Constrained and Low Rank Gaussian Process on some Manifolds

Tien-Tam Tran

▶ **To cite this version:**

Tien-Tam Tran. Constrained and Low Rank Gaussian Process on some Manifolds. Probability [math.PR]. Université Clermont Auvergne, 2023. English. NNT : 2023UCFA0108 . tel-04529284

HAL Id: tel-04529284
https://theses.hal.science/tel-04529284

Submitted on 2 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITY OF CLERMONT AUVERGNE

Ecole Doctorale Des Sciences Pour L'Ingenieur (SPI)

By

Tien – Tam TRAN

To obtain the degree of

DOCTOR OF UNIVERSITY

Specialty : Computer Science

Constrained and Low Rank Gaussian Process on Some Manifolds

Publicly defended on December 20, 2023, in front of the jury composed of:

Pr. Christophette BLANCHET-SCALLIET ,	l'École Centrale de Lyon, France	Reviewer
Pr. Hong-Van LE ,	Czech Academy of Sciences, Czech Republic	Reviewer
Pr. Shantanu H. JOSHI ,	University of California Los Angeles, USA	Member
Pr. Rodolphe LE RICHE ,	Director CNRS, France	Member
Pr. Mourad BAIOU ,	Director CNRS, France	Jury President
Pr. José BRAGA,	University Paul Sabatier, France	Invited
Pr. Chafik SAMIR ,	University of Clermont Auvergne, France	Supervisor

LIMOS - CNRS UMR 6158, 63178, Aubière, France

THESIS

Acknowledgments

First and foremost, I would like to express my gratitude to my thesis supervisor Prof. Chafik Samir for his support and guidance throughout my PhD journey. He constantly motivated me to overcome my initial challenges. I am deeply thankful for the significant amount of time he invested in helping me. I have learned a great deal from him, and I extend my heartfelt appreciation for everything.

I would like to thank my committee members Christophette Blanchet-Scalliet, Hong-Van Le, Shantanu H. Joshi, Rodolphe Le Riche, Mourad Baiou and José Braga for their precious time, shared positive insight and guidance.

I would also like to thank all members of LIMOS for their sympathies and friendships. Especially, Prof. Viet-Hung Nguyen helped me greatly during the last three years, and I would like to express my gratitude to him. I can never forget the good moments I shared with my colleagues Anis Fradi, Yann Feunteun, Roxane Jouseau, Amal Omrani and my vietnamese friends: anh Minh, Trang Ngo, Trang Vo, Hieu, Huyen, Loan, Hoang, Minh, Thao. Thank you for teaching me valuable life lessons beyond our time in the lab.

There are many professors and friends who have helped me in the past. Prof. Duc-Thai Do spent a considerable amount of time and showed great consideration when I was in Vietnam. Prof. Tien-Dzung Nguyen and Dr. Tat-Dat To provided me with significant support both in terms of finances and emotional support during my M2 in Toulouse. I also want to express my gratitude to my friends in Toulouse, and my friends in the M1 class in Vien Toan.

Finally, I would like to acknowledge the French ANR IA PhD Grants for providing financial support during my Ph.D., as well as LabEx Archimede and the Torus Foundation for financing my M2.

Abstract

We divide the thesis into three main parts and we summarize the major contributions as follows.

Low complexity Gaussian processes

Gaussian Process (GP) regression usually scales as $O(n^3)$ for computation and $O(n^2)$ for memory requirements, where n represents the number of observations. These limitations makes GP inefficient for many problems when n is large. In this thesis, we investigate the Karhunen-Loève expansion of Gaussian processes which offers several advantages over low-rank compression techniques. By truncating the Karhunen-Loève expansion, we obtain an explicit low-rank approximation of the covariance matrix (Gram matrix), greatly simplifying statistical inference when the level of truncation is small relative to n. We then provide explicit solutions for low complexity Gaussian processes.

We seek Karhunen-Loève expansions, by solving for eigenpairs of a differential operator where the covariance function serves as the Green function. We offer explicit solutions for the Matérn differential operator and for differential operators with eigenfunctions represented by classical polynomials. In the experimental section, we compare the proposed methods with alternative approaches or baseline, revealing their enhanced capability in capturing relevant patterns.

Constrained Gaussian processes

The second contribution introduces a novel approach used constrained Gaussian processes to approximate a density function based with a prior from a finite set, only few, observations. To address these constraints, our approach involves modeling the square root of unknown density function with a Gaussian process prior. In this part of the work, we adopt a truncated version of the Karhunen-Loève expansion as an approximation method. A notable advantage of this approach is that the coefficients are Gaussian and independent, with the constraints on the realized functions entirely dictated by the constraints on the random coefficients. After conditioning on both available data and constraints, the posterior distribution of the coefficients is a normal constrained to the unit sphere. This distribution is analytical intractability, which requires the use of numerical methods for approximation. To this end, we employs spherical Hamiltonian Monte Carlo (HMC). The utility and the efficiency of the proposed framework are validated through a series of experiments, with performance comparisons against alternative methods.

Transfer learning on the manifold of finite probability measures

Finally, we introduce transfer learning models in the space of probability measures on a finite set I, denoted as $\mathcal{P}+(I)$. In our formulation, we endow the space $\mathcal{P}_+(I)$ with the Fisher-Rao metric, transforming it into, a nice and easy to use, Riemannian manifold. This Riemannian manifold, $\mathcal{P}_+(I)$, holds a significant place in information geometry with a wide range of scientific and engineering applications. Within this thesis, we provide detailed formulas for geodesics, the exponential map, the log map, and the parallel transport on $\mathcal{P}_+(I)$.

Our exploration extends to statistical models on $\mathcal{P}_+(I)$, typically conducted within the tangent space of this manifold. With a comprehensive set of geometric tools, we introduce transfer learning models facilitating knowledge transfer between these tangent spaces. Detailed algorithms for transfer learning encompassing Principal Component

Analysis (PCA) and linear regression models are presented. To substantiate these concepts, we conduct a series of experiments, offering empirical evidence of their efficacy.

Keywords: Artificial intelligence; Gaussian processes; Classification; Regression; Constrained Gaussian processes; HMC sampling; Regression; Low Rand Gaussian processes; Riemannian manifold; Fisher-Rao metric; Parallel transport, Transfer learning; Statistical models;

Résumé

La thèse est divisée en trois parties principales, nous résumerons les principales contributions de la thèse comme suit.

Processus gaussiens à faible complexité

La régression par processus gaussien s'échelonne généralement en $O(n^3)$ en termes de calcul et en $O(n^2)$ en termes d'exigences de mémoire, où n représente le nombre d'observations. Cette limitation devient inapplicable pour de nombreux problèmes lorsque n est grand. Dans cette thèse, nous étudions l'expansion de Karhunen-Loève des processus gaussiens, qui présente plusieurs avantages par rapport aux techniques de compression à faible rang. En tronquant l'expansion de Karhunen-Loève, nous obtenons une approximation explicite à faible rang de la matrice de covariance (matrice de Gram), simplifiant considérablement l'inférence statistique lorsque le nombre de troncatures est faible par rapport à n.

Ensuite, nous fournissons des solutions explicites pour les processus gaussiens à faible complexité. Tout d'abord, nous cherchons des expansions de Karhunen-Loève en résolvant les paires propres d'un opérateur différentiel où la fonction de covariance sert de fonction de Green. Nous offrons des solutions explicites pour l'opérateur différentiel de Matérn et pour les opérateurs différentiels dont les fonctions propres sont représentées par des polynômes classiques. Dans la section expérimentale, nous comparons nos méthodes proposées à des approches alternatives, révélant ainsi leur capacité améliorée à capturer des motifs complexes.

Processus gaussiens contraints

Cette thèse introduit une approche novatrice utilisant des processus gaussiens contraints pour approximer une fonction de densité basée sur des observations. Pour traiter ces contraintes, notre approche consiste à modéliser la racine carrée de la fonction de densité inconnue réalisée comme un processus gaussien. Dans ce travail, nous adoptons une version tronquée de l'expansion de Karhunen-Loève comme méthode d'approximation. Un avantage notable de cette approche est que les coefficients sont gaussiens et indépendants, les contraintes sur les fonctions réalisées étant entièrement dictées par les contraintes sur les coefficients aléatoires. Après conditionnement sur les données disponibles et les contraintes, la distribution postérieure des coefficients est une distribution normale contrainte à la sphère unité. Cette distribution pose des difficultés analytiques, nécessitant des méthodes numériques d'approximation. À cette fin, cette thèse utilise l'échantillonnage Hamiltonien Monte Carlo sphérique (HMC). L'efficacité du cadre proposé est validée au moyen d'une série d'expériences, avec des comparaisons de performances par rapport à des méthodes alternatives.

Apprentissage par transfert sur la variété des mesures de probabilité finies

Finalement, nous introduisons des modèles d'apprentissage par transfert dans l'espace des mesures de probabilité sur un ensemble fini I, noté $\mathcal{P} + (I)$. Dans notre étude, nous dotons l'espace $\mathcal{P}_+(I)$ de la métrique de Fisher-Rao, le transformant en une variété riemannienne. Cette variété riemannienne, $\mathcal{P}_+(I)$, occupe une place significative en géométrie de l'information et possède de nombreuses applications. Au sein de cette thèse, nous fournissons des formules détaillées pour les géodésiques, la fonction exponentielle, la fonction logarithmique et le transport parallèle sur $\mathcal{P}_+(I)$.

Notre exploration s'étend aux modèles statistiques situés au sein de $\mathcal{P}_+(I)$, généralement réalisés dans l'espace tangent de cette variété. Avec un ensemble complet d'outils géométriques, nous introduisons des modèles d'apprentissage par transfert facilitant le

transfert de connaissances entre ces espaces tangents. Des algorithmes détaillés pour l'apprentissage par transfert, comprenant l'Analyse en Composantes Principales (PCA) et les modèles de régression linéaire, sont présentés. Pour étayer ces concepts, nous menons une série d'expériences, fournissant des preuves empiriques de leur efficacité.

Mots clés: Processus gaussiens; Processus gaussiens contraints; HMC; Régression; Variété riemannienne; Métrique de Fisher-Rao; Apprentissage par transfert

Publications

Here is the list of submitted and published papers during the thesis.

Published journal papers

1. **T.T.Tran**, Y. Feunteun, C. Samir, and J. Braga. A scalable Matérn Gaussian process for learning spatial curves distributions. Information Sciences-2022.
2. **T.T Tran**, A. Fradi and C. Samir. Learning, Inference, and Prediction on Probability Density Functions with Constrained. Information Sciences-2023.

Submitted journal papers

1. **T.T.Tran**, I. Adouani and C. Samir. Cubic Hermite Interpolators on the Space of Probability Measures. **Submitted** to: "MMA: Mathematical Methods in the Applied Sciences".

Preprint journal papers

1. **T.T.Tran**, I. Adouani and C. Samir. Transfer Learning on Space of Probability Measures.
2. A. Fradi, **T.T.Tran**, C. Samir; Reducing the Complexity of Gaussian Processes via Covariance Decomposition when Dealing with Differential Operators Based on Orthogonal Polynomials

Communications with presentation

1. **T.T.Tran**, **Poster**: Transfer Learning of Statistical Models on Riemannian Manifolds, MASCOT-NUM 2022.

List of Abbreviations

Geodesic curve (α)

constrained Gaussian process (cGP)

dimension (d)

distance function ($d(\cdot, \cdot)$)

Data set ($\mathcal{D}$)

Laplace operator (or second derivative in the Euclidean space) (Δ)

expectation ($\mathbb{E}$)

Gaussian noise (ϵ)

Riemannian metric ($\mathfrak{g}$)

Green function ($G(\cdot)$)

Hilbert space ($\mathcal{H}$)

Hamilton Monte Carlo (HMC)

interval or finite set (I)

identity matrix of size $n \times n$ ($\mathcal{I}_n$)

covariance function ($K(\cdot, \cdot)$)

Karhunen-Loève (K-L)

correlation function ($k(\cdot, \cdot)$)

integral operator ($\mathcal{K}$)

differential operator ($\mathcal{L}$)

eigenvalue, eigenfunction (λ_i, ϕ_i)

Levi Civita (LC)

mean function ($m(\cdot)$)

number of truncation (M)

general manifold ($\mathcal{M}$)

Markov chain Monter Carlo (MCMC)

number of observations (n)

the set of natural numbers ($\mathbb{N}^*$)

big O of n ($O(n)$)

open set ($\mathcal{O}$)

probability density function ($p(\cdot)$)

project Gaussian process (pGP)

principle component analysis (PCA)

probability density function (PDF)

the space of probability density function on finite set I ($\mathcal{P}_+(I)$)

data set of large size and data set of small size (P_L, P_S)

List of Figures

List of Tables

List of Algorithms

Contents

Chapter I: General introduction

In this chapter, we describe the context and motivations of the scientific problems that will be addressed in the thesis. In particular, we highlight their importance, the general formulations, and the proposed solutions for different cases. Afterward, we present our main contributions. Finally, we conclude this chapter with an outline of the rest of this manuscript.

I.1 Context and motivations

The thesis can be divided into three main parts. We start by giving an overview without details of each part.

I.1.1 Reduced-rank Gaussian processes

Gaussian processes are powerful tools for non-parametric Bayesian inference and learning, widely employed today. A Gaussian process is characterized by its mean function and covariance function. We typically set the mean function to zero for convenience, while the choice of the covariance function is determined through data-driven learning or prior knowledge. In Gaussian process regression, we assume that the unknown function is a realization of the Gaussian process, and we make predictions for unseen values using Gaussian conditioning. However, this process involves taking inverse of covariance matrix, with computational and memory requirements typically scaling as $O(n^3)$ and $O(n^2)$, respectively, where n represents the data size. This limitation becomes particularly evident when working with large datasets. For example, Gaussian processes have been extensively used in astronomy to model various phenomena, including the cosmic microwave background, active galactic nuclei, and the logarithmic flux of X-ray binaries. Unfortunately, there exist astronomical time series datasets such as NASA's Kepler Mission, K2, TESS, etc.[42], for which applying a Gaussian process model is no longer tractable.

There is many proposals that address those limitations. Most of the previous

methods attempt to approximate the inverse of the covariance matrix using reduced-rank algorithms [5, 96, 110]. On the other hand, some methods based on Variational Inference (VI) consist in finding an approximation of the posterior distribution that minimizes the Kullback-Leibler divergence [121]. The Variational Fourier Features (VFF) method, [58], combines the variational approach with Fourier features and overcomes the local weakness of VI. The Variational Orthogonal Feature (VOF) [19] method improves VFF for a broader class of covariance functions by using the Bochner's theorem. Another strategy for reducing the computational cost is to approximate the Gaussian process as a finite truncation of its Karhunen-Loève expansion.

The Karhunen-Loève expansion allows to represent a stochastic process as an infinite series of orthogonal basis functions and random coefficients. Suppose we have the Karhunen-Loève expansion of a given Gaussian process, we can efficiently compute its truncation with the help of this expansion. Thus, this can lead to a reduced computational cost $O(nM^2)$ where M represents the truncation number. However, explicit Karhunen-Loève expansions are not available for all covariance functions [30, 63]. Finding this expansion is equivalent to determining the Mercer representation (eigenfunction expansion) of the covariance function. This step requires solving the eigenvalues and eigenfunctions of the integral operator that has the covariance function as a kernel. It is very important to note the relationship between the eigenpairs of the integral operator and the differential operator with the covariance as a Green function [39]. In fact, they share the same eigenfunctions but their eigenvalues are inverses. We will exploit this relationship to solve the eigen-equations and provide expansions with eigenfunctions as bases, for several classes of covariance functions.

I.1.2 Constrained Gaussian processes

In Gaussian process models, selecting an appropriate covariance function allows us to capture the expected smoothness and likely patterns within data [98]. However, many real-world phenomena demand the introduction of additional constraints for a more realistic representation. For instance, when modeling a function, we note $f(t)$, representing a chemical concentration, it's essential that the values of f should belong to

the range of 0 to 1. This is a hard constraint that can not be relaxed. In a broader context, many previous works have imposed bound constraints such that $a \leq f(t) \leq b$, where a and b are application dependent constants with $-\infty \leq a < b \leq +\infty$. To give but a two examples, [62] provide an overview and comparison of the warped and bounded likelihood approaches and [26] discretize the global bound constraints into constraints at a finite number of selected points.

Early method to incorporate constraints and ensure that they are satisfied across the entire domain has been done with splines in the prominent work [127]. This approach approximates the Gaussian process using a finite-dimensional model based on spline functions with Gaussian random coefficients. Recently, [85] proposed a basis functions that are piecewise linear but depend on a finite set of knots to form a partition of unity. The coefficients correspond to the values of the original Gaussian process computed at the respective knots. Using this approximated process, the model can incorporate bound constraints, monotonicity constraints, and convexity constraints, which are equivalently translated into constraints on the coefficients. After conditioning with interpolation (observations) and constraints, the problem reduces to simulating the truncated multivariate normal distribution. Very recently, [84] have proposed a comparison fo several Markov chain Monte Carlo (MCMC) methods for sampling and have concluded that Hamiltonian Monte Carlo (HMC) is the most efficient sampler in this context.

To enhance the flexibility of constrained models for various applications, the introduction of new types of constraints is necessary. Typically, the posterior density is not analytically tractable, requiring a sampling method to approximate the integral. In this thesis, we will introduce a new type of constraints and a new method for sampling the posterior distribution.

I.1.3 Transfer learning

Although machine learning methods have achieved great success and have been successfully applied in many applications, their performance are still highly dependent on data, both in term of quality and quantity. Moreover collecting data is expensive and

time-consuming as well as being a crucial step. Transfer learning can assist us in reusing a well trained model or an existing data to enhance a new, albeit different but related model. This methodology is especially promising when we do not have enough data for the new model.

Transfer learning, also known as domain adaptation, focuses on transferring knowledge across domains (source domain and target domain) in order to boost the performance of the target model. There are several applications of transfer learning, including Natural Language Processing (NLP) [28, 129], text sentiment classification [128], image classification [34, 54, 80], human activity classification [55] and multi-language text classification [95]. We refer to [29, 131, 136] for an extended review.

In transfer learning, the source task and the target task need to share some relationships. However, in reality, guaranteeing such relationships can be very challenging which leads us to ask a key question: When should we transfer? In fact, there are situations where transfer learning, when applied to unrelated source and target domains, may result in unsuccessful or even harmful outcomes for the target model or population. This situation is commonly described as *negative transfer*, see for example [99]. Despite its importance, the negative transfer has not received significant attention [37, 48, 103].

For certain applications, data imposes some hard constraints as well as belonging to non-flat manifolds. For example, we will consider a study of probability density functions. In particular, each observation consists of a non-negative functions with a unit integral that belongs to a convex set without a geometric structure. In order to exploit the intrinsic properties of the underlying space, it becomes essential to extend transfer learning into a Riemannian manifold setting. In a different context [46] introduced the Model transport using parallel transport between tangent spaces of a manifold. Nevertheless, there is still more work needed in this direction of research. Subsequently, in this thesis, we develop a new transfer learning model on the manifold of finite probability measures.

I.2 Contributions

We summarize the main contributions in three main parts. First, we study reduced Gaussian processes by truncating their Karhunen-Loève expansion. This approximation provides a natural reduced-rank approximation of the covariance matrix. Applying the matrix inversion lemma, the prediction cost scales as $O(nM^2)$, and the memory requirement as $O(M^2)$ where n is the data size and M is the order of truncation. However, finding the Karhunen-Loève expansion is generally not an easy task. In this thesis, we introduce the Gaussian processes with covariance functions derived from differential operators. We consider the Matérn differential operator on a bounded domain, as well as differential operators with eigenfunctions represented by classical polynomials such as Legendre, Laguerre, Hermite, Chebyshev and Jacobi. To the best of our knowledge, this is a novelty. Through the introduction of various Gaussian process models, we approximate a wide range of functions based on different data patterns. Furthermore, we show that truncating at an appropriate order M, the inverse of the covariance matrix is more numerically stable. To assess the importance of this framework, we have conducted several and various experiments.

As a second contribution, we introduce a new type of constraint into Gaussian process models. The problem consists of approximating a probability density function based on finite set of observation points. Since a probability density function must satisfy non-negativity and have integral equal to one, the approximation needs to satisfy these conditions too. Nevertheless, it is still hard to ensure these conditions in a global setting. Hence, we exploit an isometric mapping to model the square root of the probability density function as a realization of the Gaussian process. The Gaussian process is then approximated by a truncation version of its Karhunen-Loève expansion, represented by finite sum of random coefficients and eigenfunctions that are orthonormal. This approximation, theoretically solid, allows us to incorporate both data observations and constraints into the random coefficients. After conditioning, the posterior distribution is a normal distribution restricted on the unit sphere (Fisher-Bingham distribution). This distribution has been widely studied in statistics and probability sciences. Consequently,

there are many efficients methods to numerically approximate this distribution. In this thesis, we introduce Spherical Hamilton Monte Carlo (HMC), which converges efficiently and very quickly. We give a detailed example with Matérn covariance function on bounded domain for which the eigenfunctions are sine functions. This example is given for illustration without restriction of the proposed model than can be applied, when adapted, for a large panel of applications. We have tested this configuration for various experiments and which demonstrate good performances.

Finally, we develop a new transfer learning on the space of finite probability measures, denoted $\mathcal{P}_+(I)$ where I is a finite index domain. We impose an appropriate geometric structure on $\mathcal{P}_+(I)$ with the Fisher-Rao metric to make it a Riemannian manifold. This space is one of the main topics in information geometry [8]. In this thesis, we first study the geometry of this space in detail, then we derive the explicit formulas for Christoffel symbols, geodesics, exponential map, logarithm map, and the parallel transport. Furthermore, we study the properties of some statistical models in this space. Thanks to the developed geometrical tools, we introduce transfer learning for populations and subdomains on $\mathcal{P}_+(I)$. Without loss of generality, we provide numerical solutions and algorithms for transporting the Principal Component Analysis (PCA) and manifold linear regression models. We have conducted several experiments to show the importance of the proposed framework.

I.3 Outline

The remainder of this manuscript is organized as follows. Chapter II presents some background basic notions that may be useful along this thesis. Chapter III covers low complexity Gaussian processes and explicit solutions with covariance functions derived from differential operators. In Chapter IV, we introduce the constrained Gaussian process framework for approximating a probability density function based on observations. Chapter V is dedicated to the geometry of finite probability measures and transfer learning on this space. We make a general conclusion in Chapter VI.

Chapter II: Backgrounds and basic notions

In this chapter, we provide the mathematical foundations and backgrounds necessary for our upcoming work. We begin by discussing the definition and properties of multivariate normal distributions and then introduce Gaussian processes, which are the main topics of Chapter III and Chapter IV. Next, we gather definitions and theorems from Differential Geometry, which will serve as the foundation for Chapter V. As an illustrative example, we will depict the multivariate normal distribution space as a differentiable manifold.

Organization. Section .1 provides a reminder of the definition and some important properties of the multivariate normal distribution. Section .2 offers detailed information on Gaussian processes, including their definitions, smoothness in the sense of mean square, and their existence. Section .3 presents the formulas for Gaussian process regression. Section .4 lists several results regarding covariance functions, Bochner's theorem, Mercer's theorem, and the relationship between the smoothness of the covariance and the process. Section .5 briefly introduces Differential Geometry. In the final Section .6, we present the geometry of normal distributions with the Fisher-Rao metric.

II.1 Multivariate normal distribution

In probability theory and statistics, the multivariate normal distribution (or Gaussian distribution) is widely used for continuous random variables. Gaussian distributions appear in many real world phenomena, and in many different contexts. For example, the Gaussian distribution maximizes the entropy (see Theorem 6.5.1 in [22]), or by the Central Limit Theorem the limit of the average of independent random variables is Gaussian (see Theorem 9.5.6 in [36]).

A random vector $X \in \mathbb{R}^d$ is said to have a Gaussian distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix Σ if it has the distribution function

$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right), \tag{II.1}$$

where $\Sigma = [\sigma_{ij}] \in \mathbb{R}^{d\times d}$ is a symmetric positive definite matrix and $|\Sigma|$ denotes its

determinant. The inverse matrix $\Sigma^{-1} = [\sigma^{ij}]$ is called the precision matrix, sometimes it is more convenient to work with the precision matrix than the covariance matrix. We denote by $X \sim \mathcal{N}(\mu, \Sigma)$, and call X a Gaussian vector. Gaussian random variables are completely determined by their mean and covariance matrix.

Let $X \sim \mathcal{N}(\mu, \Sigma)$, and split X into two disjoint subsets X_A and X_B. Without loss of generality, we take X_A is the first m component of X and X_B is the remaining $d - m$ components,

$$X = \begin{pmatrix} X_A \\ X_B \end{pmatrix}. \tag{II.2}$$

We also define corresponding partitions of the mean and the covariance as

$$\mu = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix}. \tag{II.3}$$

We have the following important properties.

1. **Normalization**. The probability density function $p(x|\mu, \Sigma)$ is positive on $\mathbb{R}^d$ and

$$\int_{\mathbb{R}^d} p(x|\mu, \Sigma)dx = 1. \tag{II.4}$$

2. **Marginalization**. The marginal densities

$$p(x_A) = \int_{X_B} p(x|\mu, \Sigma)dx_B, \tag{II.5}$$

$$p(x_B) = \int_{X_A} p(x|\mu, \Sigma)dx_A \tag{II.6}$$

   are Gaussian: $X_A \sim \mathcal{N}(\mu_A, \Sigma_{AA}), X_B \sim \mathcal{N}(\mu_B, \Sigma_{BB})$.

3. **Conditioning**. The conditional densities

$$p(x_A|x_B) = \frac{p(x|\mu, \Sigma)}{\int_{X_A} p(x|\mu, \Sigma)dx_A}, \tag{II.7}$$

$$p(x_B|x_A) = \frac{p(x|\mu, \Sigma)}{\int_{X_B} p(x|\mu, \Sigma)dx_B} \tag{II.8}$$

   are also Gaussian:

$$X_A|x_B \sim \mathcal{N}(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \mu_B), \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}, \tag{II.9}$$

$$X_B|x_A \sim \mathcal{N}(\mu_B + \Sigma_{BA}\Sigma_{AA}^{-1}(x_A - \mu_A), \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB}. \tag{II.10}$$

4. **Summation**. The sum of two independent Gaussian random variables with the

same dimension, $X \sim \mathcal{N}(\mu_X, \Sigma_X)$ and $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$, is also Gaussian:

$$X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \Sigma_X + \Sigma_Y). \tag{II.11}$$

5. **Linear combination**. A random vector $X \in \mathbb{R}^d$ is a Gaussian vector if and only if any linear combination of X has a univariate Gaussian distribution. This means that for any $a \in \mathbb{R}^d$ fixed, there exist $\mu_a \in \mathbb{R}$ and $\sigma_a \geq 0$ such that

$$\langle a, X \rangle = \sum_{i=1}^{d} a_i X_i \sim \mathcal{N}(\mu_a, \sigma_a^2). \tag{II.12}$$

6. **Decomposition**. A random vector $X \in \mathbb{R}^d$ is a Gaussian vector if and only if there exist $\mu \in \mathbb{R}^d$ fixed, a matrix $\Lambda \in \mathbb{R}^{n \times r}$ fixed, and a Gaussian vector $W \sim \mathcal{N}(0, \mathcal{I}_r)$ in $\mathbb{R}^r$, where $r \leq n$ and $\mathcal{I}_r$ is the identity matrix in $\mathbb{R}^r$, such that

$$X = \mu + \Lambda W. \tag{II.13}$$

In this case we have $X \sim \mathcal{N}(\mu, \Lambda\Lambda^T)$.

From the previous properties, we see that a Gaussian vector is obtained by shifting μ and a scaling Λ of a set of identically independently distributed (iid) standard normal distribution W. In general, the Gaussian vector depend on $d(d+3)/2$ parameters of μ and Σ. When the dimension d is large, the total number of parameters grows quadratically, but the distribution is intrinsically unimodal. This is a limitation of Gaussian vectors, when they need too many parameters but unable to provide a good approximation to multimodal distributions.

II.2 Gaussian processes

In this section, we reference the lecture notes [9] and the book [114]. A Gaussian process is a stochastic process that generalizes the Gaussian distribution. Conceptually, a Gaussian process can be thought of as a distribution over functions. Next, we will provide the definition of a Gaussian process.

Definition II.1

Let $(\Omega, \mathcal{A}, P)$ be a probability space and $\mathbb{T}$ is a parameter set. A stochastic process f indexed on a set $\mathbb{T}$ is a mapping of two variables

$$f : (\Omega, \mathcal{A}, P) \times \mathbb{T} \to \mathbb{R}$$

$$(\omega, t) \mapsto f(\omega, t).$$

We say that f is a Gaussian process if for any finite number of index $t_1, ..., t_n \in \mathbb{T}$, for $n \in \mathbb{N}$, the corresponding random vector $(f(w, t_1), ..., f(w, t_n))$ has the Gaussian distribution.

Similar to Gaussian distribution, a Gaussian process is completely determined by its mean function $m(t)$ and covariance function $K(t, t')$

$$m(t) = \mathbb{E}\left[f(t)\right], \tag{II.14}$$

$$K(t, t') = \mathbb{E}\left[(f(t) - m(t))(f(t') - m(t'))\right]. \tag{II.15}$$

For $\omega \in \Omega$ fixed, the function $f(\omega, t)$ depends on t only. This is a deterministic function, called a *sample path* or a *realization*. The underlying probability space will usually be ignored and we write $f(t)$ instead of $f(\omega, t)$. We denote the Gaussian process as $f \sim \mathcal{GP}(m(t), K(t, t'))$.

The index set $\mathbb{T}$ is usually the real line $\mathbb{R}$ or interval in $\mathbb{R}$, where $t \in \mathbb{T}$ is interptreted as time. It can also be a subset of $\mathbb{R}^d$ or an abstract set. [66] studies the case when $\mathbb{T}$ is the sigma algebra of a measure space (Wiener process), [88] studies the case when $\mathbb{T}$ is a separable Hilbert space, called isonormal Gaussian process. Recently, Gaussian process was generalized for the index set is the probability density functions [10, 44, 101] or a Riemannian manifold . In the following, we restrict ourselves to the case where $\mathbb{T}$ is a subset of $\mathbb{R}^d$.

For the general random process, it is hard to making inferences about its probability law from observing a single realization of the process. A common simplifying assumption is that the random process is stationary.

Definition II.2

A random process f is stationary if for all $t_1, ..., t_n \in \mathbb{T}$ and $h \in \mathbb{R}^d$, such that $t_1 + h, ..., t_n + h \in \mathbb{T}$, the finite distribution of f at $t_1, ..., t_n$ is the same as the finite distribution of f at $t_1 + h, ..., t_n + h$.

The covariance function K is said to be stationary if it only depends on $t - t'$, we write $K(t, t') = K(t - t')$ by an abuse of notation. The Proposition below gives the necessary and sufficient condition for a Gaussian process is stationary (the prove was given in [9]).

Proposition II.1

Let f be a Gaussian process on $\mathbb{T}$, then f is stationary if and only if its mean function is constant and its covariance function is stationary.

Proof Suppose $f \sim \mathcal{GP}(m(t), K(t, t'))$ is a Gaussian process with a mean function m is constant, and a covariance function K is stationary. Let $n, t_1, ..., t_n, h$ be as in the Definition II.2. Since f is a Gaussian process, $(f(t_1), ..., f(t_n))$ and $(f(t_1 + h), ..., f(t_n + h))$ are Gaussian vectors. Hence, their distributions are characterized by their mean vectors and covariance matrices. We will show that they are identical between the two Gaussian vectors. Indeed, we have $\mathbb{E}(f(t_i)) = \mathbb{E}(f(t_i + h))$ for $i = 1, ..., n$, since the mean function is constant. Hence the mean vectors are identical. We have also

$$\text{cov}(f(t_i), f(t_j)) = K(t_i - t_j) = K((t_i + h) - (t_j + h)) = \text{cov}(f(t_i + h), f(t_j + h)),$$

since the covariance function is stationary. Hence the two covariance matrices are identical.

For the reverse implication, let m and K be the mean function and covariance function. If there exist h so that $m(t + h) \neq m(t)$, then the two random vectors $f(t)$ and $f(t + h)$ do not have the same distribution. If there exist t_1, t_2, h so that $K(t_1, t_2) \neq K(t_1 + h, t_2 + h)$ then $(f(t_1), f(t_2))$ and $(f(t_1 + h), f(t_2 + h))$ do not have the same covariance matrix. ■

As stated in [114], there is no simple relationship between the covariance function of a Gaussian process and the smoothness of its realizations. However, it is possible to relate the covariance function and mean square continuity. The definition of mean square continuity is given as below.

Definition II.3

Let f be a stochastic process on $\mathbb{T} \subset \mathbb{R}^d$. We say that f is mean square continuous at $t_0 \in \mathbb{T}$ if

$$\lim_{t \to t_0} \mathbb{E}\left((f(t) - f(t_0))^2\right) = 0.$$

For a stationary Gaussian process f, we have

$$\mathbb{E}\left((f(t) - f(t_0))^2\right) = 2(K(0) - K(t - t_0)).$$

So f is mean square continuous at t_0 if and only if K is continuous at the origin, in this case f is mean square continuous everywhere. We say that f is mean square continuous if K is continuous at 0. The mean square continuity of f does not imply that its realizations are continuous. In considering on the probability space $(\Omega, \mathcal{A}, P)$, we have two other types of continuity: *Continuous sample paths with probability one* and *Almost surely continuous* [1]. We define the mean square differentiability, based on the definition of mean square continuous.

