The Kernel Trick for Nonlinear Factor Modeling

Varlam Kutateladze*

August 5, 2020

Abstract

Factor modeling is a powerful statistical technique that permits to capture the common dynamics in a large panel of data with a few latent variables, or factors, thus alleviating the curse of dimensionality. Despite its popularity and widespread use for various applications ranging from genomics to finance, this methodology has predominantly remained linear. This study estimates factors nonlinearly through the kernel method, which allows flexible nonlinearities while still avoiding the curse of dimensionality. We focus on factor-augmented forecasting of a single time series in a high-dimensional setting, known as diffusion index forecasting in macroeconomics literature. Our main contribution is twofold. First, we show that the proposed estimator is consistent and it nests linear PCA estimator as well as some nonlinear estimators introduced in the literature as specific examples. Second, our empirical application to a classical macroeconomic dataset demonstrates that this approach can offer substantial advantages over mainstream methods.

JEL Classification: C38, C53, C45
Keywords: Macroeconomic forecasting; Latent factor model; Nonlinear time series; Principal component analysis; kernel PCA; Neural networks; Econometric models

*Correspondence to: Department of Economics, University of California, Riverside, CA 92521, USA.
E-mail: varlam.kutateladze@email.ucr.edu.
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.
Declarations of interest: none.

1 Introduction

Over the past century, factor models have become an integral part of multivariate analysis and high-dimensional statistics, and have had a substantial effect on a number of different fields, including psychology (Thompson [1938]), biology (Hirzel et al. [2002]) and economics (Chamberlain and Rothschild [1983]). In economics and finance, applications range from portfolio optimization (Fama and French [1992]) and covariance estimation (Fan et al. [2013]) to forecasting (Stock and Watson [2002a]). The application that we consider is macroeconomic forecasting, where various forms of factor analysis have become state-of-the-art techniques for prediction.

The general idea behind factor analysis consists of determining a few latent variables, or factors, that drive the dependence of the entire outcomes. Factors are designed to capture the common dynamics in a large panel of data. This feature is crucial in the context of increasing availability of macroeconomic time series coupled with the inability of standard econometric methods to handle many variables. While classical econometric tools break down in such data-rich or "big data" environments, factor models help to compress a large amount of the available information into a few factors, turning the curse of dimensionality into a blessing.

Factor analysis possesses several attractive properties that justify a large amount of literature in its support. First, it effectively handles large dimensions thereby enhancing forecast accuracy in such regimes. This was demonstrated in Stock and Watson [2002a] and Stock and Watson [2002b] who used so-called diffusion indexes, or factors, in forecasting models when dealing with a large number of predictors. More recently, Kim and Swanson [2018] find that factor augmented models nearly always outperform a wide range of big data and machine learning models in terms of predictive power. Second, due to its conceptual simplicity, this methodology found its use beyond academic research. For example, the Federal Reserve Bank of Chicago constructs the Chicago Fed National Activity Index (CFNAI) simply as the first principal component of a large number of time series. Third, factor analysis aligns naturally with the dynamic equilibrium theories as well as the stylized fact of Sargent and Sims [1977] of a small number of variables explaining most of the fluctuations in macroeconomic time series. And finally, factor estimates can also be used to provide efficient instruments for augmenting vector autoregressions (VARs) (Bernanke et al. [2005]) to assist in tracing structural shocks.

Factors, however, are not observable and need to be estimated. Two classical estima-

tion strategies rely on either intertemporal or contemporaneous smoothing. The former casts the model into a state-space representation and estimates it by the maximum likelihood via the Kalman filter. The disadvantages of this approach are that it requires parametric assumptions and that it quickly becomes computationally infeasible as the number of predictor series grows [1]. Contemporaneous smoothing is a more predominant and computationally simpler way based on principal component analysis (PCA) (Pearson [1901]), nonparametric least-squares approach for estimating factors.

Forecasts are obtained via a two-step procedure. First, factors estimates are derived from the set of available time series by one of the two methods described above. Once the factors are estimated, run a linear autoregression of the variable of interest onto factor estimates and observed covariates (e.g. lagged values of the dependent variable).

We analyze a factor model that is high-dimensional, static and approximate. High-dimensional framework (Bai and Ng [2002]), as opposed to classical framework (Anderson [1984]), allows both time and cross-section dimensions to grow. Static models do not explicitly model time-dependence of factors contrary to more general dynamic counterparts (Forni et al. [2000]). Approximate factor structure (Chamberlain and Rothschild [1983]) is more flexible compared with a strict version (Ross [1976]) as it imposes milder assumptions on the idiosyncratic component.

Factor analysis is closely related to PCA, although the two are not the same (Jolliffe [1986]). It is, however, well documented that the two are asymptotically equivalent under suitable conditions (see the pervasiveness assumption in Fan et al. [2013] for a recent treatment). There are several results on consistency of PCA estimators of factors (Connor and Korajczyk [1986], Stock and Watson [2002a], Bai and Ng [2006] among others) for various forms of factor models. One of the most relevant of the results is established in Bai and Ng [2002] who derive convergence rates of such estimators for an approximate static factor model of large dimensions.

Despite its widespread use, factor modeling, and diffusion index forecasting methodology in particular, is still fundamentally limited to linear framework. Over the past two decades, leading researchers noted multiple times (e.g. see Stock and Watson [2002b], Bai and Ng [2008], Stock and Watson [2012], Cheng and Hansen [2015]) that further forecast improvements "will need to come from models with nonlinearities and/or time variation" and that "nonlinear factor-augmented regression should be considered" for forecasting.

[1] Interestingly, there has been some evidence to the contrary, see Doz et al. [2012]

While there have been multiple attempts to incorporate time dependence (see, for example, Negro and Otrok [2008], Mikkelsen et al. [2015] and Coulombe et al. [2019]), the literature on addressing nonlinearity is scarce. Yet, nonlinear time series models typically dominate their linear counterparts (Teräsvirta et al. [1994], Giovannetti [2013], Kim and Swanson [2014]). One of the first attempts to address nonlinear structure is Yalcin and Amemiya [2001] who assume errors-in-variables parametrization and have no forecasting application. The most prominent work with focus on prediction exercise is Bai and Ng [2008]. They either augment the set of predictor time series with their squares and apply standard PC to the augmented set, or use squares of principal components obtained from the original (non-augmented) set. Another closely related work is Exterkate et al. [2016] who substitute the linear second step with kernel ridge regression and discover that this leads to more accurate forecasts of the key economic indicators.

This study adds to the scarce literature on nonlinear factor models. Specifically, the factors are allowed to capture nontrivial functions of predictors. To circumvent the computational difficulties associated with such novelties, we use the kernel trick, or kernel method (Hofmann et al. [2008]), which is discussed in the next section within the diffusion index methodology context.

The rest of the paper is organized as follows. Section 2 reviews the methodology, discusses the kernel trick, kernel PCA and provides the theoretical guarantees. Section 3 outlines the forecasting models, describes the data and provides the empirical results. Section 4 concludes and discusses possible extensions. All proofs are given in the Appendix.

Notation. For a vector $\mathrm{v} \in \mathbb{R}^d$, we write its i-th element as v_i. The corresponding ℓ_p norm is $\|\mathrm{v}\|_p = \left(\sum_{i=1}^d |v_i|^p\right)^{1/p}$, which is a norm for $1 \leq p \leq \infty$. For a matrix $A \in \mathbb{R}^{m\times d}$, we write its (i,j)-th entry as $\{A\}_{ij} = a_{ij}$ and denote its i-th row (transposed) and j-th column as column vectors $A_{i\cdot}$ and $A_{\cdot j}$ respectively. Its singular values are $\sigma_1(A) \geq \sigma_2(A) \geq \ldots \geq \sigma_q(A)$, where $q = \min(m,d)$. The spectral norm is $\|A\|_2 = \max_{\mathrm{v}\neq 0} \frac{\|A\mathrm{v}\|_2}{\|\mathrm{v}\|_2} = \sigma_1(A)$. The ℓ_1 norm is $\|A\|_1 = \max_{1\leq j\leq d} \sum_{i=1}^m |u_{ij}|$ and ℓ_∞ norm is $\|A\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^d |u_{ij}|$. The Frobenius norm is $\|A\|_F = \sqrt{\langle A, A\rangle} = \sqrt{tr(A'A)} = \sqrt{\sum_{i=1}^q \sigma_i^2(A)}$. For a symmetric matrix $W \in \mathbb{R}^{d\times d}$ with eigenvalues $\lambda_1(W) \geq \lambda_2(W) \geq \ldots \geq \lambda_d(W)$, define $eig_r(W) \in \mathbb{R}^{d\times r}$ to be a matrix stacking $r \leq d$ normalized eigenvectors in the order corresponding to $\lambda_1(W), \ldots, \lambda_r(W)$. Finally, for a sequence of random variables $\{X_n\}_{n=1}^\infty$ and a sequence of real nonnegative numbers $\{a_n\}_{n=1}^\infty$, denote $X_n = O_{\mathbb{P}}(a_n)$ if

$\forall \epsilon > 0, \exists M, N > 0$ such that $\forall n > N$, $\mathbb{P}(|X_n/a_n| \geq M) < \epsilon$; and denote $X_n = o_{\mathbb{P}}(a_n)$ if $\forall \epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|X_n/a_n| \geq \epsilon) = 0$. Finally, let $\mathbf{1}_{1/T}$ be a $T \times T$ matrix of ones divided by T.

2 Methodology

2.1 Diffusion Index Models

Our goal is to accurately forecast a scalar variable Y_t, given a $T \times N$ data matrix X with tth row X_t', or $X_{t\cdot}'$. Both the number of observations T and the number of series N are typically large.

