INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD • INDIA

# New Approaches to Prediction using Functional Data Analysis

**Arnab Kumar Laha**

**Poonam Rathi**

**W.P. No.2017-08-02**

August 2017

The main objective of the working paper series of the IIMA is to help faculty members, research staff and doctoral students to speedily share their research findings with professional colleagues and test their research findings at the pre-publication stage. IIMA is committed to maintain academic freedom. The opinion(s), view(s) and conclusion(s) expressed in the working paper are those of the authors and not that of IIMA.

विद्याविनियोगाद्विकासः
IIM
AHMEDABAD

**INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD-380 015
INDIA**

# New Approaches to Prediction using Functional Data Analysis

Arnab Kumar Laha
Poonam Rathi

Production and Quantitative Methods Area
Indian Institute of Management Ahmedabad
arnab@iima.ac.in
poonamr@iima.ac.in

## Abstract

In this paper we address the problem of prediction with functional data. We discuss several new methods for predicting the future values of a partially observed curve when it can be assumed that the data is coming from an underlying Gaussian Process. When the underlying process can be assumed to be stationary with powered exponential covariance function we suggest two new predictors and compare their performance. In some real life situations the data may come from a mixture of two stationary Gaussian Processes. We introduce three new methods of prediction in this case and compare their performance. In case the data comes from a non-stationary process we propose a modification of the powered exponential covariance function and study the performance of the three predictors mentioned above using three real-life data sets. The results indicate that the KM-Predictor in which the training data is clustered using the K-Means algorithm before prediction can be used in several real life situations.

**Keywords:** Gaussian process, Powered exponential covariance function, k-Nearest Neighbours, k-Means clustering, Forecasting

# 1 Introduction

One of the most widely used application of statistics is for prediction. Apart from the natural human curiosity of knowing the future, prediction is of great value in many scientific, medical and business situations. For example, in predicting the disease course of HIV infected individual, CD4 cell counts and CD4 percentages are used as important markers. It is therefore important for treating physicians to have predictions of the values of these markers in assessing the progress of such patients,Yao et al. (2005). Erbas et al. (2012) uses functional time series analysis to model mortality due to Chronic Obstructive Pulmonary Disorder (COPD) as a function of age and forecasts COPD mortality in Australia for a twenty year period. For banks having extensive Automated Teller Machine (ATM) networks it is of interest to predict cash requirements at the ATMs so that the machines could be optimally filled with cash to prevent customer dissatisfaction as well as reduce the opportunity cost of keeping cash idle. One such example in the context of Lithuanian cards payment market is given in Laukaitis (2008). Predicting accurate traffic flow is of great importance in efficient traffic management in highways. Chiou (2012) has proposed a dynamical functional prediction method that can be applied to predict traffic flow patterns during the day using the partially observed traffic flow trajectory upto a given time. Antoniadis et al. (2016) uses a non-parametric function-valued forecast model for short-term electricity demand forecasting. Hyndman and Ullah (2007) uses a functional data approach to forecast age-specific mortality and age-specific fertility rates.

A random element $X_t$ is a generalisation of the concept of a random variable. It is a measurable map from a probability space (Ω, F, P) with values in a measurable space (E, $\xi$). In this paper we would assume that E is the space of all square integrable real valued functions on a closed bounded interval. A stochastic process indexed by a set T is a collection $X = \{X_t\}_{t \in T}$. The set $X(\omega) = \{X_t(\omega) : t \in T\}$ is called a sample path of the process. In functional data analysis (FDA) the data is represented in the form of curves unlike that in the conventional univariate or multivariate data where the observations are either scalars or vectors. There has been a large number of applications of FDA over the past few years because of better data gathering technologies such as sensors, and increase in processing power of computers. The book by Ramsay and Silverman (2002) gives some intersting case studies on FDA while the books by Hsing and Eubank (2015), Horváth and Kokoszka (2012), Zhang (2013) and Ramsay and Silverman (2005) discusses theory and methods of FDA and gives interesting insights.

We consider functional observations $X_i(t)$, $t \in T,\ i = 1 \cdots, n$ defined over $[a, b]$. It is assumed that $X_i$ are independent and identically distributed as X, drawn from $L^2([a, b])$. Suppose that the mean function be $\mu = E(X) \in L^2([a, b])$. Further, assume that the covariance function K is a continuous function on $[a, b] \times [a, b]$. Then, there exists a sequence of continuous eigenfunctions $\phi_n$ and a decreasing sequence of corresponding non-negative eigenvalues $\lambda_n$ such that

$$\int_a^b K(s,t)\phi_n(s)ds = \lambda_n \phi_n(t), \qquad \int_a^b \phi_n(s)\phi_m(s)ds = \delta_{nm}$$

Also, each functional observation can be decomposed as $X_t = \sum_{n=0}^{\infty} \eta_n \phi_n(t)$ where $(\eta_n)$ is a sequence of real valued random variables with zero mean such that $E(\eta_n \eta_m) = \lambda_n \delta_{nm}$. Moreover, $K(s,t) = \sum_{n=0}^{\infty} \lambda_n \phi_n(s)\phi_n(t)$; $s, t \in [a, b]$; where the series converges uniformly and absolutely on $(a, b)$ [Bosq (2000),(p.25)].

In this paper we assume that the functional observations are realization of a sample path of the underlying Gaussian process. A Gaussian process $\{X_t, t \in T\}$, indexed by a set T (in this paper we take T to be the set of non-negative real numbers), is a stochastic process, in which any finite linear combination of random variables $X_t$, (all defined on the same probability space), have a joint multivariate normal distribution. Equivalently, $\{X_t, t \in T\}$ is a Gaussian process, if for any choice of distinct values $t_1, \cdots, t_k \in T$, the random vector $X = (X_{t_1}, \cdots, X_{t_k})^T$ has a multivariate normal distribution with mean vector $\mu = E(X) = (E(X_{t_1}), \cdots, E(X_{t_k}))^T$ and covariance matrix $\Sigma = (Cov(X_{t_i}, X_{t_j}))_{i,j=1,\cdots,k} = (\sigma_{ij})_{i,j=1,\cdots,k}$. The mean and covariance functions of a Gaussian process are given by

$$\mu(t) = E(X_t) \quad \text{and} \quad \Sigma(s,t) = Cov(X_s, X_t) = E(X_s - E(X_s))(X_t - E(X_t))$$

respectively. A Gaussian process is completely specified by its mean function and covariance function. It is said to be stationary if and only if it's mean function is a constant and $\Sigma(s,t)$ depends only on $s - t$. The class of Gaussian processes is one of the most widely used families of stochastic processes for modeling dependent data observed over time (see for e.g. Müller and Yang (2010), Shi and Choi (2011)). Among the many desirable properties associated with the Gaussian process is the Karhunen-Loeve (KL) expansion. The KL-expansion of a centered Gaussian process $\{X_t, t \in T\}$ can be represented as [Wahba (1990)(p.5)]

$$X_t = \sum_{k=1}^{\infty} \xi_k \phi_k(t)$$

where $\xi_1, \xi_2 \cdots$ are independent, Gaussian random variables with

$$E\xi_k = 0, \quad E\xi_k^2 = \lambda_k$$

and

$$\xi_k = \int_T X_s \phi_k(s) ds,\ \Sigma(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\phi_k(t) \text{ and } \int_T\int_T \Sigma^2(s,t)\, ds\, dt < \infty$$

In this paper we propose prediction methods for the data coming from an underlying Gaussian Process. We propose two new predictors namely CE-Predictor and k-NN Predictor for data coming from a stationary Gaussian Process with powered exponential covariance function. Also, for mixture of two stationary Gaussian processes, we additionally propose two new predictors: KM-Predictor and FC-Predictor. At first, these methods cluster the training data into two classes, then the partially observed curve is classified in one of the classes and subsequently a prediction is made. When the underlying process appears to be non-stationary we modify the powered exponential covariance function and apply our predicton methods to three real life data-sets. It appears that the KM-Predictor performs quite well though not always the best in a variety of situations.

The organization of the paper is as follows: In section 2, we briefly review the prediction problem for functional data coming from an underlying stationary Gaussian Process or a mixture of Gaussian Processes and outline the proposed methods of prediction, in section 3, we discuss the parameter estimation techniques that are necessary for implementing the proposed methods in practice, in section 4 we provide comparisons of the performance of the proposed methods using simulations, in section 5 we discuss the prediction problem for non-stationary Gaussian processes using three real-life data sets as examples, and finally in section 6 we make some concluding remarks.

# 2 Prediction Problem for Gaussian Process

## 2.1 Stationary Gaussian Process

In this section, we discuss the prediction problem for GPs whose covariance functions $K$(s,t) has the powered exponential form given by

$$K(s,t) = v\exp\left(-w|s-t|^{\gamma}\right), \quad v,w>0, \quad 0<\gamma<2. \tag{1}$$

We refer the Gaussian Process with above covariance function as $\mathrm{GP_S}(\mu, v, w, \gamma)$. The $\mathrm{GP_S}(\mu, v, w, \gamma)$ has the following properties which are used later in the paper:

P.1 $\mathrm{V}(X_s) = \mathrm{K}(s,s) = v$ for all s $> 0$, where $\mathrm{V}(X_s)$ denotes the variance of $X_s$

P.2 Correlation of $(X_s, X_{s+1}) = \rho(X_s, X_{s+1}) = \exp(-w)$ for all $s > 0$

P.3 $\rho(X_s, X_{s+2}) = \exp(-w2^{\gamma})$ for all $s > 0$

Suppose $X^{(1)}, \cdots, X^{(m)}$ is a random sample of size m from the Gaussian process $\{X_t\}$ with each $X^{(i)}$, $1 \le i \le m$ being observed at the time points $\{1, \cdots, n\}$ and let $X^{(i)} = (X_1^{(i)}, \cdots, X_n^{(i)})$. Note that $(X_1^{(i)}, \cdots, X_n^{(i)})$ jointly follows a Multivariate Normal Distribution (MND) with mean $\mu_{n\times 1}$ and covariance matrix $\Sigma_{n\times n} = (\sigma_{ij})_{1\le i,j\le n}$ where $\sigma_{ij} = v\exp(-w|i-j|^{\gamma})$. Now suppose $\mathrm{X}^{\mathrm{new}}$ is a new observation from the above Gaussian process which has been observed only up to the point $n^* < n$. We want to predict the values of $\mathrm{X}^{\mathrm{new}}(\mathrm{j})$, $n^*+1 \le j \le n$ based on the observed values $X^{\mathrm{new}}(j), 1 \le j \le n^*$. Since the distribution of $X^{\mathrm{new}}$ is MND, we know that the $X_2^{\mathrm{new}}$ given $X_1^{\mathrm{new}}$ is again MND with expectation $\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(X_1 - \mu_1)$ and covariance $\Sigma_{22} - \Sigma_{21}^{-1}\Sigma_{11}\Sigma_{12}$, i.e.

$$X^{\mathrm{new}} = \begin{bmatrix} X^{\mathrm{new}}_{1(n^*\times 1)} \\ X^{\mathrm{new}}_{2((n-n^*)\times 1)} \end{bmatrix} \sim \mathrm{MND}\left( \begin{bmatrix} \mu_{1(n^*\times 1)} \\ \mu_{2(n-n^*)\times 1} \end{bmatrix}, \begin{bmatrix} \Sigma_{11(n^*\times n^*)} & \Sigma_{12(n^*\times(n-n^*))} \\ \Sigma_{21((n-n^*)\times n^*)} & \Sigma_{22((n-n^*)\times(n-n^*))} \end{bmatrix} \right)$$

We define $\hat{\mu}_2 + \hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1}(X_1 - \hat{\mu}_1)$ as the CE-Predictor of $X_2$ . The method used for obtaining the estimates $\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}_{21}$ and $\hat{\Sigma}_{11}$ are discussed later in this paper (see Section 3.1).

