Bayesian Regression and Classification Using Gaussian Process Priors Indexed by Probability Density Functions

Anis Fradi, Y. Feunteun, Chafik Samir, M. Bakloul, François Bachoc, Jean-Michel Loubes

Version of Record: https://www.sciencedirect.com/science/article/pii/S0020025520309373
Manuscript_567867544b8e1f3d75d4d8e1f0661fd0

Bayesian Regression and Classification Using Gaussian Process Priors Indexed by Probability Density Functions

A. Fradi[a], Y. Feunteun[a], C. Samir[a,b,*], M. Baklouti[a], F. Bachoc[b], J-M Loubes[b]

[a] *CNRS-LIMOS, UCA, France*

[b] *Institut de Mathématiques de Toulouse, France*

Abstract

In this paper, we introduce the notion of Gaussian processes indexed by probability density functions for extending the Matérn family of covariance functions. We use some tools from information geometry to improve the efficiency and the computational aspects of the Bayesian learning model. We particularly show how a Bayesian inference with a Gaussian process prior (covariance parameters estimation and prediction) can be put into action on the space of probability density functions. Our framework has the capacity of classifiying and infering on data observations that lie on nonlinear subspaces. Extensive experiments on multiple synthetic, semi-synthetic and real data demonstrate the effectiveness and the efficiency of the proposed methods in comparison with current state-of-the-art methods.

Keywords: **Information geometry, Learning on nonlinear manifolds, Bayesian regression and classification, Gaussian process prior, HMC sampling**

1. Introduction

In recent years, Gaussian processes on manifolds have become very popular in various fields including machine learning, data mining, medical imaging, computer vision, etc. The main purpose consists in inferring the unknown target value at an observed location on the manifold as a prediction by conditioning on known inputs and targets. The random field, usually Gaussian, and the forecast can be seen as the posterior mean, leading to an optimal unbiased predictor [1, 2].

*Corresponding author
Email address: chafik.samir@uca.fr (C. Samir)

Bayesian regression and classification models focus on the use of priors for the parameters to define and estimate a conditional predictive expectation. In this work, we consider a very common problem in several contexts of applications in science and technology: learning a Bayesian regression and classification models with Probability Density Functions as inputs.

Probability Density Functions (PDFs) are inherently infinite-dimensional objects. Hence, it is not straightforward to extend traditional machine learning methods from finite vectors to functions. For example, in functional data analysis [3] with applications in medical [4, 5], it is very common to compare/classify functions. The mathematical formulation leads to a wide range of applications where it is crucial to characterize a population or to build predictive models. In particular, multiple frameworks exist for comparing PDFs in different representations including Frobenius, Fisher-Rao, log-Euclidean, Jensen-Shannon and Wasserstein distances [6, 7, 4, 8, 1]. In this work, we extend this formulation to PDFs space $\mathcal{P}$ with the Matérn covariance functions.

There is a rich literature on statistical inference on manifolds among which the Fisher information matrix [9] has played a central role. Recently, there has been increasing interest in applying information geometry for machine learning and data mining tasks [10, 11, 12, 13]. The Fisher information matrix determines a Riemannian structure on a parametrized space of probability measures. Study of geometry of $\mathcal{P}$ with the Riemannian structure, which we call information geometry, contributes greatly to statistical inference, refer to [11, 12] for more details. Such methods are based on parametric models that are of great interest in many applications. However, aspects of PDFs other than parametric families may be important in various contexts [8, 14, 6, 15, 4, 16]. In particular, the consistency of regression and classification with PDFs inputs was established in [17, 18, 19] with the help of kernel density estimation [20]. More recently, [1] studied the dissimilarity between PDFs with the Wasserstein distance and [21] used a nonparametric framework to compare spherical populations.

The main aim of this paper is to learn a Bayesian inference on Gaussian processes. For instance, one can think of a Gaussian process as defining PDFs and inference taking place directly in the function-space. Moreover, the index space is that of PDFs when choosing the underlying metric in order to evaluate the dissimilarity between them [22]. The only drawback is that performing Kriging on PDFs space $\mathcal{P}$ is not straightforward due to its geometry. For this end, we exploit an isometric embedding by combinng the square root transform [23] and the distance induced by the Fisher-Rao

metric which make the covariance function non-degenerate and simplify the optimization process.

Gaussian processes (GPs) have been widely used to provide a probabilistic framework for a large variety of machine learning methods [24]. Optimization techniques are usually required to fit a GP model Z, that is to select a GP covariance function. For p_i and p_j in $\mathcal{P}$, the main issue would be to build a proper covariance between $Z(p_i)$ and $Z(p_j)$. In particular, this covariance can define a notion of stationarity for the process. Another important task is the classification process where we wish to assign an input PDF p_i to one of the given classes [25].

To search for the covariance function hyperparameter, we use several methods for maximizing the marginal likelihood. Our aim is then to select those optimizing performance criteria for regression and classification: The first method is based on the gradient descent for finding a local maximum of the marginal likelihood. The second method is a special case of MCMC methods, called Hamiltonian Monte-Carlo (HMC) [26]. The objective is to perform sampling from a probability distribution for which the marginal likelihood and its gradient are known. The latter has the advantage to simulate from a physical system governed by Hamiltonian dynamics.

The remainder of the paper is organized as follows. In Section 2, we introduce the problem formulation and we give a background of some Riemannian representations. Section 3 extends the usual notion of GPs indexed by finite vectors to those indexed by PDFs with theoretical results for the Matérn covariance function. We also give details of the proposed model for predicting and classifying PDFs as well as estimating the covariance function parameters. In Section 4, we present and discuss experimental results with some comparison studies. We conclude the paper in Section 5.

2. Problem formulation and geometry background

Let $p_1, \ldots, p_n$ denote a finite set of observed PDFs and $y_1, \ldots, y_n$ denote their corresponding outputs with real values (quantitative or qualitative). In this work, we focus on nonparametric PDFs restricted to be defined on $\Omega = [0, 1]$. Our main goals throughout this paper are: i) Fitting the proposed model's parameters in order to better explain the link between p_i and y_i, $i = 1, \ldots, n$, ii) evaluating the corresponding predictive expectation at an unobserved PDF $p^* \in \mathcal{P}$ and iii) studying the properties of the GP with the Matérn covariance function. In the particular case where $y_i \in \{-1, +1\}$, we will assign each unobserved PDF p^* to its predicted class after learning the model

parameters. To reach such goal, we follow the same idea of nonparametric information geometry that has been discovered by [9] and developed later in other works, see for example [27, 6, 7, 11, 28]. Thus, the notion of similarity/dissimilarity between any p_i and p_j is measured using the induced Rao distance [29, 22] between them on the underlying space. In this paper, we look at the space of PDFs as a Riemannian manifold, as detailed in the next section, which plays an important role in the proposed methods.

2.1. Riemannian structure of PDFs space

For more details about the geometric structure concerning the Fisher information metric, refer to [27, 15, 30]. For example, [27] showed that $\mathcal{P}$ with a Riemannian structure has a positive constant curvature. Furthermore, the action of orientation preserving diffeomorphism acts by isometry on $\mathcal{P}$ with respect to the Fisher information metric. We will exploit these nice properties to define an isometric embedding from $\mathcal{P}$ to $\mathcal{E}$ detailed in (11). Then, we use this embedding to construct a Gaussian process on PDFs with a proper covariance function (14) and a predictive model (31).

