University of Bonn

Research Module in

Econometrics and Statistics

TITLE OF YOUR PAPER

January 13, 2023

Lindi Li 3460570

Gewei Cao 3461232

Contents

1 Introduction

Treasury securities are government debt instruments backed by full faith and credit of the United States. In macroeconomic and financial research, the yield curve, or equivalently, the discount curve of the U.S. Treasury securities has always been an important and critical economic quantity. A yield curve shows the relationship between the time to maturity of the debt and the interest rate. However, it is unobserved and needs to be estimated precisely and robustly for investment decisions, bond return predictions, and monetary policy analysis.

In the previous literature, there are two categories in the existing models: parametric and non-parametric methods. Nelson and Siegel (1987), Svensson (1994), and Gürkaynak, Sack and Wright (2007) are the three most important examples in parametric method. They make a smoothness assumption on the parametric forms of the yield curve, for instance, the form of the function is linear. Their parameters are simply estimated by minimizing pricing errors. Apparently, this assumption can not always be true, and these methods involve less flexible algorithms and are usually used for less complex problems. There are fewer assumptions in the non-parametric method in Fama and Bliss (1987) and Liu and Wu (2021), which makes their methods more flexible. However, some of their assumptions are either unrealistic or too restrictive and may lead to overfitting and dynamic instability.

Filipović, Pelger and Ye (2022) believe that the most prominent benchmark in parametric models is misspecified. In their research paper, in contrast to non-parametric benchmarks, they develop a data-driven, non-parametric discount curve estimator which is robust to outliers and stable over time. There are several methodological and empirical contributions in this paper. First, they combine financial theory with modern machine learning and make an optimal trade-off between flexibility and smoothness for the yield curve estimation in reproducing Kernel Hilbert spaces. The reward of smoothness is earned from a closed-form solution as simple ridge regression on the kernel basis function with a ridge penalty. Second, the authors view the discount curve as a Gaussian process, which naturally gives their method a Bayesian interpretation. From this perspective, the confidence intervals for the estimated discount curve, yields, and implied fixed-income security prices can be found. Third, they demonstrate that their method strongly dominates all parametric and non-parametric benchmarks in an extensive empirical study on U.S. Treasury securi-

ties. Substantially smaller out-of-sample yield and pricing errors are obtained by using their method compared to previous works of literature.

This term paper presents the basic idea used in Filipović, Pelger and Ye (2022) and its empirical evidence and comparison to other methods for yield curve smoothing. The organization is listed as follows. In section 2, we introduce Hilbert space and kernel trick and tailor them to the concept and application of reproducing kernel Hilbert space and representer theorem. In section 3, we leverage the discount curve in a Gaussian process view, giving it a Bayesian interpretation and constructing a confidence interval. In section 4, we replicate the empirical analysis in Filipović, Pelger and Ye (2022) and provide a simulation study in comparison to other methods of yield curve estimation.

2 Reproducing Kernel Hilbert Space

2.1 Hilbert Space

The theory of Hilbert space was initiated by David Hilbert. Debnath and Mikusinski (2005) provides a detailed explanation of Hilbert space $\mathcal{H}$. It is defined as a normed space that is complete and separable with respect to the norm defined by the inner product by the relation:

$$\| f \| = \sqrt{\langle f, f \rangle}.$$

An example of a defined norm in Hilbert space (i.e, the space L_2 of square-integrable functions) can be

$$\| f \| = \left(\int_a^b f^2(t) dx \right)^{\frac{1}{2}}.$$

A normed space is a vector space N on which a norm is defined. A non-negative function $\| \cdot \|$ is a norm if and only if $\forall f, g \in N$ and $\alpha \in \mathbb{R}$:

- $\| f \| \geq 0$ and $\| f \| = 0$ iff $f = 0$;
- $\| f + g \| \leq \| f \| + \| g \|$;
- $\| \alpha f \| = | \alpha | \| f \|$

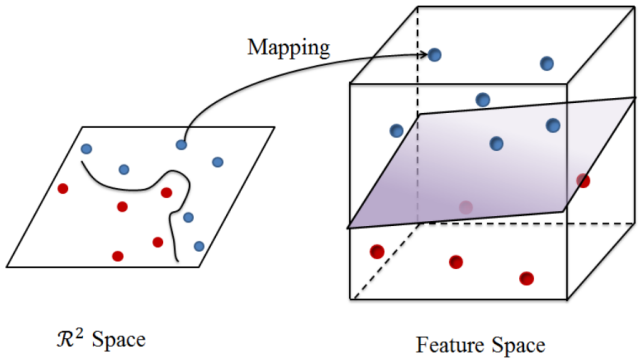
Examples of inner product $\langle a, b \rangle$ in a Hilbert space are

- a usual dot product: $\langle a, b\rangle = a'b = \sum_i a_i b_i$.
- a kernel product: $\langle a, b\rangle = k(a, b) = \psi(a)'\psi(b)$, where $\psi(a)$ may have infinite dimensions.

