2 Methodology

The basis of the methods will be the same as the methods used by Kozak et al. (2020). For this research, I will briefly mention the main steps in their methodology. Following this, their methodology will be discussed, both theoretical framework as estimation method. Finally, both the theoretical framework and estimation method of the extension will be explained.

2.1 Replication

With the data described in the Section 3, the following steps are taken by Kozak et al. (2020). First, using the assumption that the covariance matrix is known and prior economic knowledge, the distribution of both the mean of the portfolios and the Sharpe ratios of the principal components (PCs) are obtained. This allows us to estimate the estimator b of the SDF coefficients. This in turn lets us construct two penalties to implement in the maximization of the cross-sectional R^2. These penalties are in place for shrinkage and sparsity. Using K-fold cross-validation, the parameters corresponding to the penalties can be estimated. The results which follow from the shortly described steps can then be analysed.

2.1.1 Theoretical framework

In this section, the necessary explanation of the theories used in this research will be given. The first important theory is that of the fundamental pricing equation. Kozak et al. (2020) mention the conditional and unconditional asset pricing equations, however, these are derived from the fundamental pricing equation. This equation is as follows.

$$S_t = e^{-r(T-t)} E_t(M_{t,T} S_T) \quad \text{and} \quad E_t(M_{t,T}) = 1. \tag{1}$$

In this equation, S_t is an $N \times 1$ vector consisting of stock prices of N stocks (N is thus the number of stocks) at time t, r is the annual risk-free rate, continuously compounded, E_t is the expectation with the information available up until time t and $M_{t,T}$ is the stochastic discount

factor. This discount factor is intertemporal between time t and T. From this equation, the following conditional pricing equation can be derived. The derivation can be found in the Appendix in Equation (33), in which s_t is the discounted stock price.

$$E_t(M_{t,T}R_{t,T}) = 0. \tag{2}$$

Here, $R_{t,T}$ is the return in excess of the risk-free rate, which is an $N \times 1$ vector for the N stocks. This equation is also given by Kozak et al. (2020) with slightly different notations. To find the SDF $M_{t,T}$, we must use the assumption that this SDF spans the linear space of the stocks' excess returns. This is shown in the following equation.

$$M_{t,T} = 1 - b_t'(R_{t,T} - E_t(R_{t,T})). \tag{3}$$

Here b_t, which is an $N \times 1$ vector, is time-varying and denotes the SDF coefficients for $R_{t,T}$. This equation can be found in the paper by Kozak et al. (2020), however, the notation is slightly different. Now that the linear SDF formula is given, the following step is transforming this to a characteristics-based SDF.

If we have the $N \times H$ matrix of characteristics Z_t, we can compute the $H \times 1$ vector of factor returns $F_{t,T}$. Let H be the number of characteristics for each stock. We do assume in this case that the observable factors are the same as the characteristics-based factors. This simplifies the problem, as the characteristics can simply be observed. The following equations show the factor returns in relation to the excess returns and characteristics.

$$F_{t,T} = Z_t'R_{t,T} \quad \longleftrightarrow \quad R_{t,T} = (Z_tZ_t')^{-1}Z_tF_{t,T}. \tag{4}$$

The following assumption must also be made to complete the characteristics-based SDF model. Namely, that the SDF coefficients depend on the stock characteristics. For this, we must introduce new SDF coefficients b for the factor returns $F_{t,T}$. These coefficients are not time-varying and is thus a vector of dimensions $K \times 1$, where K is the number of factors. In this case, H equals K. The following equations show the relation between the two different SDF coefficients.

$$b_t = Z_tb \quad \longleftrightarrow \quad b_t' = b'Z_t'. \tag{5}$$

Now we can rewrite Equation (3) as follows;

$$M_{t,T} = 1 - b'(F_{t,T} - E(F_{t,T})). \tag{6}$$

It should be noted that for this, we use that $Z_t'(Z_t Z_t')^{-1} Z_t = 1$. We can also rewrite Equation (2) into the following, as Kozak et al. (2020) do.

$$E(M_t F_t) = 0. \tag{7}$$

Using Equations (6) and (7) and the assumptions made, we can find the estimate of the SDF coefficients b. Now the following holds for each factor $h \in \{1, ..., H\}$.

$$E_t([1 - b'(F_{t,T} - E_t(F_{t,T}))]F_{t,T}^h) = 0. \tag{8}$$

Rewriting this Equation will result in the following matrix notation for all h, for which the full derivation can be found in the Appendix in Equation (34) and Equation (35).

$$\mu = b' V_t(F_{t,T} F_{t,T}') = \Sigma_F b \quad \text{which gives} \quad b = \Sigma_F^{-1} \mu. \tag{9}$$

For the derivation of these equations, we also use that $\mu = E_t(F_{t,T})$. From Equation (9), we can see that the mean and the variance of the SDF depend on the first two moments of the factors' returns, μ and Σ_F respectively. Furthermore, for this derivation, the assumption that the conditional first two moments of the factors are constant must be made.

Following this, Kozak et al. (2020) move on to sparsity in characteristics-based factor returns, however, they do not find convincing reasons for sparsity in the characteristics of the SDF. The idea they give is instead of imposing sparsity, allowing sparsity and then evaluating the level of sparsity empirically. Their next step is to turn to PCA, driven by the findings of Kozak et al. (2018).

2.1.1.1 Principal Component Analysis

Kozak et al. (2020) give two conditions taken from Kozak et al. (2018) as to why they explore the possibility of a PC sparse SDF instead of a characteristics sparse SDF. First, the absence of near-arbitrage opportunities, from which Kozak et al. (2018) take away that factors connected to high risk largely influence covariation. Second, a few principal components (PCs) with a high variance are leading in the factors of asset returns. In this case, the few PCs with the high variance should

be able to account for the largest part of the cross-sectional variation in the expected returns.

The equations which follow are from Kozak et al. (2020). For the first step of PCA, we apply eigendecomposition to the $H \times H$ covariance matrix Σ of the factors' returns.

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

As eigendecomposition already says, the covariance matrix Σ is decomposed into an $H \times H$ matrix Q consisting of the eigenvectors of Σ, and a diagonal matrix D consisting of the H eigenvalues in decreasing order. Now we construct the $H \times 1$ vector $P_{t,T}$ of PC factors and express the SDF in terms of PC factors, from which we can clearly see that the PC factors are linear combinations of the factor returns.

$$P_{t,T} = Q' F_{t,T}, \tag{11}$$

$$M_{t,T} = 1 - b_p'(P_{t,T} - E(P_{t,T})), \quad \text{with} \quad b_p = D^{-1}E(P_t). \tag{12}$$

The next point of interest is shrinkage and the family of priors Kozak et al. (2020) construct in combination with the assumption that Σ is known. They give the following.

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

Here, τ is the trace of the covariance matrix Σ, $\tau = tr[\Sigma]$, and κ and η are constants. These constants determine the scale and shape respectively of the prior. From Harvey and Zhou (1990), we know that in the Bayesian approach, belief plays a big role in defining the probability.

In terms of PCA, we can write Equation (13) as follows, as done by Kozak et al. (2020).

$$\mu_p \sim \mathcal{N}(0, \frac{\kappa^2}{\tau}D^\eta). \tag{14}$$

Furthermore, we can easily modify this so we get the distribution of Sharpe ratios with respect to the PCs.

$$D^{-\frac{1}{2}}\mu_p \sim \mathcal{N}(0, \frac{\kappa^2}{\tau}D^{\eta-1}). \tag{15}$$

Regarding the choice of the value for the parameter η controlling the shape, Kozak et al. (2020)

provide the interpretation of different values and their implications based on previous research. Their choice falls on the setting $\eta = 2$.

Based on the assumption of the value of η, Kozak et al. (2020) obtain the following prior on SDF coefficients, namely $b \sim \mathcal{N}(0, \frac{\kappa^2}{\tau} I)$. This is independent and identically distributed.

As Kozak et al. (2020) do, we can now use the prior beliefs to get to the following posterior mean and variance of b. For this, we must assume a multivariate-normal likelihood and use its sample means $\overline{\mu}$, where the sample size is T.

$$\hat{b} = (\Sigma + \gamma I)^{-1}\overline{\mu} \quad \text{and} \quad \text{var}(b) = \frac{1}{T}(\Sigma + \gamma I)^{-1}, \quad \text{where} \quad \gamma = \frac{\tau}{\kappa^2 T}. \tag{16}$$

For a better economic interpretation, Kozak et al. (2020) modify this to fit the space of PCA. From Kozak et al. (2020), we know that the sample form of the b in Equation (9), is as follows.

$$\hat{b} = \overline{\Sigma}_F^{-1}\overline{\mu}, \quad \text{which results in} \quad \hat{b}_{P,j}^{\text{OLS}} = \frac{\overline{\mu}_{P,j}}{d_j} \quad \text{for the PCs.} \tag{17}$$

However, rotating this into the space of PCs and using Equations (16) and (11) with $\hat{b}_P = Q'\hat{b}$, we get the following.

$$\hat{b}_{P,j} = (\frac{d_j}{d_j + \gamma})\frac{\overline{\mu}_{P,j}}{d_j}. \tag{18}$$

The shrinkage of the SDF coefficients can now clearly be seen, as we already set $\gamma > 0$, resulting in $\frac{d_j}{(d_j+\gamma)} < 1$, which in turn shrinks the coefficients. From this shrinkage equation, we can clearly see that it shrinks PCs with lower eigenvalues more than those with high eigenvalues. This means low eigenvalue PCs contribute less to the volatility of the SDF.

2.1.2 Estimation

In the estimation, Kozak et al. (2020) make use of a dual-penalty method, namely the following.

$$\hat{b} = \arg\min_b (\overline{\mu} - \Sigma b)'\Sigma^{-1}(\overline{\mu} - \Sigma b) + \gamma_2 b'b + \gamma_1 \sum_{i=1}^{H} |b_i|. \tag{19}$$

This method consists of three parts, namely the minimization of the model HJ-distance from Hansen and Jagannathan (1991) and two penalties, one L^1 norm penalty and one L^2 norm penalty. This equation is in essence a single-penalty method containing only the L^2 norm penalty $\gamma_2 b'b$, to which an L^1 penalty $\gamma_1 \sum_{i=1}^{H} |b_i|$ has been added. Kozak et al. (2020) note that the single-penalty

method would result in the same solution as in Equation (16). Furthermore, they provide us with the intuition behind the L^2 norm penalty. Since we know that PCs with low eigenvalues have a rather low contribution to the volatility of the SDF, if we shrink the corresponding coefficient $\hat{b}_{P,j}$, only a small source of the volatility of the SDF will be removed, while we do reap the benefits of shrinking the coefficient. If we shrink a coefficient corresponding to a PC with a high eigenvalue by the same amount, we would obtain the same benefits in terms of the penalty. However, as we know that PCs with a high eigenvalue contribute a lot to the volatility of the SDF, shrinking the corresponding coefficient would remove a rather big source of the volatility of the SDF. This means that this penalty will shrink coefficients corresponding to PCs with lower eigenvalues more, which is what we want, as these only have a small contribution to the volatility of the SDF.