We first note that the space of PDFs defined over Ω with values in $\mathbb{R}_+$ can be viewed in different manners. The case where Ω is finite and the statistical model is parametric has been largely studied in the literature [12, 11]. In contrast, if Ω is infinite which is the case here, different modeling options have been explored [27, 31, 3, 30]. We start with the ideas developed in [27, 15, 6, 11] where $\mathcal{P}$ is an infinite dimensional smooth manifold. That is, $\mathcal{P}$ is the space of probability measures that satisfy the normalization constraint. Since we are interested in statistical PDFs analysis on $\mathcal{P}$, we need some geometrical tools [32, 33],*e.g.* geodesic. For the rest of the paper, we view $\mathcal{P}$ as a smooth manifold (1) and we impose a Riemannian structure on it with the Fisher-Rao metric (3). Let

$$\mathcal{P} = \{p : \Omega \to \mathbb{R} \,|\, p \geq 0 \text{ and } \int_\Omega p(t)dt = 1\}. \tag{1}$$

be the space of all PDFs (positive almost everywhere) including nonparametric models. We identify any tangent space of $\mathcal{P}$, locally at each p, by

$$T_p(\mathcal{P}) = \{f : \Omega \to \mathbb{R} \,|\, \int_\Omega f = 0\} \tag{2}$$

As detailed in [34, 27, 30], the tangent space contains functions that are infinitesimally differentiable. But following [32], we have a constructive method of great importance that allows one to form a local version of any arbitrary f that is continuously differentiable in a small neighborhood and null outside. Now that we have a smooth manifold and its tangent space, we can introduce a Riemannian metric. This choice is very important since it will determine the structure of $\mathcal{P}$ and consequently the covariance function of the Gaussian process. More details about the importance of the metric and the induced Riemannian structure are discussed in [7, 35, 36]. We also define and denote by $\mathcal{P}_+$ the interior of $\mathcal{P}$. For the following, we consider without justification that any probability density can be locally perturbed to be smooth enough [32]. This is true in finite dimensional cases but the generalization to infinite dimensional cases is not straightforward. Among several metrics, we are particularly interested in the Fisher-Rao metric defined, for any tangent vectors $f_1, f_2 \in T_p(\mathcal{P})$, by

$$< f_1, f_2 >_p = \int_\Omega \frac{f_1(t) f_2(t)}{p(t)} dt. \tag{3}$$

Although this metric has nice properties with an increasing interest [15, 10, 34], $\mathcal{P}$ equipped with $< ., . >_p$ is still numerically intractable. Therefore, instead of working on $\mathcal{P}$ directly, we consider a mapping from $\mathcal{P}$ to the Hilbert upper-hemisphere (positive part) around the unity $1_{\mathcal{P}}$ such that $1_{\mathcal{P}}(t) = 1$ for all t in Ω [23]. Thus, we exploit the Riemannian isometry between $\mathcal{P}$ and the upper-hemisphere to extend the notion of GPs to the space of PDFs. Indeed, we first define the map

$$\begin{aligned} \Psi : \mathcal{P} &\rightarrow \mathcal{H} \\ p &\mapsto \phi = 2\sqrt{p}, \quad (\mathtt{p} = \Psi^{-1}(\phi) = \frac{1}{4}\phi^2) \end{aligned} \tag{4}$$

where

$$\mathcal{H} = \{\phi : \Omega \rightarrow \mathbb{R} \,|\, \phi \geq 0 \text{ and } \int_\Omega \phi(t)^2 dt = 4\}. \tag{5}$$

Note that ϕ is well defined since p is nonnegative and Ψ is a Riemannian isometry from $\mathcal{P}_+$ to $\mathcal{H}$ without the boundary [33]. On the other hand, any element $\phi \in \mathcal{H}$ can be uniquely projected as $\frac{1}{2}\phi$ to have a unit norm. For simplicity and without loss of generality, we interpret $\mathcal{H}$ as the elements of unit Hilbert upper-hemisphere $\mathcal{S}_+^\infty$ up to

a multiplicative factor (2 here). From that point of view, we identify $\mathcal{H}$ with $\mathcal{S}_+^\infty$ and we define $\Psi(1_\mathcal{P}) = 1_\mathcal{H}$ to be the "unity pole" on $\mathcal{H}$. Note that $1_\mathcal{H}$ as the image of the uniform pdf $1_\mathcal{P}$ is a fixed point, *i.e.* $1_\mathcal{H} = \sqrt{1_\mathcal{P}} = 1_\mathcal{P}$. In this setup, we have

$$\|\phi\|_2^2 = \int_\Omega \phi(t)^2 dt = 1, \tag{6}$$

for any ϕ in $\mathcal{H}$ which allow us to consider $\mathcal{H}$, when equipped with the integral inner product $< .,. >_2$, as the unit upper-hemisphere (positive part). Furthermore, for arbitrary directions f_1, f_2 in $T_p(\mathcal{P})$ the Fisher-Rao metric as defined in (eq:Fisher-Rao) becomes $< .,. >_2$ as follows:

$$< f_1, f_2 >_p = < D_{f_1}\Psi, D_{f_1}\Psi >_2 . \tag{7}$$

with $D_{f_i}\Psi(p)(t) = \frac{f_i(t)}{\sqrt{p(t)}}$ for all $t \in \Omega$ and $i = 1, 2$. One of the main advantages of this formulation is to exploit the nice properties of the unit Hilbert sphere such as geodesic, exponential map, log map, and the parallel transport. For the rest of the paper, the geodesic distance $d_\mathcal{P}(p_1, p_2)$ between two PDFs p_1 and p_2 in $\mathcal{P}$, under the Fisher-Rao metric, is given by the geodesic distance $d_\mathcal{H}(\phi_1, \phi_2)$ (up to a factor 2) between their corresponding ϕ_1 and ϕ_2 on $\mathcal{H}$. We remind that the arc-length (geodesic distance) between distinct and non antipodal ϕ_1 and ϕ_2 on $\mathcal{H}$ is the angle $\beta = \arccos(< \phi_1, \phi_2 >_2)$. We also remind some geometric tools that will be needed for next sections as a lemma:

Lemma 2.1. *With $\mathcal{H}$ defined from (5) with unit norm and $T_\phi(\mathcal{H})$ its tangent space at ϕ, we have the following:*

- *The exponential map is a bijective isometry from the tangent space $T_\phi(\mathcal{H})$ to $\mathcal{H}$. For any $w \in T_\phi(\mathcal{H})$, we write*

$$\mathrm{Exp}_\phi(w) = \cos(\|w\|_2)\phi + \sin(\|w\|_2)\frac{w}{\|w\|_2}. \tag{8}$$

- *Its inverse, the log map is defined from $\mathcal{H}$ to $T_{\phi_1}(\mathcal{H})$ as*

$$\mathrm{Log}_{\phi_1}(\phi_2) = \frac{\beta}{\sin(\beta)}(\phi_2 - \cos(\beta)\phi_1). \tag{9}$$

- *For any two elements ϕ_1 and ϕ_2 on $\mathcal{H}$ the map $\Gamma : T_{\phi_1}(\mathcal{H}) \to T_{\phi_2}(\mathcal{H})$ parallel transports a vector w from ϕ_1 to ϕ_2 and is given by:*

$$\Gamma_{\phi_1 \rightarrowtail \phi_2}(w) = w - 2\frac{(\phi_1 + \phi_2)}{||\phi_1 + \phi_2||_2^2} < w, \phi_2 >_2 \tag{10}$$

For more details, we refer to [33]. As a special case, we consider the unity pole $\phi = 1_{\mathcal{H}}$ and we denote $\mathcal{E} = T_1(\mathcal{H})$ the tangent space of $\mathcal{H}$ at $1_{\mathcal{H}}$. For simplicity, we note $\mathrm{Log}_1(.)$ the log map from $\mathcal{H}$ to $\mathcal{E}$ and $\mathrm{Exp}_1(.)$ its inverse. This choice is motivated by two reasons: The numerical implementation and the fact that $1_{\mathcal{H}}$ is the center of the geodesic disc $[0, \frac{\pi}{2}[$. Indeed and since all elements are on the positive part, the exponential map and its inverse are diffeomorphisms. So, one can consider any point on $\mathcal{H}$ instead of $1_{\mathcal{H}}$ to define the tangent space, *e.g.* the Fréchet mean. However, this is without loss for the numerical precision. Furthermore we can use the properties of the log map to show that:

$$||\mathrm{Log}_1(\phi_i) - \mathrm{Log}_1(\phi_j)||_2 = d_{\mathcal{H}}(\phi_i, \phi_j) = \frac{1}{2} d_{\mathcal{P}}(p_i, p_j) \tag{11}$$

for any two p_i, p_j on $\mathcal{P}$. Note that the multiplicative factor $\frac{1}{2}$ is important to guarantee the isometry but will not have any impact on the covariance function defined in (14) as it is implicit in the hyperparameter.

