2017 International Conference on Computational Intelligence in Data Science(ICCIDS)

A Statistical Machine Learning Approach to Yield Curve Forecasting

Rajiv Sambasivan
Sourish Das
Chennai Mathematical Institute
H1, SIPCOT IT Park, Siruseri
Kelambakkam 603103
Email: rsambasivan@cmi.ac.in, sourish@cmi.ac.in

***Abstract*—Yield curve forecasting is an important problem in finance. In this work we explore the use of Gaussian Processes in conjunction with a dynamic modeling strategy, much like the Kalman Filter, to model the yield curve. Gaussian Processes have been successfully applied to model functional data in a variety of applications. A Gaussian Process is used to model the yield curve. The hyper-parameters of the Gaussian Process model are updated as the algorithm receives yield curve data. Yield curve data is typically available as a time series with a frequency of one day. We compare existing methods to forecast the yield curve with the proposed method. The results of this study showed that while a competing method (a multivariate time series method) performed well in forecasting the yields at the short term structure region of the yield curve, Gaussian Processes perform well in the medium and long term structure regions of the yield curve. Accuracy in the long term structure region of the yield curve has important practical implications. The Gaussian Process framework yields uncertainty and probability estimates directly in contrast to other competing methods. Analysts are frequently interested in this information. In this study the proposed method has been applied to yield curve forecasting, however it can be applied to model high frequency time series data or data streams in other domains.**

I. Introduction and Motivation

Accurate yield curve forecasting is of critical importance in financial applications. Investors watch the bond market closely as it is a very good predictor of future economic activity and levels of inflation. Future economic activity and levels of inflation affect prices of goods, stocks and real estate. The yield curve is a key representation of the state of the bond market. The slope of the yield curve is an important indicator of short term interest rates and is followed closely by investors. (see [1]). As a consequence, this has been the focus of considerable research. Several statistical techniques and tools commonly used in econometrics and finance have been applied to model the yield curve (see for example, [2],[3] and [4]). In this work, we took a machine learning perspective on this problem. Gaussian Processes (GP) are a widely used machine learning technique ([5]). We propose a dynamic method that uses Gaussian Processes to model the yield curve. Yield curve data can be viewed as functional data. Gaussian Process regression has been applied with great success in many domains. Results from this study suggest that Gaussian Process regression performs better than methods currently used for yield curve forecasting in the medium and long term regions of the yield curve. Achieving higher accuracy at longer term structures is more difficult than with the shorter term structures. This is because data points at the longer term structure region of the yield curve are farther apart than in the short term region. A multivariate time series based approach is also commonly used to model the yield curve. This technique had the best results in the short term region of the yield curve. This suggests that these two techniques could be used together. The multivariate time series based approach could be used for short term forecasts and the GP approach could be used for medium and long term forecasting.

The dynamic Gaussian Process method has been applied to model yield curve data in this work. However, functional data presents as a time series in many domains. For example, the hourly user requests processed at a data center could be viewed as functional data. The hourly user traffic for a day may be a variable we wish to forecast. In [6], the daily sea ice surface area in the arctic region observed in one year periods is treated as functional data. Observed sea ice surface area for years passed, could be used to forecast the sea ice surface area for a future year. This suggests that the method proposed in this study could be useful in other application domains too. This is an area of future work.The rest of this paper is organized as follows. In section II, we present an overview of relevant aspects of functional data analysis. In section III, the details of the various methods used to model yield curves, including the proposed method, are provided. In section IV, we describe the methodology for validating the performance of the the methods for yield curve forecasting. In section V, we describe the results of the study. Finally in section VI, we present the conclusions from this study.

II. Functional Data Analysis, a Review

In functional data, the data have a functional representation. Yield curve data are represented in terms of the yields associated with a set of term structures. For example, the data for this study consists of 11 terms. We have a yield associated with each term. This constitutes a map (a function) with 11

978-1-5090-5595-1/17/$31.00 ©2017 IEEE

elements between terms and yield. The i^{th} yield curve is modeled as a function that maps terms to yields:

$$y_i = f(\tau_i) + \epsilon_i$$

The function $f(\tau)$, can be represented as:

$$f(\tau) = \sum_{k=1}^{K} \beta_k \phi_k(\tau) = \boldsymbol{\phi\beta} \qquad (1)$$

we say $\boldsymbol{\phi}$ is a basis system for $f(\tau)$. That is,

$$y = \boldsymbol{\phi\beta} + \epsilon.$$

Many basis functions have been used for functional representation, each having a particular niche of applications that it is well suited to. The sine cosine functions of increasing frequencies

$$y_i = \beta_1 + \beta_2 \sin(\omega\tau) + \beta_3 \cos(\omega\tau) + \beta_4 \sin(2\omega\tau) + \beta_5 \cos(2\omega\tau) \ldots + \epsilon_i \qquad (2)$$

forms the Fourier basis, where constant $\omega = 2\pi/P$ defines the period P of oscillation of the first sine/cosine pair. A comparison of Equation 2 with Equation 1 shows that

$$\boldsymbol{\phi} = \{1, \sin(\omega\tau), \cos(\omega\tau), \sin(2\omega\tau), \cos(2\omega\tau)...\}$$

is the Fourier basis and $\beta^T = \{\beta_1, \beta_2, \beta_3, \ldots\}$ are the corresponding unknown coefficients. Examples of other basis are:

