Stripping the Discount Curve—a Robust Machine Learning Approach

Research Module of Econometrics Presentation

Lindi Li, Gewei Cao
3460570, 3461232

Department of Economics, University of Bonn

16th, January, 2023

Outline

Introduction

Fundamental finance

- **Treasury securities** are government debt instruments backed by full faith and credit of the United States
- **Discount or yield curve** shows the relationship between the time to maturity of the debt and the risk-free interest rate

Research target:

- **Yield curve** $y(x)$ How much \$1 will yield in the future time point x
- **Discount curve** $g(x)$ How much \$1 in the future time point x worth at present
- Only consider the risk-free zero-coupon bond
- Has to be estimated from sparse and noisy Treasury prices

Introduction

Previous literature:

- Parametric method
  - Nelson and Siegel (1987), Svensson (1994), and Gürkaynak, Sack and Wright (2007)
  - Smoothness assumption on the parametric forms of the yield curve
  - Parameters are estimated by minimizing pricing errors.
- Non-parametric method
  - Fama and Bliss (1987), Liu and Wu (2021)
  - Fewer assumptions in the non-parametric method, may lead to overfitting and dynamic instability

Introduction

In this research paper Filipović, Pelger and Ye (2022)
Methodological contributions:

- Develops a data-driven, non-parametric discount curve estimator
- Combines financial theory with machine-learning
- Closed-form solution as simple kernel ridge regression
- Obtain confidence intervals through Gaussian process view

Empirical contributions:

- Extensive out-of-sample study on U.S. Treasury securities
- Smaller out-of-sample yield and pricing errors
- Robust to outliers and stable over time
- Optimal trade-off between flexibility and smoothness

Outline

Hilbert Space

- A **normed space** is a vector space N on which a norm is defined.
- A **Hilbert space** is a normed space whose norm is induced by an inner product $\langle f, g \rangle$ by the relation:

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

- Examples of inner product $\langle a, b \rangle$ in 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.

Introduction to Kernel

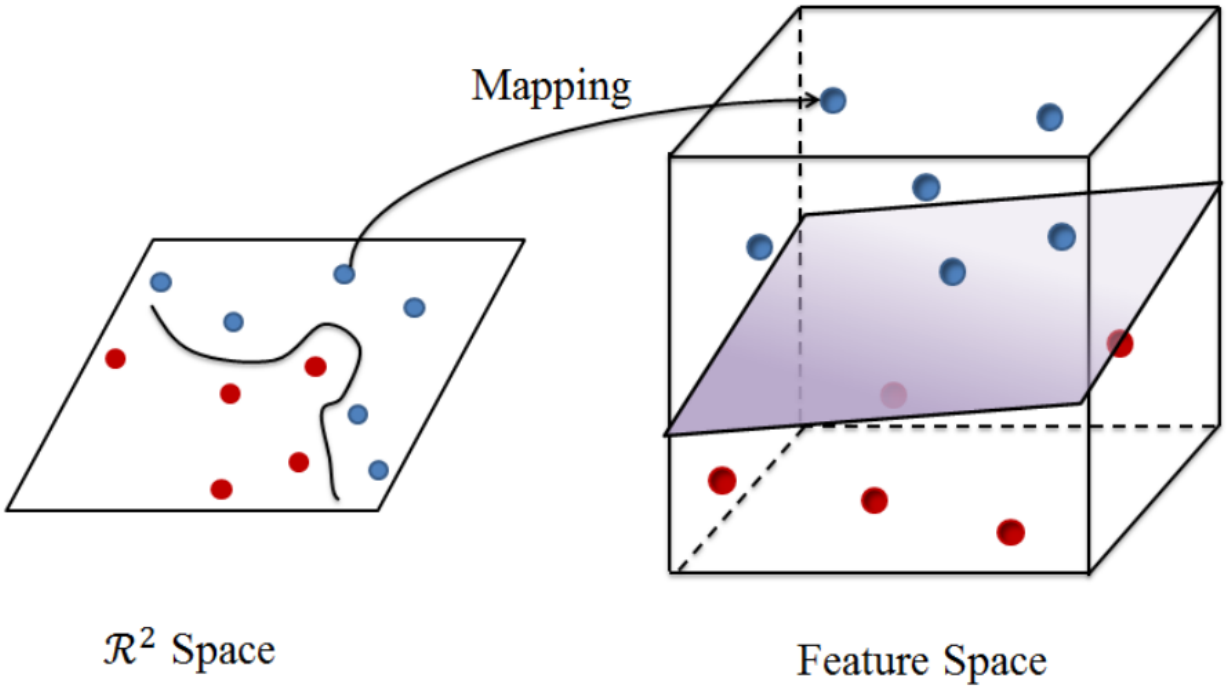


Figure: Feature map

Introduction to Kernel

A simple example to illustrate the idea of a feature map:

- In a two-dimension space, we set two vectors $\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} y_1 & y_2 \end{bmatrix}$.
- Assume there is a function $\phi(\cdot)$ defined as:

$$\phi(x) = \begin{bmatrix} x_1 x_1 & x_1 x_2 & x_2 x_1 & x_2 x_2 \end{bmatrix}$$

$$\phi(y) = \begin{bmatrix} y_1 y_1 & y_1 y_2 & y_2 y_1 & y_2 y_2 \end{bmatrix}$$

We are now successfully mapping them into a four-dimension feature space.

Introduction to Kernel

For many algorithms that study the relation in a dataset in machine learning:

- the raw data has to be explicitly transformed into feature vector representations via a user-specified feature map.
- feature map is infinite-dimensional in most cases.

Introduction to Kernel

Apply feature map in a general form of **linear regression**:

- Set an equation where $\phi(\cdot) \in \mathcal{R}^m$, a vector $\boldsymbol{x}$, and $\phi(\boldsymbol{x})$ is defined as the mapping function and assume there is a linear relation between y and $\phi(\boldsymbol{x})$:

$$\begin{aligned} y &= \phi(\boldsymbol{x})^\top w \\ &= \left[\phi_1(\boldsymbol{x}), \ldots, \phi_m(\boldsymbol{x})\right] w \end{aligned}$$

- Rewrite y and $\phi(\boldsymbol{x})$ in generalization form:

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

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

Introduction to Kernel

Apply feature map in the **regularized risk minimization problem of ridge regression**:

- We get the least-square solution w^*:

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

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

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

- Then we replace $w*$ with equation (1) in $y = \phi(x)^\top w$, we get:

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

Kernel Method

Define a kernel function:

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

$$[\phi(x)^\top \Phi^\top]_j = \phi(x)^\top \phi(x_j) = K(x, x_j) \tag{2}$$

Kernel Method:

- enables them to operate in a high-dimensional, implicit feature space without computing the coordinates of the data in that space
- simply computing the inner products between the images of all pairs of data in the feature space

A Simple Proof

Recall the simple example we used above:

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

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

Define a corresponding kernel function $K(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle^2$.
Proof: $K(\mathbf{x}, \mathbf{y})$ is the same as $\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$.

