Kernel Trick for the Cross-Section[*]

Serhiy Kozak
University of Maryland

September 23, 2022

Abstract

Characteristics-based asset pricing implicitly assumes that factor betas or risk prices are linear functions of pre-specified characteristics. Present-value identities, such as Campbell-Shiller or clean-surplus accounting, however, clearly predict that expected returns are highly non-linear functions of all characteristics. While basic non-linearities can be easily accommodated by adding non-linear functions to the set of characteristics, the problem quickly becomes infeasible once interactions of characteristics are considered. I propose a method which uses economically-driven regularization to construct a stochastic discount factor (SDF) when the set of characteristics is extended to an arbitrary—potentially infinitely-dimensional—set of non-linear functions of original characteristics. The method borrows ideas from a machine learning technique known as the "kernel trick" to circumvent the curse of dimensionality. I find that allowing for interactions and non-linearities of characteristics leads to substantially more efficient SDFs; out-of-sample Sharpe ratios for the implied MVE portfolio double.

[*] Email: sekozak@umd.edu. I thank Jacob Boudoukh, Victor DeMiguel, Shrihari Santosh, Rossen Valkanov, Michael Weber (discussant), participants at Emory, Michigan, UC San Diego, SITE Asset Pricing Theory and Computation session, and ASSA Econometric Society meetings for helpful comments and suggestions.

1 Introduction

Characteristic-based factor models have been widely used in finance to summarize the cross section of expected returns since Rosenberg (1974) and Fama and French (1992, 1993a, 1996). The main idea behind such models is that factor betas, factor risk premia, or prices of risk are functions of some pre-specified observed characteristics. In such cases, Kozak et al. (2019), Kelly et al. (2018) show that one can, equivalently, seek to explain the cross section of expected stock returns by working with characteristics-managed (or characteristics-sorted) portfolios instead of individual stock returns. It is common in the literature to construct such portfolios as *linear* sorts in any given pre-specified characteristic. Any non-linearities and interactions of characteristics are, therefore, ignored by such an approach. In this paper I argue that non-linearities and interactions of characteristics are important and develop a method of studying them that does not suffer from the curse of dimensionality.

To understand why nonlinearities and interactions are important, consider a simple model by Fama and French (2016). With clean surplus accounting, they argue that market-to-book ratio, $\frac{M}{B}$, firm's expected earnings, $\mathrm{E}[Y]$, investment, ΔB, and discount rates, r, are jointly linked by an identity:

$$\frac{M_t}{B_t} = \frac{1}{B_t}\sum_{\tau=1}^{\infty}\frac{\mathrm{E}\left(Y_{t+\tau}-\Delta B_{t+\tau}\right)}{\left(1+r\right)^{\tau}}. \tag{1}$$

In particular, Fama and French argue, that, holding everything else constant, (i) a lower value of M_t, or equivalently a higher book-to-market equity ratio, $\frac{B_t}{M_t}$, implies a higher expected return, (ii) higher expected future earnings imply a higher expected return, and (iii) higher expected growth in book equity—investment—implies a lower expected return. While, by the virtue of an identity, the book-to-market ratio, firm's expected earnings, and investment must predict future equity returns, the dependence of discount rates on these characteristics implied by the equation above is clearly highly non-linear. Moreover, as soon as we deviate from a static exercise of holding everything else constant, interactions of these variables appear. It is, therefore, unlikely that factors sorted on linear characteristics, one at a time, can summarize the cross-section of expected returns well. Non-linearities and interactions of characteristics, such as value and momentum, are potentially important.

An immediate problem that one faces in this context is the curse of dimensionality. Consider a hundred of characteristics which could be helpful in explaining the cross-section of expected returns. A naïve approach of modeling non-linearities and interactions is to include portfolios sorted on powers and interactions of the hundred original characteristics. However, even with second-order interactions of characteristics (size×mom, value×mom, etc.) one already obtains $100 \times 101/2 = 5,050$ potential factors. Allowing for the third- or higher-order interactions leads to a complete loss of tractability.

I propose a solution to the curse of dimensionality by borrowing ideas from machine learning techniques known as *kernel methods* and economic restrictions in Kozak et al. (2018); Kozak et al. (2019). I start by an observation in Kozak et al. (2018); Kozak et al. (2019) that a stochastic discount factor (SDF) can be represented by dominant principal components (PCs) of characteristics-managed portfolios. In the current context, characteristics can include any number of interactions of base characteristics, so their number is potentially very high, or even infinite. The *kernel trick* allows me to extract a large number of dominant PCs of these portfolios, even if there are infinitely many of them. Therefore, an SDF can be still well approximated by a finite number of dominant PCs.

The starting point of my method is the collection of returns on characteristics-based "features" portfolios $F_{t+1} = \Phi(Z_t)' R_{t+1}$, where R_{t+1} is a $T \times N$ matrix of returns on N stocks at times $t = 1...T$, and Z_t is a matrix of K characteristics-based instruments for each of the N stocks. A flexible non-linear transformation $\Phi\left(Z_t\right) = \left(\phi\left(Z_{t,1}\right), ..., \phi\left(Z_{t,N}\right)\right)' : \mathbb{R}^{N \times K} \to \mathbb{R}^{N \times L}$ of these characteristics rotates characteristics of any stock i, $Z_{t,i}$, into a high-dimensional (potentially infinitely-dimensional) space $\mathbb{R}^L$ of characteristic-based "features" portfolios. Such a rotation takes care of any potential variability in SDF prices of risk and thus translates a difficult conditional problem of estimating an SDF into a simpler, though potentially much higher dimensional, *unconditional* problem. The curse of dimensionality can be circumvented, however, using the kernel trick, which I discuss below.

In the next step I consider a dual PCA problem which focuses on eigenvalue decomposition of the $T \times T$ matrix FF' of returns on the features portfolios. Eigenvectors associated with largest eigenvalues of this problem are shown to be proportional to the principal components of the second-moment matrix of the features portfolio returns $F'F$.[1] In the above formulation characteristics of any two stocks enter only as inner products. The kernel trick uses a generalization of an inner product that replaces the original inner product with some non-linear function, called the *kernel*.

Suppose we start with a set of K observed stocks' i and j characteristics at times t and s, $Z_{t,i}$ and $Z_{s,j}$. For kernels $\kappa(Z_{t,i}, Z_{s,j})$ that satisfy certain regularity conditions it can be shown that there exists a mapping $\phi\left(Z_{t,i}\right) : \mathbb{R}^K \to \mathbb{R}^L$, where L is possibly infinite, and for which $\kappa(Z_{t,i}, Z_{s,j}) = \phi\left(Z_{t,i}\right)' \phi\left(Z_{s,j}\right)$. That is, the kernel is a dot product of characteristics to which the transformation $\phi(\cdot)$ has been applied. A non-trivial, arbitrary implicit transformation

[1] Connor and Korajczyk (1988) call this method "asymptotic principal components" and provide associated asymptotic theory.

function $\phi(\cdot)$ is chosen in a way that it is never calculated explicitly, allowing the possibility to use very high-dimensional $\phi(\cdot)$, since we never have to actually evaluate the data in that space. In other words, certain choices of the kernel $\kappa(\cdot,\cdot)$, which is easy to compute, lead to the exact same solution as PCA on an extended set of portfolios sorted on original characteristics, their powers and interactions of an arbitrary (potentially infinite) order. This problem can be solved at a fixed computational cost which does not increase in the order of interactions.

The final step is to combine the extracted dominant principal components into a single mean-variance efficient (MVE) portfolio, or the SDF. I rely on the method in Kozak et al. (2019) in doing so. In that paper the authors use a Bayesian prior to link mean returns on factor portfolios and their variance-covariance matrix in a way that (i) rules out near-arbitrage opportunities, and (ii) prevents portfolio weights of a marginal investor to become unbounded. The prior proves powerful in combining the multitude of factor portfolios into a single SDF. Importantly, my method is based on *economically*-driven regularization rooted in this Bayesian prior. As such, the strong economic link between expected returns and covariances allows me to obtain a robust estimate of the pricing kernel (and MVE portfolio) without overfitting the data.

Equipped with the method, I explore non-linearities and interactions of characteristics in the cross section of equity returns. First, I focus on a simple motivational example above, which includes four Fama and French (2016) factors, and the momentum factor based on Carhart (1997). I quantify the goodness of the model by the maximum out-of-sample Sharpe ratio on the optimally constructed portfolio of factors. Without interactions, the optimal portfolio is just a portfolio of five factors. With second-order interactions the number of factors increases to 20. The method allows me to increase the order of interactions to any arbitrary (potentially infinite) number at no additional computational cost. I find that higher order interactions indeed do matter and help increase the Sharpe ratio to 0.7 (from 0.2 in the case of no interactions).

Next, I consider a setting of forty anomaly characteristics and apply the same method to these data. With only second-order interactions the number of potential factors increases to almost a thousand. Third or higher order interactions make the problem completely unfeasible for the standard approach. My method, however, experiences no such shortcomings and allows me to estimate an SDF corresponding to an infinitely many interactions. Using these data I again find that allowing for interactions and non-linearities of characteristics leads to substantially more efficient SDFs and higher out-of-sample Sharpe ratios of above 3.0 for the implied MVE portfolio (relative to 1.65 in the case of no interactions).

These results survive in the full out-of-sample exercise. I split the sample at the beginning of 2005, estimate all PC rotations and SDF parameters in the pre-2005 sample and later apply these estimates to post-2004 data in the full out-of-sample sense. I find that non-linearities and interactions substantially improve the out-of-sample maximal Sharpe ratios as well. In the case of no interactions, the Sharpe ratio of the MVE portfolio is 0.5 in this period. With the radial kernel, which allows for all high-order interactions, the Sharpe ratio goes up to 1.0. While the level of Sharpe ratios drops significantly in post-2004 sample—consistent with the anomaly performance deterioration evidence documented in the literature—allowing for non-linear effects effectively doubles the out-of-sample Sharpe ratios.

The method recovers the time series of an SDF that prices equity excess returns conditionally through time, as well as conditional loadings of the SDF on every stock at each point in time. I use the SDF to infer the conditional cost of capital on any firm at any point in time non-parametrically by simply computing covariances of the firm-level realized returns with the SDF over short windows of daily data. I find that the firm-level conditional expected returns constructed in this way explain a significant fraction of variation in the firm-level realized returns. At a monthly horizon, individual firm's returns can be forecasted with an R^2 of 0.8% (relative to a benchmark of 0.2% of a constant mean). At a daily horizon, the R^2 is 0.045% (benchmark: 0.01%). This high degree of stock-level predictability aggregates to high predictability of the aggregate market index. At a monthly horizon, the equal-weighted market portfolio is predictable with an R^2 of 2.5% and a t-statistic above 3.0. The value-weighted aggregate market portfolio can be forecasted with an R^2 of 1.3% and a t-statistic of 2.5.

Related literature. The kernel trick has been widely used in the machine learning literature, especially in the context of Support Vector Machines (SVMs) and Kernel PCA (Schölkopf et al. (1997)). The application in this paper is different in several ways. Relative to SVMs, my approach is based on *economically*-driven regularization via the prior which links mean returns and covariances from Kozak et al. (2019). Relative to the Kernel PCA, the "kernel trick" is not applied directly to the data points themselves (returns), but characteristics which underly portfolio sorts. Therefore, while the method allows me to study arbitrary non-linearities and interactions in characteristics, importantly, the SDF (and the MVE portfolio) is linear in individual stock returns, that is, non-linearities appear only in variables used to sort stocks into portfolios.

Conceptually, my estimation approach is related to research on mean-variance portfolio optimization in the presence of parameter uncertainty. SDF coefficients of factors are proportional to their weights in the MVE portfolio. Accordingly, my estimator of SDF coefficients maps into constraints MVE portfolio weights studied in Brandt et al. (2009) and DeMiguel et al. (2009). My paper extends this literature by allowing for flexible non-linearities and interactions in SDF loadings.

Several recent papers argued that non-linearities and interactions should be important in the cross section of expected returns as well. Kozak et al. (2019) manually construct portfolios sorted on second-order interactions of all characteristics. Their approach is analogous to using the second-order polynomial kernel in this paper, but becomes infeasible for higher dimensions. Freyberger et al. (2017) allow for flexible non linearities in individual characteristics and show they are important. Their paper, however, ignores interactions, due to the curse of dimensionality. Gu et al. (2018) study multiple machine learning techniques to model expected returns as flexible non-linear functions of underlying characteristics. My paper instead focuses on SDF weights, which incorporate information in both means and covariances. Moreover, it relies on economically-motivated regularization based on restrictions imposed by no near-arbitrage and finite portfolio weights of marginal investors; machine learning literature typically employs purely statistical restrictions.

2 Methodology

2.1 Characterizing the SDF

2.1.1 Characteristics-based factors

One of the primary goals of empirical asset pricing is to find and characterize the empirical SDF, which summarizes the cross section of expected return on all available assets. I will focus on the projection of the "true" SDF, M_{t+1}, pricing N US stocks' excess returns R_{t+1}, on these returns:

$$M_{t+1} = 1 - b_t' \left(R_{t+1} - \mathrm{E}[R_{t+1}]\right), \tag{2}$$

where b_t is an $N \times 1$ vector of SDF coefficients. This SDF is normalized every period in a way that makes the constant term equal to unity, period by period. Additionally, note that in the above formulation I subtract *unconditional* means from factor returns. The normalization, therefore, requires that b_t absorbs variation in conditional means, $\mathrm{E}_t\left[R_{t+1}\right]$, variances, and covariances of returns.

Next, I assume that an econometrician has access to a set of K characteristics-based instruments for each of the N stocks, Z_t (with dimensions $N \times K$), that can capture all time-series and cross-sectional variation of b_t across all stocks. With no loss of generality I parameterize b_t as linear in derived (expanded) characteristics (henceforth *features*), $\Phi(Z_t)$:

$$b_t = \Phi(Z_t)b, \tag{3}$$

where $\Phi\left(Z_t\right) = \left(\phi\left(Z_{t,1}\right), ..., \phi\left(Z_{t,N}\right)\right)' : \mathbb{R}^{N\times K} \to \mathbb{R}^{N\times L}$ is an arbitrary non-linear transform of the K original instruments for each of the N stocks into L features; each of the $\phi\left(Z_{t,i}\right)$ maps K characteristics of a stock i into L features; and b is an $L \times 1$ vector of constants. For example, Z_t can contain instruments such as log-market equity, book-to-market ratio, or profitability. Non-linearities and interactions of characteristics, such as B/M × ME, value × momentum, are potentially important and can be easily accommodated via a transform $\Phi(Z_t)$.

The parametrization (3) allows me to move away from estimating SDF coefficients for each stock at each point in time, to estimating them as a single function of characteristics that applies to all stocks over time. In other words, the rotation $\Phi(\cdot)$ translates a difficult conditional problem of estimating an SDF into a simpler, though potentially much higher dimensional, *unconditional* problem.