Definition II.4

A Gaussian process f on $\mathbb{T} \subset \mathbb{R}^d$ is mean square differentiable if there exist d Gaussian processes (defined on the same probability space $(\Omega, \mathcal{A}, P)$), $\frac{\partial f}{\partial t_1}, ..., \frac{\partial f}{\partial t_d}$, such that for $k = 1, ..., d$, for all $t_0 \in \mathbb{T}$, we have

$$\lim_{h \to 0} \mathbb{E}\left(\left(\frac{f(t_0 + he_k) - f(t_0)}{h} - \frac{\partial f(t_0)}{\partial t_k}\right)^2\right) = 0,$$

with $\{e_k\}_{k=1}^d$ is the canonical basis of $\mathbb{R}^d$.

By induction, we can define the mean square differentiable of higher order. A Gaussian process f is n times mean square differentiable if it is mean square differentiable and if the d Gaussian processes $\frac{\partial f}{\partial t_1}, ..., \frac{\partial f}{\partial t_d}$ are $n-1$ times mean square differentiable. We have the following result about the smoothness of f in the sense of mean square differentiable.

Proposition II.2

Let f be a Gaussian process on $\mathbb{T} \subset \mathbb{R}^d$ with mean function m and covariance function K. Then f is n times mean square differentiable if m is n times continuously differentiable and K is $2n$ times continuously differentiable.

The probability density of the finite-dimensional Gaussian vector $(f(t_1), ..., f(t_n))$ is

given by (II.1):

$$p_{t_1,...,t_n}(x_1,...,x_n) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T\Sigma^{-1}(\mathbf{x}-\mathbf{m})\right\},$$

where $\mathbf{x}^T = (x_1,...,x_n) \in \mathbb{R}^n$, $\mathbf{m}^T = (m(t_1),...,m(t_n)) \in \mathbb{R}^n$ is the mean vector, the covariance matrix Σ has the elements $\sigma_{ij} = K(t_i,t_j)$. The finite-dimensional cumulative distribution is given by:

$$\begin{aligned} F_{t_1,...,t_n}(r_1,...,r_n) &= P\left(f(t_1) \le r_1,...,f(t_n) \le r_n\right) \\ &= \int_{-\infty}^{r_1} ... \int_{-\infty}^{r_n} p_{t_1,...,t_n}(x_1,...,x_n)dx_1...dx_n. \end{aligned} \tag{II.16}$$

We say that cumulative distributions $G_{t_1,...,t_n}(r_1,...,r_n)$ satisfy the *symmetry condition* if

$$G_{t_1,...,t_n}(r_1,...,r_n) = G_{t_{\pi(1)},...,t_{\pi(n)}}\left(r_{\pi(1)},...,r_{\pi(n)}\right)$$

for any permutation π of the index set $\{1,...,n\}$. The distributions $G_{t_1,...,t_n}$ satisfy the *compatibility condition* if

$$G_{t_1,...,t_{n-1}}(r_1,...,r_{n-1}) = G_{t_1,...,t_n}(r_1,...,r_{n-1},\infty).$$

The finite cumulative distribution of a Gaussian process satisfies the two conditions.

Proposition II.3

The distribution functions defined as in (II.16) satisfy two consistency requirements: symmetry condition and compatibility condition.

Proof See section 1.4 of [1]. ■

The existence of Gaussian process is asserted by the Kolmogorov's existence theorem.

Theorem II.1

Let $t_1,...,t_n$ be arbitrary points in $\mathbb{T}$. If a system of finite-dimensional distributions $F_{t_1,...,t_n}$ satisfies the symmetry condition and compatibility condition, then there exists on some probability space $(\Omega,\mathcal{A},P)$ a random field $f(\omega,t), t \in \mathbb{T}$ having $F_{t_1,...,t_n}$ as its finite-dimensional distributions.

Proof See page 174 of [79]. ■

II.3 Prediction

The prediction of Gaussian processes rely on Gaussian conditioning property (II.9). Gaussian Process prediction is a Bayesian method, where the first thing is constructing a prior distribution, this is equivalent to choose the mean function and covariance function, and updating this distribution by conditioning on the data to get the posterior distribution. The posterior distribution is still a Gaussian process with a new updated mean function and covariance function. We will consider two cases, noise-free observations and noisy observations.

Let a Gaussian process $f(t) \sim \mathcal{GP}(0, K(t,t')), t \in \mathbb{T}$, with zero mean function and covariance function K. Suppose we have a noise-free observations $\{(t_i, f(t_i))|i = 1, ..., n\}$ of f. We want to predict the values at n_* test point $t'_1, ..., t'_{n_*}$. Denote $T = (t_1, ..., t_n)$ is the training points, $T_* = (t'_1, ..., t'_{n_*})$ is the test points, $F = (f(t_1), ..., f(t_n))^T$ is the training outputs, and $F_* = (f(t'_1), ..., f(t'_{n_*}))^T$ is the test outputs. By the definition of Gaussian process, the joint distribution is given by

$$\begin{pmatrix} F \\ F_* \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} K(T,T) & K(T,T_*) \\ K(T_*,T) & K(T_*,T_*) \end{pmatrix}\right), \tag{II.17}$$

where $K(T,T) = [K(t_i,t_j)], t_i, t_j \in T$ is the covariance matrix of size $n \times n$, and similar for $K(T,T_*)$, $K(T_*,T)$ and $K(T_*,T_*)$ where the components is the corresponding values of covariance function. Apply the conditioning property (II.9), we have posterior distribution

$$(F_*|T_*, F = y) \sim \mathcal{N}(K(T_*,T)K(T,T)^{-1}y, K(T_*,T_*) - K(T_*,T)K(T,T)^{-1}K(T,T_*)). \tag{II.18}$$

By placing the Gaussian process prior over a underlying unknown function f, we get not only the value for the predictive test output, but we get the full predictive distribution. These distributions provide approximations by the conditional means $\hat{F}_* = K(T_*,T)K(T,T)^{-1}y$, and confidence intervals by the conditional covariances $\text{cov}(F_*)$. When the test point is equal to one training point, we can prove that the predictive value is equal to the corresponding training output (if the covariance matrix $K(T,T)$ is invertible). This means that the predictive function interpolates the data.

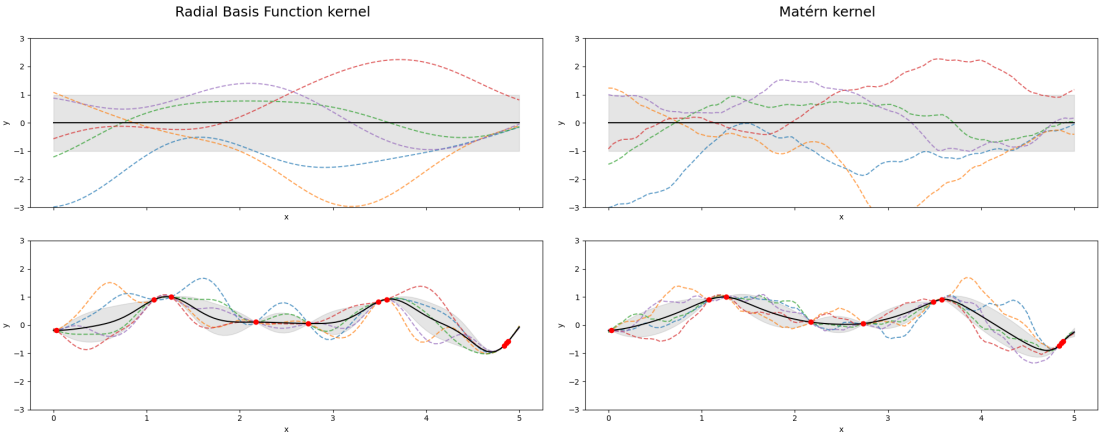


Figure II.1: The prior realizations of Gaussian processes (first row) and the posterior realizations. The black line is the graph of the predictive function in black line, the grey region is the $\pm$ standard deviation, the red points are the observations.

It is more realistic if we use the noisy observation model $y = f(t) + \epsilon$, where the noise is independent identically distributed Gaussian $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$. In this case, we have the joint distribution as

$$\begin{pmatrix} Y \\ F_* \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} K(T,T) + \sigma_n^2 \mathcal{I}_n & K(T,T_*) \\ K(T_*,T) & K(T_*,T_*) \end{pmatrix}\right), \tag{II.19}$$

where $Y = (f(t_1) + \epsilon_1, ..., f(t_n) + \epsilon_n)^T$, and $\mathcal{I}_n$ is the $n \times n$ identity matrix. Apply the conditioning theorem, we have $(F_*|T_*, Y = y) \sim \mathcal{N}(\hat{F}_*, \text{cov}(F_*))$, where

$$\hat{F}_* = K(T_*,T)\left(K(T,T) + \sigma_n^2 \mathcal{I}_n\right)^{-1} y, \tag{II.20}$$

$$\text{cov}(F_*) = K(T_*,T_*) - K(T_*,T)\left(K(T,T) + \sigma_n^2 \mathcal{I}_n\right)^{-1} K(T,T_*). \tag{II.21}$$

Figure II.1 shows examples of Gaussian prediction corresponding with Radial Basis function kernel (or Gaussian covariance) and Matérn kernel.

II.4 Covariance functions

A Gaussian process is characterized by its mean and covariance function. We usually assume that the mean function is identically zero. Hence the study of the covariance function remains important for the Gaussian process. We will consider the class of all possible stationary covariance functions. And by the Bochner's theorem, we can present the covariance functions with an unique spectral representation. Then we give the conditions for the existence of n times mean square differentiable of the process, and

the representation of its covariance function.

Definition II.5

Let n be a positive integer, and let $t_i \in \mathbb{T}$ and $c_i \in \mathbb{R}$ for $i = 1, ..., n$. Then the function K on $\mathbb{T} \times \mathbb{T}$ is said to be positive semi-definite on $\mathbb{T}$ if

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(t_i, t_j) \geq 0 \tag{II.22}$$

for all choice of n, $\{t_1, ..., t_n\}$ and $\{c_1, ..., c_n\}$. If K is stationary, we have

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(t_i - t_j) \geq 0.$$

We can easily show that any covariance function is positive semi-definite. We apply the Kolmogorov's existence theorem to prove the inverse. The arguments below follow B.V. Gnedenko [50].

Theorem II.2

The class of covariance functions coincide with the class of positive semi-definite functions.

Proof Let K be a covariance function and t_i, c_i like in Definition II.5. Then we have

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(t_i, t_j) = \text{var}\left(\sum_{i=1}^{n} c_i f(t_i)\right) \geq 0.$$

So any covariance function is a positive semi-definite function. We now prove the inverse: each positive semi-definite function K is the covariance function of some random field. The positive semi-definiteness of K ensures that any finite dimensional distribution of $(f(t_1), ..., f(t_n))$

$$p_{t_1,...,t_n}(x_1, ..., x_n) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}\right\},$$

where Σ has components $\sigma_{ij} = K(t_i, t_j)$, is a finite dimensional multivariate normal distribution. So like in Theorem II.1, we can apply Kolmogorov's existence theorem to show that the corresponding Gaussian process exists. ■

Corollary II.1

The correlation function of a random process f is defined as the function k on $\mathbb{T} \times \mathbb{T}$, where $k(t, t') = Cor(f(t), f(t'))$, representing the correlation between $f(t)$ and $f(t')$. Then, the class of correlation functions coincides with the class of positive semi-definite functions where $k(t, t) = 1$.

Proof The Corollary follows from the previous theorem and

$$k(t,t') = \frac{K(t,t')}{\sqrt{K(t,t)}\sqrt{K(t',t')}}.$$

■

We state here the well known Bochner theorem.

Theorem II.3. Bochner's Theorem

A complex-valued function K on $\mathbb{R}^d$ is positive semi-definite if and only if it is the Fourier transform of a finite nonnegative Borel measure μ on $\mathbb{R}^d$, i.e.

$$K(t) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-ix.t} d\mu(x), \quad t \in \mathbb{R}^d. \tag{II.23}$$

Proof See [132]. ■

Definition II.6

The measure μ defined in (II.23) is called the spectral measure or spectrum of the corresponding process f.

Now we discuss Mercer's theorem, which allows us to express the kernel in the series of eigenfunctions and eigenvalues of the integral operator

$$\mathcal{K}\phi = \int_{\mathbb{T}} K(x,\cdot)\phi(x)d\mu(x). \tag{II.24}$$

In general, there are an infinite number of eigenfunctions $\{\phi_i(x)\}_{i=1}^{\infty}$ and corresponding eigenvalues $\{\lambda\}_{i=1}^{\infty}$,

$$\mathcal{K}\phi_i = \lambda\phi_i. \tag{II.25}$$

Theorem II.4. Mercer's theorem

Let (Ω, μ) be a finite measure space and K be a kernel on Ω such that the integral operator $\mathcal{K}$ is positive definite, i.e.

$$\int_{\Omega\times\Omega} K(x,x')f(x)f(x')d\mu(x)d\mu(x') \geq 0, \quad \forall f \in \mathbb{L}_2(\Omega,\mu). \tag{II.26}$$

Let $\{\phi_i(x)\}_{i=1}^{\infty}$ be the normalized eigenfunctions of $\mathcal{K}$ associated with the eigenvalues $\{\lambda\}_{i=1}^{\infty}$. Then:

1. *the eigenvalues $\{\lambda\}_{i=1}^{\infty}$ are absolutely summable,*
2. *the equation*

$$K(x,x') = \sum_{i=1}^{\infty} \lambda_i\phi_i(x)\phi_i(x') \tag{II.27}$$

holds for μ^2-almost everywhere, where the series converges absolutely and uniformly μ^2-almost everywhere.

Proof See [78]. ■

With $\mathbb{T} \subset \mathbb{R}^d$, let $\hat{\mathcal{H}}$ be a subspace of $L^2(\Omega, \mathcal{A}, P)$ consisting of functions which can be represented as finite linear combinations of the form $\xi = \sum_{s\in S} c_s f(s)$ where c_s are complex coefficients and S is an arbitrary finite subset of $\mathbb{T}$. Then $\hat{\mathcal{H}}$ is a complex vector space. The inner product on $\hat{\mathcal{H}}$ is induced from $L^2(\Omega, \mathcal{A}, P)$. Namely, for $\xi = \sum_{s\in S_1} c_s f(s)$ and $\eta = \sum_{s\in S_2} d_s f(s)$,

$$(\xi, \eta) = \mathbb{E}\left(\xi\overline{\eta}\right) = \sum_{s_1\in S_1}\sum_{s_2\in S_2} c_{s_1}\overline{d_{s_2}}\mathbb{E}\left(f(s_1)f(s_2)\right).$$

Let $\mathcal{H} = \mathcal{H}(0, K)$ be the closure of the linear manifold $\hat{\mathcal{H}}$ with respect to this inner product. Here 0 in $\mathcal{H}(0, K)$ refers to the mean function of f, and K refers to its covariance function. Thus $\xi \in \mathcal{H}$ if one can find a Cauchy sequence $\xi_n \in \hat{\mathcal{H}}$ such that $\mathbb{E}|\xi - \xi_n|^2 \to 0$ as $n \to \infty$.

Definition II.7

The space $\mathcal{H}$ is called the Hilbert space generated by the random process f.

Similarly, define $\mathcal{L}(\mu)$ to be the closed linear manifold of $\hat{\mathcal{L}}(\mu)$ with

$$\hat{\mathcal{L}}(\mu) = \left\{\xi(\lambda) = \sum_{s\in S} c_s e^{i\lambda.s} | S \subset \mathbb{T}, c_s \in \mathbb{C}\right\}.$$

Let $\xi(\lambda) = \sum_{s_1\in S_1} c_{s_1} e^{i\lambda.s_1}$ and $\eta(\lambda) = \sum_{s_2\in S_2} d_{s_2} s e^{i\lambda.s_2}$ be in $\hat{\mathcal{L}}(\mu)$. The inner product of $\xi(\lambda)$ and $\eta(\lambda)$ is defined by

$$\left(\xi(\lambda), \eta(\lambda)\right)_\mu = \sum_{s_1\in S_1}\sum_{s_2\in S_2} c_{s_1}\overline{d_{s_2}} \int_{\mathbb{R}^d} e^{i\lambda.(s_1-s_2)}\mu(d\lambda),$$

with μ is spectral measure. For $\xi(\lambda)$ and $\eta(\lambda)$ in $\mathcal{L}(F)$ we define the inner product

$$(\xi, \eta)_\mu = \lim_{n\to\infty} \left(\xi_n(\lambda), \eta_n(\lambda)\right)_\mu,$$

where $\xi_n(\lambda) \to \xi(\lambda)$ and $\eta_n(\lambda) \to \eta(\lambda)$. If we identify $\sum_{s\in S} c_s f(s)$ with $\sum_{s\in S} c_s e^{i\lambda.s}$ and extend this correspondence to respective limits of such sums, we have:

Proposition II.4

The two Hilbert spaces $\mathcal{H}$ and $\mathcal{L}(\mu)$ are isometrically isomorphic.

Let us apply this correspondence to mean square differentiability of a stationary Gaussian

process f on $\mathbb{R}$. By Definition II.4 of mean square differentiability, to consider the convergence of $f_h(t) = \dfrac{f(t_0+h)-f(t_0)}{h}$ as $h \to 0$ in $\mathcal{H}$, it is equivalent to consider the convergence of $\tau_h = \dfrac{e^{i\lambda(t+h)} - e^{i\lambda t}}{h}$ as $h \to 0$ in $\mathcal{L}(\mu)$. Study the convergence of τ_h we get the theorem below.

Theorem II.5

Suppose f is a stationary Gaussian process on $\mathbb{R}$ with covariance function K.Then f is mean square differentiable if and only if $K''(0)$ exists and is finite. And, if f is mean square differentiable then f' has covariance function $-K''$.

Proof See section 2.6 of [114]. ■

By repeated application of the previous theorem, it follows that f is n-times mean square differentiable if and only if $K^{(2n)}(0)$ exists and is finite and, if so, the covariance of $f^{(n)}$ is $(-1)^n K^{(2n)}$.

Below are some examples of covariance functions where $\ell > 0$ and σ^2 are parameters:

covariance function	expression
exponential	$\sigma^2 \exp\left(-\frac{\lvert t\rvert}{\ell}\right)$
Matérn $\frac{3}{2}$	$\sigma^2\left(1+\sqrt{6}\frac{\lvert t\rvert}{\ell}\right)\exp\left(-\sqrt{6}\frac{\lvert t\rvert}{\ell}\right)$
Matérn $\frac{5}{2}$	$\sigma^2\left(1+\sqrt{10}\frac{\lvert t\rvert}{\ell}+\frac{10}{3}\frac{\lvert t\rvert^2}{\ell^2}\right)\exp\left(-\sqrt{10}\frac{\lvert t\rvert}{\ell}\right)$
Gaussian	$\sigma^2 e^{-\frac{t^2}{\ell^2}}$

The general Matérn covariance is given by (see [114])

$$K_{\ell,\nu}(t) = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(2\sqrt{\nu}\frac{|t|}{\ell}\right)^{\nu} K_\nu\left(2\sqrt{\nu}\frac{|t|}{\ell}\right), \tag{II.28}$$

with Γ the Gamma function and K_ν the modified Bessel function of second order. The covariance $K_{\ell,\nu}$ is Matérn (ℓ,ν) with ℓ the correlation length and ν the smoothness parameter.

In the regression and classification problems using a Gaussian process, we do not know its covariance function in many practical applications. Thus in order to turn Gaussian processes into powerful practical tools it is essential to develop methods that address the model selection problem. We first determine what is the type of covariance function that is more suitable from the context. Then we will use a statistical estimator to

estimate the parameter of the covariance function like the Maximum Likelihood method and Cross Validation method.

II.5 Manifold

In computer science, there are many datasets that reside on a manifold, a topological space that locally looks like an open set of the Euclidean space. For example $3D$ rotation matrices belong to the Lie group $SO(3)$ [56], normalized histograms belong to the unit sphere , the space of symmetric positive definite (SPD) matrices [91]. On the manifold, we can define the metric, called Riemannian manifold, and we can compute the distance between two points. The precise mathematical descriptions in applications are facilitated by the use of differential geometry that generalizes the Euclidean space. In this section we give briefly some definitions and notions about manifolds and differential geometry. There are many books on Differential geometry, we follow maily the books [57, 67].

Definition II.8

A manifold $\mathcal{M}$ of dimension d is a connected paracompact Hausdorff space for which every point $m \in \mathcal{M}$, there exists a neighborhood $\mathcal{O}_m$ of m that is homeomorphic to an open subset Ω of $\mathbb{R}^d$. The homeomorphism $\psi_m : \mathcal{O}_m \to \Omega$ is called a coordinate chart. An atlas is a family of charts $\{\mathcal{O}_a, \psi_a\}$, where a belongs to some index set A, such that $\{\mathcal{O}_a\}$ forms an open covering of $\mathcal{M}$.

For any chart $(\mathcal{O}_a, \psi_a)$, if $m \in \mathcal{O}_a$ and $\psi_a(m) = (x_1(m), ..., x_d(m))$ then $\mathcal{O}_a$ is called coordinate neighborhood of m, and $(x_1(m), ..., x_d(m))$ is called local coordinates of m. Having the definition of manifold, we can go further to define the differentiable structure.

Definition II.9

An atlas $\{\mathcal{O}_a, \psi_a\}, a \in A$, on a manifold is called differentiable if all the transition maps

$$\psi_b \psi_a^{-1} : \psi_a(\mathcal{O}_a \cap \mathcal{O}_b) \to \psi_b(\mathcal{O}_a \cap \mathcal{O}_b) \tag{II.29}$$

are differentiable of class C^∞. A chart is called compatible with a differentiable atlas if adding the chart to the atlas yields again a differentiable atlas. An atlas is called maximal if any chart compatible with it is already contained in

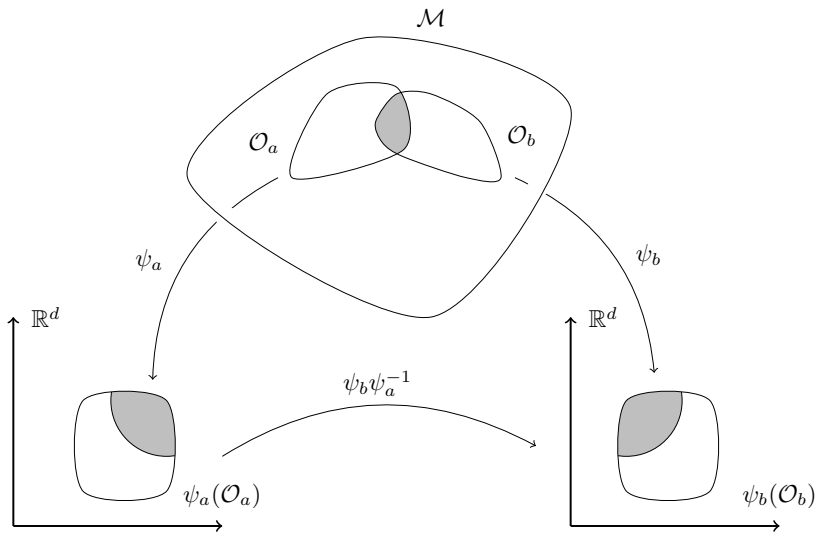


Figure II.2: Two charts and their transition maps.

it. A maximal differentiable atlas is called a differential structure. A differential manifold of dimension d is a manifold with a differentiable structure.

The maximal condition of differential structure is cumbersome to check. But it is not essential since any family of atlas can be extended in a unique way to satisfy the maximal condition. The differentiable structure allows us to define a differentiable map between manifolds.

Definition II.10

Let $\mathcal{M}$ and $\mathcal{N}$ be two differentiable manifolds with their corresponding atlases $\{\mathcal{O}_a, \psi_a\}$ and $\{\mathcal{Q}_b, \theta_b\}$. A map $h : \mathcal{M} \to \mathcal{N}$ is called differentiable if all the maps $\theta_b \circ h \circ \psi_a^{-1}$ are differentiable in the defined domain. In the special case, when $\mathcal{N}$ is $\mathbb{R}$ the differential map is called the differential function. The set of all differential functions is denoted by $C^\infty(\mathcal{M})$.

Furthermore, h is called a diffeomorphism if it is a bijection, and both h and its inverse h^{-1} are differentiable. Some manifolds usually have complex geometries. A tangent space of a manifold at given point gives an approximation of the manifold locally by a linear space. If the manifold $\mathcal{M}$ is embedded in some Euclidean space, the tangent space at m is the space of all tangent vectors at m. Where the tangent vector can be though as the velocity of a curve passing through m. In general manifold, we define the tangent

space by the equivalence class.

Definition II.11

Let $m \in \mathcal{M}$, and let $(\mathcal{O}_a, \psi_a)$ and $(\mathcal{O}_b, \psi_b)$ be two local charts of m. Define the equivalence relation of tangent vectors $v \in T_{\psi_a(m)}\mathbb{R}^d$ and $w \in T_{\psi_b(m)}\mathbb{R}^d$ as

$$(\psi_a, v) \sim (\psi_b, w) \iff w = d(\psi_b \circ \psi_a^{-1})v. \tag{II.30}$$

Define the tangent space to $\mathcal{M}$ at m as the space of equivalent classes (ψ_a, v), denoted by $T_m\mathcal{M}$. The space $T\mathcal{M}$ is defined as the disjoint union of tangent space $T_m\mathcal{M}$, for all $m \in \mathcal{M}$.

We note that $T_m\mathcal{M}$ is a vector space of dimension d. For any tangent vector $v \in T_m\mathcal{M}$, in a local coordinate $(\mathcal{O}_a, \psi_a)$ of m we can write as

$$v = \sum_{i=1}^{d} v^i \frac{\partial}{\partial x_i}, \tag{II.31}$$

where $v^i \in \mathbb{R}$, and $\frac{\partial}{\partial x_i} : h \mapsto \left(\frac{\partial h \circ \psi_a^{-1}}{\partial x_i}\right) \circ \psi_a$, for any $h \in C^\infty(\mathcal{M})$, $i = 1, ..., d$. We call m is the base point, and the set of vectors $\frac{\partial}{\partial x_i}$, for $i = 1, ..., d$, is a basis of the tangent space. The tangent vector v can be represented as a derivative at 0 of a differentiable curve γ that satisfies: γ is defined on a neighborhood of 0, $\gamma(0) = m$, and in coordinate $(\mathcal{O}_a, \psi_a)$ we have

$$\frac{d}{dt} x_i \circ \psi_a \left(\gamma(t)\right) = v_i, \quad i = 1, ..., d. \tag{II.32}$$

Let $\pi : T\mathcal{M} \to \mathcal{M}$ be the projection of the tangent vector into its base point. Then the triple $(T\mathcal{M}, \pi, \mathcal{M})$ is called the tangent bundle of $\mathcal{M}$. A smooth section of the tangent bundle is called a vector field. The space of all vector fields is denoted by $\mathfrak{X}(\mathcal{M})$. Furthermore, we can introduce the scalar product on the tangent space. That permits us to measure the lengths and the angles of tangent vectors. Then, we can evaluate the length of a differentiable curve by taking integration of the norm of its tangent vector.

Definition II.12

A Riemannian metric on a differentiable manifold $\mathcal{M}$ is given by a scalar product $\mathfrak{g}_m$ on each tangent space $T_m\mathcal{M}$, which depends smoothly on the base point m. A Riemannian manifold is a differential manifold $\mathcal{M}$ equipped with a Riemannian metric $\mathfrak{g}$, denote $(\mathcal{M}, \mathfrak{g})$.

In the definition, the metric $\mathfrak{g}$ depends smoothly on the base point means that for any two smooth vector fields V, W in $\mathfrak{X}(\mathcal{M})$, the function $\mathfrak{g}_m(V|_m, W|_m)$ is a smooth function of m. A Riemannian isometry between $(\mathcal{M}, \mathfrak{g}_\mathcal{M})$ and $(\mathcal{N}, \mathfrak{g}_\mathcal{N})$ is a diffeomorphism $h : \mathcal{M} \to \mathcal{N}$ such that the pullback metric $h^*\mathfrak{g}_\mathcal{N}$ is the same as $\mathfrak{g}_\mathcal{M}$, i.e,

$$\mathfrak{g}_\mathcal{N}(Dh(v), Dh(w)) = \mathfrak{g}_\mathcal{M}(v, w) \tag{II.33}$$

for all $v, w \in T_m\mathcal{M}$ and all $m \in \mathcal{M}$. Now let $\gamma : [c_1, c_2] \to \mathcal{M}$ be a smooth curve from a closed interval $[c_1, c_2] \subset \mathbb{R}$ into the Riemannian manifold $\mathcal{M}$. Then the length of γ is defined as

$$L(\gamma) := \int_{c_1}^{c_2} \left\| \frac{d\gamma(t)}{dt} \right\| dt, \tag{II.34}$$

where $\|\frac{d\gamma(t)}{dt}\| = \sqrt{\mathfrak{g}_{\gamma(t)}(\dot\gamma(t), \dot\gamma(t))}$ is the norm of tangent vector $\dot\gamma(t)$ in $T_{\gamma(t)}\mathcal{M}$. Here, $\dot\gamma(t)$ represents the derivative of γ with respect to t. Now, let's define the distance between two points on $\mathcal{M}$.

Definition II.13

Let $m, n \in (\mathcal{M}, \mathfrak{g})$. The distance between m and n is defined as the infimum length of the piecewise smooth curve connecting them

$$d(m, n) := \inf_{\gamma:[c_1,c_2]\to\mathcal{M}} \{L(\gamma) | \gamma \textit{ piecewise smooth curve}, \gamma(c_1) = m, \gamma(c_2) = n\}. \tag{II.35}$$

On a general manifold, the tangent space associated with different base points are different. We cannot define the derivative of a vector field as usual way by taking the limit of the ratio of two differences. Sine the difference of two tangent vectors at different base points is not well defined, because they belong to two different spaces. On manifolds, the affine connections provides us the rule to take the derivative of vector fields.

Definition II.14

An affine connection on a manifold $\mathcal{M}$ is a rule ∇ which assigns to each $V \in \mathfrak{X}(\mathcal{M})$ (first argument) a linear mapping ∇_V of the space $\mathfrak{X}(\mathcal{M})$ (second argument) into itself satisfying the following conditions:

1. *∇ is tensorial in first argument*

$$\nabla_{fV+hW} = f\nabla_V + h\nabla_W, \tag{II.36}$$

2. *∇ is linear in second argument, and satisfies the product rule*

$$\nabla_V(W+Z) = \nabla_V(W) + \nabla_V(Z), \tag{II.37}$$

$$\nabla_V(fW) = V(f)W + f\nabla_V(W), \tag{II.38}$$

for $f, h \in C^\infty(M)$, $V, W, Z \in \mathfrak{X}(M)$. The linear operator ∇_V is called covariant differentiation with respect to V.

The following lemma states the local property of the connection ∇.

Lemma II.1

Suppose $\mathcal{M}$ has the affine connection ∇. Let $\mathcal{O}$ be an open submanifold of $\mathcal{M}$. Let $V, W \in \mathfrak{X}(\mathcal{M})$. If V or W vanishes identically on $\mathcal{O}$, then so does $\nabla_V(W)$. Furthermore, if V vanishes at a point $m \in \mathcal{M}$, then so does $\nabla_V(W)$.

Proof See Section 4, Chapter 1 of [57]. ■

In local coordinate, there is a one-one relation between the affine connection and the Christoffel symbols $\Gamma^k_{i,j}$ that satisfy

$$\nabla_{\frac{\partial}{\partial x_i}} \frac{\partial}{\partial x_j} = \sum_k \Gamma^k_{i,j} \frac{\partial}{\partial x_k}. \tag{II.39}$$

The fundamental theorem of Riemannian Geometry determines a unique connection, called the Levi-Civita connection, ∇^{LC}.

Theorem II.6

On each Riemannian manifold $(\mathcal{M}, \mathfrak{g})$, there is uniquely one connection ∇^{LC} that satisfies

1. *∇^{LC} is torsion free: $\nabla^{LC}_V W - \nabla^{LC}_W V = VW - WV = [V, W]$,*
2. *∇^{LC} is metric: $\nabla^{LC}_V \mathfrak{g}(W, Z) = \mathfrak{g}(\nabla^{LC}_V W, Z) + \mathfrak{g}(W, \nabla^{LC}_V Z)$,*

where V, W and Z are vector fields.

Proof See [57]. ■

Let $\gamma : [c_1, c_2] \to \mathcal{M}$ be a curve in $\mathcal{M}$. We have the following definition of parallelism.

Definition II.15

Let $\gamma : [c_1, c_2] \to \mathcal{M}$ be a curve in $\mathcal{M}$, and let $V, W \in \mathfrak{X}(\mathcal{M})$ such that

$$V(t) = V_{\gamma(t)} = \dot{\gamma}(t). \tag{II.40}$$

Then, given an affine connection ∇ on $\mathcal{M}$, the family $W(t) = W_{\gamma(t)}$ is said to be parallel with respect to γ if

$$\nabla_V W|_{\gamma(t)} = 0, \quad \forall t \in [c_1, c_2]. \tag{II.41}$$

In the local chart $(\mathcal{O}_a, \psi_b)$ the vector fields V, W can be written by

$$V = \sum_i V^i \frac{\partial}{\partial x_i}, \quad W = \sum_i W^i \frac{\partial}{\partial x_i}, \tag{II.42}$$

where V^i, W^i are functions on $\mathcal{O}_a$. For simplicity, we write $x_i(t) = x_i(\psi(\gamma(t))$, $V^i(t) = V^i(\gamma(t)), W^i(t) = W^i(\gamma(t))$, and assume $\gamma([c_1, c_2]) \subset \mathcal{O}_a$. Then $V^i(t) = \dot{x}_i(t)$ and on the local coordinate $\mathcal{O}_a$ we have:

$$\nabla_V W = \sum_k \left(\sum_i V^i \frac{\partial W^k}{\partial x_i} + \sum_{i,j} V^i W^j \Gamma^k_{i,j} \right) \frac{\partial}{\partial x_k}.$$

So $W(t)$ is parallel with respect to γ if

$$\frac{dW^k}{dt} + \sum_{i,j} \Gamma^k_{i,j} \frac{dx_i}{dt} W^j = 0, \tag{II.43}$$

for all $k = 1, ..., d$. We say that the tangent vector $W(c_1)$ was parallel translated to $W(c_2)$, this depends also on the curve γ in general. By the parallelism, we can identify the tangent space of different base points.