Besides the L^2 norm penalty, which imposes shrinkage, sparsity must still be imposed in some way. This is where the L^1 norm penalty comes from. Kozak et al. (2020) choose to add this penalty to impose sparsity, motivated by Zou and Hastie (2005). The L^1 norm penalty is able to set some of the coefficients equal to zero because of the way it is constructed. The benefit of setting the coefficient equal to zero is in that case bigger than the cost of keeping the coefficient on a non-zero value. To implement this method, Kozak et al. (2020) make us of an algorithm provided by Zou and Hastie (2005).

The amount of shrinkage or sparsity imposed is thus dictated by the strength of the L^1 and L^2 norm penalties. Kozak et al. (2020) also mention the importance of both of the penalties together as opposed to only the L^1 norm penalty. Even only the L^2 norm penalty would perform better than the L^1 norm penalty, especially in cases where there is a high correlation between the explanatory variables. This is rather straightforward, as the L^2 norm penalty will shrink each coefficient, while still conserving the joint explanatory power. The L^1 norm penalty would disregard this and set all coefficients to zero except for one.

Now for the estimation of the penalty parameters, Kozak et al. (2020) explain that since this method uses two different penalty parameters γ_1 and γ_2, which are also different than the penalty parameter γ specified in Equation 16, these values must be set beforehand. Using a method with only an L^2 norm penalty would result in only one penalty parameter γ, which would then simply be the parameter as specified in Equation 16.

Kozak et al. (2020) show that because of their choice of $\eta = 2$ and the family of priors in Equation 13, the following equation of the root expected maximum squared Sharpe ratio is very

useful in terms of interpretation.

$$E[\mu\Sigma^{-1}\mu]^{\frac{1}{2}} = \kappa. \tag{20}$$

Here, μ and Σ are taken from the family of priors in Equation (13) and κ now represents the root expected maximum squared Sharpe ratio. Kozak et al. (2020) proceed to show that from Equation (16), we know that $\gamma = \frac{\tau}{\kappa^2 T}$, which gives us a better interpretation. If we expect the Sharpe ratio to be very high, hence a high value for κ, the parameter γ is then low, indicating a low amount of shrinkage. The reverse also holds. The prior beliefs dictate this reasoning, which could thus be altered depending on the beliefs. This makes this method rather complicated and therefore, Kozak et al. (2020) choose a different approach, which is the following. They choose to estimate γ through K-fold cross-validation. It should be noted that Equation (20) is used to indicate the strength of the L^2 norm penalty, for interpretation purposes.

This estimation method works as follows, as explained by Kozak et al. (2020). The data set is divided into K samples of equal size. Then, the $\hat{b}$ is estimated. As mentioned before, we must set the values of parameters γ_1 and γ_2, thus the estimation of the $\hat{b}$ is done for each possible value of those parameters. For the estimation, Equation (16) is used for each sample except for one, thus to K - 1 samples. Now, with the resulting model, the out of sample (OOS) R^2 of the sample that was not included in estimating $\hat{b}$, can be computed with the following equation as defined by Kozak et al. (2020).

$$R^2_{\text{OOS}} = 1 - \frac{(\overline{\mu}_2 - \overline{\Sigma}_2\hat{b})'(\overline{\mu}_2 - \overline{\Sigma}_2\hat{b})}{\overline{\mu}_2'\overline{\mu}_2}. \tag{21}$$

Here, R^2_{OOS} is the out of sample R^2 and the sample moments with a subscript 2 are those corresponding to the excluded sample. Since only one of the K samples was excluded, this method must be repeated K times, excluding a different sample each time. To choose the optimal values for the parameters γ_1 and γ_2, we must average the values found in Equation (21) across all of the K repetitions of the method and choose the parameter values that maximize this average. This average is the OOS R^2 which has been cross-validated. We base our choice of K on the value chosen by Kozak et al. (2020), which is $K = 3$.

Furthermore, Kozak et al. (2020) mention the OOS R^2 bias this method produces. However, for their research, this does not matter, as the OOS R^2 is compared between different models and different values of shrinkage and sparsity. The bias is present for each of these different models and

penalties, thus they can still compare the performance between these differences. The same holds for our research.

Finally, to compare the different methods, we compare the Sharpe ratios of the mean-variance efficient (MVE) portfolios obtained. Kozak et al. (2020) use the SDF coefficients as weights for the MVE portfolio, which we will also do. For the comparison, we determine the optimal value for the L^2 norm penalty when no sparsity is imposed, and with this, we determine the corresponding Sharpe ratio.

For the estimation, we use the MATLAB code provided by Kozak et al. (2020), where we added some code to be able to produce certain figures.

2.2 Extension

2.2.1 Theoretical framework

The central theme Kozak et al. (2020) address is dimensionality reduction. To achieve dimensionality reduction and sparsity, they consider PCA, which is a very popular method for dimensionality reduction. There are however multiple dimensionality reduction techniques. A very clear review of different techniques is provided in the paper by Van Der Maaten et al. (2009). Here they consider only convex techniques and make a distinction between methods that work if there is a full matrix available and methods that function if there is only a sparse matrix available. A full matrix is a matrix with values for each entry, even if the value is zero, while a sparse matrix does not store values of zero, instead, the entry will be empty. For simplicity, I will consider only the methods that make use of a full matrix.

The dimensionality reduction technique additional to PCA that will be used in this research is Kernel PCA (KPCA). This is very similar to the method used by Kozak (2019), which is still a working paper. In this method, the principal components are not of the covariance matrix, but rather of the kernel matrix (Van Der Maaten et al., 2009). As Van Der Maaten et al. (2009) and Ince and Trafalis (2007) mention, KPCA is a nonlinear version of PCA. This is nonlinear in the following way. For KPCA, the eigenvalues and eigenvectors are calculated based on a kernel matrix, which consists of non-linear dependencies between the different factor returns. The factor returns are thus combined in a nonlinear manner through a kernel function, resulting in a kernel matrix. Once we have this kernel matrix, the principal components of this matrix are computed in a linear manner. Nothing changes in the data generating process of the factor returns, only after we have

the factor returns, are these kernelized and used for PCA. For normal PCA, the eigenvalues and eigenvectors are calculated based on either the covariance matrix or correlation matrix. These are both linear dependencies between the different stock returns. The non-linearity of KPCA is the motivation of Kozak (2019) for using KPCA. The most important implication of using KPCA is the ability to overcome the curse of dimensionality. This is especially useful when interactions of characteristics are used, such as Kozak et al. (2020) do. While Kozak (2019) apply the kernel to the characteristics, which means that the SDF stays linear in terms of the individual stock returns, we apply the kernel directly to the portfolio factor returns. We thus create kernel matrix through nonlinear combinations of these portfolio factor returns and then apply PCA to this matrix, from which we can create the PCs, which in turn are linear combinations of the factor returns. The weights of these linear combinations thus come from nonlinear combinations of the factor returns. This is further explained in Section 2.2.1.1. Following Van Der Maaten et al. (2009), first, the kernel matrix of the data points must be computed in the following way.

$$k_{ij} = \kappa_{kernel}(R_{t,i}, R_{s,j}). \tag{22}$$

This equation gives the entries for the kernel matrix by using the kernel function κ_{kernel}. Here we use the subscript "kernel", to distinguish this κ from the previously introduced κ in Equation (13). This is a function where the input consists of two vectors. From these vectors, a nonlinear transformation is created, where the output is a single constant. In some cases, the transformation can be simply linear. It thus maps the data to a new data space. Furthermore, $R_{t,i}$ and $R_{s,j}$ are the returns of stocks i and j in excess of the risk-free rate at time t and s, thus part of the earlier defined matrix $R_{t,T}$. In our case, we can plug in the factor returns $F_{t,T}$ for each portfolio in Equation (22). If this is done for each combination of factor returns, we have an entry k_{ij} for all i and j, which means we can create the $H \times H$ kernel matrix $\mathbf{K}$. Equation (22) can be written in matrix form in the following way, where we already use $F_{t,T}$ as input instead of R_t.

$$\mathbf{K} = \mathcal{K}(F_{t,T}, F_{s,S}). \tag{23}$$

Here, the function $\mathcal{K}$ is actually a matrix consisting of the functions κ_{kernel}, which create each entry k_{ij} of the kernel matrix $\mathbf{K}$. This can be done for any kernel function, as each entry of the matrix consists of the same kernel function. The entries of the $H \times H$ kernel matrix $\mathbf{K}$ are thus $k_{ij} = \kappa_{kernel}(F_{t,T,i}, F_{s,S,j})$ as noted by Kozak (2019), where i and j denote the portfolio and both

t, T and s, S denote the time. There exist multiple kernel functions, I will use multiple kernels to also compare these, namely the linear kernel, the Gaussian kernel, the Polynomial kernel and possibly more. The linear kernel is added to evaluate whether the KPCA works correctly because, as mentioned by Van Der Maaten et al. (2009), the use of this kernel in KPCA is equivalent to normal PCA.

2.2.1.1 Kernel operations

As given by Amari and Wu (1999) and slightly modified to fit our notations, the kernel functions are defined as follows. We define $\mathbf{f}_i$ and $\mathbf{f}_j$ to be vectors of values of factor returns for portfolios i and j, where we omit the subscripts t, T and s, S for simplicity.

First, we have the Gaussian Radial Basis Function (Gaussian RBF):

$$\kappa_{kernel}(\mathbf{f}_i, \mathbf{f}_j) = \exp(-c \parallel \mathbf{f}_i - \mathbf{f}_j \parallel^2). \tag{24}$$

Here, $c = \frac{1}{2\sigma^2}$ for the Gaussian RBF, however, Kozak (2019) choose to define c as a constant with the value 0.5. We will do the same. The Polynomial kernel of degree d is as follows.

$$\kappa_{kernel}(\mathbf{f}_i, \mathbf{f}_j) = (c + \langle \mathbf{f}_i, \mathbf{f}_j \rangle)^d. \tag{25}$$

In this case, Amari and Wu (1999) choose $c = 1$, however Kozak (2019) define c as a free parameter. We will use $c = 1$ for simplicity. The Linear kernel should be the Polynomial kernel with $d = 1$, thus resulting in the following.

$$\kappa_{kernel}(\mathbf{f}_i, \mathbf{f}_j) = c + \langle \mathbf{f}_i, \mathbf{f}_j \rangle. \tag{26}$$

Weinberger et al. (2004) however, do not include the constant c in the Linear kernel. From Van Der Maaten et al. (2009), we can reason that setting $c = 0$ would result in what was mentioned before, namely that this kernel in KPCA would give the same results as normal PCA. Therefore, we set $c = 0$. We will also test for $c = 1$.

Van Der Maaten et al. (2009) then proceed to modify the kernel matrix in order to obtain a zero mean kernel function space. They do this with the following function:

$$k_{ij} = -\frac{1}{2}(k_{ij} - \frac{1}{n}\sum_l k_{il} - \frac{1}{n}\sum_l k_{jl} - \frac{1}{n^2}\sum_{lm} k_{lm}). \tag{27}$$

The intuitive interpretation of Equation 27 is as follows. The kernel function κ_{kernel} defines a space, this space is visualized in the kernel matrix. This kernel space has a mean, and thus in Equation 27, the mean of this space is subtracted from the kernel matrix. This is done individually for each entry.