Next, I plug in the parametrization (3) into (2) to obtain the following SDF:

$$M_{t+1} = 1 - b' \left(F_{t+1} - \mathrm{E}[F_{t+1}]\right), \tag{4}$$

where $F_{t+1} = \Phi(Z_t)' R_{t+1}$ is a vector of characteristics-based *factors*, formed as linear sorts on features $\Phi(Z_t)$.

Note that classical approaches to factor models correspond to a simple case of $\Phi(Z_t) = Z_t$. For instance, Fama and French (1992) specify three characteristics: (i) market weights, leading to the value-weighted aggregate market factor; (ii) market equity of each company – the "size" (SMB) factor; and (iii) book-to-market ratios – the "value" (HML) factor.[2] Their SDF is given by $M = 1 - b_1 F_M - b_2 F_{SMB} - b_3 F_{HML}$. Hou et al. (2015) proposes a similar SDF based on four factors, while Barillas and Shanken (2018) uses six factors. Similarly Kozak et al. (2019) consider portfolios as linear sorts on fifty underlying anomaly characteristics and show how to estimate an SDF with such a plethora of factors. Kozak et al. (2019) further consider the case of second-order interactions by explicitly constructing portfolios based on such interactions.

Note that time variation in b_t can be captured via time-series instruments in addition to any cross-sectional instruments. To accommodate such a case a set of factors needs to be extended to include all kronecker products of factors and time-series instruments: $\tilde{F}_{t+1} = z_t \otimes F_{t+1}$. Alternatively, we can assume that any time variation in aggregate risk prices is reflected in the cross-section and thus can be captured through higher-order interactions of only cross-sectional instruments. For example, suppose there is a momentum factor and its price of risk is driven by aggregate market/book ratio (M/B). Further, suppose value stocks become more sensitive to momentum shocks in times when the level of aggregate M/B goes up. Value and momentum characteristics, and their interaction, span the strategy of long value-winners, short value-losers stocks. Even though the momentum characteristic is cross-sectionally normalized, because the combined strategy focuses on primarily value stocks, and because their sensitivity to momentum increases as M/B goes up, the strategy to a large extent mimics the behavior of a single momentum factor but with a time-varying price of risk.

[2] Technically, Fama and French (1992) parameterize $\Phi(\cdot)$ as a step functions that delivers their long-short portfolio construction.

2.1.2 Estimating the SDF: economically motivated priors

In general, estimating the SDF in (4) is not easy when the number of factors is large. Indeed, the Markowitz portfolio approach procedure can be quite unreliable with a large number of assets. Instead, I rely on the approach proposed in Kozak et al. (2019). The basic idea of the approach lies in linking means and covariances via an economically motivated prior:

$$\mu \sim \mathcal{N}\left(0, \frac{\kappa^2}{\tau}\Sigma^2\right), \tag{5}$$

where μ is a vector of expected factor returns, $\mathrm{E}[F_t]$, and $\Sigma = \mathrm{E}\left[(F_t - \mathrm{E}\,F_t)'(F_t - \mathrm{E}\,F_t)\right]$ is their covariance matrix, $\tau = \mathrm{tr}[\Sigma]$, and κ is a constant that controls the strength of the prior.[3] Kozak et al. (2019) argue that this prior imposes two important economic restrictions: (i) absence of near-arbitrage opportunities, and (ii) finite portfolio holdings (SDF weights) of marginal investors.[4]

Combining prior with sample data on mean factor returns μ we get the posterior mean of the SDF coefficients b:

$$\hat{b} = \left(\Sigma + \gamma I_K\right)^{-1}\mu, \tag{6}$$

where $\gamma = \frac{\tau}{\kappa^2 T}$ is the penalty parameter and T is the number of time-period observations. Kozak et al. (2019) argue that this solution can be interpreted as a solution to a problem minimizing Hansen and Jagannathan (1991) distance subject to an L^2-norm penalty on $b'b$:

$$\hat{b} = \arg\min_b \left\{(\bar{\mu} - \Sigma b)'\,\Sigma^{-1}\left(\bar{\mu} - \Sigma b\right) + \gamma b'b\right\}, \tag{7}$$

or, equivalently, minimizing an OLS objective subject to a penalty on the model-implied maximum squared Sharpe ratio:

$$\hat{b} = \arg\min_b \left\{(\bar{\mu} - \Sigma b)'\left(\bar{\mu} - \Sigma b\right) + \gamma b'\Sigma b\right\}. \tag{8}$$

The penalty term in (7) effectively down-weights contributions of low-variance PCs to the overall maximal Sharpe ratio. To see this, consider a transformed SDF expressed in terms of principal components of original factors. The corresponding SDF coefficients are

[3] κ can be interpreted as the square root of expected maximal squared Sharpe ratio under the prior.

[4] They show that the lowest power on Σ that is consistent with aforementioned restrictions is two. Further, they argue that the prior in (5) is the flattest (least restrictive) Bayesian prior within the family of Normal priors which satisfies these two conditions.

given by:

$$\hat{b}_{P,j} = \left(\frac{d_j}{d_j + \gamma}\right) \frac{\bar{\mu}_{P,j}}{d_j}. \tag{9}$$

Note that contribution of each PC to SDF variance is $\left(\frac{d_j}{d_j+\gamma}\right)^2 \bar{\mu}_{P,j}^2$. This contribution is decreasing in d_j, so the estimator focuses primarily on large-variance PCs, that is, an SDF should be well approximated by dominant principal components (Kozak et al. (2018)). If we could extract sufficiently many dominant PCs of F_t, we can approximate an SDF well.

Proposition 1. *An SDF is well approximated by dominant PCs of factor portfolios. Therefore, if we could extract sufficiently many dominant PCs of F_t, we can recover an SDF which approximately prices all available assets' expected returns.*

Proof. See Kozak et al. (2019). □

2.2 The Kernel Trick

Let R_t denote an $N \times 1$ vector of excess returns on N assets and $X_t \equiv \Phi(Z_t)$ denote an $N \times L$ matrix of features. Rotate returns into managed portfolios:

$$F_{t+1} = X_t' R_{t+1},$$

where F_t is an $L \times 1$ vector of returns on rotated portfolios.

The unconditional covariance of returns (assume F_t are mean-zero) is given by:

$$\Sigma = \frac{1}{T} F'F = \frac{1}{T} \sum_{t=1}^{T} X_t' R_{t+1} R_{t+1}' X_t \tag{10}$$

$$= \text{vec}\,(X)' \, \text{diag}\,(R) \, \text{diag}\,(R)' \, \text{vec}\,(X)\,, \tag{11}$$

where F is a matrix of all stacked factors F_t', $\text{vec}\,(X)' = [X_1', X_2', ..., X_T']$ is an $L \times TN$ matrix and $\text{diag}\,(R)$ is an $NT \times T$ matrix with $R_1, R_2, ..., R_T$ on the diagonal. We are interested in extracting N dominant PCs of Σ for any given $\Phi(\cdot)$.

Note that we can instead extract PCs of FF' as in Connor and Korajczyk (1988), in which case the eigenvectors become the (scaled) principal components corresponding to the initial problem.[5] Therefore, I proceed with the eigenvalue-decomposition of a $T \times T$ matrix

[5] To see this, start with the singular-value decomposition of F, $F = UDV'$, and let $P = UD$ denote the matrix of principle component variables. Next, notice that $K \equiv FF' = UD^2U'$ and hence we can compute P from eigenvalue decomposition of K. Appendix A provides a more formal argument.

Ω:

$$\Omega = \underbrace{\operatorname{diag}(R)'}_{T\times NT}\underbrace{\operatorname{vec}(X)}_{TN\times L}\underbrace{\operatorname{vec}(X)'}_{L\times TN}\underbrace{\operatorname{diag}(R)}_{NT\times T}, \tag{12}$$

where

$$\underbrace{\operatorname{vec}(X)}_{TN\times L}\underbrace{\operatorname{vec}(X)'}_{L\times TN} = \begin{bmatrix} X_1X_1' & \cdots & X_1X_T' \\ \vdots & & \vdots \\ X_TX_1' & \cdots & X_TX_T' \end{bmatrix}_{NT\times NT},$$

where each X_tX_s' is an $N\times N$ matrix consisting of all inner products of features for each pair of stocks (i,j) at times t and s, respectively:

$$X_tX_s' = \Phi(Z_t)\Phi(Z_s)' \equiv \mathcal{K}(Z_t, Z_s), \tag{13}$$

where $\mathcal{K}(Z_t, Z_s)$ is the $N\times N$ matrix of *kernels*, $\kappa(Z_{t,i}, Z_{s,j})$ for stocks (i,j), and Z is a matrix of observed characteristics. Note that in the standard linear PCA approach Kelly et al. (2017) features exactly coincide with base characteristics, $X_{t,i} \equiv \phi(Z_{t,i}) = Z_{t,i}$. In that case, each of the kernels is given by $\kappa(Z_{t,i}, Z_{s,j}) = Z_{t,i}'Z_{s,j}$ — the inner product of observed characteristics of stocks i and j for observations given at times t and s, respectively.

In general, however, the set of features $X_t = \Phi(Z_t)$ is potentially much larger than the set of observed characteristics Z_t. Inner products of features thus involve a large—potentially infinite—number of multiplications and additions corresponding to all elements in $\phi(Z_{t,s})$. For instance, we might want to include powers of basic characteristics to capture non-linearities as in Freyberger et al. (2017). Similarly, interactions of basic characteristics can be important, such as value-momentum, value-size etc. Even if only second-order interactions are included, as in Kozak et al. (2019), the size of the characteristics space grows exponentially, making most classical techniques infeasible. Higher-order non-linearities lead to even worse curse of dimensionality.

The "kernel trick"—a popular technique in machine learning—offers an alternative approach designed to sidestep the curse of dimensionality. Note that equation (13) does not require $\phi(\cdot)$ in explicit form — they are only needed as dot products. Therefore, we are able to use these dot products without actually performing the map $\phi(\cdot)$: for some choices of a kernel $\kappa(\cdot,\cdot)$, it can be shown by methods of functional analysis that there exists a map into some dot product space $\mathbb{R}^L$ (possibly of infinite dimension) such that $\kappa(\cdot,\cdot)$ computes the dot product in the space of features $\mathbb{R}^L$.

Theorem 2.1. *(Mercer's theorem) For any positive-definite kernel κ on a space $\mathcal{X}$, there* *exists a Hilbert space $\mathcal{F}$ and a mapping ϕ: $\mathcal{X} \to \mathcal{F}$ such that*

$$\kappa(x, x') = \langle \phi(x), \phi(x') \rangle, \quad \forall x, x' \in \mathcal{X},$$

where $\langle u, v \rangle$ represents the dot product in the Hilbert space between any two points $u, v \in \mathcal{F}$.[6][7]

Proof. See Aronszajn (1950). □

Proposition 2. *Any algorithm for vectorial data that can be expressed only in terms of dot products between vectors can be performed implicitly in the feature space associated with any kernel, by replacing each dot product by a kernel evaluation.*

In other words, instead of performing the mapping $\phi(\cdot)$ of characteristics into features explicitly, we can pick a kernel—such as polynomial or radial kernels below—and simply replace the dot product of features by kernel evaluation. For certain choices of kernels, such a procedure is mathematically equivalent to performing PCA in the space of all features portfolio returns.

2.2.1 Examples of Kernels

Kernels which have successfully been used in Support Vector Machines (Schölkopf et al. (1996)) include: (i) polynomial kernels

$$\kappa(\mathbf{z}_i, \mathbf{z}_j) = (c + \langle \mathbf{z}_i, \mathbf{z}_j \rangle)^d, \tag{14}$$

for two vectors of characteristics $\mathbf{z}_i$ and $\mathbf{z}_j$ for some two stocks i and j, and a free parameter c, trading off the influence of higher-order versus lower-order terms in the polynomial; and (ii) Gaussian (radial) kernels

$$\kappa(\mathbf{z}_i, \mathbf{z}_j) = \exp\left(-c \left\|\mathbf{z}_i - \mathbf{z}_j\right\|^2\right), \tag{15}$$

where c is a constant controlling the roughness of the kernel.

[6] A Hilbert space is a vector space endowed with a dot product (a strictly positive and symmetric bilinear form), that is complete for the norm induced. $\mathbb{R}^p$ with the classic inner product is an example of a finite-dimensional Hilbert space.

[7] A function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *positive definite kernel* if it is symmetric (that is, $\kappa(x, x') = \kappa(x', x), \ \ \forall x, x' \in \mathcal{X}$), and positive definite, that is, for any $n > 0$, any choice of n objects $x_1, x_n \in \mathcal{X}$, and any choice of real number $c_1, c_n \in \mathbb{R}$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_j c_j \kappa(x_i, x_j) \geq 0.$$

A positive definite kernel κ is called a valid kernel, a Mercer's kernel, or simply kernel.

It can be shown that polynomial kernels of degree d correspond to a map $\phi(\cdot)$ into a feature space which is spanned by all products of d entries of an input pattern.

Second-order polynomial kernel. Consider a case of two stocks and two characteristics as in Schölkopf et al. (1997):

$$\langle \mathbf{z}_i, \mathbf{z}_j \rangle^2 = (z_{i1}^2, z_{i1}z_{i2}, z_{i2}z_{i1}, z_{i2}^2)(z_{j1}^2, z_{j1}z_{j2}, z_{j2}z_{j1}, z_{j2}^2)'. \tag{16}$$

In other words, in this simple case kernel evaluation is mathematically equivalent to replacing the set of two characteristics (z_{i1}, z_{i2}) for each stock with features $(z_{i1}^2, z_{i1}z_{i2}, z_{i2}z_{i1}, z_{i2}^2)$. In case of $c \neq 0$, the expanded set of features also includes original characteristics (z_{i1}, z_{i2}) with their weight relative to the second-order terms governed by c. For instance, if $c = 1$, we get:

$$\begin{aligned} (1 + \langle \mathbf{z}_i, \mathbf{z}_j \rangle)^2 &= (1 + z_{i1}z_{j1} + z_{i2}z_{j2})^2 && (17)\\ &= 1 + 2z_{i1}z_{j1} + 2z_{i2}z_{j2} + z_{i1}^2 z_{j1}^2 + z_{i2}^2 z_{j2}^2 + 2z_{i1}z_{i2}z_{j1}z_{j2} && (18)\\ &= \langle \Phi(\mathbf{z}_i), \Phi(\mathbf{z}_j) \rangle, && (19) \end{aligned}$$

where $\Phi(\mathbf{z}) \rightarrow \left\{1, \sqrt{2}z_1, \sqrt{2}z_2, z_1^2, z_2^2, \sqrt{2}z_1z_2\right\}$.

