Kernel Three Pass Regression Filter

Rajveer Jat and Daanish Padha

University of California, Riverside

July 02, 2024

Abstract

We forecast a single time series using a high-dimensional set of predictors. When predictors share common underlying dynamics, a latent factor model estimated by the Principal Component method effectively characterizes their comovements. These latent factors succinctly summarize the data and aid in prediction, mitigating the curse of dimensionality. However, two significant drawbacks arise: (1) not all factors may be relevant, and utilizing all of them in constructing forecasts leads to inefficiency, and (2) typical models assume a linear dependence of the target on the set of predictors, which limits accuracy. We address these issues through a novel method: Kernel Three-Pass Regression Filter. This method extends a supervised forecasting technique, the Three-Pass Regression Filter, to exclude irrelevant information and operate within an enhanced framework capable of handling nonlinear dependencies. Our method is computationally efficient and demonstrates strong empirical performance, particularly over longer forecast horizons.

Keywords: Forecasting, High dimension, Approximate factor model, Reproducing Kernel Hilbert space, Three-pass regression filter, Machine Learning.

1 Introduction

In recent years, the surge in high-dimensional datasets across fields like economics has ushered in new opportunities and challenges. A paramount issue is the 'curse of dimensionality,' which undermines the effectiveness of traditional finite-dimensional estimation methods. Most modeling techniques applied to high-dimensional data assume the existence of a low-dimensional structure that effectively summarizes the data. One stylized feature of high-dimensional economic datasets is the presence of high and pervasive collinearity among variables, leading researchers to posit a data-generating process that assumes all variables are a function of a few latent factors. This formulation is commonly referred to as the factor model. A vast amount of literature focuses on using this latent factor structure for forecasting applications. A typical example is found in diffusion index models (Stock & Watson (2002)), where latent factors are derived from a high-dimensional set of variables using Principal Components (hereafter, PC) method. These factors are subsequently utilized to forecast a target variable. A limitation of this PC-based factor estimation is its unsupervised nature, i.e., no information from the target variable is incorporated.

Since the primary goal is to forecast a target rather than estimate the underlying factor structure, introducing a degree of supervision can be beneficial. This can help filter out irrelevant information from the predictor set, thus enhancing the predictive accuracy. This can be done in different ways: using soft and hard thresholding methods to remove predictors with no predictive content, as in Bai & Ng (2008), or assigning varying weights to predictors based on their predictive capabilities for the target (see, for example, Huang *et al.* (2022)), or estimate the subset of factors that exhibit predictive power for the target rather than the complete set of factors that drive the predictors, as in Kelly & Pruitt (2015).

The aforementioned models, whether utilizing PCA or supervised factor models, are predicated on the convenient assumption of linearity. However, as underscored in Goulet Coulombe *et al.* (2022), non-linearity often characterizes many predictive relationships in economics, particularly over extended time horizons and within data-rich environments.

Various approaches have been proposed to integrate non-linearity into factor models. For instance, squared principal components (PCs) or principal component squared (PC^2) as seen in Bai & Ng (2008), sufficient forecasting by Fan *et al.* (2017), the kernel trick to estimate factors (Kutateladze (2022)) among others. However, these approaches have limited supervision in the prediction process, if any. For example, Fan *et al.* (2017) estimates factors through an unsupervised method (PC) and then derives sufficient indices using these PCs. Similarly, Kutateladze (2022) essentially applies kernel PCA (an unsupervised method) to estimate the set of factors driving a higher-dimensional space obtained by lifting the set of predictors through the kernel method. In Bai & Ng (2008), a very particular form of non-linearity (quadratic) is examined, which is somewhat ad hoc. Although they employ thresholding methods to reduce predictors to a smaller set, their screening method, however, may still encounter challenges in filtering relevant factors within this subset, leading to inefficient forecasts.

Our paper incorporates both non-linearity and supervision by introducing a novel kernel three-pass regression filter. Our approach essentially applies the three-pass filter (hereafter 3PRF) proposed by Kelly & Pruitt (2015) to a transformed set of predictors.

We adopt the lifting concept similar to Kutateladze (2022), but instead of employing an unsupervised method like kernel PCA, we utilize a supervised method to estimate factors relevant to the target variable.

The table below summarizes our discussion by listing some popular methods[1] in literature and how this paper is placed among them

	Linear	Non-Linear
Unsupervised	PC	kernel PCA, Sq-PC, $PC - sq$
Supervised	3PRF	This Paper

Table 1: Factor Model Based Forecasting Methods

The paper proceeds as follows. Section 2 provides a brief introduction to Kernel methods. Section 3 introduces our estimator and discusses its similarity with the estimator of Kelly & Pruitt (2015). We also list a set of assumptions that ensure the theoretical properties of our estimators, which are given in the subsequent section 4. We present our empirical results in sections 5 and 6 and conclude in section 7. Mathematical proofs and implementation details are given in the appendix.

Definitions and notations

We use $\boldsymbol{y}$ to denote the T × 1 vector of the target variable, i.e. $\boldsymbol{y} = (y_h, y_{h+1} \dots y_{t+h})$. We have N predictors with T observations for each predictor. The cross section of predictors at a time t is given by the $N \times 1$ vector $\boldsymbol{x}_t$. Similarly, the vector of temporal observations of a predictor i is given by $\mathbf{x}_i$. We stack the predictors in a $T \times N$ matrix $\boldsymbol{X}$, $\boldsymbol{X} = (\boldsymbol{x}_1', \boldsymbol{x}_2', \dots, \boldsymbol{x}_T')' = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)$. We have L proxies which we stack in a $T \times\ L$ matrix $\boldsymbol{Z} = (\boldsymbol{z}_1', \boldsymbol{z}_2', \dots, \boldsymbol{z}_T')'$. The demeaning matrix $\boldsymbol{J}_T \equiv \boldsymbol{I}_T - \frac{1}{T}\iota_T\iota_T'$, where

[1] The entries in this table are some of the most popular forecasting methods used in econometric literature. However, by no means do they form an exhaustive set.

$\boldsymbol{I}_T$ is the T-dimensional identity matrix and ι_T the T-vector of ones. For matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ of conformable dimensions, $\boldsymbol{S}_{UV} \equiv \boldsymbol{U}'\boldsymbol{J}_T\boldsymbol{V}$. For the transformed set of predictors $\varphi(\boldsymbol{X})$, $\varphi_j(\mathbf{x})$ denotes the j^{th} observation. $\varphi(\boldsymbol{X}) = (\varphi_1(\mathbf{x}), \varphi_2(\mathbf{x}), \ldots, \varphi_M(\mathbf{x})) = (\varphi(\boldsymbol{x}_1)', \varphi(\boldsymbol{x}_2)', \ldots, \varphi(\boldsymbol{x}_T)')'$. Stochastic orders are denoted by the usual O_p and o_p. For a matrix, $\boldsymbol{O}_p$ and $\boldsymbol{o}_p$ denotes the element wise stochastic order, i.e., a matrix is said to be $\boldsymbol{O}_p(1)$ or $\boldsymbol{o}_p(1)$ if all it's elements are $O_p(1)$ or $o_p(1)$ respectively.

2 Kernel Method

Let $\varphi : \boldsymbol{X} \to \mathcal{F}$ denote a transformation of the original data into a higher-dimensional space[2] containing the original set of predictors and their non-linear transformations. Methods such as principal components or the three-pass regression filter depend on the input $\boldsymbol{X}$ only through the $T \times T$ matrix of dot products $\boldsymbol{X}'\boldsymbol{X}$. Applying these methods to the transformed predictors $\varphi(\boldsymbol{X})$ would therefore require computing the inner product $\varphi(\boldsymbol{X})'\varphi(\boldsymbol{X})$. This computation can be cumbersome or infeasible[3]. Here, the kernel trick proves to be handy, allowing us to calculate inner products within the transformed space without requiring explicit knowledge of φ. A valid[4] kernel corresponds to an inner product of features $\varphi(\boldsymbol{X})$, where the analytical form of the function $\varphi(\cdot)$ may be unknown, but it is guaranteed to exist by Mercer's theorem (Appendix-A.2). Hence, using a Kernel function to compute inner products within a method is akin to performing the estimation exercise(implicitly) on the set of transformed features. In the Supplementary appendix-B.1, we illustrate how different kernel functions correspond to the inner products of transformed inputs.

[2] Precisely, we are referring to a Hilbert space where the inner product of the vectors is well-defined.

[3] When the transformed space is infinite-dimensional

[4] A positive semi-definite kernel as discussed in Appendix-A.2.

Utilizing a transformed set of predictors provides a significant advantage, as many nonlinear relationships can be reformulated as linear relationships in the appropriately transformed space. As an illustration, consider the following example. We generate two variables X and Y from uniform distribution $U[-2, 2]$. Define a binary variable z as:

$$z = \begin{cases} 1 & \text{if } X^2 + Y^2 \leq 2 \\ -1 & \text{otherwise} \end{cases}$$

As shown in the figure-1 (left), a linear boundary cannot separate the two classes of variable z. However, upon transforming the original spaces X and Y to $\varphi_1(X) = \sin^2(X)$ and $\varphi_2(Y) = \cos^2(Y)$ respectively, we find that the two classes can be easily distinguished as seen in figure-1(right). The blue points are in class 1, and the red ones are in class -1.

For the sake of simplicity, this example illustrates the transformation of a two-dimensional input $\boldsymbol{W} = (X, Y)$ into a two-dimensional feature space. The transformed space $\varphi(\boldsymbol{W})$ is typically high-dimensional and potentially infinite-dimensional. Transformation to a higher-dimensional space makes a large set of non-linear forms available, rendering the discovery of a nonlinear relationship very likely.

3 The Estimator

We delineate the three regression passes that we use to construct our forecast. The first two passes, as explained below, are not feasible in practice, whilst the eventual closed-form solutions are. Nonetheless, these steps offer valuable insights into the underlying process of our estimator and elucidate its similarity to the well-known linear three-pass

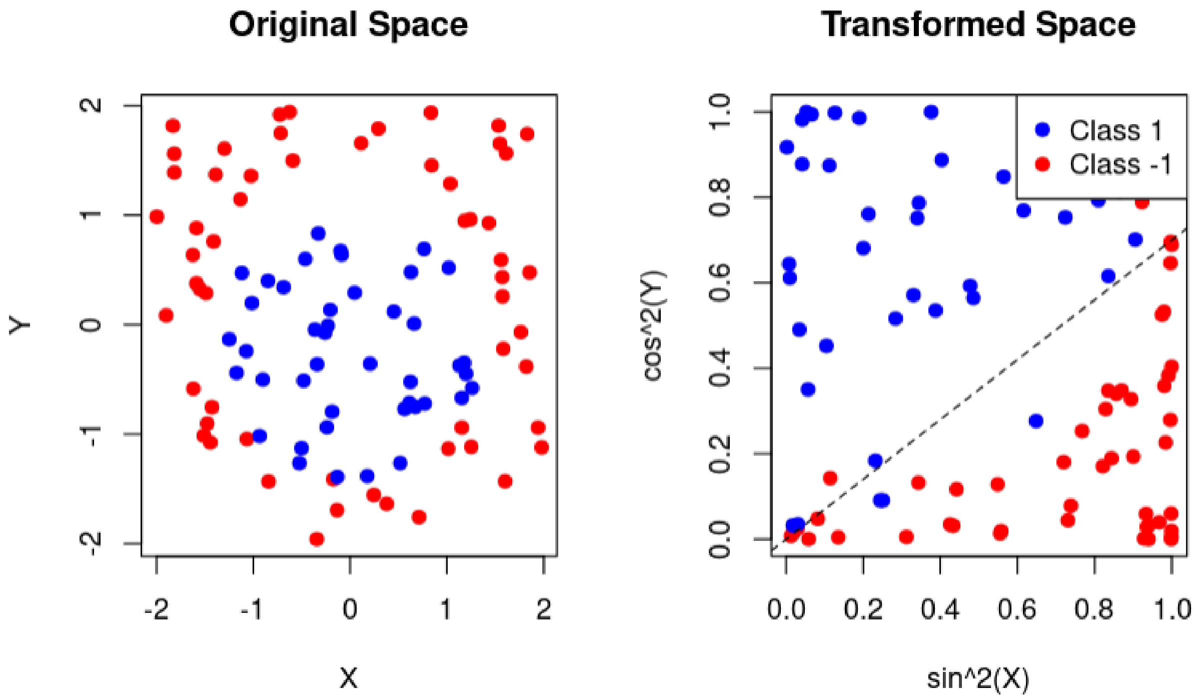


Figure 1: Non-Linear Transformation Making Classification Easy

filter proposed by Kelly & Pruitt (2015).

Below, we list the data generation process for the transformed predictor set ($\varphi(\boldsymbol{X})$), the target ($\boldsymbol{y}$), and the proxies employed for supervision ($\boldsymbol{Z}$). Given the data structure, it is easy to explain why this supervised methodology is effective in estimating the target relevant factors.

Assumption 1 *Data generating Process.*

$$\begin{aligned} \varphi(\boldsymbol{x}_t) &= \boldsymbol{\Phi}\boldsymbol{F}_t + \boldsymbol{\varepsilon}_t & \quad y_{t+h} &= \beta_0 + \boldsymbol{\beta}'\boldsymbol{F}_t + \eta_{t+h} & \quad \boldsymbol{z}_t &= \boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}\boldsymbol{F}_t + \boldsymbol{\omega}_t \\ \varphi(\boldsymbol{X}) &= \boldsymbol{F}\boldsymbol{\Phi}' + \boldsymbol{\varepsilon} & \quad \boldsymbol{y} &= \boldsymbol{\iota}_T\beta_0 + \boldsymbol{F}\boldsymbol{\beta} + \boldsymbol{\eta} & \quad \boldsymbol{Z} &= \boldsymbol{\iota}_T\boldsymbol{\lambda}_0' + \boldsymbol{F}\boldsymbol{\Lambda}' + \boldsymbol{\omega} \end{aligned}$$

where $\boldsymbol{F}_t = (\boldsymbol{f}_t', \boldsymbol{g}_t')', \boldsymbol{\Phi} = (\boldsymbol{\Phi}_f, \boldsymbol{\Phi}_g), \boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_f, \boldsymbol{\Lambda}_g),$ *and* $\boldsymbol{\beta} = (\boldsymbol{\beta}_f', \boldsymbol{0}')'$ *with* $|\boldsymbol{\beta}_f| > \boldsymbol{0}$.

$K_f > 0$ is the dimension of vector $\boldsymbol{f}_t, K_g \geq 0$ is the dimension of vector $\boldsymbol{g}_t, L > 0$ is the dimension of vector $\boldsymbol{z}_t$, and $K = K_f + K_g$.

φ maps our N-dimensional predictors $\boldsymbol{x}_t$ to a higher M-dimensional space. Assumption 1 endows this transformed set of predictors with a factor structure. An underlying factor structure among $\boldsymbol{X}$ implies the existence of a low dimensional plane, projection onto which explains maximal variation in the predictors. An equivalent interpretation of a linear factor structure on $\varphi(\boldsymbol{x}_t)$ would be the existence of a lower dimensional manifold which explains maximum variation in $\boldsymbol{x}_t$. This manifold's basis comprises a few uni-dimensional orthogonal projections of $\varphi(\boldsymbol{x}_t)$.