An alternative to using the CE-Predictor is to first use the k-nearest neighbor (k-NN) algorithm to select a subset S ($\mathrm{X}^{\mathrm{new}}$) of k observations from the sample of m observations whose values in the first $n^*$ components are closest to $X_1^{\mathrm{new}}$ using some distance measure. In this paper we have used the Euclidean distance in $\mathbb{R}^{n^*}$ as the distance measure. It is also possible to have other distances measures such as Minkowski $L^m$ distance metric, $m \ge 1$ (Arya et al. (1998)). Let $\tilde{\mu}_1, \tilde{\mu}_2, \tilde{\Sigma}_{11}$ and $\tilde{\Sigma}_{21}$ be the estimates of $\mu_1, \mu_2, \Sigma_{11}$ and $\Sigma_{21}$ based only on the elements of the subset $S(X^{\mathrm{new}})$. Then we define k-NN predictor of $X_2^{\mathrm{new}}$ is $\tilde{\mu}_2 + \tilde{\Sigma}_{21}\tilde{\Sigma}_{11}^{-1}(X_1 - \tilde{\mu}_1)$. The choice of k is important for practical applications as it determines the performance of the estimates $\tilde{\mu}_1, \tilde{\mu}_2, \tilde{\Sigma}_{21}$ and $\tilde{\Sigma}_{11}$. A very small value of k is likely to give poor estimates while too large a value of k would make the estimates similar to $\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}_{21}$ and $\hat{\Sigma}_{11}$, thus losing its adaptibility. In this paper we will vary k and study the performance of prediction in section (4.1). In this context it may be noted that the determination of the value of k in the k-NN algorithm has been studied in some detail in the literature particularly in connection with the classification problems (see for e.g. Ghosh (2006), Hall et al. (2008)).

## 2.2 Mixture of Stationary Gaussian Processes

In this section, we assume that the $X^{(1)}, \ldots, X^{(m)}$ is a random sample of size m from the mixture of Gaussian process $\{X_{kt}\}$, $k = 1, 2, \ldots, p$ with each $X^{(i)}$, $1 \leq i \leq m$ being observed at the time points $\{1, \ldots, n\}$. We assume that the covariance function for Gaussian processes $\{X_{kt}\}$ are same and equal to $K$ and they differ only in their mean functions. We are aware that assumption of equal covariance may be relaxed with additional computational complexities. Suppose as earlier we have a new observation $\mathrm{X}^{\mathrm{new}}$ which has been observed only up to the point $n^* \leq n$. In this paper, we restrict ourselves to two Gaussian processes $\{X_{1t}\}$ and $\{X_{2t}\}$ but it is straight forward to extend it further. Let the mixture of the two Gaussian processes be $\pi X_{1t} + (1-\pi)X_{2t}$ where $0 \leq \pi \leq 1$. As mentioned earlier, we assume that the covariance function for both $\{X_{1t}\}$ and $\{X_{2t}\}$ are same and equal to $K$ and they differ only in their mean functions i.e. $X_{1t} \sim \mathrm{GP}(\mu_1,\mathrm{K})$ and $X_{2t} \sim \mathrm{GP}(\mu_2,\mathrm{K})$.

We use the K-Means algorithm (Hartigan and Wong (1979)) to divide the data $\{X_1^{(1)}, \cdots, X_1^{(m)}\}$ into two clusters where $X_1^{(i)}$ denotes the first $n^*$ components of $X^{(i)}$. Then treating these two clusters as training samples coming from the two populations we obtain the estimates $\breve{\mu}_{1(1)}, \breve{\mu}_{2(1)}, \breve{\mu}_{1(2)}, \breve{\mu}_{2(2)}, \breve{\Sigma}_{21}$ and $\breve{\Sigma}_{11}$, where $\mu_{1(i)}$ represents first $n^*$ components of mean of cluster i and $\mu_{2(i)}$ represents $n^*+1$ to n components of mean of cluster i. The details of the parameter estimation method used are given in the section 3.3. A kernel based classification technique (Ferraty and Vieu (2006), pp. 113-116) is then used with the data $\{X_1^{(1)}, \cdots, X_1^{(m)}\}$ to classify $(X_1^{\text{new}})$ into one of the two populations. Then $(X_2^{\text{new}})$ is predicted as $\breve{\mu}_{2(i)} + \breve{\Sigma}_{21}\breve{\Sigma}_{11}^{-1}\left(X_1 - \breve{\mu}_{1(i)}\right)$, depending on the cluster in which observation $(X_1^{\text{new}})$ lies. We will refer to this as the KM-predictor of $(X_2^{\text{new}})$.

An alternative approach is to use the technique Funclust introduced in Jacques and Preda (2013) which is a model based clustering technique with functional data. Let the KL-expansion of the centered Gaussian process $X_{1t} - \mu_1$ be $X_{1t} - \mu_1 =$

$\sum_{k=1}^{\infty} \xi_{1k}\phi_k(t)$ and that $X_{2t} - \mu_2 = \sum_{k=1}^{\infty} \xi_{2k}\phi_k(t)$ where $\{\xi_{ik}\}_{k=1}^{\infty}$, $i = 1, 2$ are independent, Gaussian random variables with $E\xi_{ik} = 0$, $E\xi_{ik}^2 = \lambda_{ik}$ and

$$\xi_{ik} = \int_T X_s \phi_k(s) ds, \ \Sigma_i(s,t) = \sum_{k=1}^{\infty} \lambda_{ik}\phi_k(s)\phi_k(t)$$

and

$$\int_T\int_T \Sigma_i^2(s,t)\, ds\, dt < \infty, \quad i = 1, 2.$$

We truncate the above KL-expansions of the processes $X_{1t}-\mu_1$ and $X_{2t}-\mu_2$ at $q_1$ and $q_2$. The approximate density for $X_{it}-\mu_i$ is then $f_{X q_i}(x) = \prod_{k=1}^{q_i} \dfrac{1}{\sqrt{2\pi\lambda_{ik}}} \exp\left(-\dfrac{\xi_{ik}^2(x)}{2\lambda_{ik}}\right)$. Thus, the approximate density of the mixture $\pi X_{1t} + (1-\pi)X_{2t}$ is

$$f_{X_{(q_1,q_2)}}(x) = \pi \prod_{k=1}^{q_1} \frac{1}{\sqrt{2\pi\lambda_{1k}}} \exp\left(-\frac{\xi_{1k}^2(x)}{2\lambda_{1k}}\right) + (1-\pi) \prod_{k=1}^{q_2} \frac{1}{\sqrt{2\pi\lambda_{2k}}} \exp\left(-\frac{\xi_{2k}^2(x)}{2\lambda_{2k}}\right).$$

where $x$ is a random sample path of the centered Gaussian Process having the above KL-expansion.

Jacques and Preda (2013) suggests maximizing the pseudo-likelihood defined by

$$l^{(q)}(\theta; X^{(1)}, \cdots, X^{(m)}) = \prod_{i=1}^{m} \sum_{g=1}^{2} \pi_g \prod_{j=1}^{q_g} \frac{1}{\sqrt{2\pi\lambda_{jg}}} \exp\left(-\frac{1}{2}\frac{\xi_{i,j,g}^2(X^{(i)})}{\lambda_{jg}}\right)$$

with $\pi_1 = \pi$, $\pi_2 = 1 - \pi_1$ , $q = (q_1, q_2)$, $\theta = (\pi, \lambda_{11}, \cdots, \lambda_{1q_1}, \cdots, \lambda_{21}, \lambda_{2q_2})$ are the parameters which needs to be estimated and $\xi_{i,j,g} = \xi_{j,g}(X^i)$ represents the $\text{j}^{\text{th}}$ principal component score of the curve $X^{(i)}$ belonging to the group g, g = 1, 2.

The parameters are estimated by using EM like algorithm details of which can be found in Jacques and Preda (2013). We classify an observation to be in group 1 if the posterior probability of the observation lying in group 1 is greater than 0.5 otherwise it is classified to be in group 2. Ties, if any, are broken randomly. Going forward we treat these two groups as two clusters. Now treating observations in the two clusters as training samples coming from the two populations we obtain the estimates $\acute{\mu}_{1(1)}, \acute{\mu}_{2(1)}, \acute{\mu}_{1(2)}, \acute{\mu}_{2(2)}, \acute{\Sigma}_{21}$ and $\acute{\Sigma}_{11}$, where $\acute{\mu}_{1(i)}$ represents first $n^*$ components of mean of cluster i and $\acute{\mu}_{2(i)}$ represents $n^* + 1$ to n components of mean of cluster

i. The details of the parameter estimation method used are given in the next section.

For the new observation $\mathrm{X}^{\mathrm{new}}$ which has been observed only up to point $n^*$, we compute

$$P(C_1|X^{\rm new}) = \frac{f(X^{\rm new}; \acute{\mu}_{1(1)}, \acute{\Sigma}_{11})P(\hat{C}_1)}{f(X^{\rm new}; \acute{\mu}_{1(1)}, \acute{\Sigma}_{11})P(\hat{C}_1) + f(X^{\rm new}; \acute{\mu}_{1(2))}, \acute{\Sigma}_{11})P(\hat{C}_2)}$$

where f is the multivariate normal density, $P(\hat{C}_1)$ and $P(\hat{C}_2)$ is calculated as the proportion of the number of elements in cluster 1 and 2 respectively.

Similarly we calculate $P(C_2|X^{\rm new})$. We classify $X^{\rm new}$ into that cluster for which $P(C_i|X^{\rm new}), i = 1, 2$ is larger. In case of ties, we allocate the observation to one of the two clusters randomly. Suppose that the observation is classified in group i then the $X_2^{\rm new}$ is predicted as $\acute{\mu}_2 + \acute{\Sigma}_{21}\acute{\Sigma}_{11}^{-1}\,(X_1 - \acute{\mu}_1)$. We will refer to this method as the FC-predictor of $(X_2^{\rm new})$.

# 3 Parameter Estimation

## 3.1 CE-Predictor

Let the training functional observations be $X^{(i)} = (X_1^{(i)}, \cdots, X_n^{(i)}), i = 1, \cdots, m$. The estimate of $\mu$ is the mean of the training data set which is denoted as $\hat{\mu} = [\bar{\mu}_1, \cdots, \bar{\mu}_{\rm n}]^T$ where $\bar{\mu}_j = \dfrac{1}{m}\sum_{i=1}^m X_j^{(i)}$. Let the covariance function estimated using training data be $\Sigma = [\sigma_{ij}]_{n\times n}$ where $\sigma_{ij} = Cov(X_i, X_j)$. Since by P.1, $v = K(s, s)$ for all $s > 0$ we propose to estimate $v$ as

$$\hat{v} = \frac{1}{n}\sum_{i=1}^{n}\sigma_{ii} \tag{2}$$

Again since by P.2, $w = -\ln\rho(X_s, X_{s+1})$ we propose to estimate $w$ as

$$\hat{w} = -\ln\left(\frac{1}{n-1}\sum_{i=1}^{n-1}\frac{\hat{\sigma}_{i,i+1}}{\sqrt{\hat{\sigma}_{i,i}\hat{\sigma}_{i+1,i+1}}}\right) = -\ln\left(\frac{1}{n-1}\sum_{i=1}^{n-1} r_{i,i+1}\right) \tag{3}$$

where $r_{i,i+1} = \frac{\hat{\sigma}_{i,i+1}}{\sqrt{\hat{\sigma}_{i,i}\hat{\sigma}_{i+1,i+1}}}$ is the estimated correlation between $X_i$ and $X_{i+1}$. Similarly, by using P.3 the parameter $\gamma$ can be estimated as

$$\hat{\gamma} = \frac{1}{\ln 2} \ln \left( -\frac{\ln \frac{1}{n-2} \sum_{i=1}^{n-2} r_{i,i+2}}{\hat{w}} \right) \tag{4}$$

where $r_{i,i+2}$ is the estimated correlation between $X_i$ and $X_{i+2}$.
To evaluate the efficiency of these estimates, we have carried out an experiment in which a sample of 200 curves are of the type

$$X_t = 4t + \frac{1}{100}\mathrm{f}(t) + \epsilon(t),$$

where t $\in \{1, 2, \cdots, 10\}$ , $\epsilon$ is a Gaussian processes with zero mean and covariance function $\sigma(s,t) = \mathrm{vexp}\left(-\mathrm{w}\left|\frac{s}{4} - \frac{t}{4}\right|^{\gamma}\right)$, and f(t) is the probability density function of a normal random variable with mean zero and standard deviation 0.001 were generated. We have taken various combinations of v, w and $\gamma$ in the above model and then estimated v, w and $\gamma$ using the method discussed in section 3.1. The mean square error (MSE) of the estimates based on 1000 simulations for each combination is reported in table 1.