Now that the kernel matrix is computed, the eigenvalues and eigenvectors of this matrix need to be computed. This is rather straightforward, as this is done in the same way as for normal PCA, except we now have a different matrix. Before, we applied eigendecomposition to the covariance matrix Σ, however, now we apply this to the kernel matrix $\mathbf{K}$. This results in the following.

$$\mathbf{K} = Q^* D^* Q^{*\prime} \quad \text{with} \quad D^* = \text{diag}(d_1^*, d_2^*, ..., d_H^*). \tag{28}$$

Here, we obtain different eigenvectors and eigenvalues compared to normal PCA. In the equation above, we have the $H \times H$ kernel matrix $\mathbf{K}$, which is decomposed into an $H \times H$ matrix Q^* and a diagonal matrix D^*. The matrix Q^* consists of the eigenvectors and D^* consists of the H eigenvalues in decreasing order. This now gives us PC factors, which we can use to express the SDF in terms of these new PC factors $P_{t,T}^*$.

$$P_{t,T}^* = Q^{*\prime} F_{t,T}, \tag{29}$$

$$M_{t,T} = 1 - b_p'(P_{t,T}^* - E(P_{t,T}^*)), \quad \text{with} \quad b_p = D^{-1}E(P_t^*). \tag{30}$$

Furthermore, the unconditional fundamental pricing equation, Equation (7), remains the same. As mentioned before, we can now see how the SDF has not changed much, only in terms of how the PCs are constructed. Only the characteristics have been used for the kernel function and this function, in turn, created the kernel matrix, which was used for creating the PCs. The SDF in terms of the individual stock returns changes slightly, where there is some non-linearity. This is shown in the following equivalent equations, for which we have used equations (4), (29) and (30).

$$\begin{aligned} M_{t,T} = 1 - b_p'(P_{t,T}^* - E(P_{t,T}^*)) \quad &\longleftrightarrow \quad M_{t,T} = 1 - b_p'(Q^{*\prime} F_{t,T} - E(Q^{*\prime} F_{t,T})) \\ &\longleftrightarrow \quad M_{t,T} = 1 - b_p'(Q^{*\prime} Z_t' R_{t,T} - E(Q^{*\prime} Z_t' R_{t,T})). \end{aligned} \tag{31}$$

At first glance, the last part of Equation 31 seems linear in terms of the returns $R_{t,T}$, however

this is not completely correct. The matrix Q^* consists of the eigenvectors of the kernel matrix $\mathbf{K}$. Multiplying $R_{t,T}$ with $Q^{*\prime}$ is a linear operation in terms of the returns, however the kernel matrix $\mathbf{K}$, from which we obtain Q^*, consists of nonlinear dependencies between the factors returns $F_{t,T}$. These factor returns are linear combinations of the returns $R_{t,T}$, as can be seen in Equation (4). Therefore, in the construction of the SDF for which we have used KPCA, there is one nonlinear part in terms of the factor returns. This is because we apply the kernel to the factor returns, creating nonlinear dependencies between the factor returns. To summarize, the SDF itself is a linear operation on the returns $R_{t,T}$, however in the steps taken to obtain the eigenvectors, which are in Q^*, nonlinear operations are applied to the factor returns $F_{t,T}$. In terms of the implications this has, it should not change much for the interpretability of the SDF. In the end, the operation performed on the returns $R_{t,T}$ in the SDF is linear, meaning that a linear combination of different stock returns can easily be constructed to obtain a well performing portfolio. The nonlinear part is only necessary in determining the values for Q^*, which is the part of the linear operation in the new SDF.

Kozak (2019) provide us with the algorithm containing the previous theoretical information. The steps we will take are slightly different. First, we use kernel functions to move from the current data space to the kernel space. After this, change the current kernel space to be zero mean. Now apply PCA to the new zero mean kernel space. This gives us PCs and we further follow the previous steps in terms of SDF explained in Section 2.1.1.

2.2.2 Estimation

The estimation for the extension does not differ much from the estimation for the replication. The same methods are used as those explained in Section 2.1.2. The only difference is in the theoretical framework, which has influence on the implementation of those methods, not necessarily on the estimation. In terms of implementation, we do exactly what we explained in Sections 2.2.1 and 2.2.1.1. This means, first applying different kernels to the factor returns, then centering the resulting kernel matrices, and finally applying PCA. Furthermore, we compare the Sharpe ratios in the same way as explained in Section 2.1.2. Finally, with regards to the code we use, we implement the methods of the extension, however the framework of the rest of the code is still the MATLAB code provided by Kozak et al. (2020).

3 Data

For dimensionality reduction to be fruitful, the data for this research must be of large dimensionality. There should be a high amount of characteristics on which the factors will be based. Kozak et al. (2020) use multiple data sets to perform their research. First, they use 25 ME/BM-sorted portfolios which originate from Fama and French (1993), then they use their own set of 50 characteristics which are tied to anomalies and finally they use 68 financial ratios from Wharton Research Data Services (WRDS) in combination with 12 portfolios. The first two of these data sets are made available by Kozak et al. (2020) and are also used in this research. For the 25 Fama and French (1993) portfolios, the daily returns are available ranging from July 1926 to December 2017. Furthermore, Kozak et al. (2020) orthogonalize this data. For this, they use the Center for Research in Security Prices (CRSP) value-weighted index return in combination with the β values obtained from the full sample. They do this with the following equation.

$$F_{t,T} = \tilde{F}_{t,T} - \beta R_{m,t,T}. \tag{32}$$

Here, $F_{t,T}$ is the $H \times 1$ vector of abnormal factor returns, $\tilde{F}_{t,T}$ is the $H \times 1$ vector of the raw portfolio returns and $R_{m,t,T}$ is the CRSP value-weighted index return at that specific point time interval between times t and T, which is just a 1×1 value. Furthermore, β is the $K \times 1$, which is equivalent to $H \times 1$ in our case, vector consisting of the β values obtained from the full sample. After this orthogonalization, Kozak et al. (2020) proceed to modify the portfolio returns in terms of the standard deviations. This rescaling makes the new standard deviations the same as "...the in-sample standard deviation of the excess return on the aggregate market index." (Kozak et al., 2020, p.280). Besides the 25 Fama and French (1993) portfolios, we also have the set of 50 anomaly portfolios, for which daily data is available from November 1973 to December 2017. Kozak et al. (2020) take a few steps with regards to processing the raw data. We will take the same steps, which are mainly a rank transformation and a normalization. These operations ensure a correct data set for the purpose of the research of Kozak et al. (2020) and thus also this paper. Namely, the focus on the cross-section with regards to the predictions of returns. Furthermore, more obvious reasons are the removal of the effects of outliers and the conservation of the same leverage with regards to all the portfolios. Finally, the factor returns are again orthogonalized with respect to the CRSP value-weighted index return in the same way as previously explained. We do not deviate from Kozak et al. (2020) with regards to these transformations. Finally, the definitions of the variables

are explained very clearly by Kozak et al. (2020) and can be found in their Internet Appendix, together with the annualized mean returns.

4 Results

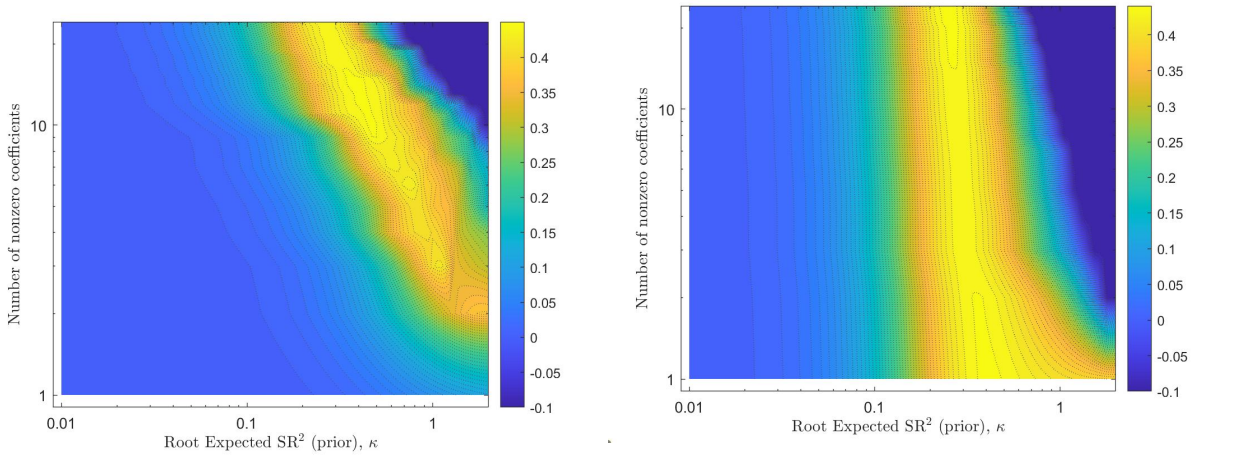
4.1 Replication

First, we replicated the results found by Kozak et al. (2020). For this, we considered two data sets, namely the daily returns of the 25 ME/BM-sorted portfolios originating from Fama and French (1993) and their own set of 50 anomaly portfolios. These results can be found in the following two sections. As this part is a replication, the figures provided are almost exactly the same as those provided by Kozak et al. (2020). Furthermore, the comparison of the Sharpe ratios can be found in Section 4.2, as we evaluate the results of the replication and the extension simultaneously for convenience.

4.1.1 25 Fama and French (1993) portfolios

For the following figures, Figure 1 and Figure 3, either the raw characteristics are used or PCA is applied to both data sets. These figures show a mapping of the OOS R^2 values corresponding to different levels of shrinkage and sparsity. This means that the dual-penalty method has been used here, and the levels of shrinkage and sparsity actually correspond to certain values of the penalty parameters γ_1 and γ_2. These levels of shrinkage and sparsity can be seen on the horizontal axis and the vertical axis respectively, which have a logarithmic scale. For the horizontal axis, the far left corresponds to a very high amount of shrinkage, while the far right corresponds to no shrinkage. The vertical axis is similar, where values low on the axis correspond to a high amount of sparsity, while values high on the axis correspond to a low amount of sparsity. This is expressed in the number of nonzero coefficients.

In Figure 1a, we see the mapping obtained from the raw 25 Fama and French (1993) portfolios. Here we see that the section with the highest OOS R^2 values is diagonal, indicating sparsity and shrinkage substitute each other. If there is no shrinkage, we can see that a high amount of sparsity reaches high OOS R^2 values. If there is no sparsity, we can see there must be a substantial amount of shrinkage to reach high OOS R^2 values. Furthermore, when no shrinkage is imposed, including 2 or 3 portfolios is a combination that reaches one of the highest OOS R^2 values. This is as expected, as Kozak et al. (2020) already mention that they know the following from Lewellen et al. (2010).