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

□

Introduction to Kernel

Properties of kernel function: 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 $\sum_i \sum_j \alpha_i \alpha_j k(x_i, x_j) \geq 0, \forall \alpha_i, \alpha_j \in \mathbb{R}, x \in \mathbb{R}^{\mathbb{D}}, \mathbb{D} \in \mathbb{Z}^+$.

Example: Gaussian kernel

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

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**.

Outline

Representer Theorem

- **Good news**: 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.
- **Bad news**: It is extremely difficult to study many popular kernels since their Hilbert spaces are known to be infinite-dimensional in most cases, especially for the purpose of machine learning.
- **Representer Theorem** contributes to simplifying the regularized risk-minimization problem by reducing the infinite-dimensional space to a finite-dimensional space.

Representer theorem

Set

- 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 $\mathcal{H}$

Schölkopf, Herbrich and Smola (2001) find the function f^* in the RKHS $\mathcal{H}$ satisfying:

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

where

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

Note that Ω is a non-decreasing regularizer and y is a vector of y_i.

Representer Theorem

Representer Theorem: the solution to

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

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

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

If Ω is strictly increasing, all solutions apply to this form.

Fundamental Problem of Machine Learning

- How to measure the similarity between samples?
  - Hofmann, Schölkopf and Smola (2008) wrote that the advantage of using a kernel as a similarity measure is that it allows us to construct algorithms in dot product spaces.
- How to weigh the value of each sample?
  - The higher the similarity of the sample to our point of interest x, the more the sampling weights.

Interpretation of the solution in the Representer theorem

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

where α_i the **weighted value** of each sample and k is the **similarity measure**.

Outline

A Simple Example of Using Representer Theorem–Kernel Ridge Regression

- Set an empirical data $(y_1, x_1), \ldots, (y_n, x_n)$, where $i = 1, \ldots, N$
- Assume $y = g(x)$ in RKHS
- Add a penalty term Ω
- We want to estimate the function $g(\cdot)$ to minimize

$$\min_{g \in \mathcal{H}} \sum_{i=1}^{N} (y_i - g(x_i))^2 + \Omega \parallel g \parallel_{\mathcal{H}}^2 \tag{3}$$

A Simple Example of Using Representer Theorem–Kernel Ridge Regression II

According to **Representer theorem**, the optimization problem always has a solution of the form

$$g(\cdot)^* = \sum_{i=1}^{N} \alpha_i k(\cdot, x_i), \tag{4}$$

and recall the reproducing property of RKHS, we have

$$g(x) = \langle g(\cdot), k(\cdot, x) \rangle_{\mathcal{H}}. \tag{5}$$

Also, it is obvious that

$$\| g \|^2 = \langle g(\cdot), g(\cdot) \rangle. \tag{6}$$

A Simple Example of Using Representer Theorem–Kernel Ridge Regression III

We then substitute (5), (6) for (3),

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

and plug (4) in (3) and get

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

$$\Rightarrow \min_{\alpha}\sum_{i=1}^{N}(y_i - \sum_{j=1}^{N}\alpha_j k(x_j, x_i))^2 + \Omega\sum_{i=j}^{N}\sum_{i=1}^{N}\alpha_j\alpha_i k(x_i, x_j) \qquad (7)$$

A Simple Example of Using Representer Theorem–Kernel Ridge Regression IV

Rewrite (7) in the matrix form:

$$\| y_i - K\mathbf{a} \|^2 + \Omega \mathbf{a}' K \mathbf{a}. \tag{8}$$

By differentiation and setting the first-order derivative of (8) to zero, we get:

$$\alpha^* = (K + \Omega I_n)^{-1} y. \tag{9}$$

Settings in Filipović, Pelger and Ye (2022)

- Observe prices $P_1, P_2, P_3, ..., P_M$ for M bonds
- Observe $M \times N$ cash flow matrix C for M bonds in N days
- C_{ij} denotes cash flow in day x_j for bond i
- Time series of days $0 < x_1 < x_2 < x_3 < .... < x_N$
- Fundamental value of bond i

$$P_i^g = \sum_{j=1}^{N} C_{ij} g(x_j)$$

- Observed price

$$P_i = P_i^g + \epsilon_i$$

Settings in Filipović, Pelger and Ye (2022) II

- Model the discount curve $g(x)$ as

$$g = p + h$$

  - p is some exogenous prior curve: $[0, \infty) \to \mathbb{R}$ with $p(0) = 1$. In the paper, they assume that the constant $p = 1$.
  - h is a hypothesis function which is optimally chosen from a RKHS $\mathcal{H}$ consisting of functions $h : [0, \infty) \to \mathbb{R}$.

Minimization

- Weighted mean square

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

- Penalize smoothness

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

- Trade-off

$$\min_{g \in G_{\alpha,\delta}} \{\sum_{i=1}^{M} \omega_i (P_i - P_i^g)^2 + \lambda ||g||^2_{\alpha,\delta}\}$$

$$\min_{g \in G_{\alpha,\delta}} \{\sum_{i=1}^{M} \omega_i (P_i - C_i(p(x) + h(x))^2 + \lambda ||g||^2_{\alpha,\delta}\}$$

Minimization II

- Weight

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

where D_i is the duration of bond i

Duration

Mishkin and Eakins (2006): Duration is a weighted average of the maturities of the cash payments.

Solved Discount Function

Recall the minimize function (3) and its solution (9) in our simple example:

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

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

In Filipović, Pelger and Ye (2022), it is the trade-off function:

$$\min_{g\in G_{\alpha,\delta}} \{\sum_{i=1}^{M} \omega_i (P_i - P_i^g)^2 + \lambda ||g||_{\alpha,\delta}^2\}$$

Solved Discount Function II

Solve the function in an analogous way, we get h given by

$$\hat{h} = k(\cdot, x)^\top \beta,$$

and the discount function $g(x)$ is given by

$$\hat{g}(x) = 1 + \sum_{j=1}^{N} k(x, x_j)\beta_j$$

and the parameter—the weight of different kernel functions:

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

where

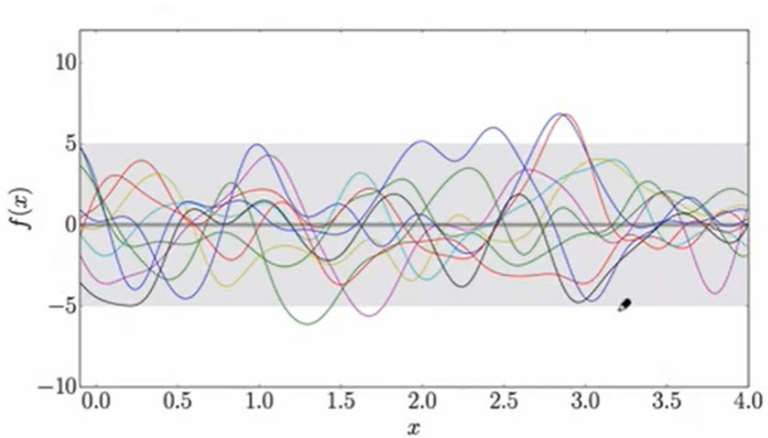
$$\Lambda = diag(\frac{\lambda}{\omega_1}, ..., \frac{\lambda}{\omega_M})$$

Outline

Gaussian Process Idea

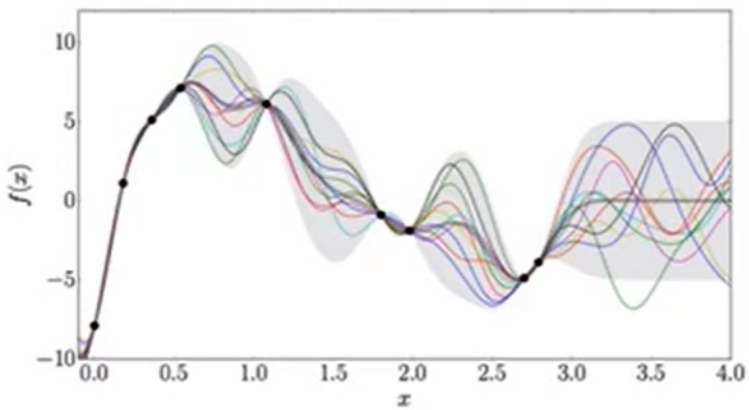
How can we do statistical inference for such a non-parametric estimation?

Figure: Prior[1]



10 samples from the prior distribution using a squared exponential kernel (Eq. 1) with $l = 0.2$ and $\sigma = 2.5$. The dark grey line indicates $m(x)$, the gray area indicates the 95% confidence region, i.e. $m(x) \pm 2\sqrt{K(x,x)} = m(x) \pm 2\sigma$.

Figure: Posterior[1]



10 samples from the posterior distribution using a squared exponential kernel (Eq. 1) with $l = 0.2$ and $\sigma = 2.5$. The dark grey line indicates $\mathbb{E}[f(x)]$, the grey area indicates the 95% confidence region.

[1] https://www.youtube.com/watch?v=30CvM72VXnct=113s

Gaussian Process Intuition

- Gaussian Process (GP) $f(\boldsymbol{x}) \sim \mathcal{N}(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}^T))$
- GP $y = f(\boldsymbol{x}) + \epsilon \sim \mathcal{N}(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}^T) + \Sigma^{\epsilon})$
- Bayesian rule for parametric estimation
  $p(\boldsymbol{y}^*|\boldsymbol{x}^*, \boldsymbol{x}, \boldsymbol{y}) = \int_\theta p(\boldsymbol{y}^*, \theta|\boldsymbol{x}^*, \boldsymbol{x}, \boldsymbol{y})d\theta = \int_\theta p(\boldsymbol{y}^*|\theta, \boldsymbol{x}^*)p(\theta|\boldsymbol{x}, \boldsymbol{y})d\theta$
- Bayesian rule for non-parametric estimation
  $p(f(\boldsymbol{x}^*)|\boldsymbol{x}^*, \boldsymbol{x}, \boldsymbol{y}) = \int p(f(\boldsymbol{x}^*)|f, \boldsymbol{x}^*)p(f|\boldsymbol{x}, \boldsymbol{y})df$
- Bayesian updating: posterior with given data

Gaussian Process Settings

- given data $\mathbf{D} = \{(\boldsymbol{x_i}, y_i)\}_{i=1}^{M}$, and $\boldsymbol{x_i} \in \mathbb{R}^N$
- find the distribution of $f^*(x)$, $f(\boldsymbol{x_i}) : \mathbb{R}^N \to \mathbb{R}$
- prediction function is: $y_i = f(\boldsymbol{x_i}) + \epsilon_i$ and $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$
- $\boldsymbol{y} \in \mathbb{R}^M$ and $\boldsymbol{X}$ is $M \times N$ matrix.
- $\epsilon \sim \mathcal{N}(\mathbf{0}, \Sigma^\epsilon)$, and $\Sigma^\epsilon = diag(\sigma_1^2, \sigma_2^2, \sigma_3^2, ..., \sigma_M^2)$
- $n \times N$ matrix $\mathbf{Z}$ and predicted value $f^*(\boldsymbol{z}) \in \mathbb{R}^n$
- $$\begin{bmatrix} f^*(\boldsymbol{z}) \\ \boldsymbol{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_{f^*(\boldsymbol{z})} \\ \mu_{\boldsymbol{y}} \end{bmatrix} , \begin{bmatrix} \boldsymbol{K}_{Z,Z} & \boldsymbol{K}_{Z,X} \\ \boldsymbol{K}_{X,Z} & \hat{\boldsymbol{K}}_{X,X} \end{bmatrix} \right)$$

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

Gaussian Process Posterior Distribution

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

Rasmussen and Williams (2006)

Gaussian Process in Filipović, Pelger and Ye (2022)

- Estimate discount function $g(\boldsymbol{z})$ given a vector of different maturties $\boldsymbol{z} = (z_1, z_2, ..., z_n)$
- Gaussian Process $g(\boldsymbol{z}) \sim \mathcal{N}(m(\boldsymbol{z}), k(\boldsymbol{z}, \boldsymbol{z}^T))$
- Bayesian updating for given price P, corresponding cash flow matrix C, and time to maturities x
- $m^{post}(x) = m(x) + k(x, \boldsymbol{x})^\top C^\top (C\boldsymbol{K}C^\top + \sum^{\epsilon})^{-1}(P - Cm(\boldsymbol{x}))$
- $k^{post}(x, y) = k(x, y) - k(x, \boldsymbol{x})^\top C^\top (C\boldsymbol{K}C^\top + \sum^{\epsilon})^{-1} Ck(\boldsymbol{x}, y)$
- where $\Sigma^{\epsilon} = diag(\sigma_1^2, \sigma_2^2, \sigma_3^2, \ldots, \sigma_M^2)$ where diagonal elements all satisfy $\omega_i = \frac{\lambda}{\sigma_i^2}$
- distribution of bound price: $Cg(\boldsymbol{x}) \sim \mathcal{N}(Cm^{post}(\boldsymbol{x}), Ck^{post}(\boldsymbol{x}, \boldsymbol{x}^\top)C^\top)$

Outline

Empirical Basic Concepts

- **Yield curve** $y(x)$ What \$1 will yield in the future time point x
- **Discount curve** $g(x)$ What \$1 in the future time point x worth at present
- Aim : estimate $g(x)$ and corresponding $y(x)$
- $y(x) = -\frac{1}{x} ln(g(x))$

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

Yield to Maturity

Mishkin and Eakins (2006): The interest rate equates the present value of cash flows received from a debt instrument with its value today.