The infeasible three-passes are summarized in 3 below.

Stage-1	
Pass	Description
1.	Run time series regression of $\varphi_j(\mathbf{x})$ on $\boldsymbol{Z}$ for $j = 1, \ldots, M$, $\varphi_j(\boldsymbol{x}_t) = \tilde{\phi}_{0,j} + \boldsymbol{z}_t'\tilde{\boldsymbol{\phi}}_j + \hat{v}_{1jt}$, retain slope estimate $\tilde{\boldsymbol{\phi}}_j$.
2.	Run cross section regression of $\varphi(\boldsymbol{x}_t)$ on $\tilde{\boldsymbol{\phi}}$ for $t = 1, \ldots, T$, $\varphi_j(\boldsymbol{x}_t) = \tilde{\boldsymbol{\phi}}_j'\hat{\boldsymbol{F}}_t + \hat{v}_{2jt}$, retain slope estimate $\hat{\boldsymbol{F}}_t$.
3.	Run time series regression of y_{t+h} on predictive factors $\hat{\boldsymbol{F}}_t$, $\hat{y}_{t+h} = \hat{\beta}_0 + \hat{\boldsymbol{F}}'\hat{\boldsymbol{\beta}}$, delivers the forecast.

Table 2: Kernel 3PRF

These three passes rely on the fact that the correlation between the transformed $\varphi(\boldsymbol{X})$ and the proxies is only due to target relevant factors. Therefore, pass 1 of the regression asymptotically yields a rotation of the relevant-factor loadings of the j^{th} predictor. Cross-sectional covariance between these loadings and the predictors, across t, is solely affected

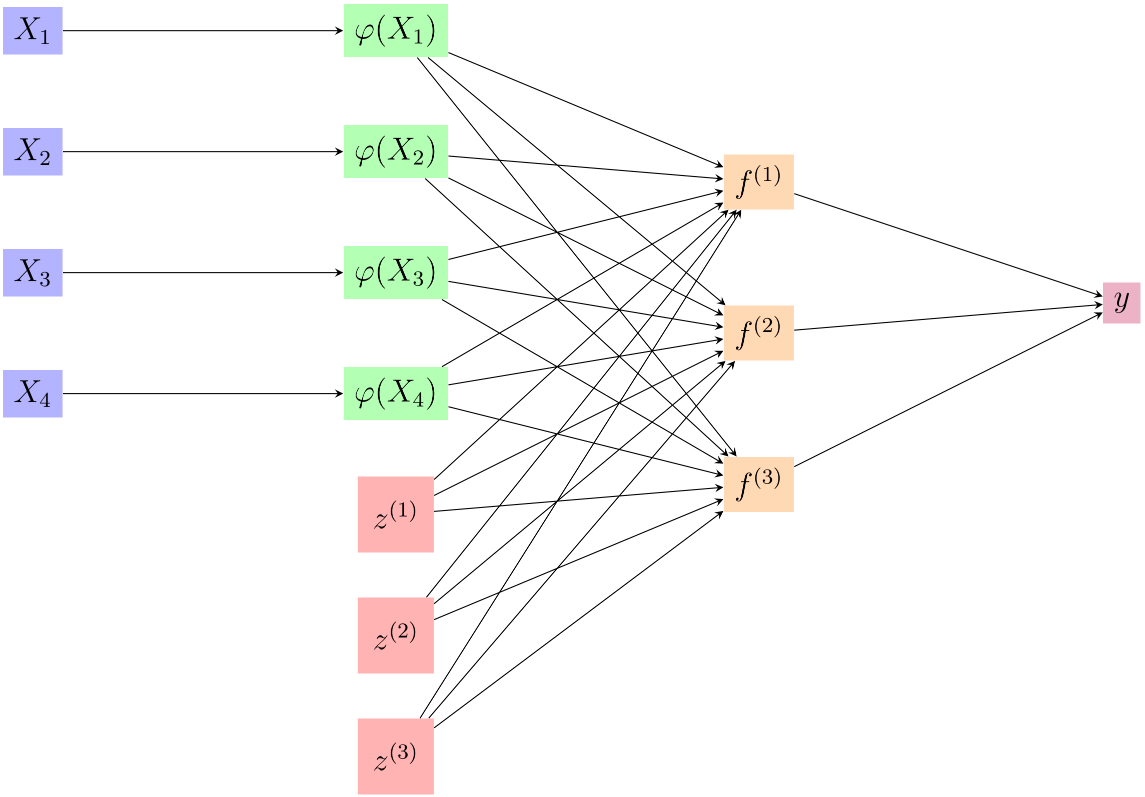


Figure 2: Implementation of the Three Pass regression filter for the case T=4 and L=3 relevant factors. The variables $z^{(1)} \dots z^{(3)}$ and $f^{(1)} \dots f^{(3)}$ are the vectors representing the time series of the respective variables. X_s, (resp $\varphi(X_s)$) represents the cross section of $\boldsymbol{X}$ (resp $\varphi(\boldsymbol{X})$) in period s.

by the target relevant factor(s). Hence, pass 2 of this process traces the factor(s) out as a slope parameter. The last pass involves regressing the target variable on the estimated factor(s). Although these three passes offer valuable insights into the mechanics of our process, they are infeasible in practice due to the unavailability of the transformed inputs $\varphi(\boldsymbol{X})$. This is where the kernel trick proves to be useful. To see this, we note that factor(s), their predictive coefficients, and the forecast can be expressed in closed form as below,

The estimated factor(s) :

$$\begin{aligned}\hat{\boldsymbol{F}}' &= \boldsymbol{S}_{ZZ}\left(\boldsymbol{S}'_{\varphi(X)Z}\boldsymbol{S}_{\varphi(X)Z}\right)^{-1}\boldsymbol{S}'_{\varphi(X)Z}\varphi(\boldsymbol{X})' \\ &= \boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})' \\ &= \boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\mathcal{K}(\boldsymbol{X},\boldsymbol{X}')\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\mathcal{K}(\boldsymbol{X},\boldsymbol{X}')\end{aligned}$$

The estimated coefficient(s) of the factor(s) :

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \boldsymbol{S}_{ZZ}\boldsymbol{S}_{\varphi(\boldsymbol{X})Z}\boldsymbol{S}_{\varphi(\boldsymbol{X})Z}\left(\boldsymbol{S}'_{\varphi(\boldsymbol{X})Z}\boldsymbol{S}_{\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})}\boldsymbol{S}_{\varphi(\boldsymbol{X})Z}\right)^{-1}\boldsymbol{S}'_{\varphi(\boldsymbol{X})Z}\boldsymbol{S}_{\varphi(\boldsymbol{X})y}. \\ &= \left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\times \\ &\qquad \boldsymbol{Z}'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{J}_T\boldsymbol{y} \\ &= \left(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\mathcal{K}(\boldsymbol{X},\boldsymbol{X}')\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\mathcal{K}(\boldsymbol{X},\boldsymbol{X}')\boldsymbol{J}_T\mathcal{K}(\boldsymbol{X},\boldsymbol{X}')\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\mathcal{K}(\boldsymbol{X},\boldsymbol{X}')\boldsymbol{J}_T\boldsymbol{y}\end{aligned}$$

Finally, the estimated target :

$$\begin{aligned}\hat{\boldsymbol{y}} &= \iota_T\bar{y} + \boldsymbol{J}_T\hat{\boldsymbol{F}}\hat{\boldsymbol{\beta}} \\ &= \iota_T\bar{y} + \boldsymbol{J}_T\varphi(\boldsymbol{X})\boldsymbol{S}_{\varphi(\boldsymbol{X})Z}\left(\boldsymbol{S}'_{\varphi(\boldsymbol{X})Z}\boldsymbol{S}_{\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})}\boldsymbol{S}_{\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})}\right)^{-1}\boldsymbol{S}'_{\varphi(\boldsymbol{X})Z}\boldsymbol{S}_{\varphi(\boldsymbol{X})y} \\ =&\iota\bar{y} + \boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{J}_T\boldsymbol{y} \\ =&\iota\bar{y} + \boldsymbol{J}_T\mathcal{K}(\boldsymbol{X},\boldsymbol{X}')\boldsymbol{J}_T\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{J}_T\mathcal{K}(\boldsymbol{X},\boldsymbol{X}')\boldsymbol{J}_T\mathcal{K}(\boldsymbol{X},\boldsymbol{X}')\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\mathcal{K}(\boldsymbol{X},\boldsymbol{X}')\boldsymbol{J}_T\boldsymbol{y}\end{aligned}$$

These expressions are obtained by simply replacing $\boldsymbol{X}$ by $\varphi(\boldsymbol{X})$ in the three-pass regression filter of Kelly & Pruitt (2015). As evident from the expression of $\hat{\boldsymbol{F}}'$, the filtration process applied on the transformed predictor space results in a favorable scenario where the eventual estimate of the factor(s) depends upon the transformed predictors only through their dot products in the transformed space. This holds true for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{y}}$

as well.

This inner product can be computed using a suitable kernel function. Alternatively, it can be inferred that employing a positive semidefinite (psd) kernel function to calculate dot products in these derived expressions is akin to executing the three-pass filter process on the transformed set of predictor(s), which, according to Mercer's theorem, are guaranteed to exist.

The Kernel three-pass regression, like the linear 3PRF, relies on the availability of suitable proxies. Kelly & Pruitt (2015) show that such proxies can always be constructed using the target variable $\boldsymbol{y}$. That process is explained in table-3 below.

0. Initialize $\boldsymbol{r}_0 = \boldsymbol{y}$. For $k = 1, \ldots, L$. (L is the total number of proxies)
1. Define the k^{th} automatic proxy to be $\boldsymbol{r}_{k-1}$. Stop if $k = L$; otherwise proceed.
2. Compute the k3PRF for target $\boldsymbol{y}$ using cross-section $\boldsymbol{X}$ using statistical proxies 1 through k. Denote the resulting forecast $\hat{\boldsymbol{y}}_k$.
3. Calculate $\boldsymbol{r}_k = \boldsymbol{y} - \hat{\boldsymbol{y}}_k$, advance k, and go to step 1.

Table 3: Automatic Proxy-Selection Algorithm

Assumption 1 lays out the factor structure of our model. Below, we delineate a set of additional assumptions under which our model delivers consistent forecasts.

Assumption 2 *(Factors, Loadings and Residuals).*

Let $R < \infty$. For any i, s, t and some $0 < \psi \leq 1$,

1. $\mathbb{E}\left\|\boldsymbol{F}_t\right\|^4 < R, T^{-1}\sum_{s=1}^{T} \boldsymbol{F}_s \xrightarrow[T\to\infty]{p} \boldsymbol{\mu}$ *and* $T^{1/2}\left(\frac{\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}}{T} - \boldsymbol{\Delta}_F\right) = \boldsymbol{O}_p(1)$.

2. $\mathbb{E}\left\|\phi_i\right\|^4 \leq R, M^{-1}\sum_{j=1}^{M} \phi_j \xrightarrow[N\to\infty]{p} \mathbf{0},\ M^{1/2}\left(\frac{\boldsymbol{\Phi}'\boldsymbol{\Phi}}{M} - \mathcal{P}\right) = \boldsymbol{O}_p(1)$.

3. $\mathbb{E}\left(\varepsilon_{it}\right) = 0, \mathbb{E}\left|\varepsilon_{it}\right|^8 \leq R$

4. $\mathbb{E}\left(\boldsymbol{\omega}_t\right) = \mathbf{0}, \mathbb{E}\left\|\boldsymbol{\omega}_t\right\|^4 \leq R, T^{-1/2}\sum_{s=1}^{T} \boldsymbol{\omega}_s = \boldsymbol{O}_p(1)$ *and* $T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\omega} \xrightarrow[N\to\infty]{p} \boldsymbol{\Delta}_\omega$

5. *$\mathbb{E}_t(\eta_{t+h}) = \mathbb{E}(\eta_{t+h} \mid y_t, F_t, y_{t-1}, F_{t-1}, \ldots) = 0, \mathbb{E}\left(\eta_{t+h}^2\right) = \delta_\eta < \infty$, and η_{t+h} is independent of $\phi_i(m)$ and $\varepsilon_{i,t}$ for any $h > 0$.*

Assumption 2.1 requires that our factors are regular in the sense that their covariance matrix is well-behaved asymptotically. Assumption 2.2 is an adaptation from Kelly & Pruitt (2015). Since we assume a factor structure on the transformed space instead of the original predictor space, the normalization in various terms features M and not N, where M is the dimension of our transformed space. Assumptions 2.3-2.5, borrowed from Kelly & Pruitt (2015), impose regularity on various error processes.

Assumption 3 *(Dependence).*

Let $x(m)$ denote the m^{th} element of $\boldsymbol{x}$. For $R < \infty$ and any i, j, t, s, m_1, m_2

1. *$\mathbb{E}(\varepsilon_{it}\varepsilon_{js}) = \sigma_{ij,ts}, |\sigma_{ij,ts}| \le \bar{\sigma}_{ij}$ and $|\sigma_{ij,ts}| \le \tau_{ts}$, and*

   a. $M^{-1}\sum_{i,j=1}^{M} \bar{\sigma}_{ij} \le R$ b. $T^{-1}\sum_{t,s=1}^{T} \tau_{ts} \le R$

   c. $M^{-1}\sum_{i,s} |\sigma_{ii,ts}| \le R$ d. $M^{-1}T^{-1}\sum_{i,j,t,s} |\sigma_{ij,ts}| \le R$

2. $\mathbb{E}\left|M^{-1/2}T^{-1/2}\sum_{s=1}^{T}\sum_{i=1}^{M}\left[\varepsilon_{is}\varepsilon_{it} - \sigma_{ii,st}\right]\right|^4 \le R$

3. $\mathbb{E}\left|T^{-1/2}\sum_{t=1}^{T} F_t(m_1)\,\omega_t(m_2)\right|^2 \le R$

4. $\mathbb{E}\left|T^{-1/2}\sum_{t=1}^{T} \omega_t(m_1)\,\varepsilon_{it}\right|^2 \le R.$

5. $\mathbb{E}\left|T^{-1/2}\sum_{t=1}^{T} F_t(m_1)\,\varepsilon_{it}\right|^2 \le R$

6. $\mathbb{E}\left|M^{-1/2}\sum_{i=1}^{M} \phi_i(m_1)\,\varepsilon_{it}\right|^2 \le R.$

7. $\mathbb{E}\left|T^{-1/2}\sum_{t=1}^{T} F_t(m_1)\,\eta_{t+h}\right|^2 \le R$

Assumption 3.1-3.2 allow various forms of weak cross-sectional and temporal dependence between the idiosyncratic components of the transformed predictors. These assumptions characterize our ‘Approximate’ factor model. The terminology of approximate, as opposed to a strict factor model, alludes to the allowance of these weak correlations, as outlined by Chamberlain & Rothschild (1983). These assumptions are standard in the literature; see Bai (2003). Assumption 3.4-3.7 are either borrowed from or are weaker versions of Assumptions in Kelly & Pruitt (2015). They are reasonable because each of them involves a product of orthogonal series.

Assumption 4 *(Normalization).*

1. $\mathcal{P} = \boldsymbol{I}$

2. $\boldsymbol{\Delta}_F$ *is diagonal, positive definite, and each diagonal element is unique and bounded.*

Assumption 4 is a normalization assumption that is common in factor model literature. It pertains to the non-identifiability of the true factor(s). It is well known that only the vector space spanned by the factor(s) can be consistently estimated but not the factor themselves. Imposing some normalization condition for the uniqueness of solution(s) is common in literature.

