University of Bonn

Research Module in

Econometrics and Statistics

# Term Paper of Estimating Discount Curve with Robust Machine Learning

February 9, 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 in the estimation of yield curve: 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. Their parameters are 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 on the curve.

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 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 empiri-

cal study on U.S. Treasury securities. 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 theories and concepts involved in Filipović, Pelger and Ye (2022) and its empirical evidence and comparison to Nelson and Siegel (1987) for yield curve estimating. The organization is listed as follows. In section 2, we introduce Hilbert space and kernel and tailor them to the concept and application of reproducing kernel Hilbert space (RKHS) and representer theorem. In section 3, we introduce the kernel ridge regression implementation of discount curve estimating. In section 4, we leverage the discount curve in a Gaussian process view, giving it a Bayesian interpretation. In section 5, we replicate the empirical analysis in Filipović, Pelger and Ye (2022) and provide a simulation study in comparison to another method of yield curve estimation.

# 2 RKHS and Representer Theorem

## 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:

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

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

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

Examples of inner product $\langle a, b \rangle$ in a Hilbert space $\mathcal{H}$ are

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

## 2.2 Introduction to Kernel

In statistical learning, defining feature mapping $\phi(x)$ is a usual approach, which enables us to exploit more information from the raw data, just like adding polynomial terms and interaction terms in linear regression. *What distinguishes kernel methods is that they can use infinitely many features. This can be achieved as long as our learning algorithms are defined in terms of dot products between the features, where these dot products can be computed in closed form.* (Gretton, 2013). For instance, the kernel function $K(x_i, x_j)$ could be defined in the form of inner product $\langle \phi(x_i), \phi(x_j) \rangle$.

Theodoridis and Koutroumbas (2006) explained that kernel functions allow operating in a high-dimensional feature space without computing the coordinates of the data in that space, but rather the inner product between the image of a pair of data in feature space. And the inner product can keep the properties of the mapping function and of the dataset. In the Appendix, you can find an easy example that relates feature mapping and kernel function. Nevertheless, not all kernels are defined in the form of the dot product. For example, the Gaussian kernel is

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

Gaussian kernel measures 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. In general, kernels 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 inner product spaces. 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 $\mathbf{a}' K \mathbf{a} \geq 0$.

## 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 (Gretton, 2013):

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

**Theorem 1** (The Representer Theorem)**.** *(Schölkopf, Herbrich and Smola, 2001) Let k be a kernel on $\mathcal{X}$ and let $\mathcal{H}$ be its associated RKHS. Fix $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$, and consider the optimization problem*

$$\mathcal{J}(f^*) = \min_{f \in \mathcal{H}} \mathcal{J}(f), \tag{1}$$

*where*

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

*Note that $\Omega$ is a non-decreasing regularizer and $L$ depends on $f$ only though $f(x_1), \ldots, f(x_n)$. If (1) has a minimizer, then it has a minimizer of the form*

$$f^* = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot), \tag{2}$$

*where $\alpha_i \in \mathbb{R}$. Furthermore, if $\Omega$ is strictly increasing, all solutions apply to this form.*

Equation (2) shows that the solution of $f$ lies on the subspace that is expanded by $k(x_i, \cdot), \forall x_i \in \mathcal{X}$. $f^*$ is in the finite-dimensional subspace.

# 3 Discount Curve: Kernel-Ridge Regression Solution

## 3.1 Optimized Discount Curve

In Filipović, Pelger and Ye (2022), their research problem is to estimate the unobserved discount curve $g(x)$ from sparse and noisy Treasury prices. $g(x)$ is interpreted as how much $\$1$ in the future time point $x$ worth at present. The authors let:

- $P = (P_1, P_2, P_3, ..., P_M)^\top$ denotes the observed prices for $M$ bonds;
- $0 < x_1 < x_2 < x_3 < .... < x_N$, also $\mathbf{x} = (x_1, \ldots, x_N)^\top$, denote the time series of days;
- $C$ denotes observed $M \times N$ cash flow matrix, and $C_{ij}$ denotes the cash flow of security $i$ at date $x_j$.