2.2 Introduction to Kernel

2.2.1 Feature map

The motivation of the kernel method is simple. Imagine there are some blue dots and red dots on a vector space $\mathcal{R}^2$ and we want to separate them by color. As it is shown in the left-hand side figure, it is difficult to divide them through a straight line. However, we may be able to separate them easily by mapping each dot into a high-dimension feature space. The figure below shows how the feature map works:



Let's now use a simple example to illustrate the idea of a feature map. we set two vectors $\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} y_1 & y_2 \end{bmatrix}$ in a two-dimension space. Tow functions $\phi(x)$ and $\phi(y)$ are defined as:

$$\phi(x) = \begin{bmatrix} x_1x_1 & x_1x_2 & x_2x_1 & x_2x_2 \end{bmatrix}$$

$$\phi(y) = \begin{bmatrix} y_1y_1 & y_1y_2 & y_2y_1 & y_2y_2 \end{bmatrix}$$

We are now successfully mapping them into a four-dimension feature space through the two functions. To write the above example in a general form of linear regression, we first set an equation where $\phi(\cdot) \in \mathcal{R}^m$ and $\phi(x)$ is defined as the mapping function. We then assume

there is a linear relation between y and $\phi(x)$:

$$\begin{aligned} y &= \phi(x)^\top w \\ &= \begin{bmatrix} \phi_1(x) \cdots \phi_m(x) \end{bmatrix} w \end{aligned} \tag{1}$$

Y and Φ in generalization is defined by:

$$Y = \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix}^\top \tag{2}$$

$$\begin{aligned} \Phi &= \begin{bmatrix} \phi(x_1) & \cdots & \phi(x_n) \end{bmatrix}^\top \\ &= \begin{bmatrix} \phi_1(x_1) & \cdots & \phi_m(x_1) \\ \vdots & \vdots & \vdots \\ \phi_1(x_n) & \cdots & \phi_m(x_n) \end{bmatrix} \end{aligned} \tag{3}$$

Recall the regularized risk minimization problem of ridge regression. In this case, it can be re-written as:

$$\begin{aligned} w* &= \underset{w}{argmin} \sum_{i=1}^{n} (y_i - \phi(x_i)^\top w)^2 + \lambda \parallel w \parallel_2^2 \\ &= \underset{w}{argmin} \parallel Y - \Phi w \parallel_2^2 + \lambda \parallel w \parallel_2^2 \end{aligned}$$

The least-square solution can also be re-defined by:

$$w* = (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top Y \tag{4}$$

Then we replace $w*$ with equation (4) in $y = \phi(x)^\top w$, we get:

$$\begin{aligned} y_w*(x) &= \phi(x)^\top w* \\ &= \phi(x)^\top (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top Y \\ &= \underbrace{\phi(x)^\top \Phi^\top}_{1 \times n} (\underbrace{\Phi \Phi^\top}_{n \times n} + \lambda I)^{-1} Y \\ &\text{using that} (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top = \Phi^\top (\Phi \Phi^\top + \lambda I)^{-1} \end{aligned} \tag{5}$$

2.2.2 Kernel Method

In most cases, it is surprisingly difficult to know and calculate the feature function after mapping. We want to avoid computing $\phi(x)$ in an explicit way, especially when m is very large. Therefore, we define a kernel function:

$$[\Phi\Phi^\top]_{i,j} = \phi(x_i)^\top \phi(x_j) = K(x_i, x_j)$$

$$[\phi(x)^\top \Phi^\top]_j = \phi(x)^\top \phi(x_j) = K(x, x_j) \tag{6}$$

This is simply the intuition of using the kernel method. For example, the Gaussian kernel is

$$k(x_i, x_j) = e^{\frac{-\|x_i - x_j\|}{\sigma^2}},$$

Gaussian kernel means the similarity between two points where $\| x_i - x_j \|$ is the Euclidean distance between x_i and x_j, and $\sigma^2 \in \mathbb{R}^+$ is the bandwidth of the kernel function. As shown in Hofmann, Schölkopf and Smola (2008), it has the following properties:
for $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if

- k is symmetric: $k(x, y) = k(y, x)$.
- k is positive semi-definite, meaning that $\sum_i \sum_j \alpha_i \alpha_j k(x_i, x_j) \geq 0, \forall \alpha_i, \alpha_j \in \mathbb{R}, x \in \mathbb{R}^{\mathbb{D}}, \mathbb{D} \in \mathbb{Z}^+$.
- We define the corresponding kernel matrix as the matrix K with entries $k_{ij} = k(x_i, x_j)$, the sencond property of k is equivalent to saying that $\mathbf{a}' K \mathbf{a} \geq 0$.

Now we can define a function k: $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if and only if there exists a Hilbert space $\mathcal{H}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $k(x, y) = \langle \phi(x), \phi(y) \rangle$.
Recall the simple example above, instead of computing the inner product of $\langle \phi(x), \phi(y) \rangle$, we can define a corresponding kernel function $K(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle^2$. It can be easily proofed that

$K(\mathbf{x}, \mathbf{y})$ is the same as $\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$:

$$\begin{aligned} K(\mathbf{x}, \mathbf{y}) &= \langle \mathbf{x}, \mathbf{y} \rangle^2 \\ &= (x_1 y_1 + x_2 y_2)^2 \\ &= x_1^2 y_1^2 + 2 x_1 y_2 x_2 y_1 + x_2^2 y_2^2 \\ &= \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle \end{aligned}$$

2.3 Reproducing Kernel Hilbert Space

Consider a Hilbert space $\mathcal{H}$ full of real-valued functions from $\mathcal{X}$ to $\mathbb{R}$, and a mapping $\Phi : \mathcal{X} \to \mathbb{R}^{\mathcal{X}}$ defined as $x \to \Phi(x) = k_x = k(\cdot, x)$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel of $\mathcal{H}$, and $\mathcal{H}$ is a reproducing kernel Hilbert space, if:

- $\forall x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{H}$,
- $\forall x \in \mathcal{X}$, $f \in \mathcal{H}$, $\langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$, which is the reproducing property.

Paulsen and Raghupathi (2016) explained this concept in a more intuitive way—RKHS is a Hilbert space $\mathcal{H}$ with a reproducing kernel whose span is dense in $\mathcal{H}$. Moreover, an RKHS can be defined as a Hilbert space of valid functions with all evaluation functionals bounded and linear.

2.4 Representer Theorem

In the previous section, we have learned that there is always a pair of $(\mathcal{X}, k)$ as a Hilbert space or a subset of that space whenever the input domain $\mathcal{X}$ exists. Such a fact means that we are able to study the various data structures in Hilbert spaces. In the practical world, however, it is extremely difficult to study many popular kernels since their Hilbert spaces are known to be infinite-dimensional in almost every case. Especially for the purpose of machine learning, we usually prefer to solve an optimization problem in a finite-dimensional space.