Assumption 5 *(Relevant Proxies).*

1. $\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{\Lambda}_f & \boldsymbol{0} \end{bmatrix}$

2. $\boldsymbol{\Lambda}_f$ *is non-singular.*

Assumption 5 outlines the utility of using proxies. Proxies are target-relevant in the sense that they only load on the factor(s) that have any explanatory power for the target. Non-singularity of $\boldsymbol{\Lambda}_f$ ensures that none of the proxies are redundant.

4 Results

We show that our estimated forecast converges to the infeasible best in probability. To show the same, we prove some intermediate results. All the proofs are in the appendix. Define $\delta_{MT} \equiv min\{\sqrt{M}, \sqrt{T}\}$. Define $\boldsymbol{H}_f \equiv \hat{\boldsymbol{F}}_A \hat{\boldsymbol{F}}_B^{-1} \boldsymbol{\Lambda} \boldsymbol{\Delta}_F \mathcal{P}$ where, $\hat{\boldsymbol{F}}_A = T^{-1} \boldsymbol{Z}' \boldsymbol{J}_T \boldsymbol{Z}$ and $\hat{\boldsymbol{F}}_B = M^{-1} T^{-2} \boldsymbol{Z}' \boldsymbol{J}_T \varphi(\boldsymbol{X}) \varphi(\boldsymbol{X}') \boldsymbol{J}_T \boldsymbol{Z}$

Theorem 1 *If Assumption 1-5 hold, we have*

$$\hat{\boldsymbol{F}}_t - \boldsymbol{H}_f \boldsymbol{f}_t = \boldsymbol{O}_p(\delta_{MT}^{-1})$$

This theorem establishes the estimated factor(s) convergence to the true factors up to a rotation. It is well known in the literature on factor models[5], that true underlying factor(s) are not identifiable; we instead estimate a rotated version of the true factors, which preserves their span.

Define $\boldsymbol{G}_\beta \equiv \hat{\boldsymbol{\beta}}_1^{-1} \hat{\boldsymbol{\beta}}_2 \left[\boldsymbol{\Lambda} \boldsymbol{\Delta}_F \mathcal{P} \boldsymbol{\Delta}_F \mathcal{P} \boldsymbol{\Delta}_F \boldsymbol{\Lambda}' \right]^{-1} \boldsymbol{\Lambda} \boldsymbol{\Delta}_F \mathcal{P} \boldsymbol{\Delta}_F$, where
$\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{F}}_A$ and $\hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{F}}_B$

Theorem 2 *If Assumption 1-5 hold, we have*

$$\hat{\boldsymbol{\beta}} - \boldsymbol{G}_\beta \boldsymbol{\beta} = \boldsymbol{O}_p(\delta_{MT}^{-1}).$$

$$\boldsymbol{H}_f{}' \boldsymbol{G}_\beta = \boldsymbol{I}$$

This theorem establishes the convergence of the predictive coefficients to a rotation of the true coefficients. Just like in the case of factor(s), true coefficients are not identifiable

[5] This feature of inherent unidentifiability has been emphasized in Bai (2003), Kelly & Pruitt (2015) among other papers. The normalization imposed in assumption 5 is done to handle this issue.

and we instead estimate their rotation. The rates established in **Theorems** 1 and 2 differ from the rates established in Kelly & Pruitt (2015) and the reason is that our definition of rotation matrices $\boldsymbol{H}_f$ and $\boldsymbol{G}_\beta$ are different from Kelly & Pruitt (2015). (See Remark 1).

Remark 1 *As highlighted in Bai & Ng (2006) and also emphasized in Kelly & Pruitt (2015), the presence of matrices $\boldsymbol{H}_f$ and $\boldsymbol{G}_\beta$ in **Theorem** 1 and 2 highlight the fact we are essentially estimating a vector space. These theorems "pertain to the difference between $\left[\hat{\boldsymbol{F}}_t/\hat{\boldsymbol{\beta}}\right]$ and the space spanned by $[\boldsymbol{F}_t/\boldsymbol{\beta}]$". The product $\boldsymbol{H}'_f\boldsymbol{G}_\beta$ equals an identity matrix, cancelling the rotations in the estimated coefficients and the factors; thereby consistently estimating direction spanned by $\boldsymbol{\beta}'\boldsymbol{F}_t$. However, this characteristic is absent in Theorems 5 and 6 of Kelly & Pruitt (2015). The matrices $\boldsymbol{H}$ and $\boldsymbol{G}_\beta$ as defined in their paper do not necessarily yield a product that equals an identity matrix.*

Theorem 3 *If Assumption 1-5 hold, we have*

$$\hat{y}_{t+h} - \mathbb{E}_t y_{t+h} = O_p(\delta_{MT}^{-1})$$

Combining **Theorem** 1 and 2, the convergence $\hat{y}_{t+h}$ of follows directly. Our proof, unlike Kelly & Pruitt (2015) uses the convergence results for the estimated factor(s) and coefficients to obtain this result.

Remark 2 *The rates established in **Theorem** 1, 2 and 3 are different from the result in Kelly & Pruitt (2015) where the corresponding rates are $O_p(T^{-1/2})$, $O_p(T^{-1/2})$ and $O_p(N^{-1/2})$[6] respectively (see **Theorems** 4, 5 and 6 in their paper). For **Theorem** 1 and 2, the difference is explained by a different definition of the rotation matrices in our paper*

[6] For our case, it should have been $O_p(M^{-1/2})$ as per their theorem since we apply 3PRF to the transformed M-dimesnional space.

*(see **Remark** 1). For establishing the convergence of $\hat{y}_{t+h}$, their proof follows two steps. First they show that $\hat{y}_{t+h} - \tilde{y}_{t+h} = O_p(T^{-1/2})$, where $\tilde{y}_{t+h}$ is defined in their appendix. Then then they argue that $\sqrt{T}\tilde{y}_{t+h} \underset{T,N\to\infty}{\longrightarrow} \mathbb{E}_t y_{t+h}$. Since $\tilde{y}_{t+h}$[7] is $O_p(1)$, $\sqrt{T}\tilde{y}_{t+h}$ would diverge to infinity and their statement would be false. We presume that they erroneously wrote this and instead wanted to imply that $\sqrt{T}\left(\tilde{y}_{t+h} - \mathbb{E}_t y_{t+h}\right) \underset{T,N\to\infty}{\longrightarrow} 0$. However this statement is false because $\tilde{y}_{t+h} - \mathbb{E}_t y_{t+h}$ has random elements which converge to 0 at a rate which is $O_p(M^{-1/2}) + O_p(T^{-1/2}) = O_p(\delta_{MT}^{-1})$.*

5 Empirical Applications

We apply our proposed method to real-world applications, focusing on forecasting time series variables across various economic domains such as national income, finance, labor, housing, prices, etc. To assess the performance of our approach, we conduct comparative analyses against competitive methods, employing the out-of-sample R^2 performance metric as a benchmark. Out of sample R^2 is computed as:

$$R^2 = 1 - \frac{\sum_{i \in \text{test-data}} (y_i - \hat{y}_i)^2}{\sum_{i \in \text{test-data}} (y_i - \bar{y}_{\text{train}})^2}$$

It computes the out-of-sample proximity of our forecast $\hat{y}$ with the target (y) relative to a historical mean $(\bar{y})$; a positive value indicates that the forecast is better than the historical mean. Detailed explanations of performance metrics computation are provided in the Supplementary appendix-B.2.

We compare our method against six different forecasting methods. The first is the PC regression proposed by Stock & Watson (2002); which we write as *PC* in our perfor-

[7] The definition of $\tilde{y}_{t+h}$ and fact that it is $O_p(1)$ can be seen from the proof of **Theorem** 6 of Kelly & Pruitt (2015)

mance tables, *PC-Squared* (*PC-Sq*) and *Squared-PC*(*Sq-PC*) of Bai & Ng (2008), kernel PCA (*kPCA*) [Kutateladze (2022)], our linear counterpart, the *3PRF*, and autoregressive model of lag order two[8]. Some of these methods require tuning of hyper-parameters to provide the best results, we do tune them as discussed in the subsection-5.1.2.

As discussed in section 2, different kernel functions correspond to different $\varphi(\cdot)$. We use the radial basis function (RBF) kernel.

$$K_\sigma(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

where σ is a hyperparameter determined via cross-validation. The use of this kernel is justified by its strong performance in macroeconomic forecasting, as documented by Sermpinis *et al.* (2014), Exterkate *et al.* (2016), and Kutateladze (2022).

5.1 Data and Hyper-Parameter Tuning

We utilize the quarterly macroeconomic dataset, FRED-QD. It spans the time period 1959-2023. This dataset encompasses a comprehensive array of more than 250 variables, including macroeconomic (such as GDP, Consumption, and Investment), financial, labor market, housing, and industrial and manufacturing variables. We present a tabulation comprising the mnemonic codes and details of the variables in the FRED-QD dataset alongside their counterparts in the Stock-Watson dataset in Supplementary appendix-B.4 for the series we forecast in this section.

[8] We compared the performance across various autoregressive lags and found minimal differences. However, a lag of two generally showed superior performance in most cases.

5.1.1 Data Transformation

Estimation exercises involving nonstationary time series pose significant challenges. Nonstationarity is ubiquitous in economic and financial data. Nonstationary variables lack a defined population mean, and the sample standard deviation tends to diverge as the number of observations increases, see Onatski & Wang (2021) for a more detailed discussion. Generally, researchers address this by manually examining each series to identify necessary transformations before computing principal components. Hamilton & Xi (2024) offers an improved method for transforming the predictors to achieve stationarity. We use their method to make our data stationarity.

Scholars in the literature often employ sample periods devoid of structural breaks. Fan *et al.* (2023) notes that "There exist significant structural breaks for many variables around the year of the financial crisis in 2008 which makes our data non-stationary even after performing the suggested transformations". Therefore, our study focuses on the stationary period spanning from 1965 to 2007[9]. We conduct analyses on different combinations of the sample periods, including the entire available sample period from 1959 to 2023, and find no qualitative discrepancies in our findings.

In our main analysis, the sample period from 1964:Q4 to 2007:Q1 comprises $T = 170$ observations (periods) and $N = 176$ variables (predictors). While the data is initially available for around 250 series, those with missing values are removed; this leaves us with a total of 176 series. The model training and hyperparameter tuning are conducted within a rolling window framework, utilizing 70% of the total observations as the width of the rolling window. We observe qualitatively similar performance across varying window widths (50%, 60% of total data).

[9] Another indirect advantage of the choice of this sample period is that it gives us the number of samples less than the number of predictors ($T < N$), hence a truly high-dimensional scenario to test our method in.

5.1.2 Hyper-parameter Tuning

Our methodology incorporates the kernel as a fundamental element of the estimation process. The kernel function includes a hyperparameter that necessitates optimization. Concurrently, a similar hyperparameter requires tuning in the context of a competitor method, namely kernel PCA. Thus, we employ an identical tuning procedure for both methodologies. We adopt a RBF kernel, which relies on a single hyperparameter, denoted as σ, for our specific applications. We partition the data into two folds and conduct cross-validation to determine the optimal tuning parameter. Further elaboration on the algorithm employed for this purpose is provided in the Supplementary appendix-B.5.

Furthermore, among our competitive methodologies, where factors are computed as PCs, we are required to specify the number of factors. To address this, we employ the eigenvalue ratio test method proposed by Ahn & Horenstein (2013). This method computes the ratio of each eigenvalue to its predecessor and selects the number of principal components corresponding to the index where this ratio attains its maximum value. We employ a single factor throughout our analyses in both the 3PRF and kernel 3PRF models. This choice is often prudent within the 3PRF setting, as elucidated by Kelly & Pruitt (2015), who highlight instances where a single factor can effectively represent a multi-factor system. When factors exhibit the same variances[10], a single proxy achieves optimal performance, and even when variances are not identical but closely aligned, one factor estimated through a single auto proxy typically explains a significant portion of the variation[11], rendering residual variation minimal. While we assessed the performance of our estimator with varying numbers of factors, we consistently observed that a single factor predominates, thus, we report results based on this configuration.

[10] See appendix section A.7.2 in Kelly & Pruitt (2015)

[11] See simulations in appendix A.7.3 in Kelly & Pruitt (2015)

5.2 Forecasting Using Theory Guided Proxies

The primary objective of this subsection is to establish the viability of theory-guided proxies in forecasting using our method. We also compare with the linear benchmark, i.e., 3PRF of Kelly & Pruitt (2015). A more extensive performance evaluation will be presented in subsequent subsections, where we used the auto-proxy method discussed in table-3 to construct forecasts using Kernel 3PRF.

5.2.1 Forecasting GDP Using Investment and Consumption

We construct GDP forecasts using Consumption and investment as proxies and report the results in Table-4. This exercise proves the efficacy of K3PRF over 3PRF while employing theory-guided proxies. Furthermore, our method outperforms the nearest competitive method Kelly & Pruitt (2015).

Proxy	3PRF	k3PRF
Consumption and Investment	0.621	0.768
Investment	0.627	0.748
Consumption	0.589	0.760

Table 4: One-period Ahead Out-of-Sample R^2 for National Income

5.2.2 Forecasting Inflation using Quantity Theory of Money

We reproduce the theory-guided proxy example discussed in Kelly & Pruitt (2015). Δ(Price level) i.e. inflation is our target variable for forecasting. The results for one-period ahead inflation forecasts are presented in Table-5. The quantity theory of money equation states that:

$$\frac{\Delta(\text{Money supply}) \times \Delta(\text{Velocity of money})}{\Delta(\text{Real Product})} = \Delta(\text{Price level})$$

Proxy	3PRF	k3PRF
GDP and Money Supply	0.265	0.265
GDP	0.037	0.037
Money Supply	0.350	0.355

Table 5: One-period Ahead Out-of-Sample R^2 for Inflation

The results indicate that the theory-guided proxies effectively capture inflation dynamics, yielding performance comparable to that of the closest competitor. It is important to emphasize again that this analysis focuses on one-step-ahead forecasts, which are not the primary strength of our methodology. The purpose of presenting these results is solely to demonstrate the workings of the procedure through the theory-guided proxies.

5.3 Comparative Forecasting Plots

To visually demonstrate the enhanced performance of kernel 3PRF compared to its linear counterpart, we provide comparative performance plots across four distinct types of economic series spanning various domains: macroeconomic series (*Exports*), price series (*GDP Deflator*), manufacturing series (*Industrial Production*), and financial series (*S&P 500 Index*) in figure-5.3 and 4. Plots of all other series on different forecast horizons are given in the Supplementary appendix-B.6.