Proposition II.5

Let m and n be two points in $\mathcal{M}$, and let γ be a curve segment from m to n. The corresponding parallel translation with respect to γ induces an isomorphism between $T_m\mathcal{M}$ and $T_n\mathcal{M}$.

Proof See Proposition 5.2. in [57]. ■

We can see that the equation involves V and W only through their values on the curve. The following definition of the geodesic depends on the connection through the parallelism.

Definition II.16

Let $\gamma : [c_1, c_2] \to \mathcal{M}$ be a curve in $\mathcal{M}$. The curve γ is called a geodesic if the family of tangent vector $\dot{\gamma}(t)$ is parallel with respect to γ. A geodesic γ is called maximal if it is not a proper restriction of any geodesic.

In a local coordinate neighborhood, the geodesic satisfies

$$\frac{d^2 x_k}{dt^2} + \sum_{i,j} \Gamma^k_{i,j} \frac{dx_i}{dt} \frac{dx_j}{dt} = 0, \quad k = 1, ..., d. \tag{II.44}$$

This means that the geodesic is a curve parallel to itself, that we call also autoparallel curve. Given the initial conditions (initial point and initial velocity), the geodesic is uniquely defined.

Proposition II.6

Let $\mathcal{M}$ be a differential manifold with an affine connection. Let $m \in \mathcal{M}$ and let $v \neq 0$ in the tangent space $T_m\mathcal{M}$. Then there exists a unique maximal geodesic γ on $\mathcal{M}$ such that

$$\gamma(0) = m, \quad \dot{\gamma}(0) = v. \tag{II.45}$$

The following theorem shows a topological relationship between the tangent space and the manifold.

Theorem II.7

Let $\mathcal{M}$ be a manifold with an affine connection, and let $m \in \mathcal{M}$. For any $v \in T_m\mathcal{M}$, let γ be the geodesic with $\gamma(0) = m$ and $\dot{\gamma}(0) = v$. Then there exists an open neighborhood $\mathcal{O}_0$ of 0 in the tangent space $T_m\mathcal{M}$ and an neighborhood $\mathcal{O}_m$ of m in $\mathcal{M}$ such that the mapping $v \mapsto \gamma(1)$ is a diffeomorphism between $\mathcal{O}_0$ onto $\mathcal{O}_m$.

This brings us to the definition of the exponential map and the log map.

Definition II.17

The mapping $v \mapsto \gamma(1)$ defined in the theorem is called the exponential mapping at m, denoted by $\exp_m$. Its inverse is called the logarithm, denoted by $\log_m$.

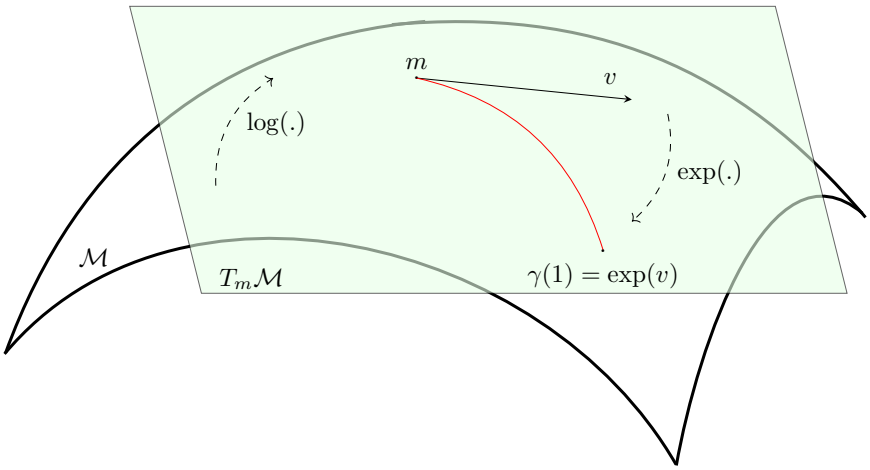


Figure II.3: Exponential map and logarithmic map.

II.6 The geometry of Normal distributions

In this section, we give an example of manifold to illustrate the previous section. We consider the statistical model of all Gaussian distributions

$$\mathcal{S} = \{p(x|\mu, \Sigma) = \mathcal{N}(\mu, \Sigma)| \quad \theta = (\mu, \Sigma) \in \mathbb{R}^d \times \mathbb{S}^d_{++}\}, \tag{II.46}$$

with the Fisher-Rao metric. This space was widely studied and has many applications [93]. Define the mapping $\psi : \mathcal{S} \to \mathbb{R}^{d'}, \quad d' = d(d+3)/2$, as

$$\psi(\mathcal{N}(\mu, \Sigma)) = \theta = ((\mu_i)_{i=1,...,d}, (\sigma_{ij})_{i\leq j}), \tag{II.47}$$

where $\mu = (\mu_1, ..., \mu_d)^T$ and $\Sigma = (\sigma_{ij})_{i,j=1,...,d}$. We see that ψ is a one-to-one map between $\mathcal{S}$ and a subset of $\mathbb{R}^{d'}$. Considering $(\mathcal{S}, \psi)$ as a global chart, so there is a corresponding differentiable structure on $\mathcal{S}$ where $(\mathcal{S}, \psi)$ is a coordinate system. This shows $\mathcal{S}$ is a differential manifold of dimension d'. On the coordinate system $(\mathcal{S}, \psi)$, we define a basis of the set of vector field $\mathfrak{X}(\mathcal{S})$ by

$$\frac{\partial}{\partial \mu_i}, i = 1, ..., d; \quad \frac{\partial}{\partial \sigma_{ij}}, i \leq j \leq d. \tag{II.48}$$

We then identify these basis vector fields with the vectors and symmetric matrix

$$\frac{\partial}{\partial \mu_i} \leftrightarrow e_i \in \mathbb{R}^d; \quad \frac{\partial}{\partial \sigma_{ij}} \leftrightarrow E_{ij}, i \leq j, \tag{II.49}$$

where $(e_i)_{i=1,...,d}$ is the canonical basis of $\mathbb{R}^d$, and E_{ij} is the basis of symmetric matrix space

$$E_{ij} = \begin{cases} \mathbb{1}_{(i,i)}, & i = j, \\ \mathbb{1}_{(i,j)} + \mathbb{1}_{(j,i)}, & i \neq j. \end{cases} \tag{II.50}$$

In which $\mathbb{1}_{(i,j)}$ is the $d \times d$ matrix with 1 in the (i, j) component and zero elsewhere. Any vector field $X \in \mathfrak{X}(\mathcal{S})$ can be decomposed as

$$X = \sum_{i=1}^{d} X^i e_i + \sum_{i \leq j} \bar{X}^{ij} E_{i,j}, \tag{II.51}$$

where $X^i, \bar{X}^{ij} : \mathcal{S} \to \mathbb{R}$ are smooth functions on $\mathcal{S}$.

A natural Riemannian structure on $\mathcal{S}$ can be provided by the Fisher information matrix

$$\mathfrak{g}(\theta) = [g_{ij}(\theta)] = Cov(\nabla \log \mathcal{N}(\mu, \Sigma)). \tag{II.52}$$

We have $\log \mathcal{N}(\mu, \Sigma) = -\frac{d}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| - \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)$. Suppose μ and Σ depend on θ, taking partial derivative with respect to θ_i we have

$$\begin{aligned} \frac{\partial \log \mathcal{N}(\mu, \Sigma)}{\partial \theta_i} =& -\frac{1}{2} \text{trace} \left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} \right) + \frac{1}{2}(x-\mu)^T \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} \Sigma^{-1}(x-\mu) \\ &+ \left(\frac{\partial \mu}{\partial \theta_i} \right)^T \frac{\partial}{\partial \mu_i} \Sigma^{-1}(x-\mu). \end{aligned} \tag{II.53}$$

By computing $\mathbb{E}\left(\frac{\partial \log \mathcal{N}(\mu,\Sigma)}{\partial \theta_i} \frac{\partial \log \mathcal{N}(\mu,\Sigma)}{\partial \theta_j} \right)$ directly ([94]), we get the closed formula

$$g_{ij}(\theta) = \frac{\partial \mu}{\partial \theta_i} \Sigma^{-1} \frac{\partial \mu}{\partial \theta_j} + \frac{1}{2} \text{trace}(\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j}). \tag{II.54}$$

In the basis of the vector field, we have

$$\begin{aligned} \mathfrak{g}\left(\frac{\partial}{\partial \mu_i}, \frac{\partial}{\partial \mu_j} \right) &= \mathfrak{g}(e_i, e_j) = e_i^T \Sigma^{-1} e_j = \sigma^{ij}, \quad i, j = 1, ..., d, \\ \mathfrak{g}\left(\frac{\partial}{\partial \mu_i}, \frac{\partial}{\partial \sigma_{kl}} \right) &= \mathfrak{g}(e_i, E_{kl}) = 0, \quad i, k, l = 1, ..., d, \\ \mathfrak{g}\left(\frac{\partial}{\partial \sigma_{ij}}, \frac{\partial}{\partial \sigma_{kl}} \right) &= \mathfrak{g}(E_{ij}, E_{kl}) = \frac{1}{2} \text{trace}(\Sigma^{-1} E_{ij} \Sigma^{-1} E_{kl}) \\ &= \sigma^{il} \sigma^{jk} + \sigma^{ik} \sigma^{jl}, \quad i, j, k, l = 1, ..., d. \end{aligned}$$

Let $X = \sum_{i=1}^{d} X^i e_i + \sum_{i \leq j} \bar{X}^{ij} E_{i,j}$ and $Y = \sum_{i=1}^{d} Y^i e_i + \sum_{i \leq j} \bar{Y}^{ij} E_{i,j}$ be tangent vectors at θ. Then be The inner product of X and Y is

$$\langle X, Y \rangle_\theta = X_\mu^T \Sigma^{-1} Y_\mu + \frac{1}{2} \text{trace} \left(\Sigma^{-1} X_\Sigma \Sigma^{-1} Y_\Sigma \right), \tag{II.55}$$

where $X_\mu = (X^1, ..., X^d)^T, Y_\mu = (Y^1, ..., Y^d)^T$ are the tangent vectors in $\mathbb{R}^d$ and $X_\Sigma = \sum_{i\leq j} \bar{X}^{ij} E_{i,j}$, $Y_\Sigma = \sum_{i\leq j} \bar{Y}^{ij} E_{i,j}$ are the symmetric matrices.

The space $\mathcal{S}$ with the Fisher-Rao metric is a Riemannian manifold, called Fisher-Rao Gaussian. The induced Riemannian geodesic distance $\rho_{\mathcal{N}}(\cdot, \cdot)$ is called Fisher-Rao distance:

$$d_{\mathcal{S}}(\mathcal{N}(\theta_1), \mathcal{N}(\theta 2)) = \inf_{\gamma}\{\text{Length}(\gamma) \mid \gamma(0) = \mathcal{N}(\theta_1), \gamma(1) = \mathcal{N}(\theta_2)\}, \tag{II.56}$$

where γ is the piecewise smooth curve connecting the two distributions, and the length is defined as

$$\text{Length}(\gamma) = \int_0^1 \sqrt{\langle \dot{\gamma}(t), \dot{\gamma}(t)\rangle_{\theta(t)}} dt. \tag{II.57}$$

The Christoffel symbols of the Levi-Civita connection ∇^{LC} were given in [107]. The corresponding geodesic curves $(\mu(t), \Sigma(t))$ satisfy the equations

$$\begin{cases} \dfrac{d^2\mu}{dt^2} - \left(\dfrac{d\Sigma}{dt}\right)\Sigma^{-1}\left(\dfrac{d\mu}{dt}\right) = 0 \\ \dfrac{d^2\Sigma}{dt^2} + \left(\dfrac{d\mu}{dt}\right)\left(\dfrac{d\mu}{dt}\right)^T - \left(\dfrac{d\Sigma}{dt}\right)\Sigma^{-1}\left(\dfrac{d\Sigma}{dt}\right) = 0. \end{cases} \tag{II.58}$$

The closed form for the geodesic and distance are not known in general. But they were explicitly given in the cases where μ is constant, Σ is constant, Σ is diagonal or in the one dimensional case.

II.6.1 The submanifold $\mathcal{S}_\Sigma$ where Σ is constant

The statistical manifold $\mathcal{S}_\Sigma = \{p(x|\mu, \Sigma) \mid \theta = (\mu), \Sigma = \Sigma_0 \text{ fixed}\}$ is a submanifold of $\mathcal{S}$ of dimension d. The Fisher information matrix is given by $\mathfrak{g}(\mu) = [g_{ij}(\mu)] \in \mathbb{R}^{d\times d}$ where

$$g_{ij}(\mu) = \frac{\partial\mu}{\partial\theta_i}\Sigma^{-1}\frac{\partial\mu}{\partial\theta_j}. \tag{II.59}$$

The geodesics and distance are given in [7, 93]. Let $\theta_0 = \mu_0$ and $\theta_1 = \mu_1$, then the geodesic curve $\gamma(t)$ in $\mathcal{S}_\Sigma$ connecting θ_0 and θ_1 is given by

$$\gamma(t) = ((1-t)\mu_0 - t\mu_1, \Sigma_0)\,. \tag{II.60}$$

The Fisher-Rao distance is given by

$$d_\Sigma(\theta_0, \theta_1) = \sqrt{(\mu_1 - \mu_0)^T\Sigma_0^{-1}(\mu_1 - \mu_0)}. \tag{II.61}$$

II.6.2 The submanifold $\mathcal{S}_\mu$ where μ is constant

The statistical manifold $\mathcal{S}_\mu = \{p(x|\mu,\Sigma) \mid \theta = (\Sigma), \mu = \mu_0 \text{ fixed}\}$ is of dimension $d(d+1)/2$. The Fisher information matrix in this case is given by $\mathfrak{g}(\Sigma) = [g_{ij}(\Sigma)]$ where

$$g_{ij}(\Sigma) = \frac{1}{2}\text{trace}(\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_i}\Sigma^{-1}\frac{\partial\Sigma}{\partial\theta_j}), \quad i,j = 1,...,d(d+1)/2. \tag{II.62}$$

The geodesics and distance has been studied in [87]. Let $\theta_0 = \Sigma_0$ and $\theta_1 = \Sigma_1$, then the geodesic curve is given by

$$\gamma(t) = \left(\mu_0, \Sigma_0^{1/2}\exp\left(t\log\left(\Sigma_0^{-1/2}\Sigma_1\Sigma_0^{-1/2}\right)\right)\Sigma_0^{1/2}\right). \tag{II.63}$$

The Fisher-Rao distance is given by

$$d_\mu(\theta_1,\theta_2) = \sqrt{\frac{1}{2}\sum_{i=1}^{d}\log^2(\lambda_i)}, \tag{II.64}$$

where $0 < \lambda_1 \leq ... \leq \lambda_d$ are the eigenvalues of $\Sigma_0^{-1/2}\Sigma_1\Sigma_0^{-1/2}$.

II.6.3 The one dimensional case

Now we consider the one dimensional case. The univariate Gaussian distribution

$$p(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-|x-\mu|^2}{2\sigma^2}\right) \tag{II.65}$$

is parametrized by the half upper plane of $\mathbb{R}^2$

$$H = \{(\mu,\sigma) \in \mathbb{R}^2 | \sigma > 0\}. \tag{II.66}$$

The information matrix is

$$[g_{ij}(\mu,\sigma)] = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}. \tag{II.67}$$

The Christoffel symbols are given by

$$[\Gamma^1_{ij}] = \begin{pmatrix} 0 & -\frac{1}{\sigma} \\ -\frac{1}{\sigma} & 0 \end{pmatrix}, \quad [\Gamma^2_{ij}] = \begin{pmatrix} \frac{1}{2\sigma} & 0 \\ 0 & -\frac{1}{\sigma} \end{pmatrix}. \tag{II.68}$$

Consequently, the geodesic equations are the following

$$\begin{cases} \frac{d^2\mu}{dt^2} - \frac{2}{\sigma}\frac{d\mu}{dt}\frac{d\sigma}{dt} = 0 \\ \frac{d^2\sigma}{dt^2} + \frac{1}{2\sigma}\left(\frac{d\mu}{dt}\right)^2 - \frac{1}{\sigma}\left(\frac{d\sigma}{dt}\right)^2 = 0. \end{cases} \tag{II.69}$$

To solve for this ODE system, we first separating and integrating the first equation

$$\frac{\ddot{\mu}}{\dot{\mu}} = \frac{2\dot{\sigma}}{\sigma} \leftrightarrow \frac{d}{dt}\ln\dot{\mu} = 2\frac{d}{ds}\ln\sigma \leftrightarrow \dot{\mu} = A\sigma^2, \tag{II.70}$$

where A is constant. There are two cases.

1. The case $A = 0$. It follows that $\mu =$ constant, which corresponds to vertical lines. The second equation of the system is reduced to

$$\frac{\ddot{\sigma}}{\dot{\sigma}} = \frac{\dot{\sigma}}{\sigma}. \tag{II.71}$$

Integrating we find $\sigma(t) = Be^{Ct}$, with B, C are constants. Hence the geodesics are

$$\begin{cases} \mu & = \mu_0, \\ \sigma(t) & = Be^{Ct}, \end{cases} \tag{II.72}$$

where μ_0, B, C are constants.

2. The case $A \neq 0$. Substituting $\dot{\mu} = A\sigma^2$ in second equation, we obtain

$$\sigma\ddot{\sigma} + \frac{A^2}{2}\sigma^4 - (\dot{\sigma})^2 = 0. \tag{II.73}$$

To solve for σ, we put $u = \dot{\sigma}$. Then the equation becomes

$$\sigma\frac{du}{d\sigma}u + \frac{A^2}{2}\sigma^4 - u^2 = 0. \tag{II.74}$$

Multiplying by the factor $1/\sigma^3$ leads to the exact differential equation

$$\frac{u}{\sigma^2}du + \left(\frac{A^2}{2}\sigma - \frac{u^2}{\sigma^3}\right)d\sigma = 0. \tag{II.75}$$

The solution is

$$\frac{u^2}{2\sigma^2} + \frac{A^2\sigma^2}{4} = \frac{E}{2}, \tag{II.76}$$

where E is positive constant. Replacing $u = \dot{\sigma}$ and solve for σ we get

$$\sigma = \sqrt{\frac{2E}{A^2}}\frac{1}{\cosh\left(\sqrt{E}(t+t_0)\right)}. \tag{II.77}$$

In order to solve for μ, we integrate $\dot{\mu} = A\sigma^2$ and obtain

$$\mu = \frac{2E}{A}\int \frac{1}{\cosh^2\left(\sqrt{E}(t+t_0)\right)}dt = \frac{2\sqrt{E}}{A}\tanh\left(\sqrt{E}(t+t_0)\right) + F \tag{II.78}$$

in conclusion, the solution in this case is

$$\begin{cases} \mu(t) & = \frac{2\sqrt{E}}{A}\tanh\left(\sqrt{E}(t+t_0)\right) + F, \\ \sigma(t) & = \frac{\sqrt{2E}}{|A|}\frac{1}{\cosh\left(\sqrt{E}(t+t_0)\right)}, \end{cases} \tag{II.79}$$

This satisfies the equation

$$\frac{A^2}{4E}(\mu(t) - F)^2 + \frac{A^2}{2E}\sigma(t)^2 = 1. \tag{II.80}$$

So in the plane (μ, σ) the geodesic is the ellipse with center $(F, 0)$, the width $4\sqrt{E}/|A|$ and the height $2\sqrt{2E}/|A|$. With the boundary conditions, we can find the values of the constants. We remark that the formula for geodesic on Poincaré half-plane H is known to Atkinson and Mitchell [7] and also Stoker [116]. But they use different system of coordinates. The geodesics for this model are circular arcs perpendicular to the real axis and straight vertical lines ending on the real axis.

Chapter III: Gaussian processes based on Classical Polynomial

In this chapter, we propose new data-driven statistical regression models with low complexity. First, we introduce new Gaussian processes where the covariance functions have explicit Mercer's representation. Second, we truncate the infinite sum of Karhunen-Loève expansion to employ it in regression models with low computational cost scaling: $O(nM^2)$ for inference and $O(M^3)$ for learning, instead of $O(n^3)$ for a canonical Gaussian process, where n is large in comparison to M ($n >> M$). Moreover, we develop an implementation that requires a negligible memory $O(M^2)$ instead of $O(nM)$. Finally, we demonstrate the robustness and the practical interest of the proposed methods with simulation and real studies. An extensive set of comparisons is explored to further investigate their efficiency against some state-of-the-art methods.

Organization. This chapter is organized as follows. Section .1 presents a general introduction. Section .2 provides background information on Gaussian processes regression. In Section .3, we discuss the low complexity Gaussian processes and highlight their main advantages in terms of computational complexity. Section .4 presents the proposed solutions for several differential operators with orthogonal polynomial bases. The experimental results are presented and discussed in Section .5. Finally, we provide a comprehensive discussion and conclusion in Section .6.

III.1 Introduction

Gaussian processes are powerful and flexible statistical models that have gained significant popularity in the field of econometrics, shape analysis, signal processing, data science, machine learning, etc [3, 43, 44, 98, 125]. However, modeling with Gaussian processes may also suffer from some computational challenges. When the number of observations n increases, the computational complexity for inference and learning grows significantly and incurs $O(n^3)$ computational cost which is unfeasible for many modern

problems [86]. Another limitation of Gaussian processes is the memory scaling $O(n^2)$ in a direct implementation. Significant efforts have been dedicated to the development of asymptotically efficient or approximate computational methods for modeling with Gaussian processes. Various approximations and scalable algorithms, such as sparse Gaussian processes [108] and variational inference [27], have been developed to make Gaussian processes applicable to larger dataset. The book [98] dedicated the whole chapter (Chapter 8) to describe a number of approximation methods.

Usually, certain approximations, as demonstrated in [30, 65], involve a sort of reduced-rank Gaussian processes that rely on approximating the covariance function. Most of these approximations typically reduce the complexity to $O(nM^2)$ and the storage to $O(nM)$ with $M << n$. For example, [134] addressed the computational challenge of working with large-scale dataset by approximating the covariance matrix, which is often required for computations involving kernel methods. In addition, [47] proposed a FFT-based method for stationary covariances as a technique that leverage the Fast Fourier Transform (FFT) to efficiently compute and manipulate covariance functions in the frequency domain. The link between state space models (SSM) and Gaussian processes inference has been explored by [109]. This could avoid the cubic complexity in time using Kalman filtering inference methods [71]. Recently, [110] presented a novel method for approximating covariance functions as an eigenfunction expansion of the Laplace operator defined on a compact domain. More recently, [52] introduced a reduced-rank algorithm for Gaussian processes regression with a numerical scheme.

In this chapter, we consider the Karhunen-Loève (K-L) expansion of a Gaussian process with many advantages over other low-rank compression techniques [49]. First, it allows us to represent a Gaussian process as a series of basis functions and random coefficients. By selecting a subset of the most significant basis functions according to the more important eigenvalues, the rank of the Gaussian process can be reduced. This is particularly useful when dealing with big data, as it can help alleviate computational and storage requirements. Second, the K-L decomposition can be particularly useful for modeling the noise component of a Gaussian process. By analyzing the eigenvalues corresponding to the eigenfunctions, one can identify the level of contribution of each

eigenfunction to the noise component. This information can aid in noise modeling, estimation, and separation from the clean Gaussian process. Finally, the K-L decomposition provides a natural framework for model selection and regularization in Gaussian process modeling. By truncating the decomposition to a subset of significant eigenfunctions, one can prevent overfitting. This regularization can improve the generalization capability of the Gaussian process and mitigate the impact of noise or irrelevant features.

The K-L expansion of a Gaussian process is the optimal representation in the $\mathbb{L}^2$-sense, but the K-L expansions are available only for Gaussian processes with some covariance functions [63]. Instead of solving difficult integral equations for eigenpairs, we aim to exploit differential operators with orthogonal polynomials acting as eigenfunctions in contrast to previous works on K-L expansions. This choice is crucial because polynomials are designed to be numerically stable and well-conditioned, which will lead to more accurate and stable computations, especially in the presence of round-off errors. Moreover, orthogonal polynomials often possess convenient integration and differentiation properties. These properties facilitate efficient calculations involving the interpolated functions, making them highly advantageous for applications that require hard computations. Overall, Gaussian processes decomposition with orthogonal polynomials provides numerical stability, faster convergence and accurate approximation [2]. Their use in Gaussian processes for machine learning has been virtually nonexistent. The most existing researches are only based on analysis of integral operators and numerical approximations for computing K-L expansions [49].

III.2 Canonical Gaussian processes regression

In this section, we remind Gaussian process prediction for convenience. A one-dimensional Gaussian process defined on an index set $\mathbb{T} \subseteq \mathbb{R}$ is a stochastic process in which the marginal variables for any finite set in $\mathbb{T}$ follows a Gaussian distribution. In a regression task, a nonparametric function f is assumed to be a realization of a stochastic Gaussian process whereas the likelihood term holds from observations corrupted by a

noise according to the canonical form

$$\begin{cases} y_i = f(t_i) + \epsilon_i; \quad i = 1, \dots, n \\ f \sim \mathcal{GP}(0, K(t,s)) \end{cases} \tag{III.1}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma_n^2)$ is a Gaussian noise. Given a training dataset $\mathcal{D} = (\mathbf{t}, \mathbf{y}) = (t_i, y_i)_{i=1}^n$, the posterior distribution over $\mathbf{f} = f(\mathbf{t}) = (f(t_1), \dots, f(t_n))^T$ is also Gaussian: $p(\mathbf{f}|\mathcal{D}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. From Bayes' rule, we state that the mean and the covariance posterior are expressed as

$$\boldsymbol{\mu} = \mathbf{K}(\mathbf{K} + \sigma_n^2 \mathcal{I}_n)^{-1}\mathbf{y}, \tag{III.2}$$

$$\boldsymbol{\Sigma} = \left(\mathbf{K}^{-1} + \frac{1}{\sigma_n^2}\mathcal{I}_n\right)^{-1}, \tag{III.3}$$

where $\mathbf{K} = [K(t_i, t_j)]_{i,j=1}^n$ is the prior covariance matrix and $\mathcal{I}_n$ is the $n \times n$ identity matrix. The predictive distribution at any test input $t_\star$ can be computed in closed-form as $f(t_\star)|\mathcal{D}, t_\star \sim \mathcal{N}\Big(\hat{f}_\star, \text{var}(f_\star)\Big)$, with

$$\hat{f}_\star = \mathbf{k}(t_\star)^T(\mathbf{K} + \sigma_n^2 \mathcal{I}_n)^{-1}\mathbf{y}, \tag{III.4}$$

$$\text{var}(f_\star) = K(t_\star, t_\star) - \mathbf{k}(t_\star)^T(\mathbf{K} + \sigma_n^2 \mathcal{I}_n)^{-1}\mathbf{k}(t_\star), \tag{III.5}$$

where $\mathbf{k}(t_\star) = [K(t_i, t_\star)]_{i=1}^n$.

The covariance function $K(\cdot, \cdot)$ usually depends on a set of hyperparameters, denoted by θ_k, that needs to be estimated from the training dataset. The log marginal likelihood for Gaussian process regression serves as an indicator of the degree to which the selected model accurately captures the observed patterns. The log marginal likelihood is typically used for model selection and optimization. Let $\Theta = (\theta_k, \sigma_n^2)$ denote the model hyperparameters then the log marginal likelihood $\log p(\mathbf{y}|\mathbf{t}, \Theta)$ is given by

$$l(\Theta) = -\frac{1}{2}\log|\mathbf{K} + \sigma_n^2 \mathcal{I}_n| - \frac{1}{2}\mathbf{y}^T(\mathbf{K} + \sigma_n^2 \mathcal{I}_n)^{-1}\mathbf{y} - \frac{n}{2}\log(2\pi). \tag{III.6}$$

Here, $|\cdot|$ denotes the determinant. The goal is to estimate the hyperparameter Θ that maximizes the log marginal likelihood. This can be achieved using different methods, such as gradient-based algorithm [12], where the gradient vector with respect to the

hyperparameter is

$$\frac{\partial l(\Theta)}{\partial \theta_k} = \frac{1}{2}\mathbf{y}^T(\mathbf{K}+\sigma_n^2\mathcal{I}_n)^{-1}\frac{\partial \mathbf{K}}{\partial \theta_k}(\mathbf{K}+\sigma_n^2\mathcal{I}_n)^{-1}\mathbf{y} - \frac{1}{2}\text{tr}\left((\mathbf{K}+\sigma_n^2\mathcal{I}_n)^{-1}\frac{\partial \mathbf{K}}{\partial \theta_k}\right), \quad \text{(III.7)}$$

$$\frac{\partial l(\Theta)}{\partial \sigma_n^2} = \frac{1}{2}\mathbf{y}^T(\mathbf{K}+\sigma_n^2\mathcal{I}_n)^{-1}(\mathbf{K}+\sigma_n^2\mathcal{I}_n)^{-1}\mathbf{y} - \frac{1}{2}\text{tr}\left((\mathbf{K}+\sigma_n^2\mathcal{I}_n)^{-1}\right). \quad \text{(III.8)}$$

The weakness of inferring the posterior mean, the mean prediction, or learning the hyper-parameters from the log marginal likelihood is the need to inverse the $n \times n$ Gram matrix $\mathbf{K} + \sigma_n^2\mathcal{I}_n$. This operation usually costs $O(n^3)$, which limits the applicability of standard Gaussian processes when the sample size n increases significantly. Furthermore, the memory requirements for Gaussian process regression scale with a computational complexity of $O(n^2)$.

III.3 Low complexity Gaussian processes

One of the main advantages of a Gaussian process is that it can be represented as a series expansion involving a complete set of deterministic basis functions with corresponding random coefficients. Let the inner product in $\mathbb{L}^2(\mathbb{T}, \rho)$ be

$$\left\langle \phi, \psi \right\rangle = \int_{\mathbb{T}} \phi(t)\psi(t)\rho(t)dt, \quad \text{(III.9)}$$

where $\rho(t)$ is a positive weight function such that $\int_{\mathbb{T}} \rho(t)dt < \infty$. Consider a linear integral operator $\mathcal{K} : \mathbb{L}^2(\mathbb{T}, \rho) \mapsto \mathbb{L}^2(\mathbb{T}, \rho)$ with kernel K, expressed in terms of the inner product, as

$$\mathcal{K}\phi = \int_{\mathbb{T}} K(\cdot, t)\phi(t)\rho(t)dt. \quad \text{(III.10)}$$

The following spectral theorem states the general result of an operator on a Hilbert space.

Theorem III.1. Spectral theorem

Let $\mathcal{H}$ be a separable infinite-dimensional Hilbert space, and let A be a compact self-adjoint operator on $\mathcal{H}$. Then there exists a sequence of real eigenvalues $\{\lambda_j\}$ with $\lambda_j \to 0$ as $j \to \infty$, and an orthonormal basis of $\{\phi_j\}$ of eigenvectors with $A\phi_j = \lambda_j\phi_j$ for all $j \geq 1$.

Proof See the book [38]. ■

In our case, the operator $\mathcal{K}$ is compact and self-adjoint with respect to the inner product defined in (III.9), since $\left\langle \mathcal{K}\phi, \psi \right\rangle = \left\langle \mathcal{K}\psi, \phi \right\rangle$, allowing us to apply the spectral

theorem for $\mathcal{H} = \mathbb{L}^2(\mathbb{T}, \rho)$. Consequently, there exists an orthonormal set of basis functions $\{\phi_j\}_{j=1}^{\infty}$ in the weighted space $\mathbb{L}^2(\mathbb{T}, \rho)$, that is,

$$\int_{\mathbb{T}} \phi_j(t)\phi_l(t)\rho(t)dt = \delta_{jl}, \tag{III.11}$$

and a set of real eigenvalues $\{\lambda_j\}_{j=1}^{\infty}$. If $\mathcal{K}$ is positive and bounded then it admits absolutely summable positive eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. According to Mercer's Theorem II.4, the covariance function has the series expansion

$$K(t, s) = \sum_{j=1}^{\infty} \lambda_j \phi_j(t)\phi_j(s). \tag{III.12}$$

The eigenvalues $\{\lambda_j\}_{j=1}^{\infty}$ and eigenfunctions $\{\phi_j\}_{j=1}^{\infty}$ can be obtained from the integral operator and the solution is provided by the Fredholm integral equation

$$\mathcal{K}\phi_j(t) = \lambda_j \phi_j(t), \quad \forall t \in \mathbb{T}. \tag{III.13}$$

Now, the Gaussian process $f \sim \mathcal{GP}(0, K(\cdot, \cdot))$ can be decomposed using a series of eigenfunctions and random coefficients, as described in Karhunen-Loève [130].

Theorem III.2. Karhunen-Loève (K-L)

Let f be a nonparametric function on $\mathbb{T}$ modeled with a Gaussian process of a covariance function $K(\cdot, \cdot)$. Then, for all $t \in \mathbb{T}$ the function f can be written as

$$f(t) = \sum_{j=1}^{\infty} a_j \phi_j(t), \quad \textit{with} \quad a_j \overset{ind}{\sim} \mathcal{N}\left(0, \lambda_j\right) \tag{III.14}$$

where $\{\lambda_j\}_{j=1}^{\infty}$ and $\{\phi_j\}_{j=1}^{\infty}$ are eigenvalues and eigenfunctions of the integral operator $\mathcal{K}$ defined in (III.10).