(a) *25 Fama and French (1993) portfolios*

(b) *Principal Components of the 25 Fama and French (1993) portfolios*

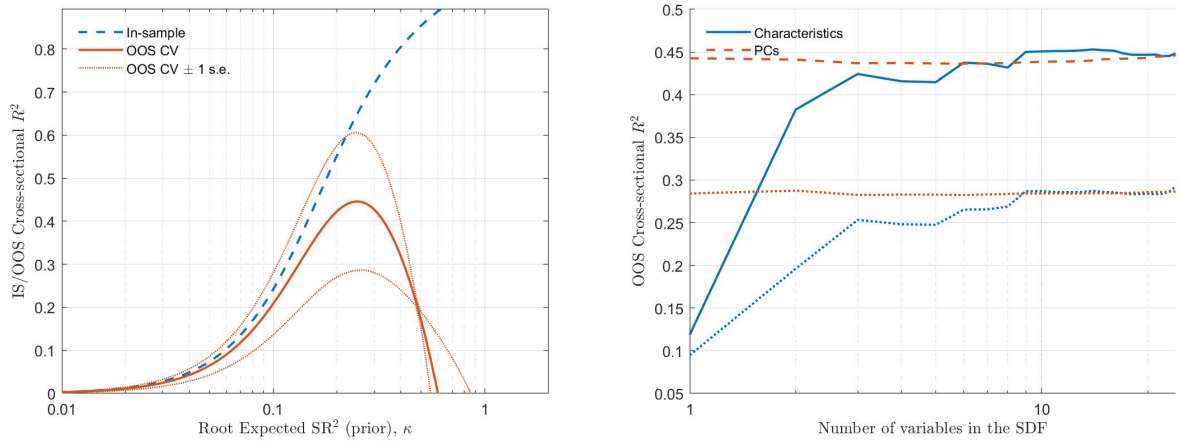
Figure 1

Mappings of the OOS R^2 from the dual-penalty method for the Fama and French (1993) 25 ME/MB-sorted portfolios (full sample from July 1926 to December 2017), where either no PCA is applied, or PCA is applied. The x-axis specifies the amount of shrinkage through the root expected SR^2, or κ. The y-axis specifies the sparsity, thus the number of nonzero coefficients. For each combination of shrinkage and sparsity, the value of the OOS R^2 is given in colour, for which the scale is on the right side of the figure.

The structure of the 25 Fama and French (1993) portfolio returns is such that a linear combination of only a few portfolios, which differ in SMB and HML factor loadings, could span the SDF. This is what we see in Figure 1a. Furthermore, the worst OOS R^2 values are obtained when there is no sparsity or shrinkage imposed (top right).

For the case in which PCA is used, the expectation based on Kozak et al. (2020) and Kozak et al. (2018), is that even more sparsity can be found. Kozak et al. (2018) found that the first two PCs are almost the same as the SMB and HML factors. Figure 1b shows us the results for the case in which PCA is applied. We can clearly see that including only 1 PC, already results in a OOS R^2 value close to the maximum, if the right amount of shrinkage is applied. Furthermore, including 2 PCs in combination with the right amount of shrinkage is able to obtain the highest OOS R^2. Furthermore, the area with the highest OOS R^2 values is nearly vertical, meaning that adding more PCs does not necessarily decrease the OOS R^2 and the amount of shrinkage stays rather constant along this area of high OOS R^2 values. As Kozak et al. (2020) mention, this is because the L^2 norm penalty, which indicates the shrinkage, shrinks the PCs that already have a low variance. For PCA, the idea is that only a few PCs account for most of the variance, so all the other PCs have a low variance, and if these are shrunk, their value is even closer to nothing. Therefore, adding or

removing these PCs does not make a noticeable difference.



(a) *Fixed on no sparsity* **(b)** *Variable sparsity, optimal shrinkage*

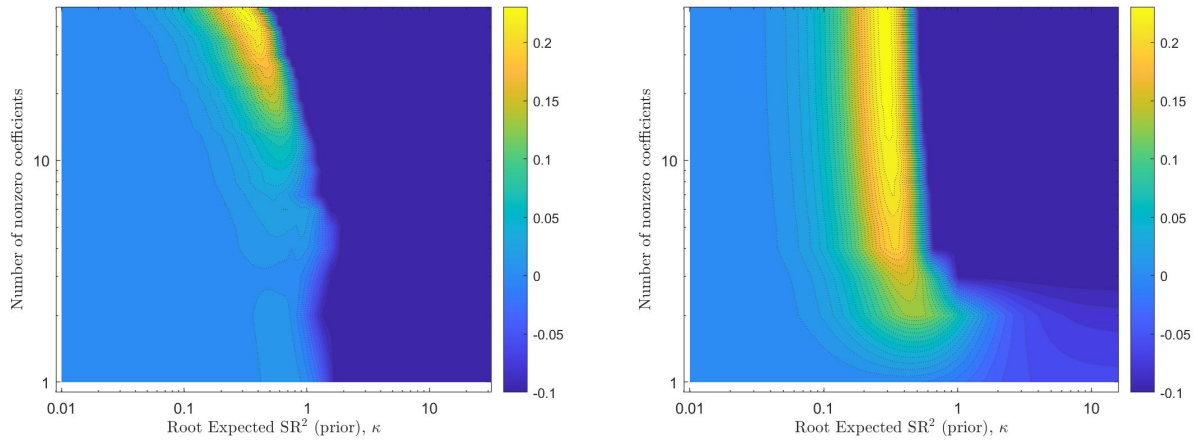
Figure 2

Optimal sparsity (L^1) and shrinkage (L^2) for the 25 Fama and French (1993) portfolios (full sample from July 1926 to December 2017). In the left figure, the values of the in-sample cross-sectional R^2 (blue dashed line) and the OOS cross-sectional R^2 (obtained from cross-validation) on the y-axis (red solid line) are plotted corresponding to each value of shrinkage (L^2) on the x-axis. In this case, no sparsity is imposed. Furthermore, two lines representing one standard error above and below the OOS R^2 from the cross-validation are provided (red dotted lines). In the right figure, the OOS cross-sectional R^2 on the y-axis is plotted again, however now against the sparsity on the x-axis. In this case, the shrinkage (L^2) differs for each point, as for each value of sparsity, the optimal value of shrinkage is chosen to obtain the highest OOS R^2. These lines are given for the both the raw characteristics (blue solid line) and the PCs (red dashed line). Furthermore, two lines are added representing one standard error below the OOS R^2 (red and blue dotted lines).

In Figure 2a, we focus specifically on the case of no sparsity, thus the top edge of Figure 1a. Here, we see the optimal OOS R^2 is found for a shrinkage indicator κ value of approximately 0.25, slightly higher than the value found by Kozak et al. (2020) (higher κ value means less shrinkage). The most important part, as Kozak et al. (2020) mention, is that this figure illustrates how the in-sample cross-sectional R^2 differs from the OOS cross-sectional R^2. With no shrinkage, the in-sample values would indicate a wrong explanatory power of the SDF for the expected returns out of sample. In Figure 2b, we focus on the optimal OOS R^2 values per variable included in the SDF. This means that for each variable included, the highest OOS R^2 is chosen across all possible shrinkage values (optimal L^2). These are thus the values in Figure 1a and Figure 1b along the optimal (most yellow) area. Now we can clearly see that including only 1 PC results in a very high OOS R^2. Furthermore, including 2 portfolios for the raw characteristics case, already obtains a high OOS R^2, but including 3 portfolios almost reaches the highest OOS R^2. Clearly, the method used by Kozak et al. (2020)

performs well in imposing sparsity where it is possible and applying the right amount of shrinkage. Furthermore, these findings are in line with Fama and French (1993).

4.1.2 50 anomaly portfolios



(a) *50 anomaly portfolios* **(b)** *Principal Components of the 50 anomaly portfolios*

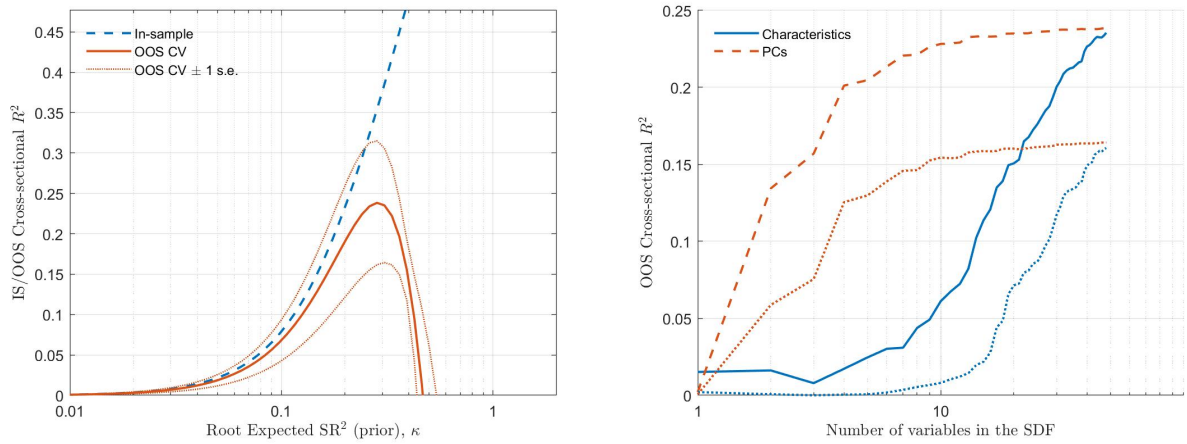
Figure 3

Mappings of the OOS R^2 from the dual-penalty method for the 50 anomaly portfolios of Kozak et al. (2020) (full sample from November 1973 to December 2017), where either no PCA is applied, or PCA is applied. The x-axis specifies the amount of shrinkage through the root expected SR^2, or κ. The y-axis specifies the sparsity, thus the number of nonzero coefficients. For each combination of shrinkage and sparsity, the value of the OOS R^2 is given in colour, for which the scale is on the right side of the figure.

In Figure 3a, we see the mapping obtained from the 50 anomaly portfolios. This figure is a lot different than Figure 1a obtained for the 25 Fama and French (1993) portfolios. The only similarity is that if no sparsity or shrinkage is imposed, the worst OOS R^2 values are obtained. Figure 3a shows us that for the 50 anomaly portfolios, the area with the highest OOS R^2 values is slightly diagonal, however, it is very small. This means that shrinkage and sparsity are not as interchangeable as for the 25 Fama and French (1993) portfolios and that almost no sparsity should be imposed for high OOS R^2 values. For these high values, there should be a substantial amount of shrinkage. If we increase the sparsity, the OOS R^2 values will decline rapidly. This tells us something about the structure of the 50 anomaly portfolios. The structure is the opposite of the structure of the 25 Fama and French (1993) portfolios. For these 50 anomaly portfolios, we need almost all the portfolios to capture the SDF, indicating each portfolio substantially contributes to the OOS R^2 of the SDF. Imposing sparsity is therefore not a good idea, so for this data set, a characteristics-sparse SDF

would not work in terms of pricing performance.

—



(a) *Fixed on no sparsity* **(b)** *Variable sparsity, optimal shrinkage*

Figure 4

Optimal sparsity (L^1) and shrinkage (L^2) for the 50 anomaly portfolios (full sample from November 1973 to December 2017). In the left figure, the values of the in-sample cross-sectional R^2 (blue dashed line) and the OOS cross-sectional R^2 (obtained from cross-validation) on the y-axis (red solid line) are plotted corresponding to each value of shrinkage (L^2) on the x-axis. In this case, no sparsity is imposed. Furthermore, two lines representing one standard error above and below the OOS R^2 from the cross-validation are provided (red dotted lines). In the right figure, the OOS cross-sectional R^2 on the y-axis is plotted again, however now against the sparsity on the x-axis. In this case, the shrinkage (L^2) differs for each point, as for each value of sparsity, the optimal value of shrinkage is chosen to obtain the highest OOS R^2. These lines are given for the both the raw characteristics (blue solid line) and the PCs (red dashed line). Furthermore, two lines are added representing one standard error below the OOS R^2 (red and blue dotted lines).