| v | w | $\gamma$ | MSE($\hat{v}$) | MSE($\hat{w}$) | MSE($\hat{\gamma}$) |
|---|---|---|---|---|---|
| 2 | 0.5 | 1 | 0.02143 | 0.0027 | 0.0018 |
| 2 | 1 | 2 | 0.0166 | 0.0072 | 0.0004 |
| 1 | 1 | 0.5 | 0.0026 | 0.0124 | 0.0041 |
| 1.5 | 2 | 2 | 0.0070 | 0.0309 | 0.0007 |
| 2 | 1 | 1 | 0.0135 | 0.0103 | 0.0025 |
| 0.5 | 1 | 2 | 0.0009 | 0.0070 | 0.0004 |

Table 1: Mean Square Error of $\hat{v}$, $\hat{w}$ and $\hat{\gamma}$ for different values of v, w and $\gamma$ as obtained in section 3.1

Now using $(\hat{v}, \hat{w}, \hat{\gamma})$ we get $\hat{\Sigma} = \hat{\sigma}_{st} = \hat{v}\exp\left(-\hat{w}|s-t|^{\gamma}\right)$. Writing $\hat{\mu}_1 = [\bar{\mu}_1, \cdots, \bar{\mu}_{\mathrm{n}^*}]^T$ and $\hat{\mu}_2 = [\bar{\mu}_{\mathrm{n}^*+1}, \cdots, \bar{\mu}_{\mathrm{n}}]^T$ and decomposing
$\hat{\Sigma} = \begin{bmatrix} \hat{\Sigma}_{11(n^* \times n^*)} & \hat{\Sigma}_{12(n^* \times (n-n^*))} \\ \hat{\Sigma}_{21((n-n^*) \times n^*)} & \hat{\Sigma}_{22((n-n^*) \times (n-n^*)} \end{bmatrix}$, we use these estimates for computing the CE Predictor.

## 3.2 k-NN Predictor

Let the set S as defined in section 2.1 be $\left\{(Y_1^{(1)}, \cdots, Y_n^{(1)})^T, \cdots, (Y_1^{(k)}, \cdots, Y_n^{(k)})^T\right\}$. We calculate the mean of set S which we denote as $\tilde{\mu} = [\bar{\mu}_1, \cdots, \bar{\mu}_{\mathrm{n}}]$. We estimate v, w and $\gamma$ using the observations from set S as discussed in section 3.1 and form the estimated covariance function $\tilde{\Sigma}_{n \times n}$. We decompose the same as earlier to get $\tilde{\Sigma}_{11}$ and $\tilde{\Sigma}_{21}$. The other parameters required for predicting k-NN Predictor are $\tilde{\mu}_1 = [\bar{\mu}_1, \cdots, \bar{\mu}_{\mathrm{n}*}]^T$ and $\tilde{\mu}_2 = [\bar{\mu}_{\mathrm{n}*\,+1}, \cdots, \bar{\mu}_{\mathrm{n}}]^T$ .

## 3.3 KM-Predictor

Suppose the two clusters obtained by using the K-Means algorithm as discussed in the section 2.2 be $\{Y^{(1)}, \cdots, Y^{(k)}\}$ and $\{Z^{(1)}, \cdots, Z^{(l)}\}$ , $k + l = m$. Let $\breve{\mu}_{1(1)}$ be mean of $([Y_1^{(1)}, \cdots, Y_{n^*}^{(1)}]^T, \cdots, [Y_1^{(k)}, \cdots, Y_{n^*}^{(k)}]^T)$,and $\breve{\mu}_{2(1)}$ be mean of $([Y_{n^*+1}^{(1)}, \cdots, Y_n^{(1)}]^T, \cdots,$ $[Y_{n^*+1}^{(k)}, \cdots, Y_n^{(k)}]^T)$. Similarly we obtain the estimates of mean of second cluster $\breve{\mu}_{1(2)}$ and $\breve{\mu}_{2(2)}$ based on $\{Z^{(1)}, \cdots, Z^{(l)}\}$.

These two clusters of observations are subsequently treated as training samples from two populations. We assume that the two populations have the same covariance function and they differ only in their mean function. Therefore we base our estimates of v, w and $\gamma$ on the information from both the populations.

The pooled estimated of v is

$$\breve{v} = \frac{1}{2n}\left[\sum_{i=1}^{n} \sigma_{ii(1)} + \sum_{j=1}^{n} \sigma_{jj(2)}\right] \tag{5}$$

where $\Sigma_{(1)} = [\sigma_{ij(1)}]_{n \times n}$ and $\sigma_{ij(1)} = Cov(Y_i, Y_j)$ and $\Sigma_{(2)} = [\sigma_{ij(2)}]_{n \times n}$ where $\sigma_{ij(2)} = Cov(Z_i, Z_j)$.

The pooled estimated of w is

$$\breve{w} = \frac{-\ln\left(\frac{1}{n-1}\sum_{i=1}^{n-1} r_{i,i+1(1)}\right) - \ln\left(\frac{1}{n-1}\sum_{i=1}^{n-1} r_{i,i+1(2)}\right)}{2} \tag{6}$$

where $r_{i,i+1(1)} = \frac{\hat{\sigma}_{i,i+1(1)}}{\sqrt{\hat{\sigma}_{i,i(1)}\hat{\sigma}_{i+1,i+1(1)}}}$ is the estimated correlation between $Y_i$ and $Y_{i+1}$ and $r_{i,i+1(2)} = \frac{\hat{\sigma}_{i,i+1(2)}}{\sqrt{\hat{\sigma}_{i,i(2)}\hat{\sigma}_{i+1,i+1(2)}}}$ is the estimated correlation between $Z_i$ and $Z_{i+1}$.

Also, pooled $\gamma$ can be estimated using the formula

$$\breve{\gamma} = \frac{\frac{1}{\ln 2}\ln\left(-\frac{\ln \frac{1}{n-2}\sum_{i=1}^{n-2} r_{i,i+2(1)}}{\hat{w}}\right) + \frac{1}{\ln 2}\ln\left(-\frac{\ln \frac{1}{n-2}\sum_{i=1}^{n-2} r_{i,i+2(2)}}{\hat{w}}\right)}{2} \tag{7}$$

where $r_{i,i+2(1)}$ is the estimated correlation between $Y_i$ and $Y_{i+2}$, $r_{i,i+2(2)}$ is the estimated correlation between $Z_i$ and $Z_{i+2}$,

Using the estimates $\breve{v}$, $\breve{w}$ and $\breve{\gamma}$ we find covariance function $\breve{\Sigma}_{n\times n}$ and decompose it as described in (section 3.1) earlier to get $\breve{\Sigma}_{11}$ and $\breve{\Sigma}_{21}$. Then these are used in obtaining KM-Predictor.

## 3.4 FC-Predictor

Let the two clusters obtained by using the Funclust be $\{Y^{(1)}, \cdots, Y^{(k)}\}$ and $\{Z^{(1)}, \cdots, Z^{(l)}\}$, $k+l = m$. Let the mean of $([Y_1^{(1)}, \cdots, Y_{n^*}^{(1)}]^T \cdots, [Y_1^{(k)}, \cdots, Y_{n^*}^{(k)}]^T)$ and $([Y_{n^*+1}^{(1)}, \cdots, Y_n^{(1)}]^T, \cdots, [Y_{n^*+1}^{(k)}, \cdots, Y_n^{(k)}]^T)$ be $\acute{\mu}_{1(1)}$ and $\acute{\mu}_{2(1)}$ respectively. Similarly we obtain the estimates of mean of second cluster $\acute{\mu}_{1(2)}$ and $\acute{\mu}_{2(2)}$ based on $\{Z^{(1)}, \cdots, Z^{(l)}\}$. The covariance function $\acute{\Sigma}_{n\times n}$ is estimated by using $\breve{v}$,$\breve{w}$ and $\breve{\gamma}$ which are evaluated by using equation 5,6 and 7 respectively. $\acute{\Sigma}_{n\times n}$ is decomposed as described in (section 3.1)to get $\acute{\Sigma}_{11}$ and $\acute{\Sigma}_{21}$ which are used in obtaining FC-Predictor.

# 4 Simulations

This section is devoted to compare the different methods suggested in section 2. For the situation where underlying data comes from a stationary Gaussian process, we compare the CE-Predictor and k-NN Predictor along with the prediction obtained from the best ARIMA model using the auto.arima() command in the forecast package of R. In the mixture of two populations case we compare the k-NN Predictor, KM-Predictor, FC-Predictor with the prediction obtained from the best ARIMA model. We measure the predictive performance of a method using the Mean Square Prediction Error (MSPE) criterion. MSPE is the average square difference between the actual and predicted values at the time points $n^* > n$, where prediction is evaluated. It measures how close the forecasts are in comparison to the actual values. It is defined as

$$\text{MSPE}(t) = \frac{1}{q}\sum_{i=1}^{q}\left[X^i(t) - \hat{X}^i(t)\right]^2$$

where q represents the number of prediction instances, $X^i(t)$ represents the actual value at time t in the i$^{th}$ instance and $\hat{X}^i(t)$ represents the point forecast for the same. A method which gives lower MSPE on a test data set is considered to be better than another one which has higher MSPE.

## 4.1 Comparison of Methods for a Stationary Gaussian Process

A sample of 200 curves was used as a training set for development of the CE-Predictor and k-NN Predictor. Each of these curves are observed at 10 equidistant points $t = 1, 2, \cdots, 10$. In experiment 1, each of these curves are of the type

$$X_t = 4t + \frac{1}{100}\mathrm{f}(t) + \epsilon(t)$$

where t $\in \{1, 2, \cdots, 10\}$ , $\epsilon$ is a Gaussian processes with zero mean and covariance function $\sigma(s,t) = 2 * \exp\left(-0.5\left|\frac{s}{4} - \frac{t}{4}\right|^2\right)$, and f(t) is the probability density function of a normal random variable with mean zero and standard deviation 0.001. We take $n^* = 7$ and predict $(X^{\text{new}}(8), X^{\text{new}}(9), X^{\text{new}}(10))'$. The testing is done with a fresh set of 100 curves drawn from the same population. Since k-NN Predictor depends on the choice of k we compare the performance of this predictor by varying $k = 10, 15, 20$ and 25. The MSPE is calculated for all the methods with q = 100 for each of the forecast periods separately. The whole process is then repeated 1000 times and MSPE for each of the forecast periods separately is reported in table 2 below.

| Method | $\mathrm{X}^{\text{new}}(8)$ | $\mathrm{X}^{\text{new}}(9)$ | $\mathrm{X}^{\text{new}}(10)$ |
|---|---:|---:|---:|
| CE-Predictor | 0.00012 | 0.00306 | 0.02462 |
| 10-NN Predictor | 0.005 | 0.0833 | 0.37583 |
| 15-NN Predictor | 0.00321 | 0.06101 | 0.30984 |
| 20-NN Predictor | 0.00233 | 0.04824 | 0.2651 |
| 25-NN Predictor | 0.00178 | 0.03939 | 0.23012 |
| Best ARIMA | 0.047015 | 0.30618 | 0.98532 |

Table 2: Comparison of MSPE of different methods for three periods as discussed in experiment 1

We observe from table 2 that the CE-Predictor performs best across the three periods. We also note that the accuracy of the k-NN Predictor increases with increase in k across all the three time periods. In practical situations, using a trial data one may experiment with multiple values of k and form a plot of MSPE for different values of k and choose the one after which reduction i MSPE is not significant. In this case, all predictors perform better than the Best ARIMA method. This is to be expected since the Best ARIMA method does not benefit from the information contained in the training data set.