Example:
A bond with cash flows from years 1 to 10 are ($110.0, $121.0, $133.1, $146.4, $161.1, $177.2, $194.9, $214.4, $235.8, $259.4), and its price is $1000. Then YTM of this bound is the solution for

$$\$1000 = \frac{\$110.0}{(1+YTM)} + \frac{\$121.0}{(1+YTM)^2} + \frac{133.1}{(1+YTM)^3} + \ldots\ldots + \frac{259.4}{(1+YTM)^{10}}$$

Then $YTM = 10\%$

Empirical Basic Concepts

Difference between YTM and yield rate

Different from yield to maturity (YTM), the yield curve denotes what \$1 will yield in the future time x without any risk, so it is the risk-free interest rate. On the other hand, YTM is the annualized interest rate for any specific bond, it is the "average" interest rate for a bond. Only if a bond only makes 1 payment at its maturity, and without noise in pricing, then the yield rate equals YTM.

Yield Rate

Yield rate as a risk-free rate is widely used in finance and economics modeling and investment decision.

Empirical Data

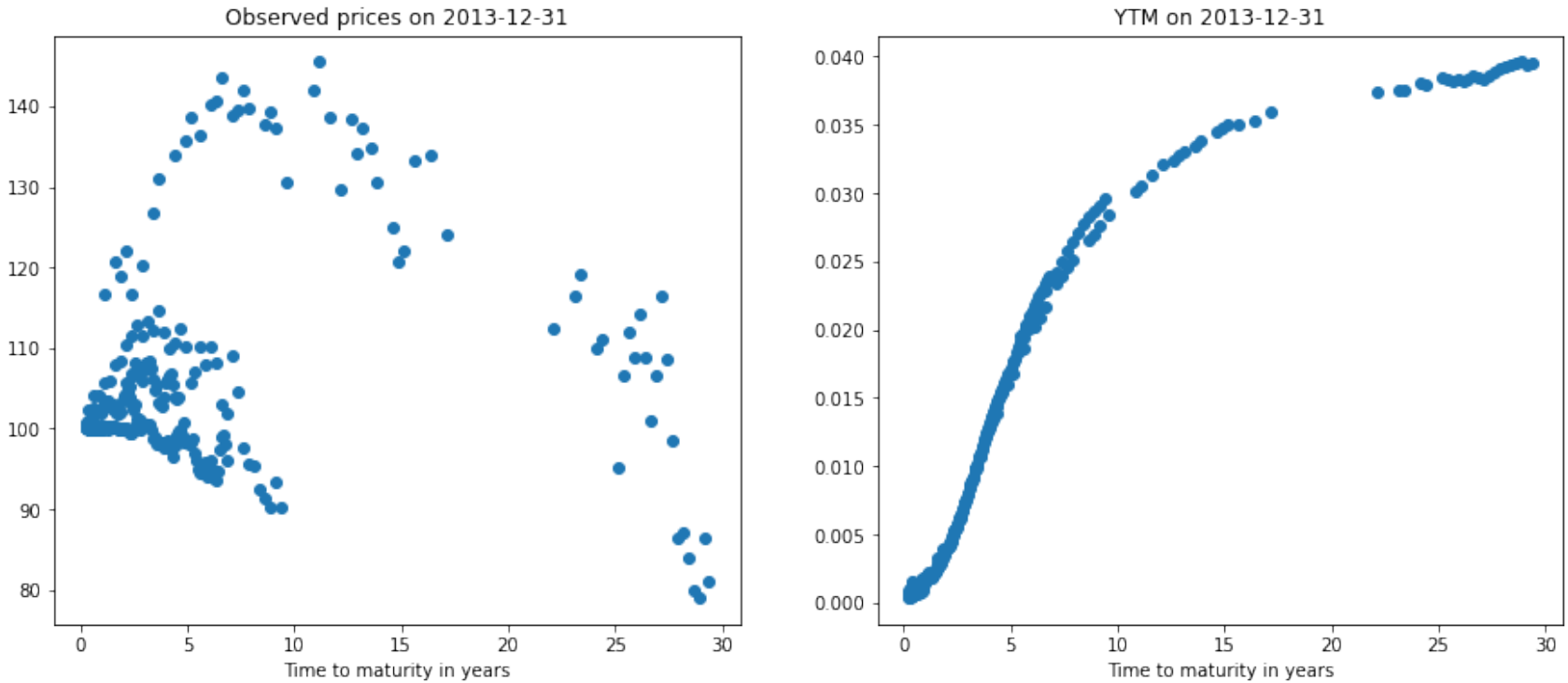


Figure: Observed Price and YTM

Empirical Data

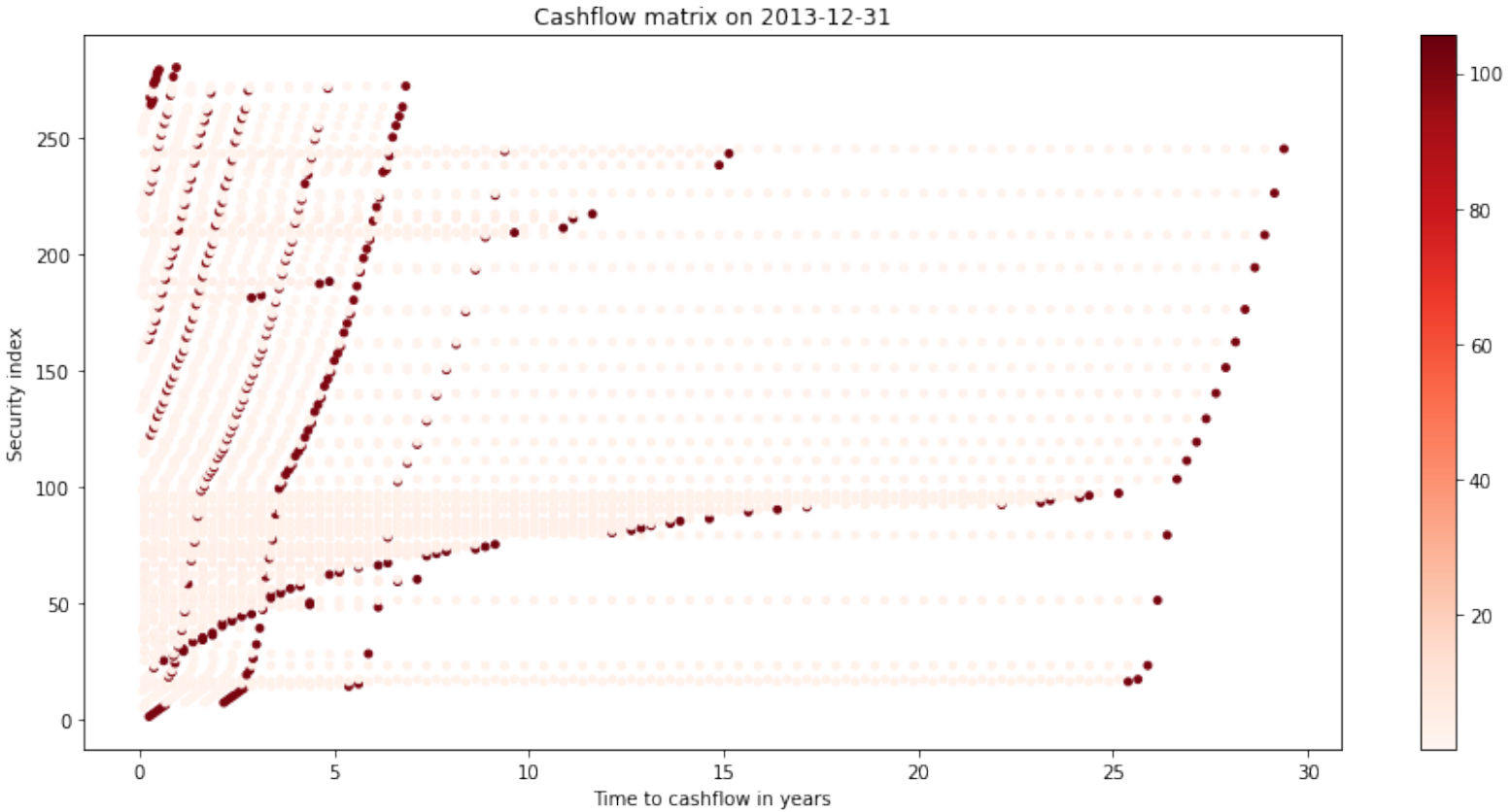


Figure: Observed Cash Flow

Empirical NSS model

Nelson–Siegel–Svensson (NSS) model is a widely used parametric method to estimate the discount curve, proposed by Nelson and Siegel (1987) and Svensson (1994)
NSS can fit the observed YTM directly, and obtain the yield curve, but this property is demanding in bond types.

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

Empirical Estimated Results

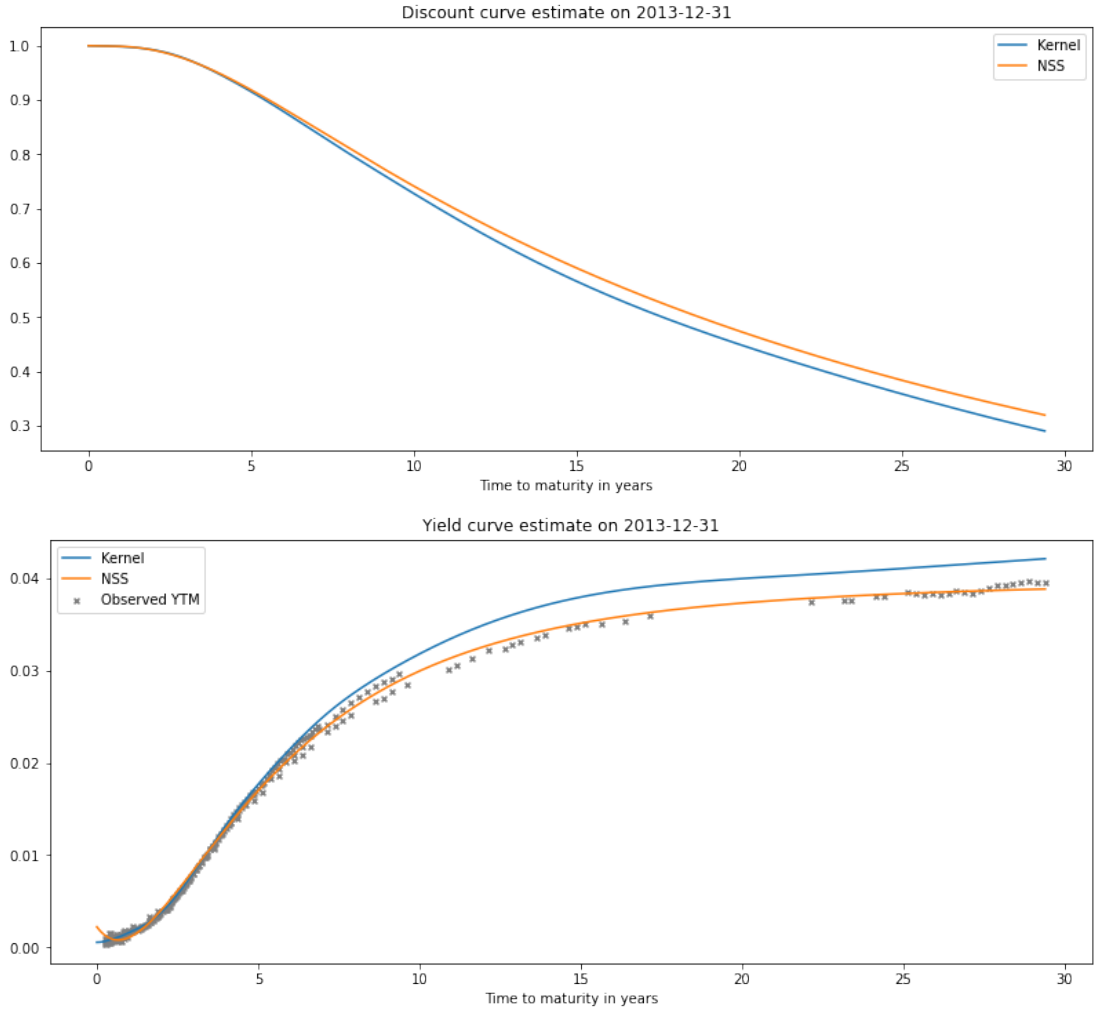
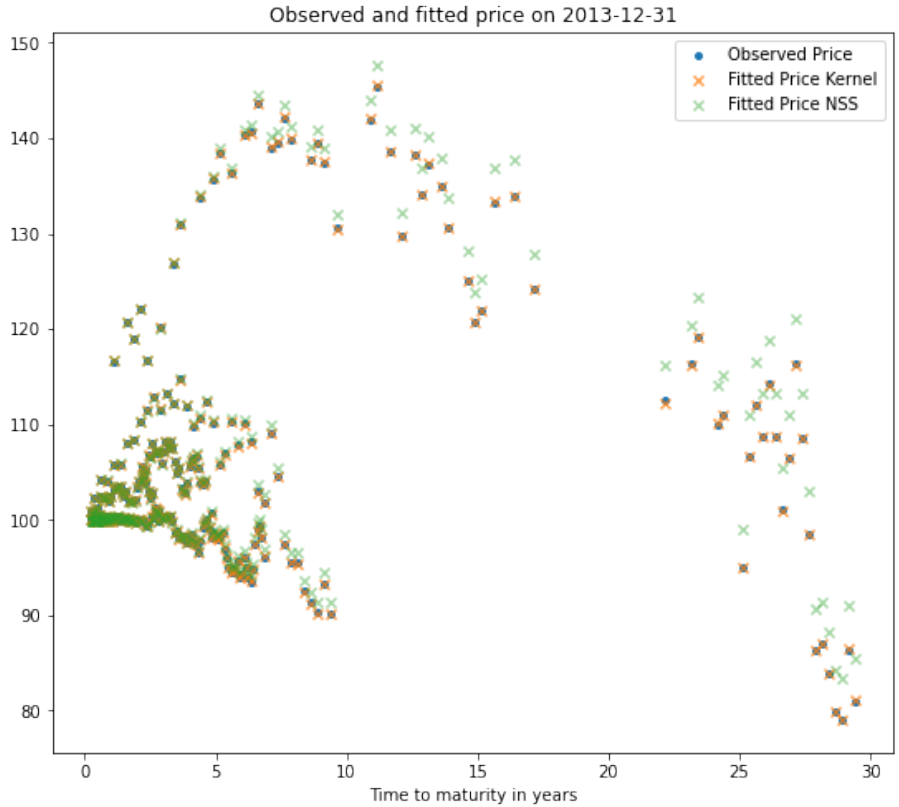