The no-arbitrage pricing relation is defined as:

$$\begin{aligned} P_i &= P_i^g + \epsilon_i \\ &= \sum_{j=1}^{N} C_{ij} g(x_j) + \epsilon_i, \end{aligned}$$

where the error $\epsilon_i$ is the deviations from fundamental value due to market imperfections and data errors.

The authors need $g(0) = 1$ because of the properties of the discount curve. It is convenient to model the discount curve as:

$$g = p + h$$

for some exogenous prior curve $p$: $[0,\infty) \to \mathbb{R}$ with $p(0) = 1$, and a hypothesis function $h$ optimally chosen from a RKHS $\mathcal{H}$ consisting of functions $h : [0,\infty) \to \mathbb{R}$ with an initial value $h(0) = 0$. In this paper, a special case where $p$ is always constant to be 1 is assumed.
The estimation of discount curve $g(x)$ naturally starts with minimizing the pricing errors for some exogenous weights $w_i$:

$$\min_g \left\{ \sum_{i=1}^{M} \omega_i (P_i - P_i^g)^2 \right\}.$$

The main problem here is that the number of observed prices $M \approx 300$ is substantially smaller than the number of maturity dates $N \approx 10,000$. Hence, the authors impose regularizing assumptions to limit the number of parameters by ridge regression. Meanwhile, the smoothness assumption of the discount curve is motivated by economic principles. Similar bonds should have similar payoffs and the limits to arbitrage require a sufficiently smooth curve. Therefore, the smoothness problem is formulated by:

$$||g||_{\alpha,\delta} = (\int_0^\infty (\delta g'(x)^2 + (1-\delta) g''(x)^2) e^{\alpha x} dx)^{1/2}.$$

The discount curve $g$ is chosen from the set $\mathcal{G}_{\alpha,\delta}$ of twice weakly differentiable functions $g : [0,\infty) \to \mathbb{R}$ with $g(0) = 1$. This is a general measure of smoothness given by the weighted average of the first and second derivatives of $g$. $e^{\alpha x}$ is the weight function that allows the smoothness measure to be maturity-dependent. And $\delta \in [0,1]$ is the shape parameter that allows a trade-off between two forces. The first force is given by penalizing $g'(x)^2$ to avoid oscillations and make the curve tense. The second force is given by penalizing $g''(x)^2$ to avoid kinks and make the curve more straight.
Now the estimation problem is expressed by the optimization problem

$$\min_{g \in \mathcal{G}_{\alpha,\delta}} \left\{ \sum_{i=1}^{M} \omega_i (P_i - P_i^g)^2 + \lambda ||g||_{\alpha,\delta}^2 \right\}, \tag{3}$$

and it trades off the weighted mean squared pricing error against the smoothness of $g$. The $\omega_i$ in $(3)$ is set to be inversely proportional to the squared duration $D_i$ of bond $i$, that is

$$\omega_i = \frac{1}{M} \frac{1}{(D_i P_i)^2}.$$

This implies that the authors believe that the pricing error in short-term bonds is more important than in long-term bonds in the estimation of discount curve $g(x)$.

The solution to the optimization problem $(3)$ is obtained by using the definition of RKHS and the representer theorem. For every function in an RKHS that minimizes an objective function can be written as a linear combination of the reproducing kernel. It effectively simplifies an infinite-dimensional optimization problem into a finite one. Thus, the optimization problem $(3)$ has a unique solution $\hat{g}$, given in closed form by and the discount function $g(x)$ is given by

$$\begin{aligned}\hat{g}(x) &= p + \hat{h} \\ &= 1 + \sum_{j=1}^{N} k(x, x_j)\beta_j,\end{aligned}$$

and the parameter—the weight of different kernel functions:

$$\boldsymbol{\beta} = C^\top (C\boldsymbol{K}C^\top + \Lambda)^{-1}(P - C\mathbf{1}),$$

where $\Lambda = diag(\frac{\lambda}{\omega_1}, ..., \frac{\lambda}{\omega_M})$. A general solution for kernel ridge regression could be found in the Appendix.