Consider the following baseline model, known as a Diffusion Index (DI) model:

$$\underset{1\times 1}{Y_{t+h}} = \underset{1\times r}{\beta_F'} \underset{r\times 1}{F_t} + \underset{1\times p}{\beta_W'} \underset{p\times 1}{W_t} + \underset{1\times 1}{\epsilon_{t+h}}, \tag{1}$$

$$\underset{N\times 1}{X_t} = \underset{N\times r}{\Lambda} \underset{r\times 1}{F_t} + \underset{N\times 1}{e_t}\,. \tag{2}$$

Equation 1 is a linear forecasting model, where Y_{t+h} is the value of the target variable h periods in the future, F_t is the vector of r factors at time t, W_t is a vector of p observed covariates (e.g. an intercept and lags of Y_{t+h}), ϵ_{t+h} is a disturbance term. Equation 2 specifies the factor model, where X_t is vector of N candidate predictor series, Λ is a loading matrix for r common driving forces in F_t, e_t is an idiosyncratic disturbance; and $t = 1, \ldots, T$. The latter equation can be rewritten in matrix form

$$\underset{T\times N}{X} = \underset{T\times r}{F} \underset{r\times N}{\Lambda'} + \underset{T\times N}{e}, \tag{3}$$

where $X = [X_1, \ldots, X_T]'$ and $F = [F_1, \ldots, F_T]'$. Throughout the paper it is assumed that all series are weakly stationary and variables in X have been standardized.

If the above set of equations is augmented with transition equations for F_t, we obtain a dynamic factor model which is estimated by the Kalman filter as discussed above. Let us instead focus on a nonparametric estimation approach as suggested in Stock and Watson [2002a]. The goal at first stage is to solve

$$\begin{aligned} &\arg\min_{F,\Lambda} \ \|X - F\Lambda'\|_F^2 \\ &N^{-1}\Lambda'\Lambda = I_r, \quad F'F \text{ diagonal}, \end{aligned} \tag{4}$$

where the restrictions are in place for identifying the unique solution (up to a column sign change). It is well known that the estimator of factor loadings $\widehat{\Lambda}$ is given by the r eigenvectors associated with largest eigenvalues of $X'X$, while $\widehat{F} = X\widehat{\Lambda}$. This estimator $\widehat{F}$ is equivalent to principal component (PC) scores derived from the matrix X. Once we have an estimate of F, the second stage involves least squares estimation of equation 1 with F substituted with its estimate.

It is clear that the standard PC estimator reduces the dimensionality of X linearly: F_t represents the projection of X_t onto r eigenvector directions exhibiting the most variation. However, if there is a nonlinearity in X, that is if the true lower dimensional representation is a nonlinear submanifold in the original space, such linear projections will be inaccurate. There are several ways to take into account a possible nonlinearity. For example, Bai and Ng [2008] propose a squared principal components (SPCA) procedure, which applies the standard PCA algorithm to the matrix X augmented by its square, that is $[X, X^2]$. Although this procedure supposedly leads to additional forecasting gains, it is limited by the second-order features of the data.

Other nonlinear dimension reduction techniques include Laplacian eigenmaps, local linear embeddings (LLE), isomaps and a number of others. In this paper we use the approach that applies the kernel trick to the standard PCA, so-called kernel PCA (kPCA) (Scholkopf et al. [1999]). This algorithm can be shown to contain a number of widely used dimensionality reduction methods, including the ones listed above (Hofmann et al. [2008]). While it permits modeling a set of nonlinearities rich enough for successful applications in nontrivial pattern recognition tasks such as face recognition (Kim et al. [2002]), the algorithm does not involve any iterative optimization.

2.2 Kernel Method

The kernel method implicitly maps the original data nonlinearly into a high-dimensional space, known as a feature space, $\varphi(\cdot) : \mathcal{X} \to \mathcal{F}$. This space can in fact be infinite-dimensional which would seemingly prohibit any calculations. However, the trick is precisely in avoiding such calculations. The focus is instead on similarities between any two transformed data points $\varphi(\mathrm{x_i})$ and $\varphi(\mathrm{x_j})$ in the feature space[2] as measured by $\varphi(\mathrm{x_i})'\varphi(\mathrm{x_j})$, calculating which would at first sight require the knowledge of the functional form of

[2]Formally, the feature space is thus a Hilbert space, that is a vector space with a dot product defined on it.

$\varphi(\cdot)$. The solution is to use a kernel function $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which would output the inner product in the feature space without ever requiring the explicit functional form of $\varphi(\cdot)$. Moreover, a valid kernel function guarantees the existence of a feature mapping $\varphi(\cdot)$ although its analytic form may be unknown. The only requirement for this is positive-definiteness of the kernel function (Mercer's condition, see A.9), specifically, $\int\int f(\mathrm{x_i})k(\mathrm{x_i},\mathrm{x_j})f(\mathrm{x_j})d\mathrm{x_i}d\mathrm{x_j} \geq 0$, for any square-integrable function $f(\cdot)$.

This kernel function forms a Gram matrix K, known as a Kernel matrix, elements of which are inner products between transformed training examples, that is $\{K\}_{ij} = k(\mathrm{x_i},\mathrm{x_j}) = \varphi(\mathrm{x_i})'\varphi(\mathrm{x_j})$. What makes the kernel trick useful is the fact that many models can be written exclusively in terms of dot products between data points. For example, the ridge regression coefficient estimator can be formulated as $X'(XX' + \lambda I_T)^{-1}Y$[3], and hence the prediction for a test example x_* is $\hat{Y} = \mathrm{x}_*'X'(XX' + \lambda I_T)^{-1}Y$, where both $\{\mathrm{x}_*'X'\}_i = \mathrm{x}_*'X_{i\cdot}$ and $\{XX'\}_{ij} = X_{i\cdot}'X_{j\cdot}$ depend exclusively on inner products between the covariates. This property allows us to apply the kernel method by substituting dot products between original variables with their nonlinear kernel evaluations, that is dot products between transformed variables. Hence, an alternative form is $k_*'(K + \lambda I_T)^{-1}Y$, where k_* with $\{k_*\}_i = \{\varphi(\mathrm{x}_*)'\varphi(X)'\}_i$ is a vector of similarities between the test example and training examples in the feature space. In terms of the time complexity, the algorithm needs to invert a $T \times T$ matrix instead of inverting an $N \times N$ matrix.

The key advantage of the kernel method is that it effectively permits using a linear model in a high-dimensional nonlinear space, which amounts to applying a nonlinear technique in the original space. As a toy example, consider a classification problem shown in Figure 1, where the true function separating the two classes is a circle of radius .5 around the origin. The left panel depicts observations from two classes which are not linearly separable in the original two-dimensional space. Applying a simple polynomial kernel of degree 2, $k(\mathrm{x_i},\mathrm{x_j}) = (\mathrm{x_i}'\mathrm{x_j})^2$, implicitly corresponds to working in the feature space depicted on the right panel, since for $\varphi(\mathrm{x_i}) = (x_{i1}^2,\ \sqrt{2}x_{i1}x_{i2},\ x_{i2}^2)'$ we have $\varphi(\mathrm{x_i})'\varphi(\mathrm{x_j}) = (\mathrm{x_i}'\mathrm{x_j})^2$. In this toy example, a linear classifier could perfectly separate the observations in the right panel of Figure 1.

While all valid kernel functions are guaranteed to have the corresponding feature space, in many cases it is implicit and infinite-dimensional, as, for instance, for the

[3]This can be derived by solving the dual of the ridge least squares optimization problem, however a simpler approach would be to apply the matrix identity $(B'C^{-1}B + A^{-1})^{-1}B'C^{-1} = AB'(BAB' + C)^{-1}$ to the usual ridge estimator $(X'X + \lambda I_N)^{-1}X'Y$.

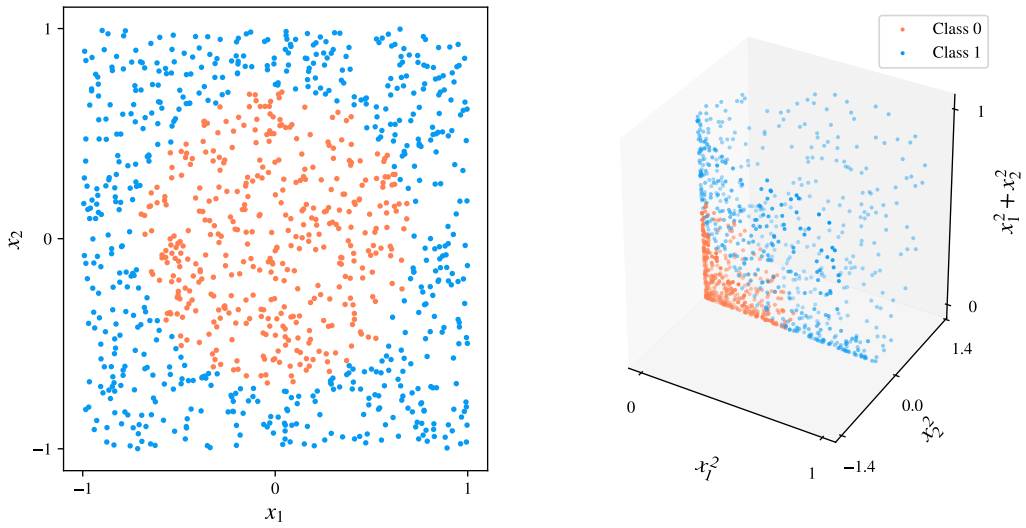

Figure 1: Kernel trick illustration for a toy classification example. Left: observations from two classes in the original space, not linearly separable. Right: observations in the feature space, linearly separable. See the details in the text.

radial basis function (RBF) kernel $k(\mathrm{x_i}, \mathrm{x_j}) = e^{-\gamma \left\| \mathrm{x_i} - \mathrm{x_j} \right\|_2^2}$ (see A.2). Again, luckily, the knowledge of the feature mapping is not required.