Figure: Estimated Discount Curve and Yield Curve

Empirical Fitted Results



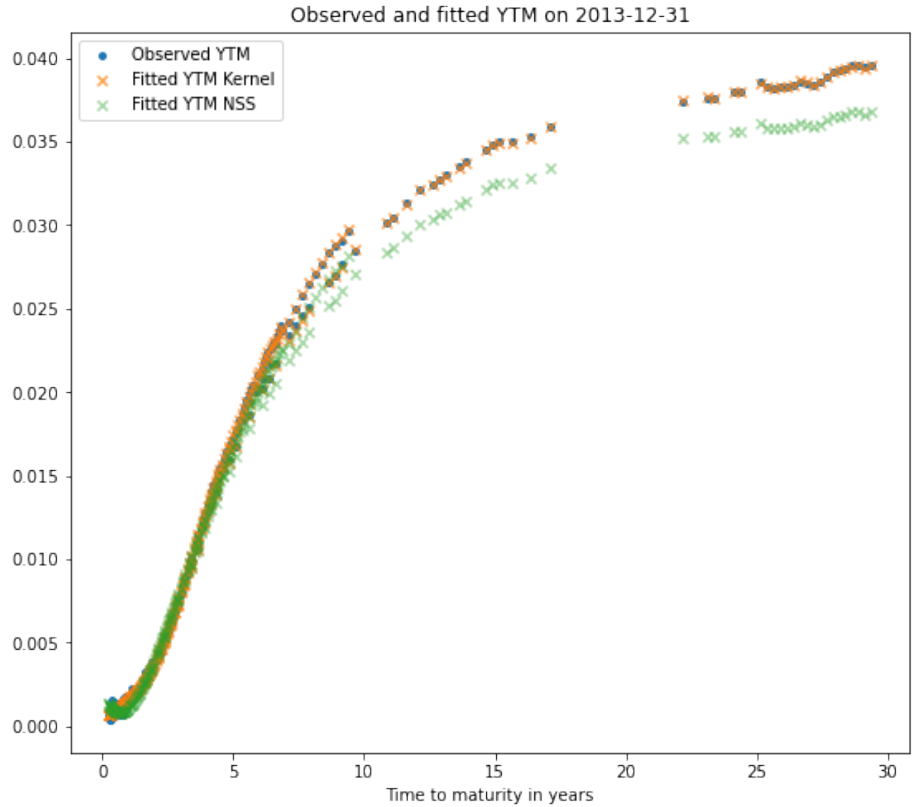


Figure: Fitted Price and corresponding YTM

Empirical Simulation

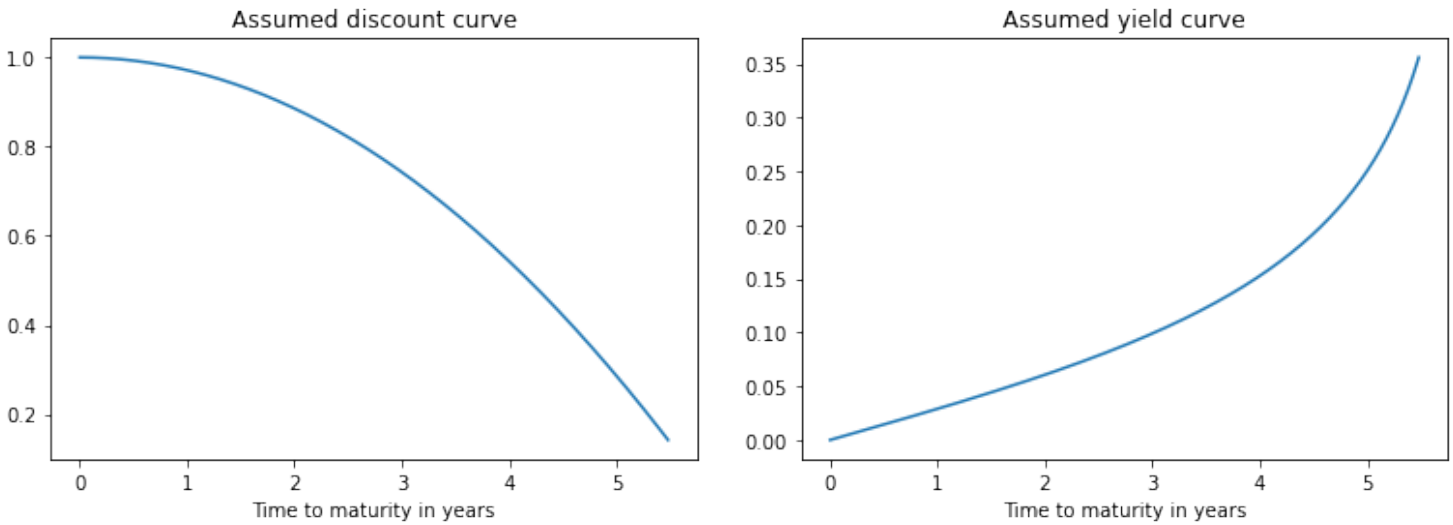


Figure: Simulated Discount Curve and Yield Curve

The function of the discount curve:

$$g(x) = -\frac{x^2}{35} + 1, \text{ and } y(x) = -\frac{1}{x} ln(g(x))$$

Empirical Simulation DGP

- Total 2,000 days, around 5 years
- 200 bonds, each has 10 cash payments, payment date and amount are random
- Maximum possible maturities are 500 days, 1000 days, 1500 days, and 2000 days, each for 50 bounds, respectively
- Each cash flow payment is randomly sampled from a uniform distribution within $[100, 200]$
- 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 $\epsilon \sim \mathcal{N}(0, 4)$