3. Gaussian Processes on PDFs

In this section, we focus on constructing GPs on $\mathcal{P}$. A GP Z on $\mathcal{P}$ is a random field indexed by $\mathcal{P}$ so that $(Z(p_1), \ldots, Z(p_n))$ is a multivariate Gaussian vector for any $n \in \mathbb{N}\backslash\{0\}$ and $p_1, \ldots, p_n \in \mathcal{P}$. A GP is completely specified by its mean function and its covariance function. We define a mean function $m : \mathcal{P} \to \mathbb{R}$ and the covariance function $C : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ of a real process Z as

$$m(p_i) = \mathbb{E}\big[Z(p_i)\big]. \tag{12}$$

$$C(p_i, p_j) = \mathbb{E}\big[(Z(p_i) - m(p_i))(Z(p_j) - m(p_j))\big]. \tag{13}$$

Thus, if a GP is assumed to have zero mean function ($m \equiv 0$), defining the covariance function completely defines the process behavior. In this paper, we assume that the GPs are centered and we only focus on the issue of constructing a proper covariance function C on $\mathcal{P}$.

3.1. Constructing covariance functions on $\mathcal{P}$

A covariance function C on $\mathcal{P}$ must satisfy the following conditions. For any $n \in \mathbb{N}\backslash\{0\}$ and $p_1, \ldots, p_n \in \mathcal{P}$, the matrix $\mathbf{C} = [C(p_i, p_j)]_{i,j=1}^n$ is symmetric nonnegative definite. Furthermore, C is called non-degenerate when the above matrix is invertible whenever $p_1, \ldots, p_n$ are two-by-two

distinct [1]. The strategy that we adopt to construct covariance functions is to exploit the full isometry map $\mathrm{Log}_1(.)$ to $\mathcal{E}$ given in (11). That is, we construct covariance functions of the form

$$C(p_i, p_j) = K(\|\mathrm{Log}_1(\phi_i) - \mathrm{Log}_1(\phi_j)\|_2), \tag{14}$$

where $K : \mathbb{R}^+ \to \mathbb{R}$.

Proposition 1. *Let $K : \mathbb{R}^+ \to \mathbb{R}$ be such that $K(u_i, u_j) = K(\|u_i - u_j\|_2)$ is a covariance function on $\mathcal{E}$ and C as defined as in (14). Then*

1. *C is a covariance function.*
2. *If $[K(\|u_i - u_j\|_2)]_{i,j=1}^n$ is invertible, then C is non-degenerate.*

A closely related proof when dealing with Cumulative Density Functions (CDFs) is given in [1]. In practice, we can select the function K from the Matérn family, letting for $t \geq 0$

$$K_\theta(t) = \frac{\delta^2}{\Gamma(\nu)2^{\nu-1}}\Big(\frac{2\sqrt{\nu t}}{\alpha}\Big)^\nu K_\nu\Big(\frac{2\sqrt{\nu t}}{\alpha}\Big), \tag{15}$$

where K_ν is a modified Bessel function of the second kind and Γ is the gamma function. We note $\theta = (\delta^2, \alpha, \nu) \in \Theta$ where $\delta^2 > 0$ is the variance parameter, $\alpha > 0$ is the correlation length parameter and $\nu = \frac{1}{2} + k (k \in \mathbb{N})$ is the smoothness parameter. The Matérn form [37] has the desirable property that GPs have realizations (sample paths) that are k times differentiable [38], which prove its smoothness as function of ν. As $\nu \to \infty$, the Matérn covariance function approaches the squared exponential form, whose realizations are infinitely differentiable. For $\nu = \frac{1}{2}$, the Matérn takes the exponential form. From Proposition 1, the Matérn covariance function defined by $C(p_i, p_j) = K_\theta(\|\mathrm{Log}_1(\phi_i) - \mathrm{Log}_1(\phi_j)\|_2)$ is indeed non-degenerate.

3.2. Regression on $\mathcal{P}$

Having set out the conditions on the covariance function, we can define the regression model on $\mathcal{P}$ by

$$y_i = Z(p_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{16}$$

where Z is a zero mean GP indexed by $\mathcal{P}$ with a covariance function in the set $\{C_\theta; \theta \in \Theta\}$ and $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \gamma^2)$. Here γ^2 is the observation noise variance, that we suppose to be known for simplicity. Moreover, we note $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\mathbf{p} = (p_1, \ldots, p_n)^T$ and $\mathbf{v} = (v_1, \ldots, v_n)^T =$

$(\mathrm{Log}_1(\Psi(p_1)),\ldots,\mathrm{Log}_1(\Psi(p_n)))^T$. The likelihood term is $\mathbb{P}(\mathbf{y}|Z(\mathbf{p})) = \mathcal{N}(Z(\mathbf{p}),\gamma^2 I_n)$ where I_n is the identity matrix. Moreover, the prior on $Z(\mathbf{p})$ is $\mathbb{P}(Z(\mathbf{p})) = \mathcal{N}(0,\mathbf{C}_\theta)$ with $\mathbf{C}_\theta = [K_\theta(\|v_i - v_j\|_2)]_{i,j=1}^n$. We use the product of likelihood and prior terms to perform the integration yielding the log-marginal likelihood

$$l_r(\theta) = -\mathbf{y}^T(\mathbf{C}_\theta + \gamma^2 I_n)^{-1}\mathbf{y} - \log|\mathbf{C}_\theta + \gamma^2 I_n| - \frac{n}{2}\log 2\pi. \tag{17}$$

Let $\theta = \{\theta^j\}_{j=1}^3 = (\delta^2,\alpha,\nu)$ denote the parameters of the Matérn covariance function K_θ. The partial derivatives of $l_r(\theta)$ with respect to θ^j are

$$\frac{\partial l_r(\theta)}{\partial\theta^j} = \frac{1}{2}\mathbf{y}^T\mathbf{C}_\theta^{-1}\frac{\partial\mathbf{C}_\theta}{\partial\theta^j}\mathbf{C}_\theta^{-1}\mathbf{y} - \mathrm{tr}\big[\mathbf{C}_\theta^{-1}\frac{\partial\mathbf{C}_\theta}{\partial\theta^j}\big]. \tag{18}$$

For an unobserved PDF p^* and by deriving the conditional distribution, we arrive at the key predictive equation

$$\mathbb{P}(Z(p^*)|\mathbf{p},\mathbf{y},p^*) = \mathcal{N}(\mu(p^*),\sigma^2(p^*)), \tag{19}$$

with

$$\begin{cases} \mu(p^*) = \mathbf{C}_\theta^{*T}(\mathbf{C}_\theta + \gamma^2 I_n)^{-1}\mathbf{y}, \\ \sigma^2(p^*) = C_\theta^{**} - \mathbf{C}_\theta^{*T}(\mathbf{C}_\theta + \gamma^2 I_n)^{-1}\mathbf{C}_\theta^*, \end{cases} \tag{20}$$

where $\mathbf{C}_\theta^* = K_\theta(\mathbf{v},v^*)$ and $C_\theta^{**} = K_\theta(v^*,v^*)$ for $v^* = \mathrm{Log}_1(\Psi(p^*))$. As we have introduced GP regression indexed by PDFs, we will present GP classifier in the next section.

3.3. Classification on $\mathcal{P}$

For the classification part, we focus on the case of binary outputs, i.e., $y_i \in \{-1,+1\}$. We first adapt the Laplace approximation to GPc indexed by PDFs in Section 3.3.1. We also give the approximate marginal likelihood and the Gaussian predictive distribution in Section 3.3.2.

