Learning a Gaussian Process Model on the Riemannian Manifold of Non-decreasing Distribution Functions

Chafik Samir, Jean-Michel Loubes, Anne-Françoise Yao, François Bachoc

Learning a Gaussian Process Model on the Riemannian Manifold of Non-decreasing Distribution Functions[*]

Chafik Samir
University of Clermont Auvergne
Jean-Michel Loubes
Institut de Mathématiques de Toulouse
Anne-Françoise Yao
University of Clermont Auvergne
François Bachoc
Institut de Mathématiques de Toulouse

September 4, 2019

Abstract

In this work, we consider the problem of learning regression models from a finite set of functional objects. In particular, we introduce a novel framework to learn a Gaussian process model on the space of Strictly Non-decreasing Distribution Functions (SNDF). Gaussian processes (GPs) are commonly known to provide powerful tools for non-parametric regression and uncertainty estimation on vector spaces. On top of that, we define a Riemannian structure of the SNDF space and we learn a GP model indexed by SNDF. Such formulation enables to define an appropriate covariance function, extending the Matérn family of covariance functions. We also show how the full Gaussian process methodology, namely covariance parameter estimation and prediction, can be put into action on the SNDF space. The proposed method is tested using multiple simulations and validated on real-world data.

Keywords: Gaussian process; Riemannian manifold; Functional data.

1 Introduction

In this paper, we consider the problem of learning regression models from a finite set of functional objects. This problem has become very common in several contexts of appli-

[*] The authors thank the ANITI program (Artificial Natural Intelligence Toulouse Institute) and the ANR Project RISCOPE (Risk-based system for coastal flooding early warning). JM Loubes acknowledges the funding by DEEL-IRT and C. Samir acknowledges the funding by CNRS Prime.

cations, including science and technology. For example in functional data analysis and medical data it is very common to compare two objects (functions, curves, surfaces, volumes, etc.) in order to find optimal correspondences between their representations. This methodology, is usually refereed to as statistical shape analysis in [13, 24, 18]. The mathematical formulation leads to a wide range of applications when studying temporal or spatial changes to characterize a population or to build predictive models [11, 26]. In particular, we are interested in studying variations corresponding to domain deformations in observed objects. For instance, the human heart beating of the same person during a cycle can be different under different circumstances. Hence any regression model should take into account the domain (timing) difference (deformation) in observations when studying them. Consequently, such models will become more realistic, efficient, and parsimonious. Many authors have studied registration methods using dynamic time warping models or semi-parametric deformation models, see for instance [9, 19, 10].

Gaussian process regression has been successfully applied in many fields. It has been introduced in [15] by Kolmogrov in the 1940s and applied for multivariate regression starting from 1960s. As a supervised learning process, we will refer to a Kriging approach without restriction to any area of research such as geostatistics, time series, etc. The usual Kriging procedure consists in assuming that we observe a random process $Z = (Z_I)$ indexed by an object I living on a compact space $\mathcal{E}$, also called the index space. Hence predicting unobserved values of the process leads to estimating conditional expectation which can be done as soon as a covariance, between the process observed at different locations, can be defined. Actually for I_i and I_j in $\mathcal{E}$, the main issue is to build a proper covariance between Z_{I_i} and Z_{I_j}. In particular, this covariance can define a notion of stationarity for the process. In this work, we consider the case where for any $i = 1, \dots, N$ I_i is a nonlinear deformation of a common pattern I^*. In this framework, we assume that I_i can be written as $I_i = I^* \circ F_i$ with F_i being a strictly non-increasing distribution, and I being a one-dimensional real-valued function.

In order to capture deformations between observed functions I's and perform optimal predictions for unobserved data, we thus consider a Gaussian Process on the space of distributions functions $\mathcal{F}$ where $Z \sim GP(m, C)$ is defined by a mean function $m : \mathcal{F} \to \mathbb{R}$ and a covariance function $C : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$. To reach such goal we will present properties for the Gaussian process on $\mathcal{F}$ using isometric mappings defined in Section 2.2. Indeed, the regression problem on the space of strictly non-decreasing distribution functions will be re-defined as follows. Given a finite set of observations $\{(F_i, y_i) \in \mathcal{F} \times \mathbb{R}, \quad 1 \leq i, \leq n\}$, define a regression model and an estimate of the conditional expectation $\mathbb{E}[y|y_1, \dots, y_n]$, for a new pair (F, y)

Applications. In this section, we describe some typical applications and explain why they require regression on distribution functions. We consider some applications which belong to the case when the observed curves are real-valued functions (functional data) defined on an interval of the real line. We point out that the choice of application is independent of the proposed method. In particular, we are interested in studying variability in different groups and utilizing full function patterns from such analysis for subsequent classification. An added difficulty in the problem at hand is the fact that the observed objects do not have the same parametrization of the domain. Without loss of generality, we consider that their respective parametrization are deformations (as a convolution with a distribution function) of the unit interval $[0, 1]$. In other words, different individuals will present different amounts of time, and thus, it becomes important to focus on deformations rather than the common pattern. Thus, we require a comprehensive statistical framework for analysis of functional data that allows for statistically analyzing variabilities. There exists a large body of literature on statistical analysis of functions; see for example [20, 14, 26, 23]. When restricting to the analysis of functions that require temporal alignment, the literature is somehow limited [10, 17, 16, 12, 24]. Another class of methods are variations of the Dynamic Time Warping (DTW) algorithm that was first applied for speech recognition [22] and has been extended to other engineering and computer science areas [8]. The main differences between our suggested method and those in the literature are: (i) deformations are continuous in our formulations instead of discrete vectors, (ii) we present the asymptotic properties of the regression model, and (iii) our formulation can be extended for other domains' deformations.