In order to avoid the inversion of the $n \times n$ matrix $\mathbf{K} + \sigma_n^2 \mathcal{I}_n$, we use the approximation scheme presented above and project the Gaussian process to a truncated set of M basis functions. The truncated version of f at an arbitrary order $M \in \mathbb{N}^*$ is given by

$$f_M(t) = \sum_{j=1}^{M} a_j \phi_j(t) \tag{III.15}$$

with an approximation error $e_M(t) = \sum_{j=M+1}^{\infty} a_j \phi_j(t)$. The canonical Gaussian process regression model adapted to the truncated Gaussian process becomes

$$\begin{cases} y_i = f_M(t_i) + \epsilon_i, \quad i = 1, \ldots, n, \\ f_M \sim \mathcal{GP}(0, K_M(t, s)), \end{cases} \tag{III.16}$$

where $K_M(t, s) = \mathbb{E}(f_M(t) f_M(s)) = \sum_{j=1}^{M} \lambda_j \phi_j(t)\phi_j(s)$. The following proposition

proves the convergence.

Proposition III.1

1) The approximation $K_M(\cdot,\cdot)$ converges uniformly to $K(\cdot,\cdot)$ when $M \to \infty$, i.e.,

$$\lim_{M\to\infty} \Big(\sup_{t,s\in\mathbb{T}} |K(t,s) - \sum_{j=1}^{M} \lambda_j \phi_j(t)\phi_j(s)| \Big) = 0 \tag{III.17}$$

2) The mean integrated squared error (MISE) of f_M tends to 0 as $M \to \infty$.

Proof The proof of 1) follows from Mercer's theorem [118], while here we solely present the proof of 2). The MISE of f_M also known as the $\mathbb{L}^2$ risk function is given by

$$\begin{aligned} MISE &= \mathbb{E}\Big(||f - f_M||^2_{\mathbb{L}^2}\Big) \\ &= \mathbb{E}\Big(||e_M||^2_{\mathbb{L}^2}\Big) \\ &= \mathbb{E}\Big(\int_{\mathbb{T}} (\sum_{j=M+1}^{\infty} a_j \phi_j(t))^2 dt\Big) \\ &= \mathbb{E}\Big(\sum_{j=M+1}^{\infty} a_j^2 \int_{\mathbb{T}} \phi_j(t)^2 dt\Big) \\ &= \mathbb{E}\Big(\sum_{j=M+1}^{\infty} a_j^2\Big) \\ &= \sum_{j=M+1}^{\infty} \lambda_j, \end{aligned} \tag{III.18}$$

which tends to 0 as $M \to \infty$ since λ_j are absolutely summable. ■

The convergence of the Mercer's decomposition depends hardly on the eigenvalues and the differentiability of the covariance function. [118] showed that the speed of the uniform convergence varies in terms of the decay rate of eigenvalues and demonstrated that for a 2β times differentiable covariance $K(\cdot,\cdot)$ the truncated covariance $K_M(\cdot,\cdot)$ approximates $K(\cdot,\cdot)$ as $O\Big((\sum_{j=M+1}^{\infty} \lambda_j)^{\frac{\beta}{\beta+1}}\Big)$. For infinitely differentiable covariances the latter is $O\Big((\sum_{j=M+1}^{\infty} \lambda_j)^{1-\varepsilon}\Big)$ for any $\varepsilon > 0$. To summarize, smoother covariance functions tend to exhibit faster convergence, while less smooth or non-differentiable covariance functions may exhibit slower or no convergence.

The resulting approximation fall into the class of reduced-rank approximations based on approximating the covariance matrix $\mathbf{K}$ with a matrix $\hat{\mathbf{K}} = [K_M(t_i,t_j)]_{i,j=1}^n = \mathbf{\Phi\Lambda\Phi}^T$, where $\mathbf{\Lambda}$ is a $M \times M$ diagonal matrix eigenvalues such that $\mathbf{\Lambda}_{jj} = \lambda_j$ and $\mathbf{\Phi}$

is a $n \times M$ matrix eigenfunctions such that $\mathbf{\Phi}_{ij} = \phi_j(t_i)$. Note that the approximate covariance matrix $\hat{\mathbf{K}}$ is ill-conditioned if λ_1/λ_M is large or if the observation points t_i are too closed to each other [23]. This lead to large numerical error when inverting $\hat{\mathbf{K}}$. By Theorem II.2, a bivariate function $K(\cdot,\cdot)$ is a covariance function if and only if it is positive semi-definite. The following proposition states that the truncation covariance $K_M(\cdot,\cdot)$ is well define a covariance function.

Proposition III.2

Let $M \in \mathbb{N}^*$ be the order of truncation. Let λ_j and ϕ_j be eigenvalues and eigenfunctions of the integral operator $\mathcal{K}$, for $j = 1, \ldots, M$. If $K(\cdot,\cdot)$ is positive semi-definite then $K_M(\cdot,\cdot)$ is also positive semi-definite.

Proof Let $N \in \mathbb{N}^*, \{t_1, \ldots, t_N\} \subset \mathbb{T}$ and $\{c_1, \ldots, c_N\} \in \mathbb{R}^N$ be as in Definition II.5. From (III.12), we have

$$\begin{aligned}\sum_{i=1}^{N}\sum_{l=1}^{N} c_i c_l K_M(t_i,t_l) &= \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{l=1}^{N} c_i c_l \lambda_j \phi_j(t_i)\phi_j(t_l) \\ &= \sum_{j=1}^{M}\lambda_j \sum_{i=1}^{N}\sum_{l=1}^{N} c_i c_l \phi_j(t_i)\phi_j(t_l) \\ &= \sum_{j=1}^{M}\lambda_j \left(\sum_{i=1}^{N} c_i \phi_j(t_i)\right)^2 \geq 0.\end{aligned}$$

In the above equality, we have used the fact that if $K(\cdot,\cdot)$ is positive semi-definite then all eigenvalues λ_j are nonnegative. ■

Now, we show how our novel regression model that utilizes Gaussian processes decomposition technique is able to achieve low complexity. We write down the expressions needed for both inference and hyperparameters learning and discuss the computational requirements. Applying the matrix inversion lemma [51] we re-rewrite the predictive distribution (III.4–III.5) as

$$\hat{f}_\star = \phi_\star^T(\mathbf{\Phi}^T\mathbf{\Phi} + \sigma_n^2\mathbf{\Lambda}^{-1})^{-1}\mathbf{\Phi}^T\mathbf{y} \tag{III.19}$$

$$\text{var}(f_\star) = \sigma_n^2\phi_\star^T(\mathbf{\Phi}^T\mathbf{\Phi} + \sigma_n^2\mathbf{\Lambda}^{-1})^{-1}\phi_\star \tag{III.20}$$

where $\phi_\star$ is an M-dimensional vector with the j-th entry being $\phi_j(t_\star)$. When the number of observations is higher than the number of required basis functions ($n >> M$) the use of this approximation is advantageous. Thus, any prediction mean evaluation is

dominated by the cost of constructing $\mathbf{\Phi}^T\mathbf{\Phi}$, which means that the method has an overall asymptotic computational complexity of $O(nM^2)$.

The approximate log marginal likelihood updated with the model (IV.8) satisfies

$$\begin{aligned} l(\Theta) &= -\frac{1}{2}\log|\mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^T + \sigma_n^2\mathcal{I}_n| - \frac{1}{2}\mathbf{y}^T(\mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^T + \sigma_n^2\mathcal{I}_n)^{-1}\mathbf{y} - \frac{n}{2}\log(2\pi) \\ &= -\frac{1}{2}(n-M)\log\sigma_n^2 - \frac{1}{2}\log|\mathbf{\Phi}^T\mathbf{\Phi} + \sigma_n^2\mathbf{\Lambda}^{-1}| - \frac{1}{2}\sum_{j=1}^{M}\log\lambda_j \\ &\quad - \frac{1}{2\sigma_n^2}(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi} + \sigma_n^2\mathbf{\Lambda}^{-1})^{-1}\mathbf{\Phi}^T\mathbf{y}) - \frac{n}{2}\log(2\pi) \end{aligned} \tag{III.21}$$

After the initial cost needed for inferring the prediction mean (III.19) evaluating the approximate log marginal likelihood has $O(M^3)$ complexity needed to inverse the $M \times M$ matrix $\mathbf{\Phi}^T\mathbf{\Phi} + \sigma_n^2\mathbf{\Lambda}^{-1}$. In practice, if the sample size n is large it is preferable to cache the result of $\mathbf{\Phi}^T\mathbf{\Phi}$ causing a memory requirement scaling as $O(M^2)$.

III.4 Explicit solutions for low complexity Gaussian processes

In this section, we describe explicit solutions of the low complexity Gaussian process (LCGP) with covariances derived from differential operators. In this chapter, we focus on the construction of covariance functions that incorporate orthogonal polynomials as eigenfunctions for two main reasons:

i) On the one hand, polynomials can approximate a wide range of functions with various degrees of complexity. They can be adjusted to predict different data patterns and can capture both linear and nonlinear relationships [25],

ii) On the other hand, polynomial regression is a well-established technique that extends linear regression by incorporating polynomial terms. It allows for more flexible modeling and can capture complex relationships between predictors and the response variable.

The connection between a differential operator denoted by $\mathcal{L}$ and the integral operator $\mathcal{K}$ has been largely used, see for example [39]. We follow the same idea and find the differential operator $\mathcal{L}$ such that the covariance function K plays the role of its Green's function

$$(\mathcal{L}K)(t,s) = \delta(t-s), \quad \text{for all } t,s \in \mathbb{T}, \tag{III.22}$$

where $\delta(\cdot)$ denotes the Dirac delta function. If $\{\lambda_j\}_{j=1}^{\infty}$ and $\{\phi_j\}_{j=1}^{\infty}$ refer to the eigenvalues and eigenfunctions of the integral operator $\mathcal{K}$, interchange integration and differentiation we have

$$\begin{aligned}\lambda_j \mathcal{L}\phi_j(t) &= \mathcal{L}\mathcal{K}\phi_j(t) \\ &= \int_{\mathbb{T}} \mathcal{L}K(t,s)\phi_j(s)\rho(s)ds \\ &= \int_{\mathbb{T}} \delta(t-s)\phi_j(s)\rho(s)ds \\ &= \phi_j(t)\rho(t).\end{aligned}$$

Finally, we get

$$\mathcal{L}\phi_j(t) = \frac{1}{\lambda_j}\phi_j(t)\rho(t). \tag{III.23}$$

This implies that $\left(\frac{1}{\lambda_j}, \phi_j\right)$ is eigenpair of $\rho^{-1}\mathcal{L}$, or the eigenvalues of $\mathcal{K}$ correspond to reciprocal eigenvalues of $\rho^{-1}\mathcal{L}$, while the corresponding eigenfunctions still the same [6, 53].

Now, suppose we have eigenvalues and normalized eigenfunctions of $\mathcal{L}$, denoted as (γ_j, ϕ_j), satisfying $\mathcal{L}\phi_j(t) = \gamma_j\phi_j(t)$. To incorporate the weight function ρ as in (III.23), we only need to consider $\rho\mathcal{L}$ instead of $\mathcal{L}$. This gives us the relationship:

$$\rho(t)\mathcal{L}\phi_j(t) = \gamma_j\phi_j(t)\rho(t). \tag{III.24}$$

Next, using the Mercer decomposition (III.12), we define $K(t,s) = \sum_{j=1}^{\infty}\gamma_j^{-1}\phi_j(t)\phi_j(s)$. Then, $K(t,s)$ is associated with $\rho\mathcal{L}$ as its Green function. This approach is applicable to a wide range of differential operators with corresponding integral operators that are positive and bounded. Detailed explanations are provided in the following sections.

III.4.1 Matérn covariance function

We choose one among the interesting operators on $\mathbb{L}^2([0,1])$, called the Matérn differential operator [16, 133], defined by

$$\mathcal{L}_{Ma} = \left(\varepsilon - \frac{d^2}{dt^2}\right)^{\alpha}, \tag{III.25}$$

depending on $\varepsilon \geq 0$ a scale parameter, and $\alpha \in \mathbb{N}$ a smoothness parameter. In which ε means ε times identity operator. Whittle [133] in 1963 discovered that, the Matérn covariance function is a unique stationary solution to (III.25) in the case of Euclidean

space $\mathbb{R}^d$. On the bounded domain $[0,1]$ with zero boundary conditions, we can verify (see [23]) that the corresponding eigenvalues and eigenfunctions of $\mathcal{L}_{Ma}$ are given by

$$\gamma_j = \left(\varepsilon + j^2\pi^2\right)^{\alpha}, \text{ and } \phi_j(t) = \sqrt{2}\sin(j\pi t). \tag{III.26}$$

By Mercer's theorem, we construct a covariance function as

$$K(t,s) = 2\sum_{j=1}^{\infty}\left(\varepsilon + j^2\pi^2\right)^{-\alpha}\sin(j\pi t)\sin(j\pi s). \tag{III.27}$$

Let $f \sim \mathcal{GP}(0, K(\cdot,\cdot))$, the K-L expansion of f is given by

$$f(t) = \sqrt{2}\sum_{j=1}^{\infty} a_j \sin(j\pi t), \quad a_j \sim \mathcal{N}(0, 1/\gamma_j). \tag{III.28}$$

Then we approximate f by

$$f_M(t) = \sqrt{2}\sum_{j=1}^{M} a_j \sin(j\pi t). \tag{III.29}$$

So, the approximation error is

$$e_M(t) = \sqrt{2}\sum_{j=M+1}^{\infty} a_j \sin(j\pi t). \tag{III.30}$$

We have the following proposition that shows the convergence of f_M.

Proposition III.3

If we approximate f by f_M (III.29) then the MISE of f_M tends to zero as M tends to infinity.

Proof Indeed, we have

$$\begin{aligned} MISE &= \mathbb{E}\|e_M\|^2_{\mathbb{L}^2(I)} = \sum_{j=M+1}^{\infty}\lambda_j \\ &= \sum_{j=M+1}^{\infty}\left(\varepsilon + j^2\pi^2\right)^{-\alpha}. \end{aligned} \tag{III.31}$$

Therefore, $MISE \to 0$ as $M \to \infty$. ■

III.4.2 Legendre Polynomials

We recall that the classic Legendre operator $\mathcal{L}_{Le}$ defined on $\mathbb{L}^2([-1,1])$ is given by

$$\mathcal{L}_{Le} = -(1-t^2)\frac{d^2}{dt^2} + 2t\frac{d}{dt}. \tag{III.32}$$

The eigenvalues are $\{\gamma_j = j(j+1)\}_{j=1}^{\infty}$ and eigenfunctions are Legendre polynomial $\{\phi_j(t) = P_j(t)/\|P_j\|_{\mathbb{L}^2}\}_{j=1}^{\infty}$ with

$$P_j(t) = \frac{1}{2^j j!}\frac{d^j}{dt^j}(t^2-1)^j, \tag{III.33}$$

and $\|P_j\|^2_{\mathbb{L}^2} = \frac{2}{2j+1}$. Next, we construct the covariance function

$$K(t,s) = \sum_{j=1}^{\infty} \frac{(2j+1)}{2j(j+1)} P_j(t)P_j(s). \tag{III.34}$$

We can check that $K(\cdot,\cdot)$ is square-integrable with the orthogonality of Legendre polynomials ([6, 33])

$$||K||^2_{\mathbb{L}^2(\mathbb{T}\times\mathbb{T})} = \sum_{j=1}^{\infty} \left[\frac{(2j+1)}{2j(j+1)}\right]^2 \frac{1}{(j+\frac{1}{2})^2} = \sum_{j=1}^{\infty} \left[\frac{1}{j(j+1)}\right]^2 < \infty.$$

Let $f \sim \mathcal{GP}(0, K(\cdot,\cdot))$, then $f(t) = \sum_{j=1}^{\infty} a_j\phi_j(t)$. We approximate f by

$$f_M(t) = \sum_{j=1}^{M} \sqrt{\frac{2j+1}{2}} a_j P_j(t), \tag{III.35}$$

with the approximation error

$$e_M(t) = \sum_{j=M+1}^{\infty} \sqrt{\frac{2j+1}{2}} a_j P_j(t). \tag{III.36}$$

We have the following proposition that shows the convergence of f_M.

Proposition III.4

If we approximate f by f_M (III.35) then the MISE of f_M tends to zero as M tends to infinity.

Proof We have

$$MISE = \mathbb{E}\|e_M(t)\|^2_{\mathbb{L}^2([-1,1])} = \sum_{j=M+1}^{\infty} \lambda_j = \sum_{j=M+1}^{\infty} \frac{1}{j(j+1)} = \frac{1}{M+1}. \tag{III.37}$$

From (III.37), $MISE \to 0$ as $M \to \infty$. ■

III.4.3 Laguerre Polynomials

As a second example, we consider the operator

$$\mathcal{L}_{La} = t\frac{d^2}{dt^2} + (1-t)\frac{d}{dt} \tag{III.38}$$

operating on $\mathbb{L}^2([0,\infty),\rho)$, where $\rho(t) = e^{-t}$ is the weight function. The operator $\mathcal{L}_{La}$ has eigenvalues $\gamma_j = -j$, and eigenfunctions from Laguerre polynomials $\{\phi_j = L_j(t)\}_{j=1}^{\infty}$ with

$$L_j(t) = \frac{e^t}{j!}\frac{d^j}{dt^j}\left(e^{-t}t^j\right).$$

We can check that $\{L_j\}_{j=1}^{\infty}$ is an orthonormal basis in $\mathbb{L}^2([0,\infty),\rho)$ [6].

Since $\mathcal{L}_{La}$ has negative eigenvalues, we consider the operator $\mathcal{L}^2_{La}$ with eigenvalue $\gamma_j = j^2$ and unchanged eigenfunction $\phi_j = L_j$. We construct the covariance function

defined by

$$K(t,s) = \sum_{j=1}^{\infty} \frac{1}{j^2} L_j(t) L_j(s). \tag{III.39}$$

We then approximate f by

$$f_M(t) = \sum_{j=1}^{M} a_j L_j(t). \tag{III.40}$$

Similarly, we have the following proposition that shows the convergence of f_M.

Proposition III.5

The MISE of f_M (III.40) tends to zero as M tends to infinity.

Proof We have

$$MISE = \mathbb{E}\|e_M(t)\|^2_{L^2([0,+\infty),e^{-t})} = \sum_{j=M+1}^{\infty} \lambda_j = \sum_{j=M+1}^{\infty} \frac{1}{j^2}. \tag{III.41}$$

From (III.41) we see that $MISE \to 0$ as $M \to \infty$. ■

III.4.4 Hermite Polynomials

In this example, we consider the operator

$$\mathcal{L}_{He} = \frac{d^2}{dt^2} - 2t\frac{d}{dt}, \tag{III.42}$$

defined on $\mathbb{L}^2(\mathbb{R}, \rho)$, for $\rho(t) = e^{-t^2}$. The operator $\mathcal{L}_{He}$ has eigenvalues $\gamma_j = -2j$ and eigenfunctions from Hermite polynomials $\{\phi_j = H_j(t)/\|H_j\|\}_{j=1}^{\infty}$, where

$$H_j(t) = (-1)^j e^{t^2} \frac{d^j}{dt^j} e^{-t^2}, \tag{III.43}$$

and $\|H_j\|^2_{\mathbb{L}^2(\mathbb{T},\rho)} = \sqrt{\pi} 2^j j!$. Like in Laguerre polynomial, we consider the operator $\mathcal{L}^2_{He}$ with eigenvalues $\gamma_j = (2j)^2$ and the same eigenfunctions. By Mercer's theorem, we construct the covariance function as

$$K(t,s) = \sum_{j=1}^{\infty} \frac{1}{\sqrt{\pi} 2^j (2j)^2 j!} H_j(t) H_j(s). \tag{III.44}$$

The truncated version of f is

$$f_M(t) = \sum_{j=1}^{M} \frac{1}{\sqrt{\sqrt{\pi} 2^j j!}} a_j H_j(t). \tag{III.45}$$

Similarly, we have the following proposition.

Proposition III.6

The MISE of f_M (III.45) tends to zero as M tends to infinity.

Proof Indeed,

$$MISE = \sum_{j=M+1}^{\infty} \frac{1}{(2j)^2} \to 0 \quad \text{as } M \to \infty. \tag{III.46}$$

■

III.4.5 Chebyschev Polynomials

Now we consider the operator

$$\mathcal{L}_{Ch} = (1-t^2)\frac{d^2}{dt^2} - t\frac{d}{dt}$$

acts on the weighted space $\mathbb{L}^2((-1,1),\rho)$, where $\rho(t) = \frac{1}{\sqrt{1-t^2}}$. The operator $\mathcal{L}_{Ch}$ has eigenvalues $\gamma_j = -j^2$ and eigenfunctions from Chebyshev polynomials $T_j(t) = \cos(j \arccos t)$ [33]. Furthermore, $\{T_j\}_{j=1}^{\infty}$ forms a sequence of orthogonal polynomials in $\mathbb{L}^2(\mathbb{T},\rho)$, and $\|T_j\|^2_{\mathbb{L}^2(\mathbb{T},\rho)} = \frac{\pi}{2}$. Let the normalized eigenfunction $\phi_j = T_j(t)/\|T_j\|$.

Sine γ_j is negative, we consider the operator $-\mathcal{L}_{Ch}$ with eigenvalues $\gamma_j = j^2$ and the same eigenfunctions. By Mercer's theorem, we construct the covariance function as

$$K(t,s) = \sum_{j=1}^{\infty} \frac{2}{\pi j^2} T_j(t) T_j(s). \tag{III.47}$$

Then, the truncated version of f is given by

$$f_M(t) = \sum_{j=1}^{M} \sqrt{\frac{2}{\pi}} a_j T_j(t). \tag{III.48}$$

Similarly, we have the following proposition.

Proposition III.7

The MISE of f_M (III.48) tends to zero as M tends to infinity.

Proof Indeed,

$$MISE = \sum_{j=M+1}^{\infty} \frac{1}{j^2} \to 0 \quad \text{as } M \to \infty. \tag{III.49}$$

■

III.4.6 Jacobi Polynomials

As the last example, we consider the differential operator

$$\mathcal{L}_{Ja} = (t^2-1)\frac{d^2}{dt^2} + (\alpha - \beta + (\alpha+\beta+2)t)\frac{d}{dt}, \tag{III.50}$$

where α and β are parameters. The operator $\mathcal{L}_{Ja}$ has eigenvalues $\gamma_j = j(j+\alpha+\beta+1)$ and eigenfunctions from Jacobi polynomials $J_j^{\alpha,\beta}(t)$. Jacobi polynomials are orthogonal in the space $\mathbb{L}^2((-1,1),\rho)$, where $\rho(t) = (1-t)^\alpha(1+t)^\beta$, and given by

$$J_j^{\alpha,\beta}(t) = \frac{(-1)^j}{2^j j!}(1-t)^{-\alpha}(1+t)^{-\beta}\frac{d^j}{dt^j}\left((1-t)^{j+\alpha}(1+t)^{j+\beta}\right). \tag{III.51}$$

The norm of $J_j^{\alpha,\beta}$ are given by

$$\|J_j^{\alpha,\beta}(t)\|_{\mathbb{L}^2}^2 = \int_{-1}^{1}\left(J_j^{\alpha,\beta}(t)\right)^2\rho(t)dt = \frac{2^{\alpha+\beta+1}\Gamma(j+\alpha+1)\Gamma(j+\beta+1)}{(2j+\alpha+\beta+1)j!\Gamma(j+\alpha+\beta+1)}. \tag{III.52}$$

As before, let the normalized eigenfunction of $\mathcal{L}_{Ja}$ be $\phi_j(t) = J_j^{\alpha,\beta}(t)/\|J_j^{\alpha,\beta}(t)\|$, and let covariance function

$$K(t,s) = \sum_{j=1}^{\infty}\frac{1}{j(j+\alpha+\beta+1)}\phi_j(t)\phi_j(s). \tag{III.53}$$

The truncated version of $f \sim \mathcal{GP}(0, K(\cdot,\cdot)$ is given by

$$f_M(t) = \sum_{j=1}^{M} a_j\phi_j(t) = \sum_{j=1}^{M}\frac{a_j J_j^{\alpha,\beta}(t)}{\|J_j^{\alpha,\beta}(t)\|}. \tag{III.54}$$

We can also check that the MISE of f_M tends to zero as M tends to infinity.

Covariance	Operator	Domain	ρ	γ_j	ϕ_j	$\|\|\phi_j\|\|_{\mathbb{L}^2}$
Matérn	$\mathcal{L}_{Ma}$	$[0,1]$	1	$\left(\varepsilon + j^2\pi^2\right)^\alpha$	$\sqrt{2}\sin(j\pi t)$	1
Legendre	$\mathcal{L}_{Le}$	$[-1,1]$	1	$j(j+1)$	$\frac{1}{2^j j!}\frac{d^j}{dt^j}(t^2-1)^j$	$\sqrt{\frac{2}{2j+1}}$
Laguerre	$\mathcal{L}_{La}^2$	$[0,\infty)$	e^{-t}	j^2	$\frac{e^t}{j!}\frac{d^j}{dt^j}\left(e^{-t}t^j\right)$	1
Hermite	$\mathcal{L}_{He}^2$	$\mathbb{R}$	e^{-t^2}	$4j^2$	$(-1)^j e^{t^2}\frac{d^j}{dt^j}e^{-t^2}$	$\sqrt{\sqrt{\pi}2^j j!}$
Chebyshev	$-\mathcal{L}_{Ch}$	$(-1,1)$	$(1-t^2)^{-1/2}$	j^2	$\cos(j\arccos t)$	$\sqrt{\frac{\pi}{2}}$
Jacobi	$\mathcal{L}_{Ja}$	$(-1,1)$	$(1-t)^\alpha(1+t)^\beta$	$j(j+\alpha+\beta+1)$	$J_j^{\alpha,\beta}(t)$	$\|J_j^{\alpha,\beta}(t)\|$

Table III.1: Different operators and their corresponding eigenpairs.

In Table III.1, we provide a summary for each class of differential operator, including the domain, the weight function ρ, the eigenvalues γ_j, the eigenfunctions ϕ_j, and their respective norms. Figure III.1 illustrates the behavior of the eigenvalues $\lambda_j = \frac{1}{\gamma_j}$ as the index j varies from 1 to 40. It is evident that the eigenvalues of all covariance functions converge to zero. Matérn and Jacobi, in particular, exhibit much faster convergence to zero compared to the other cases.

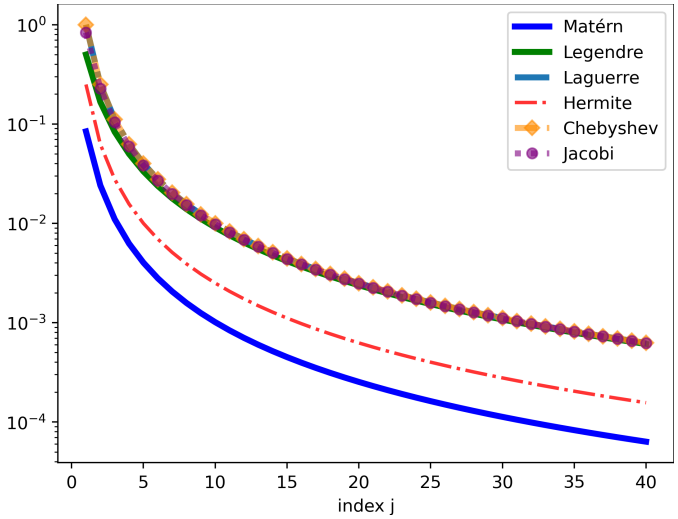


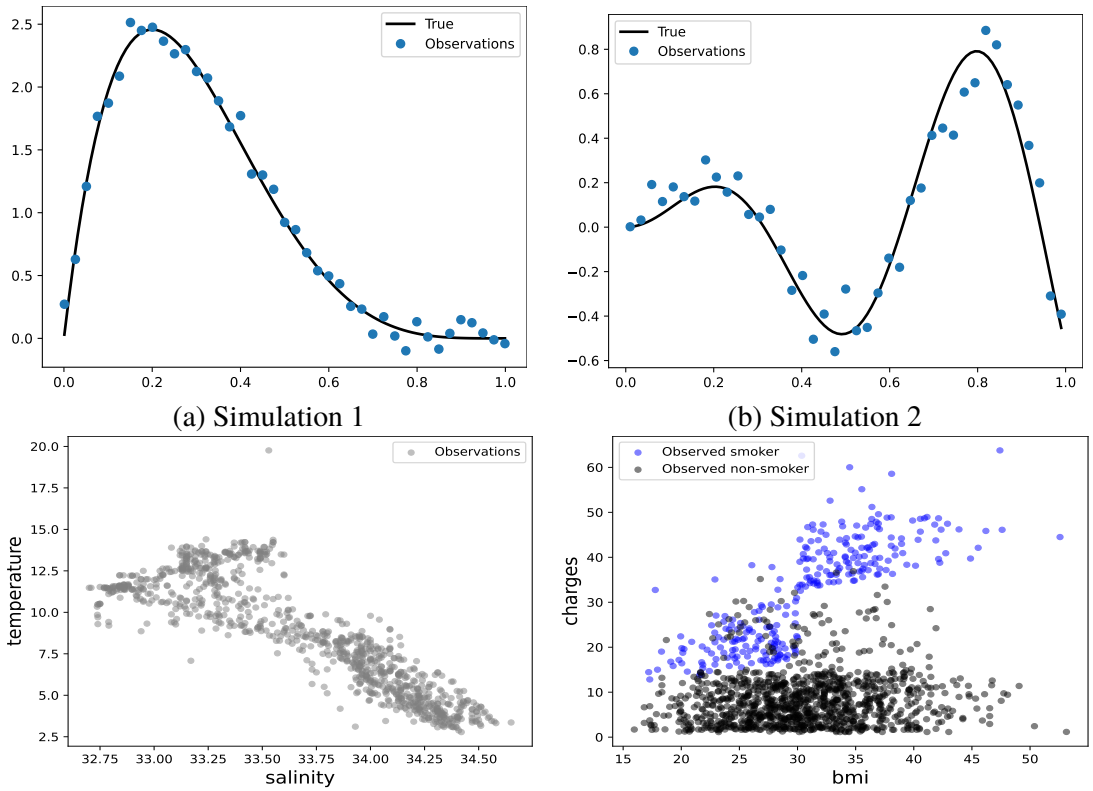
Figure III.1: The eigenvalues λ_j of different operators. Where we let: $\varepsilon = 2, \alpha = 1$ for Matérn, and $\alpha = -0.5, \beta = -0.3$ for Jacobi.

III.5 Experiments

In this section, we assess the effectiveness of the proposed methods by conducting evaluations on multiple dataset. We will compare their performance with some state-of-the-art methods. The comparative analysis will enable us to gain insights into the strengths and weaknesses of our approach and determine its competitiveness.

In Table III.1, differences in the domains for each covariance are observed. Through a change of variables, we can transform the basis to be defined on the open interval $\mathbb{T} = (0, 1)$. The details are provided in Table III.2, where the new basis function φ_j is the normalized function ϕ_j multiplied by the square root of the Jacobian of the transformation map. It's worth noting that we can choose other transformation maps; Table III.2 provides only explicit examples. For Laguerre and Hermite, the formula for φ_j is additionally multiplied by the square root of the weight function ρ to mitigate boundary effects, as the polynomial tends to infinity when the variable approaches infinity.

In this section, we use λ_j and φ_j from Table III.2 to create the covariance functions, and then we apply Gaussian process models as discussed earlier in our regression problems. We let $f_M(t)$ defined by $f_M(t) = \sum_{j=1}^{M} a_j \varphi_j(t)$, for $a_j \overset{\text{ind}}{\sim} \mathcal{N}\left(0, \lambda_j\right)$. Then $f \sim \mathcal{GP}(0, K_M(t, s)$, where $K_M(t, s) = \sum_{j=1}^{M} \lambda_j \varphi_j(t) \varphi_j(s)$. Sine all λ_j are positive,



(a) Simulation 1 (b) Simulation 2

(c) CalCOFI (d) MCP

Figure III.2: Some observations from different datasets.

$K_M(\cdot,\cdot)$ is positive semidefinite. Hence $f_M(t)$ well defines a Gaussian process on $\mathbb{T}$.

Covariance	Transformation map	$\frac{du}{dt}$	λ_j	φ_j
Matérn	$u(t) = t$	1	$\left(\varepsilon + j^2\pi^2\right)^{-\alpha}$	$\phi_j(t)$
Legendre	$u(t) = 2t - 1$	2	$\frac{1}{j(j+1)}$	$\sqrt{2j+1}\phi_j(u(t))$
Laguerre	$u(t) = \frac{t}{1-t}$	$\frac{1}{(1-t)^2}$	$\frac{1}{j^2}$	$\frac{e^{-u(t)/2}}{1-t}\phi_j(u(t))$
Hermite	$u(t) = \log\left(\frac{t}{1-t}\right)$	$\frac{1}{t(1-t)}$	$\frac{1}{4j^2}$	$\frac{e^{-u(t)^2/2}\phi_j(u(t))}{\sqrt{\sqrt{\pi}2^j j!}\sqrt{t(1-t)}}$
Chebyshev	$u(t) = 2t - 1$	2	$\frac{1}{j^2}$	$\frac{2\phi_j(u(t))}{\sqrt{\pi}}$
Jacobi	$u(t) = 2t - 1$	2	$\frac{1}{j(j+\alpha+\beta+1)}$	$\frac{\sqrt{2}}{\|J_j^{\alpha,\beta}(u(t))\|}J_j^{\alpha,\beta}(u(t))$

Table III.2: Transformation maps and new eigenpairs.

III.5.1 Data

Simulations. In this study, we examine two parametric functions: a beta density function represented by $f(t) = \mathcal{B}(t|a = 2, b = 5)$ (Simulation 1), and a quasi-periodic function satisfying $f(t) = t\sin(10t)$ (Simulation 2). Both functions are defined on the unit interval $(0, 1)$. For these experiments, we generated a total of 140 observations.

Out of these, we allocated 40 observations for training and the remaining for test. The inputs points t_i are uniformly distributed on $(0, 1)$. To introduce variability and simulate real-world conditions, each observed point was calculated as $y_i = f(t_i) + \epsilon_i$, where ϵ_i represents Gaussian noise drawn from $\mathcal{N}(0, \sigma_n^2 = 0.1)$. This procedure allows us to evaluate the performance of our models using noisy data. Figure III.2 (a)-(b) shows both true parametric functions and the noisy observations.