- **Nelson-Siegel Basis**: $\boldsymbol{\phi} = \{1, \frac{1-e^{-\lambda\tau}}{\lambda\tau}, \frac{1-e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\}$.
- **Exponential Basis**: $\boldsymbol{\phi} = \{1, e^{\lambda_1 t}, e^{\lambda_2 t} ...\}$
- **Gaussian Basis**: $\boldsymbol{\phi} = \{1, e^{-\lambda(t_1-c)^2}, e^{-\lambda(t_2-c)^2} ...\}$

A. *Parameter Learning with Bayesian Method*

Once the basis has been picked, the β's in Equation 1 need to be determined. Several methods of parameter estimation are available and are used (Ordinary Least Squares, Penalized Least Squares etc). In this work we used a Bayesian method, accordingly. Parameter learning in the methods mentioned above involved learning an optimal representation by minimizing a loss function. These approaches posit that there is a fixed unique set of parameters associated with the functional representation of the yield curve. A contrasting methodology, the Bayesian methodology treats these parameters differently. In Bayesian methodology, the unknown parameters are assumed to be random variables with valid probability measure on the parameter space.

B. *Gaussian Processes*

Consider the model:

$$\boldsymbol{y} = f(t) + \boldsymbol{\epsilon}$$

Where:

$$\boldsymbol{\epsilon} \sim \boldsymbol{N}(0, \sigma_\epsilon^2 \boldsymbol{I}). \text{ This implies } \boldsymbol{y} \sim \boldsymbol{N}(f(t), \sigma_\epsilon^2 \boldsymbol{I}).$$

The function $f(t)$ has the following representation:

$$f(t) = \boldsymbol{\phi\beta} = \sum_{k=1}^{\infty} \phi_k(t)\beta_k,$$

We want to estimate $\boldsymbol{\beta}$. We adopt a Bayesian methodology, so we assume β's are uncorrelated random variables and $\phi_k(t)$ are known deterministic real-valued functions. Then due to **Kosambi-Karhunen-Loeve** theorem, $f(t)$ is a **stochastic process**. If we assume $\boldsymbol{\beta} \sim \boldsymbol{N}(\boldsymbol{0}, \sigma_\epsilon^2 \boldsymbol{I})$, then $f(t) = \boldsymbol{\phi\beta}$ follows a Gaussian process and the induced process on $f(t)$ is known as '**Gaussian Process Prior**'. The prior on $\boldsymbol{\beta}$:

$$p(\boldsymbol{\beta}) \propto \exp\Bigg(-\frac{1}{2\sigma_\epsilon^2}\boldsymbol{\beta}^T\boldsymbol{\beta}\Bigg).$$

The induced prior on $f = \boldsymbol{\phi\beta}$:

$$p(f) \propto \exp\Bigg(-\frac{1}{2\sigma_\epsilon^2}\boldsymbol{\beta}^T\boldsymbol{\phi}^T\boldsymbol{K}^{-1}\boldsymbol{\phi\beta}\Bigg),$$

where the prior mean and covariance of f are given by(see [5]):

$$\begin{aligned} \mathbf{E}[f] &= \phi E[\boldsymbol{\beta}] = \phi\boldsymbol{\beta}_0 = \mathbf{0}, \\ \mathbf{cov}[f] &= \mathbf{E}[f.f^T] = \phi.\mathbf{E}[\boldsymbol{\beta}.\boldsymbol{\beta}^{\mathbf{T}}]\phi^T = \sigma_\epsilon^2\phi.\phi^T = \mathbf{K}. \end{aligned}$$

An alternative generic formulation of the model is:

$$\begin{aligned} f(\tau) &= \mu(t) + W(t), \\ \boldsymbol{y} &= \mu(t) + W(t) + \boldsymbol{\epsilon}, \end{aligned}$$

where $W(\tau) \sim \boldsymbol{N}(\boldsymbol{0}, \mathbf{K})$ and $\mu(\tau)$ is a parametric function. If there are m many points then,

$$\begin{aligned} f &\sim \boldsymbol{N}_m(\mu(\tau), \mathbf{K}), \quad \boldsymbol{\epsilon} \sim \boldsymbol{N}_m(0, \sigma_\epsilon^2 \boldsymbol{I}_m) \\ \boldsymbol{y} &\sim \boldsymbol{N}_m(f(\tau), \mathbf{K} + \sigma_\epsilon^2 \boldsymbol{I}). \qquad (3) \end{aligned}$$