Empirical Simulation Data

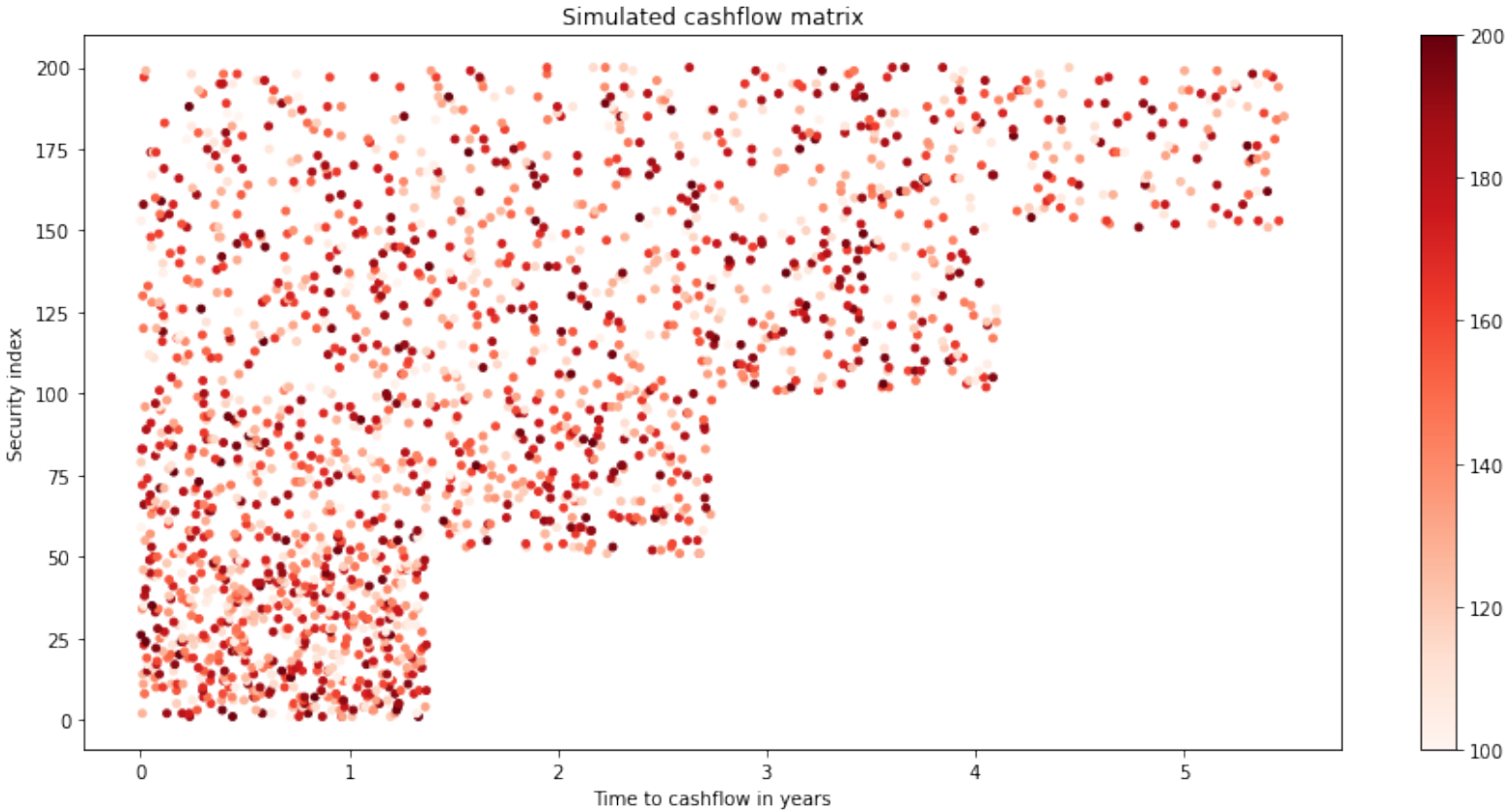


Figure: Simulated Cash Flow

Empirical Simulation Data

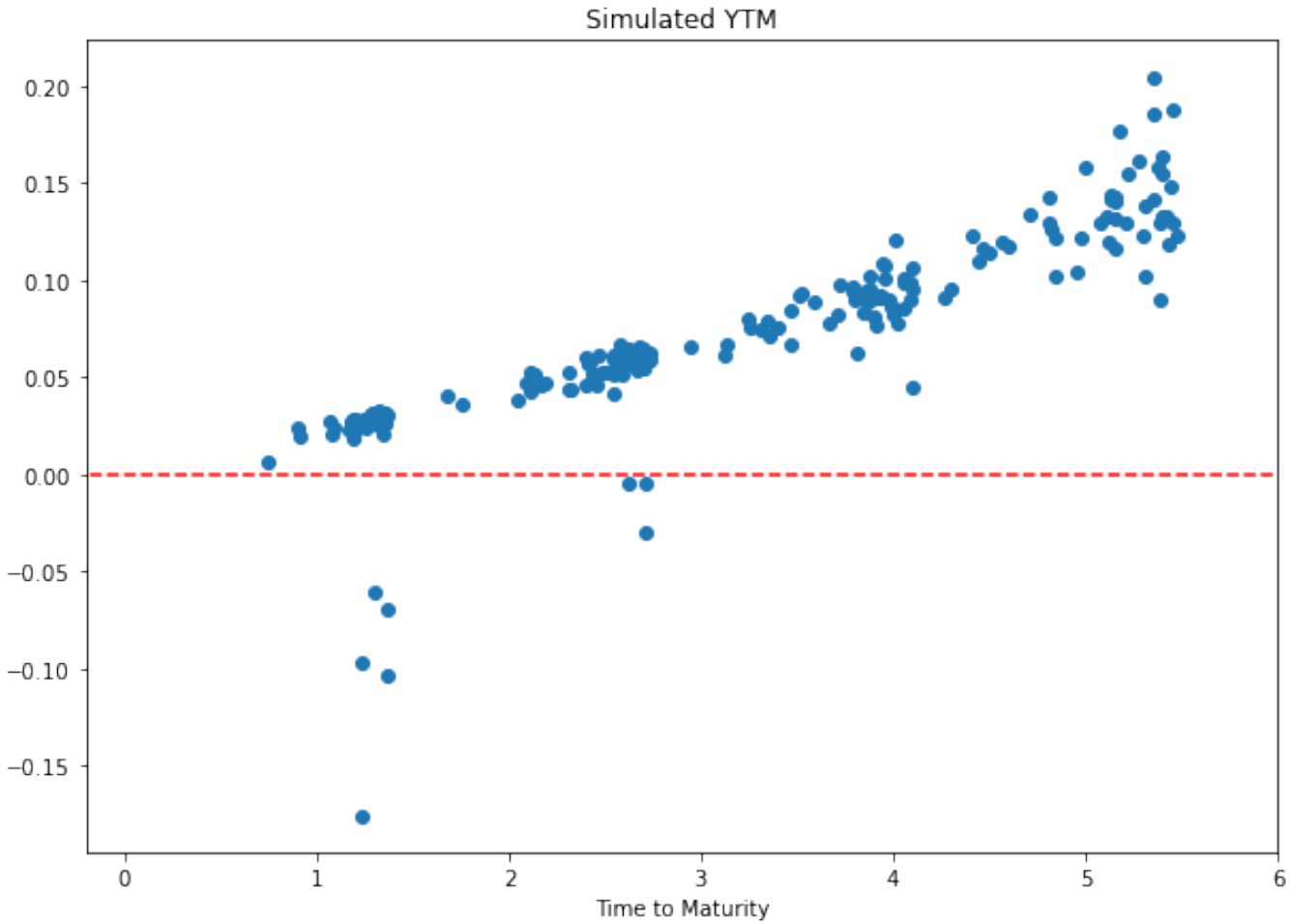
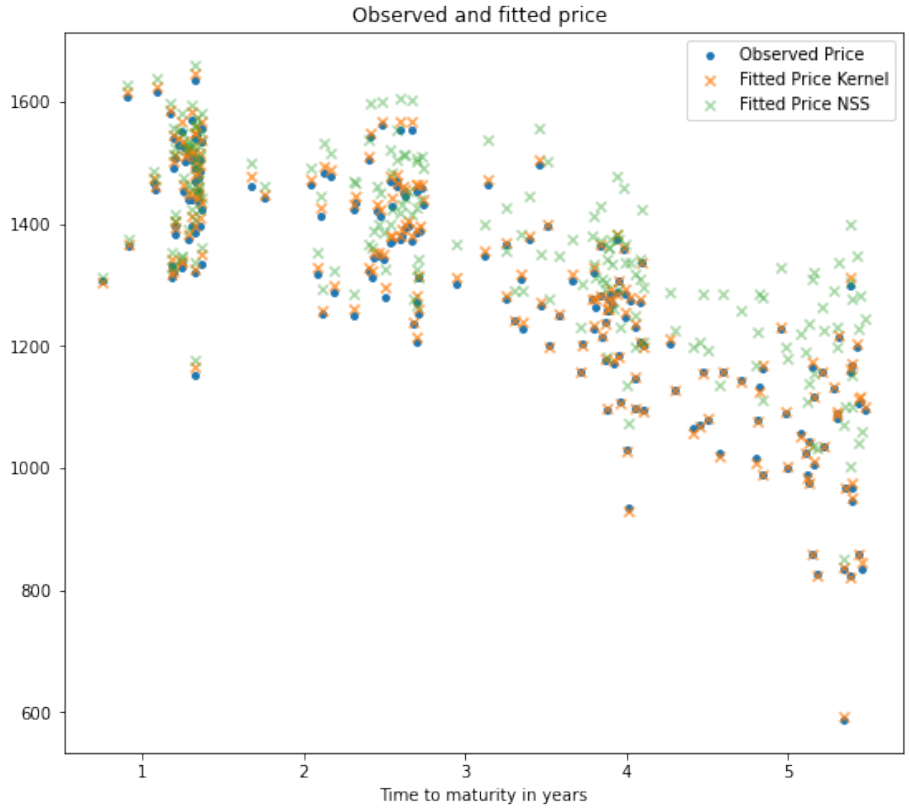


Figure: Simulated Yield to Maturity

Empirical Simulation Results