## 4.2 Comparison of Methods for a Mixture of two Gaussian Processes

### 4.2.1 Comparison of k-NN Predictor, KM-Predictor and Best ARIMA

In experiment 2, motivated by Alonso et al. (2012), a sample of 400 curves are used as a training set, with mixing proportions $\pi_1 = \pi_2 = 0.5$ for development of the k-NN predictor, KM-Predictor and FC-Predictor. Each of these curves are observed at 10 equidistant points $t = 1, 2, \cdots, 10$. We consider the following two functional data generating models or each of these curves are of the type

$$X_{1t} = t + \epsilon^{(1)}(t), \quad \text{or} \quad X_{2t} = t + 10 + \epsilon^{(2)}(t) \tag{8}$$

where t $\in \{1, 2, \cdots, 10\}$ , $\epsilon^{(k)}, k = 1, 2$ are Gaussian processes with zero mean and covariance function $\sigma(s, t) = 2 * \exp\left(-0.5\left|\frac{s}{4} - \frac{t}{4}\right|^2\right)$. We take $n^* = 7$ and predict $(X^{\text{new}}(8), X^{\text{new}}(9), X^{\text{new}}(10))'$. The testing is done with a fresh set of 200 curves, 100 curves each from both population. The figure 1 plots some simulated curves from both population.

Since k-NN Predictor depends on the choice of k we compare the performance of this predictor by varying $k = 10, 15, 20$ and 25. The MSPE is calculated for all the methods with q = 200 for each of the forecast periods separately. The whole process is then repeated 1000 times and the average value of the MSPE for each of the forecast periods separately is reported in table 3 below.

In experiment 3, a sample of 400 curves were used as a training set, the mixing proportions $\pi_i$'s are chosen to be equal to 0.5. Each of these curves are observed at 10 equidistant points $t = 1, 2, \cdots, 10$. We consider the following two functional data generating models for each of these curves

$$X_{1t} = 4t + \epsilon^{(1)}(t), \quad \text{and} \quad X_{2t} = 4|t - 5| + \epsilon^{(2)}(t) \tag{9}$$

where t $\in \{1, 2, \cdots, 10\}$ , $\epsilon^{(k)}, k = 1, 2$ are Gaussian processes with zero mean and covariance function $\sigma(s, t) = 2 * \exp\left(-0.5\left|\frac{s}{4} - \frac{t}{4}\right|^2\right)$. The figure 2 shows some

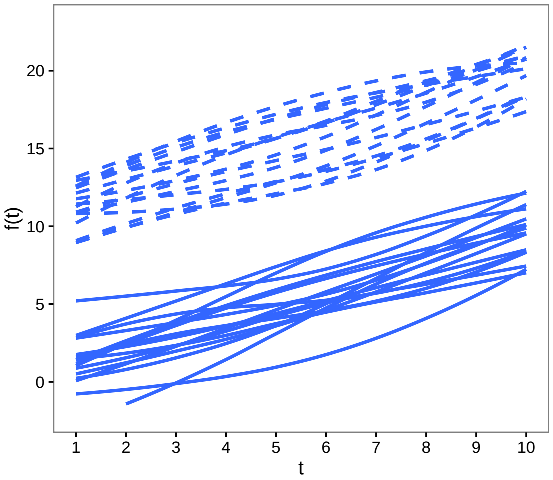

Figure 1: Plots of sample curves for $X_{1t}$ (smooth line) and $X_{2t}$ (dashed line) from experiment 2, equation 8 for t $\in \{1, \cdots, 10\}$

| Method | $X^{new}(8)$ | $X^{new}(9)$ | $X^{new}(10)$ |
|---|---|---|---|
| 5-NN Predictor | 0.01218 | 0.1452 | 0.51399 |
| 10-NN Predictor | 0.0005 | 0.08331 | 0.37741 |
| 15-NN Predictor | 0.00322 | 0.06137 | 0.31168 |
| 20-NN Predictor | 0.00232 | 0.04819 | 0.2652 |
| 25-NN Predictor | 0.00178 | 0.03944 | 0.2305 |
| KM-Predictor | 0.00013 | 0.003 | 0.026 |
| Best ARIMA | 0.052 | 0.321 | 1.016 |

Table 3: Comparison of MSPE of different methods for three periods as discussed in experiment 2

simulated curves from both populations. We take $n^* = 7$ and predict $X^{new} = (X^{new}(8), X^{new}(9), X^{new}(10))'$. The testing is done with a fresh set of 200 curves, 100 curves each from both population. The MSPE is calculated for all the methods with q = 200 for each of the forecast periods separately. The whole process is then repeated 1000 times and the average value of the MSPE for each of the forecast periods separately is reported in table 4 below.

From both experiment 2 and 3 we observe that the KM-Predictor performs the best across the three periods. All the predictors perform better than the Best ARIMA method, which may be expected since Best ARIMA method does not benefit from

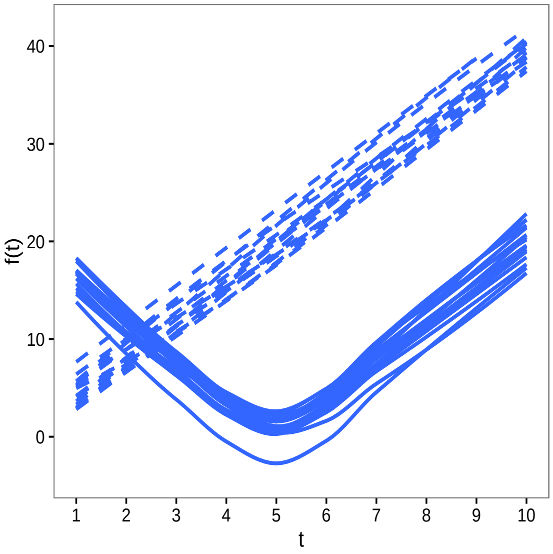

Figure 2: Plots of sample curves for $X_{1t}$ (dashed line) and $X_{2t}$ (smooth line) from experiment 3, equation 9

.

| Method | $X^{\text{new}}(8)$ | $X^{\text{new}}(9)$ | $X^{\text{new}}(10)$ |
|---|---|---|---|
| 10-NN Predictor | .00496 | 0.08260675 | 0.3746607 |
| 25-NN Predictor | 0.001770926 | 0.03914277 | 0.2292558 |
| KM-Predictor | 0.0001299241 | 0.0032155687 | 0.02557286 |
| Best ARIMA | 9.975927 | 31.92362 | 66.47435 |

Table 4: Comparison of MSPE of different methods for three periods as discussed in experiment 3

the information contained in the training data set.

### 4.2.2 Comparison of k-NN Predictor, KM Predictor, FC Predictor and Best ARIMA

In this section, we report the results of a small simulation experiment comparing the FC-predictor with k-NN predictor, KM-predictor and the Best ARIMA predictor using the same set of curves as in section 4.2.1. Functional data generating models as in experiment 2 is considered with a sample of 50 curves as a training set with mixing proportion $\pi_1 = \pi_2 = 0.5$. It is tested on a fresh set of 50 curves, 25 curves each from both population. The whole process is then repeated 50 times and the average value of the MSPE for each of the forecast periods separately is reported in table 5 below. We refer to this as experiment 4.

The reason for doing a much smaller simulation experiment compared to that reported in the earlier section is mainly computational. We encountered several computational issues namely the likelihood becoming extremely small hitting machine 0, the Funclust algorithm failing to converge possibly due to large size of training sample or due to slow convergence of the EM-like algorithm. This indicates that lot of improvement in computational efficacy is required for Funclust algorithm on which the FC-Predictor is based to be useful in practice.

| Method | $X^{\text{new}}(8)$ | $X^{\text{new}}(9)$ | $X^{\text{new}}(10)$ |
|---|---|---|---|
| 25-NN Predictor | 0.0001534 | 0.00361192 | 0.0293429 |
| KM-Predictor | 0.00010548 | 0.00276592 | 0.02320982 |
| FC-Predictor | 0.0002 | 0.0055 | 0.0425 |
| Best ARIMA | 0.04772418 | 0.2985654 | 0.95431502 |

Table 5: Comparison of MSPE of different methods for three periods as discussed in experiment 4

We observe from above experiment that KM-Predictor performs best across four methods. For comparison we perform one more experiment with the functional data generating model as in experiment 3. This simulation is done under same conditions as in experiment 4. Table 6 shows the MSPE for 1 period forecast, 2 period forecast, 3 period forecast for 50 iterations by all four methods. We refer to this experiment as experiment 5.

| Method | $X^{\text{new}}(8)$ | $X^{\text{new}}(9)$ | $X^{\text{new}}(10)$ |
|---|---|---|---|
| 25-NN Predictor | 0.00013426 | 0.00356004 | 0.03005374 |
| KM-Predictor | 0.00010888 | 0.00282346 | 0.02350198 |
| FC-Predictor | 0.00010888 | 0.00282346 | 0.02350198 |
| Best ARIMA | 9.57670348 | 30.93669896 | 64.77716212 |

Table 6: Comparison of MSPE of different methods for three periods as discussed in experiment 5

The performance in terms of MSPE is equally good for KM-Predictor and FC-Predictor.

# 5 Prediction of Non-Stationary Gaussian Processes

With real life data it is often seen that the assumption of stationarity fails to hold. The variances and covariances are seen to vary over time. Other non-stationary behaviors that are often seen are trends, cycles or some combinations of these. The difficulty of modeling and forecasting non-stationary Gaussian processes (GPNS) are well known. In this section, we illustrate the performance of our prediction methods in the case of non-stationary Gaussian Process with three real life data-sets, namely booking curves, growth data and temperature data.

## 5.1 Real data examples

In the first example, we analyze a booking position data of air conditioned 3-tier coaches (AC III) of Gujarat Mail (train no 12902) of Indian Railways which runs from the city of Ahmedabad (capital of the state Gujarat in Western India) to the city of Mumbai (capital of the state of Maharashtra and located on the western coast of India). The data is collected manually starting from 27 November 2014 till 28 February 2015 from the Indian Railways (www.irctc.co.in). During this time, the reservations opened 60 days before the journey date. The total number of seats available for booking under normal conditions is 325 in the AC III category for this train. Booking positions were noted at 9 p.m. on each day for train number 12902 departing from Ahmedabad for 34 consecutive days starting from 26 January, 2015 till 28 February, 2015. This led to 34 functional observations $\mathrm{f}_i$, $i = 1, \cdots, 34$ with $\mathrm{f}_1$ being the observation for 26 January 2015, $\mathrm{f}_2$ being the observation for 27 January 2015 etc. Each observation $\mathrm{f}_i$ is of the type ${}_id_1, {}_id_2, \cdots, {}_id_{60}$ where ${}_id_j$ is the booking position of the train on the j-th day (with day 1 being the day when the booking opens). While recording the data the following convention was used when the status is shown as RAC and Waitlist:

$${}_id_j = \begin{cases} \text{x} & \text{if x number of seats is Available} \\ \text{- x} & \text{if status is RAC with number x} \\ \text{-47-x} & \text{if status is Waitlist with number x} \end{cases}$$

(The number 47 was arrived at by observing that the highest RAC number shown during the 34 day data collection period was 47). The figure 3 shows some of these booking curves.

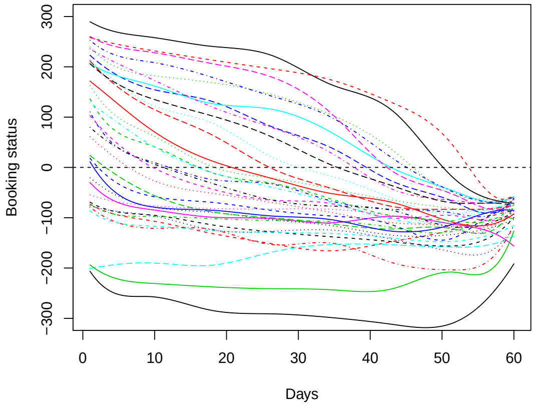

Figure 3: Daily booking position of Gujarat Mail from day of start of booking till departure. Each curve represents a different departure date

In the second example, we consider growth data for females (Tuddenham and Snyder (1954)). The growth dataset comes from Berkeley growth study and is available in the fda package of R. In this dataset, the heights of 54 girls are measured at 31 ages, and these range from 1 to 18 years, unequally spaced. Measurements were taken every three months until two years of age, every year until eight years of age, and then every six months from eight to eighteen years of age. Figure (4) shows growth curves for 24 girls.