The likelihood function is given by: $L(f|\boldsymbol{y}, \boldsymbol{\phi}, \sigma^2) \propto (\sigma_\epsilon^2)^{-m/2} \exp\Bigg(-\frac{1}{2\sigma_\epsilon^2}(\boldsymbol{y}-f)^T[\mathbf{K} + \sigma_\epsilon^2\boldsymbol{I}]^{-1}(\boldsymbol{y}-f)\Bigg)$, The negative log-likelihood function can then be expressed as:

$$l(f) \propto \frac{1}{2\sigma_\epsilon^2}(\boldsymbol{y}-f)^T[\mathbf{K} + \sigma_\epsilon^2\boldsymbol{I}]^{-1}(\boldsymbol{y}-f).$$

The corresponding negative log-posterior function is:

$$p(f) \propto \frac{1}{2\sigma_\epsilon^2}\Bigg((\boldsymbol{y}-f)^T[\mathbf{K} + \sigma_\epsilon^2\boldsymbol{I}]^{-1}(\boldsymbol{y}-f) + f^T\boldsymbol{K}^{-1}f\Bigg).$$

Hence the induced penalty matrix in the Gaussian process prior is identity matrix. It looks like weighted least square method with L_2 penalty $P(f) = f^T\boldsymbol{K}^{-1}f$. The posterior distribution over these functions is computed by applying Bayes theorem. The posterior is used to make predictions. The estimated value of y for a given t is the mean (expected) value of the functions sampled from the posterior at that value of t. The expected value of the estimate at t_* is given by:

$$\begin{aligned} \hat{f}(t_*) &= E(f|t_*, \boldsymbol{y}) \qquad (4) \\ &= \mu(t_*) + K(t_*, t).[K(t,t) + \sigma_\epsilon^2.\boldsymbol{I}]^{-1}.(\boldsymbol{y} - \mu(t)) \qquad (5) \end{aligned}$$

The variance of the estimate at t_* is given by

$$cov(f_*) = K(t_*, t_*) - K(t_*, t).[K(t,t) + \sigma_\epsilon^2.\boldsymbol{I}]^{-1}.K(t, t_*) \tag{6}$$

III. Forecasting Methods

In this section we discuss the methods use to forecast yield curves. This includes the proposed dynamic Gaussian Process method.

A. Nelson-Siegel Model

The Nelson-Siegel model [7], [3] specifies the yield curve as:

$$y(\tau) = \beta_1 + \beta_2\left(\tfrac{1-e^{-\lambda\tau}}{\lambda\tau}\right) + \beta_3\left(\tfrac{1-e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\right) + \epsilon(\tau), \quad \epsilon(\tau) \sim N(0, \sigma_\epsilon^2) \tag{7}$$

where $y(\tau)$ is the yield at maturity τ. The three factors β_1, β_2 and β_3 are denoted as level, slope and curvature of slope respectively. Parameter λ controls exponentially decaying rate of the loadings for the slope and curvature.
These factors have the following econometric interpretations:

- The factor β_1 captures the strength of the long term component of the yield curve.
- The factor β_2 captures the strength of the short term component of the yield curve.
- The factor β_3 captures the strength of the medium term component of the yield curve.

The goodness-of-fit of the yield curve is not very sensitive to the specific choice of λ [7]. Therefore [3] treated λ as a known quantity. The factors of the Nelson-Siegel model need to be estimated from the data for the yield curve. Yield curve data are instances of a type of data called functional data. When this technique is applied to a successive yield curves, there could be a pattern in the evolution of the coefficients for the Nelson-Siegel model over time. Section III-C provides a mathematical framework to abstract this problem.

B. Multivariate Time Series Forecasting

A common method to model yield curve data is to use a Vector Auto-Regressive model to represent the yields for the term structures. ([8]). An auto-regressive model of order k is represented by:

$$\boldsymbol{y}_i(\tau) = \beta_0 + \beta_1.\boldsymbol{y}_{i-1}(\tau) + \ldots + \beta_k\boldsymbol{y}_{i-k}(\tau) \tag{8}$$

Equation 8 represents a regression of the i^{th} yield curve on the previous k yield curves. A model selection criterion, like the Bayesian Information Criterion is used to determine the optimal order, k, for the data. Forecasting is then performed using the optimal model. The results of modeling are presented in section V.