Gaussian (radial) kernel. Interestingly, the radial kernel in (15) corresponds to a map $\phi(\cdot)$ into an infinitely-dimensional feature space.

For simplicity, suppose $c = \frac{1}{2}$. The Gaussian kernel can then the expanded as,

$$\exp\left(-\frac{||\mathbf{z}_i - \mathbf{z}_j||^2}{2}\right) = C\Big\{1 - \underbrace{\frac{\langle \mathbf{z}_i, \mathbf{z}_j \rangle}{1!}}_{1^{st}\text{-order}} + \underbrace{\frac{\langle \mathbf{z}_i, \mathbf{z}_j \rangle^2}{2!}}_{2^{nd}\text{-order}} - \underbrace{\frac{\langle \mathbf{z}_i, \mathbf{z}_j \rangle^3}{3!}}_{3^{rd}\text{-order}} + \dots\Big\}$$

where, $C = \exp\left(-\frac{1}{2}||\mathbf{z}_i||^2\right) \exp\left(-\frac{1}{2}||\mathbf{z}_j||^2\right)$.

From the calculations for polynomial kernels above, we know that $\langle \mathbf{z}_i, \mathbf{z}_j \rangle^n$ will yield n-order terms. Since the Gaussian has an infinite series expansion, we get terms of all orders till infinity.

Note that a kernel imposes certain restrictions on weights of characteristics-managed portfolios. These weights cannot be chosen fully flexibly but can be controlled by the kernel parameter c. Simple algebra shows that c effectively shifts weights between higher- and lower-terms in the features mapping. For instance, for polynomial kernels, as $c \rightarrow 0$, a kernel contains only higher-order terms, as we saw in the second-order polynomial case above.

Similarly, as $c \to \infty$, a kernel puts the entire weight on the first-order terms and thus effectively collapses to the linear kernel.

2.2.2 Centering

Recall $\mathcal{K}\left(Z_t, Z_s\right) \equiv \Phi\left(Z_t\right)\Phi\left(Z_s\right)'$, where $X_t \equiv \Phi\left(Z_t\right) = \left(\phi\left(Z_{t,1}\right), ..., \phi\left(Z_{t,N}\right)\right)'$. In practice, to preserve the interpretation of features as long-short portfolios, we want to center each feature in the cross-section, that is, subtract the mean across all N observations.

In particular, we are looking for a kernel sub-matrix based on de-meaned characteristics,

$$\tilde{\mathcal{K}} \equiv \tilde{\mathcal{K}}(Z_t, Z_s) = \left(\Phi_t - \bar{\Phi}_t\right)\left(\Phi_s - \bar{\Phi}_s\right)',$$

where $\bar{\Phi} = M\Phi$ and $M = \frac{1}{N}\mathbf{1}_N\mathbf{1}_N'$.

It turns out that we can compute $\tilde{\mathcal{K}}$ even when the mapping Φ is infinitely-dimensional by double-centering the kernel matrix $\mathcal{K}$:

$$\begin{aligned} \tilde{\mathcal{K}} = \mathcal{K} - \Phi_t\bar{\Phi}_t' - \bar{\Phi}_s\Phi_s' + \bar{\Phi}_t\bar{\Phi}_s' &= \mathcal{K} - \mathcal{K}M - M\mathcal{K} + M\mathcal{K}M \qquad (20) \\ &= (I - M)\,\mathcal{K}\,(I - M)\,. \qquad (21) \end{aligned}$$

2.3 Approximate SDF

Going back to equation (12), and using centered features, we obtain:

$$\Omega = \begin{bmatrix} R_1'\tilde{\mathcal{K}}\left(Z_1, Z_1\right)R_1 & \cdots & R_1'\tilde{\mathcal{K}}\left(Z_1, Z_T\right)R_T \\ \vdots & & \vdots \\ R_T'\tilde{\mathcal{K}}\left(Z_T, Z_1\right)R_1 & \cdots & R_T'\tilde{\mathcal{K}}\left(Z_T, Z_T\right)R_T \end{bmatrix}_{T\times T}, \qquad (22)$$

where $\tilde{\mathcal{K}}$ denotes a double-centered kernel matrix given by $\tilde{\mathcal{K}} = (I - M)\,\mathcal{K}\,(I - M)$ with $M = \frac{1}{N}\mathbf{1}\mathbf{1}'$. Note that Ω does not depend on the number of features we consider. For a fixed T we can therefore solve problems corresponding to infinitely dimensional spaces of features $X_{t,s}$.

The last step is to compute eigenvectors of Ω, which are the datapoints projected on the respective principal components, that is, eigenvectors coincide with PCs of the original problem.

2.4 The Algorithm

1. Map K observed stock's i characteristics at time t, $Z_{t,i}$, into an L-dimensional space of features $X_{t,i} \equiv \phi\left(Z_{t,i}\right)$, where $\phi\left(Z_{t,i}\right) : \mathbb{R}^{K} \rightarrow \mathbb{R}^{L}$; L is high dimensional, possibly infinite, since interactions of characteristics are important. Note, however, that $\phi\left(\cdot\right)$ is never calculated explicitly.

2. Rotate the problem (of finding largest N PCs of F_t) in a way that only inner products of $\phi(\cdot)$ need to be computed.

3. Replace the inner products by a *kernel.* The kernel allows me to operate in a high-dimensional, implicit feature space without ever computing the implied features in that space — the "kernel trick".

4. Extract largest N PCs using the *kernelized* matrix – these correspond to PCs of all managed portfolios based on $\phi\left(Z_{t,i}\right)$.

5. Use these PCs as input to the algorithm in Kozak et al. (2019) to construct an SDF. Note that this SDF is equivalent to an SDF constructed from an expanded set of features, which include original characteristics, as well as their non-linear transformations and interactions, possibly infinitely many. Also note that while the constructed SDF allows for non-linearities in characteristics, it is still linear in returns, that is, it is just a portfolio of original underlying equities.

3 Results

3.1 Simulations

I simulate a factor model where loadings β depend on non-linear functions of base (simulated) characteristics. I then use and compare four methods to extract latent factors: (i) standard PCA using the cross-section of individual stocks; (ii) PCA using the cross-section of managed portfolios constructed as linear functions of base characteristics; (iii) PCA using the cross-section of managed portfolios that include non-linear functions/interactions used to obtain β — this should work well by design; (iv) a kernel-based PCA method in this paper—henceforth CK-PCA (Characteristics-Kernel PCA)— with the Gaussian kernel.

Concretely, I simulate 20 years of data and a cross section of 1,000 stocks. I assume a single factor model (e.g., CAPM). An econometrician has access to five characteristics which he believes could contain useful signals for betas. The true (simulated) model is given by:

$$R_{t+1,i} = \beta_{t,i} \times F_{t+1} + \epsilon_{t+1}, \tag{23}$$

where $\beta_{t,i} = c_{t,i}^{(1)} \times c_{t,i}^{(2)} \times c_{t,i}^{(3)}$. That is, the true beta is an interaction of three of the five characteristics an econometrician has access to. The remaining two characteristics are useless.

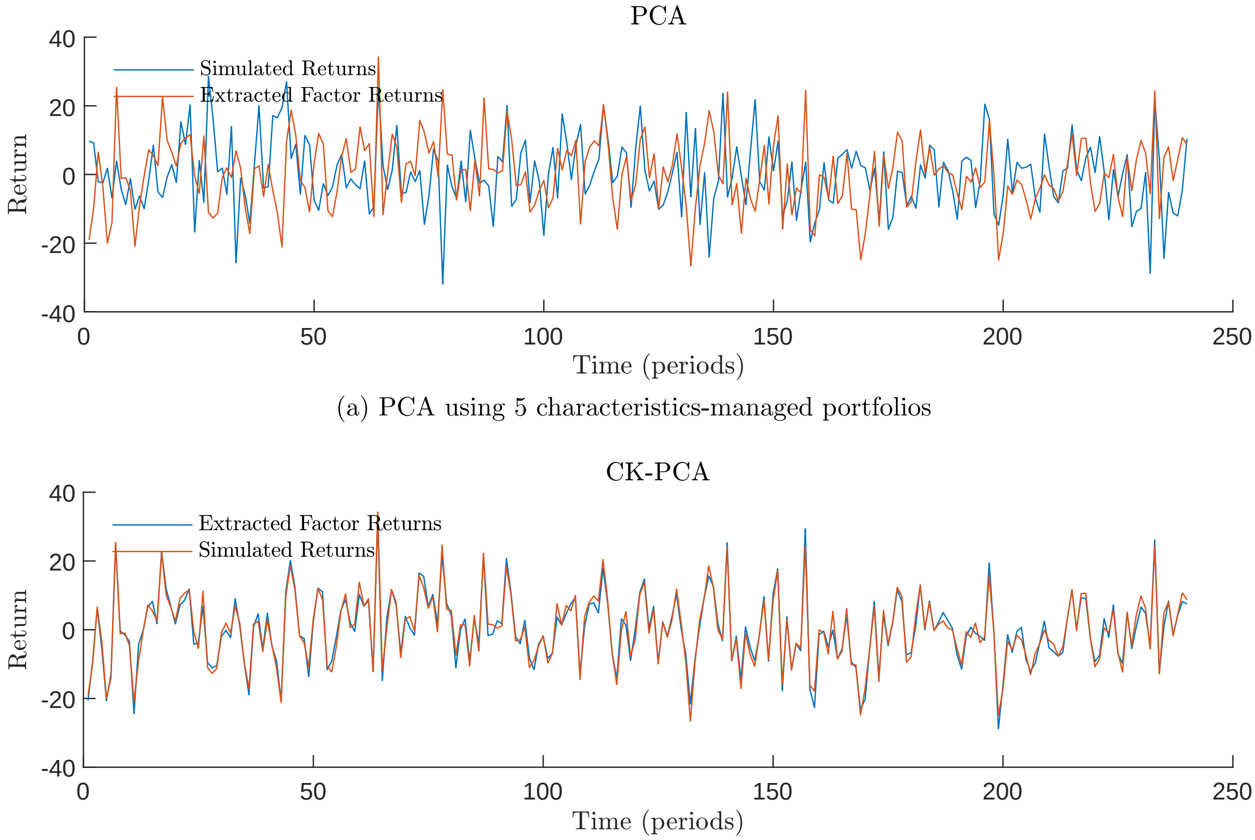
In such a setup using one characteristic at a time—assuming linearity in characteristics—would not typically recover the underlying factor or its corresponding betas. To verify this conjecture, as well as test the ability of the CK-PCA method in addressing this problem, I use each of the four methods mentioned above to recover F_{t+1} and $\beta_{t,i}$.

3.1.1 Recovered factor

Method (i) fail to recover the time-series of the factor F_t. Method (iii), which sorts stocks into portfolio based on the product of the first three characteristics, works by design.

Next, I investigate the ability of the PCA applied to portfolios sorted linearly on characteristics (method (ii) from above) and CK-PCA method (method (iv), this paper) to recover the true factor. Figure 1 shows the time series of the true (simulated) factor and the recovered factor using two methods: (1) PCA applied to five managed portfolios sorted linearly on each of the characteristics (method (ii) from above) in Panel (a); and (2) CK-PCA using Gaussian (radial) kernel in Panel (b).

Panel (a) of the figure shows that PCA applied to portfolios sorted linearly on characteristics, one at a time, as in Kelly et al. (2018); Kozak et al. (2019) is unable to find the



(a) PCA using 5 characteristics-managed portfolios

(b) CK-PCA (this paper) using Gaussian kernel

Figure 1: **Simulated and extracted factor returns.** The figure plots simulated and recovered factor returns based on the model in (23) using two methods: (a) PCA using the cross-section of managed portfolios constructed as linear functions of base characteristics, and (b) an agnostic CK-PCA (Characteristics-Kernel PCA) method that uses the Gaussian kernel.

correct factor. The CK-PCA method in Panel (b), on the other hand, works well in recovering the true factor. The correlation between the true factor and the factor extracted using this method is 0.98 (it is only 0.13 for the method in Panel (a)). The time-series of returns on the true factor (blue) and the extracted factor (red) in Panel (b) Figure 1 match very closely.

3.2 Empirical analysis

3.2.1 Data

I start with the universe of U.S. firms in CRSP. I construct two independent sets of characteristics. The first set relies on characteristics underlying the four factors from Fama and French (2015), excluding the value-weighted market, and the momentum factor from Carhart (1997). The second set is based on forty equity characteristics underlying common "anomalies" in the literature, constructed as in Kozak et al. (2019).

In order to focus exclusively on the cross-sectional aspect of return predictability, remove the influence of outliers, and keep constant leverage across all portfolios, I perform certain normalizations of characteristics that define our characteristics-based factors. First, similarly to Asness et al. (2014); Freyberger et al. (2017); Kozak et al. (2019), I perform a simple rank transformation for each characteristic. For each characteristic i of a stock s at a given time t, denoted as $c_{s,t}^{i}$, I sort all stocks based on the values of their respective characteristics $c_{s,t}^{i}$ and rank them cross-sectionally (across all s) from 1 to n_t, where n_t is the number of stocks at t for which this characteristic is available.[8] I then normalize all ranks by dividing by n_t+1 to obtain the value of the rank transform:

$$rc_{s,t}^{i} = \frac{\text{rank}\left(c_{s,t}^{i}\right)}{n_t+1}. \tag{24}$$

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

$$z_{s,t}^{i} = \frac{\left(rc_{s,t}^{i} - \bar{rc}_{t}^{i}\right)}{\frac{1}{n_t}\sum_{s=1}^{n_t}\left|rc_{s,t}^{i} - \bar{rc}_{t}^{i}\right|}, \tag{25}$$

where $\bar{rc}_{t}^{i} = \frac{1}{n_t}\sum_{s=1}^{n_t} rc_{s,t}^{i}$. The resulting zero-investment long-short portfolios of transformed characteristics $z_{s,t}^{i}$ are insensitive to outliers and have the average absolute weight equal to unity. Finally, I combine all transformed characteristics $z_{s,t}^{i}$ for all stocks into a matrix of instruments Z_t.[9] Interaction with returns, $F_t = Z_{t-1}'R_t$, then yields one factor for each characteristic.

To ensure that the results are not driven by very small illiquid stocks, I exclude small-cap stocks with market caps below 0.01% of aggregate stock market capitalization at each point in time.[10] In all of our analysis I use *daily* returns from CRSP for each individual stock. Using daily data allows me to estimate second moments much more precisely than with monthly data and focus on uncertainty in means while largely ignoring negligibly small

[8] If two stocks are "tied", I assign the average rank to both. For example, if two firms have the lowest value of c, they are both assigned a rank of 1.5 (the average of 1 and 2). This preserves any symmetry in the underlying characteristic.

[9] If $z_{s,t}^{i}$ is missing I replace it with the mean value, zero.