## 3.2 Example of Solving the Kernel Function

In this paper, we are actually studying 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

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

for some measurable weight function $\omega : [0, \infty) \to [1, \infty)$ and shape parameter $\delta \in (0, 1)$. With additional assumption in the Filipović, Pelger and Ye (2022)'s *appendix A.2*, we can extend it to the case $\delta \in \{0, 1\}$.

The Filipović, Pelger and Ye (2022)'s *Lemma 3* assumes that for any fixed $y \geq 0$, exists a

solution $\phi$ of the linear differential equation

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

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

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

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 $\phi$, and we could solve $\phi$ by giving different value of $\delta$ and $\omega$. Below is an example of how to solve *equation (8)* in Filipović, Pelger and Ye (2022):
The weight function is $\omega(x) = e^{\alpha x}$, if $\alpha = 0, \delta = 1$, then $\phi = 1_{[0,y]}$, and $k(x, y) = \psi(x) = \int_0^x 1_{[0,y]}dt = min\{x, y\}$.

# 4 Gaussian Process and Bayesian Perspective

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

## 4.1 Noised Gaussian Process

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 $\boldsymbol{K}_{X,X} = k(\boldsymbol{x}, \boldsymbol{x}^\top)$ 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 $\boldsymbol{z} = (\boldsymbol{z_1}, \boldsymbol{z_2}, \boldsymbol{z_3}, \boldsymbol{.....}, \boldsymbol{z_n})^\top$.
- we assume **y** and $f^*(\boldsymbol{z})$ follow a $(M+n)$ multivariate normal distribution(MVN):

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

  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

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

## 4.2 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 $\boldsymbol{x} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{.....}, \boldsymbol{x_M})^\top$, and for each $\boldsymbol{x_i}, i = 1, 2, ..., M, x_i \in \mathbb{R}^N$, then we can obtain a series of $f(\boldsymbol{x}) = (f(\boldsymbol{x_1}), f(\boldsymbol{x_2}), ..., f(\boldsymbol{x_M}))^\top$. We assume that the series of f(**x**) follows a multivariate normal distribution which is:

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

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 $k(\boldsymbol{x_i}, \boldsymbol{x_i}^\top)$. This makes the distribution of $f(\boldsymbol{x})$ to be called Gaussian Process (GP). Note that if we add a noise term $\epsilon \sim \mathcal{N}(\mathbf{0}, \Sigma^{\epsilon})$, then our prior distribution of $y = f(\boldsymbol{x}) + \epsilon \sim$

$\mathcal{N}(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}^\top) + \Sigma^\epsilon)$ 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 $\theta$ 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 $\theta$. The estimated $\theta$ value is determined by training data $D$. This is to say that we update our parameter $\theta$ 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, $\theta$ could be substituted by function $f(\cdot)$. One can show that the joint distribution of $(f(\boldsymbol{x}^*), \boldsymbol{y})^\top$ follows a multivariate normal distribution, 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 Rasmussen and Williams (2006).

## 4.3 Gaussian Process in Filipović, Pelger and Ye (2022)

In Filipović, Pelger and Ye (2022), 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 $\mathcal{N}(m(\boldsymbol{z}), k(\boldsymbol{z}, \boldsymbol{z}^\top))$. 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 Filipović, Pelger and Ye (2022)'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 $Cg(\boldsymbol{x}) \sim \mathcal{N}(Cm^{post}(\boldsymbol{x}), Ck^{post}(\boldsymbol{x}, \boldsymbol{x}^\top)C^\top)$.

Note that authors assume the variance-covariance matrix of error term $\Sigma^\epsilon = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_M^2)$

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 Filipović, Pelger and Ye (2022)'s *equation (5)*.