C. Forecasting the Yield Curve through Nelson-Siegel Parameters

The Dynamic Nelson-Siegel (DNS) model [7], [3] for yield curve has the following representation:

$$y_t(\tau_j) = \beta_{1t} + \beta_{2t}\left(\frac{1-e^{-\lambda\tau_j}}{\lambda\tau_j}\right) + \beta_{3t}\left(\frac{1-e^{-\lambda\tau_j}}{\lambda\tau_j} - e^{-\lambda\tau_j}\right) + \epsilon_t(\tau_j),$$
$$\beta_{it} = \theta_{0i} + \theta_{1i}\beta_{i,t-1} + \eta_i, \quad i = 1, 2, 3$$

here:

- $\epsilon_t(\tau_j) \sim N(0, \sigma_\epsilon^2)$
- $\eta_i \sim N(0, \sigma_\eta^2)$,
- $t = 1, 2, \ldots, T$ represents the time steps in days
- $j = 1, 2, \ldots, m$ represents the term structure or maturity
- $y_t(\tau)$ is the yield for maturity τ (in months) at time t.

The three factors β_{1t}, β_{2t} and β_{3t} are denoted as level, slope and curvature of slope respectively. Parameter λ controls exponentially decaying rate of the loadings for the slope and curvature. The goodness-of-fit of the yield curve is not very sensitive to the specific choice of λ [7]. Therefore [3] chose λ to be known. In practice, λ can be determined through grid-search method. There are eight static parameters $\boldsymbol{\theta} = (\theta_{01}, \theta_{02}, \theta_{03}, \theta_{11}, \theta_{12}, \theta_{13}, \sigma_\epsilon^2, \sigma_\eta^2)$ in the model. In matrix notation the DNS model can be presented as

$$\boldsymbol{\beta}_t = \theta_0 + \boldsymbol{Z}\boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t, \tag{9}$$
$$\boldsymbol{y}_t = \boldsymbol{\phi}\boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t, \tag{10}$$

where $\boldsymbol{y}_t = \begin{pmatrix} y_t(\tau_1) \\ y_t(\tau_2) \\ \vdots \\ y_t(\tau_m) \end{pmatrix}_{m\times1}$,

$$\boldsymbol{\phi} = \begin{pmatrix} 1 & f_1(\tau_1) & f_2(\tau_1) \\ 1 & f_1(\tau_2) & f_2(\tau_2) \\ \vdots & \vdots & \vdots \\ 1 & f_1(\tau_m) & f_2(\tau_m) \end{pmatrix}_{m\times3},$$

$$\boldsymbol{\beta}_t = \begin{pmatrix} \beta_{0t} \\ \beta_{1t} \\ \beta_{2t} \end{pmatrix}_{3\times1}, \boldsymbol{\epsilon}_t = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{pmatrix}_{m\times1},$$

such that:

- $f_1(\tau_j) = \left(\frac{1-e^{-\lambda\tau_j}}{\lambda\tau_j}\right)$
- $f_2(\tau_j) = \left(\frac{1-e^{-\lambda\tau_j}}{\lambda\tau_j} - e^{-\lambda\tau_j}\right)$, $j = 1, 2, ..., m$. The index j represents the term structure or maturity. There are 11 term structures for this study ($m = 11$)
- $\theta_0 = \begin{pmatrix} \theta_{01} \\ \theta_{02} \\ \theta_{03} \end{pmatrix}$
- $\boldsymbol{Z} = \begin{pmatrix} \theta_{11} & 0 & 0 \\ 0 & \theta_{12} & 0 \\ 0 & 0 & \theta_{13} \end{pmatrix}$

Note that $\boldsymbol{\epsilon}_t \sim \boldsymbol{N}_m(0, \sigma_\epsilon^2\boldsymbol{I}_m)$ and $\boldsymbol{\eta}_t \sim \boldsymbol{N}_3(0, \sigma_\eta^2\boldsymbol{I}_3)$. Note that (9) is *system equation* and (10) is *observation* *equation*. [2] suggest that the factors of the Nelson-Siegel model be estimated using a least squares procedure. The data for each yield curve produces a set factors associated with the Nelson-Siegel representation of the yield curve. The dataset is a collection of yield curves. Therefore sequential application of the least squares procedure would yield a set of Nelson-Siegel factors. The evolution of these factors can be represented using a Vector Auto-Regressive model. A model selection methodology like the Bayesian Information Criterion can be used to determine the optimal lag order for the model. Once an optimal model structure has been determined, forecasting is performed using the optimal model.