This is where the representer theorem is useful. It contributes to simplifying the regularized risk-minimization problem by reducing the infinite-dimensional space to a finite-dimensional space of optimal coefficients and provides provisions for kernels in training data in machine learning.

In the simplest form of machine learning, in order to predict x, the algorithm collects the samples in the training set $\mathcal{X}$ that are similar to x, and then takes the weighted value of these samples as the predict value of x. Here comes the questions:

- How to measure the similarity between samples?
- How to weigh the value of each sample?

In general, the higher the similarity of the sample to our point of interest x, the more the sampling weights. To evaluate the similarity between two observations, a kernel is defined as a function of two input patterns $k(x_i, x_j)$, mapping onto a real-valued output. Hofmann, Schölkopf and Smola (2008) wrote that the advantage of using such a kernel as a similarity measure is that it allows us to construct algorithms in dot product spaces.
Suppose we are given a nonempty set $\mathcal{X}$, a positive definite real-valued kernel k on $\mathcal{X} \times \mathcal{X}$, a training sample $(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X} \times \mathbb{R}$, a real-valued function f in Hilbert space $\mathcal{H}$. As explained in Schölkopf, Herbrich and Smola (2001), we can find the function f^* in the RKHS $\mathcal{H}$ satisfying:

$$\mathcal{J}(f^*) = \underset{f \in \mathcal{H}}{min} \mathcal{J}(f),$$

where

$$\mathcal{J}(f) = L_y(f(x_1), \cdots, f(x_n)) + \Omega(\| f \|^2_{\mathcal{H}}).$$

Note that Ω is a non-decreasing regularizer and y is a vector of y_i.
Representer theorem: the solution to

$$\underset{f \in \mathcal{H}}{min}[L_y(f(x_1), \cdots, f(x_n)) + \Omega(\| f \|^2_{\mathcal{H}})]$$

can be written in a simpler version, which takes the form:

$$f^* = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot),$$

where α_i the weighted value of each sample and k is the similarity measure. If Ω is strictly increasing, all solutions apply to this form.

2.5 Example of Using Representer Theorem–Kernel Ridge Regression

Suppose we are given empirical data $(y_1, x_1), \ldots, (y_n, x_n)$, where $i = 1, \ldots, N$. Assume $y = g(x)$ in RKHS. In order to avoid extremely high variance, we impose an additional assumption that a smoother curve with fewer oscillations is preferred. We utilize regularization to simplify the function and satisfy the additional assumption by adding a penalty term Ω. We want to estimate the function $g(\cdot)$ to minimize

$$\min_{g \in \mathcal{H}} \sum_{i=1}^{N} (y_i - g(x_i))^2 + \Omega \parallel g \parallel_{\mathcal{H}}^2 \tag{7}$$

Luckily, the representer theorem already tells us that the least squared problem always has a solution of the form

$$g(\cdot)^* = \sum_{i=1}^{N} \alpha_i k(\cdot, x_i), \tag{8}$$

and according to reproducing property of RKHS, we have

$$g(x) = \langle g(\cdot), k(\cdot, x) \rangle_{\mathcal{H}}. \tag{9}$$

Also, it is obvious that

$$\parallel g \parallel^2 = \langle g(\cdot), g(\cdot) \rangle. \tag{10}$$

We then substitute (9), (10) for (7),

$$\min_{g \in \mathcal{H}} (y_i - \langle g(\cdot), k(\cdot, x) \rangle)^2 + \Omega \langle g(\cdot), g(\cdot) \rangle,$$

and plug (8) in (7) and get

$$\min_{\alpha} \sum_{i=1}^{N} (y_i - \langle \sum_{j=1}^{N} \alpha_j k(\cdot, x_j), k(\cdot, x_i) \rangle)^2 + \Omega \langle \sum_{i=1}^{N} \alpha_i k(\cdot, x_i), \sum_{j=1}^{N} \alpha_j k(\cdot, x_j) \rangle$$

$$\Rightarrow \min_{\alpha} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{N} \alpha_j k(x_j, x_i))^2 + \Omega \sum_{i=j}^{N} \sum_{i=1}^{N} \alpha_j \alpha_i k(x_i, x_j) \tag{11}$$

Remember the corresponding kernel matrix as the matrix K with entries $k_{ij} = k(x_i, x_j)$ is equivalent to saying that $\mathbf{a}'K\mathbf{a} \geq 0$, we can then rewrite (11) as:

$$\| y_i - K\mathbf{a} \|^2 + \Omega \mathbf{a}' K \mathbf{a}. \tag{12}$$

By differentiation and setting the first order derivative of (12) to zero, we get:

$$\alpha^* = (K + \Omega I_n)^{-1} y. \tag{13}$$

2.6 Example of Solving the Kernel Function

In this paper, we study in the RKHS $\mathcal{H} = \mathcal{H}_{\omega,\delta}$ consisting of differentiable functions $h : [0, \infty) \to \mathbb{R}$ of the form $h(x) = \int_0^x h'(t)dt$ with continuous derivatives, $h'(x) = h'(0) + \int_0^x h''(t)dt$ for integrable h'', and with finite norm

$$\langle h, h \rangle = \| h \|_{\omega,\delta} = \left(\int_0^\infty (\delta h'(x)^2 + (1-\delta) h''(x)^2) \omega(x) dx \right)^{\frac{1}{2}} \tag{14}$$

for some measurable weight function $\omega : [0, \infty) \to [1, \infty)$ and shape parameter $\delta \in (0, 1)$. With additional assumption in the research paper's appendix A.2, we can extend it to the case $\delta \in \{0, 1\}$.