Similar as for the 25 Fama and French (1993) portfolios, we now focus specifically on the case of no sparsity and on the optimal OOS R^2 values per value of sparsity. The two figures in Figure 4 are the same as the figures previously in Figure 2 with respect to the construction of the figures, the only difference is the data set used. In Figure 4a we see a similar difference between in-sample and OOS as we found in Figure 2a. Again, the in-sample cross-sectional R^2 is misleading in terms of explanatory power OOS of the SDF for the expected returns. We do see that the highest OOS R^2 is obtained for a shrinkage indicator κ value of approximately 0.28, slightly lower than the value found by Kozak et al. (2020). In the right figure, Figure 4b, we can clearly see the poor performance of the characteristics based SDF for different levels of sparsity. Including 10 characteristics-based factors does not even reach half the maximum OOS R^2. For the PCs, including 2 already gives a decent OOS R^2 and including 4 PCs almost reaches the maximum OOS R^2.

Table 1

The 11 largest (most contributing) SDF factors for the 50 anomaly portfolios, both raw characteristics based factors and the Principal Components. Given are the coefficient estimates b and the absolute values of their t-statistics. These values are obtained for the optimal value of shrinkage (prior root expected SR^2). A distinction is made between the values obtained for the characteristics-based SDF (the raw 50 anomaly portfolios) and the values obtained for the PC based method. Furthermore, as Kozak et al. (2020) do, the values are sorted on the t-statistics in a descending order.

50 anomaly portfolios			Principal Components of the 50 anomaly portfolios		
	b	t-stat		b	t-stat
Industry relative reversals (low vol.)	-0.879	3.527	PC 4	1.014	4.249
Industry momentum-reversals	0.483	1.945	PC 1	-0.537	3.081
Industry relative 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

Table 1 provides the SDF factors that contribute the most to the SDF for the 50 anomaly portfolios. In Table 1 we can see a few points of interest. First, the most important coefficients for the characteristics-based SDF do not differ much in value. Also, only the t-statistic corresponding to the Industry relative reversals (low vol.) is quite large, the rest is much lower and not very different from each other. This shows what we already found in Figures 3 and 4, that only a few of these portfolios would not be enough to obtain a high OOS R^2 for the SDF. As Kozak et al. (2020) mention, the joint significance almost all of the 50 anomaly portfolios is what has a good explanatory power for the SDF. Furthermore, as mentioned previously, we can see that 4 PCs have significantly different values from 0 based on the t-statistics at a significance level of 5%. These are PC4, PC1, PC2 and PC9, which as mentioned before, can obtain a very high OOS R^2 if only these PCs are included in the SDF. That these PCs have high coefficients and contribute the most to the explanatory power of the SDF comes from what was mentioned before in Section 2.1.1. More shrinkage is imposed on low eigenvalue PCs, which do not contribute much to the volatility of the SDF, which means that the remaining high coefficient PCs, are those PCs that contribute much to the volatility of the SDF. Besides, since PCs are linear combinations of the factor returns, it is not unexpected that only a few PCs have a large explanatory power, as the different factor returns corresponding to the 50 anomaly portfolios have a large joint significance. A linear combination

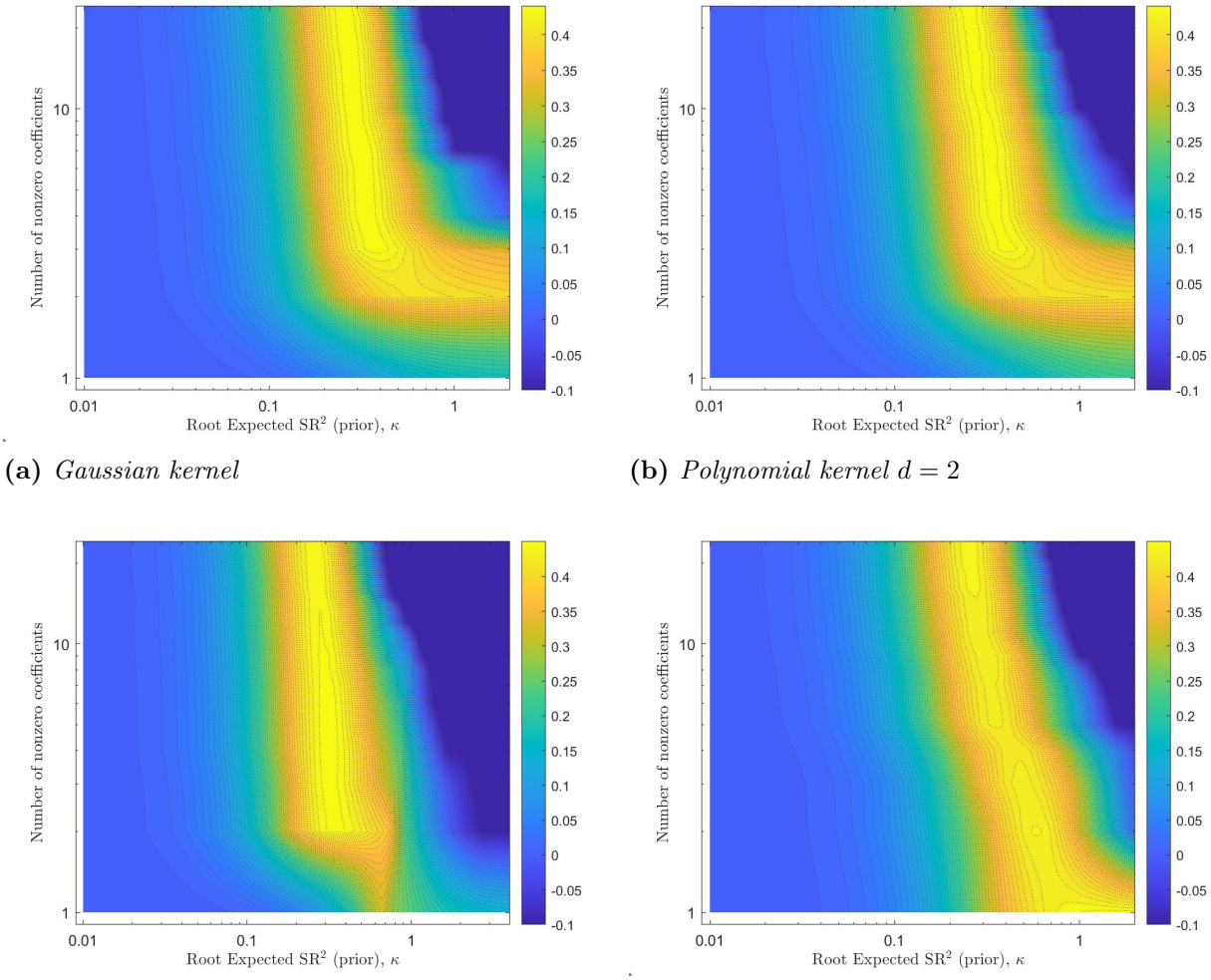
would then be much better in terms of explanatory power, which are thus the PCs. Lastly, there are some small differences with the results found by Kozak et al. (2020) regarding the sign corresponding to the PC based SDF coefficients. For PC9, PC15, PC17, PC11, PC13 and PC7, we have a different sign. This does not have any implications for the optimal OOS R^2 that was obtained, as we know that we have the same findings as Kozak et al. (2020).

4.2 Extension

For this part, we use the exact same data, however, the method has slightly changed. Instead of PCA, Kernel PCA is used to obtain the results. For the following figures, Figure 5 and Figure 6, the Gaussian kernel, the Polynomial kernel with $d = 2$ and the Linear kernel are used for both data sets, the Fama and French (1993) 25 ME/MB-sorted portfolios and the 50 anomaly portfolios of Kozak et al. (2020). These figures show a mapping of the OOS R^2 values corresponding to different levels of shrinkage and sparsity. Again, the dual-penalty method has been used here, and the levels of shrinkage and sparsity actually correspond to certain values of the penalty parameters γ_1 and γ_2.

4.2.1 25 Fama and French (1993) portfolios

In Figure 5 we can see the OOS R^2 values for the Fama and French (1993) data set. First, in Figure 5a, the Gaussian kernel has been used to obtain these values. Here, the highest OOS R^2 can be found when 3 or more PCs are included and when there is some shrinkage. For higher of lower amounts of shrinkage, the OOS R^2 rapidly decreases. We do see however, that if only 2 PCs are included, a decrease in the amount of shrinkage does not nearly as much decrease the OOS R^2 as when 3 PCs are included. Furthermore, the yellow area is rather vertical, indicating a constant degree of shrinkage, regardless of the number of PCs included, except for 1 or 2 PCs included. Similar results are found for the Polynomial kernel with $d = 2$, which can be seen in Figure 5b. The only difference is that the most yellow area is slightly shifted to the right and more narrow. Furthermore, the Polynomial kernel with $d = 5$ in Figure 5c yields improved results in terms of sparsity, where including 2 PCs can already obtain the highest OOS R^2. In this case, less shrinkage variation is possible to obtain this high OOS R^2. If we compare these two figures with Figure 1b, the results are not as expected. The use of KPCA increased the number of PCs needed to obtain a similar OOS R^2. The amount of shrinkage however, is rather similar in both cases. Furthermore, we included the Linear kernel, Figure 5d, which should in theory obtain the same results as regular PCA. Note that $c = 1$, as the mapping obtained for $c = 0$ seemed to perform worse in terms of



(a) *Gaussian kernel*

(b) *Polynomial kernel* $d = 2$

(c) *Polynomial kernel* $d = 5$

(d) *Linear kernel* $c = 1$

Figure 5

Mappings of the OOS R^2 *from the dual-penalty method for the Fama and French (1993) 25 ME/MB-sorted portfolios (full sample from July 1926 to December 2017), where the Gaussian kernel, the Polynomial kernel (*$d = 2$ *and* $d = 5$*) and the Linear kernel (*$c = 1$*) are used. The x-axis specifies the amount of shrinkage through the root expected* SR^2*, or* κ*. The y-axis specifies the sparsity, thus the number of nonzero coefficients. For each combination of shrinkage and sparsity, the value of the OOS* R^2 *is given in colour, for which the scale is on the right side of the figure.*

OOS R^2 when shrinkage and sparsity are imposed. The mapping for $c = 0$ can be found in Figure 7 in the Appendix. There are large similarities between Figure 5d and Figure 1b. In both cases, including only 1 PC obtains one of the highest OOS R^2 values and increasing the number of PCs included, does not change the OOS R^2 much. There are however some small differences. In Figure 1b, the area containing the highest OOS R^2 values is nearly exactly vertical. In Figure 5d this area is also quite vertical, but not nearly as vertical. This means that as the number of included PCs increases, the amount of shrinkage must be increased to obtain similar OOS R^2 values. This

difference could be due to the regularization of the covariance for the regular PCA. Regular PCA makes use of the covariance matrix to obtain the eigenvectors and eigenvalues. KPCA uses the kernel matrix to obtain the eigenvectors and eigenvalues. We have followed the method of Kozak et al. (2020) for the estimation of regular PCA, in which they do not use the covariance matrix, but a regularized covariance matrix. This could be a reason for the differences between Figure 1b and Figure 5d.