2.3 Nonlinear Modeling and kPCA

Suppose there is a nonlinear function $\varphi(\cdot) : \mathbb{R}^N \to \mathbb{R}^M$, where $M \gg N$ is very large (often infinitely large), mapping each observation to a high-dimensional feature space, $X_t \to \varphi(X_t)$. For now we consider M to be finite for simplicity of exposition, so the original $T \times N$ data matrix X can be represented as a $T \times M$ matrix $\Phi = [\varphi(X_1), \ldots, \varphi(X_T)]'$ in the transformed space, which may not be observable. Infinite-dimensional case induces several complications and is considered later.

Both the original X and its transformation Φ are assumed to be demeaned. The latter requirement is simple to incorporate in the kernel matrix despite the mapping being unobserved. Specifically, supposing the original (non-demeaned) transformation is $\widetilde{\Phi}$, the kernel associated with demeaned features is

$$K = (I_T - \mathbf{1}_{1/T})\widetilde{\Phi}\widetilde{\Phi}'(I_T - \mathbf{1}_{1/T})' = \widetilde{K} - \mathbf{1}_{1/T}\widetilde{K} - \widetilde{K}\mathbf{1}_{1/T} + \mathbf{1}_{1/T}\widetilde{K}\mathbf{1}_{1/T}, \tag{5}$$

where $\widetilde{K} = \widetilde{\Phi}\widetilde{\Phi}'$ is based on the original $\widetilde{\Phi}$.

Our modeling of nonlinearity is through the feature mapping $\varphi(\cdot)$. This function

replaces the original variables of interest in equation (2) with their transformations,

$$\underset{M\times 1}{\varphi(X_t)} = \underset{M\times r}{\Lambda_\varphi}\ \underset{r\times 1}{F_{\varphi,t}} + \underset{M\times 1}{e_{\varphi,t}}\,, \tag{6}$$

where the subscript φ indicates the association with the transformation. By stacking these into a $T \times M$ matrix Φ we can rewrite the minimization problem (4) as

$$\begin{aligned} &\underset{F_\varphi,\Lambda_\varphi}{\arg\min}\ \left\|\Phi - F_\varphi\Lambda_\varphi'\right\|_F^2 \\ &N^{-1}\Lambda_\varphi'\Lambda_\varphi = I_r, \quad F_\varphi'F_\varphi \text{ diagonal.} \end{aligned} \tag{7}$$

Note that solving this directly through the eigendecomposition of $\Phi'\Phi$ is generally infeasible, since $\Phi'\Phi$ is $M \times M$ dimensional. Even if the dimension M was not prohibitive, the map $\varphi(\cdot)$ is unknown for interesting problems rendering any computation dependent on Φ or $\Phi'\Phi$ alone impossible. Fortunately, it is possible to reformulate this problem in terms of the $T \times T$ Gram matrix $K = \Phi\Phi'$.

While we are assuming M to be prohibitively large but finite, the following decomposition generalizes to infinite dimensions. Starting from the "infeasible" eigendecomposition of the unknown covariance matrix of Φ

$$\frac{\Phi'\Phi}{T}V_\varphi^{[i]} = \lambda_i^c V_\varphi^{[i]}, \qquad i = 1, \ldots, M, \tag{8}$$

where the eigenvalues $\lambda_i^c = \lambda_i(\frac{\Phi'\Phi}{T})$ satisfy $\lambda_1^c \geq \lambda_2^c \geq \ldots \geq \lambda_T^c$ and $\lambda_j^c = 0$ for $j > T$ (assuming $M \geq T$) and $V^{[i]}$ is an M-dimensional eigenvector associated with the ith eigenvalue λ_i^c.

The key is to observe that each $V_\varphi^{[i]}$ can be expressed as a linear combination of features

$$V_\varphi^{[i]} = \frac{\Phi'\Phi}{\lambda_i^c T}V_\varphi^{[i]} \equiv \Phi' A^{[i]}, \qquad i = 1, \ldots, M, \tag{9}$$

where $A^{[i]} = \frac{\Phi V_\varphi^{[i]}}{\lambda_i^c T} = \left[\alpha_1^{[i]}, \ \ldots\ , \alpha_T^{[i]}\right]'$ is a vector of weights which is determined next. Plugging this back into (8) yields

$$\lambda_i^c \Phi' A^{[i]} = \frac{\Phi'\Phi}{T}\Phi' A^{[i]}, \qquad i = 1, \ldots, M. \tag{10}$$

Finally, premultiplying equation (10) to the left by Φ and removing $K = \Phi\Phi'$ from both

sides we obtain

$$\frac{K}{T}A^{[i]} = \lambda_i^c A^{[i]}, \qquad i = 1, \dots, M, \tag{11}$$

hence the ith vector of weights $A^{[i]}$ corresponds to an eigenvector of a finite-dimensional Gram matrix K associated with the ith largest eigenvalue $\lambda_i(\frac{K}{T}) = \lambda_i^c$ with $\lambda_j(\frac{K}{T}) = 0$ for $j > T$.

Notice that while solving the eigenvalue problem of $\frac{K}{T}$ allows to compute $A^{[i]}$, we are still unable to obtain the vector $V^{[i]} = \Phi' A^{[i]}$ since Φ may not be known. However, the main object of interest is recoverable: to calculate principal component projections, we project the data onto (unknown) eigenspace,

$$\underset{T\times 1}{\widehat{F}_{\varphi}^{[i]}} = \Phi V^{[i]} = \Phi\Phi' A^{[i]} = KA^{[i]}, \qquad i = 1, \dots, M. \tag{12}$$

Stacking estimated factors corresponding to the first r eigenvalues, define a $T \times r$ matrix $\widehat{F}_{\varphi} = \left[\widehat{F}_{\varphi}^{[1]}, \dots, \widehat{F}_{\varphi}^{[r]}\right]$, where the subindex r is dropped to simplify the notation. We refer to the factors constructed this way as *kernel factors*.

A similar alternative solution that only involves the Gram matrix has been known in econometrics since at least Connor and Korajczyk [1993]. In particular, for a given r the optimization problem in (7) with the identification constraints $T^{-1}F_{\varphi}'F_{\varphi} = I_r$ and diagonal $\Lambda_{\varphi}'\Lambda_{\varphi}$, has the solution $\widetilde{F}_{\varphi} = \sqrt{T}eig_r(\Phi\Phi') = \sqrt{T}A_r$, where $A_r = \left[\widehat{A}^{[1]}, \dots, \widehat{A}^{[r]}\right]$. Hence, the kPCA estimator is equivalent to the latter premultiplied by $\frac{\Phi\Phi'}{\sqrt{T}} = \frac{K}{\sqrt{T}}$. Note, however, that both estimators yield the same predictions when passed to the main forecasting equation (1) as they have identical column spaces. This idea is summarized in the following proposition.

Proposition 2.1. *Estimators $\widehat{F}_{\varphi} = \Phi\Phi' eig_r(\Phi\Phi')$ and $\widetilde{F}_{\varphi} = \sqrt{T}eig_r(\Phi\Phi')$ produce the same projection matrix.*

Importantly, we also establish that certain commonly used kernels allow the kernel factor estimator to incorporate the usual PC estimator. The following proposition demonstrates that RBF and sigmoid kernels allow to nest (a constant multiple of) the PC estimator for limiting values of the hyperparameter.

Proposition 2.2. *For a column-centered matrix $X \in \mathbb{R}^{T\times N}$, let $\widehat{F}_{\varphi} = Keig_r(K)$ be the kernel factor estimator and $\widehat{F} = Xeig_r(X'X)$ be the usual linear PCA factor estimator. Then*

$$\exists s = \pm 1, \text{ such that } \lim_{\gamma \to 0} c\gamma^{-1} \widehat{F}_\varphi L^{-1/2} = s\widehat{F}, \quad \forall r = 1, \ldots, \min\{T, N\},$$

where $K = \widetilde{K} - \mathbf{1}_{1/T}\widetilde{K} - \widetilde{K}\mathbf{1}_{1/T} + \mathbf{1}_{1/T}\widetilde{K}\mathbf{1}_{1/T}$*,* L *is a diagonal matrix of* r *largest eigenvalues of* XX' *sorted in nonincreasing order. Furthermore,*
(a) *for RBF kernel* $\widetilde{k}_{ij} = e^{-\gamma \|X_{i\cdot} - X_{j\cdot}\|_2^2}$ *we have* $c = 2^{-1}$*,*
(b) *for sigmoid kernel* $\widetilde{k}_{ij} = \tanh(c_0 + \gamma X_{i\cdot}' X_{j\cdot})$ *we have* $c = (1 - \tanh^2(c_0))^{-1}$*, where* c_0 *is an arbitrary (hyperparameter) constant.*

Proof. See appendix A.3. □

Proposition 2.2 states that the (properly scaled) kernel factor matrix converges pointwise to its PCA analog (up to a sign flip) as the value of the hyperparameter γ nears zero. That is, in the limit the two factor estimators are constant multiples of one another and hence produce the same forecasts. This ability to mimic the linear estimator is important for certain applications. For example, in macroeconomic forecasting it is notoriously difficult to beat linear models in a short horizon prediction exercise.