5.4 Forecasting Aggregate Macroeconomic Variables

An astute economic decision, such as monetary policy formulation, hinges upon well-informed anticipations of future trends in macroeconomic and financial data. Consequently, forecasting macroeconomic variables emerges as a pivotal pursuit for economists. Quoting Federal Reserve of New York's website, Kim & Swanson (2014) notes, "In formulating the nation's monetary policy, the Federal Reserve considers a number of factors,

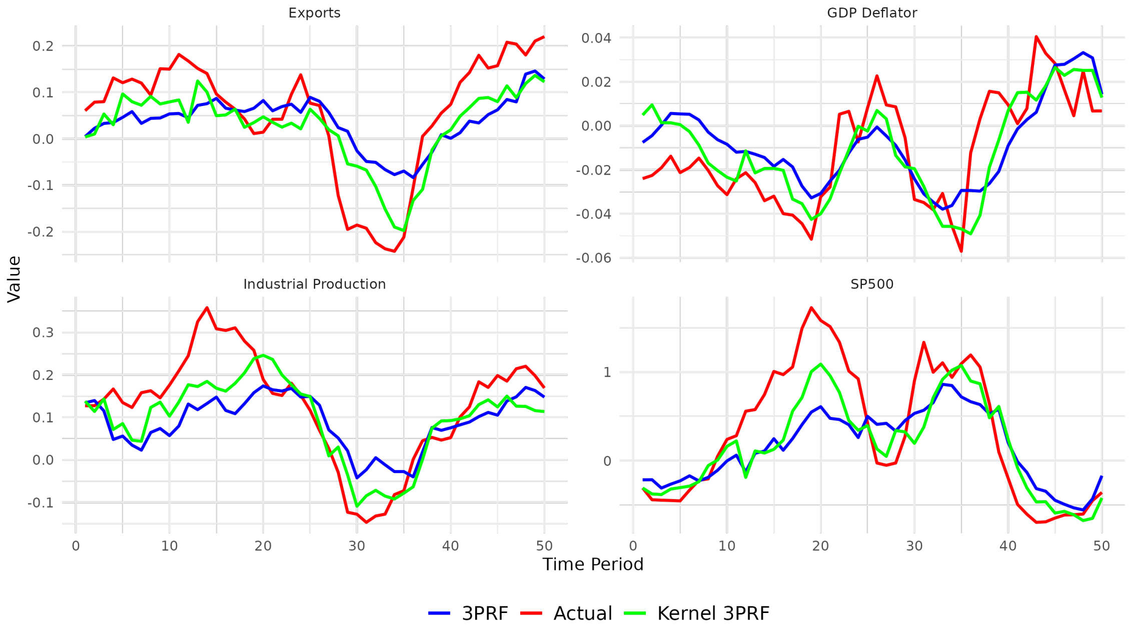


Figure 3: Short Horizon (One period ahead) Forecasting: Comparative Performance

including the economic and financial indicators which follow, as well as the anecdotal reports compiled in the Beige Book. Real Gross Domestic Product (GDP); Consumer Price Index (CPI); Nonfarm Payroll Employment Housing Starts; Industrial Production/Capacity Utilization; Retail Sales; Business Sales and Inventories; Advance Durable Goods Shipments, New Orders and Unfilled Orders; Lightweight Vehicle Sales; Yield on 10-year Treasury Bond; S&P 500 Stock Index; M2". We, therefore, aim to forecast some of these crucial indicators in this paper. We compare the performance of our model against the competitors. This section forecasts seven macro series: GDP, Consumption, Investment, Exports, Imports, Fixed Investment, and Industrial Production (Final).

To present the results in an organized manner, we create two tables. In Table-6, we display the forecasting performance for three series: GDP, Consumption, and Investment, which we informally refer to as 'Group-I'. Table-7 presents a comparative analysis

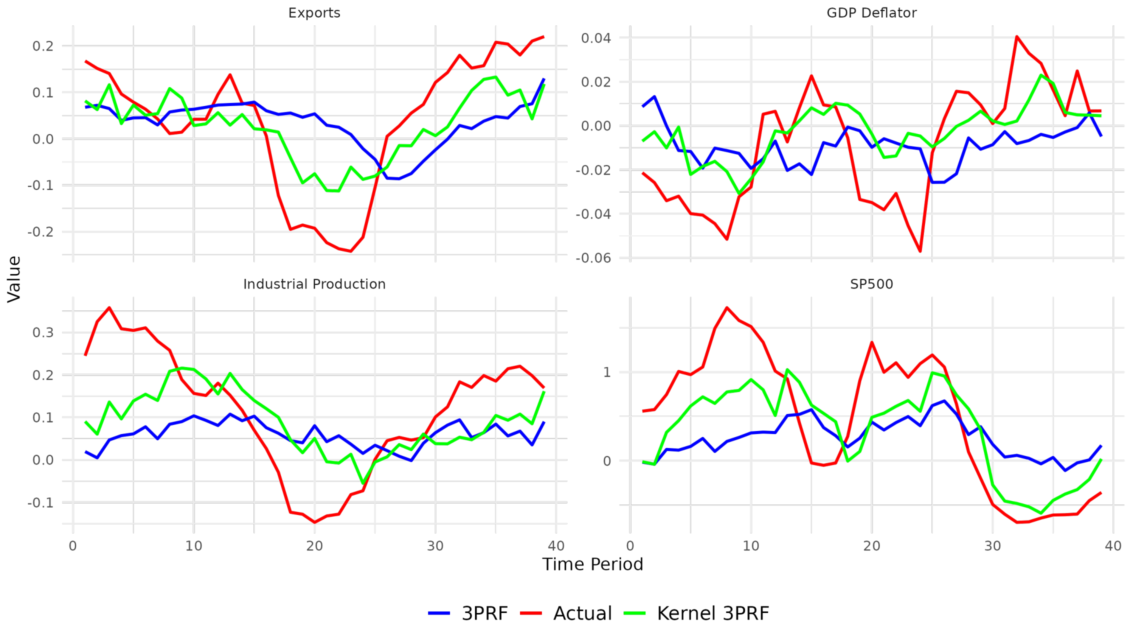


Figure 4: Long Horizon (Twelve periods ahead) Forecasting: Comparative Performance

of forecasting performance for 'Group-II' macro variables[12]: Exports, Imports, Fixed Investments, and Industrial Production (Final Index). As defined earlier in the text, the reported numbers in the tables represent out-of-sample R^2 values across various forecast horizons ranging from one period ahead to twelve periods ahead.

Results highlight a secular observation that among various unsupervised forecasting methodologies, PC, Squared-PC, PC-squared, and non-linear unsupervised approaches such as kernel PCA, none exhibit superior performance compared to our proposed method across any forecast horizon for the seven series under consideration. While the supervised linear forecasting model 3PRF demonstrates improved performance relative to the unsupervised techniques, it still falls short of outperforming our non-linear supervised approach. Notably, the autoregressive (AR) model emerges as the sole con-

[12] Variables' FRED-QD code and description can be found in Supplementary appendix-B.4

GDP									
	Method	h=1	h=2	h=3	h=4	h=6	h=8	h=10	h=12
	AR(2)	0.929	0.906	0.843	0.719	0.302	-0.216	-0.555	-0.724
	PC	0.717	0.650	0.575	0.492	0.311	0.130	-0.001	-0.075
	Sq-PC	0.615	0.593	0.552	0.488	0.290	0.076	-0.092	-0.166
	PC-Sq	0.773	0.733	0.676	0.594	0.398	0.175	0.008	-0.063
	kPCA	0.638	0.589	0.528	0.464	0.322	0.204	0.060	0.063
	3PRF	0.667	0.619	0.561	0.493	0.341	0.193	0.130	0.201
	k3PRF	0.808	0.788	0.757	0.701	0.603	0.544	0.608	0.434
Consumption									
	Method	h=1	h=2	h=3	h=4	h=6	h=8	h=10	h=12
	AR(2)	0.957	0.943	0.892	0.805	0.485	0.015	-0.375	-0.557
	PC	0.573	0.554	0.504	0.430	0.238	0.038	-0.093	-0.155
	Sq-PC	0.546	0.541	0.499	0.428	0.235	0.025	-0.137	-0.206
	PC-Sq	0.611	0.637	0.628	0.596	0.412	0.161	-0.041	-0.128
	kPCA	0.433	0.419	0.369	0.319	0.143	0.076	0.039	0.181
	3PRF	0.589	0.547	0.501	0.464	0.386	0.196	0.169	0.326
	k3PRF	0.713	0.730	0.720	0.741	0.770	0.747	0.275	0.496
Investment									
	Method	h=1	h=2	h=3	h=4	h=6	h=8	h=10	h=12
	AR(2)	0.830	0.807	0.711	0.546	0.087	-0.450	-0.544	-0.586
	PC	0.516	0.393	0.300	0.231	0.149	0.089	0.030	0.011
	Sq-PC	0.398	0.348	0.297	0.238	0.099	-0.022	-0.083	-0.065
	PC-Sq	0.605	0.488	0.391	0.296	0.186	0.090	0.017	0.044
	kPCA	0.479	0.390	0.317	0.272	0.196	0.030	-0.016	-0.013
	3PRF	0.597	0.484	0.429	0.369	0.273	0.111	0.083	0.176
	k3PRF	0.760	0.640	0.478	0.605	0.433	0.199	0.169	0.389

Table 6: h-period ahead out of sample R^2 of Macro Variables : Group-I

tender capable of surpassing our method in the shorter horizons, albeit only marginally and for a few series. However, our method significantly outperforms all the competitors across longer horizons. Therefore, our method emerges as a dependable and preferred forecasting framework across all forecast horizons in macroeconomic prediction tasks.

5.5 Forecasting Labor Market and Price Variables

This analysis aims to forecast key labor market and price variables. Within the labor market category, we focus on unemployment rates and total non-farm employment (Non-farm Emp). We examine the GDP Deflator and the Consumer Price Index (CPI) for

Exports									
	Method	h=1	h=2	h=3	h=4	h=6	h=8	h=10	h=12
	AR(2)	0.928	0.926	0.906	0.863	0.723	0.522	0.409	0.302
	PC	0.353	0.306	0.248	0.193	0.123	0.107	0.106	0.109
	Sq-PC	0.275	0.249	0.215	0.183	0.120	0.056	0.008	-0.013
	PC-Sq	0.399	0.326	0.243	0.166	0.073	0.066	0.113	0.194
	kPCA	0.027	0.033	0.033	0.270	0.142	-0.002	-0.044	0.130
	3PRF	0.535	0.523	0.459	0.389	0.223	0.137	0.109	0.092
	k3PRF	0.724	0.705	0.641	0.602	0.546	0.575	0.600	0.631
Imports									
	Method	h=1	h=2	h=3	h=4	h=6	h=8	h=10	h=12
	AR(2)	0.969	0.964	0.951	0.931	0.845	0.710	0.577	0.460
	PC	0.417	0.380	0.343	0.306	0.233	0.154	0.072	0.006
	Sq-PC	0.395	0.373	0.341	0.299	0.194	0.079	-0.005	-0.046
	PC-Sq	0.477	0.462	0.438	0.398	0.306	0.182	0.060	0.000
	kPCA	0.421	0.389	0.348	0.311	0.241	0.081	0.064	0.033
	3PRF	0.546	0.506	0.468	0.436	0.394	0.347	0.322	0.338
	k3PRF	0.777	0.783	0.790	0.786	0.749	0.411	0.388	0.558
Fixed Invest.									
	Method	h=1	h=2	h=3	h=4	h=6	h=8	h=10	h=12
	AR(2)	0.905	0.881	0.818	0.682	0.224	-0.267	-0.467	-0.605
	PC	0.490	0.384	0.290	0.220	0.134	0.088	0.042	0.016
	Sq-PC	0.401	0.352	0.293	0.231	0.095	-0.024	-0.077	-0.064
	PC-Sq	0.595	0.492	0.385	0.314	0.208	0.104	0.030	0.068
	kPCA	0.498	0.407	0.315	0.250	0.167	0.039	-0.034	0.007
	3PRF	0.525	0.454	0.389	0.348	0.251	0.122	0.127	0.226
	k3PRF	0.736	0.659	0.426	0.578	0.265	0.235	0.261	0.359
IP : Final									
	Method	h=1	h=2	h=3	h=4	h=6	h=8	h=10	h=12
	AR(2)	0.830	0.807	0.711	0.546	0.087	-0.450	-0.544	-0.586
	PC	0.516	0.393	0.300	0.231	0.149	0.089	0.030	0.011
	Sq-PC	0.398	0.348	0.297	0.238	0.099	-0.022	-0.083	-0.065
	PC-Sq	0.605	0.488	0.391	0.296	0.186	0.090	0.017	0.044
	kPCA	0.479	0.390	0.317	0.272	0.196	0.030	-0.016	-0.013
	3PRF	0.597	0.484	0.429	0.369	0.273	0.111	0.083	0.176
	k3PRF	0.760	0.640	0.478	0.605	0.433	0.199	0.169	0.389

Table 7: h-period ahead out of sample R^2 of Macro Variables : Group-II

price variables. The GDP Deflator offers insights into overall inflation at the macroeconomic level, while the CPI captures inflation experienced by consumers at a more disaggregated level. The results of this analysis are summarized in Table 8.

Our relative forecast performance results are qualitatively similar to those of aggregate macroeconomic series forecasting.

Nonfarm Emp

Method	h=1	h=2	h=3	h=4	h=6	h=8	h=10	h=12
AR(2)	0.992	0.961	0.864	0.693	0.170	-0.429	-0.881	-1.079
PC	0.786	0.728	0.604	0.435	0.057	-0.219	-0.258	-0.146
Sq-PC	0.528	0.498	0.440	0.361	0.167	-0.024	-0.109	-0.098
PC-Sq	0.836	0.795	0.679	0.510	0.131	-0.139	-0.210	-0.110
kPCA	0.832	0.790	0.702	0.587	0.370	0.196	0.112	0.059
3PRF	0.765	0.731	0.712	0.662	0.407	0.312	0.264	0.229
k3PRF	0.929	0.895	0.846	0.768	0.556	0.444	0.441	0.584

Unemp Rate

Method	h=1	h=2	h=3	h=4	h=6	h=8	h=10	h=12
AR(2)	0.963	0.927	0.847	0.721	0.378	0.011	-0.150	-0.196
PC	0.810	0.853	0.849	0.809	0.648	0.426	0.255	0.133
Sq-PC	0.825	0.852	0.849	0.821	0.686	0.457	0.251	0.097
PC-Sq	0.798	0.849	0.851	0.820	0.687	0.497	0.304	0.225
kPCA	0.610	0.664	0.672	0.675	0.647	0.562	0.440	-0.035
3PRF	0.913	0.914	0.863	0.802	0.638	0.475	0.402	0.471
k3PRF	0.924	0.937	0.903	0.846	0.674	0.508	0.459	0.390

GDP Deflator

Method	h=1	h=2	h=3	h=4	h=6	h=8	h=10	h=12
AR(2)	0.797	0.774	0.740	0.657	0.418	0.146	0.156	0.077
PC	0.444	0.276	0.056	-0.184	-0.408	-0.347	-0.221	-0.057
Sq-PC	0.299	0.145	-0.035	-0.168	-0.245	-0.230	-0.192	-0.108
PC-Sq	0.431	0.268	0.104	-0.039	-0.106	-0.038	-0.111	-0.182
kPCA	-0.032	0.247	-0.021	0.008	0.003	0.004	0.029	-0.023
3PRF	0.584	0.496	0.426	0.243	0.174	0.279	0.300	0.155
k3PRF	0.667	0.632	0.563	0.476	0.479	0.413	0.197	0.512

CPI

Method	h=1	h=2	h=3	h=4	h=6	h=8	h=10	h=12
AR(2)	0.704	0.706	0.620	0.565	0.397	0.211	0.062	-0.038
PC	0.660	0.535	0.364	0.154	-0.163	-0.252	-0.248	-0.173
Sq-PC	0.410	0.296	0.161	0.049	-0.055	-0.156	-0.200	-0.173
PC-Sq	0.649	0.512	0.353	0.186	-0.019	-0.087	-0.187	-0.228
kPCA	0.440	0.380	-0.050	0.189	-0.043	-0.024	0.042	-0.006
3PRF	0.641	0.566	0.487	0.352	0.192	0.241	0.255	0.141
k3PRF	0.676	0.612	0.541	0.463	0.469	0.434	0.349	0.477

Table 8: Out of Sample R^2 of Labor Market and Price Variables

5.6 Forecasting Housing and Financial Variables

We evaluate the relative performance of our method across several key indicators: Privately Owned Housing Starts (*HStart*), Privately Owned Housing Starts in the Western Census region (*HStart-W*), GS-1 (Treasury Bills), GS-10 (Treasury Notes), and the S&P 500 Index. The first two indicators pertain to the housing market, while the latter three belong to the financial market. These financial variables are listed in ascending order of volatility.