# 5 Empirical Study

## 5.1 Background knowledge in finance

In this paper, we estimate the discount curve and the yield curve. The discount curve $g(x)$ gives us how much \$1 in the future date $x$ is worth at present. In other words, in the future time $x$, any amount of assets $\$A$ is worth $\$Ag(x)$ at present. In addition, the yield rate in date $x$ is how much interest \$1 will yield in the future time $x$. In other words, \$1 will worth $\$(1 + y(x))$ in the future time $x$. All the yield rates and discount rates refer to the risk-free rate. Furthermore, note that the discount curve $g(x)$ and the yield curve $y(x)$ are equivalent: $y(x) = -\frac{1}{x}ln(g(x))$. The proof can be found in Appendix.

Meanwhile, we will use the concept of yield to maturity (YTM). Here we must clarify the difference between the yield rate in the yield curve and YTM for any specific bond. Mishkin and Eakins (2006) defined YTM as *The interest rate equating the present value of cash flows received from a debt instrument with its value today*. This means for each bond, whether it is a risk-free treasury bond, we always can solve the YTM for this bond. However, yield rate and discount rate in the context of Filipović, Pelger and Ye (2022) refers to risk-free rates and only a zero-coupon Treasury bond has the same interest rate (YTM) as yield rate. Nevertheless, the bond data used in our study contains no zero-coupon Treasury bond, so in the context of this research, YTM and yield rate are different.

## 5.2 Data

In this paper, we only use a part of bond data which are provided by Filipović, Pelger and Ye (2022) to train and test our models, which are the price data and cash flow data observed on December 31st, 2013. The bond types are specified in Filipović, Pelger and Ye (2022). Figure 1 (left) shows the distribution of bond price $P$ against time to maturity. One can find

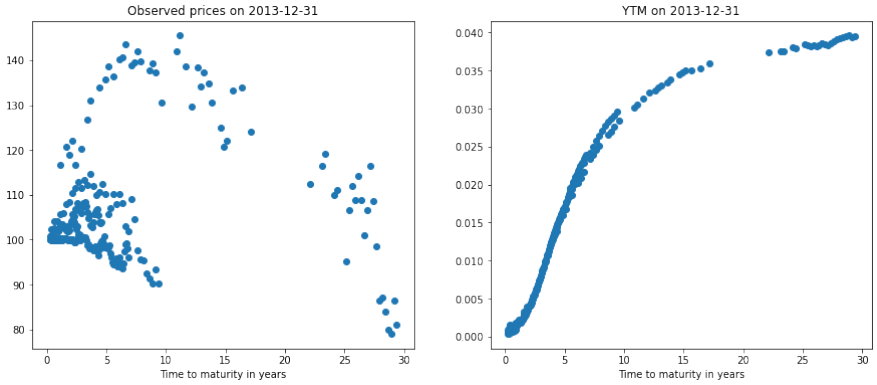


Figure 1: Price distribution and Yield to Maturity

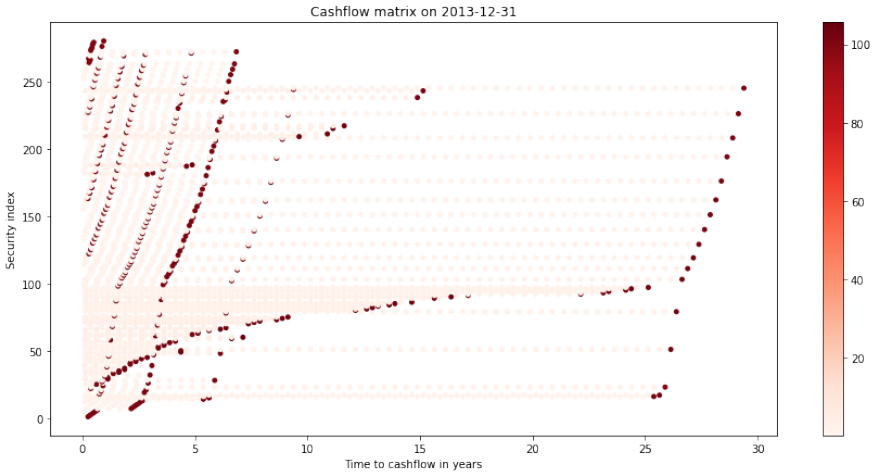


Figure 2: Cash flow

that a big part of bounds has maturity within 10 years, and our data lacks of bounds whose maturity is around 20 years. This means our data is unbalanced and sparse in some time windows. The yield to maturity (YTM) is shown in figure 1 (right).