In the last example, we consider temperature data of a city located in western India, Ahmedabad. The maximum and minimum temperature of Ahmedabad for the period 1961 to 2012 was provided by the Indian Meteorological department, Ministry of Earth Sciences, Government of India. The average temperature of a month was computed by averaging the maximum and minimum temperature for a month, (World Meteorological Organisation (2012)). The observation of each year is taken as a functional observation. The figure (5) illustrates temperature curves for all 52 years.

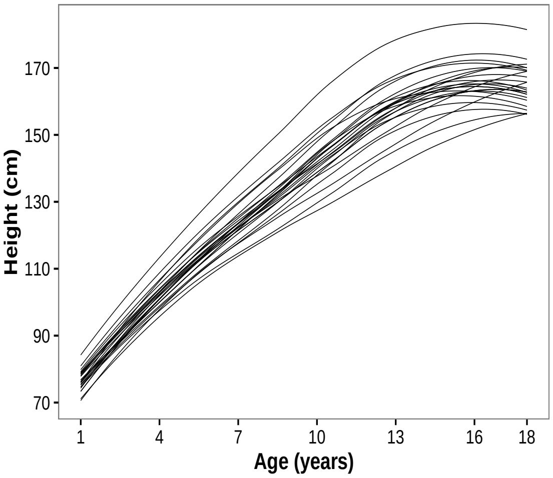

Figure 4: Height of girls in the Berkeley growth study represented as smooth curve. Each curve represents a different girl.

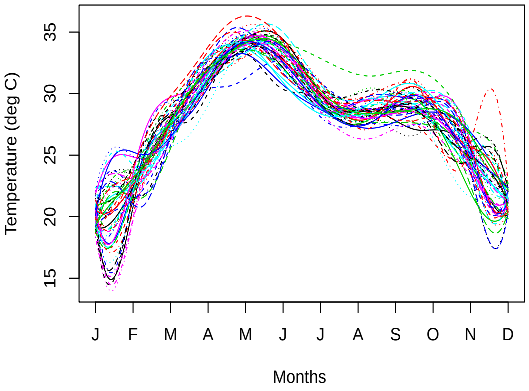

Figure 5: Average monthly temperature curves of the Ahmedabad city represented as smooth curve. Each curve represents a different year

## 5.2 Modeling of non-stationary Gaussian Process

In both examples 1 and 2, we observed that the variance is not constant over time. Thus we need a method to incorporate this information in our covariance function specification. We suggest the following covariance function which can accommodate time varying variance:

$$K(s,t) = \sqrt{d\left(\frac{s}{c}\right)^{\alpha-1}\left(1+\frac{s}{c}\right)^{-\alpha-\beta}}\sqrt{d\left(\frac{t}{c}\right)^{\alpha-1}\left(1+\frac{t}{c}\right)^{-\alpha-\beta}}\exp\left(-w|s-t|^{\gamma}\right) \tag{10}$$

It may be noted that in this specification the variance is modeled using a modification of the unnormalized density function of a beta distribution of the second kind (Johnson and Samuel (1995), p. 325) which can take a wide variety of shapes for various choices of $\alpha$, $\beta$, $c$ and $d$. Figure **??** shows some of these shapes.

We denote a Gaussian process with the above covariance function $K(s,t)$ as $\text{GP}_{\text{NS}}(d,c,\alpha,\beta,w,\gamma)$. The parameters $(d,c,\alpha,\beta,w,\gamma)$ needs to be estimated in practical applications. As before, we assume that each observation is a random sample path from $\text{GP}_{\text{NS}}(d,c,\alpha,\beta,w,\gamma)$ which has been observed at a few discrete points and all the observations are mutually independent.

Let $X^{(i)} = (X^{(i)}(t_1), X^{(i)}(t_2), \cdots, X^{(i)}(t_n))$ denote the $i^{\text{th}}$ functional observation observed at time points $t_1, t_2, \cdots, t_n$; $i = 1, \cdots, m$. We estimate $w$ and $\gamma$ using equation 3 and equation 4. Let $V = (V(t_1), V(t_2), \ldots, V(t_n))$ denote the estimated variances of the $\text{GP}_{\text{NS}}$ process at time points $(t_1, t_2, \ldots, t_n)$. The $d, c, \alpha, \beta$ are estimated by minimizing

$$\sqrt{\sum_{1=1}^{n}\left(d\left(\frac{t_s}{c}\right)^{\alpha-1}\left(1+\frac{t_s}{c}\right)^{-\alpha-\beta} - V(t_s)\right)^2}$$

for the parameters $d, c, \alpha$ and $\beta$. Among the several modern numerical approaches to global optimization of functions the Differential Evolution (DE), Storn and Price (1997) approach is one of the most promising. In the R software, the package DEoptim provides the functionality of optimizing a given function using the DE algorithm. We use the DEoptim function in R for obtaining the estimates of $d, c, \alpha$ and $\beta$. Mullen (2014) reports that the performance of the DEoptim algorithm is good across a variety of benchmark problems.

An alternative approach is to obtain the maximum likelihood estimate (MLE) of the parameters. While we have discussed this briefly below we have not used the same because of the computational difficulties associated with it.
As earlier, assume that the curves $X_1, \cdots, X_m$ is a random sample from $\text{GP}_{\text{NS}}(d,c,\alpha,\beta,w,\gamma)$

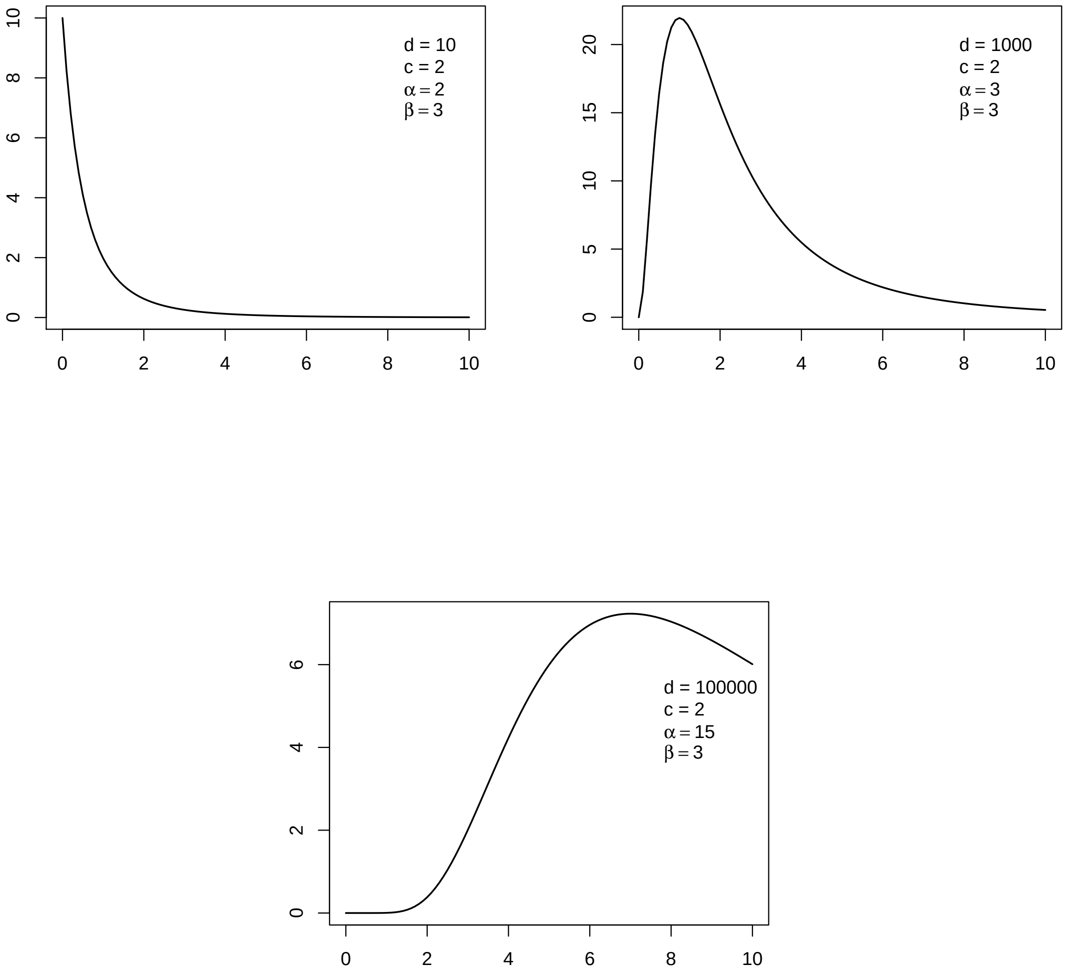

Figure 6: Shapes of beta function by varying parameters $\alpha$, $\beta$, $c$ and $d$ as discussed in section

and each of these curves are observed at same time points $t_1, \cdots, t_n$. The log likelihood of the data is given by:

$$l_0 = l(\mu, d, c, \alpha, \beta, w, \gamma) = \sum_{i=1}^{m} \left( \frac{-n}{2} ln 2\pi - \frac{1}{2} ln|\Sigma| - \frac{1}{2}(X_i - \mu)' \Sigma^{-1} (X_i - \mu) \right) \quad (11)$$

where $\mu = (\mu_{t_1}, \cdots, \mu_{t_n})$, $\Sigma$ is the $n \times n$ matrix whose $(g, h)^{th}$ element is

$$\sqrt{d \left(\frac{t_g}{c}\right)^{\alpha-1} \left(1 + \frac{t_g}{c}\right)^{-\alpha-\beta}} \sqrt{d \left(\frac{t_h}{c}\right)^{\alpha-1} \left(1 + \frac{t_h}{c}\right)^{-\alpha-\beta}} \exp\left(-w|t_g - t_h|^{\gamma}\right)$$

Differentiating equation (11) with respect to $\mu$ and equating to zero yields $\hat{\mu}(t_l) = \frac{\sum_{i=1}^{m} X_i(t_l)}{m}$, $1 \le 1 \le n$.

Writing $\Sigma = \left[\sigma_{ij}\right]_{m \times m}$ and $l_0 = l_0(\mu, \sigma_{11}, \sigma_{12}, \cdots, \sigma_{1m}, \sigma_{22}, \sigma_{23}, \cdots, \sigma_{2m}, \cdots, \sigma_{mm})$ and $\sigma_{ij} = \sigma_{ij}(d, c, \alpha, \beta, w, \gamma)$

$$\frac{\partial l_0}{\partial w} = \sum_{i,j=1}^{n} \frac{\partial l_0}{\partial \sigma_{ij}} \frac{\partial \sigma_{ij}}{\partial w} \quad (12)$$

$$\frac{\partial \sigma_{ij}}{\partial w} = -\sigma_{ij} |t_i - t_j|^{\gamma} \quad (13)$$

and by using Smith (1978),

$$\frac{\partial l_0}{\partial \sigma_{ij}} = \frac{1}{2} \mathrm{tr} \left\{ \left[ -m\Sigma^{-1} + \Sigma^{-1} \left( \sum_{k=1}^{m} (X_k - \mu)(X_k - \mu)' \right) \Sigma^{-1} \right] \left[ \frac{\partial \Sigma}{\partial \sigma_{ij}} \right] \right\} \quad (14)$$

The values from equation (13) and equation (14) are substituted in equation (12) and the resulting expression is equated to zero. Similarly $\frac{\partial l_0}{\partial \alpha}$, $\frac{\partial l_0}{\partial \beta}$, $\frac{\partial l_0}{\partial \gamma}$, $\frac{\partial l_0}{\partial c}$, and $\frac{\partial l_0}{\partial d}$ are calculated and are all equated to 0. These six equations can then be solved simultaneously to obtain the MLE of these six parameters. As can be seen from 7, the analytical solutions of these equations are not tractable. While numerical techniques can be used to obtain the solutions of these equations, these too are not straightforward. Alternatively, one may attempt to maximize the log-likelihood function directly using a global optimization algorithm which in many cases can be computationally very expensive.