3.3.1. Approximation of the posterior

The likelihood is the product of individual likelihoods $\mathbb{P}(\mathbf{y}|Z(\mathbf{p})) = \prod_{i=1}^n \mathbb{P}(y_i|Z(p_i))$ where $\mathbb{P}(y_i|Z(p_i)) = \sigma(y_i Z(p_i))$ and $\sigma(.)$ refers to the sigmoid function satisfying $\sigma(t) = \frac{1}{1+\exp(-t)}$. As for regression, the prior law of GPc is $\mathbb{P}(Z(\mathbf{p})) = \mathcal{N}(0,\mathbf{C}_\theta)$. From the Bayes' rule, the posterior distribution of $Z(\mathbf{p})$ satisfies

$$\mathbb{P}(Z(\mathbf{p})|\mathbf{y}) = \frac{\mathbb{P}(\mathbf{y}|Z(\mathbf{p}))\times\mathbb{P}(Z(\mathbf{p}))}{\mathbb{P}(\mathbf{y}|\mathbf{p},\theta)}, \propto \mathbb{P}(\mathbf{y}|Z(\mathbf{p}))\times\mathbb{P}(Z(\mathbf{p})), \tag{21}$$

where $\mathbb{P}(\mathbf{y}|\mathbf{p},\theta)$ is the exact marginal likelihood. The log-posterior is simply proportional to $\log \mathbb{P}(\mathbf{y}|Z(\mathbf{p})) - \frac{1}{2}Z(\mathbf{p})^T\mathbf{C}_\theta^{-1}Z(\mathbf{p})$. For the Laplace approximation, we approximate the posterior given in (21) by a Gaussian distribution. We can find the maximum a posterior (MAP) estimator denoted by $\hat{Z}(\mathbf{p})$, iteratively, according to

$$Z^{k+1}(\mathbf{p}) = (\mathbf{C}_\theta + \mathbf{W})^{-1}(\mathbf{W}Z^k(\mathbf{p}) + \nabla\mathbb{P}(\mathbf{y}|Z^k(\mathbf{p}))), \tag{22}$$

where $\mathbf{W}$ is a $n \times n$ diagonal matrix with entries $\mathbf{W}_{ii} = \frac{\exp(-\hat{Z}(p_i))}{(1+\exp(-\hat{Z}(p_i)))^2}$. Using the MAP estimator, we can specify the Laplace approximation of the posterior by

$$\hat{\mathbb{P}}(Z(\mathbf{p})|\mathbf{p},\mathbf{y}) = \mathcal{N}(\hat{Z}(\mathbf{p}), (\mathbf{C}_\theta^{-1} + \mathbf{W})^{-1}). \tag{23}$$

3.3.2. Predictive distribution

We evaluate the approximate marginal likelihood denoted by $\hat{\mathbb{P}}(\mathbf{y}|\mathbf{p},\theta)$ instead of the exact marginal likelihood $\mathbb{P}(\mathbf{y}|\mathbf{p},\theta)$ given in the denominator of (21). Integrating out $Z(\mathbf{p})$, the log-marginal likelihood is approximated by

$$l_c(\theta) = -\frac{1}{2}\hat{Z}(\mathbf{p})^T\mathbf{C}_\theta^{-1}\hat{Z}(\mathbf{p}) + \log p(\mathbf{y}|\hat{Z}(\mathbf{p})) - \frac{1}{2}\log\left|I_n + \mathbf{W}^{\frac{1}{2}}\mathbf{C}_\theta\mathbf{W}^{\frac{1}{2}}\right|. \tag{24}$$

The partial derivatives of $l_c(\theta)$ with respect to θ^j satisfy

$$\frac{\partial l_c(\theta)}{\partial\theta^j} = \frac{\partial l_c(\theta)}{\partial\theta^j}|_{\hat{Z}(\mathbf{p})} + \sum_{i=1}^{n}\frac{\partial l_c(\theta)}{\partial\hat{Z}(p_i)}\frac{\partial\hat{Z}(p_i)}{\partial\theta^j}. \tag{25}$$

The first term, obtained when we assume that $\hat{Z}(\mathbf{p})$ (as well as $\mathbf{W}$) does not depend on θ, satisfies

$$\frac{\partial l_c(\theta)}{\partial\theta^j}|_{\hat{Z}(\mathbf{p})} = \frac{1}{2}\hat{Z}(\mathbf{p})^T\mathbf{C}_\theta^{-1}\frac{\partial\mathbf{C}_\theta}{\partial\theta^j}\mathbf{C}_\theta^{-1}\hat{Z}(\mathbf{p}) - \frac{1}{2}\mathrm{tr}\big[(\mathbf{C}_\theta + \mathbf{W}^{-1})^{-1}\frac{\partial\mathbf{C}_\theta}{\partial\theta^j}\big]. \tag{26}$$

The second term, obtained when we suppose that only $\hat{Z}(\mathbf{p})$ (as well as $\mathbf{W}$) depends on θ, is determined by

$$\frac{\partial l_c(\theta)}{\partial\hat{Z}(p_i)} = -\frac{1}{2}\big[(\mathbf{C}_\theta^{-1} + \mathbf{W})^{-1}\big]_{ii}\frac{\partial^3\log p(\mathbf{y}|\hat{Z}(\mathbf{p}))}{\partial^3\hat{Z}(p_i)}, \tag{27}$$

and

$$\frac{\partial\hat{Z}(\mathbf{p})}{\partial\theta^j} = \big(I_n + \mathbf{C}_\theta\mathbf{W}\big)^{-1}\frac{\partial\mathbf{C}_\theta}{\partial\theta^j}\nabla\log p(\mathbf{y}|\hat{Z}(\mathbf{p})). \tag{28}$$

Given an unobserved PDF p^*, the predictive distribution at $Z(p^*)$ is given by

$$\hat{\mathbb{P}}(Z(p^*)|\mathbf{p},\mathbf{y},p^*) = \mathcal{N}(\mu(p^*),\sigma^2(p^*)), \tag{29}$$

with

$$\begin{cases} \mu(p^*) = \mathbf{C}_\theta^{*T} \mathbf{C}_\theta^{-1} \hat{Z}(\mathbf{p}), \\ \sigma^2(p^*) = C_\theta^{**} - \mathbf{C}_\theta^{*T} (\mathbf{C}_\theta + \mathbf{W}^{-1})^{-1} \mathbf{C}_\theta^*. \end{cases} \tag{30}$$

Finally, using the moments of prediction, the predictor for $y^* = +1$ is

$$\pi(p^*) = \int_{\mathbb{R}} \sigma(Z^*) \hat{\mathbb{P}}(Z^* | \mathbf{p}, \mathbf{y}, p^*) dZ^*, \tag{31}$$

where we note $Z^* = Z(p^*)$ for simplicity.

3.4. Covariance parameters estimation

The marginal likelihoods for both regression and classification depend on the covariance parameters controlling the stationarity of the GP. To show potential applications of this framework, we explore several optimization methods in Section 3.4.1 and Section 3.4.2.

3.4.1. log-marginal likelihood gradient

In the marginal likelihood estimation, the parameters are obtained by maximizing the log-marginal likelihood with respect to θ, i.e., finding

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, l_l(\theta), \tag{32}$$

where $l_l(\theta)$ is given in (17) by $l_r(\theta)$ for regression or $l_c(\theta)$ in(24) for classification. We summarize the main steps in Algorithm 1.

Algorithm 1: Gradient descent.

Require: log-marginal likelihood l_l and its gradient ∇l_l

Ensure: $\hat{\theta}$

1: **repeat**
2: $\nabla l_l(\theta(k)) = \{\frac{\partial l_l(\theta(k))}{\partial \theta^j}\}_{j=1}^3$ from (18) or (25)
3: Find the step-size λ (e.g., by backtracking line search)
4: Evaluate $\theta(k+1) = \theta(k) - \lambda \nabla l_l(\theta(k))$
5: Set $k = k + 1$
6: **until** $||\nabla l_l||_2$ is small enough or a maximum iterations is reached

3.4.2. HMC sampling

Generally, the marginal likelihoods are non-convex functions. Indeed, conventional optimization routines may not find the most probable candidate leading to a lost of robustness and uncertainty quantification. To deal with such limitations, we use weak prior distributions for δ^2 and α whereas ν is simply estimated by cross-validation [39]:

$$\mathbb{P}(\delta^2, \alpha) = \mathbb{P}(\delta^2) \times \mathbb{P}(\alpha), \tag{33}$$

with δ^2 and α being independent. Following [40], δ^2 will be assigned a half-Cauchy (positive-only) prior, *i.e.* $\mathbb{P}(\delta^2) = \mathcal{C}(0, b_{\delta^2})$ and α an inverse gamma, *i.e.* $\mathbb{P}(\alpha) = \mathcal{IG}(a_\alpha, b_\alpha)$. Consequently, the log-marginal posterior is proportional to

$$l_p(\delta^2, \alpha) = l_l(\theta) + \log \mathbb{P}(\delta^2) + \log \mathbb{P}(\alpha). \tag{34}$$