Figure 2 shows the cash flow distribution, in other words, the cash flow matrix $C$, 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 unbalanced and are sparsely distributed around 20 to 30 years of maturity.

## 5.3 Estimated results

We use the Nelson–Siegel–Svensson (NSS) model to compare with our RKHS estimated discount curve and yield curve. NSS model is a parametric estimation model, in this study

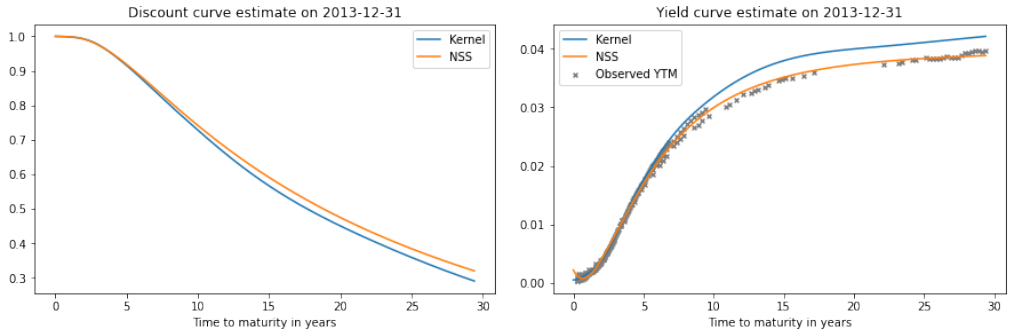


Figure 3: Fitted Discount Curve & Yield Curve

it fits the YTM to infer the yield curve. The formula of NSS is shown below.

$$y(x) = \beta_1 + \beta_2\Big(\frac{1 - exp(\frac{-x}{\lambda_1})}{\frac{x}{\lambda_1}}\Big) + \beta_3\Big(\frac{1 - exp(\frac{-x}{\lambda_1})}{\frac{x}{\lambda_1}} - exp(\frac{-x}{\lambda_1})\Big) + \beta_4\Big(\frac{1 - exp(\frac{-x}{\lambda_2})}{\frac{x}{\lambda_2}} - exp(\frac{-x}{\lambda_2})\Big)$$