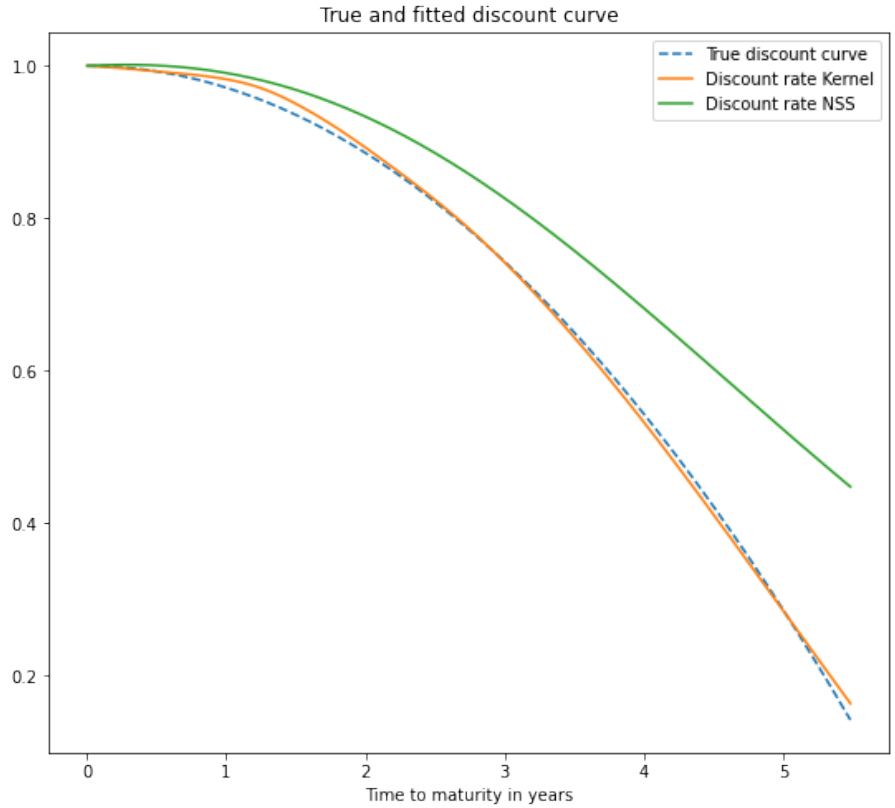


Figure: Fitted results with simulated data

References I

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.

References II

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.

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.".