D. Forecast with Dynamic Gaussian Process Prior Model

Here we introduce dynamic Gaussian process prior model. The observation equation is

$$\boldsymbol{y}_t \quad = \quad \mu_t(\tau) + \boldsymbol{\epsilon}_t,$$

where $\boldsymbol{y}_t$ and $\boldsymbol{\epsilon}_t$ are defined as in (10), $\mu_t(\tau)$ is the mean function. The system equation is defined as

$$\mu_t(\tau) \quad = \quad \mu_{t-1}(\tau) + W_t, \tag{11}$$

where $W_t(\tau) \sim \boldsymbol{N}_m(\mathbf{0}, \mathbf{K}_{t-1})$, where $\mathbf{K}_{t-1} = K(\tau, \tau'|\rho_{t-1})$, ρ_{t-1} is the hyper-parameter estimated at $t-1$. The key notion here is that given the data $\boldsymbol{Y}_t = (\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_1)$ inference about μ_t and prediction about $\boldsymbol{y}_{t+1}$ can be carried via Bayes theorem, which can be expressed as

$$\mathbb{P}(\mu_t(\tau)|\boldsymbol{Y}_t) \propto \mathbb{P}(\boldsymbol{y}_t|\mu_t(\tau), \boldsymbol{Y}_{t-1}) \times \mathbb{P}(\mu_t(\tau)|\boldsymbol{Y}_{t-1}). \tag{12}$$

Note that the expression on the left of equation (12) is the *posterior process* of $\mu(\tau)$ at time t, whereas the first and second expression on the right side of (12) is the *likelihood* and *prior process* of $\mu(\tau)$, respectively. Suppose the posterior process at time point $t-1$ is the

$$\mu_{t-1}|\boldsymbol{Y}_{t-1} \sim \boldsymbol{N}_m\left(\hat{\mu}_{t-1}(\tau), \hat{\mathbf{K}}_{t-1}\right),$$

where $\hat{\mu}_{t-1}(\tau)$ is the posterior mean function and $\hat{\mathbf{K}}_{t-1}$ is the posterior covariance function of the process at the time-point $(t-1)$. Following the structure of the GP regression model as presented in (3) and (11), the prior predictive process at time point t is

$$\mu_t|\boldsymbol{Y}_{t-1} \quad \sim \quad \boldsymbol{N}_m\left(\hat{\mu}_{t-1}(\tau), \hat{\mathbf{K}}_{t-1}\right),$$

the likelihood function is

$$\boldsymbol{y}_t|\mu_t(\tau), \boldsymbol{Y}_{t-1} \sim \boldsymbol{N}_m(\mu_t(\tau), \sigma_t^2 \boldsymbol{I}_m),$$

and the marginal likelihood function is

$$\boldsymbol{y}_t|\boldsymbol{Y}_{t-1} \quad \sim \quad \boldsymbol{N}_m(\hat{\mu}_{t-1}(\tau), \hat{\mathbf{K}}_{t-1} + \sigma_{t-1}^2 \boldsymbol{I}_m). \tag{13}$$

Note that in (13) the $\mu_{t-1}(\tau)$ is a measurable under the σ-field generated by $\boldsymbol{Y}_{t-1}$. We can estimate the hyper-parameters $\theta_t = (\rho_{t-1}, \sigma_{t-1})$, using a optimization procedure to maximize the marginal-likelihood (13). Let's assume $\hat{\theta}_{t-1}$ is the estimated hyper-parameter estimated by optimizing the marginal likelihood (13). We can then provide an estimate for the observation at time t using the expected value of $\boldsymbol{y}_t|\boldsymbol{Y}_{t-1}$ (obtained from 13). This is:

$$\begin{aligned}\hat{\mu}_t(\tau*) &= \mathbb{E}(\mu_t(\tau*)|\boldsymbol{Y}_{t-1}) \\ &= K(\tau*, \tau|\hat{\rho}_{t-1}).[K(\tau, \tau|\hat{\rho}_{t-1}) + \hat{\sigma}_{t-1}^2.\boldsymbol{I}]^{-1}.\boldsymbol{y}_{t-1}(\tau).\end{aligned}$$

Once we have obtained the observation at time t, we can update the posterior process over $\boldsymbol{y}_t$ as:

$$\boldsymbol{y}_t(\tau)|\boldsymbol{Y}_t \sim \boldsymbol{N}_m(\hat{\mu}_{t.updated}(\tau), \hat{\mathbf{K}}_{t.updated}),$$

where the corresponding covariance function is

$$\hat{\mathbf{K}}_{t.updated} = K(\tau_*, \tau_*|\hat{\rho}_t) - K(\tau_*, \tau|\hat{\rho}_t).[K(\tau, \tau|\hat{\rho}_t) + \hat{\sigma}_t^2.\boldsymbol{I}]^{-1}.K(\tau, \tau_*|\hat{\rho}_t),$$

and the mean function or the expected value of μ_t at $\tau*$ is

$$\begin{aligned}\hat{\mu}_{t.updated}(\tau*) &= \mathbb{E}(\hat{\mu}_t(\tau*)|\boldsymbol{Y}_t) \\ &= \hat{\mu}_t(\tau*) + K(\tau*, \tau|\hat{\rho}_t).[K(\tau, \tau|\hat{\rho}_t) + \hat{\sigma}_t^2.\boldsymbol{I}]^{-1}.(\boldsymbol{y}_{t+1} - \hat{\mu}_t(\tau)).\end{aligned}$$

The details of an algorithmic implementation of this procedure is provided section III-D1

1) The Dynamic Gaussian Process Algorithm: There are two distinct phases of the algorithm. These correspond to the time steps $t = 0$ (the first yield curve in the dataset) and $t > 0$ (the subsequent yield curves in the dataset). The details of each of these phases is provided below.