## 5.3 Results

While dealing with real life datasets it is essential to check for presence of outliers. The outlier in functional setting is described as follows: A curve is an outlier if it has been generated by a stochastic process with a different distribution than the rest of the curves (Febrero et al. (2007)). In general, outliers in a functional dataset can arise due to measurement, recording and typing errors which can be corrected whenever possible to detect or they may be data curves that come from a distribution other than rest of the curves. The concept of functional depth has been developed to measure the centrality of a given curve. Functional depth has been defined by various authors, out of which we use the sample Fraiman and Muniz depth (FMD), Fraiman and Muniz (2001). This is defined as

$$\mathrm{SFMD}_n(X^{(i)}) = \sum_{j=2}^{n} \triangle_j \left[ 1 - \left| \frac{1}{2} - \frac{1}{n} \sum_{k=1}^{m} I(X^{(k)}(t_j) \leq X^{(i)}(t_j)) \right| \right]$$

where I(.) is the indicator function, $X^{(i)}, i = 1, \cdots, m$ are the sample curves (assumed to belong to $C[a,b]$, the space of continuous functions defined on the interval $[a,b] \subset \mathbb{R}$), $t_1, \ldots, t_n$ are the points at which $X^{(i)}, i = 1, \cdots, m$ are observed and $\Delta_j = t_j - t_{j-1}$.

If an outlier is present in the data set, then it would have much lower value of FMD compared to the other observations. This fact is used for detection of outliers. In R software, outliers.depth.trim function in fda.usc package performs outlier detection using FMD. This function is used by us to detect presence of outliers in all the three datasets. In the railway booking position data discussed in the first example this method identified two outliers which were removed before proceeding with further analysis. In the second example where growth data of girls is studied this method identified one outlier which was also removed before analyzing the data further. In the temperature data of Ahmedabad discussed in the third example this method did not indicate presence of any outlier.

*Booking position data*: For reasons discussed in remark, we do not work with $f_i$ but with $g_i = ({}_id_5, {}_id_{10}, \cdots, {}_id_{40})$. We treat each observation $g_i$ as a discretely observed random sample path from $\mathrm{GP}_{\mathrm{NS}}(d, c, \alpha, \beta, w, \gamma)$. In the railway booking position dataset which had 32 observations after deletion of outliers, we use 28 of them to train the model and use the remaining four observations to test the prediction accuracy of the methods. For each of the four observations we predict the booking positions on 35th and 40th day since beginning of booking, using the knowledge of the booking positions on $5^{\text{th}}, 10^{\text{th}}, \ldots, 30^{\text{th}}$ days. We compare the predictions obtained using the 10-NN, KM and FC predictors with the Best ARIMA forecasts obtained using the auto.arima function in the forecast package of R software. The results are given in table 7. Also, the predictions are plotted in figure 7 where different types of line show prediction using different methods. For comparing the prediction accuracy of the methods in quantitative terms we use the MSPE which is computed using all the eight points for which predictions were obtained. The results are shown in the column with heading Booking Position in table 10. We find the 10-NN Predictor performs best in terms of having minimum MSPE among the four methods under consideration.

| Method | ${}_{29}d_{35}$ | ${}_{29}d_{40}$ | ${}_{30}d_{35}$ | ${}_{30}d_{40}$ | ${}_{31}d_{35}$ | ${}_{31}d_{40}$ | ${}_{32}d_{35}$ | ${}_{32}d_{40}$ |
|---|---|---|---|---|---|---|---|---|
| Actual value | 160 | 139 | -161 | -162 | -95 | -103 | 66 | 19 |
| 10-NN Predictor | 170 | 134 | -166 | -171 | -98 | -105 | 64 | 31 |
| KM Predictor | 176 | 138 | -159 | -160 | -96 | -104 | 75 | 43 |
| FC Predictor | 173 | 144 | -170 | -175 | -103 | -111 | 74 | 52 |
| Best ARIMA | 193 | 180 | -177 | -191 | -101 | -111 | 80 | 61 |

Table 7: Comparison of predictions obtained using different methods for the booking position data, section 5.3

Remark: In the training samples of full booking curves, the number of samples ($n = 28$) is much smaller than the number of parameters ($p = 60$) to be estimated. It is known that when $n < p$ the covariance matrix is not positive definite and hence it cannot be inverted to compute the inverse of the covariance matrix, which is required in many applications like in estimating the maximum likelihood estimator. Hence, we reduce the number of parameters to eight i.e. we take the observation on every fifth day till 40 days. However, it is also known that when $n > p$, the eigenstructure tends to be systematically distorted, resulting in ill-conditioned estimators for covariance except when $\frac{p}{n}$ is extremely small(Won et al. (2009)).

*Growth data*: In the growth data for girls dataset we had 53 observations after deletion of the outlier. We used 49 of them to train the model and used the remaining four observations to test the prediction accuracy. For each of the four observations we predicted the height at age 17.5 and 18 years given that we know her height at 29 time points till 17 years of age. We compare the predictions obtained using 10-NN, KM and FC predictors. The Best ARIMA method is not suitable for use here as the

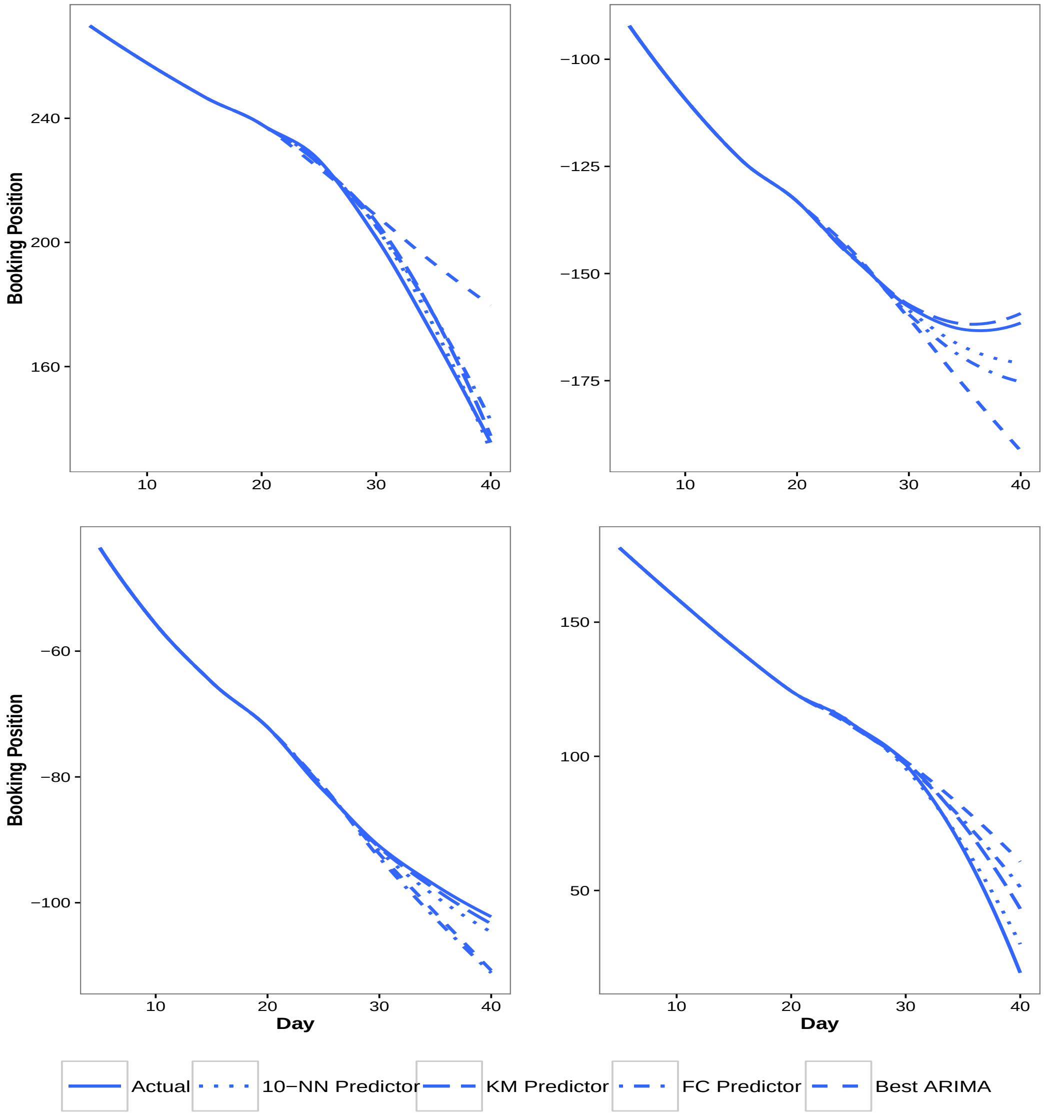

Figure 7: Prediction of Booking position for four days on 35th and 40th day. Each plot represents prediction of a day using all four methods. The smooth line shows the actual values till 30th day and then different lines show prediction using different methods.

29 time points where the height of a girl are measured are not equally spaced. The results are given in table 8. Also, the predictions are plotted in figure 8. The MSPE of the three methods are shown in the column with heading Growth in table 10. We find that KM-Predictor performs best in terms of having minimum MSPE among three methods under consideration.

| Method | ${}_{50}d_{17.5}$ | ${}_{50}d_{18}$ | ${}_{51}d_{17.5}$ | ${}_{51}d_{18}$ | ${}_{52}d_{17.5}$ | ${}_{52}d_{18}$ | ${}_{53}d_{17.5}$ | ${}_{53}d_{18}$ |
|---|---|---|---|---|---|---|---|---|
| Actual value | 173.1 | 173.5 | 166.3 | 166.8 | 168.4 | 168.6 | 168.9 | 169.2 |
| 10-NN Predictor | 172.80 | 172.97 | 166.09 | 166.15 | 168.47 | 168.73 | 168.74 | 168.98 |
| KM Predictor | 173.15 | 173.28 | 166.10 | 166.23 | 168.73 | 168.85 | 168.97 | 169.05 |
| FC Predictor | 172.59 | 172.48 | 166.15 | 166.29 | 168.20 | 168.12 | 168.42 | 168.28 |

Table 8: Comparison of predictions obtained using different methods for the growth data, section 5.3

*Temperature data*: The temperature data of Ahmedabad city had 52 observations, out of which 48 observations were used to train the model and the remaining four observations were used to test the prediction accuracy. In this case, we treat the 12 monthly average temperature readings for a given year as observations from a random sample path of $\mathrm{GP_S}(\mu, v, w, \gamma)$ observed at time points $t = 1, \cdots, 12$. Further, we assume that the 52 random sample paths corresponding to the years 1961-2012 are mutually independent.