When sampling from continuous variables, HMC can prove to be a more powerful tool than usual MCMC sampling. We define the Hamiltonian as the sum of a potential energy and a kinetic energy:

$$E((\theta^1, \theta^2), (s^1, s^2)) \quad = \quad E^1(\theta^1, \theta^2) + E^2(s^1, s^2) - l_p(\theta^1, \theta^2) + \frac{1}{2}\sum_{j=1}^{2} {s^j}^2, \tag{35}$$

which means that $(s^1, s^2) \sim \mathcal{N}(0, I_2)$. Instead of sampling from $\exp\big(l_p(\theta^1, \theta^2)\big)$ directly, HMC operates by sampling from the distribution $\exp\big(-E((\theta^1, \theta^2), (s^1, s^2))\big)$. The differential equations are given by

$$\frac{d\theta^j}{dt} = \frac{\partial E}{\partial s^j} = s^j \quad \text{and} \quad \frac{ds^j}{dt} = -\frac{\partial E}{\partial \theta^j} = -\frac{\partial E^1}{\partial \theta^j}, \tag{36}$$

for $j = 1, 2$. In practice, we can not simulate Hamiltonian dynamics exactly because of time discretization. To maintain invariance of the Markov chain, however, care must be taken to preserve the properties of volume conservation and time reversibility. The leap-frog algorithm, summarized in Algorithm 2, maintains these properties [41].

Algorithm 2: Leap-frog.

1: **for** $k = 1, 2, \ldots$ **do**
2: $s^j(k + \frac{\lambda}{2}) = s^j(k) - \frac{\lambda}{2}\frac{\partial}{\partial \theta^j}E^1(\theta^1(k), \theta^2(k))$ where λ is a finite step-size
3: $\theta^j(k + \lambda) = \theta^j(k) + \lambda s^j(k + \frac{\lambda}{2})$
4: $s^j(k + \lambda) = s^j(k + \frac{\lambda}{2}) - \frac{\lambda}{2}\frac{\partial}{\partial \theta^j}E^1(\theta^1(k + \lambda), \theta^2(k + \lambda))$
5: **end for**

We thus perform a half-step update of the velocity at time $k + \frac{\lambda}{2}$, which is then used to compute $\theta^j(k + \lambda)$ and $s^j(k + \lambda)$. A new state $((\theta^1(N), \theta^2(N)), (s^1(N), s^2(N)))$ is then accepted with the probability

$$\min\Big(1, \frac{\exp\big(-E((\theta^1(N), \theta^2(N), (s^1(N), s^2(N))\big)}{\exp\big(-E((\theta^1(1), \theta^2(1), (s^1(1), s^2(1))\big)}\Big). \tag{37}$$

We summarize the HMC sampling in Algorithm 3.

Algorithm 3: HMC sampling.

Require: log-marginal posteriors l_p and its gradient ∇l_p

Ensure: $\hat{\theta}$

1: Sample a new velocity from a Gaussian distribution $(s^1(1), s^2(1)) \sim \mathcal{N}(0, I_2)$

2: Perform N leapfrog steps to obtain the new state $(\theta^1(N), \theta^2(N))$ and velocity $(s^1(N), s^2(N))$ from Algorithm 2

3: Perform accept/reject of $(\theta^1(N), \theta^2(N))$ with acceptance probability defined in (37).

4. Experimental Results

In this section, we test and illustrate the proposed methods using synthetic, semi-synthetic and real data. For all experiments, we study the empirical results of a Gaussian process indexed by PDFs for both regression and classification.

Baselines. We compare results of GP indexed by PDFs (GPP) where the parameters are estimated by gradient descend (G-GPP) and HMC (HMC-GPP) to: Functional Linear Model (**FLM**) [5] for regression, Nonparametric Kernel Wasserstein (**NKW**) [42] for regression, A GPP based on the Wasserstein distance (**W-GPP**) [43, 1] for classification, and a GPP based on the Jensen-Shannon divergence (**JS-GPP**) [8] for classification.

Performance metrics. For regression, we illustrate the performance of the proposed framework in terms of root mean square error (RMSE) and negative log-marginal likelihood (NLML). For classification, we consider accuracy, area under curve (AUC) and NLML.

4.1. Regression

Dataset. We first consider a synthetic dataset where we observe a finite set of functions simulated according to (16) as $Z(p_i) = h(< \sqrt{p_i}, \sqrt{\tilde{p}} >_2) = 0.5 < \sqrt{p_i}, \sqrt{\tilde{p}} >_2 + 0.5$. In this example, we

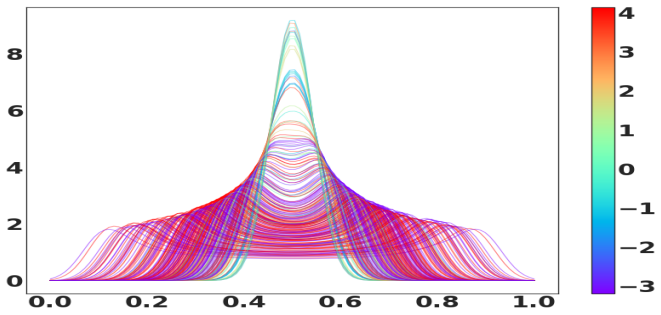

Figure 1: Examples of PDFs input for regression. The output with continuous value in $[-3, 4]$ is illustrated by a colorbar.

consider a truncated Fourier basis (TFB) with random Gaussian coefficients to form the original functions satisfying $g_i(t) = \delta_{i,1}\sqrt{2}\sin(2\pi t) + \delta_{i,2}\sqrt{2}\cos(2\pi t)$ with $\delta_{i,1}, \delta_{i,2} \sim \mathcal{N}(0, 1)$. We also take $\tilde{g}(t) = -0.5\sqrt{2}\sin(2\pi t) + 0.5\sqrt{2}\cos(2\pi t)$. We suppose that $\tilde{p}$ and p_is refer to the corresponding PDFs of $\tilde{g}$ and g_is estimated from samples using the nonparametric kernel method (bandwidths were selected using the method given in [20]). Examples of $n = 100$ estimates are displayed in Fig. 1 with colors depending on their output levels.

Regression results. Focusing on RMSE, we summarize all results in Table 1. Accordingly, the proposed G-GPP gives better precision than FLM. On the other hand, HMC-GPP substantially outperforms NKW with a significant margin. As illustrated in Table 2, we note that the proposed methods are more efficient than the baseline FLM when maximizing the log-marginal likelihood. Again, this is a very simple explanation on how the quality of GPP strongly depends on parameters estimation method. In addition, G-GPP stated in Algorithm 1 is very effective from a computational point of view.

Table 1: Regression: RMSE as a performance metric.

G-GPP		HMC-GPP		FLM		NKW	
mean	std	mean	std	mean	std	mean	std
0.07	0.03	0.13	0.31	0.10	0.04	0.28	0.01

Table 2: Regression: negative log-marginal likelihood as a performance metric.

G-GPP		HMC-GPP		FLM	
mean	std	mean	std	mean	std
73.28	1.14	**21.89**	5.32	329.66	6.52

4.2. Classification

In this section, we perform some extensive experiments to evaluate the proposed methods using a second category of datasets.

4.2.1. Datasets for classification

Synthetic datasets. We consider a dataset of two synthetic PDFs of beta and inverse gamma distributions. This choice is very crucial for many reasons since beta is defined on $[0, 1]$, parametrized by two positive parameters, and has been widely used to represent a large family of PDFs with finite support in various fields. Increasingly, the inverse gamma plays an important role to characterize random fluctuations affecting wireless channels [44]. In both examples, the covariance matrix with $\mathbb{L}^2$ distance and Total Variation TV-distance have a very low rank. We performed this experiment by simulating $n = 200$ pairs of PDFs slightly different for the two classes. Each observation represents a density when we add a random white noise. We refer to these datasets as Beta and InvGamma, see random examples in Fig. 2 (a&b). We also illustrate the Fréchet mean for each class. The search of the mean is performed using a gradient approach detailed in [6].