1.1 Contribution:

The goal of this study is to develop a new set of measures that can enhance classification of functional observations based on distribution functions as element of a Riemannian manifold. To that end, we show how Gaussian process regression works in this setting. Beyond the previous methods, the main contribution of this paper can be summarized in:

- We consider the problem of prediction from a finite set of an observed pattern I^* where the randomness is caused by deformations of its domain. When the deformations are strictly nondecreasing distributions in a space $\mathcal{F}$, we propose to consider a Gaussian process regression on $\mathcal{F}$ with a covariance defined on a Riemannian manifold.
- Since a Gaussian process is determined by its mean function m and covariance function C, we focus on C and we extend the Matérn covariance functions on the SNDF space.
- We study the asymptotic properties of the Gaussian process model, by showing a general microergodicity result.

The rest of this paper is organized as follows. Section 2 describes our framework for Gaussian process models indexed by distributions as well as a reminder for tools needed for our formulation. Section 3 extends the Matérn covariance function to this context, and provides the microergodicity result. Section 4 presents experimental results. In particular, we compare the classification accuracy using different simulations and real medical data. Conclusions are proposed in section 4 while all proofs are postponed to the Appendix.

2 Proposed Method

In this section, we first formulate the problem and then introduce the tools for the manifold Gaussian process regression model.

2.1 Problem Formulation

Let $I_1, \dots, I_n$ denote a finite set of n observed objects that are non-linear deformations of a specific pattern $I^\star$. By deformation, we mean that for all $i = 1, \dots, N$ there is a unique distribution $F_i \in \mathcal{F}$ such that

$$I_i = I^\star \circ F_i : \Omega = [0,1] \to \mathbb{R}^d.$$

In this study, we will focus on strictly nondecreasing distribution functions belonging to the space $\mathcal{F} = \{F : \Omega = [0,1] \mapsto [0,1], F(0) = 0, F(1) = 1, \quad f > 0\}$, letting $f = \dot{F}$, has been studied to solve statistical shape analysis problem with various applications in medical imaging, computer vision, and mechanics. The space $\mathcal{F}$ can be viewed as a Lie group without topological structures which acts onto the space of objects Is on the right as follows:

$$(I, F) = I \circ F$$

Thus, the notion of dissimilarity between any two objects $I_i = (I^\star \circ F_i)$ and $I_j = (I^\star \circ F_j)$ must be measured with respect to the deformation between these two objects, namely using a proper distance between the two distribution functions F_i and F_j. To reach such goal one needs to consider $\mathcal{F}$ as a Riemannian manifold by putting a Riemannian structure on it in order to define a geodesic distance for our study.

2.2 Background and Space of Representations

We endow $\mathcal{F}$ with the Fisher-Rao metric so that, for any $F \in \mathcal{F}$ and $T_F(\mathcal{F})$ being the tangent space to $\mathcal{F}$ at F we have:

$$< g_1, g_2 >_F = \int_0^1 \dot{g}_1(t)\dot{g}_2(t)\frac{1}{f(t)}dt.$$

for any $g_1, g_2 \in T_F(\mathcal{F})$. Note that $\mathcal{F}$ is now a nonlinear manifold due to boundary conditions and that this metric defines a Riemannian structure on it. As mentioned above, performing Kriging on $\mathcal{F}$ directly is not straightforward. In this work we will use a mapping from $\mathcal{F}$ to another Riemannian manifold and will exploit the isometry to extend the notion of Gaussian process to the space of strictly increasing functions. Indeed, we map each distribution F to the square root of its derivative, the corrresponding density function as follows:

$$\begin{array}{rcl} \Psi : \mathcal{F} & \to & \mathcal{H} \\ F & \mapsto & \phi = \Psi(F) = \sqrt{f}. \end{array}$$

Here $\mathcal{H} = \{\phi : \Omega = [0,1] \to \mathbb{R}, \int_0^1 \phi(t)^2 dt = 1, \phi > 0\}$. Note that $\phi = \sqrt{f}$ is well defined since $f > 0$ by definition and that $\Psi(id_{\mathcal{F}}) = 1$ is a constant function. Following the definition of the metric given above, we assume that the new space of ϕ, denoted by $\mathcal{H}$, is a subset of $\mathbb{L}^2[0,1]$. Then, since

$$\|\phi\|_2^2 = \int_0^1 \phi(t)\phi(t)dt = 1,$$

$\mathcal{H}$ is a subset of the unit Hilbert sphere with a Riemannian structure that will be useful later to ease the analysis of SNDFs. We remind that the geodesic distance on the sphere is given by the length of the connecting arc and that the parallel transport is a rotation. The reader can refer to [6, 7] for more details. Furthermore, Ψ is an isometry with the following inverse:

$$\begin{array}{rcl} \Psi^{-1} : \mathcal{H} & \to & \mathcal{F} \\ \phi & \mapsto & \left(t \to \int_0^t \phi^2(s)ds = \int_0^t f(s)ds \right). \end{array}$$