Figure 2 illustrates the implicit procedure for obtaining r kernel factors. The selected kernel function induces nonlinearity $\varphi(\cdot)$ on each element of the input layer, N-dimensional observations $X_{1\cdot}, \ldots, X_{T\cdot}$. Next, pairwise similarities between high-dimensional vectors $\varphi(X_{1\cdot}), \ldots, \varphi(X_{T\cdot})$ are computed, with $k(X_{i\cdot}, \cdot) = \Big[\varphi(X_{i\cdot})'\varphi(X_{1\cdot}),\ \ldots\ , \varphi(X_{i\cdot})'\varphi(X_{T\cdot})\Big]'$. Finally, each factor is obtained as a linear combination of these inner products with the weights given as a solution to the eigenvalue problem discussed above. Of course, the kernel PCA algorithm does not explicitly nonlinearize the inputs as the kernel trick allows us to immediately calculate similarities and avoid the expensive high-dimensional computation. However, as mentioned earlier, the existence of such implicit nonlinearities is guaranteed by Mercer's theorem (A.9). The full algorithm is presented in Appendix A.1. As opposed to standard feedforward neural networks, there is no iterative training involved and no peril of being trapped in local optima. On the other hand, it has been noted that certain kernels, including RBF and sigmoid, allow extracting features of the same type as the ones extracted by neural networks (Scholkopf et al. [1999]). The two necessary steps involve evaluation of similarities in the kernel matrix and solving its eigenvalue problem. The complexity is thus dependent only on the sample size.

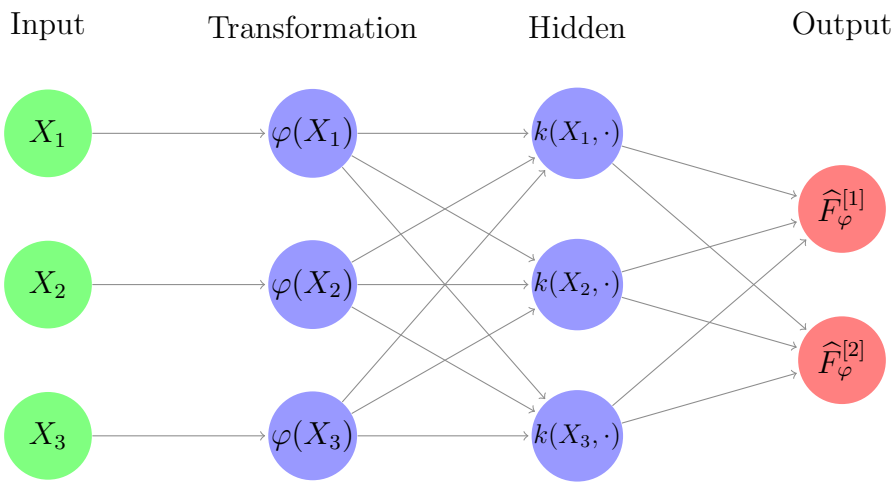

Figure 2: Neural network interpretation of kernel PCA, illustrated for the case $T = 3$, $r = 2$. Each observation is nonlinearly transformed and the inner products are computed. The output units are kernel factors with $\widehat{F}_\varphi^{[j]} = \sum_{i=1}^{T} \alpha_i^{[j]} k(X_i, \cdot)$ which linearly combine these dot products, with the weight estimate calculated as the eigenvector of the kernel matrix K.

2.4 Theory

Depending on the choice of the kernel, the induced feature space could be either finite or infinite. A polynomial kernel considered earlier generates a finite-dimensional feature space, consisting of a set of polynomial functions over the inputs. In this simple case, the eigenspace associated with the kernel factor estimator in equation (12) can generally be consistently estimated within the framework of Bai [2003]. Particularly, proposition 2.1 and the following theorem in Bai [2003] immediately imply $\sqrt{M}$-consistency of $\widetilde{F}_\varphi$.

For the model associated with equation (6):

Assumption A: There exists a constant $c_1 < \infty$ independent of M and T, such that

(a) $\mathbb{E}\,\|F_{\varphi,t}\|_F^4 \leq c_1$ and $T^{-1}F_\varphi' F_\varphi \xrightarrow{p} \Sigma_F > 0$, where Σ_F is a non-random positive definite matrix;

(b) $\mathbb{E}\,\|\Lambda_{\varphi,i\cdot}\|_F \leq c_1$ and $N^{-1}\Lambda_\varphi'\Lambda_\varphi \xrightarrow{p} \Sigma_\Lambda > 0$, where Σ_Λ is a non-random positive definite matrix.

(c1) $\mathbb{E}(e_{\varphi,it}) = 0$, $\mathbb{E}\,|e_{\varphi,it}|^8 \leq c_1$;

(c2) $\mathbb{E}(e_{\varphi,s}' e_{\varphi,t}/M) = \gamma_M(s,t)$, $|\gamma_M(s,s)| \leq c_1\ \forall s$, $T^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}|\gamma_M(s,t)| \leq c_1$, $\sum_{s=1}^{T}\gamma_M(s,t)^2 \leq M\ \forall t, T$;

(c3) $\mathbb{E}(e_{\varphi,it}e_{\varphi,jt}) = \tau_{ij,t}$ with $|\tau_{ij,t}| \leq |\tau_{ij}|$ for some τ_{ij} and $\forall t$; and $M^{-1}\sum_{i=1}^{M}\sum_{j=1}^{M}|\tau_{ij}| \leq c_1$;

(c4) $\mathbb{E}(e_{\varphi,it}e_{\varphi,js}) = \tau_{ij,ts}$, $(MT)^{-1}\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{t=1}^{T}\sum_{s=1}^{T}|\tau_{ij,ts}| \leq c_1$;

(c5) $\mathbb{E}\left|M^{-1/2}\sum_{i=1}^{M}\left(e_{\varphi,is}e_{\varphi,it}-\mathbb{E}(e_{\varphi,is}e_{\varphi,it})\right)\right|^4 \leq c_1\ \forall t,s$;
(d) $\mathbb{E}\left(\frac{1}{M}\sum_{i=1}^{M}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}F_{\varphi,t}e_{\varphi,it}\right\|_F^2 \leq c_1\right)$;
(e) $T^{-1}F_\varphi'F_\varphi = I_r$ and $\Lambda_\varphi'\Lambda_\varphi$ is diagonal with distinct entries.

Part (a) is standard in factor model literature. Part (b) ensures pervasiveness of factors in the sense that each factor has non-negligible effect on the variability of covariates. This is crucial for asymptotic identification of the common and idiosyncratic components. Parts (c·) partially permit time-series and cross-section dependence in the idiosyncratic component, as well as heteroskedasticity in both dimensions. Possible correlation of $e_{\varphi,it}$ across i sets up the model to have an approximate factor structure. Part (d) allows weak dependence between factors and idiosyncratic errors and (e) is for identification.

The following theorem establishes consistency of kernel factors, for kernels inducing finite nonlinearities, in a large-dimensional framework.

Theorem 2.1 (Theorem 1 Bai and Ng [2002], adapted)**.** *Suppose the kernel function induces finite dimensional nonlinearity, i.e. for $N < \infty$, $\varphi(\cdot) : \mathbb{R}^N \to \mathbb{R}^M$, where $M := M(N)$ is such that $N \leq M(N) < \infty$, and Assumption A holds. Then for any fixed $r \geq 1$, as $M, T \to \infty$*

$$\delta_{NT}^2\left\|\widetilde{F}_{\varphi,t} - H'F_{\varphi,t}\right\|_F^2 = O_p(1), \quad \forall t = 1,\ldots,T,$$

where $\delta_{NT} = \min\{\sqrt{M},\sqrt{T}\}$, $\underset{r\times r}{H} = \frac{\Lambda_\varphi^{0\prime}\Lambda_\varphi^0}{M}\frac{F_\varphi^{0\prime}\widetilde{F}_\varphi}{T}V_{MT}^{-1}$, $\underset{r\times r}{V_{MT}}$ is a diagonal matrix of r largest eigenvalues of $\frac{\Phi\Phi'}{MT}$, $\widetilde{F}_{\varphi,t}$ and $F_{\varphi,t}$ are t-th rows in $\widetilde{F}_\varphi = \sqrt{T}eig_r(\Phi\Phi')$ and F_φ, respectively.

Proof. See appendix A.4. □

Some comments are in order. First, the theorem states that the squared differences between the proposed factor estimator and (a rotation of) the true factor vanish as $M, T \to \infty$. While true factors themselves are not identifiable unless additional assumptions are imposed (see Bai and Ng [2013]), identification of the latent space spanned by factors is just as good as exact identification for forecasting purposes. Second, since for any given number of original variables N the dimension of the transformed space M is fixed and finite, the growth in M is only possible through N and thus the limit on M implies one on N. Third, this result does not imply uniform convergence in t. Lastly,

the results suggest the possibility of $\sqrt{T}$-consistent estimation of the forecasting equation with respect to its conditional mean.

Unfortunately, this result does not generalize to the most interesting kernels (e.g. RBF) inducing infinite-dimensional Hilbert spaces, rendering the traditional approach unsuitable for establishing theoretical properties. Hence we turn to a functional analytic framework which allows rigorous treatment of infinite-dimensional spaces and which has been the classical framework for analyzing statistical properties of functional PCA, and kernel PCA in particular, in machine learning literature.

As will be shown later, it turns out that we can still show that the estimator concentrates around its population counterpart. We briefly present the necessary terms for understanding this result, without aiming to be exhaustive. A sufficiently detailed introduction to the analysis in Hilbert spaces can be found in Blanchard et al. [2006], while a classical reference for operator perturbation theory is Kato [1952].

One of the first major investigations of the statistical properties of kernel PCA can be found in Shawe-Taylor et al. [2002]. The study provides concentration bounds on the sum of eigenvalues of the kernel matrix towards that of corresponding (infinite-dimensional) kernel operators. This permits to characterize the accuracy of kPCA in terms of the reconstruction error, that is the ability to preserve the information about a high-dimensional input in low dimensions. This is of significant interest for certain types of applications, such as pattern recognition. Blanchard et al. [2006] further extend the results of the aforementioned study and improve the bounds on eigenvalues using tools of perturbation theory.