Semi-synthetic dataset. Data represent clinical growth charts for children from 2 to 12 years [5]. We refer to this dataset as Growth. We simulate the charts from centers for disease control and prevention [45] through the available quantile values. The main goal is to classify observations by gender. Each simulation represents the size growth (the increase) of a child according to his age (120 months). We represent observations as nonparametric PDF and we display some examples in Fig. 2 (c). For each class: girls (red) and boys (blue) we show the Fréchet mean in black.

Real dataset. The first public dataset consists of 1500 images representing maize leaves [46] with specific textures whereas the goal is to distinguish healthy and non-healthy plants. We refer to this dataset as Plants. Motivated by this application, we first represent each image with its wavelet-deconvolved version and form a high-dimensional vector of 262144 components. Fig. 3 illustrates an example of two original images (left): a healthy plant (top) and a plant with disease (bottom),

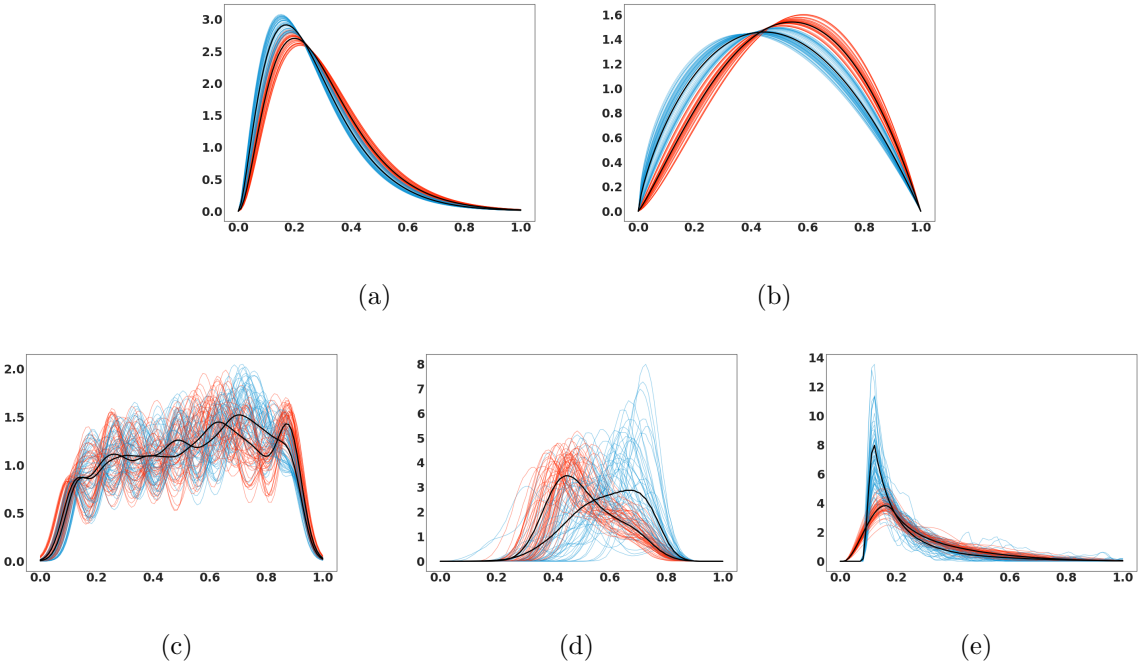

Figure 2: Synthetic PDFs for (a) InvGamma and (b) Beta with class 1 (red) and class 2 (blue). Semi-synthetic PDFs for (c) Growth with girls (red) and boys (blue). Real PDFs for (d) Temp with uninfected (red) and infected (blue). Real PDFs for (e) Plants with disease (red) and healthy (blue). The Fréchet mean for each class in black.

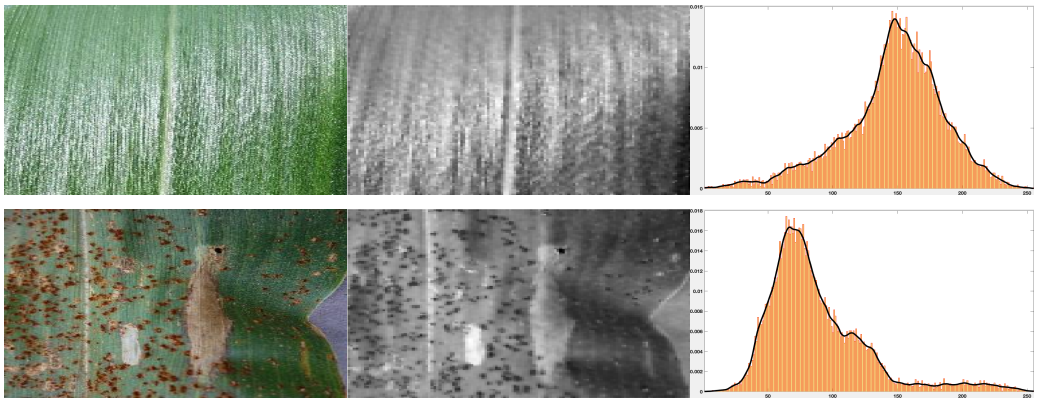

Figure 3: Two examples from maize plants dataset where (top) is a healthy leaf and (bottom) is a leaf with disease. For each class: an original image (left), the extracted features (middle), and the normalized histogram (right).

their wavelet-deconvolved versions (middle), and the corresponding histograms (right). We also

display PDFs from histograms for each example in Fig. 3 (right column in black).

A second real dataset with 1717 observations gives the body temperature of dogs. For this dataset, temporal measures of infected and uninfected dogs are stored during 24 hours. The infection by a parasite is suspected to cause persistent fever despite veterinary medicine [47]. The main goal is to learn the relationship between the infection and a dominant pattern from temporal temperatures. We display some examples of infected (blue) and uninfected (red) in Fig. 2 (c) and we refer to this dataset as Temp. The PDF estimates were obtained using an automatic bandwidth selection method described in [20]. We illustrate some examples of PDFs from real datasets in Fig. 2 (d&e). We remind that high-dimensional inputs make traditional machine learning techniques fail to solve the problem at hand. However, the spectral histograms as marginal distributions of the wavelet-deconvolved image can be used to represent/classify original images [48]. In fact, instead of comparing the histograms, a better way to compare two images (here a set of repetitive features) would be to compare their corresponding densities.

4.2.2. Classification results

We learn the model parameters from 75% of the dataset whereas the rest is kept for test. This subdivision has been performed randomly 100 times. The performance is given as a mean and the corresponding standard deviation (std) in order to reduce the bias (class imbalance and sample representativeness) introduced by the random train/test split.

Results on synthetic datasets. We summarize all evaluation results on synthetic datasets in Fig. 4 (a&b). Accordingly, one can observe that both HMC-GPP, W-GPP and JS-GPP reach the best accuracy values for InvGamma with a little margin for the proposed HMC-GPP. On the other hand, G-GPP and HMC-GPP heavily outperform W-GPP and JS-GPP for Beta. Again, this simply shows how each optimization method impacts the quality of the predictive distributions.

Results on semi-synthetic data. We summarize all results in Fig. 4 (c) where we show accuracy and AUC values on the Growth dataset as boxplots from 100 tests. One can observe that G-GPP gives the best accuracy with a significant margin. Note that we have used 10^3 HMC iterations in Algorithm 3. Furthermore, we set the "Burn-in" and "Thinning" in order to ensure a fast convergence of the Markov chain and to reduce sample autocorrelations.