One of the main advantages of this formulation is to exploit the nice properties of the sphere:

- **Geodesic distance.** Let F_1, F_2 be any two elements $\in \mathcal{F}$ and let $\xi \in [0,1]$, then the geodesic between F_1 and F_2 at time instant ξ is given by the Ψ inverse of the geodesic arc between ϕ_1 and ϕ_2:

  $$\eta(\xi) = \frac{1}{\sin(\beta)} \left[\sin(\beta - \beta\xi)\phi_1 + \sin(\beta\xi)\phi_2 \right]$$

  where $\beta = \arccos\left(< \sqrt{f_1}, \sqrt{f_2} >_2\right)$.

- **The exponential map.** Let ϕ be any element in $\mathcal{H}$ and w its tangent vector $w \in T_\phi(\mathcal{H})$, then the exponential map exp is defined as an isometry from $\mathcal{H}$ to its tangent space $T_\phi(\mathcal{H})$ by:

  $$w \quad \mapsto \quad \exp_\phi(w) = \cos(\|w\|)\phi + \frac{\sin(\|w\|)}{\|w\|} w.$$

- **Log map.** As the inverse of the exponential map from $\phi_j \in \mathcal{H}$ to $T_\phi(\mathcal{H})$ is given by $\log_\phi$:

  $$\phi_j \quad \mapsto \quad w_j = \log_\phi(\phi_j) = \frac{\beta}{\sin(\beta)}(\phi_j - \cos(\beta)\phi).$$

Therefore and as a special case, we note $\mathcal{E} = T_1(\mathcal{H})$ the tangent space of $\mathcal{H}$ at the constant function one and $\mathcal{V}$ the space of functions v such that $v - 1$ belongs to $\mathcal{E}$:

$$\mathcal{V} = \{v \in \mathbb{L}(\Omega, \mathbb{R}) : \quad \int_0^1 v(t) = 1, \quad \|v\| \leq \frac{\pi}{2}\}.$$

3 Gaussian Processes on $\mathcal{F}$

Gaussian Processes (GP) are used to provide a probabilistic framework for a large variety of machine learning methods. We refer for instance to [21] and references therein. In this paper, they enable to optimally predict an unobserved value y associated to a deformed curve $I^\star \circ F$, from observed values $y_1, \dots, y_n$ corresponding to deformed curves $I^\star \circ F_1, \dots, I^\star \circ F_n$. Here, $F, F_1, \dots, F_n$ belong to $\mathcal{F}$, so that we focus on constructing Gaussian processes on $\mathcal{F}$.

A Gaussian process Z on $\mathcal{F}$ is a random field indexed by $\mathcal{F}$ so that $(Z(F_1), \dots, Z(F_n))$ is a Gaussian vector for any $n \in \mathbb{N}$, $F_1, \dots, F_n \in \mathcal{F}$. We point out that a Gaussian process Z is characterized by its mean function $m : \mathcal{F} \to \mathbb{R}$ and by its covariance function $C : \mathcal{F}^2 \to \mathbb{R}$. The fact that these two functions characterize the Gaussian process is a simplicity benefit, and is one of the reasons for the popularity of Gaussian processes.

In this paper, we consider Gaussian processes with zero mean function and focus on the issue of constructing a proper covariance function for distribution functions in $\mathcal{F}$.

3.1 Constructing Covariance Functions on $\mathcal{F}$

A covariance function C on $\mathcal{F}$ must satisfy the following conditions. For any $n \in \mathbb{N}$, $F_1, \dots, F_n \in \mathcal{F}$, the matrix $[C(F_i, F_j)]_{1 \leq i,j, \leq n}$ is symmetric non-negative definite. Furthermore, C is called non-degenerate when the above matrix is invertible whenever $F_1, \dots, F_n$ are two-by-two distincts [4].

The strategy we adopt to construct covariance functions is to exploit the isometric map $\log_1$, based on the tangent space of $\mathcal{H}$ at 1. That is, we construct covariance functions of the form

$$C(F_1, F_2) = K(|| \log_1(\phi_1) - \log_1(\phi_2)||), \tag{1}$$

where $||.||$ is the Euclidean norm in the Hilbert space $\mathcal{E}$ and with $K : \mathbb{R}^+ \to \mathbb{R}$.

We remark that it is common to define covariance functions C_d on $\mathbb{R}^d$ of the form $C_d(v_1, v_2) = K(||v_1 - v_2||)$ [21]. These covariance functions are called isotropic. In the next proposition, we show that, for any K so that C_d is a (non-degenerate) covariance function for any $d \in \mathbb{N}$, then C is also a (non-degenerate) covariance function. The next proposition has been partially addressed in [5]. For the sake of completeness, we give a complete statement and proof here.