The Lemma 3 assumes that for any fixed $y \geq 0$, exists a solution ϕ of the linear differential equation

$$\delta \phi \omega - (1-\delta)(\phi' \omega)' = 1_{[0,y]} \tag{15}$$

and for $\psi \in \mathcal{H}_{\omega,\delta}, \psi(x) = \int_0^x \phi(t)dt$, then for $h \in \mathcal{H}_{\omega,\delta}$ with $h'(x) = 0$ for $x > n$ for some finite n, we can write

$$\langle \psi, h \rangle_{\omega,\delta} = \int_0^\infty (\delta \psi'(x) h'(x) + (1-\delta) \psi'' h''(x)) \omega(x) dx \tag{16}$$

according to the definition for any $h \in \mathcal{H}_{\omega,\delta}$. The assumption of sloution exists and Lemma 4 could give us that $\langle \psi, h \rangle = h(y)$ which implies that $k(\cdot, y) = \psi$ by the reproducing property. Then $k(x, y) = \psi(x)$, and remind that $\psi(x) = \int_0^x \phi(t)dt$, then we can find the form of $k(x, y)$ if we know the form of ϕ, and we could solve ϕ by giving different value of δ and ω. Below is an example of how to solve the research paper's equation 8.

In the research paper, the weight function $\omega(x) = e^{\alpha x}$, if $\alpha = 0, \delta = 1$, then $\phi = 1_{[0,y]}$, and $k(x,y) = \psi(x) = \int_0^x 1_{[0,y]} dt = min\{x, y\}$.

3 Gaussian Process and Bayesian Perspective

After estimating the discount function $g(\boldsymbol{x})$, we want to do inference with this function, and because of the non-parametric estimation, we cannot calculate the statistical distribution of parameters, hence here we use Gaussian Process to estimate the distribution of discount function and construct a confidence interval.

3.1 Definition

We have data $\mathbf{D} = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^M$, and assume that mean of y is 0. We want to find the distribution of $f^*(x)$.

Assume that the true form of prediction function is: $y_i = f(\boldsymbol{x_i}) + \epsilon_i$ and $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$. Here we have an M dimensional dependent variable **y**, and a $M \times N$ dimensional independent variable **X**, where M is the number of observations, and N is the dimension of x, i.e. $\boldsymbol{x_i} \in \mathbb{R}^N$. The function $f(\boldsymbol{x_i}) : \mathbb{R}^N \to \mathbb{R}$ takes vector $\boldsymbol{x_i} \in \mathbb{R}^N$. Let $\boldsymbol{K}_{X,X} = k(\boldsymbol{x}, \boldsymbol{x}^T)$ which is the matrix of $k(\boldsymbol{x_i}, \boldsymbol{x_j})$. Thus, **K** is a $M \times M$ matrix.

The assumption of the Gaussian Process is listed as follows:

- for a given vector **y**, and its corresponding data **X**, where vector $\boldsymbol{y} \in \mathbb{R}^M$ and $\boldsymbol{X}$ is $M \times N$ matrix.
- for **y** and **X** data, the error term $\epsilon \sim \mathcal{N}(\mathbf{0}, \Sigma^\epsilon)$, and $\Sigma^\epsilon = diag(\sigma_1^2, \sigma_2^2, \sigma_3^2, ..., \sigma_M^2)$.
- we have arbitrary $n \times N$ matrix **Z** and predicted value $f^*(\boldsymbol{z}) \in \mathbb{R}^n$, where $\boldsymbol{z} = (\boldsymbol{z_1}, \boldsymbol{z_2}, \boldsymbol{z_3}, ..., \boldsymbol{z_n})^T$.
- we assume **y** and $f^*(\boldsymbol{z})$ follow a $(M+n)$ multivariate normal distribution(MVN):

$$\begin{bmatrix} f^*(\boldsymbol{z}) \\ \boldsymbol{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_{f^*(\boldsymbol{z})} \\ \mu_{\boldsymbol{y}} \end{bmatrix} , \begin{bmatrix} \boldsymbol{K}_{Z,Z} & \boldsymbol{K}_{Z,X} \\ \boldsymbol{K}_{X,Z} & \hat{\boldsymbol{K}}_{X,X} \end{bmatrix} \right) \tag{17}$$

where $\hat{\boldsymbol{K}}_{X,X} = \boldsymbol{K}_{X,X} + \Sigma^\epsilon$.

Then given data **y**, **X** and **Z**, according to the conditional distributions of the multivariate normal distribution (proof see Rasmussen and Williams (2006)), we have the posterior distribution

$$f^*(\boldsymbol{z})|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z} \sim \mathcal{N}(\mu_{f^*(\boldsymbol{z})} + \boldsymbol{K}_{Z,X}\hat{\boldsymbol{K}}_{X,X}^{-1}(\boldsymbol{y} - \mu_{\boldsymbol{y}}), \boldsymbol{K}_{Z,Z} - \boldsymbol{K}_{Z,X}\hat{\boldsymbol{K}}_{X,X}^{-1}\boldsymbol{K}_{X,Z}) \tag{18}$$

3.2 Intuition behind Gaussian Process

The idea behind this process is that, assume our interested function is $f(x)$, $f(\boldsymbol{x}) : \mathbb{R}^N \to \mathbb{R}$, and we have an arbitrary vector of independent variable $\boldsymbol{x} = (\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_M})^T$, and for each $\boldsymbol{x_i}, i = 1, 2, ..., M, x_i \in \mathbb{R}^N$, then we can obtain a series of $f(\boldsymbol{x}) = (f(\boldsymbol{x_1}), f(\boldsymbol{x_2}), f(\boldsymbol{x_3}), ..., f(\boldsymbol{x_M}))^T$ We assume that the series of f(**x**) follows a multivariate normal distribution which is:

$$f(\boldsymbol{x}) \sim \mathcal{N}(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}^T)) \tag{19}$$

This is the prior distribution of our function $f(x)$, here we have a set of infinite functions that follow this distribution, their mean is the function $\mu(\boldsymbol{x_i})$, and the variance of them is $k(\boldsymbol{x_i}, \boldsymbol{x_i}^T)$. This makes the distribution of $f(\boldsymbol{x})$ to be called Gaussian Process (GP). Note that if we add a noise term $\epsilon \sim \mathcal{N}(\boldsymbol{0}, \Sigma^\epsilon)$, then our prior distribution of $y = f(\boldsymbol{x}) + \epsilon \sim \mathcal{N}(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}^T) + \Sigma^\epsilon)$ is also a Gaussian Process. Here we use the kernel matrix to denote the variance-covariance matrix because the kernel value represents how near two data points in the space are, with this property we can obtain a smooth function.