Results on real data. We further investigate whether our proposed methods can be used with real data. Fig. 5 (a&b) shows the boxplots of accuracy and AUC values for Temp and Plants,

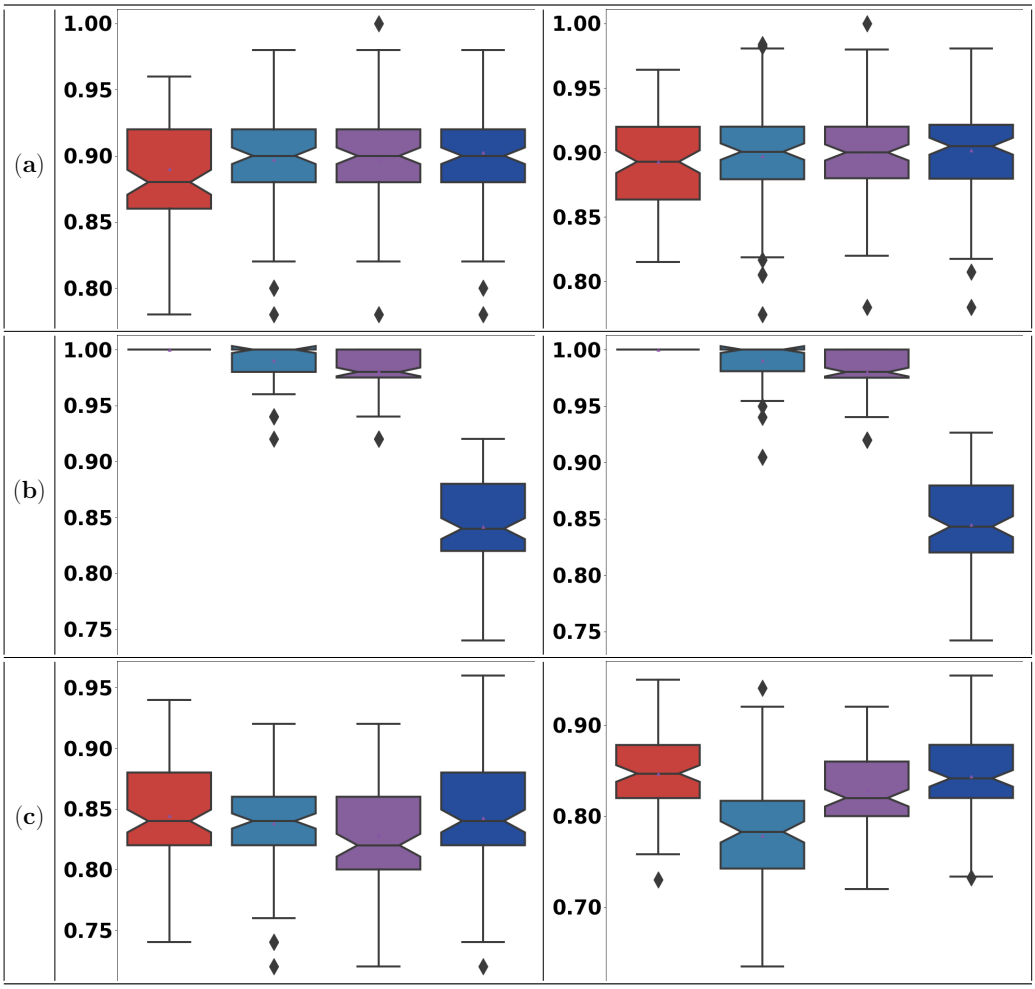

Figure 4: Boxplots of the classification accuracy (left) and AUC (right) on synthetic and semi-synthetic datasets: (a) InvGamma, (b) Beta, and (c) Growth. In each subfigure, the performance is given for different methods: G-GPP (red), HMC-GPP (light blue), W-GPP (violet), and JS-GPP (dark blue).

respectively. In short, we highlight that the proposed methods successfully modeled these datasets with improved results in comparison with W-GPP.

Fortunately, the experiments have shown that the problem of big iterations, usually needed to simulate the Markov chains for complex inputs is partially solved by considering the proposed HMC sampling (Algorithm 3). In closing, we can state that the leap-frog algorithm (Algorithm 2), based on Hamiltonian dynamics, allows us to early search the best directions giving the best minimum of

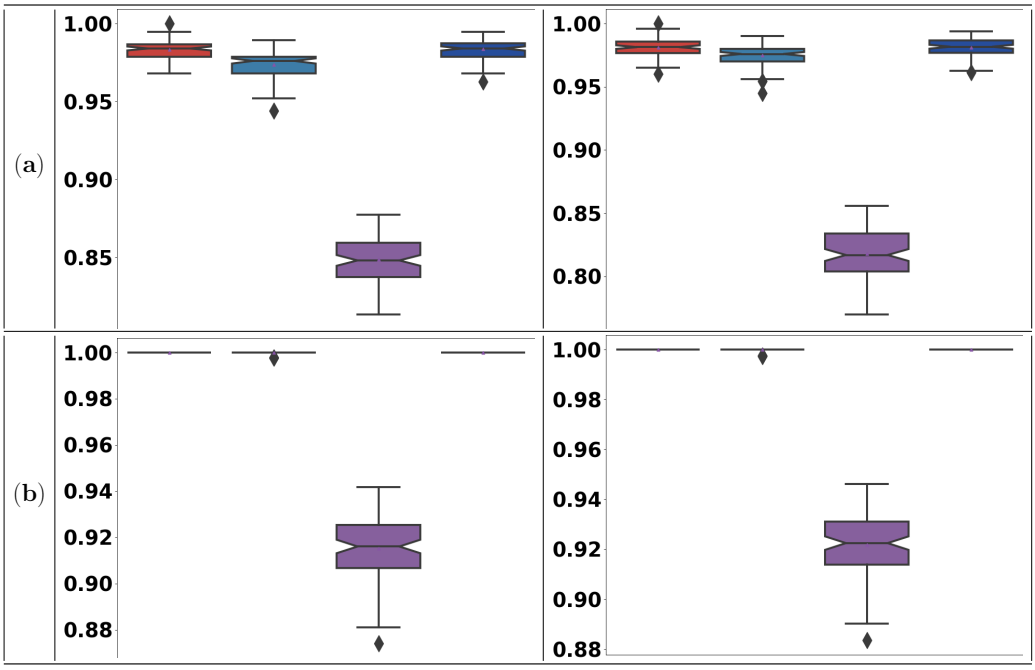

Figure 5: Boxplots of the classification accuracy (left) and AUC (right) on real datasets: (a) Temp and (b) Plants. In each subfigure, the performance is given for different methods: G-GPP (red), HMC-GPP (light blue), W-GPP (violet), and JS-GPP (dark blue).

the Hamiltonian defined in (35).

4.2.3. Summary of all classification results

Table 3: Classification: negative log-marginal likelihood as a performance metric.

Datasets	**Synthetic**				**Semi-synthetic**		**Real data**			
	InvGamma		**Beta**		**Growth**		**Temp**		**Plants**	
Method	mean	std	mean	std	mean	std	mean	std	mean	std
G-GPP	**30.50**	2.43	**4.41**	0.06	68.03	3.43	**98.66**	0.73	98.65	0.72
HMC-GPP	105.35	0.22	105.28	0.21	**61.65**	2.24	105.36	0.22	**9.33**	0.21
JS-GPP	32.2	2.38	42.87	2.73	62.0	3.02	116.65	4.13	10.26	0.12

We also confirm all previous results from Table 3, which summarizes the mean and the std of NLML values for all datasets. These clearly show that at least one of the proposed methods (G-GPP or HMC-GPP) better minimizes the NLML than JS-GPP. This brings more quite accurate estimates,

which prove the predictive power of our approaches.

5. Conclusion

In this paper, we have introduced a novel framework to extend Bayesian learning models and Gaussian processes when the index support is identified with the space of probability density functions (PDFs). We have detailed and applied different numerical methods to learn regression and classification models on PDFs. Furthermore, we showed new theoretical results for the Matérn covariance function defined on the space of PDFs. Extensive experiments on multiple and varied datasets have demonstrated the effectiveness and efficiency of the proposed methods in comparison with current state-of-the-art methods.

Acknowledgements

This work was partially funded by the French National Centre for Scientific Research.

References

[1] F. Bachoc, F. Gamboa, J.-M. Loubes, N. Venet, A Gaussian process regression model for distribution inputs, IEEE Transactions on Information Theory 64 (2018) 6620–6637.

[2] M. Abt, W. Welch, Fisher information and maximum-likelihood estimation of covariance parameters in Gaussian stochastic processes, The Canadian Journal of Statistics 26 (1998) 127–137.

[3] A. Srivastava, E. Klassen, Functional and shape data analysis, Springer-Verlag, New York, USA, 2016.