Real data. In this part, we conduct a real study using two challenging dataset. The first dataset comprises more than 864000 observations collected by the California Cooperative Oceanic Fisheries Investigationsm (CalCOFI). It investigates the ecological aspects surrounding the collapse of the sardine population off the coast of California, which is recognized as the longest and most comprehensive time series of oceanographic and larval fish data worldwide. It encompasses abundance data for over 250 fish species' larvae, as well as larval length frequency data, egg abundance data for important commercial species, and oceanographic data. Data collected at depths up to 500 meters includes: temperature, salinity, oxygen, phosphate, silicate, nitrate and nitrite, chlorophyll, phytoplankton biodiversity, etc. In this experiment, we are specifically targeting climate change indicators on the California coast when we keep 1000 observations among data illustrating the temperature (^oC) as function of the salinity (ppt). Some examples are given in Figure III.2 (c).

The second dataset used in this study pertains to Medical Cost Personal (MCP) and was sourced from demographic statistics provided by the US Census Bureau [113]. It primarily focuses on the cost of treatment, which is influenced by various factors, including age, sex, body mass index (BMI), and smoking status. Specifically, this chapter examines the relationship between treatment costs (charges in thousand dollars) and the BMI factor for both smokers and non-smokers. The dataset consists of 1338 observations, with 1064 corresponding to non-smokers and 274 pertaining to smokers, see Figure III.2 (d). For both real data a random split of 50% is allocated for training purposes, while the remaining 50% is set aside for evaluation. This partition ensures a balanced distribution of data for model training and comprehensive assessment of model performance.

Table III.3: Results of LCGP on Simulation 1.

Operator	MSE	R-squared	NLML
Matérn	3.11×10^{-3}	0.9962	13.67
Legendre	3.13×10^{-3}	0.9961	18.02
Laguerre	3.12	-2.81	20.48
Hermite	4.54×10^{-3}	0.9944	11.08
Chebyshev	3.47×10^{-3}	0.9957	15.11
Jacobi	3.12×10^{-3}	0.9961	20.71

Table III.4: Results of LCGP on Simulation 2.

Operator	MSE	R-squared	NLML
Matérn	7.82×10^{-3}	0.9426	8.79
Legendre	6.07×10^{-3}	0.9554	14.89
Laguerre	0.5	-2.67	12.47
Hermite	6.39×10^{-3}	0.9530	7.32
Chebyshev	5.69×10^{-3}	0.9582	12.49
Jacobi	5.73×10^{-3}	0.9579	12.85

Table III.5: Results of LCGP on CalCOFI data.

Operator	MSE	R-squared	NLML
Matérn	1.5186	0.8588	4635.63
Legendre	1.4988	0.8607	3755.24
Laguerre	2.41	0.7761	4017.89
Hermite	1.5095	0.8597	4430.93
Chebyshev	1.4991	0.8606	3744.62
Jacobi	1.4993	0.8606	3758.00

Table III.6: Results of LCGP on MCP data.

Operator	MSE	R-squared	NLML
Matérn	0.2354	0.3791	354.48
Legendre	0.2188	0.3958	329.57
Laguerre	0.2419	0.3686	338.93
Hermite	0.2548	0.3544	339.69
Chebyshev	0.2197	0.3943	327.84
Jacobi	0.2203	0.3938	329.68

III.5.2 Results

To evaluate the performance of the proposed methods some commonly used metrics for evaluating the performance of the regression model include:

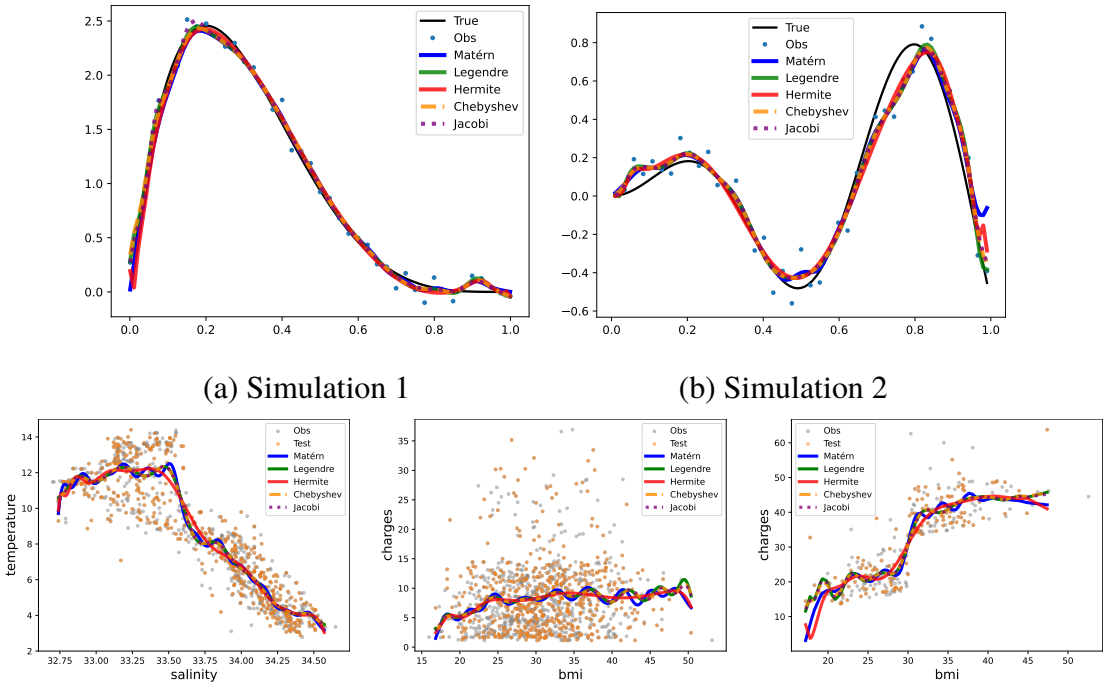
- MSE: the mean squared error as the average squared difference between the predicted values and the true values.
- R-squared: the coefficient of determination as a statistical measure used in regression analysis that represents the proportion of the variance in the dependent variable that is predictable from the independent variable.
- NLML: the negative log marginal likelihood which is a commonly used loss function defined as the negative of the average log marginal likelihood of the data given the model parameters.

It should be noted that a learning step should employed to determine the optimal hyperparameter for Matérn and Jacobi, whereas for other operators, only the noise variance estimation was necessary since the associated truncated covariance does not depend on any hyperparameter. But in this experiment study, we keep $\varepsilon = 2, \alpha = 1$ for Matérn and $\alpha = -0.5, \beta = -0.3$ for Jacobi.

Table III.3 and Table III.4 present the prediction results of the proposed method on simulations using different operators: Matérn, Legendre, Laguerre, Hermite, Chebyshev, and Jacobi. It is evident that all models can predict the function with a small mean squared error (MSE) except for Laguerre. The best performer for Simulation 1 is Matérn, while for Simulation 2, it is Chebyshev. The excellent performance of Matérn on simulated data can be attributed to its suitability for approximating parametric functions defined on the interval $(0, 1)$, as Matérn operators are specifically designed for this interval. However, according to the NLML criterion, Hermite outperforms others and is the best choice for both simulations.

Table A.1 and Table III.6 showcase the prediction results with real data. For the MCP data, we compute the average criteria (MSE, R-squared, NLML) for both smokers and non-smokers. Among the various operators, Legendre consistently outperforms others in real datasets. These results suggest that Legendre demonstrates greater flexibility in capturing various patterns and structures within real data, effectively modeling both

short-range and long-range dependencies. On the contrary, Chebyshev exhibits the smallest values of NLML in both experiments.



(a) Simulation 1 (b) Simulation 2

(c) CalCOFI (d) MCP: non-smokers (e) MCP: smokers

Figure III.3: Illustration of the prediction results with LCGP.

In Figure III.3 (a)-(b), we provide an illustration of predicting the true parametric function from simulations. In contrast, Figure III.3 (c)-(d)-(e) displays predictions from real datasets. Laguerre is not included in the figures for the sake of clarity. We observe that different types of polynomial eigenfunctions exhibit distinct advantages in prediction. We remind that Matérn and Hermite exhibit a boundary effect, where they become zero at the endpoints. They perform better when the true functions satisfies these conditions.

III.5.3 Comparison

We compare the proposed LCGP with several baseline methods to determine if there are significant performance differences. The baseline methods include: i) simple linear regression, ii) polynomial regression generating polynomial features from input data, iii) standard Gaussian process regression (as described in Section .2), and iv) neural network (NN) model. We provide some details about the NN architecture with multiple

Table III.7: Results of different methods on Simulation 1.

Method	MSE	R-squared
Linear regression	0.2871	0.6497
Polynomial regression	0.2571	0.6863
Standard GP	5.26×10^{-3}	0.9935
NN	0.2255	0.7248
LCGP	3.11×10^{-3}	0.9962

Table III.8: Results of different methods on Simulation 2.

Method	MSE	R-squared
Linear regression	0.1245	0.0862
Polynomial regression	0.1147	0.1578
Standard GP	8.12×10^{-3}	0.9403
NN	0.1153	0.1532
LCGP	5.69×10^{-3}	0.9582

Table III.9: Results of different methods on CalCOFI data.

Method	MSE	R-squared
Linear regression	2.4813	0.76940
Polynomial regression	1.8685	0.8263
Standard GP	1.4997	0.8606
NN	1.7866	0.8339
LCGP	1.4988	0.8607

Table III.10: Results of different methods on MCP data.

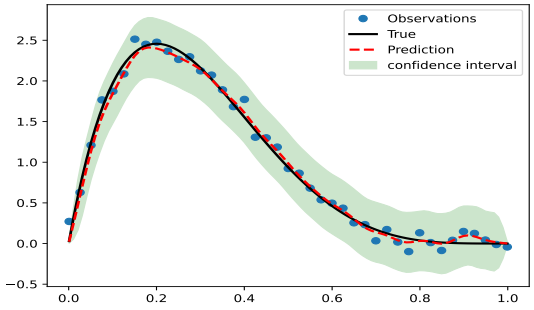
Method	MSE	R-squared
Linear regression	0.2752	0.3322
Polynomial regression	0.2816	0.3251
Standard GP	0.2294	0.3836
NN	0.2757	0.3319
LCGP	0.2188	0.3958

hidden layers. It consists of three hidden layers with 32, 64, and 64 units respectively, followed by an output layer with a single unit. The rectified linear unit (ReLU) activation function is used in the hidden layers to introduce non-linearity to the model. It is worth noting that Table III.7–III.8–III.9–III.10 represent results of standard Gaussian process with Matérn covariance.

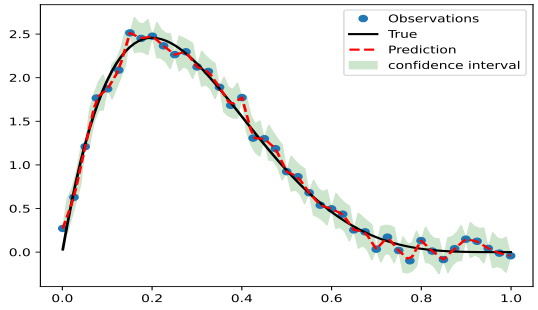
In Figure III.4 we illustrate the prediction results of the proposed method against the standard GP for simulated data (Simulation 1 and Simulation 2) as well as the baseline NN for real data (CalCOFI and MCP). The proposed method outperforms the baseline methods in terms of MSE and R-squared on both simulated and real dataset. For simulations study, the proposed method consistently achieves lower MSE values and higher R-squared scores, indicating its superior accuracy and predictive power. Furthermore, when applied to real-world dataset, the proposal demonstrates its ability to capture complex patterns and provide more precise predictions, leading to significantly improved MSE and R-squared values compared to the baseline methods. These results emphasize

the effectiveness and robustness of the proposed method in accurately modeling and predicting nonparametric regression tasks, making it a promising choice for various applications requiring high-performance regression models.

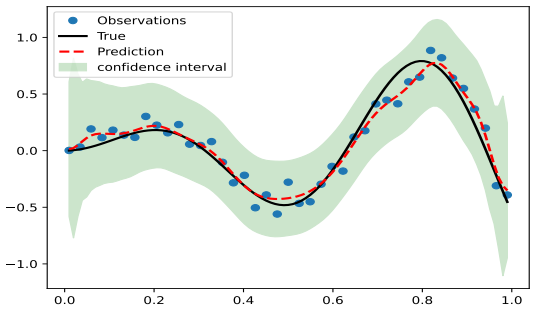
By focusing on the results of standard GP, particularly for real data, we stated that the shape parameter estimation ε is very large (around 10^5) which rends the prior covariance matrix $\mathbf{K}$ almost zeros and singular (ill-conditioned). This gives a zero mean prediction $\hat{f}_\star$ and therefore a significant noise variance estimation σ_n^2. A large noise variance leads to larger diagonal elements in the Gram matrix $\mathbf{K} + \sigma_n^2 \mathcal{I}_N$, as the noise variance contributes to the diagonal entries, i.e., $\mathbf{K} + \sigma_n^2 \mathcal{I}_N \approx \sigma_n^2 \mathcal{I}_N$. An ill-conditioned matrix causes numerical instability in the inversion process, potentially leading to inaccurate or unreliable predictions. It can make the model more sensitive to noise and result in overfitting. Fortunately, the LCGP overcomes this issue when reducing the rank of the standard GP. See Appendix A for more details.



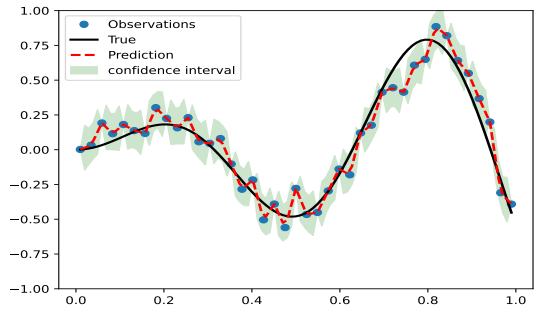
Simulation 1: LCGP



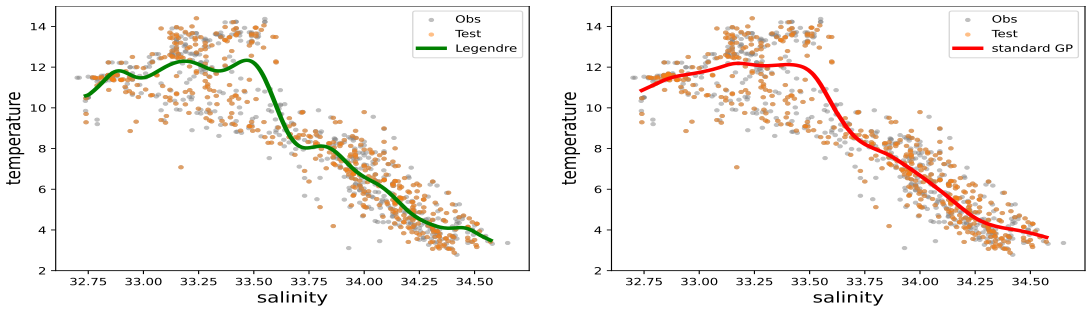
Simulation 1: Standard GP



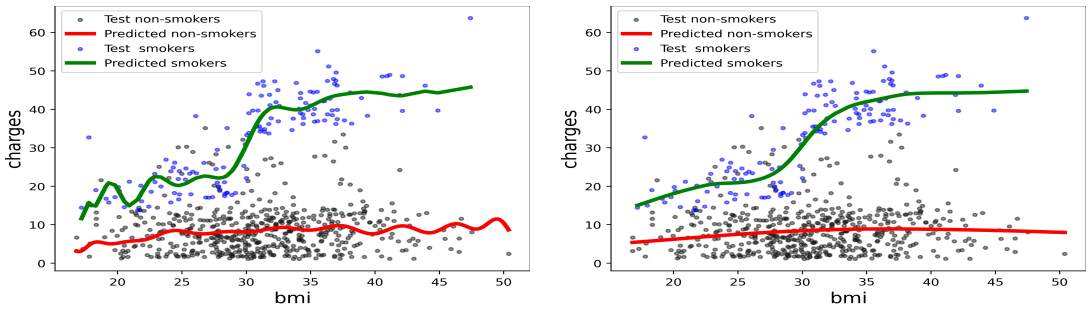
Simulation 2: LCGP



Simulation 2: Standard GP



CalCOFI: LCGP CalCOFI: Standard GP



MCP: LCGP MCP: Standard GP

Figure III.4: The prediction results of the proposed method (left) versus comparative methods (right).

III.6 Conclusion

In this chapter, we have proposed a new statistical regression model with a Gaussian process prior to infer, predict and learn nonparametric functions. The proposed methods have the ability to approximate some Gaussian processes using the eigen-expansion of some differential operators. We have showed different configurations where orthogonal polynomials are optimal bases with a simple implementations and closed-form expressions. To summarize, the main advantage consists in overcoming the limitations of standard Gaussian processes: a computational cost of $O(nM^2)$ for inference and $O(M^3)$ for learning. We have evaluated the proposed model with various simulated and real data. The experimental results and comparisons demonstrate high efficiency, low computational cost, and analytical simplicity against some existing methods.

Chapter IV: Constrained Gaussian processes to predict Probability density functions

In this chapter, we introduce a new framework based on a Gaussian process prior to learn, infer and predict nonparametric probability density functions. Our proposed method uses constrained Gaussian process to ensure that the output is a valid probability density function. In particular, the Gaussian process is approximated by truncating the Karhunen-Loève expansion, this allows us to incorporate both the knowledge of the data and the constraints. The formulation leads to constrained spherical Gaussian processes, which can be approximated by an efficient solution based on the spherical Hamiltonian Monte Carlo (HMC) sampling. We test and evaluate different strategies with extensive experiments on both simulations and real dataset.

Organization. This chapter is organized as follows. Section .1 presents a general introduction. In section .2, we introduce constrained Gaussian process models. Section .3 presents the posterior distribution of coefficients and algorithms to approximate them. Section .4 provides an example with Matérn covariance function on a bounded domain. We showcase various experiments in .5. Finally, we conclude this chapter at .6.

IV.1 Introduction

Over the last decades, probability density functions (PDFs) have become one of the main objects in information sciences and statistical modeling, with applications in many fields of science and engineering. For example, PDFs obtained from brain fMRI between voxels in a region of interest [92], warping functions in computer vision [112], income distributions [77], population pyramids for countries [31], etc. In practice, it is often not possible to directly observe the full PDF but only few values. In such situation, the main challenge would be predicting the PDF at unobserved points. In information sciences, there are many methods for predicting a function based on available observations. To cite but few, linear models, neural network, kernel method, see more details in [15]. We

should say that the problem at hand is different from the problem estimating a distribution from a data set. Where the later problem is very well known, and have been studied extensively [15, 35, 122–124].

Gaussian processes have been successfully used for both regression and classification problems [98]. In Gaussian process regression, the mean value at an unobserved point is predicted based on the conditional expectation from data (observed points) while the covariance provides the uncertainty interval, where both of them are explicit as we have seen in the previous chapters. Unfortunately, the standard Gaussian process regression is not straightforwardly applicable to the problem at hand since the underlying function may not be a PDF. Therefore, imposing hard conditions, related to the output function, on a Gaussian process model can lead to more realistic uncertainty quantification, but it requires further investigation. In this chapter, we introduce a new constrained Gaussian process model ensuring that the output is a PDF. Another challenging problem would be the nonexistence of explicit formulas for conditional expectation and covariance. We then overcome those limitations by considering sampling methods with constrained Gaussian process. Thus, depending on the constraints, we can use one among several sampling methods [84, 85, 117].

Constrained Gaussian process models have recently gained attention and have been applied to various goals [117]. In particular, a constrained Gaussian process model has been presented as a solution for sampling the multivariate normal distribution on some domains where the properties depend on the constraints. In [26], the authors consider a Gaussian process model with a set of inequality constraints for which the global constraints are approximated at a finite set of points. Hence, this formulation transforms the Gaussian process into a truncated multinormal distribution [119]. In [84, 85], the authors propose a finite-dimensional approximation of a Gaussian process using a basis of hat functions. Consequently, the constraints on the Gaussian process are directly translated to constraints on the coefficients and the output function satisfies the constraints everywhere in the index space. The coefficients are approximated by sampling from the truncated multi-normal distribution [83, 117]. Back to the problem on hand, the PDF constraints include that the integral is equal to one. This is a different type of

constraint that has not been addressed in previous works. To the best of our knowledge, the first work that modeled an unknown Square Root Density Function (SRDF) with a Riemannian structure has been introduced in [112]. More recently, [60] has used the Karhunen-Loève decomposition of the Gaussian process prior and used Hamilton Monte Carlo (HMC) [81] to sample from the posterior distribution. According to our case, we make use of the Spherical HMC [21].

Main contributions, we provide a unified framework based on a constrained Gaussian process prior for inferring PDFs manifold from observations such that the predicted solutions are inherently PDFs. We first model the SRDF with a Gaussian process prior giving a PDF modeled with a χ^2-process prior. One of the interesting properties of a Gaussian process is that it can be decomposed by Karhunen-Loève expansion as a series of random coefficients where the basis elements are eigenfunctions [30, 63]. From this decomposition, a Gaussian process regression model can be viewed as a linear regression model of an infinite number of covariates. By using the truncated version of a Gaussian process (finite-dimensional Gaussian process) the SRDF constraint is then maintained by the finite set of random coefficients. After conditioning on observations, the posterior distribution of the coefficients becomes a multivariate normal distribution that is restricted to the sphere. In some previous works, this distribution was called the Fisher-Bingham distribution and has been widely studied in directional statistics [69]. There are several ways to sample from this distribution, for example rejection sampling [75] and Gibbs sampler [59]. In this work, we consider a spherical HMC on embedded manifolds [21] for which the numerical solutions are tractable and efficient.

IV.2 Constrained Gaussian Processes

We have seen how the Gaussian processes are used in regression models. By using the Gaussian process prediction, the output function of the model does not need to satisfy any condition. Let $f(\cdot) \sim \mathcal{GP}(0, K(\cdot,\cdot))$ be a Gaussian process indexed on $\mathbb{T}$. Where we let the mean function $m(\cdot) = 0$ for convenience, and let denote covariance function as $K(\cdot,\cdot)$. For the rest of the chapter, $\mathbb{T}$ is assumed to be a subset of $\mathbb{R}$ and will be denoted by I.

IV.2.1 Gaussian Process Regression

We remind the standard case where f is not constrained as detailed in the previous Chapter. Assume that we have a finite set of noisy observations of f: $\mathcal{D} = (t_i, y_i)_{i=1}^n$, with inputs $t_i \in I$ and outputs $y_i = f(t_i) + \epsilon_i \in \mathbb{R}$, where $\epsilon_i \sim \mathcal{N}(0, \sigma_n^2)$ is a Gaussian noise. Given an unobserved input $t_\star \in I$, the predictive distribution of $f(t_\star)$ is Gaussian, denoted by $f(t_\star) \mid \{\mathcal{D}, t_\star\} \sim \mathcal{N}\Big(\hat{f}_\star, \text{var}(f_\star)\Big)$, with

$$\hat{f}_\star = \mathbf{k}(t_\star)^T (\mathbf{K} + \sigma^2 \mathcal{I}_n)^{-1} \mathbf{y}, \tag{IV.1}$$

$$\text{var}(f_\star) = K(t_\star, t_\star) - \mathbf{k}(t_\star)^T (\mathbf{K} + \sigma^2 \mathcal{I}_n)^{-1} \mathbf{k}(t_\star), \tag{IV.2}$$

where $\mathbf{K} = [K(t_i, t_j)]_{i,j=1}^n$ is the covariance matrix, $\mathbf{k}(t_\star) = [K(t_i, t_\star)]_{i=1}^n$, $\mathcal{I}_n$ is the $n \times n$ identity matrix and $\mathbf{y} = (y_1, \ldots, y_n)^T$. We predict $f(t_\star)$ to be its mean $\hat{f}_\star$ in (IV.1), whereas, $\text{var}(f_\star)$ in (IV.2) gives us the uncertainty interval. Gaussian process models can then be seen as function approximations based on kernels (covariance functions) [18, 40, 115, 127].

The Gaussian processes have a nice property that after conditioning on the data $\mathcal{D}$, the posterior is also Gaussian processes with explicit mean function and covariance function. But if we incorporate the constraints into the framework by conditioning both on $\mathcal{D}$ and the constraints, the posterior is no longer a Gaussian Process. So at the unseen point $t_\star$, we do not have the explicit formulas for the mean and covariance.

One way to control the output of the framework is to approximate the Gaussian process by a finite-dimensional Gaussian process. In [84, 85], the authors approximate a Gaussian process by a finite-dimensional Gaussian process with a basis of functions, and the coefficients follow normal distribution computed from the original Gausssian process. The benefit is that the constraints of the finite-dimensional Gaussian process are equivalent to the constraints of the random coefficients. This frameworks permit to introduce the constraints: boundedness, monotonicity, and convexity. The posterior distribution of the coefficients is the truncated multinormal distribution after a linear transformation. Then the truncated multinormal can be approximated using MCMC algorithms. Several methods of sampling were considered in [84], the authors concluded that HMC is an efficient sampler for the proposed framework.

In the following, we consider different constraints for the considered problem. We also approximate the Gaussian process, but by truncating the K-L expansion. The approximated finite-dimensional Gaussian has a basis functions derived from the eigen-functions, and the coefficients are normal distributions. The posterior distribution of the coefficients is normal distribution restricted on the sphere (Fisher-Bingham distribution). We then use HMC on the sphere to approximate the posterior distribution.

IV.2.2 The SRDF modeled with a finite-dimensional Gaussian process

A probability density function (PDF) p on I is a nonnegative function such that its integration on I equals to 1. We note $\mathcal{P}$ is the space of all PDFs on I such that

$$\mathcal{P} = \left\{ p : I \to \mathbb{R} \mid p(t) \geq 0, \forall t \in I, \text{ and } \int_I p(t)dt = 1 \right\}. \tag{IV.3}$$

The space $\mathcal{P}$ is a convex set but it is not a linear vector space. By imposing the Fisher-Rao metric, the space $\mathcal{P}$ becomes a Riemannian manifold. But working directly with PDFs is difficult due to the nonnegative and integral constraints [112]. Instead, we can work with other representations, for example: cumumlative distribution function [45], or log density function. In this chapter, we work with Square Root Density Function (SRDF) representation. By the definition SRDF is nonnegative, but we neglect the sign, and we denote the space of SRDFs by $\mathcal{S}$

$$\mathcal{S} = \left\{ q : I \to \mathbb{R} \middle| \int_I q^2(t)dt = 1 \right\}. \tag{IV.4}$$

So for each $p \in \mathcal{P}$ there are two function $q, -q \in \mathcal{S}$ such that $p = q^2 = (-q)^2$. The advantages of working with $\mathcal{S}$ are twofold. First we can release the sign constraint to consider the whole space $\mathcal{S}$, because we will take the square of SRDF in the final step. Second the geometry of $\mathcal{S}$ is the unit sphere in $\mathbb{L}^2$, which is more familiar and easier to work with than $\mathcal{P}$.

As we have seen that a Gaussian process can be decomposed by the Karhunen-Loève (K-L) expansion [30, 63]. If $f(\cdot) \sim \mathcal{GP}(0, K(\cdot,\cdot))$ then we can write

$$f(t) = \sum_{j=1}^{\infty} a_j \phi_j(t), \quad \text{with } a_j \overset{\text{ind}}{\sim} \mathcal{N}\Big(0, \lambda_j\Big), \tag{IV.5}$$

where $\lambda_j, \phi_j(t)$ are eigenvalue and normalized eigenfunction of the integral operator $\mathcal{K}$ with kernel K. The K-L expansion has many applications in Functional Data Analysis

(FDA) [11, 14, 92], Finance [106], and Machine Learning [76]. Along this chapter, the truncated version of f at an order M is given by

$$f_M(t) = \sum_{j=1}^{M} a_j \phi_j(t), \tag{IV.6}$$

with an approximation error $e_M(t) = \sum_{j=M+1}^{\infty} a_j \phi_j(t)$. Now let f_M is a SRDF. By (IV.6), $f_M \in \mathcal{S}$, or f_M^2 is a PDF, if and only if the coefficients $a_1, \ldots, a_M$ satisfy the spherical constraint $\sum_{j=1}^{M} a_j^2 = 1$, [45] .

Proposition IV.1

Let $f_M(t) = \sum_{j=1}^{M} a_j \phi_j(t)$, where $\{\phi_j(t)\}_{j=1}^{\infty}$ is an orthonormal basis of $\mathbb{L}^2(I)$ and $a_j \in \mathbb{R}$ for $j = 1, ..., M$. Then f_M^2 is a PDF if and only if $\sum_{j=1}^{M} a_j^2 = 1$.

Proof Suppose f_M^2 is a PDF, we have

$$\int_0^1 f_M^2(t)dt = \int_0^1 \left(\sum_{j=1}^{M} a_j \phi_j(t) \right)^2 dt = \sum_{j=1}^{M} a_j^2 = 1, \tag{IV.7}$$

where $\phi_1, ..., \phi_M$ is a subset of the orthonormal set $\{\phi_j(t)\}_{j=1}^{\infty}$. Conversely, if $\sum_{j=1}^{M} a_j^2 = 1$ then $f_M^2(t) \geq 0$ for all $t \in [0, 1]$, and $\int_0^1 f_M^2(t)dt = 1$ by (IV.7). This completes the proof. ■

When dealing with a truncated Gaussian process, the regression model on f_M substantially becomes

$$y_i = f_M(t_i) + \epsilon_i, \quad i = 1, ..., n. \tag{IV.8}$$

The covariance function of f_M is the truncated version of Mercer's representation of $K(\cdot, \cdot)$ with the same truncation order M

$$K_M(t, s) = \mathbb{E}[f_M(t) f_M(s)] = \sum_{j=1}^{M} \lambda_j \phi_j(t) \phi_j(s). \tag{IV.9}$$

Therefore, the covariance matrix depending on $K_M(\cdot, \cdot)$ is

$$\mathbf{K}_M = [K_M(t_i, t_j)]_{i,j=1}^{n} = \Phi \Lambda \Phi^T, \tag{IV.10}$$

where $\Phi = \begin{bmatrix} \phi_1(t_1) & \ldots & \phi_M(t_1) \\ \vdots & \ddots & \vdots \\ \phi_1(t_n) & \ldots & \phi_M(t_n) \end{bmatrix} \in \mathbb{R}^{n \times M}$ and $\Lambda = \begin{bmatrix} \lambda_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \lambda_M \end{bmatrix} \in \mathbb{R}^{M \times M}$.

The proposition of the previous chapter shows that $K_M(\cdot, \cdot)$ in (IV.9) is positive

semi-definite. We remind that the covariance matrix $\mathbf{K}_M$ in (IV.10) is ill-conditioned if the ratio λ_1/λ_M is large or if the observation points t_j are too close to each other [23]. This lead to many numerical errors when we try to take the inverse matrix in both (IV.1) and (IV.2).

IV.3 Posterior Distribution

In this section, we give details of a SRDF modeled with a Gaussian process in a regression framework. We suppose that we have a set of corresponding eigenvalues $\{\lambda_j\}_{j=1}^{\infty}$ and orthonormal eigenfunctions $\{\phi_j(t)\}_{j=1}^{\infty}$ in an explicit form. Then, the constrained Gaussian process model is

$$\begin{cases} y_i = f_M(t_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_n^2), \quad i = 1, \dots, n, \\ \int_0^1 f_M(t)^2 dt = 1, \end{cases} \tag{IV.11}$$

These equations are directly translated into random coefficients as

$$\begin{cases} \mathbf{y} = \Phi A + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_n^2 \mathcal{I}_n), \\ \|A\|_2^2 = 1, \end{cases} \tag{IV.12}$$

where $\mathbf{y} = (y_1, ..., y_n)^T$, $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T$ and $A = (a_1, ..., a_M)^T$. Since $a_j \overset{\text{ind}}{\sim} \mathcal{N}\left(0, \lambda_j\right)$ then we have $A \sim \mathcal{N}(0, \Lambda)$ where $\Lambda = \begin{bmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_M \end{bmatrix}$. By the property of jointly multivariate normal [84, 98], we have $A|\{\Phi A + \boldsymbol{\epsilon} = \mathbf{y}\} \sim \mathcal{N}(\mu, \Sigma)$, where

$$\mu = \Lambda \Phi^T [\mathbf{K}_M + \sigma_n^2 \mathcal{I}_n]^{-1} \mathbf{y}, \quad \text{and} \quad \Sigma = \Lambda - \Lambda \Phi^T [\mathbf{K}_M + \sigma_n^2 \mathcal{I}_n]^{-1} \Phi \Lambda, \tag{IV.13}$$

in which $\mathbf{K}_M = \Phi \Lambda \Phi^T$ by (IV.10). On the other hand, we can view the resulting model as a standard linear regression with basis $\phi_1(t), \dots, \phi_M(t)$ and parameters $A = (a_1, \dots, a_M)$.