Although the theoretical discussion and the mathematical approaches developed in this literature are extremely valuable, our interest is not in kPCA's ability to reconstruct a given observation. The kernel factor estimator only reduces the dimensionality and passes it to the next stage, without ever going through the reconstruction phase. As the dimensionality is reduced by projecting observations onto the eigenspace – the space spanned by eigenfunctions of the true covariance operator with largest eigenvalues – the interest is in convergence of empirical eigenfunctions towards the true counterparts. Importantly, the proximity of eigenvalues does not guarantee that underlying eigenspaces will also be close.

We now briefly introduce the technical background for understanding our result. Let $\mathcal{H}$ be an inner product space, that is a linear vector space endowed with an inner product

$\langle \cdot, \cdot \rangle_{\mathcal{H}}$, commonly denoted together as $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$. An inner product space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ is a Hilbert space if and only if the norm induced by the inner product $\|\cdot\|_{\mathcal{H}} = \langle \cdot . \cdot \rangle_{\mathcal{H}}^{1/2}$ is complete[4]. We will drop the subscript $\mathcal{H}$ to simplify the notation.

A function $A : \mathcal{F} \to \mathcal{G}$, where $\mathcal{F}, \mathcal{G}$ are vector spaces over $\mathbb{R}$, is a linear operator if $\forall \alpha_1, \alpha_2 \in \mathbb{R}, f, g \in \mathcal{F}, A(\alpha_1 f + \alpha_2 g) = \alpha_1 (Af) + \alpha_2 (Ag)$. A linear operator $L : \mathcal{H} \to \mathcal{H}$ is Hilbert-Schmidt if $\sum_{i \geq 1} \|Le_i\|^2 = \sum_{i,j \geq 1} \langle Le_i, e_j \rangle^2 < \infty$, where $\{e_i\}_{i=1}$ is an orthonormal basis of $\mathcal{H}$.

Let X be a random variable taking values on a general probability space $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if there exists a real-valued Hilbert space and a measurable feature mapping $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, x' \in \mathcal{X},\ k(x, x') = \langle \phi(x), \phi(x') \rangle$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel of $\mathcal{H}$ and $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS), if k satisfies (i) $\forall x \in \mathcal{X}, k(\cdot, x) \in \mathcal{X}$, and the reproducing property (ii) $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \langle f, k(\cdot, x) \rangle = f(x)$. Every function in RKHS can be written as a linear combination of features, $f(\cdot) = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)$. An important result from Aronszajn [1950] guarantees the existence of a unique RKHS for every positive definite k.

Assume that $\mathbb{E}\, \phi(X) = 0$ and $\mathbb{E}\, \|\phi(X)\|^2 < \infty$. A unique covariance operator on $\phi(X)$, $\Sigma = \mathbb{E}\, \phi(X) \otimes \phi(X)$, satisfying $\langle g, \Sigma h \rangle_{\mathcal{H}} = \mathbb{E}\, \langle h, \phi(X) \rangle_{\mathcal{H}} \langle g, \phi(X) \rangle_{\mathcal{H}}$, $\forall g, h \in \mathcal{H}$, always exists (Theorem 2.1 in Blanchard et al. [2006]) and is a positive, self-adjoint trace-class operator. Denote $\widehat{\Sigma} = \frac{1}{T} \sum_{i=1}^{T} \phi(X_i) \otimes \phi(X_i)$ to be its empirical counterpart. Finally, an orthogonal projector in $\mathcal{H}$ onto a closed subspace V is an operator Π_V such that $\Pi_V^2 = \Pi_V$ and $\Pi_V = \Pi_V^*$.

We now lay out two lemmas which lead to the result. Lemma (2.1) bounds the difference between the true and empirical covariance operators.

Lemma 2.1 (Difference between sample and true covariance operators)**.**
Assume random variables $X_1, \ldots, X_T \in \mathcal{X}$ are independent and $\sup_{x \in \mathcal{X}} k(x, x) \leq \bar{k}$, then

$$\mathbb{P}\left(\left\| \widehat{\Sigma} - \Sigma \right\| \geq \left(1 + \sqrt{\frac{\epsilon}{2}} \right) \frac{2\bar{k}}{\sqrt{T}} \right) \leq e^{-\epsilon}.$$

Proof. See appendix A.6. □

Note that both sigmoid and RBF kernels are bounded and hence satisfy the requirement of the above lemma. Lemma (2.2) is an operator perturbation theory result and is adapted from Koltchinskii and Giné [2000] and Zwald and Blanchard [2006]:

[4]Specifically, the limits of all Cauchy sequences of functions must be in the Hilbert space.

Lemma 2.2. *Let A, B be two symmetric linear operators. Denote the distinct eigenvalues of A as $\mu_1 > \ldots > \mu_k > 0$ and let Π_i be the orthogonal projector onto the i-th eigenspace. For a positive integer $p \leq k$ define $\delta_p(A) := \min\{|\mu_i - \mu_j| : 1 \leq i < j \leq p+1\}$. Assuming $\|B\| < \delta_p(A)/4$, then*

$$\|\Pi_i(A) - \Pi_i(A+B)\| \leq \frac{4\,\|B\|}{\delta_i(A)}.$$

Proof. See appendix A.7. □

Finally, the following theorem bounds the difference between empirical and true eigenvectors.

Theorem 2.2. *Denote the i-th eigenvectors of $\widehat{\Sigma}$ and Σ as $\widehat{\psi}_i$ and ψ_i respectively. Then, under the assumptions of Lemma 2.1 and 2.2, as $T \to \infty$ we have*

$$\left\|\widehat{\psi}_i - \psi_i\right\| = o_p(1)$$

Proof. See appendix A.8. □

Some comments are in order. First, note that we require the eigenvalues to be distinct, a well-known restriction (similar to $sin(\theta)$ theorem of Davis and Kahan [1970]), since it is impossible to identify eigenspaces with the same eigenvalues. Second, Theorem 2.2 suggests that the eigenspace estimated by kernel PCA will concentrate close to the true eigenspace. Our kernel factor estimator simply projects onto that eigenspace and hence the precision is expected to increase as $T \to \infty$. Third, this rate does not address the case when variables exhibit dependence, although the exercise in the next section is indicative of some form of concentration, which suggests that the theoretical assumptions might be too conservative. Lastly, it may be possible to obtain a sharper bound since the proof relies on crude inequalities (e.g. triangle inequality).

3 Empirical Evaluation

3.1 Forecasting Models

This subsection discusses specific forms of equations (1) and (2) that are used for forecasting. Autoregressive Diffusion Index (ARDI) model is specified as

$$Y_{t+h} = \beta_0^h + \sum_{p=1}^{P_t^h} \beta_{Y,p}^h Y_{t-p+1} + \sum_{m=1}^{M_t^h} \underset{1\times R_t^h}{\beta_{F,m}^h{}'} \underset{R_t^h\times 1}{F_{t-m+1}} + \epsilon_{t+h}, \tag{13}$$

where superscript h indicates dependence on the time horizon. Note, P_t^h, M_t^h, R_t^h are the number of lags of the target variable, number of lags of factors, number of factors respectively. These three parameters are estimated simultaneously for each time period and time horizon using BIC. Since the true factors are unknown, we instead plug in the estimates from the factor equation, which is discussed next.

Three factor equation specifications are considered,

$$\underset{N\times 1}{X_t} = \Lambda F_t + e_t, \tag{14}$$

$$\underset{2N\times 1}{X_{*,t}} = \Lambda_* F_{*,t} + e_{*,t}, \tag{15}$$

$$\underset{M\times 1}{\varphi(X_t)} = \Lambda_\varphi F_{\varphi,t} + e_{\varphi,t}. \tag{16}$$

Factors in equation (14) are estimated by PCA. Equation (15) is similar, replacing the left-hand side with an augmented vector $X_{*,t} = [X_t, X_t^2]$. This procedure was dubbed as squared principal components (SPCA) in Bai and Ng [2008]. Finally, the last equation applies nonlinearity induced by the selected kernel and is estimated by kPCA.

Forecasts using PCA and SPCA are produced in three steps. First, we extract three factors from the set transformed and standardized predictors using one of the two methods. Second, three parameters are determined according to BIC for each out-of-sample forecasting period and each prediction horizon: the number of lags of the target variable P_t^h, the number of factors M_t^h, the number of lags of the factors K_t^h. Third, the forecasting equation is estimated by least squares and forecasts are produced. The procedure for predicting with kPCA is similar, except there is an additional step where the value of the hyperparameter is specified, and the estimation is instead made in accordance with Algorithm A.1.

3.2 Data and Forecast Construction

As an empirical investigation, we examine whether using kernel factors leads to improved performance in forecasting several key macroeconomic indicators. We use a large dataset from FRED-MD (McCracken and Ng [2016]), which has become one of the classical datasets for empirical analysis of big data. Its latest release consists of 128 monthly US variables running from $1959:01$ through $2020:04$, 736 observations in total. Following previous studies, we set $1960:01$ as the first sample, leaving 724 observations. Since the models presented in this study require stationary series, each of the variables undergo a transformation to achieve stationarity. The decision on a particular form of transformation is generally dependent on the outcome of a unit root test, which is known to lack power in finite samples. So instead, following McCracken and Ng [2016], all interest and unemployment are assumed to be $I(1)$, while price indexes are assumed to be $I(2)$. The transformations applied to each series are described in supplemental materials.

We aim to predict a single time series from this dataset by utilizing the remaining variables. The series to be predicted include 8 variables characterizing different aspects of the economy. Specifically, we take one series from each of the eight variable "groups" in the dataset. The summary is provided in Table A1 in Appendix.