We expected that KPCA would improve the results in terms of sparsity and shrinkage. These results do not necessarily confirm these expectations, as we find slightly worse results with respect to normal PCA. An important reason could be that KPCA works very well in cases of high dimensionality. As mentioned in Section 2.2.1, KPCA is able to overcome the curse of dimensionality. In this case, the dimensionality might not be large enough, meaning KPCA would not necessarily perform better in terms of obtaining a high OOS R^2 when shrinkage and sparsity are imposed. Because KPCA is meant for high dimensionality, this could even mean that the results in a low dimensionality case could be worse, as we are accounting for high dimensionality, which might not be the case.

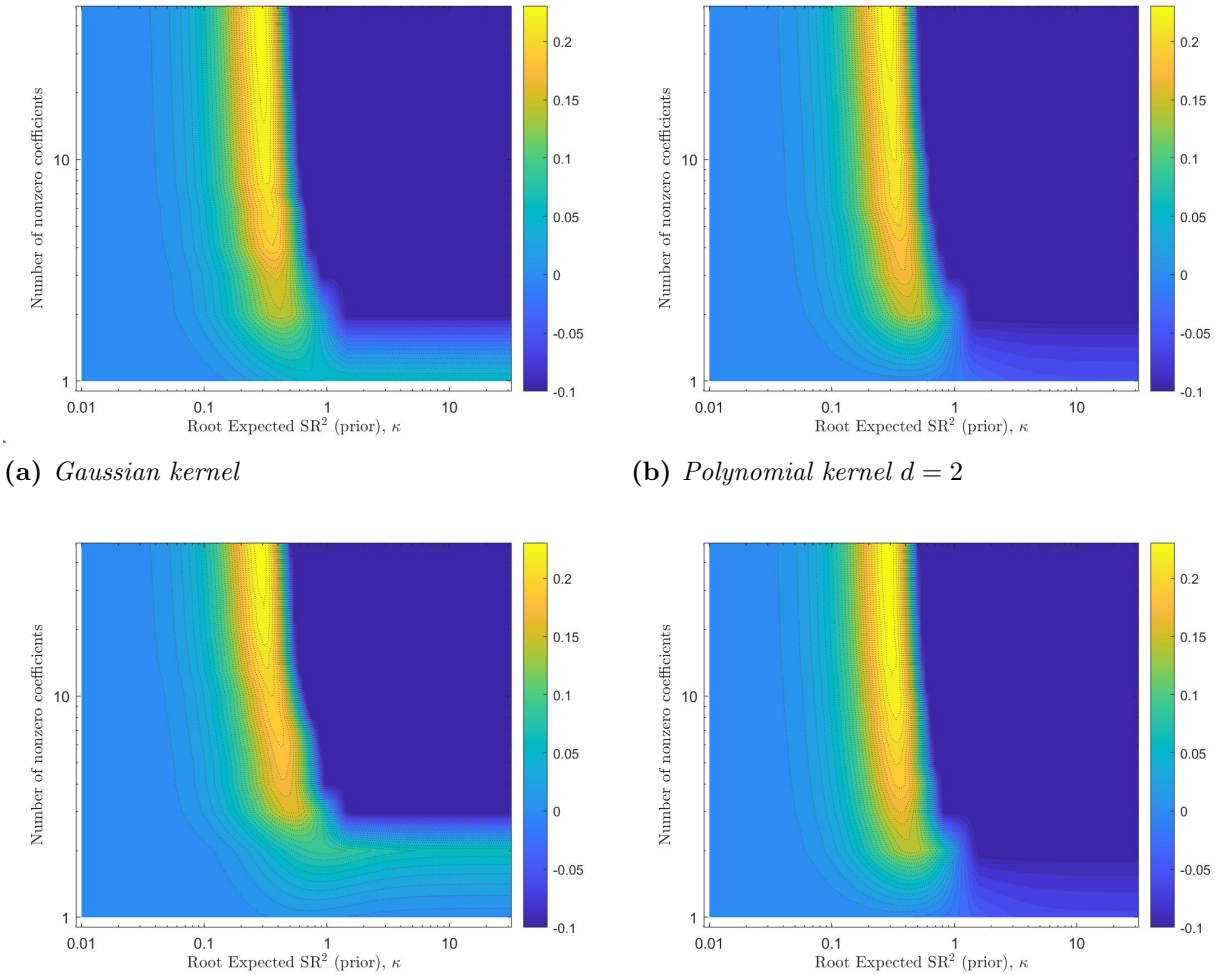
Additionally, we know from Lettau and Pelger (2020) that for regular PCA, including the mean of the factor returns could improve the results with regards to explanatory power, specifically for those factors that do not contribute much to the variance. For the regular PCA in this research, we make use of the covariance matrix, which demeans the factor returns. For the KPCA, we demean the kernel matrix to create a zero mean kernel function space, as mentioned before in Section 2.2.1.1. This is however slightly different, as we demean the information in the kernel matrix, which consists of the nonlinear kernel functions with the factor returns as input. The results obtained when we omit the demeaning operation can be found in the Appendix in Figure 5, and provide slightly better results regarding the values of the OOS R^2 for higher levels of sparsity. Especially the Polynomial kernel with $d = 5$ performs very good. However, for consistency and comparison with the results of Kozak et al. (2020), our focus is on the results obtained from a demeaned kernel matrix.

Furthermore, we compared the Sharpe ratios obtained from the different methods used, namely with and without PCA and with the different kernels. Here, we do not impose sparsity and determine the optimal amount of shrinkage, for which we then compute the Sharpe ratio obtained from the resulting MVE portfolio, as explained in Section 2.1.2. We estimate the L^2 norm penalty based on the entire sample, as explained in Section 2.1.2, but also on only a part of the entire sample. Kozak et al. (2020) also do this for their asset pricing tests, however we will only do this for the comparison

of the Sharpe ratios. The withheld sample is the data from the January 2005 to December 2017. Kozak et al. (2020) withhold a sample, because this would create a "...pure OOS test." (Kozak et al., 2020, p.289). The optimal amount of shrinkage in the sample for which it is estimated, is not necessarily the optimal value for a different sample Kozak et al. (2020) point out. The results can be found in Table 3 in the Appendix, since the results are almost identical for the different methods. We do see that the values for the Sharpe ratio, the OOS R^2 and the κ increase (shrinkage decreases) when we use a withheld sample. To test whether the Sharpe ratio estimation was not independent of the method used, we increased the value of d to 20 for the Polynomial kernel, which resulted in slightly lower values. However, these differences are not noteworthy. An explanation for the almost identical values for each of the different methods could be that because there is no sparsity imposed, the different methods do not necessarily distinguish themselves from each other. PCA and KPCA perform well when sparsity is imposed, but in this case, no sparsity is imposed.

4.2.2 50 anomaly portfolios

In Figure 6 we can see the OOS R^2 values for the 50 anomaly portfolios data set of Kozak et al. (2020). Here, we can clearly see that for each kernel used, the results are very similar. For each kernel, the Gaussian kernel in Figure 6a, the Polynomial kernel with $d = 2$ and $d = 5$ in Figures 6b and 6c, and the Linear kernel with $c = 0$ in Figure 6d, the amount of shrinkage needed for high values of OOS R^2 is very similar. There are however some differences, for instance the number of PCs that must be included to obtain the highest OOS R^2 value. For the Gaussian kernel and the Linear kernel in Figures 6a and 6d, 11 PCs with the right amount of shrinkage can obtain the highest OOS R^2. For the Polynomial kernels in Figures 6b and 6c, 12 PCs are needed. Furthermore, for the Gaussian kernel and the Linear kernel, 4 PCs can obtain a rather high OOS R^2 with respect to the highest OOS R^2 that can be obtained. For the Polynomial kernel with $d = 2$, a similar value can be obtained if 5 PCs are included, and when $d = 5$, this even increases to 6. Comparing these results with the results found in Figure 3b, the differences are very small. The use of KPCA does not necessarily improve the results compared to regular PCA. If we compare the results of the Linear kernel in Figure 6d with the results obtained with regular PCA in Figure 3b, which in theory should be the same, the differences are very small. These differences mainly manifest themselves for the lower OOS R^2 values and the corresponding amount of sparsity and shrinkage. As mentioned before, this could be because instead of using the covariance matrix for PCA, a regularized covariance matrix is used.



(a) *Gaussian kernel*

(b) *Polynomial kernel* $d = 2$

(c) *Polynomial kernel* $d = 5$

(d) *Linear kernel* $c = 0$

Figure 6

Mappings of the OOS R^2 *from the dual-penalty method for the 50 anomaly portfolios of Kozak et al. (2020) (full sample from November 1973 to December 2017), where the Gaussian kernel, the Polynomial kernel (*$d = 2$ *and* $d = 5$*) and the Linear kernel (*$c = 0$*) are used. The x-axis specifies the amount of shrinkage through the root expected* SR^2*, or* κ*. The y-axis specifies the sparsity, thus the number of nonzero coefficients. For each combination of shrinkage and sparsity, the value of the OOS* R^2 *is given in colour, for which the scale is on the right side of the figure.*

Besides these figures, Table 2 provides the SDF factors that contribute the most to the SDF for the 50 anomaly portfolios. The values found in Table 2 are similar to those found for the PCs in Table 1 in Section 4.1.2, especially for the Linear kernel, which was as expected. These values are almost identical, the small differences could be because of, as already mentioned, the regularization of the covariance matrix for normal PCA. For each kernel, we find that the b coefficients of the 4 PCs with the highest t-statistic are statistically different from 0 based on the t-statistics at a significance level of 5%, as we also found for the case of normal PCA. The 4 PCs for which this holds do differ

Table 2

The 11 largest (most contributing) SDF factors for the 50 anomaly portfolios, for each of the different kernels used (Gaussian kernel, Polynomial kernel with $d = 2$ and with $d = 5$, Linear kernel with $c = 0$). Given are the coefficient estimates b and the absolute values of their t-statistics. These values are obtained for the optimal value of shrinkage (prior root expected SR^2). Furthermore, the values are sorted on the t-statistics in a descending order.

Gaussian			Poly $d = 2$			Poly $d = 5$			Lin $c = 0$		
	b	t-stat		b	t-stat		b	t-stat		b	t-stat
PC 5	0.913	3.848	PC 5	0.972	4.099	PC 5	-0.811	3.408	PC 5	1.005	4.245
PC 2	0.589	3.372	PC 2	-0.546	3.132	PC 3	0.512	2.343	PC 1	-0.552	3.166
PC 12	-0.647	2.588	PC 3	0.485	2.227	PC 2	-0.430	2.313	PC 2	-0.505	2.344
PC 3	0.494	2.258	PC 11	-0.482	1.927	PC 18	-0.505	1.993	PC 10	-0.537	2.147
PC 20	0.333	1.310	PC 12	0.410	1.635	PC 15	-0.400	1.585	PC 7	0.353	1.438
PC 8	0.282	1.148	PC 17	-0.344	1.358	PC 11	-0.373	1.482	PC 11	0.317	1.266
PC 14	-0.259	1.030	PC 7	-0.225	0.915	PC 14	0.351	1.405	PC 17	0.298	1.176
PC 6	-0.235	0.962	PC 10	0.227	0.903	PC 4	-0.299	1.325	PC 4	0.227	1.037
PC 21	-0.179	0.702	PC 4	-0.189	0.856	PC 13	0.304	1.198	PC 15	0.202	0.800
PC 22	-0.168	0.661	PC 20	0.201	0.789	PC 10	-0.299	1.187	PC 23	0.157	0.613
PC 7	0.139	0.567	PC 18	-0.182	0.714	PC 16	0.257	1.024	PC 8	-0.141	0.572

for each different kernel used. Furthermore, we know that for each of the kernels used, including 4, 5 or 6 PCs can already obtain a very high OOS R^2. Similar as we reasoned in Section 4.1.2, this corresponds with the expectations of the application of PCA and the contribution of each PC to the volatility of the SDF. This also holds for KPCA, as the kernel does not change the interpretation of the PCs in the SDF, as these are still linear transformations of the original stock returns.