Proposition IV.2

The posterior distribution of the coefficients A is given by

$$A|\{\Phi A + \boldsymbol{\epsilon} = \mathbf{y}\} \sim \mathcal{N}\left(\frac{1}{\sigma_n^2} \left(\frac{1}{\sigma_n^2} \Phi^T \Phi + \Lambda^{-1} \right)^{-1} \Phi^T y, \left(\frac{1}{\sigma_n^2} \Phi^T \Phi + \Lambda^{-1} \right)^{-1} \right). \tag{IV.14}$$

Proof Indeed, we only need to prove the mean and the covariance matrix of the two distributions in (IV.13) and (V.8) are identical. Hence, we have

$$\begin{aligned}\frac{1}{\sigma_n^2}\Phi^T\left(\mathbf{K}_M+\sigma_n^2\mathcal{I}_n\right) &= \frac{1}{\sigma_n^2}\Phi^T\left(\Phi\Lambda\Phi^T+\sigma_n^2\mathcal{I}_n\right)\\ &= \left(\frac{1}{\sigma_n^2}\Phi^T\Phi+\Lambda^{-1}\right)\Lambda\Phi^T,\end{aligned} \tag{IV.15}$$

by definition of $\mathbf{K}_M$. Multiplying (IV.15) by $\left(\frac{1}{\sigma_n^2}\Phi^T\Phi+\Lambda^{-1}\right)^{-1}$ from the left and by $(\mathbf{K}_M+\sigma_n^2\mathcal{I}_n)^{-1}$ from the right we get

$$\frac{1}{\sigma_n^2}\left(\frac{1}{\sigma_n^2}\Phi^T\Phi+\Lambda^{-1}\right)^{-1}\Phi^T=\Lambda\Phi^T\left(\mathbf{K}_M+\sigma_n^2\mathcal{I}_n\right)^{-1}. \tag{IV.16}$$

which completes the proof of the mean. The equivalence of the covariances is directly followed by the matrix inversion lemma [98] such that

$$\begin{aligned}\left(\frac{1}{\sigma_n^2}\Phi^T\Phi+\Lambda^{-1}\right)^{-1} &= \Lambda-\Lambda\Phi^T\left(\Phi\Lambda\Phi^T+\sigma_n^2\mathcal{I}_n\right)\\ &= \Lambda-\Lambda\Phi^T(\mathbf{K}_M+\sigma_n^2\mathcal{I}_n)=\Sigma.\end{aligned} \tag{IV.17}$$

■

From the spherical condition (second condition of (IV.12)), the coefficients follow the Fisher-Bingham distribution [69]. In the literature, there are several methods to sample from this distribution on the sphere [59, 75]. In this work, we make use of the Hamiltonian Monte Carlo (HMC) [21] on the sphere embedded in the Euclidean space. Hence, we do not need a coordinate system but the orthogonal projection to the tangent space at some base points as well as the geodesic flow. Thus, HMC considers the Hamiltonian dynamics according to the Hamiltonian function $H(A,v)$ with the sum of the potential energy $U(A)$ and the kinetic energy $K(v)$

$$H(A,v)=U(A)+K(v). \tag{IV.18}$$

for which $U(A)$ is the minus of log distribution and $K(v)=\frac{1}{2}v^Tv$ where v is regarded as velocity on the tangent space of the sphere. In HMC, the goal is to solve for (A,v) from the following Hamilton's equations

$$\begin{cases}\dot{A}=\nabla_v H(A,v),\\ \dot{v}=-\nabla_A H(A,v).\end{cases} \tag{IV.19}$$

In practice, (IV.19) was numerically solved by the leapfrog integrator.

Algorithm 1: HMC on embedded sphere.

Input : Current state A_0, Data set $\mathcal{D}$, Potential function U, its gradient ∇U, number of iterations T, step size τ.

Output: Next state $\hat{A}$

Sample $v \sim \mathcal{N}(0, \mathcal{I}_M)$

Put $v = v - A_0 A_0^T v$

Evaluate $H_0 = U(A_0) + \frac{1}{2} v^T v$

Put $A = A_0$

for $i = 1$ *to* T **do**

  Put $D = \nabla U(A) - A A^T \nabla U(A)$;

  $v = v - \frac{\tau}{2} D$;

  $A = A \cos(\|v\|\tau) + \frac{v}{\|v\|} \sin(\|v\|\tau)$;

  $v = -A\|v\| \sin(\|v\|\tau) + v \cos(\|v\|\tau)$;

  Put $D = \nabla U(A) - A A^T \nabla U(A)$;

  $v = v - \frac{\tau}{2} D$;

end

Evaluate $H_T = U(A) + \frac{1}{2} v^T v$;

Sample u uniformly on $(0, 1)$;

if $u < \exp(-H_T + H_0)$ **then**

  Return $\hat{A} = A$;

  else Return $\hat{A} = A_0$

end

In Algorithm 1, the steps 8 and 9 update the state (A, v) by the geodesic flow in the sphere with a time interval τ. In our formulation, the potential function is proportional to the minus of the log-posterior on A giving $U(A) = \frac{1}{2}(A - \mu)^T \Sigma^{-1} (A - \mu)$. By Algorithm 1, we can generate a spherical sample of coefficients from the posterior distribution and use its mean to predict the unknown function. We summarize all steps

to approximate the coefficients in Algorithm 2.

Algorithm 2: Estimate the coefficients.

Input : Data set $\mathcal{D}$, truncation order M, number of samples S.

Output: Approximate $\hat{A}$ and $\hat{f}_M(t)^2$.

Let $A = (a_1, ..., a_M)^T$, and $f_M = \sum_{j=1}^{M} a_j \phi_j(t) = \langle (\phi_1(t), ..., \phi_M(t)), A \rangle$.

Evaluate the posterior mean and the conditional covariance of $A|\{\Phi A + \boldsymbol{\epsilon} = \mathbf{y}\}$

$$\mu = \frac{1}{\sigma_n^2}\left(\frac{1}{\sigma_n^2}\Phi^T\Phi + \Lambda^{-1}\right)^{-1}\Phi^T\mathbf{y}, \qquad \Sigma = \left(\frac{1}{\sigma_n^2}\Phi^T\Phi + \Lambda^{-1}\right)^{-1}.$$

Let $U(A) = \frac{1}{2}(A-\mu)^T\Sigma^{-1}(A-\mu)$, and $\nabla U(A) = \Sigma^{-1}(A-\mu)$.

Run HMC in Algorithm 1 to get S samples: $A_1, \ldots, A_S$.

Return the mean value $\hat{A}$ from the sample, and the PDF estimation: $\hat{f}_M(t)^2$.

The mean value $\hat{A}$ in Algorithm 2 can be the Karcher mean or simply

$$\hat{A} = \frac{\sum_{A_k \in Sample} A_k}{\| \sum_{\hat{A}_k \in Sample} \hat{A}_k \|}. \tag{IV.20}$$

We therefore estimate the SRDF by $\hat{f}_M = \left\langle (\phi_1(t), ..., \phi_M(t)), \hat{A} \right\rangle$, and unknown PDF by $\hat{f}_M^2$, or we approximate it by the mean function

$$\hat{f}_M(t)^2 = \frac{1}{S} \sum_{A_k \in Sample} \langle (\phi_1(t), ..., \phi_n(t)), A_k \rangle^2 . \tag{IV.21}$$

IV.4 The framework with sine function

In this section, we validate the proposed method from a particular class of covariance functions: the family of Matérn covariance functions on bounded domain $I = [0, 1]$. From the last chapter, we make use of the Matérn covariance function as a Green function of the following Matérn differential operator (III.25):

$$\mathcal{L}_{Ma} = \left(\varepsilon - \frac{d^2}{dt^2}\right)^{\alpha},$$

where $\varepsilon \geq 0$ and α are parameters. By taking the boundary conditions equal zero, the eigenvalues and eigenfunctions of the equivalent Sturm-Liouville problem

$$\mathcal{L}_{Ma}\phi = \lambda^{-1}\phi, \quad \phi^{(2i)}(0) = \phi^{(2i)}(1) = 0, \text{ for } i = 0, ..., \alpha - 1. \tag{IV.22}$$

are $\lambda_j = \left(\varepsilon + j^2\pi^2\right)^{-\alpha}$, and $\phi_j(t) = \sqrt{2}\sin(j\pi t)$ [23]. Therefore, the Matérn covariance function can be written as

$$K(t,s) = 2\sum_{j=1}^{\infty}\left(\varepsilon + j^2\pi^2\right)^{-\alpha}\sin(j\pi t)\sin(j\pi s). \tag{IV.23}$$

If $f(\cdot) \sim \mathcal{GP}(0, K(\cdot,\cdot))$ then the K-L expansion of f is

$$f(t) = \sqrt{2}\sum_{j=1}^{\infty} a_j \sin(j\pi t), \quad a_j \overset{\text{ind}}{\sim} \mathcal{N}\left(0, \left(\varepsilon + j^2\pi^2\right)^{-\alpha}\right). \tag{IV.24}$$

implying that f can be approximated by

$$f_M(t) = \sqrt{2}\sum_{j=1}^{M} a_j \sin(j\pi t). \tag{IV.25}$$

With the explicit expression of f_M, we can use Algorithm 1, and 2 to approximate a PDF based on the data $\mathcal{D}$. In the next section, we will show the efficiency of this framework for various experiments.

This framework can be applied for multidimensional case, we turn our attention to a $d-$dimensional input space with a rectangular domain $I = [0,1]^d$. In this case, a truncated Matérn covariance function is given as

$$K_M(t,t') = \sum_{j_1,\ldots,j_d=1}^{n} \lambda_{j_1,\ldots,j_d}\phi_{j_1,\ldots,j_d}(t)\phi_{j_1,\ldots,j_d}(t'), \quad (t,t') \in I^2. \tag{IV.26}$$

with the eigenfunctions and eigenvalues satisfying

$$\phi_{j_1,\ldots,j_d}(t) = \prod_{k=1}^{d}\phi_{j_k}(t_k) = 2^{\frac{d}{2}}\prod_{k=1}^{d}\sin(j_k\pi t_k), \quad t = (t_1,\ldots,t_d) \in I \tag{IV.27}$$

and

$$\lambda_{j_1,\ldots,j_d} = \left(\varepsilon + \pi^2\sum_{k=1}^{d} j_k^2\right)^{-\alpha} \tag{IV.28}$$

respectively, see [13] for the proof of a similar approach.

IV.5 Experimental results

In this section, we validate the proposed method with several and various experiments in order to test, evaluate and compare using different configurations:

- **Experiment 1**: Tuning the truncation number or checking how the constrained Gaussian process (cGP) performs according to the truncation number M?
- **Experiment 2**: Tuning the observation number or checking how cGP performs according to the observation number n?

- **Experiment 3**: Comparison between cGP and an Unconstrained Gaussian process (uGP)?
- **Experiment 4**: Comparison between Gaussian process and a Projected Gaussian process (pGP)?
- **Experiment 5**: Results and discussion on a variety of synthetic and real dataset.
- **Experiment 6**: Comparison with a neural networks (NN) based method.
- **Experiment 7**: Results with the multivariate formulation.

For the experimental part, the regression model is used with a Matérn covariance function

$$\begin{cases} y_i = f_M(t_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_n^2), \quad i = 1, \ldots, n \\ f_M(t) = \sqrt{2} \sum_{j=1}^{M} a_j \sin(j\pi t) \\ a_j \overset{\text{ind}}{\sim} \mathcal{N}\left(0, \left(\varepsilon + j^2\pi^2\right)^{-\alpha}\right) \\ \sum_{j=1}^{M} a_j^2 = 1. \end{cases}$$

Under these assumptions f_M^2 is the true unknown PDF that we want to estimate:

(i) We consider that we have n observations,

(ii) we apply cGP to estimate $\bar{f}_M$,

(iii) we evaluate the error between $\hat{f}_M^2$ and f_M^2 using different criteria.

We remind that the hyperparameters of the Matérn covariance function (ε, α) are usually fitted by maximizing the marginal likelihood function [83, 84]. In this section, we only focus on the truncation number M as well as the sample size n.

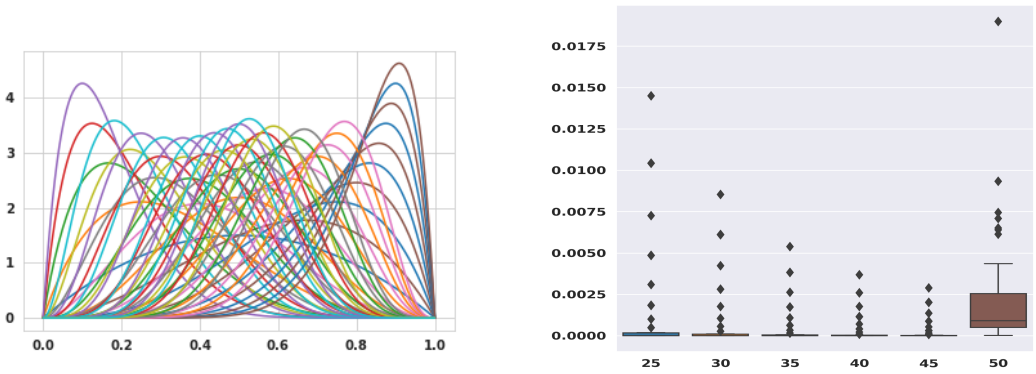
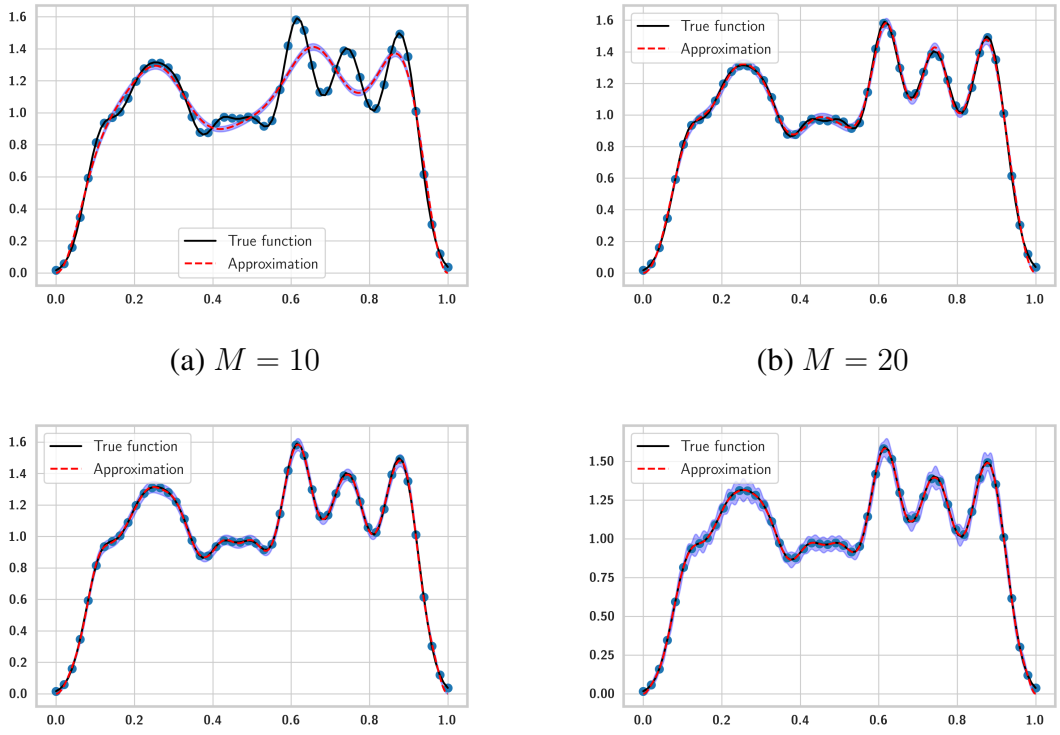


Figure IV.1: Some illustrations of the true functions f_M^2 (left) and the Boxplots of Integrated Square Error (ISE) between the true functions f_M^2 and their approximations $\hat{f}_M^2$ with different truncation orders: $M = 25, 30, \ldots, 50$ (right).

IV.5.1 Tunning the parameters

IV.5.1.1 Tunning the truncation order

In Figure IV.1 (left), we plot 50 examples of PDFs $f_M(t)^2$ generated from beta distributions. For each PDF, we observe $n = 50$ points $f_M(t_i)^2$ $(i = 1, \ldots, n)$ associated to t_i that are uniformly spaced in $I = [0, 1]$. The number of truncation M varies in $\{25, 30, ..., 50\}$. For these experiments, we apply HMC as detailed in Algorithm 1 and the resulting coefficients are updated as detailed in Algorithm 2 with the number of sample $S = 10000$. Consequently, we evaluate the mean function given in (IV.21) as an approximation $\hat{f}^2_M$ of each PDF. Figure IV.1 (right) shows the boxplot of the Integrated Square Error (ISE) between the true PDFs f^2_M and their approximates $\hat{f}^2_M$ when varying the truncation order. Note that the proposed cGP is able to minimize the ISE criteria. The best results are obtained when $25 \leq M \leq 45$.



(a) $M = 10$ (b) $M = 20$

(c) $M = 30$ (d) $M = 50$

Figure IV.2: The true PDF (black) from which we observe some points and use them for training (blue). The approximated PDF at some unobserved points (red) and the confidence region (cyan) with different truncation orders $M = 10, 20, \ldots, 50$.

As a second example, we consider a nonparametric PDF f_M and different approxi-

mates $\hat{f}_M$ with an increasing M, see Figure IV.2. We show the confidence region at some unobserved points $t_\star$ satisfying $\hat{f}_M(t_\star)^2 \pm 2\hat{\sigma}(t_\star)$ where $\hat{\sigma}(t_\star)$ is the empirical standard deviation. Accordingly, the approximation is bad for $M = 10$ but gets better when $20 \leq M \leq 30$ with a small confidence region. When $M = 50$, the approximation is still good but the confidence region becomes more larger which makes more uncertainty in our approximations. Additionally, the computational time increases for $M = 50$. Based on these results, the truncation number will be fixed to $M = 30$ throughout the rest of the experiments. This choice is motivated by the previous experiments and is not worth a theoretical or general justification.

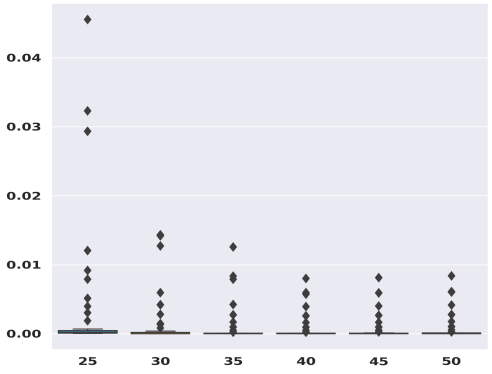
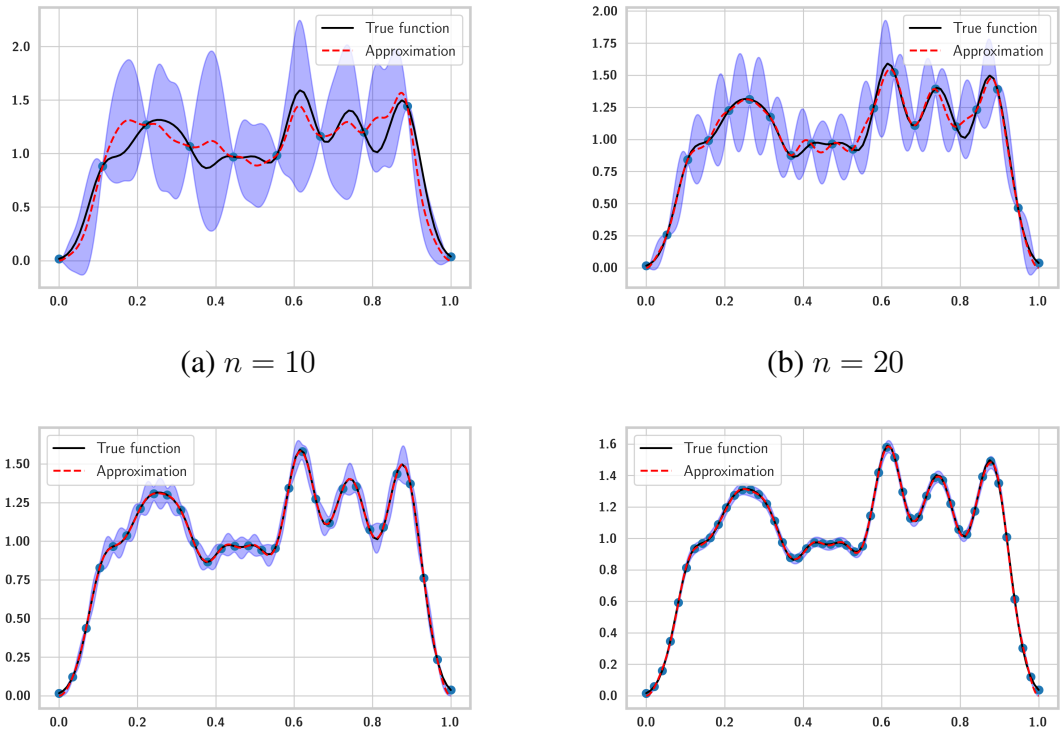


Figure IV.3: The Boxplot of ISE between f_M and $\hat{f}_M$ when $n = 25, 30, \ldots, 50$ and $M = 30$.

IV.5.1.2 Tunning the number of observations

In the second experiment, the truncation number is fixed to be $M = 30$, while the number of observations n is varied. Figure IV.3 shows a boxplot of the ISE between the true function f_M and its approximate $\hat{f}_M$ with $n = 25, 30, \ldots, 50$. Based on these results, one can state that the ISE criteria decrease when n is increasing but without a significant margin. In Figure IV.4 we plot both the true PDF f_M and $\hat{f}_M$. At the first glance, the confidence region $\hat{f}_M(t_\star)^2 \pm 2\hat{\sigma}(t_\star)$ is smaller when n increases. As expected, we can conclude that the uncertainty decreases when the data size n is increasing which seems to give more robustness to our method.

IV.5.2 Comparison with variant Gaussian process



(a) $n = 10$ (b) $n = 20$

(c) $n = 30$ (d) $n = 50$

Figure IV.4: The true PDF (black) and observations used for training (blue). The approximated PDF at some unobserved points (red) and the confidence region (cyan) with different data sizes $n = 10, 20, \ldots, 50$.

IV.5.2.1 Comparison with unconstrained Gaussian process

In this section, we compare the constrained Gaussian process (cGP) against the unconstrained Gaussian process (uGP). We remind that the prediction in uGP does not claim any spherical constraint and is performed from (IV.1) and (IV.2). We consider the true PDF as a normalized version of

$$g(t) = (1-t)\left(1 + \sin\left(-\frac{\pi}{2} + 10\pi t\right)\right), \quad t \in I.$$

From Figure IV.5, we observe that both cGP and uGP can approximate the PDF efficiently with a small advantage to cGP. However, the biggest flow of uGP is that the approximation may not be a PDF. In fact, from this example the integral of the approximate is equal to 0.729 while cGP predicts an approximation with an integral equal to 1.

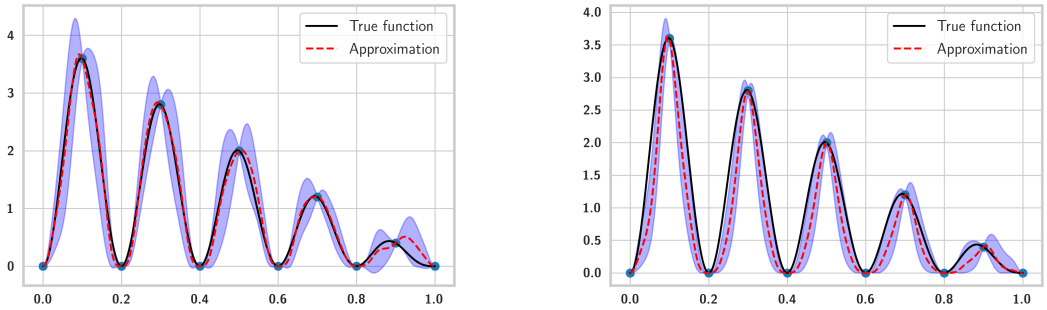


Figure IV.5: The true PDF (black) and the observations (blue). The approximated PDF at some unobserved points (red) and the confidence region (cyan) for cGP (left) and uGP (right).

IV.5.2.2 Comparison with normalized uGP

In the previous example, we showed how a uGP without the spherical constraint leads to a bad approximation: A solution that is not guaranteed to be a PDF. To fulfill such constraint one can, for example, project the estimated function to the space of PDFs by normalizing it to have the integral one. From the literature, this technique has been widely used and refereed as a projected Gaussian process (pGP). For simulations, we consider a true PDF g_l^2, such that $g_l(t) = \sqrt{2}\sum_{j=1}^{l} b_j \sin(j\pi t)$, and $B = (b_1, ..., b_l)$ are random coefficients in $\mathbb{S}^{l-1} \subset \mathbb{R}^l$. To get observed PDFs from the proposed model we use the following steps:

1. Choose the base point $\mathbb{1} = \frac{1}{\sqrt{l}}(1, \ldots, 1) \in \mathbb{S}^{l-1}$, for simplification. Any other valid point on the sphere can also be used.
2. Define the tangent space on the sphere locally at $\mathbb{1}$ denoted $T_{\mathbb{1}}\mathbb{S}^{l-1}$ which is a linear space of all elements $v = (v_1, ..., v_l) \in \mathbb{R}^l$ satisfying $\sum_{j=1}^{l} v_j = 0$.
3. Define the exponential map at $\mathbb{1}$, that maps any direction $v \in T_{\mathbb{1}}\mathbb{S}^{l-1}$ into a point $B = (b_1, ..., b_l)$ on the sphere $\mathbb{S}^{l-1}$ such that

$$\begin{aligned} \exp_{\mathbb{1}} : & T_{\mathbb{1}}\mathbb{S}^{l-1} \to \mathbb{S}^{l-1} \\ & v \mapsto B = \cos(\|v\|_2)\mathbb{1} + \sin(\|v\|_2)\frac{v}{\|v\|_2} \end{aligned} \tag{IV.29}$$

   where $\|.\|_2$ denotes the usual Euclidean norm in $\mathbb{R}^l$.
4. Compute $g_l(t) = \sqrt{2}\sum_{j=1}^{l} b_j \sin(j\pi t)$ for each v and let g_l^2 as a true PDF.

Below we consider two different strategies to generate v in $T_{\mathbb{1}}\mathbb{S}^{l-1}$ and consequently to generate PDFs randomly:

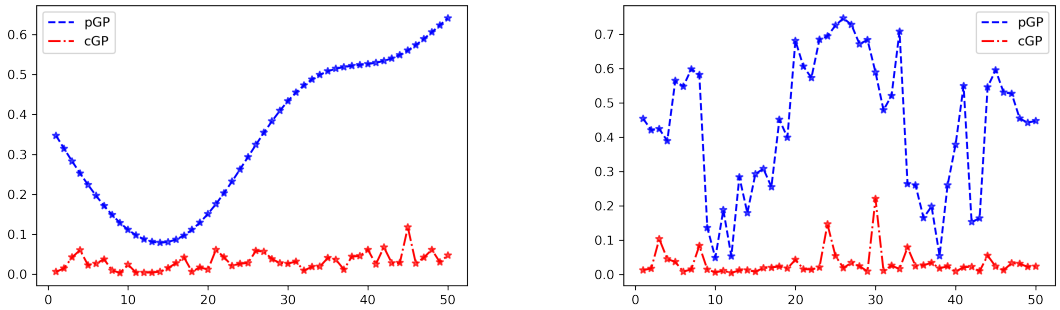


Figure IV.6: Comparison between pGP and cGP: The ISE for several PDFs between 1 and 50 along a geodesic arc as an increasing variation of distance (left) and along a circle as an increasing variation of the angle at a fixed radius (right).

- First, chose a unit direction v such that $\|v\|_2 = 1$ then get a set of T true PDFs as $g_l^k(t)^2 = \left(\sqrt{2}\sum_{j=1}^{l} b_j^k \sin(j\pi t)\right)^2$ for $k = 1, ..., T$ according to

$$B^k = (b_1^k, \ldots, b_l^k) = \exp_{\mathbb{1}}\left(\frac{kv}{2T}\right). \tag{IV.30}$$

  This means that the functions g_l^k are generated along a geodesic arc. With a small value of k the coefficients of the corresponding function g_l^k is close to $\mathbb{1}$, but with a large value of k it is far from $\mathbb{1}$. Especially, we focus on the case when the true function is far from the base point (center).

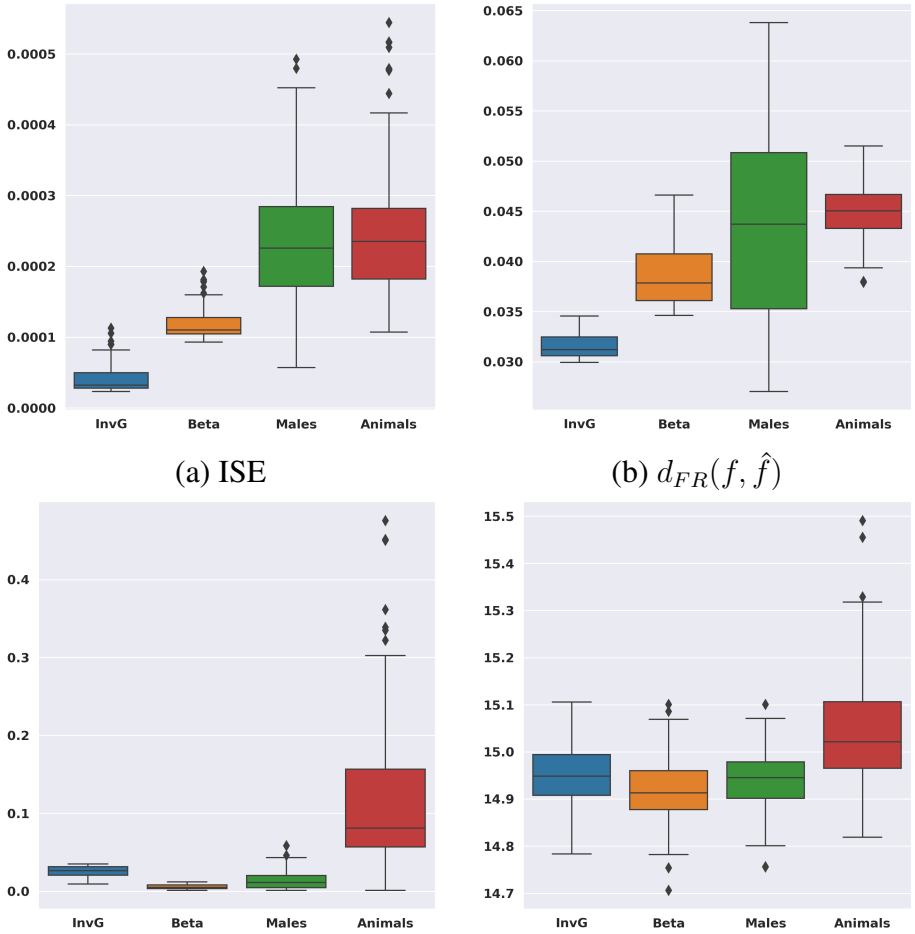
- Second, we fix a unit direction v and randomly choose v_k such that the angle between v and v_k is $\theta_k = 2k\pi/T$. Similarly, we generate a set of T true PDFs $(g_l^k)^2$ using the corresponding coefficients $B^k = \exp_{\mathbb{1}}(v_k)$. Here, the distance is fixed but the angle varies around the center.

For the second example, we let $l = 20$ and $T = 50$, $M = 30$ and $n = 25$. We summarize the ISE results in Figure IV.6. We state that cGP performs much better in general.

IV.5.3 Simulation studies and results on real dataset

In this section, we validate the cGP model on four dataset denoted InvG, Beta, Males and Animals. InvG and Beta are synthetic PDFs generated from the beta and inverse-gamma distributions. On the other hand, Males and Animals are PDFs of real data representing growth in a male group and temperatures of cats observed during several days of the disease, respectively. Each test has been performed with $M = 30$ and $n = 25$. In

addition to the ISE, we consider the Fisher-Rao distance to measure the error between the true PDF f_M^2 and its approximate $\hat{f}_M^2$ as $d_{FR}(f_M^2, \hat{f}_M^2) = 2\cos^{-1}(\int_I \sqrt{f_M^2(t)\hat{f}_M^2(t)}dt)$.



(a) ISE (b) $d_{FR}(f, \hat{f})$

(c) $U(\bar{A})$ (d) The mean of $U(\hat{A}_k)$

Figure IV.7: Results of cGP using different metrics and various dataset.

Each dataset contains $T = 100$ PDFs in the experiments. From Figure IV.7 (a) and (b), it is notable that the ISE and the Fisher-Rao distance are very small. The results for Males and Animals are significant allowing our model to be successfully applied for real applications. In Figure IV.7 (c) and (d), we display the potential function at the estimated coefficient $U(\hat{A})$ and the mean of all potentials $U(A_k)$ where A_k denotes the k-th sample of the HMC sampling. The difference between these two quantities confirms that the successful optimization using spherical HMC. In fact the much smaller values of $U(\hat{A})$ means that the predicted values are very close to the data observations.

IV.5.4 Comparison with Neural Network

In this section, we compare the proposed approach with a Neural Network-based method (NN). Indeed, many layers inside a NN are parameterized, i.e., have associated weights and biases that are optimized during the training step. In general, the principle of NN is that signals travel from the first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times. For training NN, we consider the back propagation algorithm to adjust the parameters (model weights) with a gradient-based method for minimizing the loss function. Since we are dealing with a SRDF, the coefficients are normalized during the training task.

Now, we give more details about the architecture of the NN model. We keep the same configurations as detailed in Section .5.3. The NN was implemented with 5 linear layers. The number of nodes in the input layer is equal to the number of observations: $n = 25$, the number of nodes in the output layer is equal to the truncation number (number of coefficients: $M = 30$), and the number of nodes in the hidden layers is equal to 500, 1000, and 200. In the forward function, we normalize the output to have a unit norm before returning the coefficients allowing the NN model to fit coefficients belonging to the unit sphere. Finally, the NN model uses MSE loss and Adam optimization with a learning rate equal to 2×10^{-3}. After the NN is fully trained, the NN model takes the input is the observations and output the coefficients of the prediction function.

Figure IV.8 (left) displays an example of a true function and its approximate using NN. Figure IV.8 (right) summarizes all ISE values when using the NN-based method to predict PDFs from 4 different dataset. We have randomly selected and used 70% of the dataset for training and the rest is kept for test. We remind that the same data have been used to test cGP and results are summarized in Figure IV.7 (a). We have also used the same colors to make the comparison between results easy to follow. Accordingly, we can easily notice that cGP performs better than NN. For NN model, it needs a good training data set to perform well.