Time step $t = 0$:

1) **Hyper-parameter Estimation:** Estimate hyper-parameters, $\hat{\theta}_0$, of the Gaussian Process $\boldsymbol{y}_0 \sim \boldsymbol{N}_m(\mathbf{0}, \mathbf{K} + \boldsymbol{\sigma_0^2 \mathbf{I_m}})$. Here $\mathbf{K} = K(\tau, \tau*|\rho_0)$ and $\theta_0 = (\rho_0, \sigma_0)$ are the hyper-parameters at time $t = 0$. The hyper-parameters are obtained by maximizing the marginal log-likelihood using an optimization algorithm (gradient descent, conjugate gradient descent etc.)
2) **Predict:** Provide an estimate of the yield for time step $t = 1$ using:

$$\begin{aligned}\hat{\mu}_0(\tau*) &= \mathbb{E}(\boldsymbol{y}_1(\tau*)|\boldsymbol{Y}_0) \\ &= K(\tau*, \tau|\hat{\rho}_0).[K(\tau, \tau|\hat{\rho}_0) + \hat{\sigma}_0^2.\boldsymbol{I}]^{-1}.\boldsymbol{y}_0(\tau).\end{aligned}$$

The predictive interval for time point 1 can be provided using following distribution

$$\mu_1|\boldsymbol{Y}_0 \sim \mathbf{N_m}(\hat{\mu}_\mathbf{0}, \hat{\mathbf{K}}_\mathbf{0}).$$

3) **Update**:
   a) Update the posterior covariance function as:

$$\hat{\mathbf{K}}_{updated} = K(\tau_*, \tau_*|\hat{\rho}_0) - K(\tau_*, \tau|\hat{\rho}_0).[K(\tau, \tau|\hat{\rho}_0) + \hat{\sigma}_0^2.\boldsymbol{I}]^{-1}.K(\tau, \tau_*|\hat{\rho}_0).$$

   b) Update the posterior mean function as:

$$\begin{aligned}\hat{\mu}_{updated}(\tau*) &= \mathbb{E}(\mu_0(\tau*)|\boldsymbol{Y}_0) \\ &= K(\tau*, \tau|\hat{\rho}_0).[K(\tau, \tau|\hat{\rho}_0) + \hat{\sigma}_0^2.\boldsymbol{I}]^{-1}(y_1 - \hat{\mu}_0)\end{aligned}$$

   which is the mean function associated with the **Hyper-parameter Estimation** step for time step $t = 1$.

Time step $t \geq 1$:

1) **Hyper-parameter Estimation:** Estimate hyper-parameters $\hat{\theta}_t$ of the Gaussian Process $\boldsymbol{y}_t|\boldsymbol{Y}_{t-1} \sim \boldsymbol{N}_m(\hat{\mu}_{updated}(\tau), \mathbf{K} + \sigma_t^2 \boldsymbol{I}_m)$. Here $\mathbf{K} = K(\tau, \tau * |\rho_t)$ and $\theta_t = (\rho_t, \sigma_t)$ are the hyper-parameters at time step t. The hyper-parameters are obtained by maximizing the marginal log-likelihood using an optimization algorithm.
2) **Predict:** Provide an estimate of the yield for time step $t + 1$ using:

$$\begin{aligned}\hat{\mu}_t(\tau*) &= \mathbb{E}(\boldsymbol{y}_{t+1}(\tau*)|\boldsymbol{Y}_t) \\ &= K(\tau*, \tau|\hat{\rho}_t).[K(\tau, \tau|\hat{\rho}_t) + \hat{\sigma}_t^2.\boldsymbol{I}]^{-1}.\boldsymbol{y}_t(\tau).\end{aligned}$$

   The predictive interval for time point $t + 1$ can be provided using following process

$$\boldsymbol{y}_{t+1}|\boldsymbol{Y}_t \sim \boldsymbol{N}_m(\hat{\mu}_t, \hat{\mathbf{K}}_t).$$