Again, our expectations regarding the performance of KPCA are not confirmed. Again, as mentioned in Section 4.2.1, the focus of KPCA on a high dimensionality situation could be a possible explanation. The KPCA results found for this data set, the 50 anomaly portfolios, are almost identical to the results found with normal PCA. This is better than what we found for the 25 Fama and French (1993) portfolios. This could be because there is less redundancy between the different anomaly portfolios than for the 25 Fama and French (1993) portfolios, or because the dimensionality is higher, which would be beneficial for KPCA. The redundancy between the 25 Fama and French (1993) portfolios could mean that applying a non-linear kernel unnecessarily complicates the relations between the different portfolios, while normal PCA does not do this.

Furthermore, we include the OOS R^2 mappings obtained from KPCA where we do not demean the kernel matrix in the Appendix in Figure 9. The results do not change as much as they did for the 25 Fama and French (1993) portfolios. This could again be because of the redundancy between the 25 Fama and French (1993) portfolios, which makes it beneficial to add the mean in the kernel.

However, as mentioned in Section 4.2.1, our focus is on the results obtained when we demean the kernel matrix.

Finally, we compared the Sharpe ratios obtained from the MVE portfolios in the same way as explained in Section 4.2.1. Here, we do find different values, as we make use of a different data set, however, the relative results between the different methods are the same as found in 4.2.1. The values are almost identical across all methods and they increase in the case of a withheld sample. We do see a slight change in the case of the entire sample for the PCs, however, this is not a noteworthy change (magnitude of 0.0001). Furthermore, we tested the Polynomial kernel with $d = 20$ again, resulting in slightly different values, in this case slightly higher values.

5 Conclusion and discussion

The main idea of this paper was to test whether the cross-sectional performance of an SDF could be improved with the use of different dimensionality reduction techniques, taking into account the dimensionality of the data. Kernel PCA has been applied in addition to the already used regular PCA by Kozak et al. (2020). Different kernels, specifically the Gaussian kernel, the Polynomial kernel and the Linear kernel, have been applied to two data sets provided by Kozak et al. (2020). We replicated the results provided by Kozak et al. (2020) and thereby come to the same conclusion regarding their findings. Imposing sparsity on a characteristics-based SDF does not work well, unless there is redundancy amongst the characteristics-based factors. This is the case for the 25 Fama and French (1993) portfolios. Imposing sparsity on the SDF for the 50 anomaly portfolios does not prove to perform well, as there is not much redundancy between those portfolios. When PCA is applied, an SDF sparse in PCs does perform well out of sample in terms of R^2. As Kozak et al. (2020) mention, the problem still makes use of all the portfolio factor returns to estimate the PCs, however, a PC-sparse SDF could help with regards to the interpretation of the SDF.

Regarding the application of KPCA, the results are not entirely as expected. This method either obtains worse results in terms of the OOS R^2, or in the best case similar results, depending on the data set it is applied to. We do find that for the portfolios that exhibit redundancy, namely the 25 Fama and French (1993) portfolios, KPCA provides us with worse results than for the 50 anomaly portfolios. For the 50 anomaly portfolios, the results are almost identical to those obtained with regular PCA. Depending on the redundancy between the different factor returns in the data set, KPCA thus performs rather good in terms of OOS R^2, however not better than regular PCA.

A possible reason could be that the dimensionality of the data set is not high enough for KPCA to perform as expected. As mentioned, KPCA is able to overcome the curse of dimensionality. Furthermore, we find that the Gaussian kernel and the Polynomial kernel exhibit very similar results. For the 25 Fama and French (1993) portfolios, the Polynomial kernel with $d = 5$ does distinguish itself from the Gaussian kernel, however these differences almost completely disappear for the 50 anomaly portfolios. Again, the dimensionality of the data could be a possible explanation. A higher dimensionality could show larger differences between these two kernels. KPCA with the Linear kernel was added to test the method, as in theory, this should produce the same results as for regular PCA. The results were very similar, however not identical. The regularization of the covariance matrix for regular PCA could be the reason for this. Finally, the MVE portfolios constructed for the optimal L^2 norm penalty with no sparsity give almost identical results in terms of Sharpe ratio for each method. This is reasonable, as regular PCA and KPCA improve the OOS R^2 compared to the raw characteristics-based SDF when we do impose sparsity.

For future research, there are multiple limitations which give possibilities for improvements. First, in this paper, interactions between different characteristics have not been included when estimating the SDF. This would increase the dimensionality, and possibly improve the results obtained with KPCA. Second, the kernel has been applied to the factor returns, but could be applied to the raw characteristics data, as Kozak (2019) do, to compare the differences. Furthermore, regarding the kernel matrix, more focus could be directed to including the mean and not demeaning the kernel matrix. Third, the Sharpe ratios could be compared in a characteristics- and PC-sparse SDF model instead of only an SDF with shrinkage applied. Finally, a method which has not been used here, but is interesting to examine in future research, is Maximum Variance Unfolding, as found in Van Der Maaten et al. (2009) and applied by Weinberger et al. (2004). This method does not simply choose a kernel for KPCA, but fits (learns) a kernel specifically designed to fit the data, which could provide improved results.

References

Amari, S.-i., & Wu, S. (1999). Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, *12*(6), 783–789.

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.

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

Harvey, C. R., & Zhou, G. (1990). Bayesian inference in asset pricing tests. *Journal of Financial Economics*, *26*(2), 221–254.

Ince, H., & Trafalis, T. B. (2007). Kernel principal component analysis and support vector machines for stock price prediction. *Iie Transactions*, *39*(6), 629–637.

Jagannathan, R., & Wang, Z. (2002). Empirical evaluation of asset-pricing models: A comparison of the sdf and beta methods. *The Journal of Finance*, *57*(5), 2337–2367.

Kan, R., & Zhou, G. (1999). A critique of the stochastic discount factor methodology. *The Journal of finance*, *54*(4), 1221–1248.

Kozak, S. (2019). Kernel trick for the cross-section. *Available at SSRN 3307895*.

Kozak, S., Nagel, S., & Santosh, S. (2018). Interpreting factor models. *The 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.

Lettau, M., & Pelger, M. (2020). Estimating latent asset-pricing factors. *Journal of Econometrics*, *218*(1), 1–31.

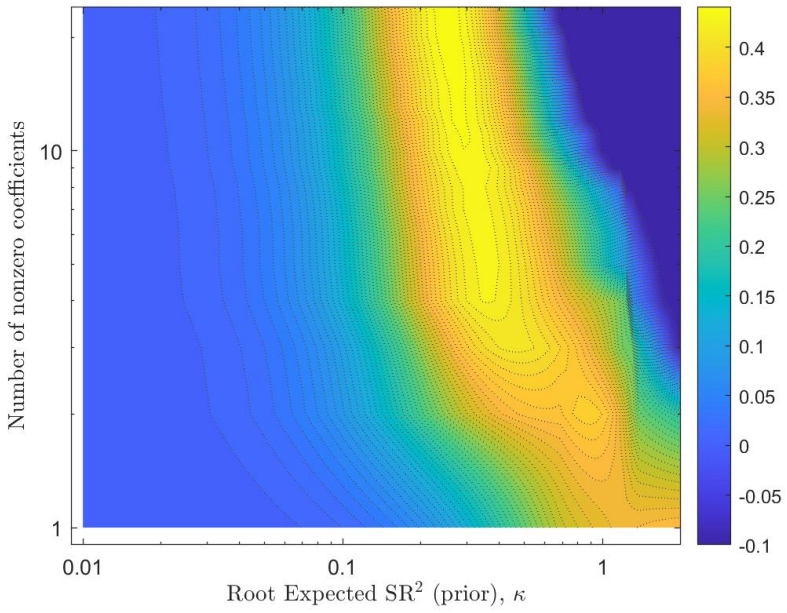
Lewellen, J., Nagel, S., & Shanken, J. (2010). A skeptical appraisal of asset pricing tests. *Journal of Financial economics*, *96*(2), 175–194.

Van Der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: A comparative. *J Mach Learn Res*, *10*(66-71), 13.

Weinberger, K. Q., Sha, F., & Saul, L. K. (2004). Learning a kernel matrix for nonlinear dimensionality reduction. *Proceedings of the twenty-first international conference on Machine learning*, 106.

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.

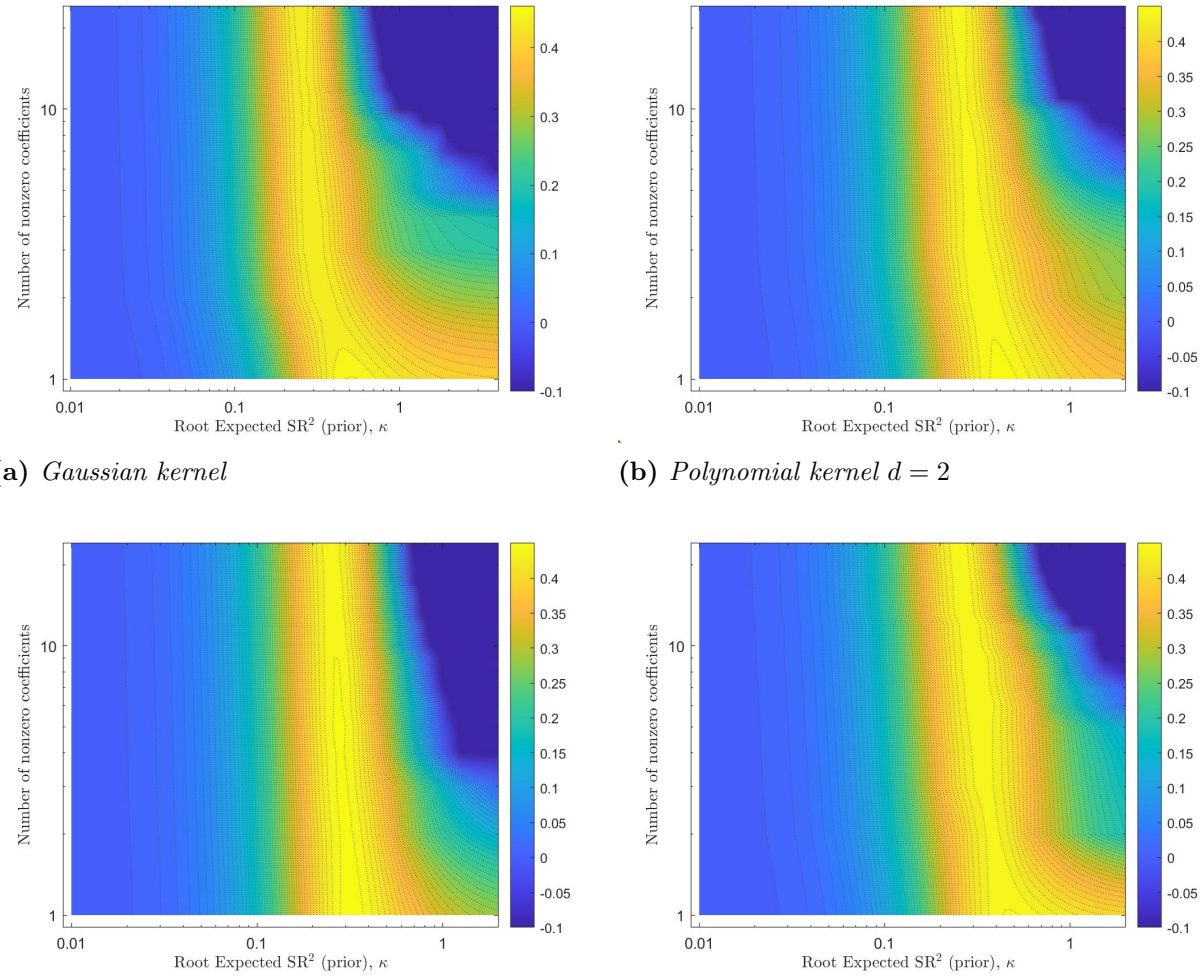
Appendix



.5

Figure 7

Mapping of the OOS R^2 from the dual-penalty method for the Fama and French (1993) 25 ME/MB-sorted portfolios (full sample from July 1926 to December 2017), where we use the Linear kernel with $c = 0$. The x-axis specifies the amount of shrinkage through the root expected SR^2, or κ. The y-axis specifies the sparsity, thus the number of nonzero coefficients. For each combination of shrinkage and sparsity, the value of the OOS R^2 is given in colour, for which the scale is on the right side of the figure.