[10] For example, for an aggregate stock market capitalization of $20tn, I keep only stocks with market caps above $2bn.

uncertainty in covariance estimates (with exceptions as noted below). I adjust daily portfolio weights on individual stocks within each month to correspond to a monthly-rebalanced buy-and-hold strategy during that month. Table 1 in the Appendix shows the annualized mean returns for the anomaly portfolios.

3.2.2 Constructing an SDF

I use the algorithm in Section 2.4 to construct an SDF (or an MVE portfolio) based on the set of anomaly portfolios. In particular, I use the CK-PCA method to construct the $T \times T$ kernel matrix Ω in (22). Next, I compute T dominant eigenvectors of this matrix. As explained earlier, these (scaled) eigenvectors exactly coincide with the T largest principal components of a variance-covariance matrix of returns formed on all non-linear functions and interactions of characteristics underlying the specific kernel function. Next, I use these T largest PCs as an input to the method in Kozak et al. (2019) to construct an SDF. The scaling of the PCs is preserved, so the shrinkage method pays more attention to PCs with higher variance. At each point the output of the method is a single time series of an SDF, or, equivalently, returns on the mean-variance efficient portfolio, which aggregates information in all (implied) characteristics-sorted portfolios.

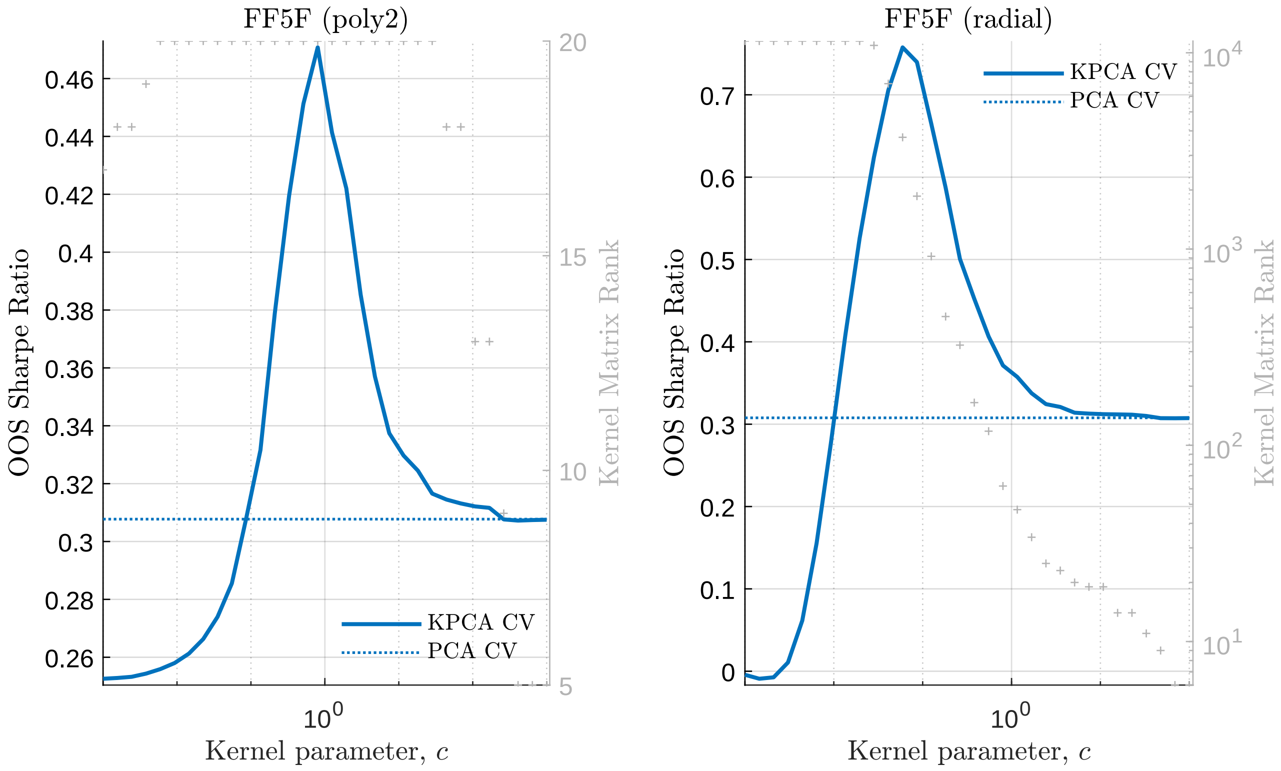
The method requires choosing two parameters: (i) κ (or, equivalently, γ) in the prior in (5) and (7), which has an economic interpretation of the root expected squared Sharpe ratio under the prior, and (ii) a kernel-specific parameter, such as c in polynomial kernel in (14) or σ^2 in radial kernel in (15). I pick both parameters using the K-fold cross validation.

Specifically, for γ, I divide the historic data into $K = 5$ equal sub-samples. Then, for each possible γ, I compute a vector of SDF coefficients, $\hat{b}$, by applying (6) to $K - 1$ of these sub-samples. I evaluate the "out-of-sample" (OOS) fit of the resulting model on the single withheld subsample. Consistent with the penalized objective (8), I compute the OOS R-squared as

$$R^2_{\text{oos}} = 1 - \frac{\left(\bar{\mu}_2 - \overline{\Sigma}_2 \hat{b}\right)' \left(\bar{\mu}_2 - \overline{\Sigma}_2 \hat{b}\right)}{\bar{\mu}_2' \bar{\mu}_2}, \tag{26}$$

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

Lastly, I pick the kernel specific parameter in order to maximize the out-of-sample Sharpe ratio of an SDF implied by the given kernel, for an optimal choice of γ above.



(a) Polynomial kernel, 2nd order

(b) Radial kernel

Figure 2: **Cross-validated Sharpe ratios (Fama-French-Carhart 5 factors).** Maximum cross-validated Sharpe ratios delivered by a kernel for a specific choice of a kernel parameter, denoted as c. Each point on the blue solid line corresponds to an SDF with a parameter γ selected optimally via cross validation, for a given value of the kernel parameter c. The dotted line shows the level of the cross-validated Sharpe ratio for the linear kernel (method (ii) – PCA on characteristics-managed portfolios), which does not depend on c. Panel (a) uses the polynomial kernel of the second order. Panel (b) uses the Gaussian (radial) kernel.

3.2.3 Five Fama-French-Carhart factors

Cross-validated Sharpe ratios implied by the optimal SDF. In Figure 2 I plot maximum cross-validated Sharpe ratios delivered by a kernel for a specific choice of a kernel parameter, denoted as c. Each point on the blue solid line corresponds to an SDF with a parameter γ selected optimally via cross validation, for a given value of the kernel parameter c. The dotted line shows the level of the cross-validated Sharpe ratio for the linear kernel (method (ii) – PCA on characteristics-managed portfolios), which does not depend on c. Recall that c controls the weight on higher-order terms relative to the weight on lower-order terms. In particular, high levels of c approximately corresponds to the linear kernel, as the higher-order terms are ignored. Figure 2 shows that for high values of c the cross-validated Sharpe ratios of the non-linear kernel indeed converge to that of the linear one. Similarly, low values of c correspond to a kernel that puts most weight on higher-order terms (e.g., for polynomial kernel of the second order, only second-order terms are used as $c \to 0$). Panel (a) of Figure 2 shows that such a kernel underperforms relative to the linear one.

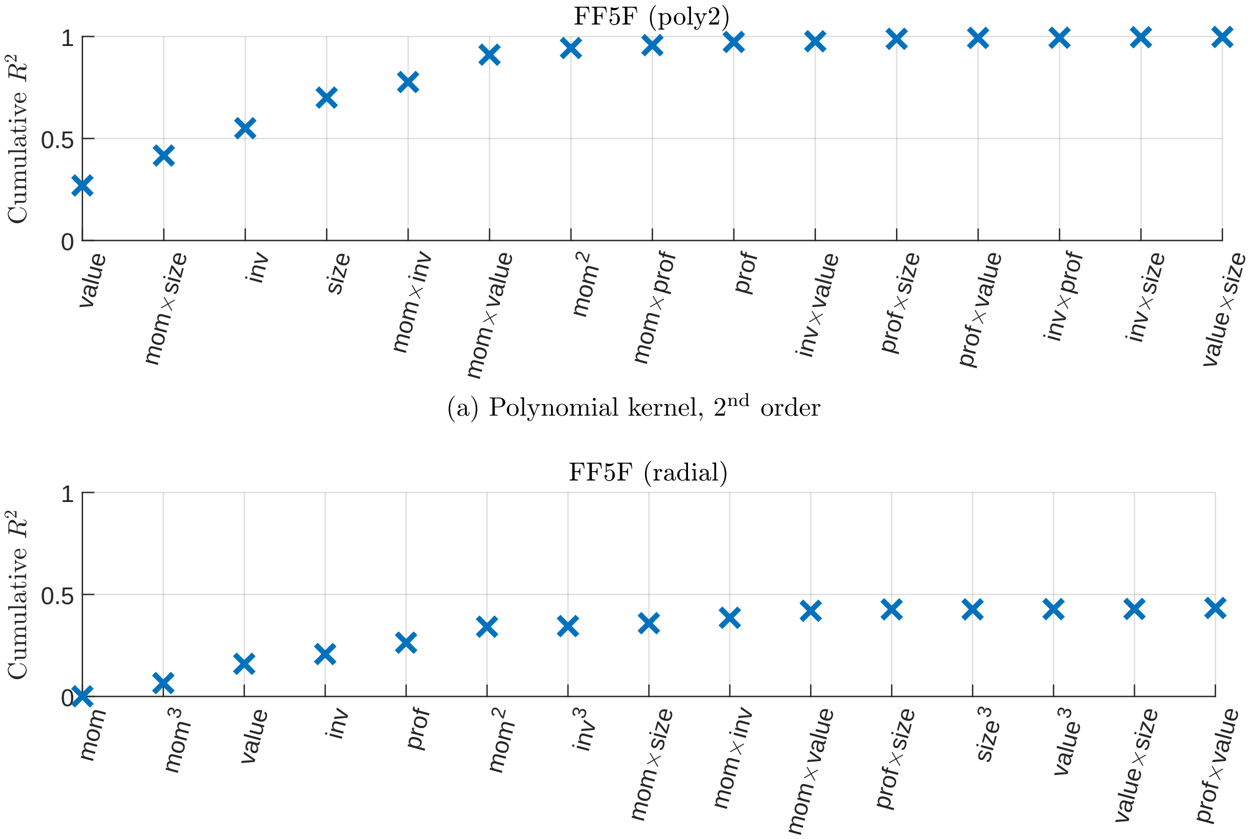
Panel (a) of the figure uses the polynomial kernel of the second order. Recall that this kernel is equivalent to PCA on characteristics-managed portfolios where the set of characteristics is expanded to include all interactions and second powers of base characteristics. The panel shows that including second-order terms improves the cross-validated Sharpe ratios from around 0.31 to 0.47.

Grey "+" markers depict an empirical rank of the kernel matrix Ω in (22). In the case when c is large, which approximately corresponds to the linear kernel, we indeed see that the rank converges to the number of base characteristics (five for Fama-French-Carhart factors). For very low values of c—when only second-order terms are present, the rank converges to the number of second-order terms, $5 \times 6/2 = 15$. For medium range values of c both first- and second-order terms are present, for the total rank of $15 + 5 = 20$.

Panel (b) uses the radial kernel. Recall that this kernel implicitly allows for an infinite number of interactions. The kernel more than doubles the cross-validates Sharpe ratios: they increase to about 0.75 relative to the linear kernel. As expected, for large values of c the rank of the kernel matrix converges to the number of base characteristics – five. For small values of c, however, any arbitrary interactions are allowed for, so the rank of the kernel matrix is maximal and equals the total number of time-series observations – roughly 12,000.

Which features matter most? Figure 3 explores the importance of each characteristic and their interactions/non-linearities in the final SDF. First, I construct an optimal SDF corresponding to a given kernel: second-order polynomial kernel in Panel (a) and Gaussian (radial) kernel in Panel (b). Second, I project this SDF onto a set of managed portfolio returns (scaled to have same volatility) based on original characteristics, their second-order interactions, second and third powers of characteristics. Lastly, I sort characteristics based on the absolute magnitude of an SDF coefficient on its managed portfolio. I report these characteristics, as well as the cumulative R^2 when this characteristics is added to an SDF.

For the second-order polynomial kernel the projection trivially achieves the R-squared of one, since the set of variables includes all first- and second-order transforms of characteristics. Even for a small number of factors, below 15, the cumulative R-squared approximately equals one, as can be seen from Panel (a) of Figure 3. However, this is no longer the case for the



(a) Polynomial kernel, 2^{nd} order

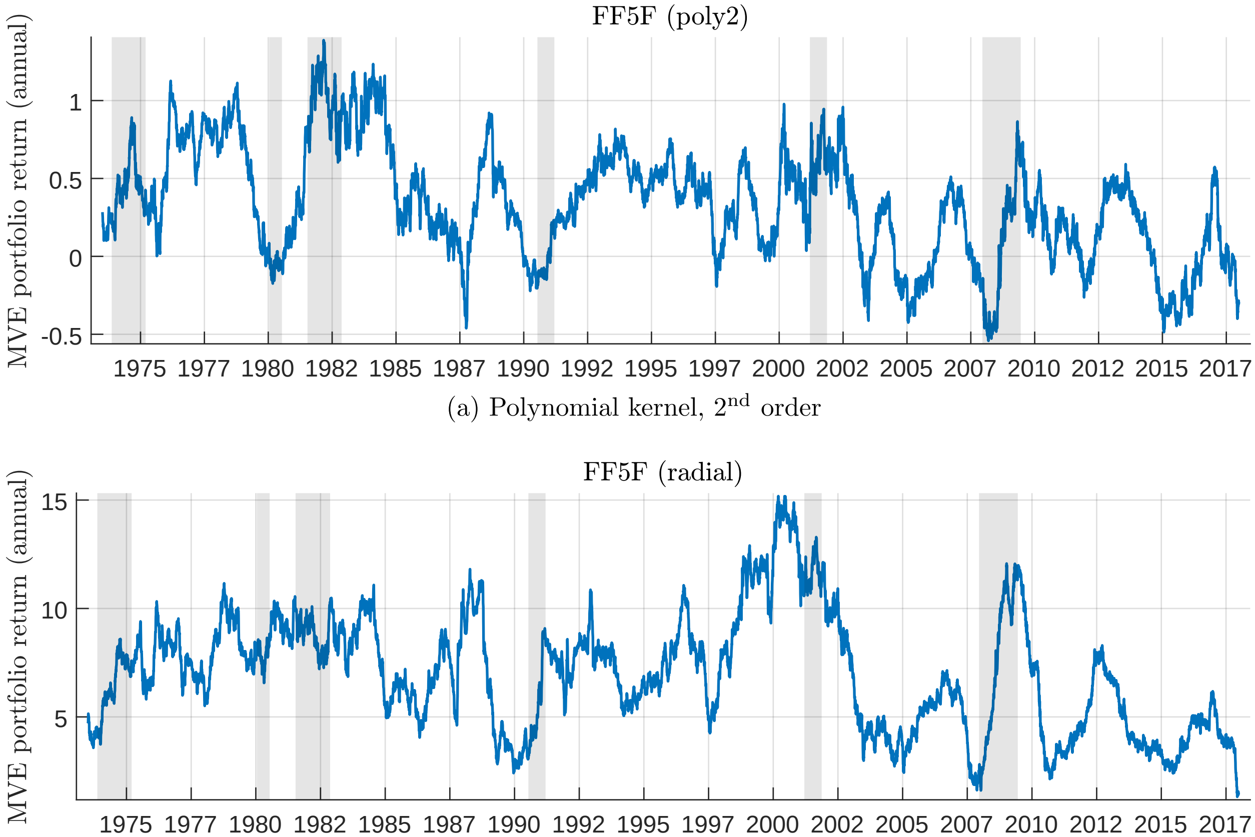
(b) Radial kernel

Figure 3: **Important features (Fama-French-Carhart 5 factors).** The figure shows characteristics that correspond to largest SDF coefficients and their contribution to the cumulative R^2 when added to an SDF. Panel (a) uses the polynomial kernel of the second order. Panel (b) uses the Gaussian (radial) kernel.

radial kernel – the R-squared of the projection is not equal to one, and remains significantly lower than one for a small number of included features. For instance, with fifteen features it remains below 0.5, as can be seen from Panel (b) of the figure.