3) **Update:**
   a) Update the posterior covariance function as:

$$\hat{\mathbf{K}}_{updated} = K(\tau_*, \tau_*|\hat{\rho}_t) - K(\tau_*, \tau|\hat{\rho}_t).[K(\tau, \tau|\hat{\rho}_t) + \hat{\sigma}_t^2.\boldsymbol{I}]^{-1}.K(\tau, \tau_*|\hat{\rho}_t)$$

   b) Update the posterior mean function for term $\tau*$ as:

$$\begin{aligned}\hat{\mu}_{updated}(\tau*) &= \mathbb{E}(\mu_t(\tau*)|\boldsymbol{Y}_t) \\ &= \hat{\mu}_t(\tau*) + K(\tau*, \tau|\hat{\rho}_t).[K(\tau, \tau|\hat{\rho}_t) + \hat{\sigma}_t^2.\boldsymbol{I}]^{-1}.(\boldsymbol{y}_t - \hat{\mu}_t(\tau)),\end{aligned}$$

   which is mean function for the **Hyper-parameter Estimation** step of the subsequent iteration.

The covariance function to use with the algorithm is a modeling decision and is problem specific. See [9] and [5] for guidelines. For the data used in this study a combination of a linear kernel and a squared exponential (Radial Basis Function) kernel produced good results.

Remark 1: Note that in this dynamic process, the posterior of last time $(t - 1)$ is being considered as a prior-predictive process for next time point t.

Remark 2: In this algorithm, we are estimating hyper-parameter at every stage. This is feasible because m is small in our case. However, this may not be the possible, in many practical problems.

E. Relationship between Dynamic Gaussian Process and Optimal Bayesian Filter

[10] proposes a Bayesian framework for dynamic models where the prior (π) at time step t, has a power law form:

$$\pi(\boldsymbol{y}_t) \propto p(\boldsymbol{y}_t|\boldsymbol{Y}_{t-1}, \delta_t), \tag{14}$$

where:

- $\boldsymbol{y}_t$ represents our prior at time step t
- δ_t represents the power at time step t

The main idea behind the power filter approach is to propagate information from one time step to the next. The posterior density at stage t is given by:

$$\pi(\boldsymbol{y}_t) \propto f(\boldsymbol{y}_t|\boldsymbol{Y}_{t-1}).\left[\pi_{t|t-1}(\boldsymbol{y}_t)\right]^{\delta_t} \quad such\ that\ 0 \leq \delta_t \leq 1, \tag{15}$$

where:

- $f(\boldsymbol{y}_t|\boldsymbol{Y}_{t-1})$ represents the likelihood
- $\left[\pi_{t|t-1}(\boldsymbol{y}_t)\right]^{\delta_t}$ represents the prior

When $\delta_t = 1$ and the likelihood and the prior are Multivariate Gaussian, then we obtain the Dynamic GP. [11] develop the power filter for dynamic generalized linear models. They show that the power filter model yields an efficient information processing rule in a dynamic model setting.

IV. Validation Methodology

The data for this study came from the website of the US Department of Treasury ([12]). The data represents over 10 years of yield curve data (February 2006 through February 2017). A rolling window was used to train and test the performance of the methods on the proposed dataset. The details of this procedure are as follows. Starting with the yield curve data for the first day, we select a batch of data to be used for training the method used for yield curve forecasting. We then use the developed model to score the first data point after the batch of data points used as training data. For training the next batch, we remove the first data point and include the first test point in the training set. As we repeat this process, we move through the dataset, forecasting one test point at a time. Forecasting using the multivariate time series method for either the Nelson-Siegel parameters or the term yield forecasts themselves are not Bayesian methods. We used 250 days of data for training for these methods. This corresponds to about a year of data. This implies one year of data is used to train the time series methods to forecast a test point. Gaussian Process regression is a Bayesian method. Section III-D provides the details of training and forecasting using the Dynamic Gaussian Process Model .

V. Results and Discussion

This study examines data over a ten year period. The Root Mean Square Error was used as the metric to assess the performance of the method. The Root Mean Square Error is defined by:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{i=N} \sum_{\tau=1}^{\tau=11} (\hat{\boldsymbol{y}}[\tau, i] - y[\tau, i])^2}{N}}, \tag{16}$$

where:

- $\hat{\boldsymbol{y}}[\tau, i]$ is the estimated yield for day i associated with term τ
- $\boldsymbol{y}[\tau, i]$ is the actual yield for day i associated with term τ

- N is the number of yield curves that are estimated using the procedure
- There are 11 terms in each yield curve.