Proposition 1. *Let $K : \mathbb{R}^+ \to \mathbb{R}$ be such that, for any $d \in \mathbb{N}$, the function K : $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ defined by $K(u, v) = K(||u - v||)$ is a covariance function. Let C be defined as in (1). Then C is a covariance function.*

Furthermore, assume that for any $d \in \mathbb{N}$ and for any pairwise different $u_1, \dots, u_n \in \mathbb{R}^d$, the matrix $(K(\|u_i - u_j\|)_{\{i,j\}}$ is invertible. Then C is non-degenerate.

In practice, we can select the function K from the Matérn family [25], letting for $t \geq 0$

$$K_\theta(t) = \frac{\sigma^2 (\alpha t)^\nu}{\mathcal{F}(\nu) 2^{\nu - 1}} K_\nu(\alpha t)$$

for $\theta = (\sigma^2, \alpha, \nu) \in (0, \infty)^3$. We obtain the Matérn covariance functions C_θ on $\mathcal{F}$ defined by $C_\theta(F_1, F_2) = K_\theta(|| \log_1(\phi_1) - \log_1(\phi_2)||)$ for $F_1, F_2 \in \mathcal{F}$. These functions C_θ are indeed covariance functions and are non-degenerate from Proposition 1, and from the fact that the Matérn covariance function is non-degenerate on $\mathbb{R}^d$ for any $d \in \mathbb{N}$.

3.2 Asymptotic Properties

Consider a parametric set of covariance functions $\{C_\theta; \theta \in \Theta\}$, with $\Theta \subset \mathbb{R}^p$ and where, for $\theta \in \Theta$ and $F_1, F_2 \in \mathcal{F}$, $C_\theta(F_1, F_2) = K_\theta(|| \log_1(\phi_1) - \log_1(\phi_2)||)$ with $K_\theta : \mathbb{R}^+ \to \mathbb{R}$.

We can see a Gaussian process Z as an application from $(\Omega, \mathcal{M}) \times \mathcal{F}$ to $\mathbb{R}$, with $(\Omega, \mathcal{M})$ a measurable space. For any $\theta \in \Theta$, we can consider a probability measure $\mathbb{P}_\theta$ on Ω so that $Z : (\Omega, \mathcal{M}, \mathbb{P}_\theta) \times \mathcal{F} \to \mathbb{R}$ is a Gaussian process with mean function zero and covariance function C_θ.

Following [25], we say that the covariance parameter θ is microergodic if, for any $\theta_1 \neq \theta_2$ so that $\theta_1, \theta_2 \in \Theta$, the measures $\mathbb{P}_{\theta_1}$ and $\mathbb{P}_{\theta_2}$ are orthogonal.

For Gaussian processes indexed by a fixed bounded subset of $\mathbb{R}^d$, for $d \in \mathbb{N}$, microergodicity is an important concept. Indeed, it is a necessary condition for consistent estimators of θ to exist under fixed-domain asymptotics [25], and a fair amount of work has been devoted to showing microergodicity or non-microergodicity of parameters, for various models of covariance functions [25, 27, 1]. In this section, we extend these types of results to Gaussian processes indexed on $\mathcal{F}$. In the next theorem, we show that the covariance parameter θ is microergodic under very mild conditions.

Theorem 1. *Assume that there does not exist $\theta_1, \theta_2 \in \Theta$, with $\theta_1 \neq \theta_2$, so that $t \to K_{\theta_1}(t) - K_{\theta_2}(t)$ is constant on $[0, \pi/2]$. Then the covariance parameter θ is microergodic.*

In particular, Theorem 1 applies to the Matérn family of covariance functions described above.

3.3 Covariance Parameter Estimation and Prediction

Consider a data set of labeled objects of the form $(I^\star \circ F_1, y_1), \dots, (I^\star \circ F_n, y_n)$, with $F_1, \dots, F_n \in \mathcal{F}$ and $y_1, \dots, y_n \in \mathbb{R}$. We adopt the point of view of Gaussian processes and assume that, for $i = 1, \dots, n$, $y_i = Z(F_i) + \epsilon_i$, where Z is a Gaussian process on $\mathcal{F}$ and with $(\epsilon_1, \dots, \epsilon_n)^t \sim \mathcal{N}(0, \rho I_n)$, independently of ϵ. Here ρ is the observation noise variance, that we assume to be known for simplicity.

Assume that Z has mean function zero and covariance function in the set $\{C_\theta; \theta \in \Theta\}$, for $\Theta \subset \mathbb{R}^p$. Then, θ can be selected by the maximum likelihood, with

$$\hat{\theta} \in \underset{\theta \in \Theta}{\operatorname{argmin}} \log \det(R_\theta + \rho I_n) + y^t (R_\theta + \rho I_n)^{-1} y \tag{2}$$

where $R_\theta = [C_\theta(F_i, F_j)]_{1 \leq i,j \leq n}$ and with $y = (y_1, \dots, y_n)^t$. We remark that alternative estimation techniques exist, for instance cross validation [2, 3, 28]. Then, for any