(a) *Gaussian kernel*

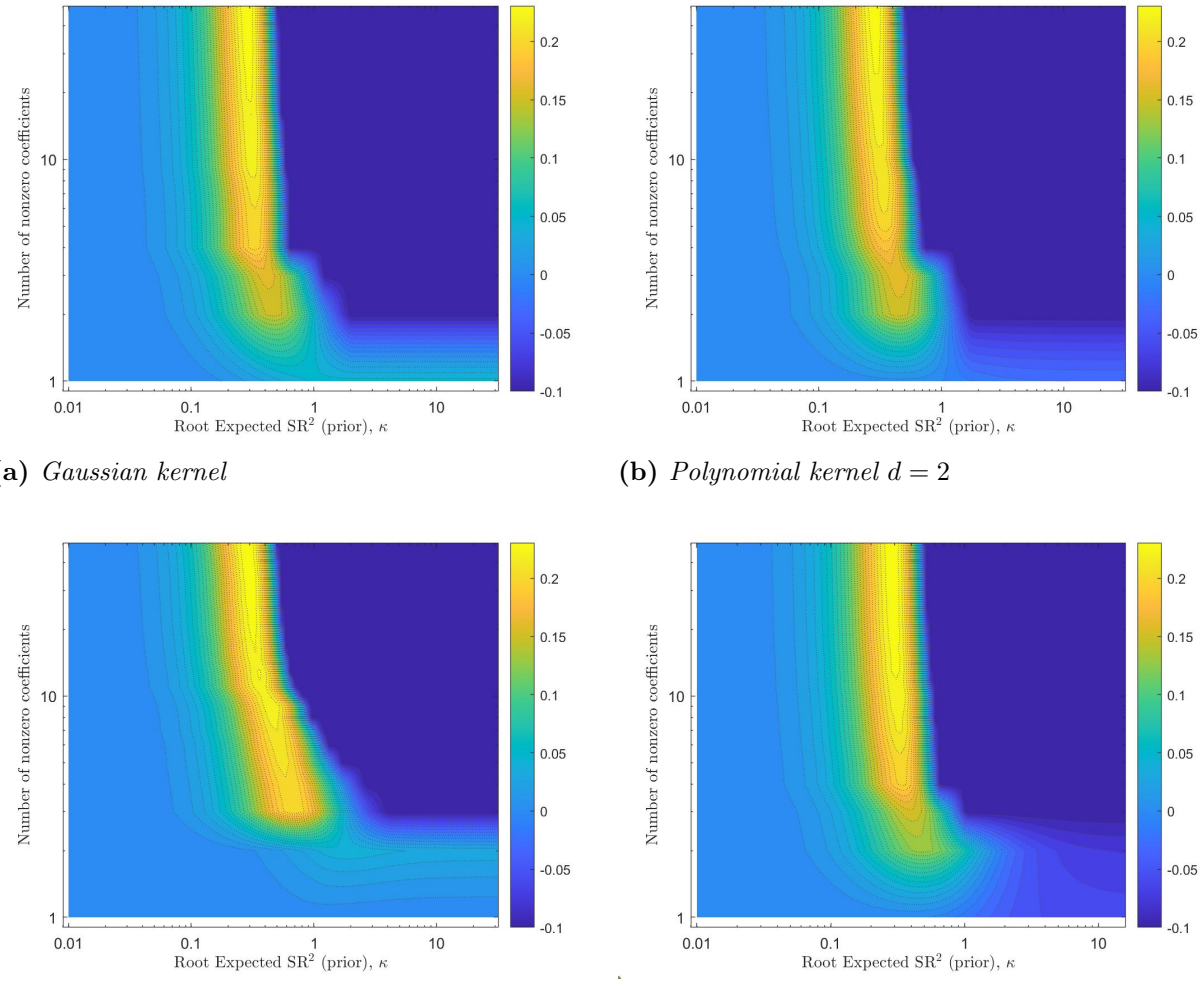
(b) *Polynomial kernel* $d = 2$

(c) *Polynomial kernel* $d = 5$

(d) *Linear kernel* $c = 1$

Figure 8

Mappings of the OOS R^2 *from the dual-penalty method for the Fama and French (1993) 25 ME/MB-sorted portfolios (full sample from July 1926 to December 2017), where the Gaussian kernel, the Polynomial kernel (*$d = 2$ *and* $d = 5$*) and the Linear kernel (*$c = 1$*) are used. For these mappings, we have not demeaned the kernel matrix in the estimation. The x-axis specifies the amount of shrinkage through the root expected* SR^2*, or* κ*. The y-axis specifies the sparsity, thus the number of nonzero coefficients. For each combination of shrinkage and sparsity, the value of the OOS* R^2 *is given in colour, for which the scale is on the right side of the figure.*



(a) *Gaussian kernel*

(b) *Polynomial kernel* $d = 2$

(c) *Polynomial kernel* $d = 5$

(d) *Linear kernel* $c = 0$

Figure 9

Mappings of the OOS R^2 *from the dual-penalty method for the 50 anomaly portfolios of Kozak et al. (2020) (full sample from November 1973 to December 2017), where the Gaussian kernel, the Polynomial kernel (*$d = 2$ *and* $d = 5$*) and the Linear kernel (*$c = 0$*) are used. For these mappings, we have not demeaned the kernel matrix in the estimation. The x-axis specifies the amount of shrinkage through the root expected* SR^2*, or* κ*. The y-axis specifies the sparsity, thus the number of nonzero coefficients. For each combination of shrinkage and sparsity, the value of the OOS* R^2 *is given in colour, for which the scale is on the right side of the figure.*

Table 3

The Sharpe ratio, OOS cross-sectional R^2 and the κ for the MVE constructed per method for the 25 Fama and French (1993) portfolios. This is done for the entire data sample (July 1926 - December 2017) and for the true out of sample performance, where we withhold a sample (January 2005 - December 2017).

	Entire sample			Withheld test sample		
	SR	OOS CS R^2	κ	SR	OOS CS R^2	κ
Raw	0.4044	0.4462	0.2481	0.4758	0.4948	0.2913
PC	0.4044	0.4462	0.2481	0.4758	0.4948	0.2913
Gauss	0.4044	0.4462	0.2481	0.4758	0.4948	0.2913
Poly $d = 2$	0.4044	0.4462	0.2481	0.4758	0.4948	0.2913
Poly $d = 20$	0.4044	0.4462	0.2481	0.4751	0.4946	0.2964
Lin $c = 1$	0.4044	0.4462	0.2481	0.4758	0.4948	0.2913

Table 4

The Sharpe ratio, OOS cross-sectional R^2 and the κ for the MVE constructed per method for the 50 anomaly portfolios. This is done for the entire data sample (November 1973 - December 2017) and for the true out of sample performance, where we withhold a sample (January 2005 - December 2017).

	Entire sample			Withheld test sample		
	SR	OOS CS R^2	κ	SR	OOS CS R^2	κ
Raw	1.3321	0.2383	0.2829	2.0721	0.2795	0.3330
PC	1.3320	0.2384	0.2860	2.0721	0.2795	0.3330
Gauss	1.3321	0.2383	0.2829	2.0721	0.2795	0.3330
Poly $d = 2$	1.3321	0.2383	0.2829	2.0721	0.2795	0.3330
Poly $d = 20$	1.3321	0.2383	0.2829	2.0760	0.2782	0.3450
Lin $c = 0$	1.3321	0.2383	0.2829	2.0721	0.2795	0.3330

Derivations

$$
\begin{aligned}
S_t &= e^{-r(T-t)} E_t(M_{t,T} S_T) \qquad \text{and} \quad E_t(M_{t,T}) = 1 \\
\longrightarrow 1 &= E_t(M_{t,T} \tfrac{S_T e^{-rT}}{S_t e^{-rt}}) \\
&= E_t(M_{t,T}(1 + \tfrac{S_T - s_t}{s_t})) \\
&= E_t(M_{t,T}(1 + R_{t,T})) \\
&= E_t(M_{t,T}) + E_t(M_{t,T} R_{t,T}) \\
&= 1 + E_t(M_{t,T} R_{t,T}) \\
\longrightarrow 0 &= E_t(M_{t,T} R_{t,T})
\end{aligned}
\tag{33}
$$

The tilde on variables indicates demeaned variables, for example:

$$\tilde{F}_{t,T} = F_{t,T} - E_t(F_{t,T})F_{t,T} - \mu \quad \text{for} \quad \mu = E_t(F_{t,T}) \tag{34}$$

$$\begin{aligned}
&E_t([1 - b'(F_{t,T} - E_t(F_{t,T}))]F^h_{t,T}) && = && 0 \\
&E_t((1 - b'\tilde{F}^h_{t,T})) + E_t(M_{t,T}E_t(F^h_{t,T})) && = && 0 \\
&E_t((1 - b'\tilde{F}^h_{t,T})) + E_t(M_{t,T})E_t(F^h_{t,T}) && = && 0 \\
&E_t(\tilde{F}^h_{t,T} - b'\tilde{F}_{t,T}\tilde{F}^h_{t,T}) + E_t(F^h_{t,T}) && = && 0 \\
&E_t(F^h_{t,T} - b'\tilde{F}_{t,T}\tilde{F}^h_{t,T}) && = && 0 \\
&E_t(F^h_{t,T}) && = && b'E_t(\tilde{F}_{t,T}\tilde{F}^h_{t,T}) \\
&\mu_h && = && b'Cov_t(F_{t,T}F^h_{t,T}) \\
& && \text{thus} && \\
&\mu = b'V_t(F_{t,T}F'_{t,T}) = \Sigma_F b && \longrightarrow && b = \Sigma_F^{-1}\mu
\end{aligned} \tag{35}$$