As seen from Table-9, forecasting *HStart*(total) proves to be a difficult problem. While most forecasting methods do not beat the historical average, our method performs better than all other methods at all horizons. It is relatively easy to forecast housing in the western census region, and our method performs better than all other methods except for a few cases. We find similar patterns in financial market variables, thereby omitting discussion.

6 Comprehensive Forecasting Analysis

To enhance the robustness of our empirical analysis, we conducted comparative assessments of our method against competing methods across all 176 series within our dataset. This entailed selecting each series as the target and repeating the comparative analysis for every series in our dataset.

6.1 Description of Comparisons

Our investigation encompasses the comparative performance of models across a total of $176 \times 8 = 1408$ target-horizon combinations. The results of these comparisons, indicating the percentage of instances where a particular method demonstrated superior performance, are presented in Supplementary appendix-B.7. For example, if a method emerged as the best performer in 704 out of 1408 combinations, it would be represented by a value of 50 in the table. Essentially, we list the relative frequency of the occurrence of the best performance of a given method.

While the preceding frequency comparisons provide insight into the number of times

HStart

Method	h=1	h=2	h=3	h=4	h=6	h=8	h=10	h=12
AR(2)	0.048	-0.029	-0.140	-0.216	-0.380	-0.157	-0.131	-0.105
PC	-1.360	-0.799	-0.317	-0.052	0.172	0.259	0.086	0.085
Sq-PC	-1.226	-0.688	-0.196	0.095	0.314	0.453	0.183	0.100
PC-Sq	-1.473	-0.936	-0.371	-0.004	0.278	0.188	-0.176	-0.024
kPCA	-0.199	-0.074	-0.157	0.244	0.408	-0.101	-0.325	0.101
3PRF	0.092	0.272	0.064	-0.223	-0.391	-0.205	-0.220	-0.653
k3PRF	0.138	0.204	0.231	0.245	0.230	0.253	0.116	0.073

HStart-W

Method	h=1	h=2	h=3	h=4	h=6	h=8	h=10	h=12
AR(2)	0.571	0.540	0.495	0.386	0.003	-0.397	-0.402	-0.787
PC	0.326	0.433	0.481	0.516	0.405	0.169	-0.070	-0.182
Sq-PC	0.201	0.318	0.356	0.372	0.184	-0.053	-0.248	-0.323
PC-Sq	0.359	0.402	0.414	0.459	0.310	0.033	-0.135	-0.244
kPCA	0.287	0.336	0.379	0.442	0.447	-0.135	-0.147	-0.062
3PRF	0.571	0.475	0.231	0.084	-0.031	0.094	0.260	0.253
k3PRF	0.586	0.464	0.207	0.554	0.178	0.141	0.160	0.463

GS-1

Method	h=1	h=2	h=3	h=4	h=6	h=8	h=10	h=12
AR(2)	0.915	0.862	0.796	0.645	0.184	-0.270	-0.304	-0.336
PC	0.687	0.487	0.261	0.055	-0.163	-0.124	-0.033	0.139
Sq-PC	0.306	0.201	0.090	-0.012	-0.145	-0.131	-0.074	0.011
PC-Sq	0.674	0.448	0.243	0.059	-0.162	-0.119	0.051	0.163
kPCA	0.635	0.472	0.282	0.119	0.029	-0.018	0.166	0.114
3PRF	0.856	0.735	0.615	0.501	0.449	0.329	0.241	0.349
k3PRF	0.873	0.806	0.782	0.699	0.381	0.224	0.428	0.605

GS-10

Method	h=1	h=2	h=3	h=4	h=6	h=8	h=10	h=12
AR(2)	0.783	0.766	0.667	0.540	0.237	-0.022	0.057	0.136
PC	0.446	0.327	0.148	0.017	-0.122	-0.177	-0.329	-0.378
Sq-PC	0.312	0.247	0.124	0.069	-0.016	-0.065	-0.194	-0.327
PC-Sq	0.421	0.292	0.200	0.155	0.117	0.032	-0.083	-0.608
kPCA	0.457	0.402	-0.098	0.246	-0.039	-0.022	0.082	0.035
3PRF	0.615	0.469	0.268	0.012	0.168	0.403	0.294	0.044
k3PRF	0.621	0.499	0.405	0.401	0.345	0.272	0.161	0.566

S&P 500

Method	h=1	h=2	h=3	h=4	h=6	h=8	h=10	h=12
AR(2)	0.953	0.943	0.912	0.866	0.697	0.456	0.277	0.272
PC	0.388	0.318	0.224	0.121	-0.019	-0.001	0.107	0.201
Sq-PC	0.265	0.214	0.152	0.089	0.023	0.061	0.136	0.192
PC-Sq	0.387	0.287	0.167	0.048	-0.079	0.034	0.220	0.295
kPCA	-0.064	-0.067	-0.039	-0.031	0.094	0.038	0.091	0.558
3PRF	0.706	0.687	0.636	0.566	0.453	0.458	0.489	0.523
k3PRF	0.812	0.791	0.736	0.654	0.565	0.586	0.674	0.781

Table 9: Out of Sample R^2 of Housing and Financial Variables

each method proved superior to others, they do not measure the extent to which the best-performing method surpassed its nearest competitor. In other words, while method A may marginally outperform method B on one forecast horizon, method B might exhibit a considerable advantage over method A on another horizon. Then, the aforementioned frequency comparison may not depict the full picture. To account for this, we introduce a notion of '*Tolerance*' level. We call a method 'best' under tolerance level ϵ if the out-of-sample R^2 of a method is within ϵ percentage lower than the best method's performance [13]. Therefore, for a non-zero tolerance, it is possible to have multiple 'best' methods.

The "All Horizons" set of rows summarizes all 1408 comparisons, i.e. encompassing all horizons and all series. Recognizing that forecast objectives may vary in time horizon, we scrutinize comparative performances in short- and long-run contexts. The "Short-run" rows incorporate horizons $h = 1, 2, 3, 4$, comprising 708 (calculated as 176 × 4) combinations, while the "Long-run" row includes horizons $h = 6, 8, 10, 12$, similarly amounting to 708 combinations. Additionally, the portion labeled as "Excluding AR" excludes the auto-regressive method and compares the remaining methods across all 1408 combinations. For more granular analysis, we report comparative performance numbers for each forecast horizon h. These numbers are reported in the Supplementary appendix-B.7.

It is important to note that multiple 'best' methods may exist for a non-zero tolerance level, resulting in the sums of rows (in Table 5 in Supplementary appendix-B.7) exceeding 100 percent. However, for a tolerance level of zero, the rows sum to 100 percent.

Analysis	Tolerance(%)	Methods						
		AR(2)	**PC**	**Sq-PC**	**PC-Sq**	**kPCA**	**3PRF**	**k3PRF**
All Horizons								
	0	48.22	0.21	0.85	1.42	2.98	6.47	39.56
	5	50.07	1.14	1.35	1.99	3.34	9.16	43.54
	10	52.41	2.27	2.13	3.34	4.26	13.07	48.37
	20	55.68	5.68	3.69	7.74	6.75	23.30	62.57
Short-run								
	0	84.09	0.14	0.43	0.57	0.43	1.70	12.64
	5	87.07	1.42	0.71	1.56	0.57	5.11	18.75
	10	90.77	3.27	1.70	3.84	1.28	9.23	26.14
	20	94.32	8.38	3.41	10.37	3.55	20.03	48.72
Long-run								
	0	12.36	0.28	1.28	2.27	5.54	11.79	66.48
	5	13.07	0.85	1.99	2.41	6.11	13.21	68.32
	10	14.06	1.28	2.56	2.84	7.24	16.90	70.60
	20	17.05	2.98	3.98	5.11	9.94	26.56	76.42
Excluding AR								
	0	-	1.42	1.56	2.84	5.47	13.00	75.71
	5	-	2.84	2.06	4.76	5.75	17.97	78.76
	10	-	5.26	3.27	7.74	7.03	25.99	81.53
	20	-	11.08	5.89	14.35	11.43	41.34	86.08

Table 10: Distribution of Best Forecasting Methods Across All Series (Percentage)

6.2 Results

We present the results in table-10. The findings presented above yield several noteworthy observations. First, it is evident that unsupervised forecasting techniques, including PCR, Squared-PC, PC-squared, and kernel PCA, exhibit inferior performance across the majority of scenarios when compared to our method. Second, our method, kernel 3PRF, demonstrates unequivocal superiority in longer-horizon forecasting endeavors. Third, our method is unequivocally superior across all horizons when autoregressive (AR) method is excluded. Our method does not outperform AR in the short term, but its performance remains competitive, often closely trailing the best short-run autoregressive method. This can be seen by increasing the tolerance level. The instances where our method can be labeled as 'best' increase rapidly as we increase the tolerance level.

[13] For example, if the AR model is the best for a of series y_ℓ and horizon h_0 with a $R^2 = 0.60$. For tolerance=5, another method will also be considered 'best' if its $R^2 \geq 0.60(1 - 5/100) = 0.57$

7 Conclusion

Building upon the three-pass regression filter by Kelly & Pruitt (2015), we introduce a new forecasting method, kernel three-pass regression filter. Through extensive empirical exercises, we show that this approach holds promise as a dependable forecasting tool. Improved performance can be attributed to two noteworthy features of our method. First, it integrates non-linear relationships by transforming input data into a higher-dimensional space, encapsulating its non-linear functions. Second, it operates as a supervised method, effectively filtering out and discarding irrelevant factors while predicting the target variable.

References

Ahn, Seung C, & Horenstein, Alex R. 2013. Eigenvalue ratio test for the number of factors. *Econometrica*, **81**(3), 1203–1227.

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

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

Bai, Jushan, & Ng, Serena. 2008. Forecasting economic time series using targeted predictors. *Journal of Econometrics*, **146**(2), 304–317.

Chamberlain, Gary, & Rothschild, Michael. 1983. Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets. *Econometrica: Journal of the Econometric Society*, 1281–1304.

Exterkate, Peter, Groenen, Patrick JF, Heij, Christiaan, & van Dijk, Dick. 2016. Nonlinear forecasting with many predictors using kernel ridge regression. *International Journal of Forecasting*, **32**(3), 736–753.

Fan, Jianqing, Xue, Lingzhou, & Yao, Jiawei. 2017. Sufficient forecasting using factor models. *Journal of econometrics*, **201**(2), 292–306.

Fan, Jianqing, Lou, Zhipeng, & Yu, Mengxin. 2023. Are latent factor regression and sparse regression adequate? *Journal of the American Statistical Association*, 1–13.

Goulet Coulombe, Philippe, Leroux, Maxime, Stevanovic, Dalibor, & Surprenant, Stéphane. 2022. How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, **37**(5), 920–964.

Hamilton, James D, & Xi, Jin. 2024. *Principal Component Analysis for Nonstationary Series*. Tech. rept. National Bureau of Economic Research.

Huang, Dashan, Jiang, Fuwei, Li, Kunpeng, Tong, Guoshi, & Zhou, Guofu. 2022. Scaled PCA: A new approach to dimension reduction. *Management Science*, **68**(3), 1678–1695.

Kelly, Bryan, & Pruitt, Seth. 2015. The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics*, **186**(2), 294–316.

Kim, Hyun Hak, & Swanson, Norman R. 2014. Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. *Journal of Econometrics*, **178**, 352–367.

Kutateladze, Varlam. 2022. The kernel trick for nonlinear factor modeling. *International Journal of Forecasting*, **38**(1), 165–177.

Onatski, Alexei, & Wang, Chen. 2021. Spurious factor analysis. *Econometrica*, **89**(2), 591–614.

Sermpinis, Georgios, Stasinakis, Charalampos, Theofilatos, Konstantinos, & Karathanasopoulos, Andreas. 2014. Inflation and unemployment forecasting with genetic support vector regression. *Journal of Forecasting*, **33**(6), 471–487.

Stock, James H, & Watson, Mark W. 2002. Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, **20**(2), 147–162.

A Technical Appendix

A.1 Proofs of Theoretical Results

Lemma 1 *Under Assumption(s) 1-3, we have the following*

1. $T^{-1/2}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\omega} = \boldsymbol{O}_p(1)$
2. $T^{-1/2}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\eta} = \boldsymbol{O}_p(1)$
3. $T^{-1/2}\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{\eta} = \boldsymbol{O}_p(1)$
4. $M^{-1/2}\varepsilon_t'\boldsymbol{\Phi} = \boldsymbol{O}_p(1)$
5. $M^{-1}T^{-1}\boldsymbol{\Phi}'\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{F} = \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right)$
6. $M^{-1}T^{-1/2}\boldsymbol{\Phi}'\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{\omega} = \boldsymbol{O}_p\left(1\right)$
7. $M^{-1/2}T^{-1/2}\boldsymbol{\Phi}\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{\eta} = \boldsymbol{O}_p(1)$
8. $M^{-1}T^{-3/2}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{F} = \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right)$

9. $M^{-1}T^{-3/2}\boldsymbol{\omega}'\boldsymbol{J}_T\varepsilon\boldsymbol{\varepsilon}'\boldsymbol{J}_T\boldsymbol{F} = \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right)$

10. $M^{-1}T^{-3/2}\boldsymbol{\omega}'\boldsymbol{J}_T\varepsilon\varepsilon'\boldsymbol{J}_T\boldsymbol{\omega} = \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right)$

11. $M^{-1}T^{-1/2}\boldsymbol{F}'\boldsymbol{J}_T\varepsilon\boldsymbol{\varepsilon}_t = \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right)$

12. $M^{-1}T^{-1/2}\boldsymbol{\omega}'\boldsymbol{J}_T\varepsilon\varepsilon_t = \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right)$

13. $M^{-1}T^{-3/2}\boldsymbol{\eta}'\boldsymbol{J}_T\varepsilon\varepsilon'\boldsymbol{J}_T\boldsymbol{F} = \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right)$

14. $M^{-1}T^{-3/2}\boldsymbol{\eta}'\boldsymbol{J}_T\varepsilon\varepsilon'\boldsymbol{J}_T\boldsymbol{F} = \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right)$

15. $T^{-1/2}\sum_t \eta_{t+h} = O_p(1)$

Proof: Proof can be seen from Kelly & Pruitt (2015), Lemma 2 in their appendix. The only difference is the omission of the matrix $\boldsymbol{J}_N$ in the various expressions. This, however, doesn't affect the rates, as can be verified from their proofs. We do not allow an intercept in pass-2 because doing so will require demeaning of the transformed predictor(s), which is not feasible.