The equation (1) could not be used directly for this data because of the nature of the monthly temperature data. It is natural to expect that the monthly temperatures of the months of January and December in a year would be correlated because both these months fall in the winter season in India. However this would not be correctly captured by the covariance function $K(s,t) = v\ \exp(-w|s-t|^{\gamma})$ since the month of January with $s = 1$ would appear to be distant from December with $t = 12$ and therefore the model would expect the temperatures of these months to have low correlation. To overcome the problem we use the following modified covariance function $K(s,t) = v\ \exp\left(-w\left(2\left|\frac{\sin(s-t)}{2}\right|\right)^{\gamma}\right)$ ,which is discussed in a different context in Solin and Särkkä (2014). The motivation behind this formulation is that, if the twelve months in a year are viewed as 12 equispaced points on the unit circle with

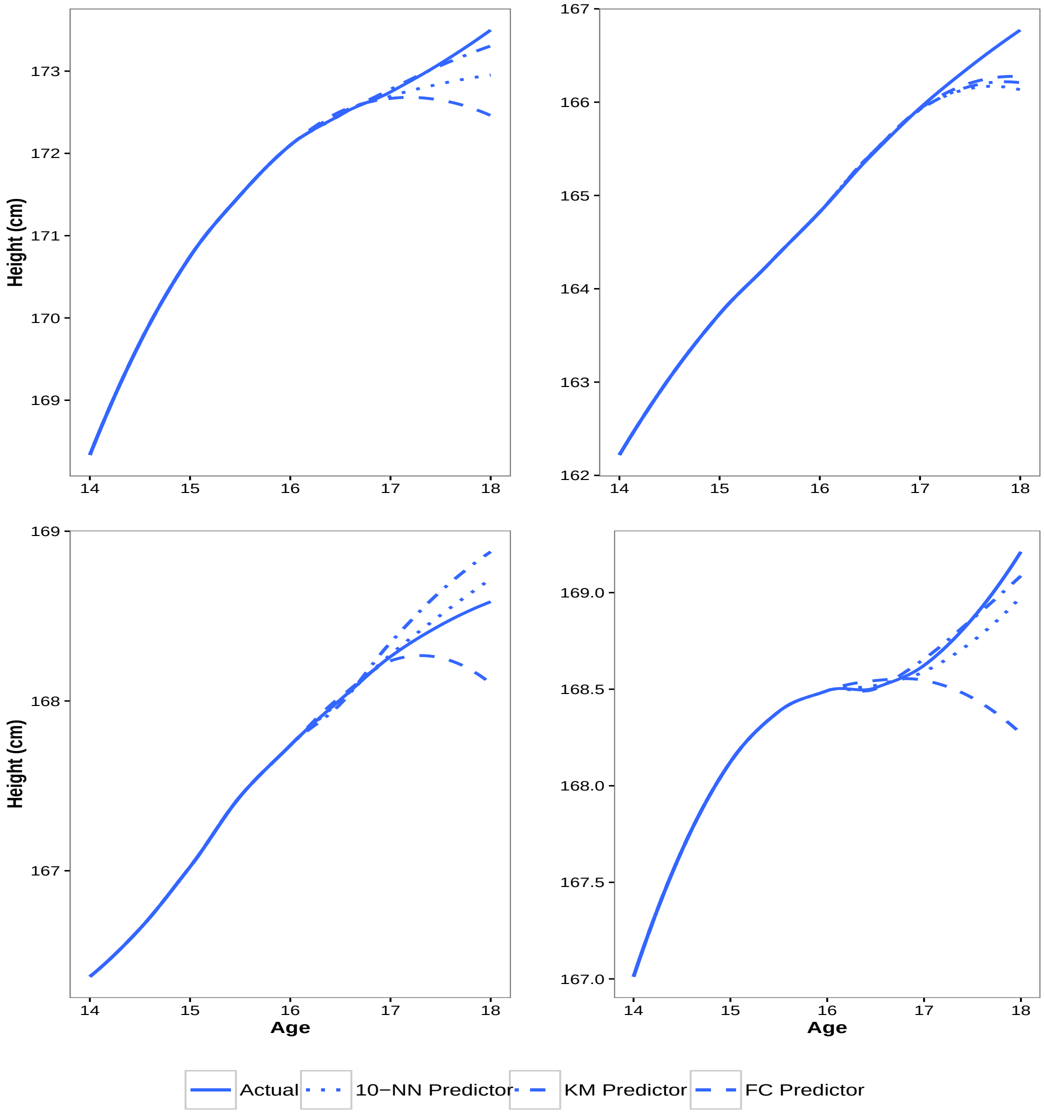

Figure 8: Prediction of height of 4 girls at the age 17.5 and 18. Each plot represents a different girl with all four methods. The smooth line shows the actual values till 17th year and then different lines show prediction using different methods.

coordinates $\left(\cos\frac{\pi \text{i}}{6}, \sin\frac{\pi \text{i}}{6}\right)$, $i = 1, \cdots, 12$ then the points representing January and December are adjacent to one another. The Euclidean distance between two points $\left(\cos\frac{\pi \text{s}}{6}, \sin\frac{\pi \text{s}}{6}\right)$ and $\left(\cos\frac{\pi \text{t}}{6}, \sin\frac{\pi \text{t}}{6}\right)$ is $2\left|\sin\frac{s-t}{2}\right|$ which is used to replace $|s-t|$ in the original definition of the covariance function, equation (1).

In this case, we estimate v using equation (2). $\gamma$ is estimated as

$$\hat{\gamma} = \frac{\ln\left(\frac{\ln\hat{\rho}_{\text{A}}}{\ln\hat{\rho}_{\text{B}}}\right)}{\ln|z_2 - z_1| - \ln|z_3 - z_1|} = \frac{\ln\left(\frac{\ln\hat{\rho}_{\text{A}}}{\ln\hat{\rho}_{\text{B}}}\right)}{\ln\left(\sqrt{2-\sqrt{3}}\right)}$$

where $\hat{\rho}_{\text{A}} = \frac{\hat{\rho}_{1,2} + \cdots \hat{\rho}_{k-1,k}}{k-1}$ and $\hat{\rho}_{\text{B}} = \frac{\hat{\rho}_{1,3} + \cdots \hat{\rho}_{k-2,k}}{k-2}$ and
$z_i = \left(\cos\left(\frac{\pi i}{6}\right), \sin\left(\frac{\pi i}{6}\right)\right)$.
$w$ is estimated as

$$\hat{w} = \frac{-\ln\hat{\rho}_{\text{A}}}{|z_2 - z_1|^{\hat{\gamma}}} = \frac{-\ln\hat{\rho}_{\text{A}}}{\ln\left(\sqrt{2-\sqrt{3}}\right)^{\hat{\gamma}}}$$

For each of the four observations we predicted the average temperatures for the months of November and December using our knowledge of the average temperatures for the months of January till October. The predictions obtained using 10-NN, KM, FC predictors and Best ARIMA were compared. The results are mentioned in Table 9. The MSPE of the three methods are shown in the column with heading Temperature in table (10). We find that KM-Predictor performs best in terms of having minimum MSPE among the three methods under consideration.

# 6 Conclusion

This paper presents several new methods for prediction of data coming from an underlying Gaussian Process. For the stationary case the powered exponential covariance function is considered whereas for the non-stationary case suitable modifications of the same are used. When the data comes from an underlying stationary Gaussian Process the CE-Predictor is seen to perform the best. In cases where it may be suspected that the data comes from a mixture of stationary Gaussian processes then the KM-Predictor is seen to perform better than the rest. Even in the non-stationary

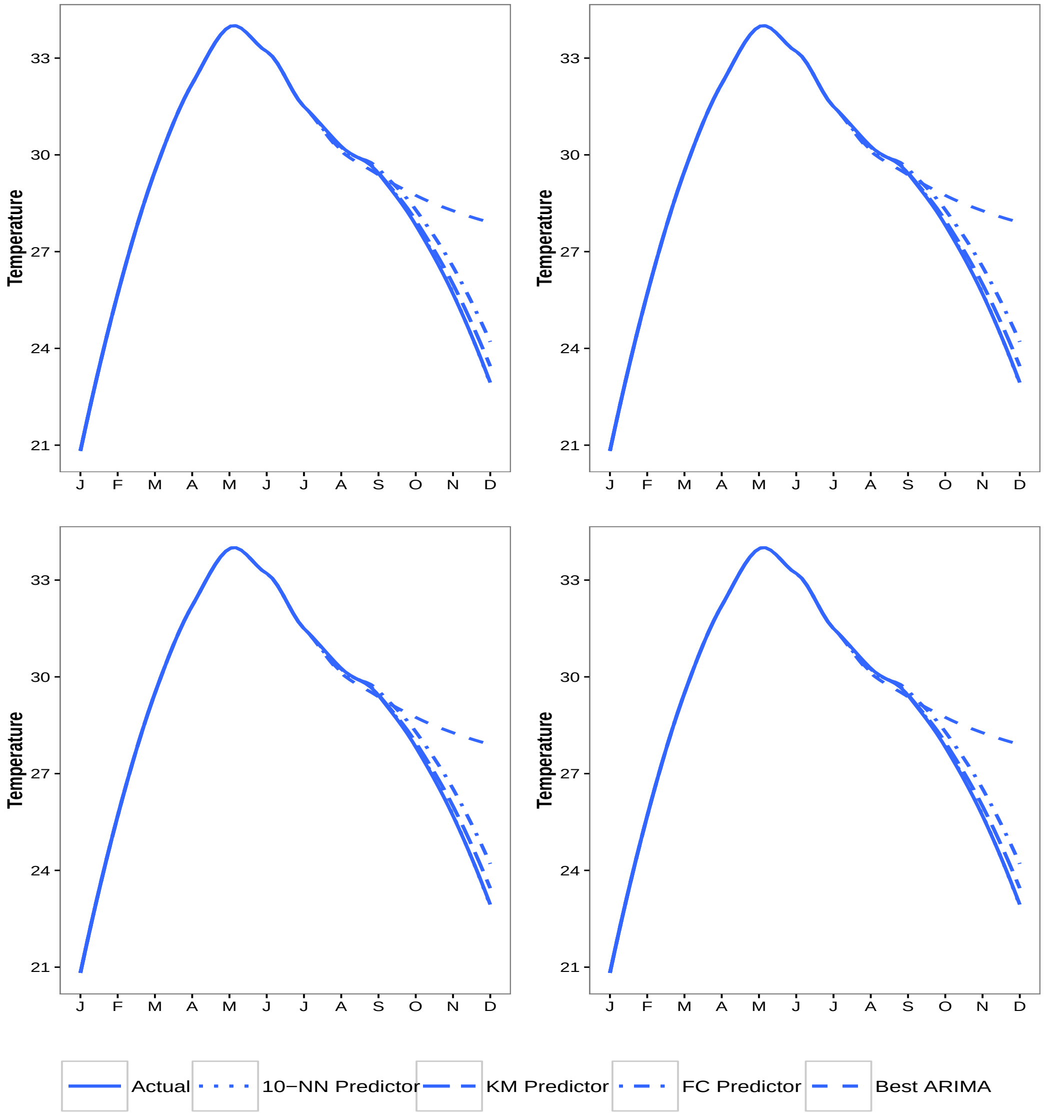

Figure 9: Prediction of temperature for November and December for 4 years. Each plot represents a different year with all four methods. The smooth line shows the actual values till October and then different lines show prediction using different methods.

| Method | 2009$d_{11}$ | 2009$d_{12}$ | 2010$d_{11}$ | 2010$d_{12}$ | 2011$d_{11}$ | 2011$d_{12}$ | 2012$d_{11}$ | 2012$d_{12}$ |
|---|---|---|---|---|---|---|---|---|
| Actual value | 25.5 | 22.70 | 25.70 | 20.55 | 26.60 | 22.25 | 23.50 | 22.20 |
| 10-NN Predictor | 25.51 | 21.97 | 25.95 | 22.17 | 24.79 | 21.31 | 25.06 | 21.32 |
| KM Predictor | 25.75 | 22.28 | 25.93 | 22.45 | 25.00 | 21.47 | 24.82 | 21.25 |
| FC Predictor | 26.22 | 22.62 | 25.21 | 21.58 | 24.21 | 21.23 | 24.98 | 21.36 |
| Best ARIMA | 28.08 | 27.91 | 30.16 | 30.59 | 28.79 | 28.65 | 28.08 | 28.04 |

Table 9: Comparison of predictions obtained using different methods for the temperature data, section 5.3.

| | Booking Position | Height | Temperature |
|---|---|---|---|
| 10-NN Predictor | 49 | 0.116 | 1.512 |
| KM Predictor | 115.5 | 0.076 | 1.378 |
| FC Predictor | 222.285 | 0.366 | 1.565 |
| Best ARIMA | 740.875 | * | 35.525 |

Table 10: Mean Square Prediction Error for Booking position, Growth and Temperature data using different methods.

case the KM-Predictor performs quite well giving the best prediction in two of the three real life data sets considered in this paper. In the remaining data set the k-NN Predictor (with k = 10) performs the best with the KM-Predictor being the second best. Based on this experience we think that the KM-Predictor can be profitably used for predicting in several real life situations such as traffic volumes in telecom networks, electricity demand at different times for a facility etc. In addition, when working with real data one may adopt an ensemble forecasting approach with the ensemble consisting of the KM-Predictor and k-NN Predictor.