A summarized view of the results for this ten year period are shown in Table I. Here the GP, MVTS and the TSNS columns represent the RMSE for various terms associated with the Dynamic Gaussian Process method, the Multivariate time series method and the Nelson Siegel based time series method respectively. An inspection of Table I shows that the Nelson Siegel model based time series does relatively poorly in comparison to the multivariate time series method and the dynamic GP. We examine the performance of these methods over three time durations - short term structures, medium term structures and long term structures. The short term structure included term structures upto 1 year. The medium term structures consists of bonds with maturities of 2 years, 3 years and 5 years. The long term structure category consists of bonds that mature at 7 years, 10 years, 20 years and 30 years. The multivariate time series appears to do well in the short term region of the yield curve while the dynamic GP method does well in the medium and long term regions of the yield curve. The above discussion is an abridged version for brevity. A detailed discussion of the results that provide performance curves for the dynamic GP as well as the multivariate time series methods in the short, medium and long term structures is available in the *arxiv* version of this work [13]

Term	GP	MVTS	TSNS
1 Month	0.104	0.088	0.121
3 Months	0.071	0.066	0.080
6 Months	0.054	0.047	0.088
1 Year	0.047	0.043	0.085
2 Years	0.052	0.055	0.088
3 Years	0.058	0.061	0.114
5 years	0.065	0.068	0.126
7 Years	0.065	0.070	0.149
10 Years	0.063	0.067	0.197
20 Years	0.061	0.065	0.977
30 Years	0.060	0.063	10.838

TABLE I
RMSE FOR TERM STRUCTURES FOR ALL METHODS

The `GPy python` package [14] was used for developing the Gaussian Process models reported in this work. The `vars R` package[15] was used to model the time series based methods to forecast the Nelson-Siegel coefficients or the yield curve term rates.

VI. Conclusion

Gaussian processes have been used for functional data analysis in several domains (see [5]). The results of this study suggest that they can be used for yield curve forecasting. The nature of yield curve data is such that there is more data in the short and medium term structure regions than the long term structure regions . This makes long term forecasts challenging. This study revealed that the proposed dynamic GP method can forecast this region of the yield curve well. Analysts could use a mix of methods to forecast the yield curve. The data for this study spans a large time interval - over ten years. The results of this study indicate that the multivariate time series approach is more accurate for forecasting the short term structures, while the proposed dynamic Gaussian Process based method is a better choice for the medium and long term structures associated with the yield curve. The proposed method has been applied to a forecasting problem in the financial domain; however, this method can be applied to other domains as well. Demand forecasting is a common business requirement. In an IT data center, we might interested in forecasting the hourly number of user requests serviced by a group of computers. The hourly energy demand might be of interest to an electrical utility company. In summary, we believe that the dynamic Gaussian Process model could be useful in other application domains too.

References

[1] B. Nielsen, *Bond Yield Curve Holds Predictive Powers Treasury Rates*, 2017. [Online]. Available: http://www.investopedia.com/articles/economics/08/yield-curve.asp

[2] F. Diebold and C. Li, "Forecasting the term structure of government bond yields," *Journal of Econometrics*, vol. 130, no. 1, pp. 337–364, 2006.

[3] Y. Chen and L. Niu, "Adaptive dynamic nelson-siegel term structure model with applications," *Journal of Econometrics*, vol. 180, no. 1, pp. 98–115, 2014.

[4] H. S. Spencer Hays and J. Z. Huang, "Functional dynamic factor models with applications to yield curve forecasting," *Annals of Applied Statistics*, vol. 6, no. 3, pp. 870–894, 2012.

[5] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[6] P. Das, A. Lahiri, and S. Das, "Understanding sea ice melting via functional data analysis," *arXiv preprint arXiv:1610.07024*, 2016.

[7] C. R. Nelson and A. F. Siegel, "Parsimonious modeling of yield curve," *The Journal of Business*, vol. 60, no. 4, pp. 473–489, 1987.

[8] F. X. Diebold and G. D. Rudebusch, *Yield Curve Modeling and Forecasting: The Dynamic Nelson-Siegel Approach*. Princeton University Press, 2013.

[9] D. Duvenaud, *Kernel Cookbook Kernel Cookbook*, 2017. [Online]. Available: http://www.cs.toronto.edu/~duvenaud/cookbook/index.html

[10] J. Smith, "The multiparameter steady model," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 256–260, 1981.

[11] S. Das and D. K. Dey, "On dynamic generalized linear models with applications," *Methodology and Computing in Applied Probability*, pp. 1–15, 2013.

[12] www.treasury.gov, *Treasury Rates Treasury Rates*, 2017. [Online]. Available: https://www.treasury.gov/resource-center/data-chart-center/interest-rates/Pages/TextView.aspx?data=yield

[13] R. Sambasivan and S. Das, "A statistical machine learning approach to yield curve forecasting," *arXiv preprint arXiv:1703.01536*, 2017.

[14] GPy, "GPy: A gaussian process framework in python," http://github.com/SheffieldML/GPy, 2012–2014.

[15] B. Pfaff, "Var, svar and svec models: Implementation within R package vars," *Journal of Statistical Software*, vol. 27, no. 4, 2008. [Online]. Available: http://www.jstatsoft.org/v27/i04/