Lemma 2 *Under Assumption(s) 1-5, we have the following*

1. $M^{-1}T^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{x}_t) = \boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{F}_t + \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right)$

2. $M^{-1}T^{-2}\boldsymbol{Z}'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{J}_T\boldsymbol{y} = \boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\beta} + \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right)$

3. $M^{-2}T^{-3}\boldsymbol{Z}'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{J}_T\boldsymbol{Z} = \boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' + \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right)$

Proof: The Proof follows directly by writing out the expressions. Item 1

$$
\begin{aligned}
M^{-1}T^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{x}_t) &= \boldsymbol{\Lambda}\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(M^{-1}\boldsymbol{\Phi}'\boldsymbol{\Phi}\right)\boldsymbol{F}_t + \boldsymbol{\Lambda}\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(M^{-1}\boldsymbol{\Phi}'\boldsymbol{\varepsilon}_t\right) \\
&+ \boldsymbol{\Lambda}\left(M^{-1}T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{\Phi}\right)\boldsymbol{F}_t + \boldsymbol{\Lambda}\left(M^{-1}T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\varepsilon_t\right) \\
&+ \left(T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(M^{-1}\boldsymbol{\Phi}'\boldsymbol{\Phi}\right)\boldsymbol{F}_t + \left(T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(M^{-1}\boldsymbol{\Phi}'\boldsymbol{\varepsilon}_t\right) \\
&+ \left(M^{-1}T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{\Phi}\right)\boldsymbol{F}_t + \left(M^{-1}T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\varepsilon\boldsymbol{\varepsilon}_t\right) \\
&= \boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{F}_t + \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right)
\end{aligned}
$$

The final line follows directly from Lemma 1 and Assumptions 2.1 and 2.2.

Item 2:

$$
\begin{aligned}
& M^{-1}T^{-2}\boldsymbol{Z}'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{J}_T\boldsymbol{y} = \boldsymbol{\Lambda}\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(M^{-1}\boldsymbol{\Phi}'\boldsymbol{\Phi}\right)\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\boldsymbol{\beta} \\
& + \boldsymbol{\Lambda}\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(M^{-1}\boldsymbol{\Phi}'\boldsymbol{\Phi}\right)\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\eta}\right) + \boldsymbol{\Lambda}\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(M^{-1}T^{-1}\boldsymbol{\Phi}'\varepsilon'\boldsymbol{J}_T\boldsymbol{F}\right)\boldsymbol{\beta} \\
& + \boldsymbol{\Lambda}\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(M^{-1}T^{-1}\boldsymbol{\Phi}'\varepsilon'\boldsymbol{J}_T\boldsymbol{\eta}\right) + \boldsymbol{\Lambda}\left(M^{-1}T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{\Phi}\right)\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\boldsymbol{\beta} \\
& + \boldsymbol{\Lambda}\left(M^{-1}T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{\Phi}\right)\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\eta}\right) + \boldsymbol{\Lambda}\left(M^{-1}T^{-2}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\varepsilon'\boldsymbol{J}_T\boldsymbol{F}\right)\boldsymbol{\beta} \\
& + \boldsymbol{\Lambda}\left(M^{-1}T^{-2}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\varepsilon'\boldsymbol{J}_T\boldsymbol{\eta}\right) + \left(T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(M^{-1}\boldsymbol{\Phi}'\boldsymbol{\Phi}\right)\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\boldsymbol{\beta} \\
& + \left(T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(M^{-1}\boldsymbol{\Phi}'\boldsymbol{\Phi}\right)\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\eta}\right) + \left(T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(M^{-1}T^{-1}\boldsymbol{\Phi}'\varepsilon'\boldsymbol{J}_T\boldsymbol{F}\right)\boldsymbol{\beta} \\
& + \left(T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{F}\right)\left(M^{-1}T^{-1}\boldsymbol{\Phi}'\varepsilon'\boldsymbol{J}_T\boldsymbol{\eta}\right) + \left(M^{-1}T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{\Phi}\right)\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{F}\right)\boldsymbol{\beta} \\
& + \left(M^{-1}T^{-1}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\boldsymbol{\Phi}\right)\left(T^{-1}\boldsymbol{F}'\boldsymbol{J}_T\boldsymbol{\eta}\right) + \left(M^{-1}T^{-2}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\varepsilon'\boldsymbol{J}_T\boldsymbol{F}\right)\boldsymbol{\beta} \\
& + \left(M^{-1}T^{-2}\boldsymbol{\omega}'\boldsymbol{J}_T\boldsymbol{\varepsilon}\varepsilon'\boldsymbol{J}_T\boldsymbol{\eta}\right) \\
& = \boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\beta} + \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right)
\end{aligned}
$$

The final line follows directly from Lemma 1 and Assumptions 2.1 and 2.2.

Item 3: Let $= \hat{\boldsymbol{F}}_{C,t} = M^{-1}T^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{x}_t)$. Then, given Lemma 2.1, standard arguments would imply that $M^{-2}T^{-3}\boldsymbol{Z}'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{J}_T\boldsymbol{Z} = \frac{\hat{\boldsymbol{F}}_C\boldsymbol{J}_T\hat{\boldsymbol{F}}_C'}{T}$ $= \boldsymbol{\Lambda\Delta}_F\mathcal{P}\left(T^{-1}\boldsymbol{F}\boldsymbol{J}_T\boldsymbol{F}\right)\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' + \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right)$. Given Assumption 2.1, we have that $\boldsymbol{\Lambda\Delta}_F\mathcal{P}\left(T^{-1}\boldsymbol{F}\boldsymbol{J}_T\boldsymbol{F}\right)\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' = \boldsymbol{\Lambda\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' + \boldsymbol{O}_p\left(T^{-1/2}\right)$.

Therefore, we have that, $M^{-2}T^{-3}\boldsymbol{Z}'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{J}_T\boldsymbol{Z} = \boldsymbol{\Lambda\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' + \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right) + \boldsymbol{O}_p\left(T^{-1/2}\right) = \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right)$.

Theorem 1 If Assumption 1-5 hold, we have

$$\hat{\boldsymbol{F}}_t - \boldsymbol{H}_f\boldsymbol{f}_t = \boldsymbol{O}_p(\delta_{MT}^{-1})$$

where $\boldsymbol{H}_f \equiv \hat{\boldsymbol{F}}_A\hat{\boldsymbol{F}}_B^{-1}\boldsymbol{\Lambda\Delta}_F\mathcal{P}$

$\hat{\boldsymbol{F}}_A = T^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}$ and

$\hat{\boldsymbol{F}}_B = M^{-1}T^{-2}\boldsymbol{Z}'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X}')\boldsymbol{J}_T\boldsymbol{Z}$

Proof:

$$\begin{aligned}
\hat{\boldsymbol{F}}_t &= T^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\left(M^{-1}T^{-2}\boldsymbol{Z}'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1}M^{-1}T^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{x}_t)\\
&= \hat{\boldsymbol{F}}_A\hat{\boldsymbol{F}}_B^{-1}\left(\boldsymbol{\Lambda\Delta}_F\mathcal{P}\boldsymbol{F}_t + \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right)\right)\\
&= \hat{\boldsymbol{F}}_A\hat{\boldsymbol{F}}_B^{-1}\boldsymbol{\Lambda\Delta}_F\mathcal{P}\boldsymbol{F}_t + \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right)\\
&= \boldsymbol{H}_f\boldsymbol{f}_t + \boldsymbol{O}_p(\delta_{MT}^{-1})
\end{aligned}$$

The second equality follows from Lemma 2.1 and the final equality uses the definition of $\boldsymbol{H}_f$.

Theorem 2 If Assumption 1-5 hold, we have

$$\hat{\boldsymbol{\beta}} - \boldsymbol{G}_\beta \boldsymbol{\beta} = \boldsymbol{O}_p(\delta_{MT}^{-1}).$$

where $\boldsymbol{G}_\beta \equiv \hat{\boldsymbol{\beta}}_1^{-1} \hat{\boldsymbol{\beta}}_2 \left[\boldsymbol{\Lambda}\boldsymbol{\Delta}_F \mathcal{P} \boldsymbol{\Delta}_F \mathcal{P} \boldsymbol{\Delta}_F \boldsymbol{\Lambda}'\right]^{-1} \boldsymbol{\Lambda}\boldsymbol{\Delta}_F \mathcal{P} \boldsymbol{\Delta}_F$,
$\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{F}}_A$ and $\hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{F}}_B$

Proof:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(T^{-1}\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1} M^{-1}T^{-2}\boldsymbol{Z}'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{J}_T\boldsymbol{Z} \\
&\times \left(M^{-2}T^{-3}\boldsymbol{Z}'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{J}_T\boldsymbol{Z}\right)^{-1} M^{-1}T^{-2}\boldsymbol{Z}'\boldsymbol{J}_T\varphi(\boldsymbol{X})\varphi(\boldsymbol{X})'\boldsymbol{J}_T\boldsymbol{y} \\
&= \hat{\boldsymbol{\beta}}_1^{-1}\hat{\boldsymbol{\beta}}_2 \left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' + \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right)\right)^{-1}\left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\beta} + \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right)\right) \\
&= \hat{\boldsymbol{\beta}}_1^{-1}\hat{\boldsymbol{\beta}}_2 \left[\left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\right)^{-1} - \left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\right)^{-1}\boldsymbol{O}_p\left(\delta_{MT}^{-1}\right)\left(\boldsymbol{O}_p(1) + \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right)\right)^{-1}\right] \times \\
&\left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\beta} + \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right)\right) \\
&= \hat{\boldsymbol{\beta}}_1^{-1}\hat{\boldsymbol{\beta}}_2 \left[\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\right]^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\beta} + \boldsymbol{O}_p\left(\delta_{MT}^{-1}\right) \\
&= \boldsymbol{G}_\beta\boldsymbol{\beta} + \boldsymbol{O}_p(\delta_{MT}^{-1})
\end{aligned}$$

where the second equality employs Lemma 2.2 and 2.3. The third equality uses the fact that for any invertible matrices $\boldsymbol{A}$ and $\boldsymbol{A}+\boldsymbol{B}$ we have $(\boldsymbol{A}+\boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{A}+\boldsymbol{B})^{-1}$, which in our case implies that,
$\left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' + \boldsymbol{O}_p(\delta_{MT}^{-1})\right)^{-1} =$
$\left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\right)^{-1} - \left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}'\right)^{-1}\boldsymbol{O}_p(\delta_{MT}^{-1})\left(\boldsymbol{\Lambda}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\mathcal{P}\boldsymbol{\Delta}_F\boldsymbol{\Lambda}' + \boldsymbol{O}_p(\delta_{MT}^{-1})\right)^{-1}$.
The last equality uses the definition of $\boldsymbol{G}_\beta$.

Theorem 3 If Assumption 1-5 hold, we have

$$\hat{y}_{t+h} - \mathbb{E}_t y_{t+h} = O_p(\delta_{MT}^{-1})$$

Proof:

$$\begin{aligned}
\hat{y}_{t+h} &= \bar{y} + \boldsymbol{J}_T \hat{\boldsymbol{F}}'_T \hat{\boldsymbol{\beta}} \\
&= \beta_0 + \bar{\boldsymbol{f}}'\boldsymbol{\beta}_f + O_p(T^{-1/2}) + \left(\boldsymbol{H}_f \boldsymbol{f}_t + \boldsymbol{O}_p(\Gamma_{NT}^{-1})\right)' \left(\boldsymbol{G}_\beta \boldsymbol{\beta} + \boldsymbol{O}_p(\delta_{MT}^{-1})\right) \\
&= \beta_0 + \bar{\boldsymbol{f}}'\boldsymbol{\beta}_f + \left(\boldsymbol{f}_t - \bar{\boldsymbol{f}}\right)' \boldsymbol{H}'_f \boldsymbol{G}_\beta \boldsymbol{\beta} + \boldsymbol{O}_p(\delta_{MT}^{-1}) \\
&= \beta_0 + \bar{\boldsymbol{f}}'\boldsymbol{\beta}_f + \left(\boldsymbol{f}_t - \bar{\boldsymbol{f}}\right)' \boldsymbol{\beta} + \boldsymbol{O}_p(\delta_{MT}^{-1}) \\
&= \beta_0 + \boldsymbol{f}'_t \boldsymbol{\beta} + \boldsymbol{O}_p(\delta_{MT}^{-1}) \\
&= \mathbb{E}_t y_{t+h} + O_p(\delta_{MT}^{-1})
\end{aligned}$$

The second equality follows from lemma 1.15. The fourth equality follows if $\boldsymbol{H}'_f \boldsymbol{G}_\beta$ is an identity matrix. This is indeed true since $\boldsymbol{H}'_f \boldsymbol{G}_\beta = \mathcal{P}\boldsymbol{\Delta}_F \boldsymbol{\Lambda}' \left[\boldsymbol{\Lambda}\boldsymbol{\Delta}_F \mathcal{P}\boldsymbol{\Delta}_F \mathcal{P}\boldsymbol{\Delta}_F \boldsymbol{\Lambda}'\right]^{-1} \boldsymbol{\Lambda}\boldsymbol{\Delta}_F \mathcal{P}\boldsymbol{\Delta}_F$. Using the arguments as in Lemma 5 and Theorem 1 of Kelly & Pruitt (2015) the RHS is an identity matrix, given assumptions 4 and 5.

A.2 Mercer's Theorem

Suppose $\mathcal{X} \subseteq \mathbb{R}^d$ is compact and kernel function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is continuous, satisfying the following conditions,

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

where $L^2(\mathcal{X}) = \left\{h : \int h^2(\mathrm{s})d\mathrm{s} < \infty\right\}$, then there exist functions $\{\varphi_i(\cdot) \in L^2(\mathcal{X}), i = 1, 2, \ldots\}$ and non-negative coefficients $\theta_1 \geq \theta_2 \geq \ldots \geq 0$ which together forms an orthonormal system in $L^2(\mathcal{X})$, i.e. $\langle \varphi_i, \varphi_j \rangle_{L^2(\mathcal{X})} = \int \varphi_i(\mathrm{x})\varphi_j(\mathrm{x})d\mathrm{x} = \mathbb{I}_{\{i=j\}}$, such that

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

B Supplementary Appendix: Algorithms, Data, and Figures

This appendix provides algorithmic details, data sources, transformation, and visual plots.

B.1 Some Popular Kernel Methods and Their Working

Many popular kernel functions exist, such as polynomial, Gaussian, and sigmoid kernels. We illustrate two of them to show that kernel function can represent the products of feature map $\varphi(\cdot)$.