The figure shows which characteristics are the most important ones for a given kernel. As the Panel (a) shows, base characteristics such as value, investment, and size are important for the polynomial kernel, as well as some their interactions with momentum. Similar characteristics are also important for the radial kernel, as can be seen from Panel (b). However, higher-order terms, such as mom^3, inv^3 become important for the construction of an SDF.

To conclude, exploiting non-linearities in the five Fama-French-Carhart characteristics does indeed allow me to recover the MVE portfolio and the corresponding pricing kernel better than in the case when linearity in characteristics is assumed. Figure 4 depicts empirical performance of the MVE portfolio implied by each of the two SDFs.



(a) Polynomial kernel, 2$^{\text{nd}}$ order

(b) Radial kernel

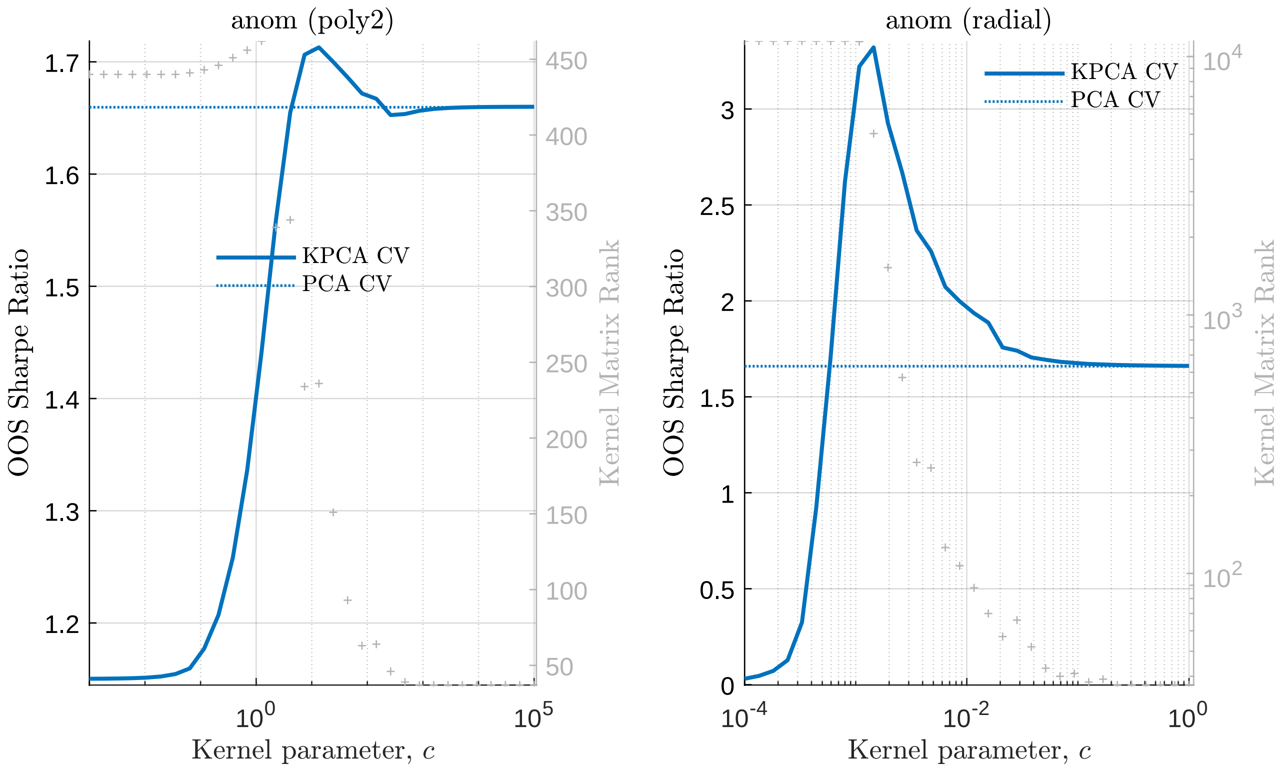
Figure 4: **MVE portfolio returns (Fama-French-Carhart 5 factors).** The figure depicts empirical performance of the MVE portfolio implied by an SDF constructed using the second-order polynomial kernel (Panel a) and using the Gaussian (radial) in Panel (b).

3.2.4 Forty anomaly factors

I now repeat the same exercise for forty anomaly portfolios – the main dataset.

Cross-validated Sharpe ratios implied by the optimal SDF. In Figure 5 I plot maximum cross-validated Sharpe ratios delivered by a kernel for a specific choice of a kernel parameter, denoted as c. The figure shows that the second-order polynomial kernel increases cross-validate Sharpe ratios only mildly. On the other hand, the Gaussian (radial) kernel in Panel (b) more than doubles cross-validated Sharpe ratios (from around 1.5 to above 3). This improvement corresponds to mid-range values of the kernel parameter c.

Which features matter most? I now investigate which characteristics or features matter most for constructing an optimal SDF or the MVE portfolio with maximal Sharpe ratios.



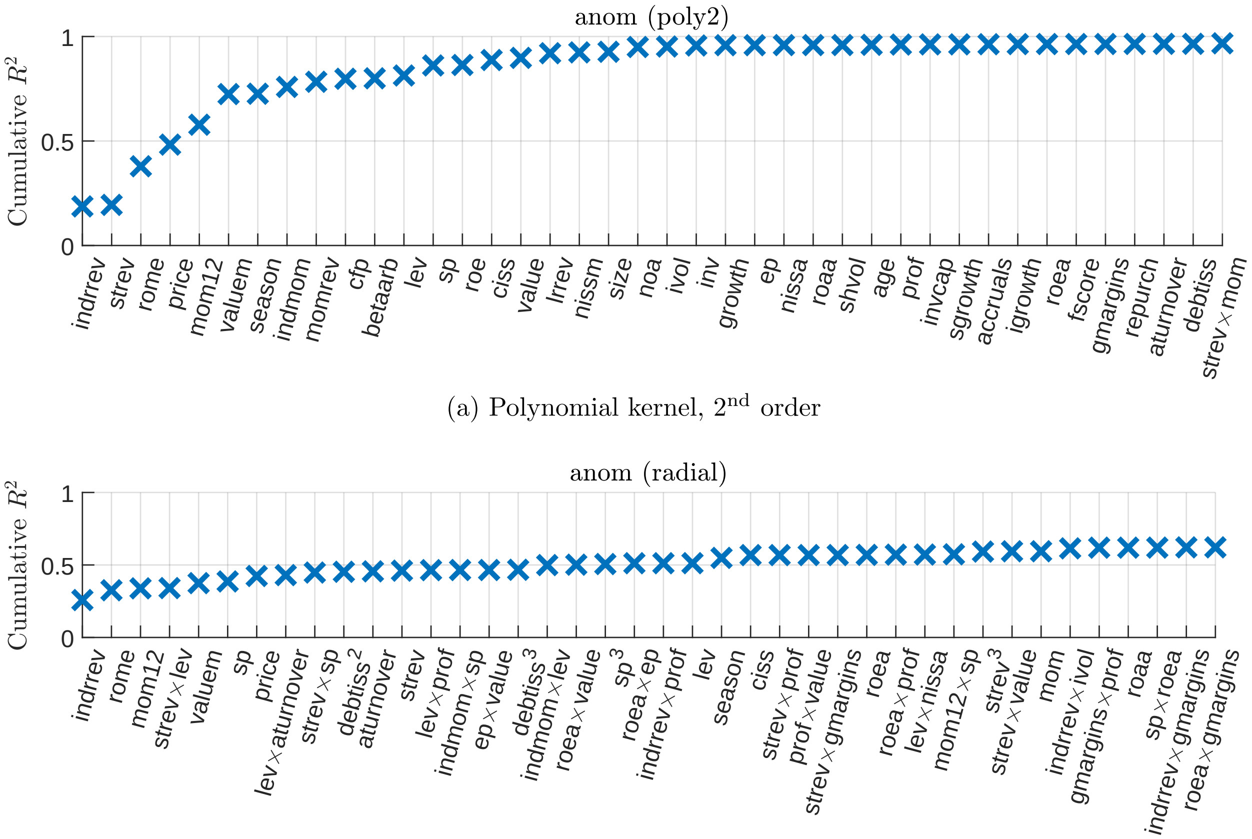
(a) Polynomial kernel, 2nd order

(b) Radial kernel

Figure 5: **Cross-validated Sharpe ratios (40 anomaly factors).** Maximum cross-validated Sharpe ratios delivered by a kernel for a specific choice of a kernel parameter, denoted as c. Each point on the blue solid line corresponds to an SDF with a parameter γ selected optimally via cross validation, for a given value of the kernel parameter c. The dotted line shows the level of the cross-validated Sharpe ratio for the linear kernel (method (ii) – PCA on characteristics-managed portfolios), which does not depend on c. Panel (a) uses the polynomial kernel of the second order. Panel (b) uses the Gaussian (radial) kernel.

In Figure 6 I project this SDF onto a set of managed portfolio returns (scaled to have same volatility) based on original characteristics, their second-order interactions, second and third powers of characteristics. Lastly, I sort characteristics based on the absolute magnitude of an SDF coefficient on its managed portfolio. I report these characteristics, as well as the cumulative R^2 when this characteristics is added to an SDF.

For the second-order polynomial kernel mostly only base linear characteristics are important, as can be seen from Panel (a) of Figure 6. Moreover, with a relatively small number of such characteristics nearly maximal cross-validated R-squared can be achieved. The situation is different for the 40 anomaly portfolios and the radial kernel. The cumulative R-squared stays around 0.6 even with more than 30 features. Moreover, many of the features with



(a) Polynomial kernel, 2nd order

(b) Radial kernel

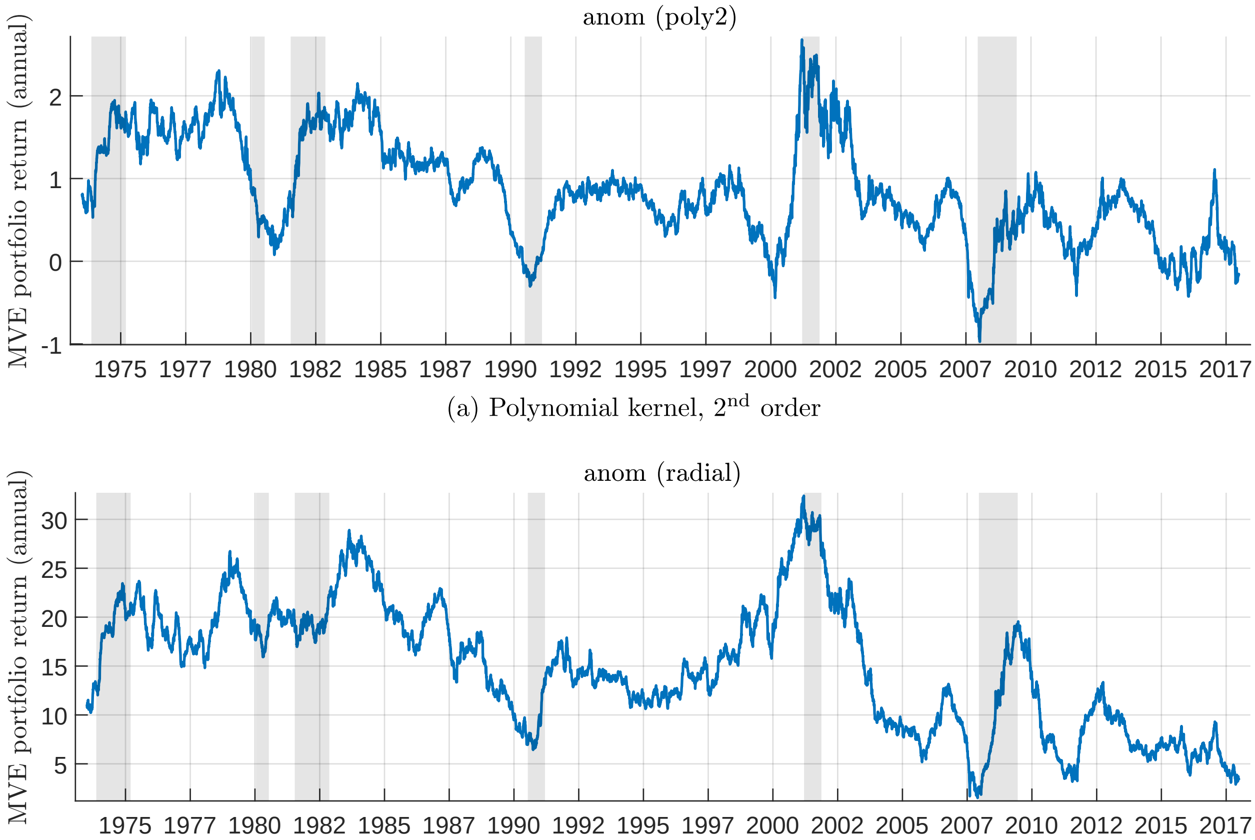
Figure 6: **Important features (40 anomaly factors).** The figure shows characteristics that correspond to largest SDF coefficients and their contribution to the cumulative R^2 when added to an SDF. Panel (a) uses the polynomial kernel of the second order. Panel (b) uses the Gaussian (radial) kernel.

largest SDF coefficients are interactions, rather than base linear characteristics.