Forecasts are constructed for $h = 1, 3, 6, 9, 12, 18, 24$ months ahead with a rolling time window, the size of which is taken to be $120 - h$. Thus, the pseudo-out-of-sample forecast evaluation period is $1970:01$ to $2020:04$, which is 604 months. We estimate 6 variants of autoregressive diffusion index models. The first model, taken as a benchmark, is a classical ARDI with PC estimates. Several studies have documented a strong performance of this model (see for example, Coulombe et al. [2019]). The second and third take SPCA and so-called PC-squared (PC^2) estimates (Bai and Ng [2008]) respectively, where the latter is identical to the first model with squares of factor estimates added in the forecasting equation. The remaining models are based on kPCA estimates with three different kernels: a sigmoid $k(\mathrm{x_i}, \mathrm{x_j}) = \tanh(\gamma(\mathrm{x_i}'\mathrm{x_j}) + 1)$, a radial basis function (RBF) $k(\mathrm{x_i}, \mathrm{x_j}) = e^{-\gamma\left\|\mathrm{x_i}-\mathrm{x_j}\right\|^2}$ and a quadratic[5] polynomial (poly(2)) kernel $k(\mathrm{x_i}, \mathrm{x_j}) = (\mathrm{x_i}'\mathrm{x_j} + 1)^d,\ d = 2$.

Optimal parameters for each model at each step, P^h, M^h, K^h and the kernel hyperparameter, are determined within the rolling window period, that is our setup only permits

[5]Polynomial kernels of lower and higher order demonstrated poor forecasting ability and are not included.

the information set that would be available at the moment of making a prediction. Specifically, P^h, M^h, K^h are selected by BIC (maximum value allowed for each is set equal to 3) for each out-of-sample period, while γ is determined over a grid of values by so-called time series cross-validation. Specifically, we consecutively predict the latest 5 available observations and select the hyperparameter that minimizes the average error. The standard cross-validation may not theoretically be fully adequate due to the presence of serial correlation in the data and several approaches were suggested to correct it (Racine [2000]).

3.3 Results

The main empirical findings are presented in Table 1. Each value in the table represents the ratio of MSPE of a given estimation method to MSPE produced by PCA estimation. The results range for 8 variables across 7 different forecast horizons. Our results are reproducible: in supplemental materials we provide the script written in Python 3.6 that generates all results within a few hours.

Table 1: Relative MSPEs for 8 variables across 7 different prediction horizons. Each value represents the ratio to out-of-sample MSPE of the autoregression augmented by diffusion indexes estimated by PCA.

	$h=1$	$h=3$	$h=6$	$h=9$	$h=12$	$h=18$	$h=24$
RPI series							
SPCA	1.0355	1.3427	1.5900	1.7746	1.1645	0.9892	1.2647
PC^2	1.0173	1.1228	1.3011	1.1790	1.2873	1.1590	1.1937
kPCA poly(2)	1.0814	1.9185	2.5423	2.2290	2.1792	1.5982	1.6674
kPCA sigmoid	1.0009	1.0104	0.9823	0.9564	0.9721	0.9895	0.9602
kPCA RBF	0.9993	1.0047	0.9874	0.9941	0.9801	0.9673	0.9774
CE16OV series							
SPCA	1.0271	1.4285	1.7018	1.4133	1.7412	2.1467	2.0262
PC^2	0.9918	1.1371	1.2390	1.3073	1.5282	1.3466	1.1950
kPCA poly(2)	1.4346	1.6811	2.8022	2.0038	1.8869	1.6520	1.3763
kPCA sigmoid	1.0065	0.9854	0.9660	0.9600	0.9754	0.9614	0.9065
kPCA RBF	0.9991	0.9804	0.9712	0.9573	0.9711	0.9661	0.9737
HOUST series							
SPCA	1.0020	1.4299	1.4810	1.6383	1.1651	1.3553	1.6741
PC^2	1.0698	1.2187	1.5125	1.5515	1.0870	0.9928	0.7737
kPCA poly(2)	1.0634	1.3674	1.6719	1.9320	1.5364	1.8500	2.5680

Continued on next page

Table 1 – *continued from previous page*

	$h = 1$	$h = 3$	$h = 6$	$h = 9$	$h = 12$	$h = 18$	$h = 24$
kPCA sigmoid	0.9996	0.9884	0.9809	0.9916	0.9861	0.8553	0.8356
kPCA RBF	0.9978	0.9838	0.9329	0.8959	0.9014	0.9389	0.9252
DPCERA3M086SBEA series							
SPCA	1.0764	1.3450	1.4787	1.5401	1.4043	1.4061	1.8804
PC^2	1.0467	1.1900	1.3065	1.1856	1.1229	1.1625	1.3006
kPCA poly(2)	1.1722	1.5608	1.7129	1.4684	1.4583	1.0932	1.3171
kPCA sigmoid	1.0021	0.9958	0.9900	0.9729	0.9721	0.9199	0.8921
kPCA RBF	0.9936	0.9942	0.9639	0.9584	0.9750	0.9792	0.9886
M1SL series							
SPCA	1.1004	1.5903	1.3759	1.2666	1.2002	1.0483	1.1778
PC^2	1.1851	1.4353	1.1629	1.1591	2.6222	2.0673	1.5914
kPCA poly(2)	1.9102	2.2750	2.4906	1.9719	1.5915	1.4239	1.4367
kPCA sigmoid	0.9994	0.9785	0.9473	0.9362	0.9577	0.9247	0.9082
kPCA RBF	0.9878	0.9921	0.9563	0.9462	0.9430	0.9556	0.9654
FEDFUNDS series							
SPCA	1.0202	1.4132	2.2041	1.3547	1.6423	2.2325	2.8836
PC^2	1.2822	1.1414	1.4214	1.6993	1.0800	1.0843	1.2805
kPCA poly(2)	1.7056	1.8629	3.4562	3.0524	1.5233	1.8725	1.5638
kPCA sigmoid	1.0072	0.9406	0.8791	0.9430	0.9930	0.9461	0.8629
kPCA RBF	1.0109	0.9696	0.9592	1.0028	0.9943	0.9370	0.8945
CPIAUCSL series							
SPCA	1.1179	2.0528	2.1043	1.2890	1.1638	1.1911	1.2676
PC^2	1.0175	1.6747	1.1324	1.8631	1.5414	1.0210	1.1420
kPCA poly(2)	0.9850	1.8597	3.9190	1.2954	1.2887	1.2227	1.3105
kPCA sigmoid	0.9909	0.9873	0.9706	0.9826	0.9686	0.9913	0.9839
kPCA RBF	0.9792	1.0166	1.0232	1.0322	0.9788	0.9476	0.9887
S&P 500 series							
SPCA	1.1137	1.3306	1.6338	1.6281	1.5476	1.8519	1.9349
PC^2	1.1150	1.2794	1.4366	1.7406	2.1577	2.0953	1.9181
kPCA poly(2)	1.6275	1.7292	2.4850	2.4820	2.2977	1.9816	1.5089
kPCA sigmoid	0.9842	0.9639	0.9460	0.9268	0.9752	0.9265	0.9637
kPCA RBF	0.9992	0.9888	0.9751	0.9722	0.9677	0.9453	0.9906

Some comments are in order. First, sigmoid and RBF kernel approaches do lead to improved forecasting accuracy, especially at medium- and long-term time horizons. The kernel method is least advantageous for one-step-ahead forecasting. The phenomenon that linearity is hard to beat in a very short horizon is rather well known in the litera-

ture. Luckily, as was shown in Proposition 2.2, kPCA is capable of mimicking a linear PCA by adjusting the kernel hyperparameter closer to 0, which often leads to the parity of the two methods in near-term forecasting. While the gains are not pronounced at $h = 1$, they become apparent at longer horizons. Results vary across variables, but the improvement is prevailing at medium-term horizons and is uniform in one-year and longer predictions. The superiority exhibited by kPCA in many cases is remarkable for macroeconomic forecasting literature.

Second, both SPCA and PC^2 perform substantially worse than a simple PCA. This result contradicts to Bai and Ng [2008], but is consistent with a recent empirical comparison of Exterkate et al. [2016]. Similar to SPCA, poly(2) kernel seeks to model the second-order features of the data and, as a result, often performs on par with SPCA.

Ultimately, note that kPCA's computational complexity is dependent on the number of time periods for estimation $120 - h$, making kPCA slightly faster in this particular exercise. Most importantly, kPCA's advantage would grow in a macroeconomic setting, where "bigger" data (i.e. larger N) is becoming the norm.

4 Concluding Remarks

In this study we have introduced a nonlinear extension of factor modeling based on the kernel method. Although our exposition mainly focused on a feature mapping $\varphi(\cdot)$ enforcing nonlinearity, it is also convenient to think of this approach as kernel smoothing in an inner product space. That is, kernel factors estimators implicitly rely on the weighted distances between original observations. This alternative viewpoint presumes that analyzing the variation in the inner product space, rather than the original space, may be more beneficial. This idea had a profound impact on machine learning and pattern recognition fields, especially as regards to support vector machines (SVMs). By using a positive definite kernel, one can be very flexible in the original space while effectively retaining the simplicity of the linear case in the high-dimensional feature space.

We have demonstrated that constructing factor estimates nonlinearly can be beneficial for macroeconomic forecasting. Specifically, the nonlinearity induced by the sigmoid and RBF kernels leads to considerable gains at medium- and long-term time horizons. This gain in performance comes at no substantial sacrifice, the algorithm remains scalable and computationally fast.

There are several possible extensions. First, it is interesting to see how the performance would change have we pre-selected the variables (targeting) before reducing the dimensionality. As shown in Bai and Ng [2008] and Bulligan et al. [2015] this generally leads to better precision. Second, the forecasting accuracy can be compared with other nonlinear dimension reduction techniques mentioned earlier, such as autoencoders. For the latter, however, one must be aware of the possibility of implicit overfitting by tuning the network architecture. This is not an issue in the current framework as there are a lot fewer parameters to specify. Third, the static factors considered here could possibly be extended to dynamic factors (Forni et al. [2000]), by explicitly incorporating the time domain, or "efficient" factors, by weighing observations by the inverse of the estimated variance.