Remind that our goal is to estimate the distribution of $f(\boldsymbol{x}^*)$ given observed training data set $D = \{\boldsymbol{x_i}, y_i\}_{i=1}^M$ and test data set $\{\boldsymbol{x_j^*}\}_{j=1}^n$. Firstly we compare our nonparametric case to a parametric case. In a parametric case, assume the parameter θ determines the form of $f_\theta(\cdot)$, according to the Bayesian rule, $p(\boldsymbol{y}^*|\boldsymbol{x}^*, \boldsymbol{x}, \boldsymbol{y}) = \int_\theta p(\boldsymbol{y}^*, \theta|\boldsymbol{x}^*, \boldsymbol{x}, \boldsymbol{y})d\theta = \int_\theta p(\boldsymbol{y}^*|\theta, \boldsymbol{x}^*)p(\theta|\boldsymbol{x}, \boldsymbol{y})d\theta$, where $\boldsymbol{y}^*$ is the prediction of given data $\boldsymbol{x}^*$, and its form of model is determmined by paramater θ. The estimated θ value is determined by training data D. This is to say that we update our parameter θ by given D, and use $p(\theta|\boldsymbol{x}, \boldsymbol{y})$ as a new prior probability, and based on this to predict posterior of $\boldsymbol{y}^*$.

Therefore, back to our GP nonparametric case, θ could be substituted by function $f(\cdot)$. One can show that the joint distribution of $(f(\boldsymbol{x}^*), \boldsymbol{y})^T$ follows a multivariate normal distribution as in the definition before, because of the assumption of GP and the property of

MVN. With the joint distribution, we want to find posterior probability: $p(f(\boldsymbol{x}^*)|\boldsymbol{x}^*, \boldsymbol{x}, \boldsymbol{y}) = \int p(f(\boldsymbol{x}^*)|f, \boldsymbol{x}^*)p(f|\boldsymbol{x}, \boldsymbol{y})df$, where $p(f|\boldsymbol{x}, \boldsymbol{y})$ is the posterior of $f(\cdot)$ given D, and is regarded as prior when estimating $p(f(\boldsymbol{x}^*)|\boldsymbol{x}^*, D)$, this process is called Bayesian updating. Fortunately, we do not need to take any integral in GP, because the posterior of $f(\boldsymbol{x}^*)$ could be calculated by the formula of conditional distribution in MVN as mentioned in the former section.

3.3 Gaussian Process in research paper

In this research paper, authors assumed discount function $g(\boldsymbol{z})$ given a vector of different maturties $\boldsymbol{z} = (z_1, z_2, ..., z_n)$ follows a MVN distribution $\mathcal{N}(m(\boldsymbol{z}), k(\boldsymbol{z}, \boldsymbol{z}^T))$. Then this is a Gaussian Process, and by Bayesian updating for given price P, corresponding cash flow matrix C, and time to maturities x, we can obtain the posterior mean and variance function in the research paper's equation (12) and equation (13). Therefore, the variance function of MVN could give us the confidence interval of $g(z)$ i.e. for each maturity time z we calculate $k^{post}(z, z)$ as its normal variance, which could help us to evaluate the precision of our prediction. Furthermore, with the posterior distribution of $g(\boldsymbol{x})$, it is implied that the coupon bond price $Cg(\boldsymbol{x}) \sim \mathcal{N}(Cm^{post}(\boldsymbol{x}), Ck^{post}(\boldsymbol{x}, \boldsymbol{x}^T)C^T)$.
Note that authors assume the variance covariance matrix of error term $\Sigma^\epsilon = diag(\sigma_1^2, \sigma_2^2, \sigma_3^2, \ldots, \sigma_M^2)$ where diagonal elements all satisfy $\omega_i = \frac{\lambda}{\sigma_i^2}$, this implies that we give a higher weight for a bond price which has less noise. In addition, we assume that the prior mean function is constant $m(x) = 1$ which assumes no time value of money. With these assumptions, the posterior mean function coincides with estimated $\hat{g}(\boldsymbol{x})$ as in the research paper's equation (5).

4 Empirical Study

We use the Nelson–Siegel–Svensson (NSS) model and naively use Spline estimated discount curve and yield curve to compare with our RKHS estimated discount curve and yield curve.