The method highlights the importance of non-linearities in characteristics when seeking for an optimal SDF, once again. In Figure 7 I show the time-series of the SDFs corresponding to the two kernels. Qualitatively the two SDFs have similar patterns; however, the SDF corresponding to the Gaussian kernel achieves much higher cross-validated Sharpe ratios.

3.2.5 Out-of-sample analysis

The analysis so far has been based purely on cross validation in a given sample. While the SDF parameters are always picked in an out-of-sample sense, the two regularization parameters κ and c are selected in sample. In addition, PC portfolios are constructed using full sample as well. I now perform a full out-of-sample evaluation of the method as follows.



(a) Polynomial kernel, 2^{nd} order

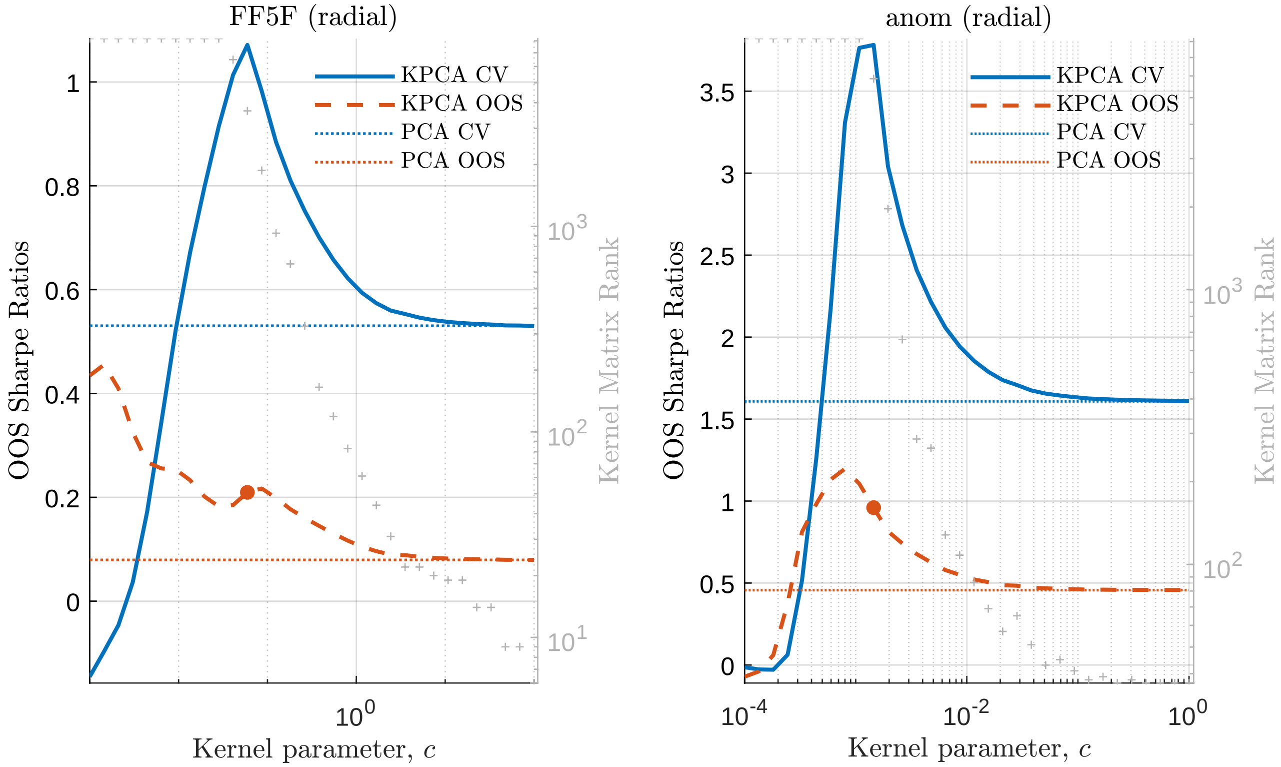
(b) Radial kernel

Figure 7: **MVE portfolio returns (40 anomaly factors).** The figure depicts empirical performance of the MVE portfolio implied by an SDF constructed using the second-order polynomial kernel (Panel a) and using the Gaussian (radial) in Panel (b).

I truncate the sample on January 1, 2005 and use the data only prior to this period to conduct the SDF estimation, which includes the construction of PC factors as well as the computation of the SDF coefficients on these factors. I use the characteristics and returns data in the post-2004 sample together with the estimates from the first half of the sample to construct the OOS SDF (or, equivalently, the MVE portfolio returns). I then empirically evaluate the performance of such an OOS MVE portfolio in the post-2004 sample.

To construct the PC portfolios I need to project returns on portfolios sorted on all features, onto corresponding eigenvectors. In turns out that this projection can be done without explicitly computing the features, analogously to the “kernel trick” idea discussed above. Appendix A provides more details on how to construct OOS PC portfolio returns.

Figure 8 depicts the maximum cross-validated Sharpe ratios in the in-sample pre-2005 period (solid) and full out-of-sample Sharpe ratios in the post-2004 period (dashed), delivered



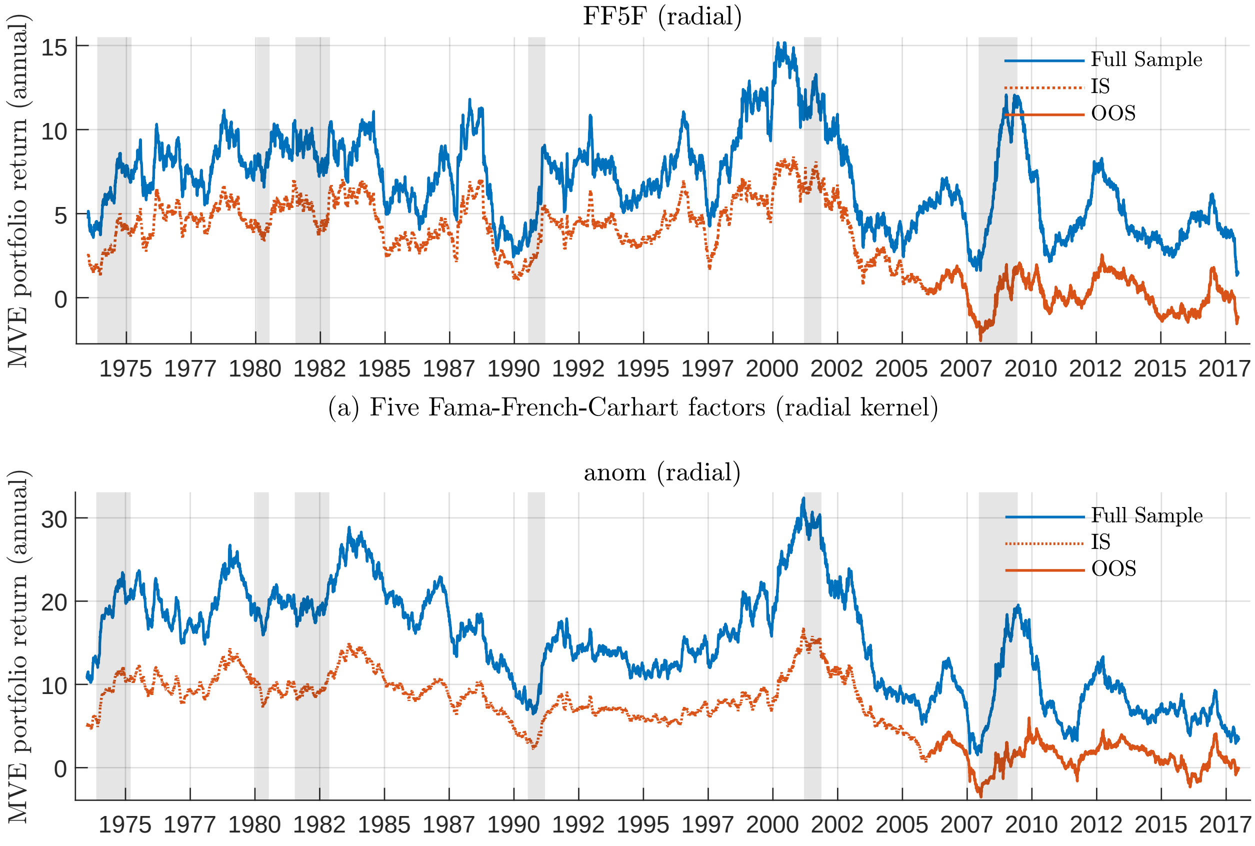
(a) Five Fama-French-Carhart factors (radial kernel)

(b) Forty anomaly factors (radial kernel)

Figure 8: **Out-of-sample Sharpe ratios.** Maximum cross-validated Sharpe ratios in the in-sample pre-2005 period (solid) and full out-of-sample Sharpe ratios in the post-2004 period (dashed), delivered by a kernel for a specific choice of a kernel parameter, denoted as c. The dotted line shows the level of the cross-validated Sharpe ratio for the linear kernel (method (ii) – PCA on characteristics-managed portfolios), which does not depend on c. Panel (a) uses the Gaussian kernel applied to five Fama-French-Carhart factors. Panel (b) applies the same kernel to forty anomaly characteristics.

by a kernel for a specific choice of a kernel parameter, denoted as c. The dotted line shows the level of the cross-validated Sharpe ratio for the linear kernel (method (ii) – PCA on characteristics-managed portfolios), which does not depend on c. Panel (a) uses the Gaussian kernel applied to five Fama-French-Carhart factors. Panel (b) applies the same kernel to forty anomaly characteristics.

The figure shows that the level of Sharpe ratios in the out-of-sample period deteriorates substantially, which is expected and is due to the overall anomaly returns deterioration in the latest part of sample (e.g., McLean and Pontiff (2016)). In spite of this, the choice of the kernel parameter c selected by cross validation in the in-sample portion of the sample generally translated well to the optimal level of c in the out-of-sample period. The plot



(a) Five Fama-French-Carhart factors (radial kernel)

(b) Forty anomaly factors (radial kernel)

Figure 9: **OOS MVE portfolio returns.** The figure depicts empirical performance of the MVE portfolio implied by an SDF constructed using the Gaussian kernel applied to five Fama-French-Carhart factors (Panel a) and forty anomaly characteristics-based factors (Panel b). The solid blue line shows the full sample estimates. Red dotted line shows the in-sample estimates. The solid red line depicts pure OOS MVE portfolio returns.

shows that for five Fama-French-Carhart characteristics using the radial kernel the out-of-sample Sharpe ratios in the latest part of the sample are around 0.2, while they are close to zero when using managed portfolios which are linear in the five base. Similarly, for the radial kernel and forty anomaly characteristics, the OOS Sharpe ratios more than doubles compared relative to the linear kernel.

Figure 9 below shows the OOS MVE portfolio returns for the two estimated SDF (solid red line) as well as their in-sample estimates (solid dotted) in the pre-2005 portion of the sample. The solid blue line shows the full-sample estimates for comparison.

3.2.6 Pricing individual stock returns

The method recovers a daily SDF and daily MVE portfolio returns. I use these estimates to compute a non-parametric estimate of conditional risk premia on individual equities. To accomplish this, I compute rolling covariance of the SDF with daily equity-level returns, and use the asset pricing equation $\mathrm{E}[MR] = 0$ to infer the implied discount rate on each stock, $\mathrm{E}_t[R] = \mathrm{cov}_t(R, -\mathrm{SDF})$. Finally, I use these estimates of discount rates to compute the predictive panel R-squared — a fraction of variation explained by my method. I compare these estimates to Kelly et al. (2018).

The time-series predictive R^2 for individual stocks is defined as follows:

$$\text{Predictive } R^2 = 1 - \frac{\mathrm{var}\left(r_{i,t} - \hat{c}_t\right)}{\sum_{i,t} r_{i,t}^2},$$

where $\hat{c}_t = \hat{\mathrm{cov}}_t(r_{i,t+1}, -\mathrm{SDF}_{t+1})$.

I compare this predictive R^2 to the benchmark which only uses information in mean returns:

$$\text{Predictive } R^2 \text{ (benchmark)} = 1 - \frac{\mathrm{var}\left(r_{i,t}\right)}{\sum_{i,t} r_{i,t}^2}.$$

I find that the firm-level conditional expected returns constructed in this way explain a significant fraction of variation in the firm-level realized returns. Monthly predictive R^2 is equal to 0.8% relative to the benchmark of only 0.2%. Daily predictive R^2 is 0.045% (benchmark: 0.01%).

3.2.7 Implied predictability of the aggregate market

I now aggregate each stock's own predictability to that of the aggregate market. That is, I construct a measure of expected returns for each stock based on covariance of stock's returns with an SDF, and then add up these fitted values using either equal or market-cap-based weights for each stock. As a result, the method uncovers strong and robust time-series predictability of the aggregate market. At a monthly horizon and daily overlapping data, returns on the aggregate equal-weighted market are predictable with an R^2 of 2.5% and a t-statistic above 3.0. Similarly, returns on the aggregate value-weighted market are also predictable with an R^2 of 1.3% and a t-statistic of 2.5.

3.2.8 Composition of the MVE portfolio in terms of individual stocks

The method allows me to recover the composition of the MVE portfolio in terms of individual stocks, or, equivalently, conditional SDF loadings (risk prices in (2)) on every stock at each point in time. These loadings can be used to construct and trade the MVE portfolio in practice, as well as compute any associated transaction costs or turnover.

[TO BE COMPLETED]

3.3 Conclusions

In this paper I argue that interactions and non-linearities in SDF loadings on characteristics are important in recovering the empirical pricing kernel. I develop a method which uses *economically*-motivated regularization and allows for arbitrary non-linearities in SDF loadings on characteristics. Relative to the linear case, such an SDF is much more efficient; the out-of-sample Sharpe ratio of the implied MVE portfolio is 3.0 (relative to 1.65 in the case of no interactions). While the method allows me to study arbitrary non-linearities and interactions in characteristics, importantly, the SDF (and the MVE portfolio) is linear in individual stock returns, that is, non-linearities appear only in variables used to sort stocks into portfolios.

The method recovers the time series of an SDF that prices equity excess returns conditionally through time, as well as conditional loadings of the SDF on every stock at each point in time. I use the SDF to infer the conditional cost of capital on any firm at any point in time non-parametrically by simply computing covariances of the firm-level realized returns with the SDF over short windows of daily data. I find that the firm-level conditional expected returns constructed in this way explain a significant fraction of variation in the firm-level realized returns. At a monthly horizon, individual firm's returns can be forecasted with an R^2 of 0.8% (relative to a benchmark of 0.2% of a constant mean). This high degree of stock-level predictability aggregates to high predictability of the aggregate market index. At a monthly horizon, the equal-weighted market portfolio is predictable with an R^2 of 2.5% and a t-statistic above 3.0. The value-weighted aggregate market portfolio can be forecasted with an R^2 of 1.3% and a t-statistic of 2.5.