Appendix

A.1 kPCA algorithm

Algorithm 1: kPCA Algorithm

Input: Observations $X_1, \ldots, X_T \in \mathbb{R}^N$, kernel function $k(\cdot,\cdot)$, dimension r.

for $i = 1, \ldots, T$ **do**
  for $j = 1, \ldots, T$ **do**
    $K_{ij} = k(X_i, X_j)$; `/* Compute similarities */`
  end
end

$K = K - 2\mathbf{1}_{1/T}K + \mathbf{1}_{1/T}K\mathbf{1}_{1/T}$; `/* Standardization */`

$[A, \Lambda] = K/T$; `/* Eigendecomposition */`

$\widehat{F}_r = KA_r$; `/* Compute factors */`

Output: $T \times r$ matrix $\widehat{F}_r$.

A.2 Decomposition of RBF kernel

Let $\mathrm{x}, \mathrm{z} \in \mathbb{R}^k$ and $k(\mathrm{x}, \mathrm{z}) = e^{-\gamma\|\mathrm{x}-\mathrm{z}\|^2}$. Then through the Tailor expansion we can write

$$
\begin{aligned}
k(\mathrm{x}, \mathrm{z}) &= e^{-\gamma\|\mathrm{x}\|^2} e^{-\gamma\|\mathrm{z}\|^2} e^{2\gamma \mathrm{x}'\mathrm{z}} \\
&= e^{-\gamma\|\mathrm{x}\|^2} e^{-\gamma\|\mathrm{z}\|^2} \sum_{j=0}^{\infty} \frac{(2\gamma)^j}{j!} (\mathrm{x}'\mathrm{z})^j \\
&= e^{-\gamma\|\mathrm{x}\|^2} e^{-\gamma\|\mathrm{z}\|^2} \sum_{j=0}^{\infty} \frac{(2\gamma)^j}{j!} \sum_{\sum_{i=1}^k n_i = j} j! \prod_{i=1}^{k} \frac{(x_i y_i)^{n_i}}{n_i!} \\
&= \sum_{j=0}^{\infty} \sum_{\sum_{i=1}^k n_i = j} \left((2\gamma)^{j/2} e^{-\gamma\|\mathrm{x}\|^2} \prod_{i=1}^{k} \frac{x_i^{n_i}}{\sqrt{n_i!}} \right) \left((2\gamma)^{j/2} e^{-\gamma\|\mathrm{z}\|^2} \prod_{i=1}^{k} \frac{y_i^{n_i}}{n_i!} \right) \\
&= \varphi(\mathrm{x})'\varphi(\mathrm{z})
\end{aligned}
$$

That is, $\varphi_j(\mathrm{x}) = \sum_{\sum_{i=1}^k n_i = j} (2\gamma)^{j/2} e^{-\gamma\|\mathrm{x}\|^2} \prod_{i=1}^{k} \frac{x_i^{n_i}}{\sqrt{n_i!}}, \quad j = 0, \ldots, \infty.$

A.3 Proof of Proposition 2.2

Proof. (a) First show that $\lim_{\gamma\to 0}(2\gamma)^{-1}K = XX'$, or $\lim_{\gamma\to 0}(2\gamma)^{-1}k_{ij} = X_i'X_j$, $\forall i, j = 1 \ldots T$, where $k_{ij} = \widetilde{k}_{ij} - \frac{1}{T}\sum_{l=1}^{T}\widetilde{k}_{il} - \frac{1}{T}\sum_{s=1}^{T}\widetilde{k}_{sj} - \frac{1}{T^2}\sum_{m,p=1}^{T}\widetilde{k}_{mp}$ and $\widetilde{k}_{ij} = e^{-\gamma\|X_i - X_j\|_2^2}$. By L'Hopital's rule we can write $\lim_{\gamma\to 0}(2\gamma)^{-1}k_{ij}$ as

$$-\frac{1}{2}\|X_i - X_j\|_2^2 + \frac{1}{2T}\sum_{l=1}^{T}\|X_i - X_l\|_2^2 + \frac{1}{2T}\sum_{l=1}^{T}\|X_l - X_j\|_2^2 - \frac{1}{2T^2}\sum_{l,m=1}^{T}\|X_l - X_m\|_2^2 .$$

Next, use the fact that X is centered, that is $X'\mathbf{1} = \mathbf{0}$ (zero column means) and hence $\sum_{l=1}^{T} X_i'X_l = \sum_{l=1}^{T} X_l'X_i = 0,\ \forall i = 1 \dots T$. This allows to simplify the above as

$$-\frac{1}{2}X_i'X_i + X_i'X_j - \frac{1}{2}X_j'X_j + \frac{1}{2}X_i'X_i + \frac{1}{T}\sum_{l=1}^{T}X_l'X_l + \frac{1}{2}X_j'X_j - \frac{1}{T}\sum_{l=1}^{T}X_l'X_l = X_i'X_j,$$

$\forall i, j = 1 \dots T$, completing the first step. Hence, since the eigenvectors are normalized, we have $\lim_{\gamma\to 0}(2\gamma)^{-1}Keig_r(K) = sXX'eig_r(XX')$ for s equal either to $+1$ or -1. Second, given the SVD decomposition of $X = UDV'$, we have $XX' = VD^2V'$ and $X'X = UD^2U'$, with $D^2 = L$. Thus, $XX'eig_r(XX')L^{-1/2} = UD = Xeig_r(X'X)$. □

(b) First show that $\lim_{\gamma\to 0}(\gamma(1-\tanh^2(c_0))^{-1}k_{ij} = X_i'X_j,\ \forall i, j = 1 \dots T$, where $k_{ij} = \widetilde{k}_{ij} - \frac{1}{T}\sum_{l=1}^{T}\widetilde{k}_{il} - \frac{1}{T}\sum_{s=1}^{T}\widetilde{k}_{sj} - \frac{1}{T^2}\sum_{m,p=1}^{T}\widetilde{k}_{mp}$ and $\widetilde{k}_{ij} = \tanh(c_0 + \gamma X_i'X_j)$. By L'Hopital's rule

$$\lim_{\gamma\to 0}(\gamma(1-\tanh^2(c_0))^{-1}k_{ij} = X_i'X_j - T^{-1}\sum_{l=1}^{T}X_i'X_l - T^{-1}\sum_{l=1}^{T}X_l'X_j + T^{-2}\sum_{l,m=1}^{T}X_l'X_m,$$

which immediately leads to the result once mean-zero property is taken into account. The second step is exactly the same as in (a). □

A.4 Proof of Theorem 2.1

Proof. The proof of Theorem 2.1 for a linear case, $\varphi(X_t) = X_t$, is available in Bai and Ng [2002]. For a general finite-dimensional $\varphi(\cdot)$ the result follows by applying the original theorem to a vector $\varphi(X_t)$ instead. □

A.5 Bounded differences inequality

Theorem (McDiarmid [1989]). *Given independent random variables $X_1, \dots, X_n \in \mathcal{X}$ and a mapping $f : \mathcal{X}^n \to \mathbb{R}$ satisfying*

$$\sup_{x_1,\dots,x_n,x_i'\in\mathcal{X}} |f(x_1, \dots, x_i, \dots, x_n) - f(x_1, \dots, x_i', \dots, x_n)| \le c_i,$$

then for all $\epsilon > 0$,

$$\mathbb{P}(f(X_1, \dots, X_n) - \mathbb{E}(f(X_1, \dots, X_n)) \ge \epsilon) \le e^{\frac{-2\epsilon^2}{\sum_{i=1}^{n} c_i^2}}.$$

A.6 Proof of Lemma 2.1

Proof. Let $\Sigma_x := \varphi(x) \otimes \varphi(x)$. Note that $\|\Sigma_x\| = k(x,x) \leq \bar{k}$, and hence

$$\sup_{x_1,\ldots,x_T,x_i' \in \mathcal{X}} \left| \left\| \frac{1}{T} \sum_{x_1 \ldots x_i \ldots x_T} \Sigma_{x_i} - \mathbb{E}\,\Sigma_X \right\| - \left\| \frac{1}{T} \sum_{x_1 \ldots x_i' \ldots x_T} \Sigma_{x_i} - \mathbb{E}\,\Sigma_X \right\| \right| \leq$$

$$\sup_{x_i \in \mathcal{X}} \frac{1}{T} \left| \|\Sigma_{x_i} - \mathbb{E}\,\Sigma_X\| \right| \leq \frac{2\bar{k}}{T}.$$

Thus, by bounded difference inequality (McDiarmid [1989]) we have

$$\mathbb{P}\left(\left\| \widehat{\Sigma} - \Sigma \right\| - \mathbb{E}\left(\left\| \widehat{\Sigma} - \Sigma \right\| \right) \geq 2\bar{k}\sqrt{\frac{\epsilon}{2T}} \right) \leq e^{-\epsilon}.$$

Finally,

$$\mathbb{E}\left(\left\| \widehat{\Sigma} - \Sigma \right\| \right) \leq \mathbb{E}\left(\left\| \widehat{\Sigma} - \Sigma \right\|^2 \right)^{1/2} = T^{-1/2}\,\mathbb{E}\left(\|\Sigma_X - \mathbb{E}(\Sigma_X)\|^2 \right)^{1/2} \leq \frac{2\bar{k}}{\sqrt{T}},$$

since $\mathbb{E}\left(\|\Sigma_X - \mathbb{E}(\Sigma_X)\|^2 \right) = \langle \Sigma_X - \mathbb{E}\left(\Sigma_X\right), \Sigma_X - \mathbb{E}\left(\Sigma_X\right) \rangle \leq 4\bar{k}^2$. □