Algorithm 1 Learning RGP

```
Input: $(I_i, y_i)$ for $i = 1 \dots n$
Output: $\hat{\theta}$, $C_{\hat{\theta}}$,$y_{\hat{\theta}}(F)$
Compute $F_i$ for every $I_i = I^* \circ F_i$ with $i = 1 \dots n$
Compute $\phi_i = \Psi(F_i)$ with $i = 1 \dots n$
Define the tangent space $T_1(\mathcal{H})$
Compute the exponential map $\log_1$ and its inverse
Compute the covariance function $K_\theta$ in eq 1
Find $\hat{\theta}$ that maximizes the likelihood in eq 2
Compute $C_{\hat{\theta}}$ and $y_{\hat{\theta}}(F)$
```

new object of the form $I^\star \circ F$ with $F \in \mathcal{F}$, the corresponding label can be predicted by $\hat{y}_{\hat{\theta}}(F)$, with

$$\hat{y}_\theta(F) = r_\theta(F)^t (R_\theta + \rho I_n)^{-1} y,$$

where $r_\theta(F) = (C_\theta(F, F_1), \dots, C_\theta(F, F_n))^t$. The prediction $\hat{y}_\theta(F)$ is the conditional expectation of $Z(\mathcal{F})$ given $y_1, \dots, y_n$, when Z has covariance function C_θ.

The above formulas for maximum likelihood and prediction can be found in [21] for instance. The simplicity of the prediction formula explains the popularity of Gaussian process models. Overall steps of the proposed method are summarized in Algorithm 1.

4 Numerical Results

We demonstrate the proposed framework for learning a Gaussian process from a finite set of domain deformations. We first represent these deformations by strictly non-decreasing distributions and then consider them as element of a Riemannian manifold using properties detailed in section 2.1. Next, we illustrate the performance of the learned model in terms of classification accuracy using synthetic data and two different real datasets: Berkeley growth study and a medical dataset from a population with arthritis.

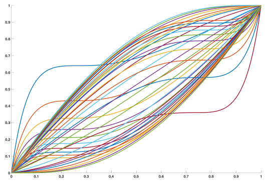
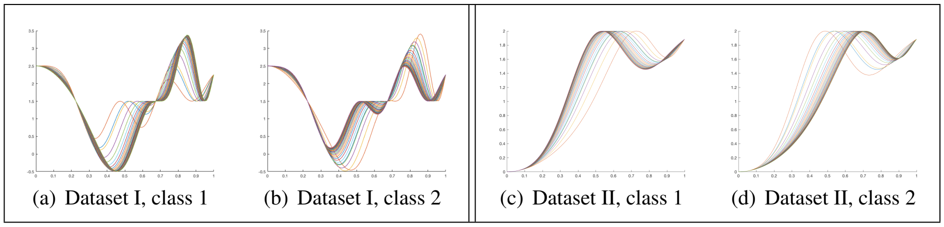


Figure 1: Examples of F_i from $[0, 1]$ to $[0, 1]$. The identity is given by the diagonal, an increase appears above the diagonal whereas a decrease appears below the diagonal.

Synthetic datasets. As a sanity test, we simulate two datasets. Some samples are displayed in Figure 2: class 1 (a) and class 2 (b) from example I and (c & d) these two



(a) Dataset I, class 1 (b) Dataset I, class 2 (c) Dataset II, class 1 (d) Dataset II, class 2

Figure 2: Synthetic data: two different synthetic datasets (a &b and c &d), each with two classes.

classes from example II. Each class contains 300 samples for both examples and have been randomly generated from a finite basis of $\mathcal{F}$, see Figure 1 for few examples of simulated F_i. As a common pattern I^* defined for $t \in [0, 1]$, we used:

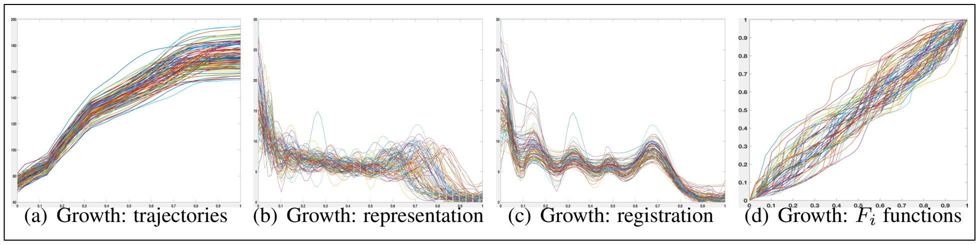
$$\begin{cases} I_1^*(t) = \frac{1}{2} + \sin(4\pi t)\cos(7t) \\ I_2^*(t) = 2(1 + \cos(8t)))\exp(-2t^2) \end{cases}$$

to provide flexible enough deformations.

Real datasets. We use two different datasets. First, we use Berkeley growth study that records the heights of children at 31 stages from 1 to 18 years (see [20]). It is a typical example of biological dynamics observed over a period of time. The dataset has been widely used as a motivating example to analyze functional data. In our context, all growth curves were represented by their first derivative functions. The common pattern I^* was given as the Fréchet mean of all derivatives. See Figure 3 top row for examples. The second dataset consists of hand force signals from a population of 80 healthy and 100 patients subject with arthritis. The medical protocol saved the hand force during a continuous period of time where the goal is to study the endurance during test. Thus, members of the healthy group are expected to hold more (less variability in F_i) than patients with pain. See Figure 4 bottom row for examples of F_i.