At the heart of the method are the rotation of individual stock returns into a high-dimensional (potentially infinitely-dimensional) space of characteristics-based "features" portfolios and the "kernel trick" technique applied to characteristics. Such a rotation takes care of any potential variability in prices of risk and thus translates a difficult conditional problem of estimating an SDF into a simpler, though potentially much higher dimensional, *unconditional* problem. The curse of dimensionality can be circumvented, however, by using the kernel trick, which substitutes the inner product of characteristics in the PCA problem with a generalized inner product—the *kernel*. The resulting procedure is equivalent to PCA in the space of "features" – characteristics that include any non-linear functions and interactions of the original characteristics. Therefore, certain choices of the kernel, which is easy to compute, lead to the exact same solution as PCA on an extended set of portfolios sorted on original characteristics, their powers and interactions of an arbitrary (potentially infinite) order. This problem can be solved at a fixed computational cost which does not increase in the order of interactions.

References

Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang (2006). The cross-section of volatility and expected returns. *Journal of Finance 61*, 259–299.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society 68*(3), 337–404.

Asness, C. and A. Frazzini (2013). The devil in hml's details. *Journal of Portfolio Management 39*(4), 49.

Asness, C. S., A. Frazzini, and L. H. Pedersen (2014). Quality minus junk. Technical report, Copenhagen Business School.

Barbee Jr, W. C., S. Mukherji, and G. A. Raines (1996). Do sales–price and debt–equity explain stock returns better than book–market and firm size? *Financial Analysts Journal 52*(2), 56–60.

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

Barry, C. B. and S. J. Brown (1984). Differential information and the small firm effect. *Journal of Financial Economics 13*(2), 283–294.

Basu, S. (1977). Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. *The Journal of Finance 32*(3), 663–682.

Bhandari, L. C. (1988). Debt/equity ratio and expected common stock returns: Empirical evidence. *The journal of finance 43*(2), 507–528.

Blume, M. E. and F. Husic (1973). Price, beta, and exchange listing. *The Journal of Finance 28*(2), 283–299.

Brandt, M. W., P. Santa-Clara, and R. Valkanov (2009). Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns. *Review of Financial Studies 22*(9), 3411–3447.

Carhart, M. M. (1997). On persistence of mutual fund performance. *Journal of Finance 52*, 57–82.

Chen, L., R. Novy-Marx, and L. Zhang (2011). An alternative three-factor model.

Connor, G. and R. A. Korajczyk (1988). Risk and return in an equilibrium apt: Application of a new test methodology. *Journal of financial economics 21*(2), 255–289.

Cooper, M., H. Gulen, and M. Schill (2008). Asset growth and the cross-section of stock returns. *Journal of Business 63*, 1609–1652.

Da, Z., Q. Liu, and E. Schaumburg (2013). A closer look at the short-term return reversal. *Management Science 60*(3), 658–674.

Daniel, K. and S. Titman (2006). Market reactions to tangible and intangible information. *Journal of Finance 61*, 1605–1643.

Datar, V. T., N. Y. Naik, and R. Radcliffe (1998). Liquidity and stock returns: An alternative test. *Journal of Financial Markets 1*(2), 203–219.

DeBondt, W. F. and R. Thaler (1985). Does the stock market overreact? *Journal of Finance 40*, 793–805.

DeMiguel, V., L. Garlappi, F. J. Nogales, and R. Uppal (2009). A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science 55*(5), 798–812.

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

Fama, E. F. and K. R. French (1993a). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*, 23–49.

Fama, E. F. and K. R. French (1993b). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*, 23–49.

Fama, E. F. and K. R. French (1996). Mulitifactor explanations of asset pricing anomalies. *Journal of Finance 51*, 55–87.

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

Fama, E. F. and K. R. French (2016). Dissecting anomalies with a five-factor model. *The Review of Financial Studies 29*(1), 69–103.

Freyberger, J., A. Neuhierl, and M. Weber (2017). Dissecting characteristics nonparametrically. Technical report, National Bureau of Economic Research.

Gu, S., B. T. Kelly, and D. Xiu (2018). Empirical asset pricing via machine learning.

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

Haugen, R. A. and L. Baker, Nardin (1996). Commonality in the determinants of expected stock returns. *Journal of Financial Economics 41*, 401–439.

Heston, S. L. and R. Sadka (2008). Seasonality in the cross-section of stock returns. *Journal of Financial Economics 87*(2), 418–445.

Hirshleifer, D., K. Hou, S. H. Teoh, and Y. Zhang (2004). Do investors overvalue firms with bloated balance sheets. *Journal of Accounting and Economics 38*, 297–331.

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

Ikenberry, D., J. Lakonishok, and T. Vermaelen (1995). Market underreaction to open market share repurchases. *Journal of financial economics 39*(2-3), 181–208.

Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. *Journal of Finance 45*, 881–898.

Jegadeesh, N. and S. Titman (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance 48*, 65–91.

Kelly, B. T., S. Pruitt, and Y. Su (2017). Instrumented principal component analysis.

Kelly, B. T., S. Pruitt, and Y. Su (2018). Characteristics are covariances: A unified model of risk and return. Technical report, National Bureau of Economic Research.

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

Kozak, S., S. Nagel, and S. Santosh (2019). Shrinking the cross-section. *Journal of Financial Economics*, Forthcoming.

Lakonishok, J., A. Shleifer, and R. W. Vishny (1994). Contrarian investment, extrapolation and risk. *Journal of Finance 49*, 1541–1578.

Lyandres, E., L. Sun, and L. Zhang (2007). The new issues puzzle: Testing the investment-based explanation. *The Review of Financial Studies 21*(6), 2825–2855.

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

Moskowitz, T. J. and M. Grinblatt (1999). Do industries explain momentum? *The Journal of Finance 54*(4), 1249–1290.

Novy Marx, R. (2013). The Other Side of Value: The Gross Profitability Premium. *Journal of Financial Economics 108*(1), 1–28.

Piotroski, J. D. (2000). Value investing: The use of historical financial statement information to separate winners from losers. *Journal of Accounting Research*, 1–41.

Pontiff, J. and A. Woodgate (2008). Share issuance and cross-sectional returns. *Journal of Finance 63*, 921–945.

Rosenberg, B. (1974). Extra-market components of covariance in security returns. *Journal of Financial and Quantitative Analysis 9*(2), 263–274.

Schölkopf, B., C. Burges, and V. Vapnik (1996). Incorporating invariances in support vector learning machines. In *International Conference on Artificial Neural Networks*, pp. 47–52. Springer.

Schölkopf, B., A. Smola, and K.-R. Müller (1997). Kernel principal component analysis. In *International Conference on Artificial Neural Networks*, pp. 583–588. Springer.

Sloan, R. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *Accounting Review 71*, 289–315.

Soliman, M. T. (2008). The use of dupont analysis by market participants. *The Accounting Review 83*(3), 823–853.

Spiess, D. K. and J. Affleck-Graves (1999). The long-run performance of stock returns following debt offerings. *Journal of Financial Economics 54*(1), 45–73.

Xing, Y. (2008). Interpreting the value effect through the q-theory: An empirical investigation. *The Review of Financial Studies 21*(4), 1767–1795.

A Derivations

PCA in feature space of characteristics requires finding eigenvalues $\lambda \geq 0$ and nonzero eigenvectors $v \in \mathcal{H}$ of the estimated covariance matrix,

$$\Sigma = T^{-1} \sum_{t=1}^{T} \Phi(Z_t)' R_{t+1} R_{t+1}' \Phi(Z_t), \tag{27}$$

of the centered and non-linearly transformed characteristics Z_t via a transform $\Phi(Z_t)$. The eigenequation $\Sigma v = \lambda v$, where v is the eigenvector corresponding to the eigenvalue $\lambda \geq 0$ of Σ, can be written in an equivalent form as

$$\langle \Phi(Z_t)' R_{t+1}, \Sigma v \rangle = \lambda \langle \Phi(Z_t)' R_{t+1}, v \rangle , \quad t = 1, 2, ..., T. \tag{28}$$

Because

$$\Sigma v = T^{-1} \sum_{t=1}^{T} \Phi(Z_t)' R_{t+1} \langle \Phi(Z_t)' R_{t+1}, v \rangle , \tag{29}$$

all solutions v with nonzero eigenvalue λ are contained in the span of $\Phi(Z_1)'R_2$, ..., $\Phi(Z_T)'R_{T+1}$. So, there exist coefficients, $\alpha_t, t = 1, 2, ..., T$, such that

$$v = \sum_{t=1}^{T} \alpha_t \Phi(Z_t)' R_{t+1} \tag{30}$$

Substituting equations (28)–(30) into (27), we get that

$$T^{-1} \sum_{j=1}^{T} \alpha_j \left\langle \Phi(Z_t)' R_{t+1}, \sum_{k=1}^{T} \Phi(Z_k)' R_{k+1} \right\rangle \langle \Phi(Z_k)' R_{k+1}, \Phi(Z_j)' R_{j+1} \rangle = \tag{31}$$

$$= \lambda \sum_{k=1}^{T} \alpha_k \langle \Phi(Z_t)' R_{t+1}, \Phi(Z_k)' R_{k+1} \rangle, \tag{32}$$

for all $t = 1, 2, ..., T$. Define the $(T \times T)$-matrix $K = (K_{ij})$, where

$$K_{ij} = \langle \Phi(x_i)' R_{i+1}, \Phi(x_j)' R_{j+1} \rangle. \tag{33}$$

Note that K will generally be a huge matrix. Then, the eigenequation above can be written as $K^2\alpha = T\lambda K\alpha$, where $\alpha = (\alpha_1, \cdots, \alpha_T)'$, or as

$$K\alpha = \tilde{\lambda}\alpha, \tag{34}$$

where $\tilde{\lambda} = T\lambda$. Note that we can express the eigenvalues and vectors, $(\tilde{\lambda}, \alpha)$, of K in terms of those, (λ, v), for Σ.

Consider now a projection of a new (out-of-sample) datapoint, given by $\{Z_t, R_{t+1}\}$. Its *nonlinear principal component scores* corresponding to Φ are given by the projection of $\Phi(Z_t)'R_{t+1} \in \mathcal{H}$ onto the eigenvectors $v_k \in \mathcal{H}$,

$$\langle v_k, \Phi(Z_t)'R_{t+1}\rangle = \lambda_k^{-1/2}\sum_{i=1}^{T}\alpha_{k,i}\langle\Phi(Z_i)'R_{i+1}, \Phi(Z_t)'R_{t+1}\rangle, \quad k = 1, 2, ..., T, \tag{35}$$

where the $\lambda_k^{-1/2}$ term is included so that $\langle v_k, v_k\rangle = 1$. Using the kernel trick, the nonlinear principal component scores of $\Phi(Z_t)'R_{t+1}$ can be expressed as

$$\langle v_k, \Phi(Z_t)'R_{t+1}\rangle = w_{k,t}'R_{t+1}, \quad k = 1, 2, ..., \tag{36}$$

The weights $w_{k,t}$ are given by

$$w_{k,t}' = \lambda_k^{-1/2}\sum_{i=1}^{T}\alpha_{k,i}R_{i+1}'\mathcal{K}(Z_i, Z_t), \quad k, t = 1, 2, ..., T, \tag{37}$$

where the kernel matrix is defined as previously, $\mathcal{K}(Z_i, Z_t) = \Phi(Z_i)\Phi(Z_t)'$, and $\alpha_{k,i}$ is the kth eigenvector of the matrix K.

Constructing out-of-sample PCs for any kernel. Therefore, PC realizations can be easily computed for new observations without knowing the transform $\phi(\cdot)$ explicitly. It only requires applying the kernel function to the new and existing T sample datapoints, and weighting these kernels using the corresponding eigenvector. For the linear kernel, for example, this construction exactly matches the PCs constructed using classic PCA on the covariance matrix. Likewise, for the second-order polynomial kernel, recovered PCs are identical to the ones based on PCA of the covariance matrix based on an expanded set of characteristics which includes all second-order terms. For the radial kernel PCs can be recovered only using this kernelized approach, however.

Equivalence between eigenvectors of FF' and PCs of $F'F$. Consider an existing datapoint at time t and the realization of the kth principal component of features (projection onto an eigenvector): $\langle v_k, \Phi(Z_t)'R_{t+1}\rangle = \lambda_k^{-1/2}\sum_i \alpha_{ki}K_{i,t} = \lambda_k^{-1/2}(K\alpha_k)_t = \lambda_k^{-1/2}(\lambda_k\alpha_k)_t \propto \alpha_{k,t}$, where $(A)_t$ stands for the tth row of A. Therefore, the kth principal component of the covariance matrix in (27) is proportional to the kth eigenvector of the matrix K.

Recovering conditional SDF risk prices, b_t. Note that equation (36) shows how to express each PC in terms of underlying individual stock returns R_t at each point in time t. We can then use the expression for $w_{k,t}$ in (37) to compute these weights for each PC used in estimation. These weights can then be translated into SDF weights in (2) using equation (6). Therefore, full conditional projections of an SDF onto the space of individual stocks returns can be recovered at each point in time t, with no linearity-in-characteristics assumption. Equivalently, the method reveals composition of the MVE portfolio in terms of individual stock returns at any t.

B Solution method

Computing the $T \times T$ kernelized matrix Ω using daily data requires a significant amount of computations. Each element of this $12,000 \times 12,000$ matrix requires computing the kernel matrix of size $N \times N$, where N is the number of stocks. Each of the elements of the latter matrix is an inner product of all characteristics on two stocks. Lastly, this problem has to be solved multiple times to cross validate the parameter c. Overall, with daily data and 32 cross-validated values, roughly 10^{18} arithmetic operations need to be performed.

Although there is a large fixed cost to solving this problem using daily data, importantly, there is no incremental cost to allowing higher-order interactions among features. Indeed, the algorithm simply requires replacing the kernel function evaluation with a different one (e.g., raising to a higher power), which have negligible impact on the overall computation time. Similarly, expanding the set of original characteristics is relatively cheap – it increases complexity linearly. For classical PCA approaches the cost is polynomial, $O(L^3)$, where L is the number of characteristics.

I solve the problem using the C++ CUDA framework for GPU computing on an Nvidia Titan V GPU. Importantly, the problem is massively parallelizable and can be very efficiently implemented on a GPU.[11] The overall runtime for the anomaly dataset using daily data and 32 cross-validation values is less than an hour.

[11] Modern GPUs, such as Titan V, can perform around 10^{13} arithmetic operations per second.