# References

Alonso, A. M., Casado, D., and Romo, J. (2012). Supervised classification for functional data: A weighted distance approach. *Computational Statistics & Data Analysis*, 56(7):2334–2346.

Antoniadis, A., Brossat, X., Cugliari, J., and Poggi, J.-M. (2016). A prediction interval for a function-valued forecast model: Application to load forecasting. *International Journal of Forecasting*, 32(3):939–947.

Arya, S., Mount, D. M., Netanyahu, N. S., Silverman, R., and Wu, A. Y. (1998). An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM (JACM)*, 45(6):891–923.

Bosq, D. (2000). *Linear processes in function spaces*. Springer, NewYork.

Chiou, J.-M. (2012). Dynamical functional prediction and classification, with application to traffic flow prediction. *The Annals of Applied Statistics*, pages 1588–1614.

Erbas, B., Ullah, S., Hyndman, R. J., Scollo, M., and Abramson, M. (2012). Forecasts of copd mortality in australia: 2006-2025. *BMC medical research methodology*, 12(1):1.

Febrero, M., Galeano, P., and González-Manteiga, W. (2007). A functional analysis of nox levels: location and scale estimation and outlier detection. *Computational Statistics*, 22(3):411–427.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media.

Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *Test*, 10(2):419–440.

Ghosh, A. K. (2006). On optimum choice of k in nearest neighbor classification. *Computational Statistics & Data Analysis*, 50(11):3113–3123.

Hall, P., Park, B. U., and Samworth, R. J. (2008). Choice of neighbor order in nearest-neighbor classification. *The Annals of Statistics*, pages 2135–2152.

Hartigan, J. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Horváth, L. and Kokoszka, P. (2012). *Inference for functional data with applications*. Springer.

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons.

Hyndman, R. J. and Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: a functional data approach. *Computational Statistics & Data Analysis*, 51(10):4942–4956.

Jacques, J. and Preda, C. (2013). Funclust: A curves clustering method using functional random variables density approximation. *Neurocomput.*, 112:164–171.

Johnson, Norman L., K. and Samuel, B. N. (1995). *Continuous Univariate Distributions, Vol 2*. John Wiley & Sons, Inc.

Laukaitis, A. (2008). Functional data analysis for cash flow and transactions intensity continuous-time prediction using hilbert-valued autoregressive processes. *European Journal of Operational Research*, 185(3):1607–1614.

Mullen, K. M. (2014). Continuous Global Optimization in R. *Journal of Statistical Software*, 60(6):1–45.

Müller, H.-G. and Yang, W. (2010). Dynamic relations for sparsely sampled gaussian processes. *Test*, 19(1):1–29.

Ramsay, J. and Silverman, B. (2002). *Applied Functional Data Analysis: Methods and case studies*. Springer.

Ramsay, J. O. and Silverman, B. (2005). *Functional data analysis*. Springer.

Shi, J. Q. and Choi, T. (2011). *Gaussian process regression analysis for functional data*. CRC Press.

Smith, D. W. (1978). A simplified approach to the maximum likelihood estimation of the covariance matrix. *The American Statistician*, 32(1):28–29.

Solin, A. and Särkkä, S. (2014). Explicit link between periodic covariance functions and state space models. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33, pages 904–912.

Storn, R. and Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359.

Tuddenham, R. D. and Snyder, M. M. (1954). Physical growth of california boys and girls from birth to eighteen years. *Publications in child development. University of California, Berkeley*, 1(2):183.

Wahba, G. (1990). *Spline models for observational data*, volume 59. SIAM, Philadelphia.

Won, J. H., Lim, J., Kim, S.-J., and Rajaratnam, B. (2009). Mmaximum Likelihood Covariance Estimation with a Condition Number Constraint . Technical report, Stanford University, Department of Statistics.

World Meteorological Organisation (2012). WMO statement on the status of the global climate in 2011.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590.

Zhang, J.-T. (2013). *Analysis of variance for functional data*. CRC Press.

# 7 APPENDIX A

In this section, we provide the partial derivatives of $l_o$ which is required for finding MLEs for parameters d,c,$\alpha$, $\beta$. For estimation of d, we need $\frac{\partial l_0}{\partial d}$ which can be found as

$$\frac{\partial l_0}{\partial d} = \sum_{i,j=1}^{n} \frac{\partial l_0}{\partial \sigma_{ij}} \frac{\partial \sigma_{ij}}{\partial d} \tag{15}$$

$$\begin{aligned}\frac{\partial \sigma_{ij}}{\partial d} = & \left\{ \frac{\left(\frac{t_i}{c}\right)^{\alpha-1}\left(\frac{t_i}{c}+1\right)^{-\alpha-\beta}\sqrt{d\left(\frac{t_j}{c}\right)^{\alpha-1}\left(\frac{t_j}{c}+1\right)^{-\alpha-\beta}}}{2\sqrt{d\left(\frac{t_i}{c}\right)^{\alpha-1}\left(\frac{t_i}{c}+1\right)^{-\alpha-\beta}}} \right. \\ & \left. + \frac{\left(\frac{t_j}{c}\right)^{\alpha-1}\left(\frac{t_j}{c}+1\right)^{-\alpha-\beta}\sqrt{d\left(\frac{t_i}{c}\right)^{\alpha-1}\left(\frac{t_i}{c}+1\right)^{-\alpha-\beta}}}{2\sqrt{d\left(\frac{t_j}{c}\right)^{\alpha-1}\left(\frac{t_j}{c}+1\right)^{-\alpha-\beta}}} \right\} \exp\left(-w|t_i-t_j|^{\gamma}\right)\end{aligned} \tag{16}$$

and by using Smith (1978),

$$\frac{\partial l_0}{\partial \sigma_{ij}} = \frac{1}{2}\mathrm{tr}\left\{ \left[ -m\Sigma^{-1} + \Sigma^{-1}\left(\sum_{k=1}^{m}\left(X_k-\mu\right)\left(X_k-\mu\right)'\right)\Sigma^{-1} \right] \left[\frac{\partial \Sigma}{\partial \sigma_{ij}}\right] \right\} \tag{17}$$

The values from 17 and 16 are substituted in 15 to get $\frac{\partial l_0}{\partial d}$

Similarly,

$$\frac{\partial l_0}{\partial c} = \sum_{i,j=1}^{n} \frac{\partial l_0}{\partial \sigma_{ij}} \frac{\partial \sigma_{ij}}{\partial c} \tag{18}$$

$$\frac{\partial \sigma_{ij}}{\partial c} = \left\{ \frac{\left(-\frac{(\alpha-1)dt_i\left(\frac{t_i}{c}\right)^{\alpha-2}\left(\frac{t_i}{c}+1\right)^{-\alpha-\beta}}{c^2} - \frac{dt_i(-\alpha-\beta)\left(\frac{t_i}{c}\right)^{\alpha-1}\left(\frac{t_i}{c}+1\right)^{-\alpha-\beta-1}}{c^2}\right)\sqrt{d\left(\frac{t_j}{c}\right)^{\alpha-1}\left(\frac{t_j}{c}+1\right)^{-\alpha-\beta}}}{2\sqrt{d\left(\frac{t_i}{c}\right)^{\alpha-1}\left(\frac{t_i}{c}+1\right)^{-\alpha-\beta}}} \right.$$

$$\left. + \frac{\sqrt{d\left(\frac{t_i}{c}\right)^{\alpha-1}\left(\frac{t_i}{c}+1\right)^{-\alpha-\beta}}\left(-\frac{(\alpha-1)dt_j\left(\frac{t_j}{c}\right)^{\alpha-2}\left(\frac{t_j}{c}+1\right)^{-\alpha-\beta}}{c^2} - \frac{dt_j(-\alpha-\beta)\left(\frac{t_j}{c}\right)^{\alpha-1}\left(\frac{t_j}{c}+1\right)^{-\alpha-\beta-1}}{c^2}\right)}{2\sqrt{d\left(\frac{t}{c}\right)^{\alpha-1}\left(\frac{t_j}{c}+1\right)^{-\alpha-\beta}}} \right\}$$

$$\exp\left(-w|t_i - t_j|^{\gamma}\right) \tag{19}$$

The value of $\frac{\partial l_0}{\partial c}$ is estimated using 17, 18 and 19.

Again,

$$\frac{\partial l_0}{\partial \alpha} = \sum_{i,j=1}^{n} \frac{\partial l_0}{\partial \sigma_{ij}} \frac{\partial \sigma_{ij}}{\partial \alpha} \tag{20}$$

$$\frac{\partial \sigma_{ij}}{\partial \alpha} = \left\{ \frac{\sqrt{d\left(\frac{t_j}{c}\right)^{\alpha-1}\left(\frac{t_j}{c}+1\right)^{-\alpha-\beta}}\left(d\left(\frac{t_i}{c}\right)^{\alpha-1}\log\left(\frac{t_i}{c}\right)\left(\frac{t_i}{c}+1\right)^{-\alpha-\beta} - d\left(\frac{t_i}{c}\right)^{\alpha-1}\log\left(\frac{t_i}{c}+1\right)\left(\frac{t_i}{c}+1\right)^{-\alpha-\beta}\right)}{2\sqrt{d\left(\frac{t_i}{c}\right)^{\alpha-1}\left(\frac{t_i}{c}+1\right)^{-\alpha-\beta}}} \right.$$

$$\left. + \frac{\sqrt{d\left(\frac{t_i}{c}\right)^{\alpha-1}\left(\frac{t_i}{c}+1\right)^{-\alpha-\beta}}\left(d\left(\frac{t_j}{c}\right)^{\alpha-1}\log\left(\frac{t_j}{c}\right)\left(\frac{t_j}{c}+1\right)^{-\alpha-\beta} - d\left(\frac{t_j}{c}\right)^{\alpha-1}\log\left(\frac{t_j}{c}+1\right)\left(\frac{t_j}{c}+1\right)^{-\alpha-\beta}\right)}{2\sqrt{d\left(\frac{t_j}{c}\right)^{\alpha-1}\left(\frac{t_j}{c}+1\right)^{-\alpha-\beta}}} \right\}$$

$$\exp\left(-w|t_i - t_j|^{\gamma}\right) \tag{21}$$

$\frac{\partial l_0}{\partial \alpha}$ is estimated using 17, 20 and 21.

Similarly,

$$\frac{\partial l_0}{\partial \beta} = \sum_{i,j=1}^{n} \frac{\partial l_0}{\partial \sigma_{ij}} \frac{\partial \sigma_{ij}}{\partial \beta} \tag{22}$$

$$\begin{aligned}\frac{\partial \sigma_{ij}}{\partial \beta} = &\left\{ -\frac{d\left(\frac{t_i}{c}\right)^{\alpha-1} \log\left(\frac{t_i}{c}+1\right)\left(\frac{t_i}{c}+1\right)^{-\alpha-\beta} \sqrt{d\left(\frac{t_j}{c}\right)^{\alpha-1}\left(\frac{t_j}{c}+1\right)^{-\alpha-\beta}}}{2\sqrt{d\left(\frac{t_i}{c}\right)^{\alpha-1}\left(\frac{t_i}{c}+1\right)^{-\alpha-\beta}}} \right. \\ &\left. - \frac{d\left(\frac{t_j}{c}\right)^{\alpha-1} \log\left(\frac{t_j}{c}+1\right)\left(\frac{t_j}{c}+1\right)^{-\alpha-\beta} \sqrt{d\left(\frac{t_i}{c}\right)^{\alpha-1}\left(\frac{t_i}{c}+1\right)^{-\alpha-\beta}}}{2\sqrt{d\left(\frac{t_j}{c}\right)^{\alpha-1}\left(\frac{t_j}{c}+1\right)^{-\alpha-\beta}}} \right\} \\ &\exp\left(-w|t_i - t_j|^{\gamma}\right)\end{aligned} \tag{23}$$

$\frac{\partial l_0}{\partial \beta}$ is estimated by substituting 17 and 23 in eq.22.