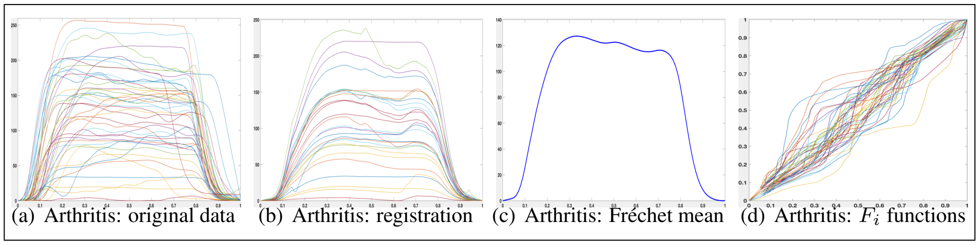
- **Hyper-parameter Tuning.** The parameters that require tuning are $\theta = (\alpha, \nu, \sigma)$. We used the gradient descent and a Newton-based optimization to search for optimal values of (α, σ) and a cross-validation on ν to find the maximum likelihood as defined in equation (2).
- **Evaluation Method.** The classification accuracy is evaluated using the mean squared prediction error (MSE).
- **Performance Comparison Results.** The accuracy rate is given for Gaussian Process on the space of (I_i) and using the proposed method on the corresponding (F_i).

First, we test the efficiency of the proposed method: We learn the model parameters from 75% of the dataset as training and use the rest for test. To compute the classification error, we first compute the estimator and its parameters from the training set. Then, given a new observation F^* for test, we apply the regression model to determine



(a) Growth: trajectories (b) Growth: representation (c) Growth: registration (d) Growth: F_i functions

Figure 3: Real data: analyzing trajectories from Berkeley growth study. The goal is to characterize the growth rate for boys and girls.



(a) Arthritis: original data (b) Arthritis: registration (c) Arthritis: Fréchet mean (d) Arthritis: F_i functions

Figure 4: Real data: functional signals from adults with and without arthritis. The goal is to learn a regression model for aided diagnostics.

its class logit(y^*), where y^* is the Gaussian process prediction. The subdivision has been performed randomly 20 times and the classification rates are given as a mean. As mentioned above, we evaluate the classification quality using MSE. All scores are summarized in Table 1 for simulated data and 2 and 3 for real data. We note that the regression using a Gaussian process on distribution functions outperforms the Gaussian process on original functions. For a fair comparison, we used the same properties of the Gaussian process such as the Matérn covariance and both gradient and Newton for parameters tuning.

Table 1: Mean squared error with two different hyper-parameters tuning methods on simulated data

Method	Gradient	Newton
Simulation I (F_i)	$3.47e-9$	$3.47e-9$
Simulation II (F_i)	$3.23e-2$	$2.01e-2$

Table 2: Mean squared error with two different hyper-parameters tuning methods on Berkley growth data

Method	Gradient	Newton
Growth (I_i)	$2.23e-1$	$1.31e-1$
Growth (F_i)	$1.10e-1$	$7.15e-2$

Table 3: Mean squared error with two different hyper-parameters tuning methods on arthritis data

Method	Gradient	Newton
Arthritis (I_i)	$1.14e-1$	$1.01e-1$
Arthritis (F_i)	$7.97e-2$	$5.29e-2$

5 Conclusion

This paper presents a novel framework for learning Gaussian process model on strictly nondecreasing distributions. With a Matérn kernel specified as the covariance function in the Gaussian process prior, we have provided the microergodicity of the covariance parameters. The proposed method was successfully tested on both synthetic and real medical data. We showed that the regression model is capable of producing highly meaningful differences on different classes of objects when using their domains' deformations only. A future direction of interest is to build theoretical extension for more complex domains where the distributions could be bivariate or even multivariate for new aspects of manifold learning.

A Proofs

Proof of Proposition 1. Let $F_1, \dots, F_n$ in $\mathcal{F}$. For $i = 1, \dots, n$, let $g_i = \log_1(\phi_i)$. Consider the matrix $\tilde{C} = (< g_i, g_j >)_{\{i,j\}}$. This matrix is a Grammian matrix in $\mathbb{R}^{n\times n}$ hence there exists a non negative diagonal matrix D and an orthogonal matrix P such that

$$\tilde{C} = PDP^{'} = PD^{1/2}D^{1/2}P^{'}.$$

Let $e_1, \dots, e_n$ be the canonical basis of $\mathbb{R}^n$. Then $e_i^t\tilde{C}e_j = u_i^t u_j$ where $u_i^t = e_i^t PD^{1/2}$. Note that the u_i's are vectors in $\mathbb{R}^n$ that depend on the $f_1, \dots, f_n$. We get that

$$< g_i, g_j >= u_i^t u_j,$$

and for any $F_1, \dots, F_n$ in $\mathcal{F}$ there are $u_1, \dots, u_n$ in $\mathbb{R}^n$ such that

$$\| \log_1(\phi_i) - \log_1(\phi_j)\| = \|u_i - u_j\|.$$

So any covariance matrix that can be written as $[K(\| \log_1(\phi_i) - \log_1(\phi_j)\|)]_{i,j}$ can be seen as a covariance matrix $[K(\|u_i - u_j\|)]_{i,j}$ on $\mathbb{R}^n$ and inherits its properties. The invertibility and non-negativity of this covariance matrix entail the invertibility and non-negativity of the first one, which proves the result. □