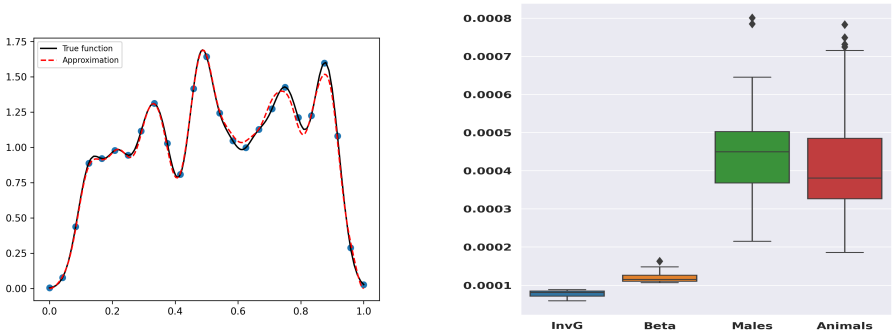
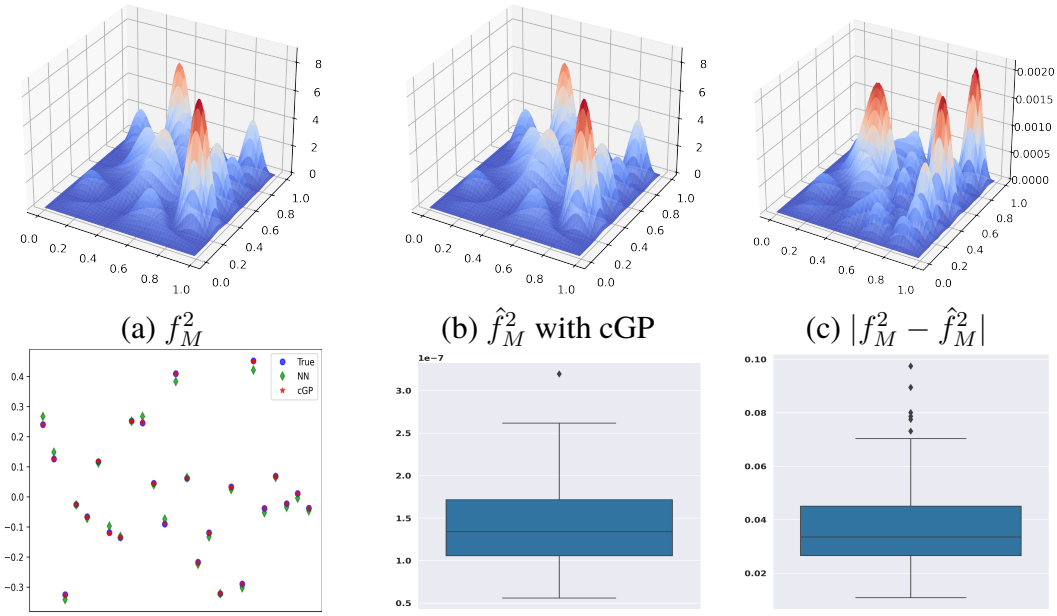


Figure IV.8: Left: The true PDF (black), the training points (blue) and the NN-based approximate (red). Right: The boxplots summarizing ISE for NN-based method on 4 dataset.

IV.5.5 The multivariate case

In this section, we show how the proposed method can be extended and applied to the multivariate case. In the following experiments, we consider bivariate PDFs defined on the unit square interval $I = [0,1]^2$. We first simulate random coefficients from $a_{j_1,j_2} \overset{\text{ind}}{\sim} \mathcal{N}\Big(0, \big(\varepsilon + (j_1^2 + j_2^2)\pi^2\big)^{-1}\Big)$ and then form the corresponding SRDF with $f_M(t) = 2\sum_{j_1,j_2=1}^{5} a_{j_1,j_2} \sin(j_1\pi t)\sin(j_1\pi t)$, see an example in Figure IV.9 (a). We display the estimate $\hat{f}_M^2$ obtained with cGP in Figure IV.9 (b) and the resulting error as the absolute value of the difference at each point in Figure IV.9 (c). In Figure IV.9 (d) we



(a) f_M^2 (b) $\hat{f}_M^2$ with cGP (c) $|f_M^2 - \hat{f}_M^2|$

(d) A and $\hat{A}$ with cGP and NN (e) ISE with cGP (f) ISE with NN

Figure IV.9: Top: An example of a true PDF (a), its approximate with cGP (b) and the difference between them (c). Bottom: The true coefficients in blue and the estimated coefficients in red (d), the boxplot summarizing the ISE$(f_n, \hat{f}_n)$ from cGP (e) and the NN (f).

display the true coefficients initially used to form the PDF $A = (a_{1,1}, a_{1,2}, \ldots, a_{5,5})^T$ and their approximations using the spherical HMC sampling and NN. Finlay, we use $T = 50$ examples as true SRDFs f_M, we estimate $\hat{f}_M^2$ and we compute ISE between them. We summarize all results as boxplots in Figure IV.9 (e). The same criteria is maintained for NN in Figure IV.9 (f). Accordingly, we state that cGP is very successful in estimating bivariate PDFs.

IV.6 Discussion and conclusion

In this chapter, we have introduced a new framework to learn, infer and predict probability density functions using constrained Gaussian Processes prior. Thanks to the formulation, the spherical constraint translates directly into a finite set of random coefficients on an appropriate basis. We have used the spherical HMC sampling to efficiently compute the mean function as a good candidate for the prediction. We have conducted various and extensive experiments to confirm the ability of our proposal for various real-world applications.

Chapter V: Transfer learning on finite probability measures

In this chapter, we propose a new powerful transfer learning method of statistical models on the space of probability measures $\mathcal{P}_+(I)$. We develop new approaches to capture the Riemannian geometry of $\mathcal{P}_+(I)$ equipped with Fisher-Rao metric. Specifically, we exploit the Levi-Civita parallel transport on $\mathcal{P}_+(I)$ that preserves the inner product. We derive explicit theoretical expression of important geometric structures on $\mathcal{P}_+(I)$ associated to the Levi-Civita connection such as: Minimal geodesic, exponential and logarithm map. We demonstrate that capturing such geometry yield a significant benefit in transfer learning of covariance matrices, PCA and linear regression models on $\mathcal{P}_+(I)$. We illustrate and discuss the effectiveness of the proposed approach with various and multiple experimental results.

Organization. This chapter is organized as follows. Section .1 presents a general introduction. In section .2, we introduce the problems that we want to address. In Section .3, we study in detail the geometry of the space of finite measures with the Fisher-Rao metric. In section .4, we show how the statistical models can be transported along the manifold. The experimental results are presented and discussed in Section .5. We conclude this Chapter at Section .6.

V.1 Introduction

Research on transfer learning has attracted more and more attention over the last years due to its importance in varied fields of machine learning and data mining areas. Today, transfer learning methods appear in many applications, most notably in computer vision [17, 82], neural networks [89], natural language processing tasks [100], sentiment analysis [120] and medical imaging [20]. The study of transfer learning is motivated by the fact that people can intelligently apply knowledge learned previously to solve new problems faster or with better solutions. The definition of transfer learning is depicted in

terms of domain and task and presented in [90]: Given a source domain $\mathcal{D}_L$ and learning task $\mathcal{T}_L$, a target domain $\mathcal{D}_S$ and learning task $\mathcal{T}_S$, transfer learning aims to help improve the learning of the target predictive function $f_S(.)$ in $\mathcal{D}_S$ using the knowledge in $\mathcal{D}_L$ and $\mathcal{T}_L$, where $\mathcal{D}_L \neq \mathcal{D}_S$, or $\mathcal{T}_L \neq \mathcal{T}_S$.

In recent years, statistical models have been extensively used for data lying on manifolds. Indeed, data points on Riemannian manifolds are fundamental objects in many fields of science and engineering including, machine learning, image and video processing, artificial vision and medical. While transfer learning methods are very well-developed on the Euclidean space $\mathbb{R}^d$ [28, 92], much effort has been directed towards efficient approaches to encompass the wide variation within the class of manifolds, in particular from methods and techniques of differential geometry. In this vein, a new class of transfer learning algorithms is developed in this chapter for data sets that intrinsically lie on the space of strictly positive probability measures on a domain I, defined by

$$\mathcal{P}_+(I) = \left\{ \mu = \sum_{i \in I} \mu_i \delta^i \mid \mu_i > 0, \quad \forall i \in I, \text{ and } \sum_{i \in I} \mu_i = 1 \right\},$$

where δ^i is the Dirac measure concentrated on i. To make the problem and the setting more general, we consider two valued datasets $P_L = \{\mu^i\}_{i=1}^L$ and $P_S = \{\nu^i\}_{i=1}^S$ in the space $\mathcal{P}_+(I)$, such that P_L is of large size and P_S is of small size. The main aim is to learn a statistical model from the data set on P_S such as covariance matrix, PCA, and regression while leveraging statistical information from the large data set P_L. Since statistical models are often expressed in tangent spaces, thus, the goal is to establish an accurate and efficient transfer learning algorithm of statistical objects between tangent spaces in a way that preserves the structure of the statistical model while aligning to the complex geometries of the underlying manifold.

In the Euclidean setting, different techniques and strategies of transfer learning have appeared in the literature, including inductive transfer learning, transductive transfer learning, and unsupervised transfer learning. A more detailed description of those methods can be found in [104]. For general Riemannian manifolds, the idea of leveraging knowledge from one class to another has been investigated in recent works. In [135], Xie et al. introduce a new framework that can transfer, not only the data but statistical models using parallel transport. Nevertheless, their choice of parallel transport suffers

from some drawbacks: Neither the inner product between two vectors nor the length of a vector is preserved. Recently, in [46], Freifeld et al. attempted to avoid the weakness of the method given in [135] and have developed a new approach that can transfer a covariance matrix using the parallel transport that respects the Riemannian metric of the manifold. In this framework, we propose to extend transfer learning strategy described in [46] such that it can be successfully applied to data lying in the space of strictly positive probability measures $\mathcal{P}_+(I)$.

Recently much attention has been focused on the space of probability measures $\mathcal{P}_+(I)$ with different metrics including Frobenius, Fisher-Rao, log-Euclidean, Jensen-Shannon and Wasserstein metrics [8, 32, 97, 126]. Works on linear regression [70, 74], estimation [64], barycenters [68, 105], have been deeply studied and led to computational advances in statistical analysis. The present chapter is an attempt to formalize a rigorous and computational approach to the transfer learning problem on the space of probability measures $\mathcal{P}_+(I)$ equipped with Fisher-Rao metric, in the context of the theory of information geometry developed in particular in [4, 8, 72, 73]. Indeed, on the Riemannian manifold $\mathcal{P}_+(I)$, each tangent vector X belongs to a tangent space $T_\mu\mathcal{P}_+(I)$ specific to its root point $\mu \in \mathcal{P}_+(I)$. Hence tangent vectors from different tangent spaces cannot be compared directly. Parallel transport is the unique mathematical tool capable of transporting vectors between tangent spaces while retaining the information they contain. In our case, we choose a metric parallel transport which asserts that the orthogonality and distance between tangents vectors in $T_\mu\mathcal{P}_+(I)$ are retained, and consequently the variance is preserved. The main contributions are summarized as follows:

1. Firstly, we develop an explicit expression of Christoffel symbols, and therefore the Levi-Civita connection associated with the Fisher-Rao metric which allows the computation of the geodesic curves joining two points on $\mathcal{P}_+(I)$. In this way, we exhibit an exact equation of the Levi-Civita parallel transport of a tangent vector along a geodesic curve joining two probability measures on $\mathcal{P}_+(I)$.
2. We apply the parallel transport to transfer learning of statistical models such as covariance matrix, PCA and linear regression models between tangent spaces of $\mathcal{P}_+(I)$. Then we illustrated the methods with several experiments.

V.2 Problem Formulation

In this chapter, we address the following problem. Given two valued datasets $P_L = \{\mu^i\}_{i=1}^L$ and $P_S = \{\nu^i\}_{i=1}^S$ in the space of strictly positive probability measures $\mathcal{P}_+(I)$, such that P_L is of large size and P_S is of small size. Our goal is to learn a statistical model from the dataset P_S such as the covariance matrix, PCA, and regression model while leveraging statistical information from the large data set P_L. Since statistical models are often expressed in tangent spaces, hence, the goal is to transfer statistical objects between tangent spaces in a way that preserves the structure of the statistical model while adapting to the structure of the manifold.

V.3 Levi-Civita Parallel Transport on $\mathcal{P}_+(I)$

In this section, we study the space of strictly positive probability measures $\mathcal{P}_+(I)$ on a given finite set I endowed with an atlas of charts forming a differentiable manifold modeled on $\mathbb{R}^n$. We compute computational tools of interest, namely Levi-Civita connection ∇^{LC}, geodesics, Exp and log maps and the Levi-Civita parallel transport. For a more detailed exposition on these concepts, we refer the reader to [61]. In the following, we denote ∇ instead of ∇^{LC} for convenience.

V.3.1 Fisher-Rao Geometry

Let a finite sample space I coded as $I = \{1, \ldots, n, n+1\}, n \in \mathbb{N}$. On I, the set of all real functions forms an algebra, denoted as $\mathcal{F}(I) = \{f : I \to \mathbb{R}\}$. Its unity function $\mathbb{1}_I$ or simply $\mathbb{1}$ is given by $\mathbb{1}(i) = 1$, for all $i = 1, \ldots, n, n+1$. A canonical basis of $\mathcal{F}(I)$ is given by

$$e_i(j) = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise,} \end{cases} \tag{V.1}$$

and hence, every $f \in \mathcal{F}(I)$ has the representation

$$f = \sum_{i \in I} f^i e_i, \tag{V.2}$$

where $f^i = f(i)$. We will denote by $\mathcal{S}(I)$ the dual space of $\mathcal{F}(I)$, the space of $\mathbb{R}$-valued linear forms on $\mathcal{F}(I)$. With the Riesz representation theorem, this vector space is interpreted as the vector space of signed measures on I, namely

$$\mathcal{S}(I) = \{\mu : \mathcal{F}(I) \to \mathbb{R} \mid \ \mu = \sum_{i\in I} \mu_i \delta^i\}, \tag{V.3}$$

where $\mu_i = \mu(e_i)$ and δ^i is considered as the Dirac measure supported at $i \in I$. It is also shown that $\mathcal{S}(I)$ is a smooth manifold, furthermore $\mathcal{S}(I)$ is isomorphic to $\mathbb{R}^{n+1}$. Besides we have a vector space isomorphism between the space $\mathcal{F}(I)$ and $\mathcal{S}(I)$, given by

$$\begin{aligned} \mathcal{F}(I) &\longrightarrow \mathcal{S}(I) \\ f &\longmapsto f\mu := \sum_{i\in I} f^i \mu_i \delta^i. \end{aligned} \tag{V.4}$$

The inverse is the Radon-Nykodym derivative with respect to μ, denoted as ϕ_μ,

$$\begin{aligned} \phi_\mu : \mathcal{S}(I) &\longrightarrow \mathcal{F}(I) \\ \nu = \sum_{i\in I} \nu_i \delta^i &\longmapsto \frac{d\nu}{d\mu} := \sum_{i\in I} \frac{\nu_i}{\mu_i} e_i. \end{aligned} \tag{V.5}$$

In particular, the tangent space at the point $\mu \in \mathcal{S}(I)$ is given by

$$T_\mu \mathcal{S}(I) = \{\mu\} \times \mathcal{S}(I). \tag{V.6}$$

Let us consider the following submanifolds of $\mathcal{S}(I)$:

$$\mathcal{S}_c(I) = \left\{ \mu = \sum_{i\in I} \mu_i \delta^i \mid \sum_{i\in I} \mu_i = c, \quad c \in \mathbb{R} \right\}$$

and

$$\mathcal{M}_+(I) = \{\mu \in \mathcal{S}(I) \mid \ \mu_i > 0, \quad \forall \quad i \in I\}$$

the space of finite strictly positive measures on I. In the following definition, we see more clearly the notion of measure on a finite space.

Definition V.1

A signed measure on a finite sample space I associated to $\mu = \sum_{i\in I} \mu_i \delta^i$ is a map $\mu : \mathcal{B}(I) \to \mathbb{R}$, defined on $\mathcal{B}(I)$ of the set of all subset of I, with:

$$\mu(A) = \sum_{i\in I} \mu_i \delta^i(A) = \sum_{i\in A} \mu_i, \quad \forall A \subset I. \tag{V.7}$$

A measure μ is called probability measure if $\mu(I) = 1$ and $\mu_i \geq 0$ for all i. Furthermore, if $\mu_i > 0$ for all i, μ is called strictly positive probability measure.

We denote by $\mathcal{P}(I)$ the space of all probability measures, and $\mathcal{P}_+(I)$ the space of strictly positive probability measures and we note

$$\mathcal{P}_+(I) = \left\{ \mu = \sum_{i\in I} \mu_i \delta^i \mid \mu_i > 0, \quad \forall i \in I, \text{ and } \sum_{i\in I} \mu_i = 1 \right\}.$$

We check at once that $\mathcal{P}_+(I) \subset \mathcal{M}_+(I) \subset \mathcal{S}(I)$. Therefore, as an open submanifold of $\mathcal{S}(I)$, $\mathcal{M}_+(I)$ has the same tangent space at the point $\mu \in \mathcal{M}_+(I)$. The tangent space of $\mathcal{P}_+(I)$ is given in the following proposition.

Proposition V.1

Consider $\mathcal{P}_+(I)$ as a submanifold of $\mathcal{S}(I)$, then the tangent space at $\mu \in \mathcal{P}_+(I)$ is given by:

$$T_\mu \mathcal{P}_+(I) = \{\mu\} \times \mathcal{S}_0(I) = \left\{ (\mu, v) \mid \mu \in \mathcal{P}_+(I) \text{ and } v = \sum_{i\in I} v_i \delta^i \in \mathcal{S}_0(I) \right\}.$$

Proof Indeed, let $\alpha(t)$ be any differential curve in $\mathcal{P}_+(I)$ that goes through μ at $t = 0$. Suppose $\alpha(t) = \sum_{i\in I} \alpha_i(t)\delta^i$, where $\alpha_i(t)$ is a differentiable function of t and $\alpha_i(0) = \mu_i$. As a curve in $\mathcal{S}(I)$, the differential of α at t is given by $\dot{\alpha}(t) = \sum_{i\in I} \dot{\alpha}_i(t)\delta^i$. The tangent vector $\dot{\alpha}(t)$ is also a measure, and

$$\sum_{i\in I} \dot{\alpha}_i(t) = \frac{d}{dt}\left(\sum_{i\in I} \alpha_i(t) \right) = \frac{d}{dt}(1) = 0, \quad \forall t. \tag{V.8}$$

This proves that $T_\mu \mathcal{P}_+(I) \subset \{\mu\} \times \mathcal{S}_0(I)$. For the converse, let any $v \in \mathcal{S}_0(I)$. We define the curve $\alpha(t) = \mu + tv$, defined for t close to 0 such that $\alpha(t) \in \mathcal{P}_+(I)$. It follows that $\dot{\alpha}(0) = v$. This proves that that $T_\mu \mathcal{P}_+(I) \supset \{\mu\} \times \mathcal{S}_0(I)$. ■

We want to study $\mathcal{P}_+(I)$ intrinsically and impose a Riemannian metric on it. To this end, we define a local coordinate map on $\mathcal{P}_+(I)$. Let U be an open set of $\mathbb{R}^n$ given by

$$U = \left\{ x = (x_1, \ldots, x_n) \in \mathbb{R}^n \mid x_i > 0, \forall i \in I, \text{ and } \sum_{i=1}^{n} x_i < 1 \right\}.$$

We define a map φ as

$$\varphi : \mathcal{P}_+(I) \longrightarrow U,$$

$$\mu = \sum_{i\in I} \mu_i \delta^i \longmapsto (\varphi^1(\mu), ..., \varphi^n(\mu)) = (x^1(\mu), ..., x^n(\mu)),$$

such that $(\varphi^1(\mu), ..., \varphi^n(\mu)) = (\mu_1, \ldots, \mu_n)$. Clearly, φ is an homomorphism and its

inverse is given by

$$\varphi^{-1} : U \longrightarrow \mathcal{P}_+(I),$$
$$(x_1, \ldots, x_n) \longmapsto \mu = \sum_{i=1}^{n} x_i \delta^i + \left(1 - \sum_{i=1}^{n} x_i\right) \delta^{n+1}.$$

The single chart $(\mathcal{P}_+(I), \varphi)$ defines a manifold structure of $\mathcal{P}_+(I)$. Given a point $\mu \in \mathcal{P}_+(I)$, let $\frac{\partial}{\partial x^i}\Big|_\mu$ be the tangent vector at μ given by

$$\frac{\partial}{\partial x^i}\Big|_\mu = \frac{\partial}{\partial x_i}\Big|_{\varphi(\mu)} \varphi^{-1} = \left(\delta^i - \delta^{n+1}\right), \quad \text{for } i = 1, \ldots, n.$$

Thus, $\left\{ \frac{\partial}{\partial x^i}\Big|_\mu, i = 1, ..., n \right\}$ define a local frame field of $T_\mu \mathcal{P}_+(I)$ at a point $\mu \in \mathcal{P}_+(I)$. Now, for any $v = \sum_{i \in I} v_i \delta^i \in T_\mu \mathcal{P}_+(I)$, it has the representation

$$v = \sum_{i=1}^{n+1} v_i \delta^i = \sum_{i=1}^{n} v_i \delta^i - \sum_{i=1}^{n} v_i \delta^{n+1} = \sum_{i=1}^{n} v_i (\delta^i - \delta^{n+1}) = \sum_{i=1}^{n} v_i \frac{\partial}{\partial x^i}, \tag{V.9}$$

since $v \in \mathcal{S}_0(I)$.

With $S(I)$ being a finite-dimensional linear space, and therefore, it can be naturally equipped with a metric. For $v, w \in T_\mu S(I)$, we define the inner product as

$$< v, w >_\mu = \mu \left(\frac{dv}{d\mu} . \frac{dw}{d\mu} \right) = \sum_i \frac{v_i w_i}{\mu_i} \tag{V.10}$$

where $\frac{dv}{d\mu} = \sum_{i \in I} \frac{v_i}{\mu_i} e_i \in \mathcal{F}(I)$, represents a simple version of the Radon–Nikodym derivative with respect to μ. This metric induces a metric on $\mathcal{M}_+(I)$. Hence, following the geometry structures in $\mathcal{M}_+(I)$ equipped with Fisher-Rao metric, we derive the corresponding one in $\mathcal{P}_+(I)$.

Definition V.2

Let μ be a probability measure in $\mathcal{P}_+(I)$. Given two tangents vectors v and w in $T_\mu \mathcal{P}_+(I)$, the Fisher-Rao metric $\mathfrak{g}_\mu : T_\mu \mathcal{P}_+(I) \times T_\mu \mathcal{P}_+(I) \to \mathbb{R}$ is given by

$$\mathfrak{g}_\mu(v, w) = \sum_{i \in I} \frac{v_i w_i}{\mu_i},$$

and $||v||_\mu = \sqrt{\mathfrak{g}_\mu(v, v)}$.

With respect to the coordinate map $(\mathcal{P}_+(I), \varphi)$, the Fisher-Rao metric $G_\mu = [g_{ij}]$ is expressed as [8]:

$$g_{ij}(\mu) = \begin{cases} \frac{1}{\mu_i} + \frac{1}{\mu_{i+1}}, & \text{if } i = j, \\ \frac{1}{\mu_{n+1}}, & \text{otherwise,} \end{cases}$$

for $i, j = 1, \cdots, n$. The inverse matrix $G_\mu^{-1} = [g^{ij}]$ has the components

$$g^{ij}(\mu) = \begin{cases} \mu_i(1-\mu_i), & \text{if } i = j, \\ -\mu_i\mu_j, & \text{otherwise.} \end{cases}$$

Our goal to make $\mathcal{P}_+(I)$ as a Riemannian manifold is now fully satisfied. Our next goal is to compute explicit expressions of geometric structures on $\mathcal{P}_+(I)$, especially, the Levi-Civita parallel transport which will be essential to define a transfer learning approach of statistical models on $\mathcal{P}_+(I)$.

V.3.2 Levi-Civita connection on $\mathcal{P}_+(I)$

Let $\mathfrak{X}(\mathcal{P}_+(I))$ denote the set of all smooth vector fields on $\mathcal{P}_+(I)$. On the Riemannian manifold $\mathcal{P}_+(I)$ with the Fisher-Rao metric. The corresponding Levi-Civita connection $\nabla : \mathfrak{X}(\mathcal{P}_+(I)) \times \mathfrak{X}(\mathcal{P}_+(I)) \to \mathfrak{X}(\mathcal{P}_+(I))$, takes vector fields X, Y, to give a new vector field, denoted $\nabla_X Y$, telling us how the vector field Y is changing in the direction X. As we know in the Chapter II that the Levi-Civita connection is metric and torsion free, for all $X, Y, Z \in \mathfrak{X}(\mathcal{P}_+(I))$, it satisfies

$$\begin{cases} X\mathfrak{g}(Y,Z) = \mathfrak{g}(\nabla_X Y, Z) + \mathfrak{g}(Y, \nabla_X Z), \\ \nabla_X Y - \nabla_Y X = [X, Y]. \end{cases} \tag{V.11}$$

In the local coordinate map $(\mathcal{P}_+(I), \varphi)$, the Levi-Civita connection is defined by the Christoffel symbols $\Gamma_{ij}^k : \mathcal{P}_+(I) \to \mathbb{R}$ such that

$$\nabla_{\partial x_i} \partial x_j = \sum_k \Gamma_{ij}^k \partial x_k. \tag{V.12}$$

The Christoffel symbols are given explicitly in the following proposition.

Proposition V.2

With respect to the local coordinate map $(\mathcal{P}_+(I), \varphi)$, the Christoffel symbols associated with the Fisher-Rao metric are given by

$$\Gamma_{ij}^k = \begin{cases} \frac{1}{2} \times \frac{x_k}{1 - \sum_{h=1}^n x_h}, & i \neq j, \\ \frac{1}{2} \times \left(\frac{x_k}{1 - \sum_{h=1}^n x_h} + \frac{x_k}{x_i} \right), & i = j \neq k, \\ \frac{1}{2} \times \left(\frac{x_k}{1 - \sum_{h=1}^n x_h} - \frac{1 - x_k}{x_k} \right), & i = j = k. \end{cases} \tag{V.13}$$

Proof The smooth functions Γ_{ij}^k are easily computed through the characterization of the Levi-Civita connection by the Koszul formula obtained from (V.11) computed for all the circular permutations of $X, Y, Z \in \mathcal{X}(\mathcal{P}_+(I))$,

$$\begin{aligned}\mathfrak{g}(\nabla_X Y, Z) =& \frac{1}{2}\{X\mathfrak{g}(Y,Z) + Y\mathfrak{g}(Z,X) - Z\mathfrak{g}(X,Y)\} \\ &+ \frac{1}{2}\{\mathfrak{g}([X,Y],Z) - \mathfrak{g}([Y,Z],X) - \mathfrak{g}([X,Z],Y)\}. \end{aligned} \tag{V.14}$$

Now, in the Koszul formula we set $X = \partial x_i, Y = \partial x_j, Z = \partial x_l$. We get

$$\Gamma_{ij}^k = \frac{1}{2}\sum_{l=1}^{n} g^{kl}\left(g_{il,j} + g_{jl,i} - g_{ij,l}\right), \quad \text{for } i,j,k \in \{1,...,n\}, \tag{V.15}$$

where $g_{il,j} = \dfrac{\partial g_{il}}{\partial x_j}$, $g_{jl,i} = \dfrac{\partial g_{jl}}{\partial x_i}$, and $g_{ij,l} = \dfrac{\partial g_{ij}}{\partial x_l}$. In the local coordinate system, the Fisher-Rao metric and its inverse are given by

$$g_{ij} = \begin{cases} \dfrac{1}{x_i} + \dfrac{1}{1-\sum_{h=1}^n x_h}, & \text{if } i = j, \\ \dfrac{1}{1-\sum_{h=1}^n x_h}, & \text{if } i \neq j, \end{cases} \tag{V.16}$$

$$g^{ij} = \begin{cases} x_i(1-x_i), & \text{if } i = j, \\ -x_i x_j, & \text{if } i \neq j, \end{cases} \tag{V.17}$$

for $i, j = 1, \ldots, n$. Now if we take the derivative of (V.16) by x_l, we get

$$g_{ij,l} = \begin{cases} -\dfrac{1}{(x_i)^2} + \dfrac{1}{(1-\sum_{h=1}^n x_h)^2}, & \text{if } i = j = l, \\ \dfrac{1}{(1-\sum_{h=1}^n x_h)^2}, & \text{otherwise.} \end{cases} \tag{V.18}$$

Replace (V.18) in (V.15), the formula follows. ■

The tangent space $T\mathcal{P}_+(I)$ is trivial, it is the product of $\mathcal{P}_+(I)$ and $\mathcal{S}_0(I)$. We can define a constant vector field as following.

Definition V.3

Let $X \in \mathcal{X}(\mathcal{P}_+(I))$ be a vector field on $\mathcal{P}_+(I)$ and let $(\mathcal{P}_+(I), \varphi)$ the local coordinate. Then X has the representation $X = \sum_{i=1}^n X_i \partial x_i$, and X is called a constant vector field on $\mathcal{P}_+(I)$ if all X_i are independent of μ.

The connection of constant vector fields is given the following theorem.

Theorem V.1

Given two constant vector fields X, Y on $\mathcal{P}_+(I)$, the Levi-Civita connection at $\mu \in \mathcal{P}_+(I)$ is given by

$$\nabla_X Y(\mu) = -\frac{1}{2}\left(\frac{dX}{d\mu}\frac{dY}{d\mu} - \mathfrak{g}_\mu(X,Y)\right)\mu. \tag{V.19}$$

Proof Let $X = \sum_{i\in I} X_i\delta^i, Y = \sum_{i\in I} Y_i\delta^i$ and $Z = \sum_{i\in I} Z_i\delta^i$ be constant vector fields on $\mathcal{P}_+(I)$. Thus, we get $[X,Y] = [Y,Z] = [X,Z] = 0$ and consequently (V.14) gives

$$\mathfrak{g}(\nabla_X Y, Z) = \frac{1}{2}\left\{X\mathfrak{g}(Y,Z) + Y\mathfrak{g}(X,Z) - Z\mathfrak{g}(X,Y)\right\}. \tag{V.20}$$

Set $\mu = \sum_{i\in I}\mu_i\delta^i \in \mathcal{P}_+(I)$, and $\alpha(t) = \mu + vt$, a curve on $\mathcal{P}_+(I)$ such that $\mu(0) = \mu$ and $\dot{\mu}(0) = v = X(\mu)$. We have

$$\begin{aligned} X\mathfrak{g}_\mu(Y,Z) &= \frac{d}{dt}\Big|_{t=0}\mathfrak{g}_{\mu(t)}(Y,Z) \\ &= \frac{d}{dt}\Big|_{t=0}\sum_{i\in I}\frac{Y_iZ_i}{\mu_i + tv_i} \\ &= -\sum_{i\in I}\frac{v_iY_iZ_i}{\mu_i^2} = -\sum_{i\in I}\frac{X_iY_iZ_i}{\mu_i^2}. \end{aligned}$$

Similarly, one obtains formulae for $Y\mathfrak{g}(X,Z)$ and $Z\mathfrak{g}(X,Y)$. Now replacing the above results in (V.20), we get

$$\begin{aligned} \mathfrak{g}_\mu(\nabla_X Y, Z) &= \frac{1}{2}\left\{-\sum_{i\in I}\frac{X_iY_iZ_i}{\mu_i^2} - \sum_{i\in I}\frac{X_iY_iZ_i}{\mu_i^2} + \sum_{i\in I}\frac{X_iY_iZ_i}{\mu_i^2}\right\} \\ &= -\frac{1}{2}\sum_{i\in I}\frac{X_iY_iZ_i}{\mu_i^2}. \end{aligned} \tag{V.21}$$

On the other hand, we have

$$\sum_{i\in I}\mathfrak{g}_\mu(X,Y)Z_i = \mathfrak{g}_\mu(X,Y)\sum_{i\in I}Z_i = 0, \tag{V.22}$$

since Z is a vector field on $\mathcal{P}_+(I)$. Then (V.21) can be written as

$$\begin{aligned} \mathfrak{g}_\mu(\nabla_X Y, Z) &= -\frac{1}{2}\sum_{i\in I}\left(\frac{X_iY_i}{\mu_i^2} - \mathfrak{g}_\mu(X,Y)\right)\mu_i\frac{Z_i}{\mu_i} \\ &= \mathfrak{g}_\mu\left(-\frac{1}{2}\left(\frac{dX}{d\mu}\frac{dY}{d\mu} - \mathfrak{g}_\mu(X,Y)\right)\mu, Z\right). \end{aligned}$$

This holds for every constant vector field Z, which completes the proof. ■

V.3.3 Geodesics curves on $\mathcal{P}_+(I)$

A geodesic curve is the autoparallel curve. By applying the formula for connection of constant vector fields, we derive the explicit formula of geodesics on $\mathcal{P}_+(I)$.

Theorem V.2

Let $\mu = \sum_{i\in I} \mu_i \delta^i$ be a probability measure in $\mathcal{P}_+(I)$ and $v \in T_\mu \mathcal{P}_+(I)$ a unit tangent vector, i.e., $||v||_\mu = 1$. Then the geodesic α that satisfies $\alpha(0) = \mu$ and $\dot\alpha(0) = v$ is given by $\alpha(t) = \sum_{i\in I} \alpha_i(t)\delta^i$ with

$$\alpha_i(t) = \left(\cos\frac{t}{2} + \frac{\dot\alpha_i(0)}{\alpha_i(0)}\sin\frac{t}{2}\right)^2 \alpha_i(0), \tag{V.23}$$

in which $\alpha_i(0) = \mu_i$ and $\dot\alpha_i(0) = v_i$, $\forall i \in I$.