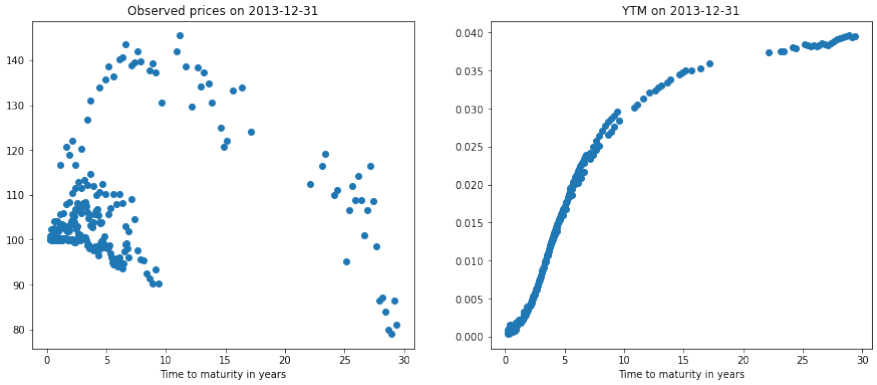


Figure 1: Price distribution and Yield to Maturity

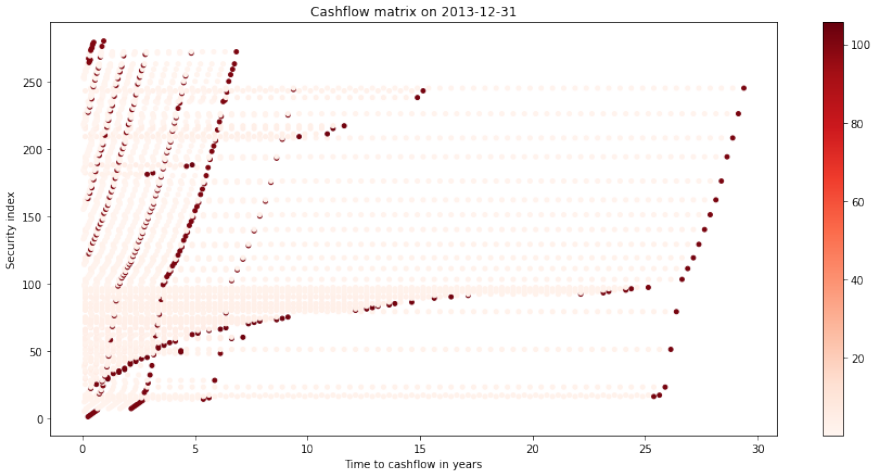


Figure 2: Cash flow

4.1 Data

In this paper, we only use a part of data to train and test our models, which are the price data and cash flow data observed in 2013, December 31st. Figure 1 shows a big part of bounds have maturity within 10 years, and our data lacks bounds about maturity around 20 years. The yield to maturity (YTM) is annualized, it is different from coupon yield.
Figure 2 shows the cash flow distribution, the dark red dots represent the last payment of a bound, which has the highest payment amount for each bound. Figure 2 also shows the cash flows are unevenly distributed.

4.2 Estimated results

Figure 3 (left) shows the estimated discount curves, within 5 years of maturity, discount curves perform similarly, but in long term, the discount rate estimated by RKSH is lower.

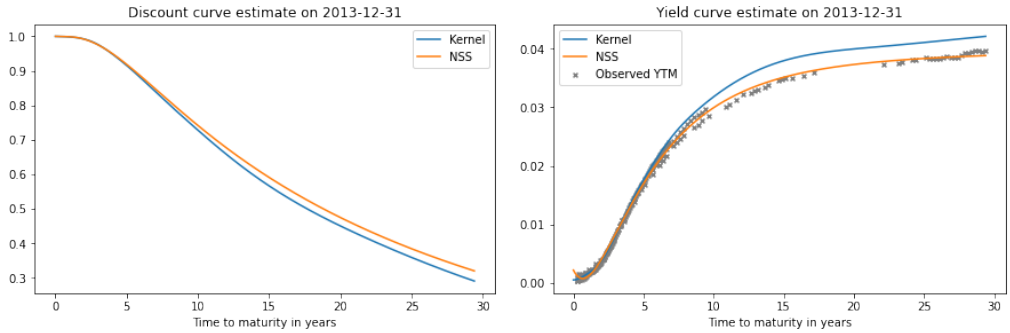


Figure 3: Fitted Discount Curve & Yield Curve

Figure 3 (right) shows the yield curves with observed YTM, in short term we can find the NSS curve is inverted, which is unnatural in a common date. Meanwhile, if we use a machine-learning model such as Spline naively, we can obtain a negative yield rate as in figure 3 (right), which makes no sense. The bias for NSS and Spline appears because both methods only fit on observed YTM, so only if all our bounds are coupon bounds, their fitted curves will make sense, but in practice, such data is hard to find.

Figure 4 shows the predicted price and fitted YTM calculated via the predicted price. One can find that the predicted results of the kernel method are very precise, (Filipović, Pelger and Ye, 2022) also showed that their method has much lower MSE compared with other methods.