Proof of Theorem 1. Let $\theta_1, \theta_2 \in \Theta$, with $\theta_1 \neq \theta_2$. Then, there exists $t^* \in [0, \pi/4]$ so that $K_{\theta_1}(0) - K_{\theta_1}(2t^*) \neq K_{\theta_2}(0) - K_{\theta_2}(2t^*)$.

For $i \in \mathbb{N}$, let $c_i : [0,1] \to \mathbb{R}$ be defined by $c_i(t) = t^* \cos(2\pi i t)$. Then, $c_i \in T_1(\mathcal{H})$. Let $\tilde{e}_i = \exp_1(c_i)$. Then, for $t \in [0,1]$

$$\tilde{e}_i(t) = \cos(t^*) + \frac{\sin(t^*)}{t^*} t^* \cos(2\pi i t) \geq \cos(t^*) - \sin(t^*) \geq 0.$$

It follows that $\tilde{e}_i \in \mathcal{H}$ and we can let $\tilde{F}_i(t) = \int_0^t \tilde{e}_i(s)^2 ds$. Letting $\bar{e}_i = \exp_1(-c_i)$, we obtain similarly that $\bar{e}_i \in \mathcal{H}$ and we let $\bar{F}_i(t) = \int_0^t \bar{e}_i(s)^2 ds$

Consider the $2n$ elements $(F_1, ..., F_{2n})$ composed by the pairs $(\tilde{F}_i, \bar{F}_i)$ for $i = 1, \dots, n$. Consider a Gaussian process Z on $\mathcal{F}$ with mean function zero and covariance function K_{θ_1}. Then, the Gaussian vector $W = (Z(F_i))_{i=1,\dots,2n}$ has covariance matrix C given by

$$C_{i,j} = \begin{cases} K_{\theta_1}(0) & \text{if } i = j \\ K_{\theta_1}(2t^*) & \text{if } i \text{ odd and } j = i+1 \\ K_{\theta_1}(2t^*) & \text{if } i \text{ even and } j = i-1 \\ K_{\theta_1}(\sqrt{2}t^*) & \text{else.} \end{cases}$$

Hence, we have $C = D + M$ where M is the matrix with all components equal to $K_{\theta_1}(\sqrt{2}t^*)$ and where D is block diagonal, composed of n blocks of size 2×2, with each block $B_{2,2}$ equal to

$$\begin{pmatrix} K_{\theta_1}(0) - K_{\theta_1}(\sqrt{2}t^*) & K_{\theta_1}(2t^*) - K_{\theta_1}(\sqrt{2}t^*) \\ K_{\theta_1}(2t^*) - K_{\theta_1}(\sqrt{2}t^*) & K_{\theta_1}(0) - K_{\theta_1}(\sqrt{2}t^*) \end{pmatrix}.$$

Hence, in distribution, $W = M + E$, with M and E independent, $M = (z, ..., z)$ where $z \sim \mathcal{N}(0, K_{\theta_1}(\sqrt{2}t^*))$ and where the n pairs (E_{2k+1}, E_{2k+2}), $k = 0, ..., n-1$ are independent, with distribution $\mathcal{N}(0, B_{2,2})$. Hence, with $\bar{W}_1 = (1/n) \sum_{k=0}^{n-1} W_{2k+1}$, $\bar{W}_2 = (1/n) \sum_{k=0}^{n-1} W_{2k+2}$ and $\bar{E} = (1/n) \sum_{k=0}^{n-1} (E_{2k+1}, E_{2k+2})^t$, we have

$$\begin{aligned} \hat{B} &:= \frac{1}{n} \sum_{i=0}^{n-1} \begin{pmatrix} W_{2i+1} - \bar{W}_1 \\ W_{2i+2} - \bar{W}_2 \end{pmatrix} \begin{pmatrix} W_{2i+1} - \bar{W}_1 \\ W_{2i+2} - \bar{W}_2 \end{pmatrix}^t \\ &= \frac{1}{n} \sum_{i=0}^{n-1} \begin{pmatrix} E_{2i+1} \\ E_{2i+2} \end{pmatrix} \begin{pmatrix} E_{2i+1} \\ E_{2i+2} \end{pmatrix}^t - \bar{E}\bar{E}^t \\ &\to^p_{n \to \infty} B_{2,2}. \end{aligned}$$