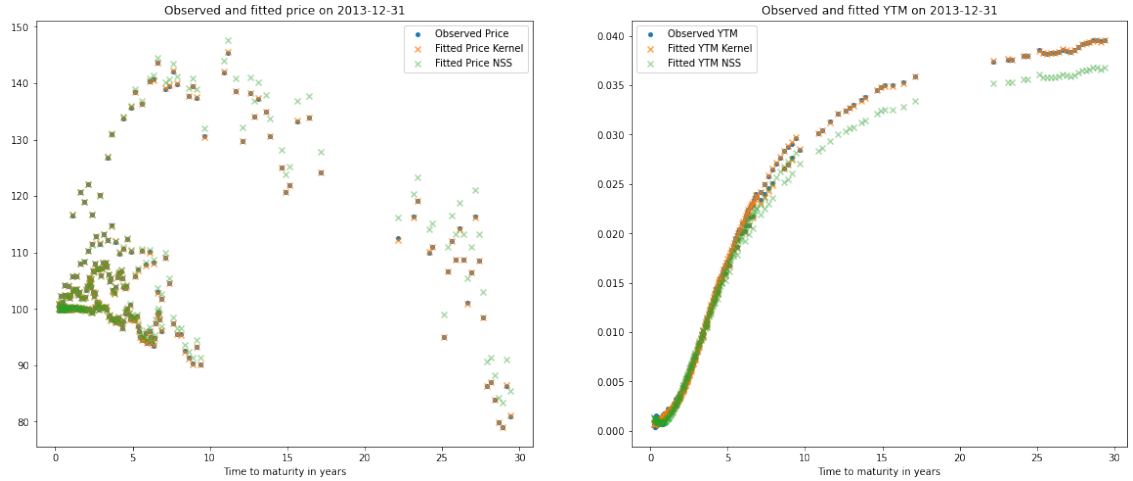


Figure 4: Fitted Results

Figure 3 (left) shows the estimated discount curves, within 5 years of maturity, discount curves perform similarly, but in long term, $g(x)$ estimated by RKSH is lower. Figure 3 (right) shows the yield curves with observed YTM, in the very short term we can find the NSS curve is inverted, which is unnatural in a common date. Meanwhile, we can observe that the NSS yield curve fits the observed YTM (gray crosses) very well, but YTM is not the yield rate, thus, the NSS yield curve is more likely biased.

Figure 4 shows the predicted price and YTM calculated via the predicted price in blue dots.

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 RMSE compared with other methods. On the contrary, NSS generates a much bigger bias in both terms.

## 5.4 Simulation

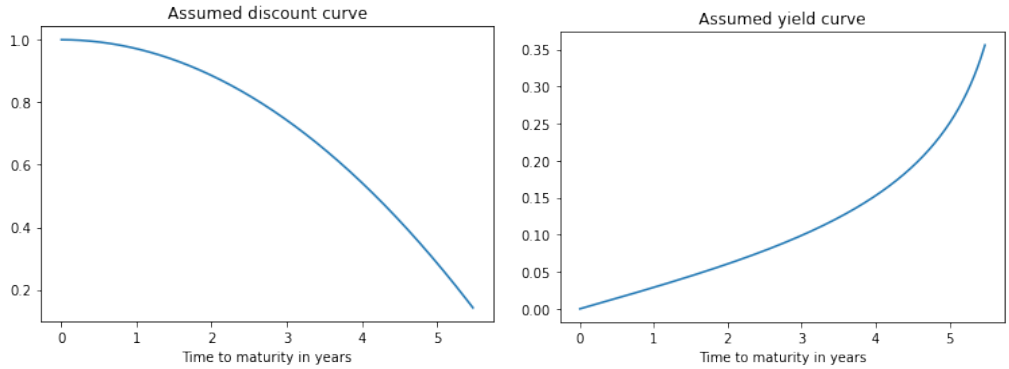


Figure 5: True Discount curve and Yield curve

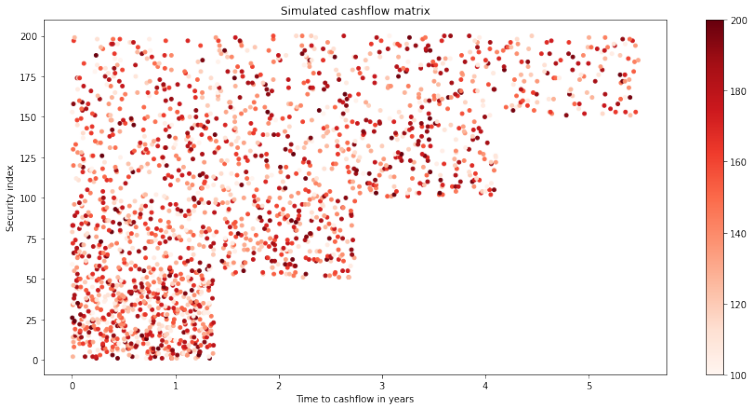


Figure 6: Simulated Cash Matrix

Finally, we can compare and test the kernel model with a simulation study. We assume a true monotone-decreasing discount curve that satisfies a monotone-increasing corresponding yield curve. The true discount curve and yield curve are $g(x) = -\frac{x^2}{35} + 1$, and $y(x) = -\frac{1}{x}ln(g(x))$, where $x$ denotes days in the future. The assumed true curves are shown in figure 5.