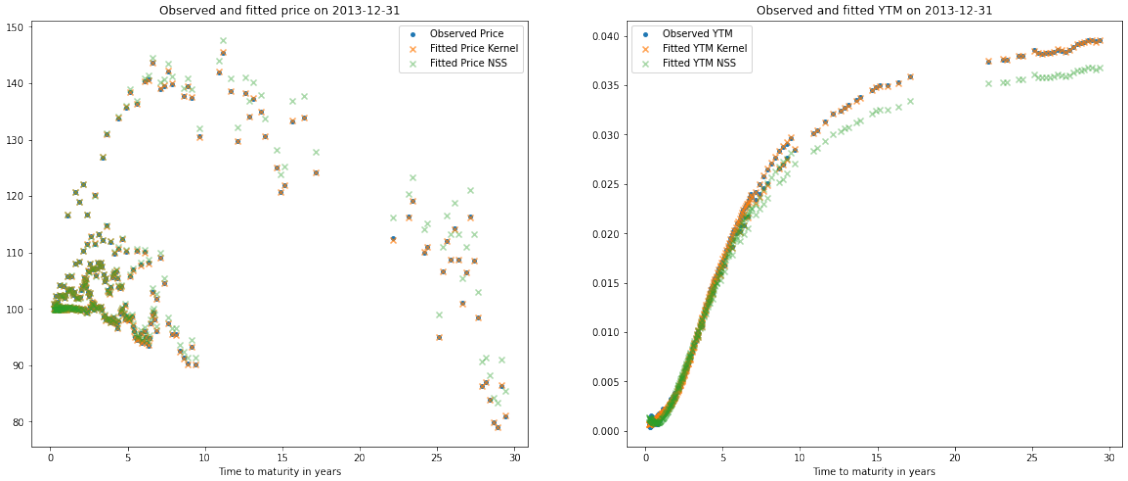


Figure 4: Fitted Yield Curve

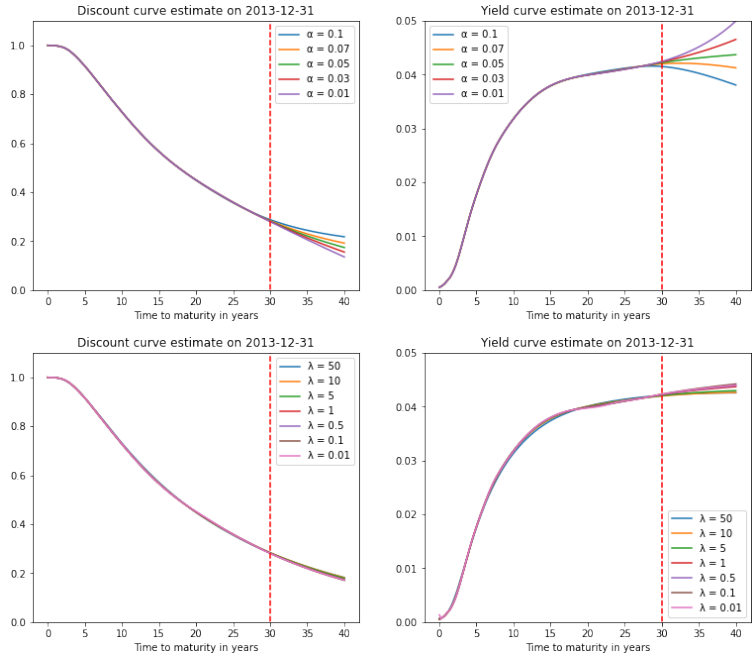


Figure 5: Extrapolation with α & λ

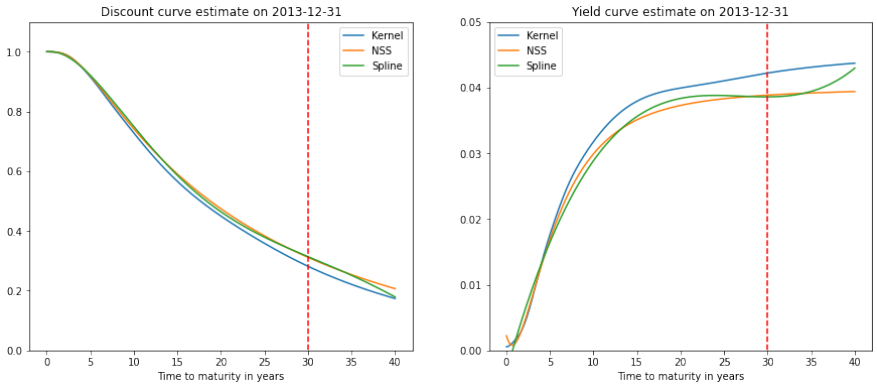


Figure 6: Extrapolation with multiple models

4.3 Extrapolation

In the extrapolation of RKHS model, we extrapolate it to 40 years, and try different hyperparameters, figure 9 shows us that for α, which shows the yield for an infinite maturity discount bound, in the sample range 30 years, different α give similar curves, but for the out-of-sample case, it varies because kernel method cannot learn it from given sample data. But for tuning parameter λ, it can be chosen.

Figure 10 shows the exploration results for different models, we can see that the curve for the kernel method is stable. Meanwhile, NSS curve is also stable because NSS is a parametric method. Nevertheless, Spline performs worse because its algorithm is designed for

interpolation.

4.4 Simulation

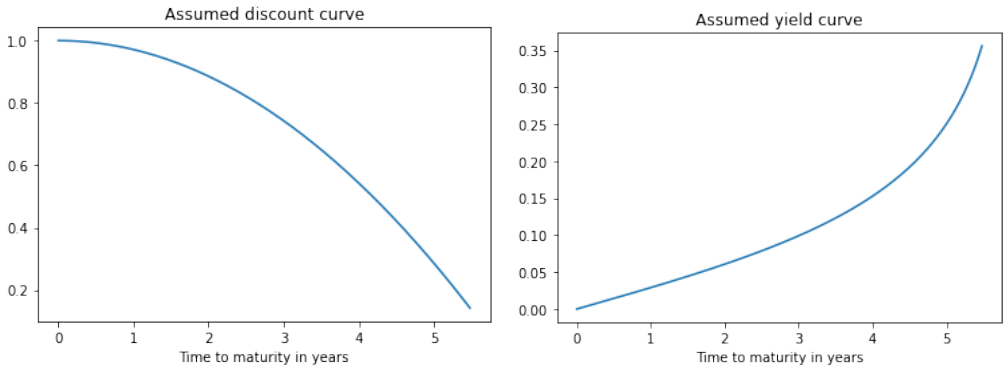


Figure 7: True Discount curve and Yield curve

Finally, we can compare and test the kernel model with a simulation study. We assume a true discount curve that satisfies a monotone-increasing transformed yield curve. It is transformed by $y(x) = -\frac{1}{x}log(g(x))$, where x denotes time. The assumed true curves are shown in figure 11.

Then we simulate the cash flow in around 5 years, for 200 bounds, each bound has 10 cash payments randomly in some time interval. In order to let the time to maturity be distributed more uniformly, some bounds have much shorter maturity, as shown in figure 12.