Hence, there exists a subsequence $n' \to \infty$ so that, almost surely $\hat{B} \to B_{2,2}$ as $n' \to \infty$. Hence, almost surely $\hat{B}_{1,1} - \hat{B}_{1,2} \to K_{\theta_1}(0) - K_{\theta_1}(2t^*)$ as $n' \to \infty$. Hence, the event $\{\hat{B}_{2,2} \to_{n' \to \infty} K_{\theta_1}(0) - K_{\theta_1}(2t^*)\}$ has probability one under $\mathbb{P}_{\theta_1}$. With the same arguments, we can show that the event $\{\hat{B}_{2,2} \to_{n'' \to \infty} K_{\theta_2}(0) - K_{\theta_2}(2t^*)\}$ has probability one under $\mathbb{P}_{\theta_2}$, where n'' is a subsequence extracted from n'. Since these two events have zero intersection, it follows that $\mathbb{P}_{\theta_1}$ and $\mathbb{P}_{\theta_2}$ are orthogonal. Hence, θ is microergodic. □

References

[1] Anderes, E.: On the consistent separation of scale and variance for Gaussian random fields. The Annals of Statistics **38**, 870–893 (2010)

[2] Bachoc, F.: Cross validation and maximum likelihood estimations of hyper-parameters of gaussian processes with model misspecification. Computational Statistics & Data Analysis **66**, 55–69 (2013)

[3] Bachoc, F.: Asymptotic analysis of covariance parameter estimation for Gaussian processes in the misspecified case. Bernoulli **24**, 1531–1575 (2018)

[4] Bachoc, F., Gamboa, F., Loubes, J.M., Venet, N.: A gaussian process regression model for distribution inputs. IEEE Transactions on Information Theory (2017)

[5] Bachoc, F., Suvorikova, A., Loubes, J.M., Spokoiny, V.: Gaussian process forecast with multidimensional distributional entries. arXiv preprint arXiv:1805.00753 (2018)

[6] Boothby, W.M.: An Introduction to Differential Manifolds and Riemannian Geometry. Academic Press, New york (1975)

[7] Dryden, L., Mardia, K.V.: Statistical Shape Analysis. John Wiley & Son (1998)

[8] Efrat, A., Fan, Q., Venkatasubramanian, S.: Curve matching, time warping, and light fields: New algorithms for computing similarity between curves. Journal of Mathematical Imaging and Vision **27**(3), 203–216 (April 2007)

[9] Gamboa, F., Loubes, J.M., Maza, E.: Semi-parametric estimation of shifts. Electronic Journal of Statistics **1**, 616–640 (2007)

[10] Gervini, D., Gasser, T.: Self-modeling warping functions. Journal of the Royal Statistical Society, Series B **66**, 959–971 (2004)

[11] Grenander, U., Miller, M., Klassen, E., Le, H., Srivastava, A.: Computational anatomy: an emerging discipline. Quarterly of applied Mathematics **4**, 617–694 (1998)

[12] James, G.: Curve alignment by moments. Annals of Applied Statistics pp. 480–501 (2007)

[13] Kendall, D.G.: Shape manifolds, procrustean metrics and complex projective spaces. Bulletin of London Mathematical Society **16**, 81–121 (1984)

[14] Kneip, A., Gasser, T.: Statistical tools to analyze data representing a sample of curves. The Annals of Statistics **20**, 1266–1305 (1992)

[15] Kolmogorov, A.N.: Wienersche spiralen und einige andere interessante kurven im hilbertschen raum. Doklady Akad. Nauk SSSR **26**, 115–118 (1940)

[16] Kurtek, S., Srivastava, A., Wu, W.: Signal estimation under random time-warpings and nonlinear signal alignment. In: Neural Information Processing Systems (NIPS) (2011)

[17] Liu, X., Müller, H.G.: Functional convex averaging and synchronization for time-warped random curves. Journal of the American Statistical Association **99**, 687–699 (2004)

[18] Michor, P.W., Mumford, D.: Riemannian geometries on spaces of plane curves. Journal of the European Mathematical Society **8**, 1–48 (2006)

[19] Ramsay, J.O., Li, X.: Curve registration. Journal of the Royal Statistical Society, Series B **60**, 351–363 (1998)

[20] Ramsay, J.O., Silverman, B.W.: Functional Data Analysis, Second Edition. Springer Series in Statistics (2005)

[21] Rasmussen, C., Williams, C.: Gaussian Processes for Machine Learning. The MIT Press, Cambridge (2006)

[22] Sakoe, H.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing **26**, 43–49 (1978)

[23] Srivastava, A., Wu, W., Kurtek, S., Klassen, E., Marron, J.S.: Registration of functional data using fisher-rao metric. arXiv: 1103.3817v2 (2011)

[24] Srivastava, A., Klassen, E.: Functional and Shape Data Analysis. Springer, New York, NY (2016)

[25] Stein, M.L.: Interpolation of Spatial Data. Springer Series in Statistics, Springer-Verlag New York (1999)

[26] Tang, R., Müller, H.G.: Pairwise curve synchronization for functional data. Biometrika **95**(4), 875–889 (2008)

[27] Zhang, H.: Inconsistent estimation and asymptotically equivalent interpolations in model-based geostatistics. Journal of the American Statistical Association **99**, 250–261 (2004)

[28] Zhang, H., Wang, Y.: Kriging and cross-validation for massive spatial data. Environmetrics: The official journal of the International Environmetrics Society **21**(3-4), 290–304 (2010)