Polynomial Kernel Let the functional mapping where $\varphi(a)$ includes a fixed term, all variables $a_1, a_2, \ldots, a_N$, and their respective squares and cross products. The kernel function $\mathcal{K}(a, b)$ assumes a simplified structure if we scale the linear and cross-product terms in $\varphi(a)$ by the constant $\sqrt{2}$. In other words, if we define

$$\begin{aligned}\varphi(a) = \Big(&1, \sqrt{2}a_1, \sqrt{2}a_2, \ldots, \sqrt{2}a_N, a_1^2, a_2^2, \\ &\ldots, a_N^2, \sqrt{2}a_1a_2, \sqrt{2}a_1a_3, \ldots, \sqrt{2}a_{N-1}a_N\Big)',\end{aligned}$$

Then, the corresponding kernel function becomes:

$$\begin{aligned}\mathcal{K}(a,b) =& \varphi(a)'\varphi(b)\\ &= 1 + 2\left(a_1b_1 + a_2b_2 + \cdots + a_Nb_N\right) + \left(a_1^2b_1^2 + a_2^2b_2^2 + \cdots + a_N^2b_N^2\right)\\ &\quad + 2\left(a_1a_2b_1b_2 + a_1a_3b_1b_3 + \cdots + a_{N-1}a_Nb_{N-1}b_N\right)\\ &= 1 + 2\left(a_1b_1 + a_2b_2 + \cdots + a_Nb_N\right) + \left(a_1b_1 + a_2b_2 + \cdots + a_Nb_N\right)^2\\ &= 1 + 2a'b + \left(a'b\right)^2 = \left(1 + a'b\right)^2\end{aligned}$$

This kernel can be generalized to a general degree d by keeping the terms of degree at most d in the expression of $\varphi(a)$. This example is also discussed in Exterkate *et al.* (2016).

Gaussian Kernel This kernel is an example of an infinite-dimensional kernel. Let $\mathrm{x}, \mathrm{z} \in \mathbb{R}^k$ and $\mathcal{K}(\mathrm{x},\mathrm{z}) = e^{-\gamma\|\mathrm{x}-\mathrm{z}\|^2}$. Then, through the Taylor expansion, we can write

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

That is, $\varphi_j(\mathrm{x}) = \sum_{\sum_{i=1}^k n_i=j}(2\gamma)^{j/2}e^{-\gamma\|\mathrm{x}\|^2}\prod_{i=1}^k\frac{x_i^{n_i}}{\sqrt{n_i!}}, \quad j = 0,\ldots,\infty$. Kutateladze (2022) use this kernel function in their paper which is based on kernel PCA.

B.2 Performance Metric: Out of Sample R^2

We employ out-of-sample R^2 relative to the historical mean as our performance metric to assess various forecasting methods alongside our own. Out-of-sample R^2 indicates goodness of fit on unseen data, providing insights into the predictive accuracy of a model. Mathematically, out-of-sample R^2 is computed as follows:

$$R^2 = 1 - \frac{\sum_{i \in \text{test-data}} (y_i - \hat{y}_i)^2}{\sum_{i \in \text{test-data}} (y_i - \bar{y}_{\text{train}})^2}$$

Here, the numerator quantifies the squared deviation between the model's predictions and the true values in the test data. At the same time, the denominator measures the deviation of the true values from the historical mean in the test data. It is important to note that we utilize the mean of the training data for the historical mean, as in real-world forecasting scenarios, access to the training mean is typically available.

It is noteworthy that out-of-sample R^2 ranges from $-\infty$ to 1, unlike in-sample R^2, which ranges from zero to one. A positive out-of-sample R^2 indicates that the forecasting method outperforms the historical average. At the same time, a negative value suggests that the forecasting method performs worse than a simple method that forecasts y_i equal to the historical average. We adopt the rolling window method to compute out-of-sample R^2, consistent with standard practices in the literature. Appendix B.3 provides a detailed exposition of our methodology.

B.3 Out of Sample Estimation

We train our model on in-sample information and then construct a sample forecast, as discussed in the algorithm below.

We have demonstrated the construction of an out-of-sample forecast (Table-11).

Step	Description
1	Take in-sample data $\{\boldsymbol{X}_{in}, \boldsymbol{y}_{in}\}$ out-of-sample predictor matrix $\boldsymbol{X}_{out}$ and proxy matrix $\boldsymbol{Z}$.
2	Compute the following two kernel matrices: $\mathbf{K}_{in} = \mathcal{K}(\boldsymbol{X}_{in}, \boldsymbol{X}_{in})$ and $\mathbf{K}_{out} = \mathcal{K}(\boldsymbol{X}_{in}, \boldsymbol{X}_{out})$
3	Estimate in and out of the sample factor matrix using the following formula: $\widehat{\boldsymbol{F}}_{in} = \big(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\big)\big(\boldsymbol{Z}'\boldsymbol{J}_T\mathbf{K}_{in}\boldsymbol{J}_T\boldsymbol{Z}\big)^{-1}\big(\boldsymbol{Z}'\boldsymbol{J}_T\mathbf{K}_{in}\big)$ $\widehat{\boldsymbol{F}}_{out} = \big(\boldsymbol{Z}'\boldsymbol{J}_T\boldsymbol{Z}\big)\big(\boldsymbol{Z}'\boldsymbol{J}_T\mathbf{K}_{in}\boldsymbol{J}_T\boldsymbol{Z}\big)^{-1}\big(\boldsymbol{Z}'\boldsymbol{J}_T\mathbf{K}_{out}\big)$ To accommodate the intercept term in pass-3, Compute a modified factor matrix estimate $\widetilde{\boldsymbol{F}}_{in} = [\mathbf{1} \quad \widehat{\boldsymbol{F}}_{in}]$ and $\widetilde{\boldsymbol{F}}_{out} = [\mathbf{1} \quad \widehat{\boldsymbol{F}}_{out}]$, where $\mathbf{1}$ is a vector of ones.
4	Estimate $\boldsymbol{\beta}$ using the following formula: $\widehat{\boldsymbol{\beta}} = \left(\widetilde{\boldsymbol{F}}'_{in}\widetilde{\boldsymbol{F}}_{in}\right)^{-1}\widetilde{\boldsymbol{F}}_{in}\boldsymbol{y}_{in}$ (this $\widehat{\boldsymbol{\beta}}$ contains intercept term as well and is estimated in-sample)
5	Obtain out-of-sample forecast: $\widehat{\boldsymbol{y}}_{t+h} = \widetilde{\boldsymbol{F}}_{out}\widehat{\boldsymbol{\beta}}$

Table 11: Algorithm: The Out-of-sample forecast by Kernel Three Pass Regression Filter

Now, we outline the rolling window procedure to obtain the out-of-sample forecast performance measured by out-of-sample R^2 in Table-12.

B.4 Data Source and Description

We use FRED-QD data. This section provides the codes of the variables we forecast in our empirical work. For detailed description details, refer to FRED website. In table-13, FRED means federal reserve economic data, and SW stands for Stock and Watson datasets.

B.5 Hyper-parameter Tuning Algorithm

The following table demonstrates our algorithm to tune hyper-parameters σ.

Step	Description
1	**Get Input Data and Parameters** We forecast h period(s) ahead w is the number of training observations. Get $T \times N$ matrix $\mathbf{X}$: matrix of predictors, and $T \times 1$ vector $\mathbf{y}$: target series.
2	**Run Rolling Windows** **Loop Begins**: j from 1 to $test\_size$ i) Set training and test using as follows: $\mathbf{y}_{train} = \mathbf{y}[(j+h):(j+w+h-1)]$ $\mathbf{X}_{train} = \mathbf{X}[j:(w+j-1)]$ and $\mathbf{X}_{test} = \mathbf{X}[(w+j)]$ ii) Train the model on $\{\mathbf{X}_{train}, \mathbf{y}_{train}\}$. Obtain $\tilde{F}_{oos}$ and $\hat{\beta}_{in}$ iii) Obtain the forecast $\widehat{y} = \tilde{F}'_{oos}\hat{\beta}_{in}$ iv) Obtain $\mathbf{y}_{pred}[j] = \widehat{y}$, $\mathbf{y}_{oos}[j] = \mathbf{y}[j+w+h]$, and $\mathbf{y}_{mean}[j] = mean(\mathbf{y}_{train})$ **Loop Ends**
3	**Compute Out-of-sample** R^2: i) Calculate the sum of squared residuals of the model $SSR_{model} = \sum_{j=1}^{test\_size} \Big(\mathbf{y}_{oos}[j] - \mathbf{y}_{pred}[j]\Big)^2$ ii) Get sum of squared residuals of historical mean $SSR_{hist} = \sum_{j=1}^{test\_size} \Big(\mathbf{y}_{oos}[j] - \mathbf{y}_{mean}[j]\Big)^2$ iii) Obtain out of sample R^2: $R^2 = 1 - \frac{SSR_{model}}{SSR_{hist}}$

Table 12: Rolling Window Procedure to Calculate Out of Sample R^2

	FRED Mnemonic	SW Mnemonic	Description
Macro			
	GDPC1	GDP	Real Gross Domestic Product, 3 Decimal (Billions of Chained 2012 Dollars)
	PCECC96	Consumption	Real Personal Consumption Expenditures (Billions of Chained 2012 Dollars)
	EXPGSC1	Exports	Real Exports of Goods & Services, 3 Decimal (Billions of Chained 2012 Dollars)
	IMPGSC1	Imports	Real Imports of Goods & Services, 3 Decimal (Billions of Chained 2012 Dollars)
	GPDIC1	Investment	Real Gross Private Domestic Investment, 3 decimal (Billions of Chained 2012 Dollars)
	FPIx	FixedInv	Real private fixed investment (Billions of Chained 2012 Dollars), deflated using PCE
	IPFINAL	IP:Final products	Industrial Production: Final Products (Market Group) (Index 2012=100)
Labor			
	PAYEMS	Emp:Nonfarm	All Employees: Total nonfarm (Thousands of Persons)
	UNRATE	Unemp Rate	Civilian Unemployment Rate (Percent)
Housing			
	HOUST	Hstarts	Housing Starts: Total: New Privately Owned Housing Units Started (Thousands of Units)
	HOUSTW	Hstarts:W	Housing Starts in West Census Region (Thousands of Units)
Price			
	GDPCTPI	GDP Defl	Gross Domestic Product: Chain-type Price Index (Index 2012=100)
	CPIAUCSL	CPI	Consumer Price Index for All Urban Consumers: All Items (Index 1982-84=100)
Finance			
	GS1	TB-1YR	1-Year Treasury Constant Maturity Rate(%)
	GS10	TB-10YR	10-Year Treasury Constant Maturity Rate (%)
	S&P 500		S&P's Common Stock Price Index: Composite

Table 13: Variable Mnemonic and Description

```
Take an appropriate range of $\sigma$ say $\sigma \in \{0.001, 0.002, 0.003, ..., 14.998, 14.999, 15\}$.
For each value of $\sigma_j$ do the following:
    0. Initialize two variables $R^2_{best} = 0$ and $\sigma_{best} = 0.001$
    1. Take training input data $\{X_{train}, y_{train}\}$ and split it into two halves:
         $\{X_{train1}, y_{train1}\}$ and $\{X_{train2}, y_{train2}\}$.
         One half works as a training set, and the other as a validation set.
    2. i) For given $\sigma_j$, train the model on $\{X_{train1}, y_{train1}\}$ and
         obtain forecast $\hat{y}_{t+h}$ on $\{X_{train2}, y_{train2}\}$.
      ii) Obtain $R^2$ from comparison of $\hat{y}_{t+h}$ and $y_{t+h}$ and call it $R^2_1$.
      iii) Repeat the procedure by flipping training and validation sets and obtain $R^2_2$.
      iv) Obtain $R^2_{\sigma_j} = \frac{R^2_1 + R^2_2}{2}$. If $R^2_{\sigma_j} > R^2_{best}$, update $\sigma_{best} = \sigma_j$ and $R^2_{best} = R^2_{\sigma_j}$.
    3. Repeat the step-1 and step-2 for all value of $\sigma_j$ and return the $\sigma_{best}$.
```

Table 14: Cross-Validation Based Hyper-Parameter Tuning Algorithm

We employ a two-fold cross-validation approach to optimize the hyperparameters. While widely used, traditional K-fold cross-validation is suboptimal for time series data due to its inherent sequential structure. Instead, for our primary analysis, we adopt a rolling window methodology. However, we resort to a fixed-window two-fold cross-validation strategy to mitigate computational expenses. Notably, we compared the computational costs and performance gains of the rolling-window tuning algorithm and the two-fold cross-validation approach.

B.6 Comparative Forecast Performance

We plot the forecasts using our method and the 3PRF method with the true value of the target series for all sixteen series discussed in the empirical application section. To save some space, we only show the plots for one, four, eight, and twelve period ahead forecasts.