[4] C. Samir, S. Kurtek, A. Srivastava, N. Borges, An elastic functional data analysis framework for preoperative evaluation of patients with rheumatoid arthritis, in: Winter Conference on Applications of Computer Vision, IEEE, Lake Placid, NY, USA, 2016, pp. 1–8.

[5] J. Ramsay, B. Silverman, Functional Data Analysis, Springer-Verlag, New York, USA, 2005.

[6] A. Srivastava, I. Jermyn, S. Joshi, Riemannian analysis of probability density functions with applications in vision, in: Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Minneapolis, USA, 2007, pp. 1–8.

[7] M. Itoh, H. Satoh, Geometry of Fisher information metric and the barycenter map, Entropy 17 (2015) 1814–1849.

[8] H. Nguyen, J. Vreeken, Non-parametric Jensen-Shannon divergence, in: Machine Learning and Knowledge Discovery in Databases, Springer International Publishing, Cham, Switzerland, 2015, pp. 173–189.

[9] C. Rao, Information and the accuracy attainable in the estimation of statistical parameters, Bulletin of Calcutta Mathematical Society 37 (1945) 81–91.

[10] S.-I. Amari, O. Barndorff-Nielsen, R. Kass, S. Lauritzen, C. Rao, Differential geometry in statistical inference, Institute of Mathematical Statistics, Hayward, CA, 1987.

[11] N. Ay, J. Jost, H. Le, L. Schwachhöfer, Information geometry, Springer, Cham, Switzerland, 2017.

[12] S.-I. Amari, Information geometry and its applications, 1st Edition, Springer, Tokyo, Japan, 2016.

[13] F. Barbaresco, Information Geometry of Covariance Matrix: Cartan-Siegel Homogeneous Bounded Domains, Mostow/Berger Fibration and Fréchet Median, Springer, Berlin, Heidelberg, 2013, Ch. 9, pp. 199–255.

[14] F. N. M. Liu, B. Vemuri, Jensen divergence-based means of SPD matrices, Springer, Berlin, Heidelberg, 2013, Ch. 6, pp. 111–122.

[15] Y. Shishido, Strong symplectic structures on spaces of probability measures with positive density function, Proceedings of the Japan Academy, Series A, Mathematical Sciences 81 (2005) 134–136.

[16] B. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, G. Lanckriet, Non-parametric estimation of integral probability metrics, in: International Symposium on Information Theory (ISIT), IEEE, Piscataway, NJ, USA, 2010, pp. 1428–1432.

[17] D. Sutherland, J. Oliva, B. Póczos, J. Schneider, Linear-time learning on distributions with approximate kernel embeddings, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16, AAAI Press, Phoenix, Arizona, 2016, pp. 2073–2079.

[18] J. Oliva, W. Neiswanger, B. Póczos, J. G. Schneider, E. P. Xing, Fast distribution to real regression, in: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, PMLR, Reykjavik, Iceland, 2014, pp. 706–714.

[19] B. Póczos, A. Singh, A. Rinaldo, L. Wasserman, Distribution-free distribution regression, in: Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, PMLR, Scottsdale, Arizona, USA, 2013, pp. 507–515.

[20] Z. Botev, J. Grotowski, D. Kroese, Kernel density estimation via diffusion, The Annals of Statistics 38 (2010) 2916–2957.

[21] Z. Zhang, E. Klassen, A. Srivastava, Robust comparison of kernel densities on spherical domains, Sankhya A: The Indian Journal of Statistics 81 (2019) 144–171.

[22] C. Atkinson, A. Mitchell, Rao's distance measure, The Indian Journal of Statistics 43 (1981) 345–365.

[23] A. Bhattacharyya, On a measure of divergence between two statistical populations defined by their probability distributions, Bulletin of the Calcutta Mathematical Society 35 (1943) 99–109.

[24] C. Rasmussen, C. Williams, Gaussian processes for machine learning, The MIT Press, Cambridge, London, 2006.

[25] D. Hernández-Lobato, J. Hernández-lobato, P. Dupont, Robust multi-class Gaussian process classification, in: Proceedings of the 24th International Conference on Neural Information Processing Systems, NIPS'11, Curran Associates, Inc., Red Hook, NY, USA, 2011, pp. 280–288.

[26] S. Duane, A. Kennedy, B. Pendleton, D. Roweth, Hybrid Monte Carlo, Physics Letters B 195 (1987) 216–222.

[27] T. Friedrich, Die Fisher-information und symplektische strukturen, Mathematische Nachrichten 153 (1991) 273–296.

[28] B. Sriperumbudur, K. Fukumizu, A. Gretton, A. Hyvärinen, R. Kumar, Density estimation in infinite dimensional exponential families (2013). arXiv:1312.3516.

[29] C. Rao, Diversity and dissimilarity coefficients: A unified approach, Theoretical Population Biology 21 (1982) 24–43.

[30] M. Bauer, M. Bruveris, P. Michor, Uniqueness of the Fisher–Rao metric on the space of smooth densities, Bulletin of the London Mathematical Society 48 (2016) 499–506.

[31] G. Pistone, C. Sempi, An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one, The Annals of Statistics 23 (1995) 1543–1561.

[32] S. Helgason, Differential geometry, lie groups, and symmetric spaces, Academic Press, New York, 1978.

[33] J. Lee, Riemannian manifolds: An introduction to curvature, Springer Science, New York, 1997.

[34] N. Cencov, Statistical decision rules and optimal inference, Translations of Mathematical Monographs, American Mathematical Society, Providence, R.I., 1982.

[35] C. Samir, P.-A. Absil, A. Srivastava, E. Klassen, A gradient-descent method for curve fitting on Riemannian manifolds, Foundations of Computational Mathematics 12 (2012) 49–73.

[36] M. Bauer, E. Klassen, S. Preston, Z. Su, A diffeomorphism-invariant metric on the space of vector-valued one-forms (2018). arXiv:1812.10867.

[37] M. Stein, Interpolation of spatial data, Springer-Verlag, New York, USA, 1999.

[38] M. Genton, W. Kleiber, Cross-covariance functions for multivariate geostatistics, Statistical Science 30 (2015) 147–163.

[39] R. Neal, Monte Carlo implementation of Gaussian process models for Bayesian regression and classification, Technical report, University of Toronto (Dept. of Statistics), Toronto, Canada, 1997.

[40] A. Gelman, Prior distributions for variance parameters in hierarchical models, Bayesian Analysis 1 (2006) 515–533.

[41] R. Neal, MCMC using Hamiltonian dynamics, Handbook of Markov Chain Monte Carlo 54 (2010) 113–162.

[42] B. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, G. Lanckriet, Non-parametric estimation of integral probability metrics, in: International Symposium on Information Theory (ISIT), IEEE, Piscataway, NJ, USA, 2010, pp. 1428–1432.

[43] A. Mallasto, A. Feragen, Learning from uncertain curves: the 2-Wasserstein metric for Gaussian processes, in: Neural Information Processing Systems (NIPS), Curran Associates, Inc., Long Beach, CA, USA, 2017, pp. 5660–5670.

[44] S. Atapattu, C. Tellambura, H. Jiang, A mixture gamma distribution to model the snr of wireless channels, IEEE Transactions on Wireless Communications 10 (2011) 4193–4203.

[45] R. Kuczmarski, C. Ogden, S. Guo, L. Grummer-Strawn, K. Flegal, Z. Mei, R. Wei, L. Curtin, A. Roche, C. Johnson, 2000 CDC growth charts for the united states: methods and development, Vital and Health Statistics 246 (2002) 1–190.

[46] C. DeChant, T. Wiesner-Hanks, S. Chen, E. Stewart, J. Yosinski, M. Gore, R. Nelson, H. Lipson, Automated identification of northern leaf blight-infected Maize plants from field imagery using deep learning, Phytopathology 107 (2017) 1426–1432.

[47] P. Kumar, A. Kumar, Haemato-biochemical changes in dogs infected with Babesiosis, in: Conference on Food Security and Sustainable Agriculture, International Journal of Chemical Studies, 2018, pp. 21–24.

[48] X. Liu, D. Wang, Texture classification using spectral histograms, IEEE Transactions on Image Processing 12 (2003) 661–670.