With this simulation setting, we can solve the discount curve and predicted price as shown

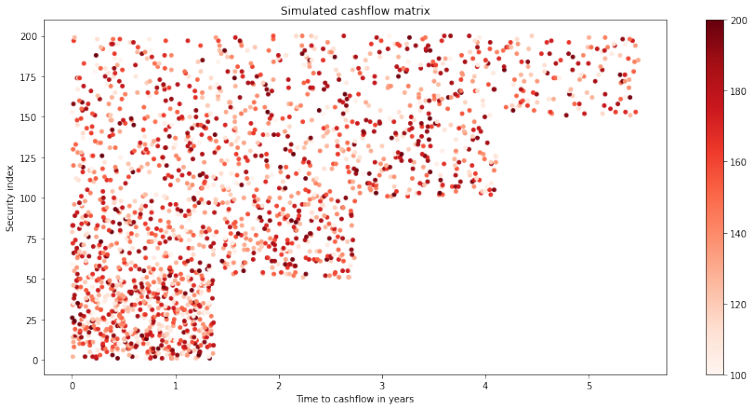


Figure 8: True Discount curve and Yield curve

in figure 13. In such a simulated case, we can find that the performance of the kernel method outperforms two others.

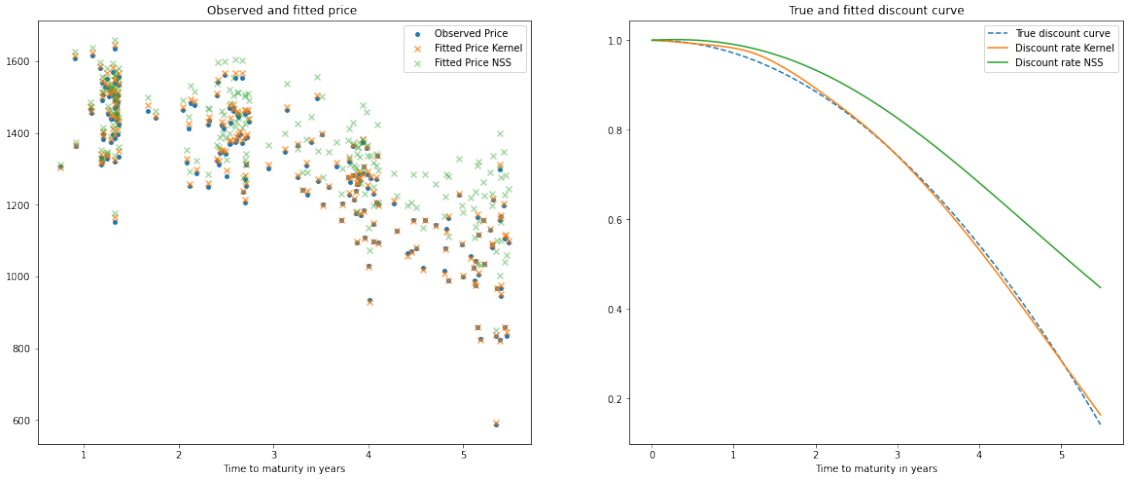


Figure 9: True Discount curve and Yield curve

5 Conclusion

In conclusion, Filipović, Pelger and Ye (2022) uses RKHS estimation to obtain a robust and effective discount curve. This RKHS model only takes cash flow and price data as input data, and solves kernel function within RKHS, as section 2.3 in Filipović, Pelger and Ye (2022), such a model can cover most of the famous yield curve estimation models, and outperforms. Liu and Wu (2021) also proposes a kernel estimation but it is local, that only fits and makes smoothness with nearby eight bounds, and it turns out that the RKHS model is globally and more robust.

Filipović, Pelger and Ye (2022) also proposes a Gaussian Process estimation for their discount curve and the yield curve, which could give such non-parametric model power for inference.

Finally, in our empirical study, we compared RKHS model with NSS and naive Spline. Our study and simulation have shown the advantage of the kernel model over the parametric model, and the dangers of using a naive machine-learning method over a complex economic problem.

In summary, the kernel method and RKHS model give us a non-parametric discount curve estimation, enabling us to exploit in the cash flow and price data directly, compared with NSS. Meanwhile, the RKHS model makes a global estimation and produces a robust result. We can use such a discount curve in the broad fields of economics and finance issues.

References

Debnath, Lokenath and Piotr Mikusinski. 2005. *Introduction to Hilbert spaces with applications*. Academic press.

Fama, Eugene F and Robert R Bliss. 1987. "The information in long-maturity forward rates." *The American Economic Review* pp. 680–692.

Filipović, Damir, Markus Pelger and Ye Ye. 2022. "Stripping the Discount Curve-a Robust Machine Learning Approach." *Swiss Finance Institute Research Paper* (22-24).

Gürkaynak, Refet S, Brian Sack and Jonathan H Wright. 2007. "The US Treasury yield curve: 1961 to the present." *Journal of monetary Economics* 54(8):2291–2304.

Hofmann, Thomas, Bernhard Schölkopf and Alexander J Smola. 2008. "Kernel methods in machine learning." *The annals of statistics* 36(3):1171–1220.

Liu, Yan and Jing Cynthia Wu. 2021. "Reconstructing the yield curve." *Journal of Financial Economics* 142(3):1395–1425.

Nelson, Charles R and Andrew F Siegel. 1987. "Parsimonious modeling of yield curves." *Journal of business* pp. 473–489.

Paulsen, Vern I and Mrinal Raghupathi. 2016. *An introduction to the theory of reproducing kernel Hilbert spaces*. Vol. 152 Cambridge university press.

Rasmussen, Carl Edward and Christopher KI Williams. 2006. "Gaussian processes for machine learning. isbn 026218253x.".

Schölkopf, Bernhard, Ralf Herbrich and Alex J Smola. 2001. A generalized representer theorem. In *International conference on computational learning theory*. Springer pp. 416–426.

Svensson, Lars EO. 1994. "Estimating and interpreting forward interest rates: Sweden 1992-1994.".

Appendix

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.