Then we simulate the cash flow in around 5 years (actually 2000 days), for 200 bonds, each 50 bonds has 10 cash payments that are randomly sampled from the time intervals $[0, 500], [0, 1000], [0, 1500], [0, 2000]$ (days). Each amount of cash payment is sampled from

uniform distribution $[100, 200]$. The cash flow matrix is shown in figure 6. The true price

$$P_i^{simulated} = \sum_{j=1}^{2000} C_{ij} g(x_j) + \epsilon_i$$

, where $C_{ij}$ is the cash flow for bond $i$ in date $x_j$, and $g(x_j)$ is the discount rate for date $x_j$, error term $\epsilon \sim \mathcal{N}(0, 4)$.

The simulated YTM is calculated by the simulated price and cash flow matrix, which is shown in figure 7. We can find that there are several negative YTM, which are unlikely to happen in reality, but in our study, we can take them as outliers.

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

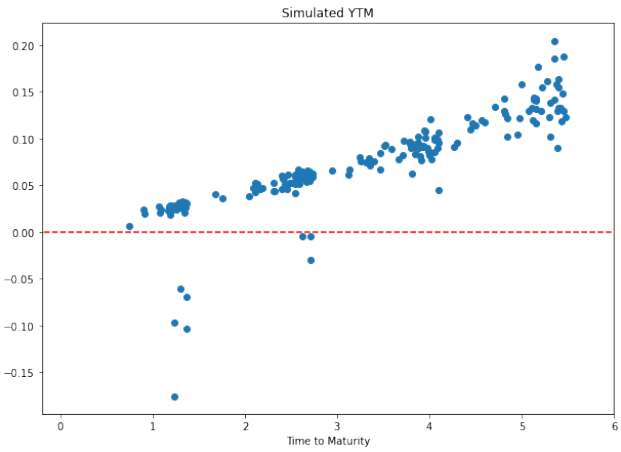


Figure 7: Simulated YTM

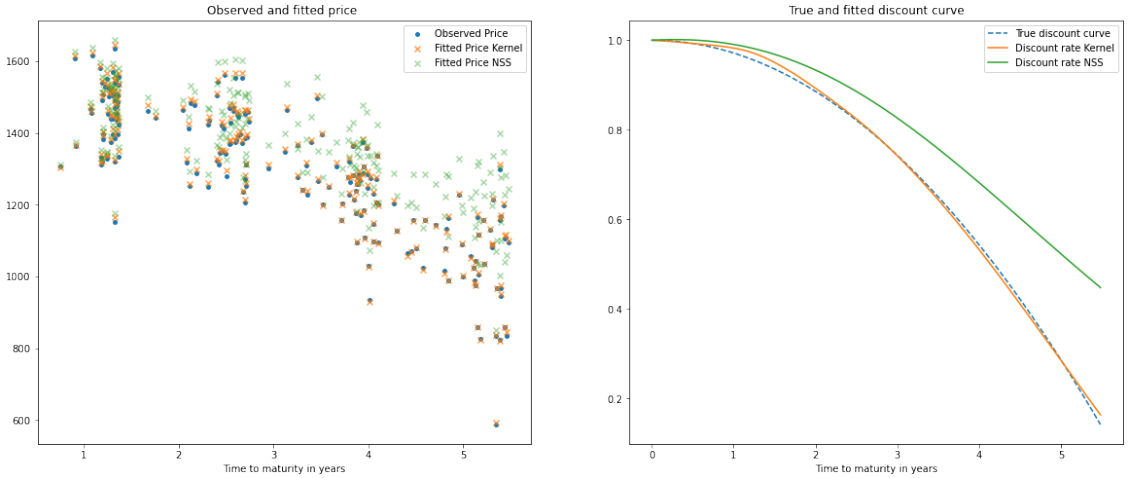


Figure 8: Simulated Result

in figure 8. In such a simulated case, we can find that the performance of the kernel method

outperforms NSS. The pricing bias for NSS is much larger than the kernel estimated price. Last but not least, the kernel-estimated discount curve is very similar to the true discount curve, but the NSS-generated discount curve biases a lot. Given the fact in figure 7 that there are several outliers, we can observe in figure 8 that the kernel estimated discount curve is slightly biased around 1-year maturity, the same time period for strong outliers. However, the kernel-estimated discount curve is still very robust and quite precise in other time periods. Filipović, Pelger and Ye (2022) also showed the robustness advantage of RKHS model over with Liu and Wu (2021) model.