C Variable definitions

C.1 Anomaly characteristics

Anomaly definitions and descriptions are based on the list of characteristics compiled by Kozak et al. (2019). All accounting variables are properly lagged. For annual rebalancing, returns from July of year t to June of year $t+1$ are matched to variables in December of $t-1$. Returns from January to June of year t are matched to variables in December of year $t-2$. Financial variables with a subscript "Dec" below are computed using the same timing convention. Flow variables (like dividends or investment) are annual totals as of the measurement date, unless otherwise specified. For monthly rebalancing, returns are matched to the latest quarterly report, lagged one month. Additional lagging (if required) is reported for each variable below individually. All subindices below are measured in months. A time subscript t refers to the time at which a portfolio is formed.

1. **Size** (*size*). Follows Fama and French (1993b). $\text{size} = \text{ME}_{\text{Jun}}$. The CRSP end of June price times shares outstanding. Rebalanced annually.

2. **Value (annual)** (*value*). Follows Fama and French (1993b). $\text{value} = \text{BE}/\text{ME}$. At the end of June of each year, we use book equity from the previous fiscal year and market equity from December of the previous year. Rebalanced annually.

3. **Gross Profitability** (*prof*). Follows Novy Marx (2013). $\text{prof} = \text{GP}/\text{AT}$, where GP is gross profits and AT is total assets. Rebalanced annually.

4. **Piotroski's F-score** (*F-score*). Follows Piotroski (2000). $\text{F-score} = 1_{\text{IB}>0} + 1_{\Delta\text{ROA}>0} + 1_{\text{CFO}>0} + 1_{\text{CFO}>\text{IB}} + 1_{\Delta\text{DTA}<0|\text{DLTT}=0|\text{DLTT}_{-12}=0} + 1_{\Delta\text{ATL}>0} + 1_{\text{EqIss}\leq 0} + 1_{\Delta\text{GM}>0} + 1_{\Delta\text{ATO}>0}$, where IB is income before extraordinary items, ROA is income before extraordinary items scaled by lagged total assets, CFO is cash flow from operations, DTA is total long-term debt scaled by total assets, DLTT is total long-term debt, ATL is total current assets scaled by total current liabilities, EqIss is the difference between sales of of common stock and purchases of common stock recorded on the cash flow statement, GM equals one minus the ratio of cost of goods sold and total revenues, and ATO equals total revenues, scaled by total assets. Rebalanced annualy.

5. **Debt Issuance** (*debtiss*). Follows Spiess and Affleck-Graves (1999). $\text{debtiss} = 1_{\text{DLTISS}\leq 0}$. Binary variable equal to one if long-term debt issuance indicated in statement of cash flow. Updated annually.

6. **Share Repurchases** (*repurch*). Follows Ikenberry et al. (1995). $\text{repurch} = 1_{\text{PRSTKC}>0}$. Binary variable equal to one if repurchase of common or preferred shares indicated in statement of cash flow. Updated annually.

7. **Share Issuance (annual)** (*nissa*). Follows Pontiff and Woodgate (2008). $\text{nissa} = \text{shrout}_{\text{Jun}}$ / $\text{shrout}_{\text{Jun}-12}$, where shrout is the number of shares outstanding. Change in real number of shares outstanding from past June to June of the previous year. Excludes changes in shares due to stock dividends and splits, and companies with no changes in shrout.

8. **Accruals** (*accruals*). Follows Sloan (1996). $\text{accruals} = \frac{\Delta\text{ACT}-\Delta\text{CHE}-\Delta\text{LCT}+\Delta\text{DLC}+\Delta\text{TXP}-\Delta\text{DP}}{(\text{AT}+\text{AT}_{-12})/2}$, where ΔACT is the annual change in total current assets, ΔCHE is the annual change in total cash and short-term investments, ΔLCT is the annual change in current liabilities, ΔDLC is the annual change in debt in current liabilities, ΔTXP is the annual change in income taxes payable, ΔDP is the annual change in depreciation and amortization, and $(\text{AT} + \text{AT}_{-12})/2$ is average total assets over the last two years. Rebalanced annually.

9. **Asset Growth** (*growth*). Follows Cooper et al. (2008). $\text{growth} = \text{AT}/\text{AT}_{-12}$. Rebalanced annually.

10. **Asset Turnover** (*aturnover*). Follows Soliman (2008). aturnover = SALE/AT. Sales to total assets. Rebalanced annually.

11. **Gross Margins** (*gmargins*). Follows Novy Marx (2013). gmargins = GP/SALE, where GP is gross profits and SALE is total revenues. Rebalanced annually.

12. **Earnings/Price** (*ep*). Follows Basu (1977). $\text{ep} = \text{IB}/\text{ME}_{\text{Dec}}$. Net income scaled by market value of equity. Updated annually.

13. **Cash Flow / Market Value of Equity** (*cfp*). Follows Lakonishok et al. (1994). $\text{cfp} = (\text{IB} + \text{DP})/\text{ME}_{\text{Dec}}$. Net income plus depreciation and amortization, all scaled by market value of equity measured at the same date. Updated annually.

14. **Net Operating Assets** (*noa*). Follows Hirshleifer et al. (2004). noa = (AT - CHE) - (AT - DLC - DLTT - MIB - PSTK - CEQ), where AT is total assets, CHE is cash and short-term investments, DLC is debt in current liabilities, DLTT is long term debt, MIB is non-controlling interest, PSTK is preferred capital stock, and CEQ is common equity. Updated annually.

15. **Investment** (*inv*). Follows Chen et al. (2011); Lyandres et al. (2007). inv = (ΔPPEGT + ΔINVT)/AT_{-12}, where ΔPPEGT is the annual change in gross total property, plant, and equipment, ΔINVT is the annual change in total inventories, and AT_{-12} is lagged total assets. Rebalanced annually, uses the full period.

16. **Investment-to-Capital** (*invcap*). Follows Xing (2008). invcap = CAPX/PPENT. Investment to capital is the ratio of capital expenditure (Compustat item CAPX) over property, plant, and equipment (Compustat item PPENT).

17. **Invetment Growth** (*growth*). Follows Xing (2008). growth = CAPX/CAPX_{-12}. Investment growth is the percentage change in capital expenditure (Compustat item CAPX).

18. **Sales Growth** (*sgrowth*). Follows Lakonishok et al. (1994). sgrowth = SALE/SALE_{-12}. Sales growth is the percent change in net sales over turnover (Compustat item SALE).

19. **Leverage** (*lev*). Follows Bhandari (1988). lev = AT/ME_{Dec}. Market leverage is the ratio of total assets (Compustat item AT) over the market value of equity. Both are measured in December of the same year.

20. **Return on Assets (annual)** (*roaa*). Follows Chen et al. (2011). roaa = IB/AT. Net income scaled by total assets. Updated annually.

21. **Return on Equity (annual)** (*roea*). Follows Haugen and Baker (1996). roea = IB/BE. Net income scaled by book value of equity. Updated annually.

22. **Sales-to-Price** (*sp*). Follows Barbee Jr et al. (1996). sp = SALE/ME_{Dec}. Total revenues divided by stock price. Updated annually.

23. **Momentum (6m)** (*mom*). Follows Jegadeesh and Titman (1993). mom = $\sum_{l=2}^{7} r_{t-l}$. Cumulated past performance in the previous 6 months by skipping the most recent month. Rebalanced monthly.

24. **Industry Momentum** (*indmom*). Follows Moskowitz and Grinblatt (1999). indmom = rank($\sum_{l=1}^{6} r_{t-l}^{\text{ind}}$). In each month, the Fama and French 49 industries are ranked on their value-weighted past 6-months performance. Rebalanced monthly.

25. **Momentum (1 year)** (*mom12*). Follows Jegadeesh and Titman (1993). mom12 = $\sum_{l=2}^{12} r_{t-l}$. Cumulated past performance in the previous year by skipping the most recent month. Rebalanced monthly.

26. **Momentum-Reversal** (*momrev*). Follows Jegadeesh and Titman (1993). momrev = $\sum_{l=14}^{19} r_{t-l}$. Buy and hold returns from $t-19$ to $t-14$. Updated monthly.

27. **Long-term Reversals** (*lrrev*). Follows DeBondt and Thaler (1985). lrrev = $\sum_{l=13}^{60} r_{t-l}$. Cumulative returns from $t-60$ to $t-13$. Updated monthly.

28. **Value (monthly)** (*valuem*). Follows Asness and Frazzini (2013). valuem = $\text{BEQ}_{-3}/\text{ME}_{-1}$. Book-to-market ratio using the most up-to-date prices and book equity (appropriately lagged). Rebalanced monthly.

29. **Share Issuance (monthly)** (*nissm*). Follows Pontiff and Woodgate (2008). nissm = shrout_{t-13} / shrout_{t-1}, where shrout is the number of shares outstanding. Change in real number of shares outstanding from $t-13$ to $t-1$. Excludes changes in shares due to stock dividends and splits, and companies with no changes in shrout.

30. **Return on Book Equity** (*roe*). Follows Chen et al. (2011). roe = $\text{IBQ}/\text{BEQ}_{-3}$, where IBQ is income before extraordinary items (updated quarterly), and BEQ is book value of equity. Rebalanced monthly.

31. **Return on Market Equity** (*rome*). Follows Chen et al. (2011). rome = $\text{IBQ}/\text{ME}_{-4}$, where IBQ is income before extraordinary items (updated quarterly), and ME is market value of equity. Rebalanced monthly.

32. **Short-term Reversal** (*strev*). Follows Jegadeesh (1990). strev = r_{t-1}. Return in the previous month. Updated monthly.

33. **Idiosyncratic Volatility** (*ivol*). Follows Ang et al. (2006). ivol = $\text{std}(R_{i,t} - \beta_i R_{M,t} - s_i \text{SMB}_t - h_i \text{HML}_t)$. The standard deviation of the residual from firm-level regression of daily stock returns on the daily innovations of the Fama and French three-factor model using the estimation window of three months. Lagged one month.

34. **Beta Arbitrage** (*beta*). Follows Cooper et al. (2008). beta = $\beta_{t-60:t-1}$. Beta with respect to the CRSP equal-weighted return index. Estimated over the past 60 months (minimum 36 months) using daily data and lagged one month. Updated monthly.

35. **Seasonality** (*season*). Follows Heston and Sadka (2008). season = $\sum_{l=1}^{5} r_{t-l\times 12}$. Average monthly return in the same calendar month over the last 5 years. As an example, the average return from prior Octobers is used to predict returns this October. The firm needs at least one year of data to be included in the sample. Updated monthly.

36. **Industry Relative Reversals** (*indrrev*). Follows Da et al. (2013). indrrev $= r_{-1} - r_{-1}^{\text{ind}}$, where r is the return on a stock and r^{ind} is return on its industry. Difference between a stocks' prior month's return and the prior month's return of its industry (based on the Fama and French 49 industries). Updated monthly.

37. **Composite Issuance** (*ciss*). Follows Daniel and Titman (2006). ciss $= \log(\frac{\text{ME}_{t-13}}{\text{ME}_{t-60}}) - \sum_{l=13}^{60} r_{t-l}$, where r is the log return on the stock and ME is total market equity. Updated monthly.

38. **Price** (*price*). Follows Blume and Husic (1973). price $= \log(\text{ME}/\text{shrout})$, where ME is market equity and shrout is the number of shares outstanding. Log of stock price. Updated monthly.

39. **Firm Age** (*age*). Follows Barry and Brown (1984). age $= \log(1 + \text{number of months since listing})$. The number of months that a firm has been listed in the CRSP database.

40. **Share Volume** (*shvol*). Follows Datar et al. (1998). shvol $= \frac{1}{3}\sum_{i=1}^{3} \text{volume}_{t-i}/\text{shrout}_t$. Average number of shares traded over the previous three months scaled by shares outstanding. Updated monthly.

Table 1: **Part I:** Mean annualized returns on anomaly portfolios, %

The table lists all basic "anomaly" characteristics used in my analysis and shows annualized mean returns on managed portfolios which are linear in characteristics. Columns (1)-(3) show mean annualized returns (in %) for managed portfolios corresponding to all characteristics in the full sample, pre-2005 sample, and post-2005 sample, respectively. All managed portfolios' returns are based on a monthly-rebalanced buy-and-hold strategy and are further rescaled to have standard deviations equal to the in-sample standard deviation of excess returns on the aggregate market index. The sample is daily from 11/01/1973 to 12/29/2017.

	(1)	(2)	(3)
	Full Sample	Pre 2005	Post 2005
1. Size	-5.0	-5.5	-3.7
2. Value (A)	8.1	11.2	0.3
3. Gross Profitability	1.5	1.6	1.1
4. F-score	4.8	7.8	-2.6
5. Debt Issuance	0.8	0.9	0.6
6. Share Repurchases	5.5	6.2	3.8
7. Net Issuance (A)	-6.8	-9.6	0.2
8. Accruals	-6.9	-9.2	-1.0
9. Asset Growth	-7.8	-10.7	-0.6
10. Asset Turnover	2.6	2.9	1.8
11. Gross Margins	-1.0	-1.1	-0.6
12. Earnings/Price	5.4	7.6	-0.2
13. Cash Flows/Price	8.8	11.9	1.1
14. Net Operating Assets	-9.3	-9.8	-7.9
15. Investment/Assets	-9.3	-11.2	-4.3
16. Investment/Capital	-3.7	-4.0	-3.2
17. Investment Growth	-6.5	-8.5	-1.4
18. Sales Growth	-4.8	-5.5	-3.0
19. Leverage	8.4	10.2	3.6
20. Return on Assets (A)	-7.9	-9.1	-4.9
21. Return on Book Equity (A)	-4.9	-6.4	-1.2
22. Sales/Price	9.6	12.0	3.6
23. Momentum (6m)	0.7	3.4	-6.1
24. Industry Momentum	5.5	8.4	-1.9
25. Momentum (12m)	5.4	9.5	-4.9
26. Momentum-Reversals	-9.3	-11.3	-4.2
27. Long Run Reversals	-8.8	-11.1	-2.9
28. Value (M)	9.9	12.3	3.9
29. Net Issuance (M)	-8.1	-10.1	-3.1
30. Return on Equity	8.2	11.9	-1.2

continued on next page...

Table 1: **Part II:** Mean annualized returns on anomaly portfolios, %

	(1)	(2)	(3)
31. Return on Market Equity	16.7	23.3	0.2
32. Short-Term Reversals	-14.5	-17.9	-6.0
33. Idiosyncratic Volatility	0.8	0.2	2.1
34. Beta Arbitrage	0.6	0.5	0.8
35. Seasonality	14.0	22.0	-6.1
36. Industry Rel. Reversals	-27.7	-35.5	-8.0
37. Composite Issuance	-7.8	-10.1	-1.8
38. Price	-11.5	-11.0	-12.6
39. Age	-1.2	-1.5	-0.5
40. Share Volume	-0.3	0.5	-2.4