A.7 Proof of Lemma 2.2

Proof. See the proof of Lemma 5.2 in Koltchinskii and Giné [2000]. □

A.8 Proof of Theorem 2.2

Proof. Since $\psi_i, \widehat{\psi}_i$ are standardized to be unit length, we have $\left\langle \psi_i, \widehat{\psi}_i \right\rangle^2 \leq 1$ by Cauchy-Schwarz inequality. Choosing eigenvector signs so that $\left\langle \psi_i, \widehat{\psi}_i \right\rangle > 0$, we have

$$\left\| \psi_i - \widehat{\psi}_i \right\|^2 = 2 - 2\left\langle \psi_i, \hat{\psi}_i \right\rangle \leq 2 - 2\left\langle \psi_i, \widehat{\psi}_i \right\rangle^2 = \left\| \Pi_i(\Sigma) - \Pi_i(\widehat{\Sigma}) \right\|^2.$$

Using Lemma 2.2,

$$\left\| \Pi_i(\Sigma) - \Pi_i(\widehat{\Sigma}) \right\| \leq 4\delta_i^{-1}(\Sigma) \left\| \widehat{\Sigma} - \Sigma \right\|,$$

and hence through Lemma 2.1 we have

$$\mathbb{P}\left(\left\| \widehat{\psi}_i - \psi_i \right\| \geq \left(1 + \sqrt{\frac{\epsilon}{2}} \right) \frac{8\bar{k}}{\sqrt{T}\delta_i(\Sigma)} \right) \leq e^{-\epsilon},$$

which implies the result. □

A.9 Mercer's Theorem

Theorem (Mercer [1909]). *Given compact* $\mathcal{X} \subseteq \mathbb{R}^d$ *and continuous* $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$*, satisfying*

$$\int_{\mathrm{y}}\int_{\mathrm{x}} K^2(\mathrm{x},\mathrm{y})d\mathrm{x}d\mathrm{y} < \infty \;\; \textit{and} \;\; \int_{\mathrm{y}}\int_{\mathrm{x}} f(\mathrm{x})K(\mathrm{x},\mathrm{y})f(\mathrm{y})d\mathrm{x}d\mathrm{y} \geq 0, \quad \forall f \in L^2(\mathcal{X}),$$

where $L^2(\mathcal{X}) = \{f : \int f^2(\mathrm{x})d\mathrm{x} < \infty\}$*, then there exist* $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ *and functions* $\{\psi_i(\cdot) \in L^2(\mathcal{X}), i = 1, 2, \ldots\}$ *forming an orthonormal system in* $L^2(\mathcal{X})$*, i.e.* $\langle \psi_i, \psi_j \rangle_{L^2(\mathcal{X})} = \int \psi_i(\mathrm{x})\psi_j(\mathrm{x})d\mathrm{x} = \mathbb{1}_{\{i=j\}}$*, such that*

$$K(\mathrm{x},\mathrm{y}) = \sum_{i=1}^{\infty} \lambda_i \psi_i(\mathrm{x})\psi_i(\mathrm{y}), \quad \forall \mathrm{x},\mathrm{y} \in \mathcal{X}.$$

A.10 Time Series

Table A1: Variables from FRED-MD dataset selected to be (individually) predicted.

Group	Fred-code	Description
Output & income	`RPI`	Real Personal Income
Labor market	`CE16OV`	Civilian Employment
Housing	`HOUST`	Housing Starts: Privately Owned
Consumption & inventories	`DPCERA3M086SBEA`	Real personal consumption
Money & credit	`M1SL`	M1 Money Stock
Interest & exchange rates	`FEDFUNDS`	Effective Federal Funds Rate
Prices	`CPIAUCSL`	CPI: All Items
Stock Market	`S&P 500`	S&P's Common Stock Price Index

References

Anderson, T. (1984). *An Introduction to Multivariate Statistical Analysis*. A Wiley publication in mathematical statistics. Wiley.

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

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, J. and Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150.

Bai, J. and Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304 – 317. Honoring the research contributions of Charles R. Nelson.

Bai, J. and Ng, S. (2013). Principal components estimation and identification of static factors. *Journal of Econometrics*, 176(1):18–29.

Bernanke, B. S., Boivin, J., and Eliasz, P. (2005). Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach*. *The Quarterly Journal of Economics*, 120(1):387–422.

Blanchard, G., Bousquet, O., and Zwald, L. (2006). Statistical properties of kernel principal component analysis. *Machine Learning*, 66:259–294.

Bulligan, G., Marcellino, M., and Venditti, F. (2015). Forecasting economic activity with targeted predictors. *International Journal of Forecasting*, 31(1):188 – 206.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304.

Cheng, X. and Hansen, B. E. (2015). Forecasting with factor-augmented regression: A frequentist model averaging approach. *Journal of Econometrics*, 186(2):280 – 293. High Dimensional Problems in Econometrics.

Connor, G. and Korajczyk, R. (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics*, 15(3):373–394.

Connor, G. and Korajczyk, R. (1993). A test for the number of factors in an approximate factor model. *Journal of Finance*, 48(4):1263–91.

Coulombe, P. G., Stevanovic, D., and Surprenant, S. (2019). How is machine learning useful for macroeconomic forecasting? Cirano working papers, CIRANO.

Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46.

Doz, C., Giannone, D., and Reichlin, L. (2012). A Quasi–Maximum Likelihood Approach for Large, Approximate Dynamic Factor Models. *The Review of Economics and Statistics*, 94(4):1014–1024.

Exterkate, P., Groenen, P. J., Heij, C., and van Dijk, D. (2016). Nonlinear forecasting with many predictors using kernel ridge regression. *International Journal of Forecasting*, 32(3):736 – 753.

Fama, E. F. and French, K. R. (1992). The cross-section of expected stock returns. *The Journal of Finance*, 47(2):427–465.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 75(4):603–680.

Forni, M., Reichlin, L., Hallin, M., and Lippi, M. (2000). The generalized dynamic-factor model: Identification and estimation. *The Review of Economics and Statistics*, 82:540–554.

Giovannetti, B. (2013). Nonlinear forecasting using factor-augmented models. *Journal of Forecasting*, 32(1):32–40.

Hirzel, A. H., Hausser, J., Chessel, D., and Perrin, N. (2002). Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? *Ecology*, 83(7):2027–2036.

Hofmann, T., Schölkopf, B., and Smola, A. J. (2008). Kernel methods in machine learning. *Ann. Statist.*, 36(3):1171–1220.

Jolliffe, I. T. (1986). *Principal Component Analysis and Factor Analysis*, pages 115–128. Springer New York, New York, NY.

Kato, T. (1952). On the perturbation theory of closed linear operators. *J. Math. Soc. Japan*, 4(3-4):323–337.

Kim, H. H. and Swanson, N. R. (2014). Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. *Journal of Econometrics*, 178:352 – 367. Recent Advances in Time Series Econometrics.

Kim, H. H. and Swanson, N. R. (2018). Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. *International Journal of Forecasting*, 34(2):339–354.

Kim, K. I., Jung, K., and Kim, H. J. (2002). Face recognition using kernel principal component analysis. *IEEE Signal Processing Letters*, 9(2):40–42.

Koltchinskii, V. and Giné, E. (2000). Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113–167.

McCracken, M. W. and Ng, S. (2016). FRED-MD: A Monthly Database for Macroeconomic Research. *Journal of Business & Economic Statistics*, 34(4):574–589.

McDiarmid, C. (1989). On the method of bounded differences. *Surveys in Combinatorics, 1989: Invited Papers at the Twelfth British Combinatorial Conference*, page 148–188.

Mercer, J. (1909). Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 209:415–446.

Mikkelsen, J. G., Hillebrand, E., and Urga, G. (2015). Maximum Likelihood Estimation of Time-Varying Loadings in High-Dimensional Factor Models. CREATES Research Papers 2015-61, Department of Economics and Business Economics, Aarhus University.

Negro, M. D. and Otrok, C. (2008). Dynamic factor models with time-varying parameters: measuring changes in international business cycles. Staff Reports 326, Federal Reserve Bank of New York.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics*, 99(1):39 – 61.

Ross, S. A. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13(3):341 – 360.

Sargent, T. and Sims, C. (1977). Business cycle modeling without pretending to have too much a priori economic theory. Working Papers 55, Federal Reserve Bank of Minneapolis.

Scholkopf, B., Smola, A., and Müller, K.-R. (1999). Kernel principal component analysis. In *Advances in kernel methods - Support vector learning*, pages 327–352. MIT Press.

Shawe-Taylor, J., Williams, C., Cristianini, N., and Kandola, J. (2002). On the eigen-spectrum of the gram matrix and its relationship to the operator eigenspectrum. In Cesa-Bianchi, N., Numao, M., and Reischuk, R., editors, *Algorithmic Learning Theory*, pages 23–40, Berlin, Heidelberg. Springer Berlin Heidelberg.

Stock, J. H. and Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Stock, J. H. and Watson, M. W. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162.

Stock, J. H. and Watson, M. W. (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics*, 30(4):481–493.

Teräsvirta, T., Tjøstheim, D., and Granger, C. W. (1994). Chapter 48 aspects of modelling nonlinear time series. In *Handbook of Econometrics*, volume 4, pages 2917 – 2957. Elsevier.

Thompson, G. H. (1938). Methods of estimating mental factors. *Nature*, 141(3562):246–246.

Yalcin, I. and Amemiya, Y. (2001). Nonlinear factor analysis as a statistical method. *Statist. Sci.*, 16(3):275–294.

Zwald, L. and Blanchard, G. (2006). On the convergence of eigenspaces in kernel principal component analysis. In Weiss, Y., Schölkopf, B., and Platt, J. C., editors, *Advances in Neural Information Processing Systems 18*, pages 1649–1656. MIT Press.