# 6 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 them. Liu and Wu (2021) also proposed a kernel estimation but it is local, that only fits and makes smoothness with nearby eight bonds, and it turns out that the RKHS model is global and more robust.

Filipović, Pelger and Ye (2022) also proposed 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. Our study and simulation have shown the advantage of the kernel model over the parametric model.

In summary, the kernel method and RKHS estimation give us a non-parametric discount curve estimation, enabling us to exploit 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 financial 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).

Gretton, Arthur. 2013. "Introduction to rkhs, and some simple kernel algorithms." *Adv. Top. Mach. Learn. Lecture Conducted from University College London* 16:5–3.

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.

Mishkin, Frederic S and Stanley G Eakins. 2006. *Financial markets and institutions*. Pearson Education India.

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.".

Theodoridis, Sergios and Konstantinos Koutroumbas. 2006. *Pattern recognition*. Elsevier.

# Appendix

## Feature Map

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. Function $\phi(\cdot)$ is 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 two-dimensional vectors into a four-dimensional feature space through function $\phi(\cdot)$.

## A simple proof kernel function

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

$$\begin{aligned} K(\mathbf{x}, \mathbf{y}) &= \langle \mathbf{x}, \mathbf{y} \rangle^2 \\ &= (x_1y_1 + x_2y_2)^2 \\ &= x_1^2y_1^2 + 2x_1y_2x_2y_1 + x_2^2y_2^2 \\ &= \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle \end{aligned}$$

## General kernel ridge regression solution–using representer theorem

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 $\mathcal{H}$. 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 $\Omega$. We want to estimate the function $g(\cdot)$ to minimize

$$\min_{g \in \mathcal{H}} \sum_{i=1}^{N} (y_i - g(x_i))^2 + \Omega \parallel g \parallel_{\mathcal{H}}^2 \tag{4}$$

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

$$g(\cdot)^* = \sum_{i=1}^{N} \alpha_i k(\cdot, x_i), \tag{5}$$

where $k(\cdot, x_i) \in \mathcal{H}$, and according to reproducing property of RKHS, we have

$$g(x) = \langle g(\cdot), k(\cdot, x) \rangle_{\mathcal{H}}. \tag{6}$$

Also, it is obvious that

$$\| g \|^2 = \langle g(\cdot), g(\cdot) \rangle_{\mathcal{H}}. \tag{7}$$

We then substitute $(6)$, $(7)$ for $(4)$,

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

and plug $(5)$ in $(4)$ and get

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

$$\Rightarrow \min_{\alpha} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{N} \alpha_j k(x_j, x_i))^2 + \Omega \sum_{i=j}^{N} \sum_{i=1}^{N} \alpha_j \alpha_i k(x_i, x_j) \tag{8}$$

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 (8) as:

$$\| y_i - K\mathbf{a} \|^2 + \Omega \mathbf{a}' K \mathbf{a}. \tag{9}$$

By differentiation and setting the first order derivative of (9) to zero, we get:

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

Thus, if we are given kernel function $k(x, y)$, we can solve this problem by calculating kernel matrix $K$.

## Equivalence of $g(x)$ and $y(x)$

Proof sketch:

for \$1, interest rate $R$ and year $x$, if in 1 year we count interest $m$ times, then \$1 in future $x$ years is $(1+\frac{R}{m})^{mx}$.

For continuous compounding let $m \to \infty$, we have $\lim_{m\to\infty}(1+\frac{R}{m})^{mx} = e^{xR}$

Here, in the risk-free case, $R = y(x)$, is what \$1 will yield

$\Rightarrow$ in the future $x$, \$1 will yield and keep value $e^{xy(x)}$,

and future \$1 will worth $g(x)$ at present

$\Rightarrow e^{xy(x)} = g(x)^{-1} \Leftrightarrow y(x) = -\frac{1}{x}ln(g(x))$