B.7 Comparative Performance on All Series For Each Horizons

Figure 5: One Period Ahead Forecasting: Comparative Performance

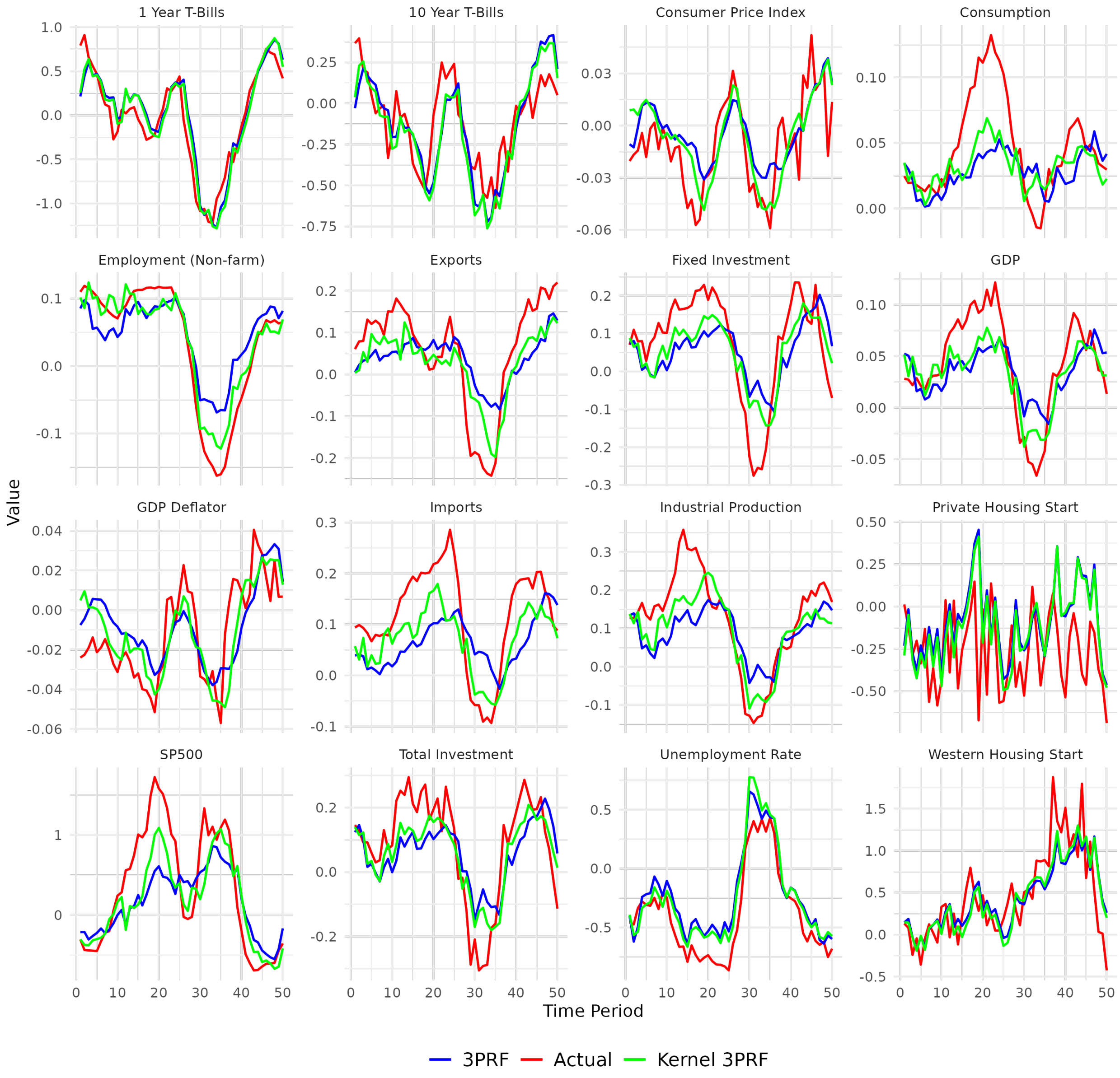


Figure 6: Four Period Ahead Forecasting: Comparative Performance

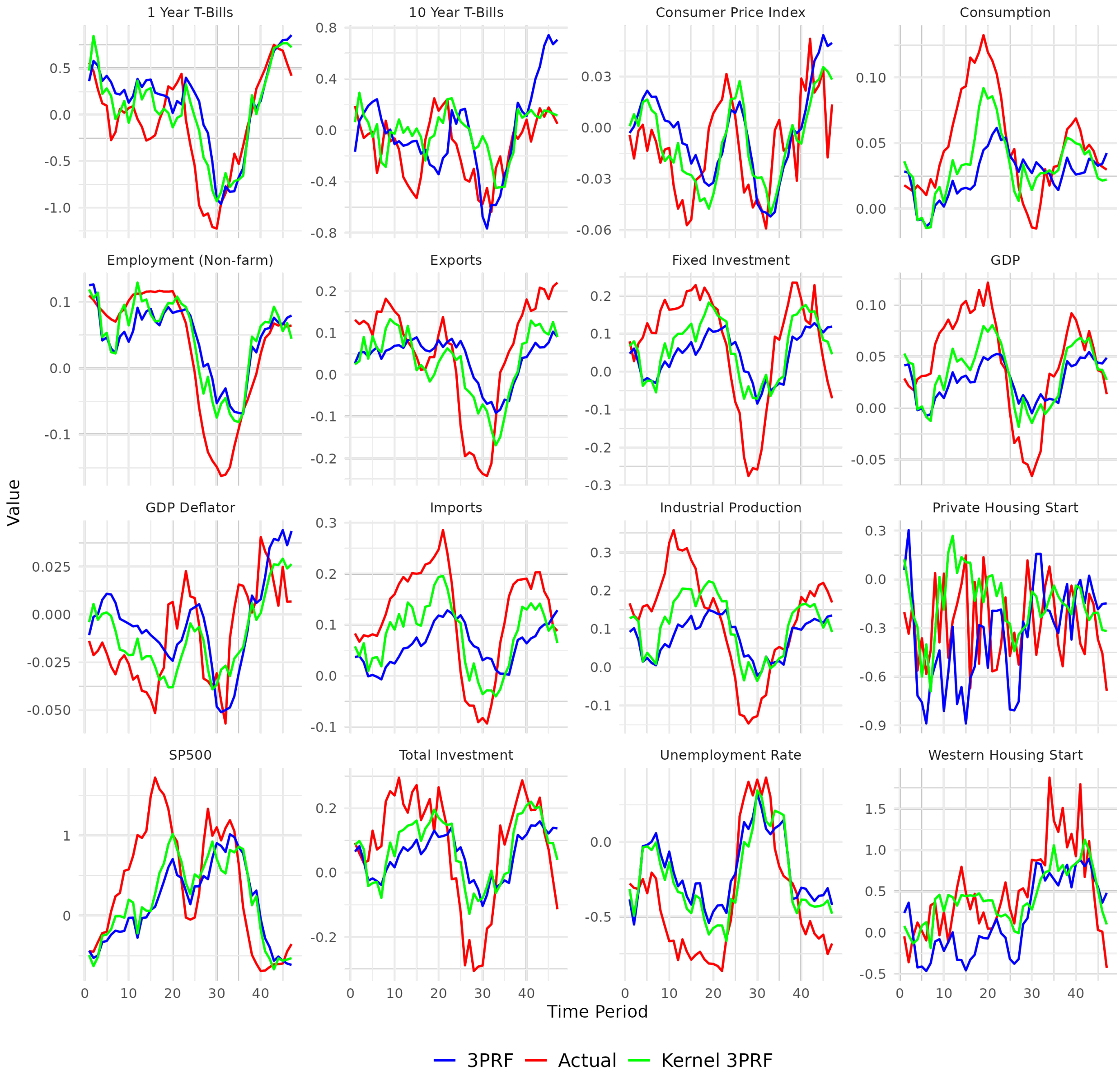


Figure 7: Eight Period Ahead Forecasting: Comparative Performance

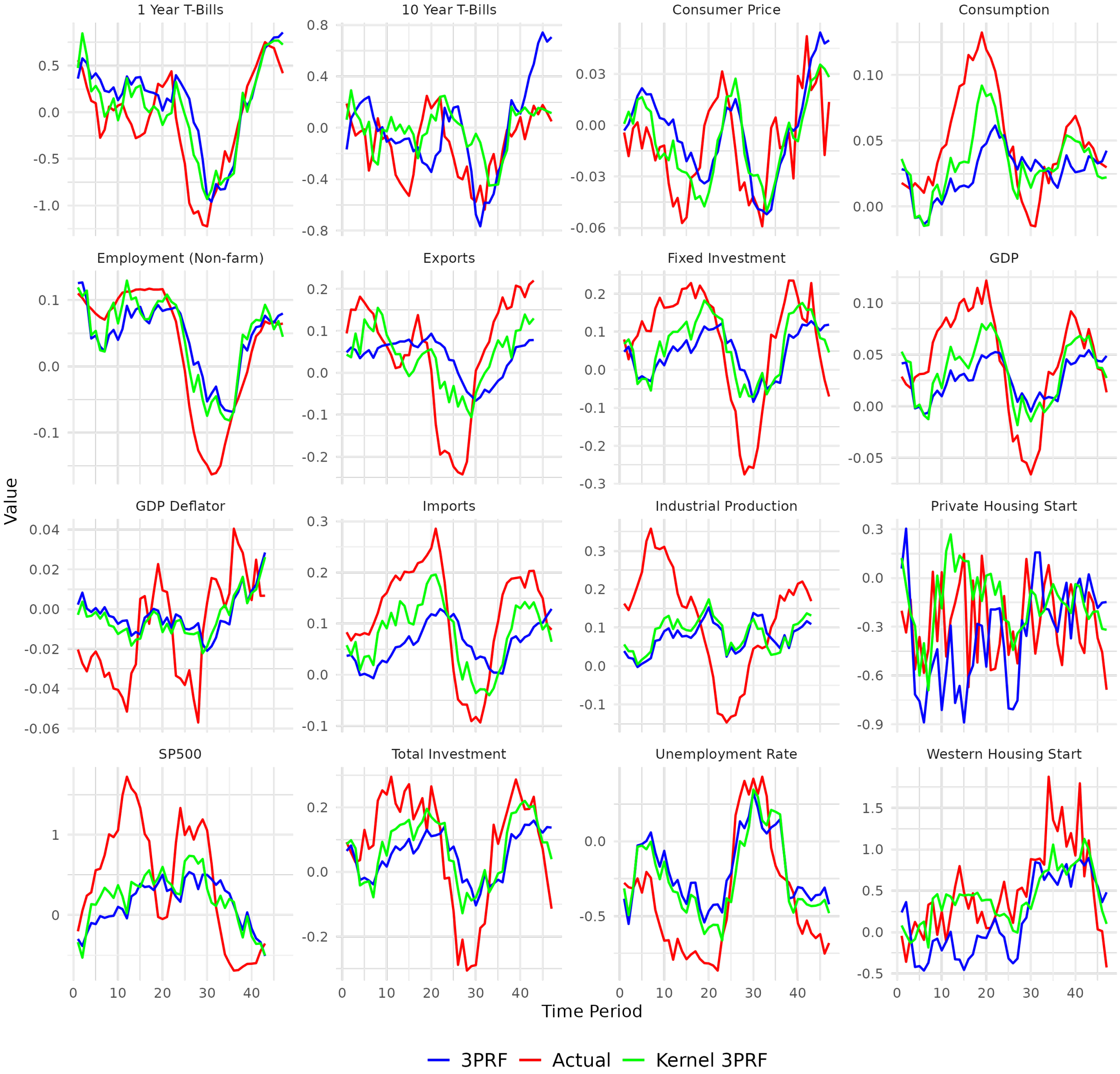
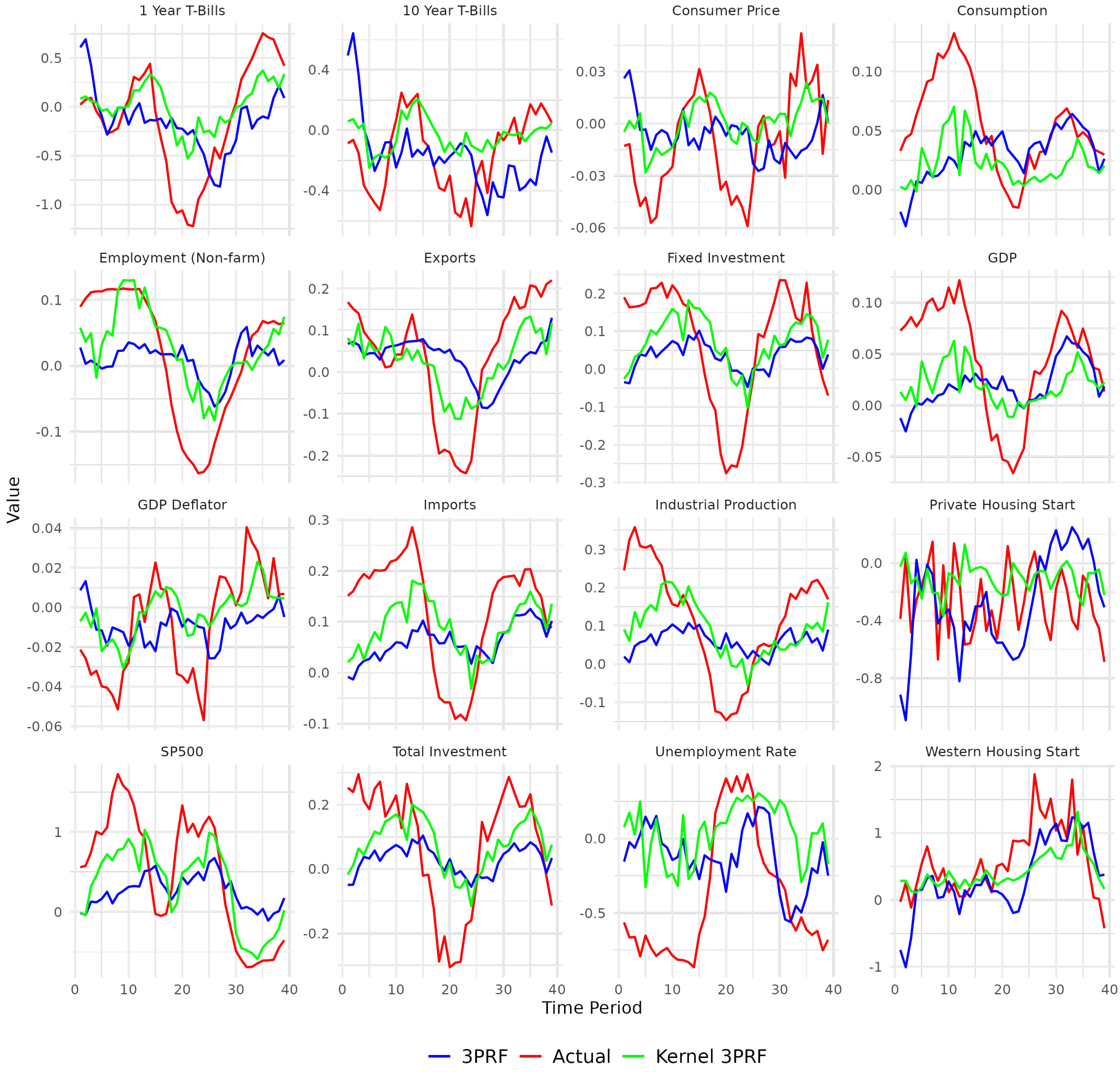


Figure 8: Twelve Period Ahead Forecasting: Comparative Performance



Analysis	**Tolerance(%)**	**Methods**						
		AR(2)	**PCA**	**Sq-PC**	**PC-Sq**	**kPCA**	**3PRF**	**k3PRF**
h=1								
	0	93.75	0.00	0.57	0.00	0.00	0.57	5.11
	5	95.45	1.70	0.00	0.57	0.57	5.11	11.93
	10	97.73	4.55	0.57	3.41	2.27	13.07	23.86
	20	97.73	13.07	3.41	14.20	4.55	23.86	49.43
h=2								
	0	93.75	0.00	0.00	0.57	0.57	1.70	3.41
	5	95.45	1.14	0.00	1.70	0.57	4.55	8.52
	10	95.45	2.84	1.70	3.98	1.70	6.82	13.64
	20	96.59	9.09	2.84	10.80	4.55	16.48	40.91
h=3								
	0	84.09	0.57	0.57	0.57	0.00	2.27	11.93
	5	89.20	1.70	1.14	2.27	0.00	3.98	19.89
	10	93.18	2.84	2.27	4.55	0.00	7.39	27.27
	20	94.89	6.25	3.41	9.66	1.70	17.61	47.16
h=4								
	0	64.77	0.00	0.57	1.14	1.14	2.27	30.11
	5	68.18	1.14	1.70	1.70	1.14	6.82	34.66
	10	76.70	2.84	2.27	3.41	1.14	9.66	39.77
	20	88.07	5.11	3.98	6.82	3.41	22.16	57.39
h=6								
	0	27.84	0.00	2.84	0.00	8.52	6.25	53.98
	5	29.55	0.00	3.98	1.70	9.66	7.95	58.52
	10	32.39	1.14	5.11	2.84	10.80	14.77	61.36
	20	40.91	5.11	6.82	7.39	14.20	25.00	70.45
h=8								
	0	9.09	0.57	1.14	2.27	7.39	9.66	69.89
	5	9.66	2.27	1.70	2.84	8.52	11.36	70.45
	10	10.23	2.84	1.70	2.84	11.36	15.34	72.16
	20	11.36	3.41	4.55	5.11	13.07	28.41	78.98
h=10								
	0	8.52	0.00	0.57	2.84	3.98	18.18	65.91
	5	8.52	0.00	1.14	2.27	3.98	19.89	67.61
	10	9.09	0.00	1.70	2.27	3.98	22.73	68.75
	20	9.66	1.14	2.27	3.98	7.95	29.55	73.30
h=12								
	0	3.98	0.57	0.57	3.41	2.27	13.07	76.14
	5	4.55	1.14	1.14	2.84	2.27	13.64	76.70
	10	4.55	1.14	1.70	3.41	2.84	14.77	80.11
	20	6.25	2.27	2.27	3.98	4.55	23.30	82.95

Table 15: Distribution of Best Forecasting Methods Across All Series in Our Data (Percentage)