Proof Let $\alpha(t) = \sum_{i\in I} \alpha_i(t)\delta^i$ and $\dot\alpha(t) = \sum_{i\in I} \dot\alpha_i(t)\delta^i$. Then for each t, we have

$$\begin{cases} \sum_{i\in I} \alpha_i(t) &= 1, \quad \text{and } \alpha_i(t) > 0, \forall i \in I, \\ \sum_{i\in I} \dot\alpha_i(t) &= 0. \end{cases} \tag{V.24}$$

Set X a constant vector field in $\mathcal{P}_+(I)$. From the condition (V.11) of Levi–Civita connection, we have

$$\mathfrak{g}_{\alpha(t)}(\nabla_{\dot\alpha(t)}\dot\alpha(t), X) = \dot\alpha(t)\left(\mathfrak{g}_{\alpha(t)}(\dot\alpha(t), X)\right) - \mathfrak{g}_{\alpha(t)}(\dot\alpha(t), \nabla_{\dot\alpha(t)}X). \tag{V.25}$$

With the properties of Levi-Civita connection, to compute $\nabla_{\dot\alpha(t)}X$, the tangent vector $\dot\alpha(t)$ can be considered as a constant vector field on $\mathcal{P}_+(I)$ when t is fixed. Therefore, applying (V.19) for $\dot\alpha(t)$ and X we get,

$$\begin{aligned} \nabla_{\dot\alpha(t)}X &= -\frac{1}{2}\left(\frac{d\dot\alpha(t)}{d\alpha(t)}\frac{dX}{d\alpha(t)} - \mathfrak{g}_{\alpha(t)}(\dot\alpha(t), X)\right)\alpha(t) \\ &= -\frac{1}{2}\sum_{i\in I}\left(\frac{\dot\alpha_i}{\alpha_i}\frac{X_i}{\alpha_i} - \sum_{j\in I}\frac{\dot\alpha_j X_j}{\alpha_j}\right)\alpha_i\delta^i. \end{aligned} \tag{V.26}$$

Taking into account (V.24), the last term in (V.25) becomes

$$\begin{aligned} \mathfrak{g}(\dot\alpha(t), \nabla_{\dot\alpha(t)}X) &= \left\langle \frac{d\dot\alpha}{d\alpha}, \frac{d\nabla_{\dot\alpha(t)}X}{d\alpha}\right\rangle_{\alpha(t)} = -\frac{1}{2}\sum_{i\in I}\frac{\dot\alpha_i}{\alpha_i}\left(\frac{\dot\alpha_i}{\alpha_i}\frac{X_i}{\alpha_i} - \sum_{j\in I}\frac{\dot\alpha_j X_j}{\alpha_j}\right)\alpha_i \\ &= -\frac{1}{2}\sum_{i\in I}\frac{\dot\alpha_i^2 X_i}{\alpha_i^2}. \end{aligned} \tag{V.27}$$

Now, we need to compute the second term in (V.25). Thus, we have

$$\dot\alpha(t)\left(\mathfrak{g}_{\alpha(t)}(\dot\alpha(t), X)\right) = \frac{d}{dt}\mathfrak{g}_{\alpha(t)}(\dot\alpha(t), X) = \sum_{i\in I}\frac{d}{dt}\left(\frac{\dot\alpha_i}{\alpha_i}\right)X_i. \tag{V.28}$$

Combining (V.27) and (V.28) in (V.25), we get

$$\mathfrak{g}_{\alpha(t)}(\nabla_{\dot\alpha(t)}\dot\alpha(t), X) = \sum_{i\in I}\left(\frac{d}{dt}\left(\frac{\dot\alpha_i}{\alpha_i}\right) + \frac{1}{2}\frac{\dot\alpha_i^2}{\alpha_i^2}\right) X_i. \tag{V.29}$$

Let's define the function $F(t)$ as

$$F(t) = -\sum_{i\in I}\left(\frac{d}{dt}\left(\frac{\dot\alpha_i}{\alpha_i}\right) + \frac{1}{2}\frac{\dot\alpha_i^2}{\alpha_i^2}\right)\alpha_i(t) = -\sum_{i\in I}\frac{d}{dt}\left(\frac{\dot\alpha_i}{\alpha_i}\right)\alpha_i(t) - \frac{1}{2}\mathfrak{g}_{\alpha(t)}(\dot\alpha(t),\dot\alpha(t)). \tag{V.30}$$

Hence, the measure

$$\nu(t) = \sum_{i\in I}\left(\frac{d}{dt}\left(\frac{\dot\alpha_i}{\alpha_i}\right) + \frac{1}{2}\frac{\dot\alpha_i^2}{\alpha_i^2} + F(t)\right)\alpha_i\delta^i \tag{V.31}$$

belongs to $T_{\alpha(t)}\mathcal{P}_+(I)$. In this way, (V.29) can be written as $\mathfrak{g}_\alpha(\nabla_{\dot\alpha}\dot\alpha, X) = \mathfrak{g}_\alpha(\nu, X)$. And since X is an arbitrary constant vector field, we get

$$\nabla_{\dot\alpha}\dot\alpha = \nu = \sum_{i\in I}\left(\frac{d}{dt}\left(\frac{\dot\alpha_i}{\alpha_i}\right) + \frac{1}{2}\frac{\dot\alpha_i^2}{\alpha_i^2} + F(t)\right)\alpha_i\delta^i. \tag{V.32}$$

Therefore, $\alpha(t) = \sum_{i\in I}\alpha_i(t)\delta^i$ is a geodesic if and only if

$$\begin{cases} \frac{d}{dt}\left(\frac{\dot\alpha_i}{\alpha_i}\right) + \frac{1}{2}\left(\frac{\dot\alpha_i}{\alpha_i}\right)^2 + F(t) = 0, \quad \forall i\in I, \\ \sum_{i\in I}\dot\alpha_i(t) = 0, \quad \forall t. \end{cases} \tag{V.33}$$

Our next goal is to solve (V.33). We may remark that if α is a geodesic then $\mathfrak{g}_{\alpha(t)}(\dot\alpha(t),\dot\alpha(t))$ is constant along $\alpha(t)$. Consequently, taking into account the assumption that $||\dot\gamma(0)||_\mu = 1$, we can assert that

$$\mathfrak{g}_{\alpha(t)}(\dot\alpha(t),\dot\alpha(t)) = \sum_{i\in I}\frac{\dot\alpha_i^2}{\alpha_i} \equiv 1. \tag{V.34}$$

Then we have

$$\sum_{i\in I}\frac{d}{dt}\left(\frac{\dot\alpha_i}{\alpha_i}\right)\alpha_i = \frac{d}{dt}\sum_{i\in I}\left(\frac{\dot\alpha_i}{\alpha_i}\alpha_i\right) - \sum_{i\in I}\frac{\dot\alpha_i^2}{\alpha_i} = -1. \tag{V.35}$$

Which translates to $F(t) = \frac{1}{2}$. Substituting this result in (V.33), we obtain

$$\frac{d}{dt}\left(\frac{\dot\alpha_i}{\alpha_i}\right) + \frac{1}{2}\left(\frac{\dot\alpha_i}{\alpha_i}\right)^2 + \frac{1}{2} = 0, \quad \forall i\in I. \tag{V.36}$$

We set $\omega_i(t) = \dfrac{\dot\alpha_i(t)}{\alpha_i(t)}$, and we rewrite equation (V.36) as

$$\frac{d}{dt}\omega_i + \frac{1}{2}\omega_i^2 + \frac{1}{2} = 0, \quad \forall i\in I, \tag{V.37}$$

The solution of this differential equation is given by $\omega_i = \tan\left(-\frac{t}{2} + \Theta^i\right)$, where Θ^i is

constant for $i \in I$. Hence, we have

$$\frac{\dot{\alpha}_i}{\alpha_i} = \tan\left(-\frac{1}{2}t + \Theta^i\right), \quad \forall i \in I,$$

and $\alpha_i(t) = \Omega^i \cos^2\left(-\frac{t}{2} + \Theta^i\right)$, where Ω^i is constant, and $i \in I$. Taking into account initial conditions, we conclude that

$$\Theta^i = \arctan\left(\frac{\dot{\alpha}_i(0)}{\alpha_i(0)}\right), \tag{V.38}$$

$$\Omega^i = \frac{\alpha_i^2(0) + \dot{\alpha}_i^2(0)}{\alpha_i(0)}. \tag{V.39}$$

Which proves the theorem. ■

When the initial vector is not normalized, the geodesic curve is given by the following corollary.

Corollary V.1

The geodesic $\alpha(t)$ with $\alpha(0) = \mu$ and $\dot{\alpha}(0) = v$, where v is a nontrivial tangent vector and not necessary unit, is given by

$$\alpha(t) = \sum_{i\in I}\left(\cos\frac{t\|v\|_\mu}{2} + \frac{v_i}{\mu_i\|v\|_\mu}\sin\frac{t\|v\|_\mu}{2}\right)^2 \mu_i\delta^i. \tag{V.40}$$

Proof Indeed, we can check that the curve is auto parallel and satisfies the initial conditions. ■

The distance between two measure is equal to the length of the geodesic segment connecting them. But instead of computing the difficult integral, we transform to compute the length of the geodesic on the sphere, which is much practical.

Proposition V.3

The Fisher Rao distance $d^{FR} : \mathcal{P}_+(I) \times \mathcal{P}_+(I) \to [0, \pi)$ between two measures $\mu, \nu \in \mathcal{P}_+(I)$ under the Fisher-Rao metric is given by

$$d^{FR}(\mu, \nu) = 2\arccos\left(\sum_{i\in I}\sqrt{\mu_i\nu_i}\right). \tag{V.41}$$

To prove Proposition V.3, we remind the following lemma [8].

Lemma V.1

Let

$$\mathbb{S}^+_{(0,2)}(I) = \left\{f \in \mathcal{F}(I) \mid f^i > 0, \forall i \in I \text{ and } \sum_{i\in I}(f^i)^2 = 4\right\}$$

be the positive sector of the sphere centered at 0 *with radius* 2*. As a submanifold of* $\mathcal{F}(I)$ *it carries the induced standard metric of* $\mathcal{F}(I)$*. That is for a given point* $f \in \mathbb{S}^+_{(0,2)}(I)$ *and two tangents vectors* $p, q \in T_f\mathbb{S}^+_{(0,2)}(I)$*, we have*

$$\langle p, q\rangle_f = \sum_{i\in I} p^i q^i. \tag{V.42}$$

Then the map Φ *given by*

$$\begin{aligned} \Phi : \mathcal{P}_+(I) &\longrightarrow \mathbb{S}^+_{(0,2)}(I) \\ \mu = \sum_{i\in I} \mu_i \delta^i &\longmapsto 2\sum_{i\in I} \sqrt{\mu_i} e_i \end{aligned} \tag{V.43}$$

is an isometry.

Proof [Proof of the lemma] It is clear that Φ is bijective. Now, let v, w be in $T_\mu\mathcal{P}_+(I)$. We have

$$\begin{aligned} \left\langle \frac{\partial\Phi}{\partial v}(\mu), \frac{\partial\Phi}{\partial w}(\mu) \right\rangle &= \left\langle \frac{d}{dt}\Phi(\mu + vt)\Big|_{t=0}, \frac{d}{dt}\Phi(\mu + wt)\Big|_{t=0} \right\rangle \\ &= \left\langle \sum_{i\in I} \frac{v_i}{\sqrt{\mu_i}} e_i, \sum_{i\in I} \frac{w_i}{\sqrt{\mu_i}} e_i \right\rangle \\ &= \sum_{i\in I} \frac{v_i w_i}{\mu_i} = \mathfrak{g}_\mu(v, w). \end{aligned}$$

■

Proof [Proof of the Proposition] By virtue of Lemma V.1, we get

$$d^{FR}(\mu, \nu) = d(\Phi(\mu), \Phi(\nu)) = 2\arccos\left(\sum_{i\in I} \sqrt{\mu_i \nu_i}\right).$$

■

We remark that the Riemannian geometry of $\mathcal{P}_+(I)$ is well known in differential geometry. However, we provide here explicit computations in special coordinates. Now we have the explicit formula of the geodesic segment connecting two measures.

Theorem V.3

Let μ, ν *be two different probability measures in* $\mathcal{P}_+(I)$*. Then there exists a unique geodesic* $\alpha : [0, l] \to \mathcal{P}_+(I)$, $t \to \alpha(t)$*, joining two points* μ *and* ν*, with* $\alpha(0) = \mu$*,* $\alpha(l) = \nu$ *and* $l = d^{FR}(\mu, \nu)$*, given by*

$$\alpha(t) = \sum_{i\in I} \left(\cos\frac{tl}{2} + \frac{d\tau}{d\mu}(i)\sin\frac{tl}{2}\right)^2 \mu_i \delta^i, \tag{V.44}$$

where τ is the tangent vector in $T_\mu\mathcal{P}_+(I)$ defined by

$$\tau = \frac{1}{\sin\frac{l}{2}}\sum_{i\in I}\left(\sqrt{\frac{d\nu}{d\mu}} - \sum_{j\in I}\sqrt{\frac{d\nu}{d\mu}}(j)\mu(j)\right)\mu_i\delta^i. \tag{V.45}$$

Proof The proof falls naturally into three steps.

Step 1 First, let us check that τ is a tangent vector in $T_\mu\mathcal{P}_+(I)$. Indeed,

$$\begin{aligned}\frac{1}{\sin\frac{l}{2}}\sum_{i\in I}\left(\sqrt{\frac{d\nu}{d\mu}}(i) - \sum_{j\in I}\sqrt{\frac{d\nu}{d\mu}}(j)\mu(j)\right)\mu_i &= \frac{1}{\sin\frac{l}{2}}\left(\sum_{i\in I}\sqrt{\frac{d\nu}{d\mu}}(i)\mu_i - \sum_{j\in I}\sqrt{\frac{d\nu}{d\mu}}(j)\mu(j)\right)\\ &= 0.\end{aligned} \tag{V.46}$$

Then, since

$$\left(\sum_{j\in I}\sqrt{\frac{d\nu}{d\mu}}(j)\mu(j)\right)^2 = \left(\sum_{j\in I}\sqrt{\mu_j\nu_j}\right)^2 = \cos^2\frac{l}{2}. \tag{V.47}$$

it follows that

$$\begin{aligned}\langle\tau,\tau\rangle_\mu &= \frac{1}{\sin^2\frac{l}{2}}\sum_{i\in I}\left(\sqrt{\frac{d\nu}{d\mu}}(i) - \sum_{j\in I}\sqrt{\frac{d\nu}{d\mu}}(j)\mu(j)\right)^2\mu_i\\ &= \frac{1}{\sin^2\frac{l}{2}}\left(\sum_{i\in I}\nu(i) - \left(\sum_{j\in I}\sqrt{\frac{d\nu}{d\mu}}(j)\mu(j)\right)^2\right)\\ &= \frac{1}{\sin^2\frac{l}{2}}\left(1-\cos^2\frac{l}{2}\right) = 1.\end{aligned} \tag{V.48}$$

hence τ is a unit tangent vector.

Step 2 Now let us examine that the curve $\alpha(t)$ defined in (V.44) satisfies $\alpha(0) = \mu$ and $\alpha(1) = \nu$. It is easily seen that for $t = 0$, $\alpha(0) = \mu$. Now for $t = l$, we have

$$\alpha(l) = \sum_{i\in I}\left(\cos\frac{l}{2} + \frac{d\tau}{d\mu}(i)\sin\frac{l}{2}\right)^2\mu_i\delta^i. \tag{V.49}$$

By (V.45) we get

$$\begin{aligned}\frac{d\tau}{d\mu}\sin\frac{l}{2} &= \sum_{i\in I}\left(\sqrt{\frac{d\nu}{d\mu}}(i) - \sum_{j\in I}\sqrt{\frac{d\nu}{d\mu}}(j)\mu(j)\right)e_i\\ &= \sum_{i\in I}\left(\sqrt{\frac{d\nu}{d\mu}}(i) - \cos\frac{l}{2}\right)e_i.\end{aligned} \tag{V.50}$$

Hence,

$$\alpha(l) = \sum_{i\in I}\left(\cos\frac{l}{2} + \sqrt{\frac{d\nu}{d\mu}}(i) - \cos\frac{l}{2}\right)^2\mu_i\delta^i = \sum_{i\in I}\nu_i\delta^i = \nu. \tag{V.51}$$

Step 3 Now we go to prove the uniqueness of the curve. Let $\mu(t) = \exp_\mu\tau t$ and $\tilde{\mu}(t) = \exp_\mu\tilde{\tau}t$ be unit speed geodesics corresponding to τ and $\tilde{\tau}$, and satisfying

$\mu(0) = \tilde{\mu}(0) = \mu$ and $\mu(l) = \tilde{\mu}(l) = \nu$. By means of Theorem V.2, we have

$$\mu(t) = \sum_{i\in I}\left(\cos\frac{t}{2} + \frac{d\tau}{d\mu}\sin\frac{t}{2}\right)^2 \mu_i\delta^i, \tag{V.52}$$

$$\tilde{\mu}(t) = \sum_{i\in I}\left(\cos\frac{t}{2} + \frac{d\tilde{\tau}}{d\mu}\sin\frac{t}{2}\right)^2 \mu_i\delta^i. \tag{V.53}$$

From later condition, we have

$$\left(\cos\frac{l}{2} + \frac{d\tau}{d\mu}(i)\sin\frac{l}{2}\right)^2 = \left(\cos\frac{l}{2} + \frac{d\tilde{\tau}}{d\mu}(i)\sin\frac{l}{2}\right)^2, \forall i \in I \tag{V.54}$$

$$\Rightarrow \cos\frac{l}{2} + \frac{d\tau}{d\mu}(i)\sin\frac{l}{2} = \pm\left(\cos\frac{l}{2} + \frac{d\tilde{\tau}}{d\mu}(i)\sin\frac{l}{2}\right), \forall i \in I. \tag{V.55}$$

Define

$$I_\pm = \left\{ i \in I \middle| \cos\frac{l}{2} + \frac{d\tau}{d\mu}(i)\sin\frac{l}{2} = \pm\left(\cos\frac{l}{2} + \frac{d\tilde{\tau}}{d\mu}(i)\sin\frac{l}{2}\right)\right\} \tag{V.56}$$

Then we have $I_- \cup I_+ = I$. Moreover $I_- \cap I_+ = \emptyset$. Indeed, if there exists $i \in I_- \cap I_+$ then

$$\nu_i = \left(\cos\frac{t}{2} + \frac{d\tau}{d\mu}\sin\frac{t}{2}\right)^2 \mu_i = 0, \tag{V.57}$$

contradict to $\nu \in \mathcal{P}_+$. Sine $0 < l < \pi$, we have

$$I_+ = \{i \in I | \tau_i = \tilde{\tau}_i\}, \tag{V.58}$$

$$I_- = \left\{ i \in I | \tau_i + \tilde{\tau}_i = -2\mu_i \cot\frac{l}{2}\right\}. \tag{V.59}$$

Suppose $I_- \neq \emptyset$, since τ and $\tilde{\tau}$ are unit tangent vectors at μ, we have

$$\sum_{i\in I_+}\tau_i + \sum_{i\in I_-}\tau_i = \sum_{i\in I_+}\tilde{\tau}_i + \sum_{i\in I_-}\tilde{\tau}_i = 0 \tag{V.60}$$

$$\Rightarrow \sum_{i\in I_-}\left(\tilde{\tau}_i + 2\mu_i\cot\frac{l}{2}\right) + \sum_{i\in I_-}\tilde{\tau}_i = 0. \tag{V.61}$$

Since (V.61) we see that if $I_- = I$, then $\cot\frac{l}{2} = 0$ contradicts to $0 < l < \pi$. So $I_- \neq I$. We have the claim below.

Claim For all $\mu \in \mathcal{P}_+(I)$ and $0 < l < \pi$. If $\tau, \tilde{\tau} \in T_\mu\mathcal{P}_+(I)$. Let

$$I_+ = \{i \in I | \tau_i = \tilde{\tau}_i\}, \tag{V.62}$$

$$I_- = \left\{ i \in I | \tau_i + \tilde{\tau}_i = -2\mu_i \cot\frac{l}{2}\right\}, \tag{V.63}$$

then $I_- = \emptyset$.

By means of the Claim, we prove the uniqueness of the geodesic (V.44) defined with the unit tangent vector (V.45). ■

The proof of the Claim will be given in Appendix B.

By the Theorem II.7 of the Chapter II (Background), we have for any $\mu \in \mathcal{P}_+(I)$ there exist a neighborhood $\mathcal{O}_0 \subset T_\mu\mathcal{P}_+(I)$ and neighborhood $\mathcal{O}_\mu \subset \mathcal{P}_+(I)$ such that the mapping $\alpha(1, \mu, v)$ is a diffeomorphism. This implies that the set $\mathcal{O}_0$ contains the vector v such that the geodesic is well defined on the interval $[0, 1]$. Furthermore, the value of the geodesic at time 1 is the exponential map. We have the explicit definition for the exponential map and the logarithm map.

Corollary V.2

Let $\mu \in \mathcal{P}_+(I)$, and let $\mathcal{O}_0, \mathcal{O}_\mu$ be the neighborhoods of 0 and μ in $T_\mu\mathcal{P}_+(I)$ and $\mathcal{P}_+(I)$ such that the exponential map $\exp_\mu$ is well defined. Then for $v \in \mathcal{O}_0$, the exponential map is given by

$$\exp_\mu(v) = \sum_{i\in I} \left(\cos \frac{\|v\|_\mu}{2} + \frac{v_i}{\mu_i \|v\|_\mu} \sin \frac{\|v\|_\mu}{2} \right)^2 \mu_i \delta^i. \tag{V.64}$$

The logarithmic map $\log_\mu$, as the inverse of $\exp_\mu$, is given by

$$\log_\mu : \mathcal{O}_\mu \longrightarrow \mathcal{O}_0$$

$$\nu \longmapsto \log_\mu(\nu) = \frac{l}{\sin \frac{l}{2}} \sum_{i\in I} \left(\sqrt{\frac{d\nu}{d\mu}}(i) - \sum_{j\in I} \sqrt{\frac{d\nu}{d\mu}}(j)\mu(j) \right) \mu_i \delta^i. \tag{V.65}$$

Proof The corollary follows directly from (V.40) when letting $t = 1$, and Theorem V.3, where $\log_\mu(\nu) = l\tau$. ■

V.3.4 Levi-Civita parallel transport on $\mathcal{P}_+(I)$

Let us consider two points $\mu, \nu \in \mathcal{P}_+(I)$, a tangent vector $v \in T_\mu\mathcal{P}_+(I)$ and a geodesic curve $\alpha : [0, l] \to \mathcal{P}_+(I)$ on $\mathcal{P}_+(I)$ such that $\alpha(0) = \mu$ and $\alpha(l) = \nu$. We would like to map v from $T_\mu\mathcal{P}_+(I) = T_{\alpha(0)}\mathcal{P}_+(I)$ to $T_\nu\mathcal{P}_+(I) = T_{\alpha(l)}\mathcal{P}_+(I)$. We introduce X, a vector field defined along the geodesic α, such that $X(\mu) = v$ and $\nabla_{\dot\alpha(t)} X(\alpha(t)) = 0$. We say that the vector field X is constant along the geodesic curve α with respect to ∇.

Definition V.4

A metric parallel transport on $\mathcal{P}_+(I)$ is the map

$$\Gamma_{\alpha(0)\rightarrowtail\alpha(t)} : T_{\alpha(0)}\mathcal{P}_+(I) \to T_{\alpha(t)}\mathcal{P}_+(I) \tag{V.66}$$

such that for every $v, w \in T_\mu\mathcal{P}_+(I)$*, and for any* $t \in [0, l]$ *we have*

$$\mathfrak{g}_{\alpha(0)}(v, w) = \mathfrak{g}_{\alpha(t)}\left(\Gamma_{\alpha(0)\to\alpha(t)}(v), \Gamma_{\alpha(0)\to\alpha(t)}(w)\right). \tag{V.67}$$

If Γ is the Levi-Civita parallel transport (corresponding with Levi-Civita connection), then Γ is metric. Rewriting equation $\nabla_{\dot\alpha(t)}X(\alpha(t)) = 0$, we conclude that computing $X(t) = X(\alpha(t))$ requires solving a linear first order differential equations on $\mathcal{P}_+(I)$ given by

$$\frac{dX_k}{dt} + \sum_{i,j}\alpha_{ij}^k\frac{d\alpha_i}{dt}X_j = 0, \quad \text{for } k = 1, \dots, n. \tag{V.68}$$

We check at once that it is difficult to solve equation (V.68) directly. Instead we will use equation (V.19).

Theorem V.4

Let μ *be a probability measure in* $\mathcal{P}_+(I)$ *and* $v \in T_\mu\mathcal{P}_+(I)$ *a unit tangent vector, i.e.,* $||v||_\mu = 1$*. Let* $\alpha : [0, l] \to \mathcal{P}_+(I)$ *be a geodesic curve such that* $\alpha(0) = \mu$ *and* $\dot\alpha(0) = v$*. The Levi-Civita parallel transport of a vector* $w \in T_\mu\mathcal{P}_+(I)$ *to* $T_{\alpha(t)}\mathcal{P}_+(I)$*, is given by*

$$\Gamma_{\mu\rightarrowtail\alpha(t)}(w) = \sum_{i\in I}\sqrt{\alpha_i(t)}\left(-F_0\sqrt{\mu_i}\left(2\sin\frac{t}{2} - 2\frac{v_i}{\mu_i}\cos\frac{t}{2}\right) + \frac{w_i}{\sqrt{\mu_i}} - 2F_0\frac{v_i}{\sqrt{\mu_i}}\right)\delta^i, \tag{V.69}$$

where $F_0 = \frac{1}{2}\mathfrak{g}_\mu(v, w)$ *is constant.*

Proof We can proceed analogously to the proof of Theorem V.2. Thus, let $\alpha(t) = \sum_{i\in I}\alpha_i(t)\delta^i$ be a geodesic curve, and define $\dot\alpha(t) = \sum_{i\in I}\dot\alpha_i(t)\delta^i$. Consider the vector field X on α defined by $X(\alpha(t)) = \sum_{i\in I}X_i(\alpha(t))\delta^i$, for $t \in [0, l]$, as the parallel transport of vector w along α. Then

$$\begin{cases}\nabla_{\dot\alpha(t)}X(t) = 0\\ X(0) = w\end{cases}, \tag{V.70}$$

where we write $X(\alpha(t))$ simply $X(t)$ when no confusion can arise. Let Y be a constant vector field (in the sense of Definition V.3) on $\mathcal{P}_+(I)$, we have

$$\mathfrak{g}_{\alpha(t)}\left(\nabla_{\dot\alpha(t)}X(t), Y\right) = \dot\alpha(t)\left(\mathfrak{g}_{\alpha(t)}(X(t), Y)\right) - \mathfrak{g}_{\alpha(t)}\left(X(t), \nabla_{\dot\alpha(t)}Y\right). \tag{V.71}$$

Applying Theorem V.1, we get

$$\nabla_{\dot\alpha} Y = -\frac{1}{2}\sum_{i\in I}\left(\frac{\dot\alpha_i}{\alpha_i}\frac{Y_i}{\gamma_i} - \sum_{j\in I}\frac{\dot\alpha_j Y_j}{\alpha_j}\right)\alpha_i\delta^i. \tag{V.72}$$

Hence the last term in (V.71) becomes

$$\begin{aligned}\mathfrak{g}_\alpha(X, \nabla_{\dot\alpha} Y) &= -\frac{1}{2}\sum_{i\in I}\frac{X_i}{\alpha_i}\left(\frac{\dot\alpha_i}{\alpha_i}\frac{Y_i}{\alpha_i} - \sum_{j\in I}\frac{\dot\alpha_j Y_j}{\alpha_j}\right)\alpha_i \\ &= -\frac{1}{2}\sum_{i\in I}\frac{X_i Y_i \dot\alpha_i}{\alpha_i^2}.\end{aligned} \tag{V.73}$$

Let us now compute the second term in (V.71). We obtain

$$\dot\alpha(t)\left(\mathfrak{g}_{\alpha(t)}(X,Y)\right) = \frac{d}{dt}\mathfrak{g}_{\alpha(t)}(X(t),Y) = \sum_{i\in I}\frac{d}{dt}\left(\frac{X_i}{\alpha_i}\right)Y_i. \tag{V.74}$$

Consequently, equation (V.71) becomes

$$\mathfrak{g}_\alpha(\nabla_{\dot\alpha} X, Y) = \sum_{i\in I}\left(\frac{d}{dt}\left(\frac{X_i}{\alpha_i}\right) + \frac{1}{2}\frac{X_i\dot\alpha_i}{\alpha_i^2}\right)Y_i. \tag{V.75}$$

Define the function $F(t)$ by

$$\begin{aligned}F(t) &= -\sum_{i\in I}\left(\frac{d}{dt}\left(\frac{X_i}{\alpha_i}\right) + \frac{1}{2}\frac{X_i\dot\alpha_i}{\alpha_i^2}\right)\alpha_i(t) \\ &= -\sum_{i\in I}\frac{d}{dt}\left(\frac{X_i}{\alpha_i}\right)\alpha_i(t) - \frac{1}{2}\mathfrak{g}_{\alpha(t)}(X(t),\dot\alpha(t)).\end{aligned} \tag{V.76}$$

Then, $\forall t\in[0,l]$, the probability measure

$$\nu(t) = \sum_{i\in I}\left(\frac{d}{dt}\left(\frac{X_i}{\alpha_i}\right) + \frac{1}{2}\frac{X_i\dot\alpha_i}{\alpha_i^2} + F(t)\right)\alpha_i\delta^i$$

belongs to $T_{\alpha(t)}\mathcal{P}_+(I)$. Thus, Equation (V.75) can be written as

$$\mathfrak{g}_\alpha(\nabla_{\dot\alpha} X, Y) = \mathfrak{g}_\alpha(\nu, Y). \tag{V.77}$$

Since Y is an arbitrary constant vector field, we get

$$\nabla_{\dot\alpha} X = \nu = \sum_{i\in I}\left(\frac{d}{dt}\left(\frac{X_i}{\alpha_i}\right) + \frac{1}{2}\frac{X_i\dot\alpha_i}{\alpha_i^2} + F(t)\right)\alpha_i\delta^i. \tag{V.78}$$

Therefore, $X(t)$ is the parallel transport of the vector w along the geodesic curve $\alpha(t)$ if and only if

$$\begin{cases}\dfrac{d}{dt}\left(\dfrac{X_i}{\alpha_i}\right) + \dfrac{1}{2}\dfrac{X_i\dot\alpha_i}{\alpha_i^2} + F(t) = 0, \quad \forall i\in I, \\ X(0) = w.\end{cases} \tag{V.79}$$

Our next concern will be to solve equation (V.79). We remind that $\mathfrak{g}_{\alpha(t)}(X(t),\dot\alpha(t)) = \mathfrak{g}_{\alpha(0)}(X(0),\dot\alpha(0))$. Moreover

$$\sum_{i\in I}\frac{d}{dt}\left(\frac{X_i}{\alpha_i}\right)\alpha_i = \frac{d}{dt}\sum_{i\in I}\left(\frac{X_i}{\alpha_i}\alpha_i\right) - \sum_{i\in I}\left(\frac{X_i\dot\alpha_i}{\alpha_i}\right) = -\mathfrak{g}_{\alpha(0)}(X(0),\dot\alpha(0)). \tag{V.80}$$

Which gives that $F(t)$ is a constant function and $F(t) = F_0 = \frac{1}{2}\mathfrak{g}_{\alpha(0)}(X(0), \dot{\alpha}(0))$. Hence, substituting this result in equation (V.79) we get

$$\frac{d}{dt}\left(\frac{X_i}{\alpha_i}\right) + \frac{1}{2}\frac{X_i\dot{\alpha}_i}{\alpha_i^2} + F_0 = 0, \quad \forall i \in I. \tag{V.81}$$

Set $\omega_i = \dfrac{X_i}{\alpha_i}$. Equation (V.81) can be written as

$$\frac{d}{dt}\omega_i + \frac{1}{2}\frac{\dot{\alpha}_i}{\alpha_i}\omega_i + F_0 = 0, \quad \forall i \in I. \tag{V.82}$$

Solution of the first order differential equation (V.82) is given by

$$\omega_i(t) = \frac{1}{\sqrt{\alpha_i(t)}}\left(-F_0\sqrt{\alpha_i(0)}\left(2\sin\frac{t}{2} - 2\frac{\dot{\alpha}_i(0)}{\alpha_i(0)}\cos\frac{t}{2}\right) + \Theta_i\right), \text{ for } \Omega_i \text{ constant}, i \in I. \tag{V.83}$$

Therefore,

$$X_i = \sqrt{\alpha_i(t)}\left(-F_0\sqrt{\alpha_i(0)}\left(2\sin\frac{t}{2} - 2\frac{\dot{\alpha}_i(0)}{\alpha_i(0)}\cos\frac{t}{2}\right) + \Theta\right), \quad \text{for } \Theta_i \text{ constant}, i \in I. \tag{V.84}$$

According to the initial conditions, it follows that

$$\Theta = \frac{w_i}{\sqrt{\mu_i}} - 2F_0\frac{v_i}{\sqrt{\mu_i}}. \tag{V.85}$$

We conclude that

$$X_i(t) = \sqrt{\alpha_i(t)}\left(-F_0\sqrt{\mu_i}\left(2\sin\frac{t}{2} - 2\frac{v_i}{\mu_i}\cos\frac{t}{2}\right) + \frac{w_i}{\sqrt{\mu_i}} - 2F_0\frac{v_i}{\sqrt{\mu_i}}\right), \quad i \in I. \tag{V.86}$$

and it is easy to check that, $\forall t \in [0, l]$, $X(t) = \sum_{i\in I} X_i(t)\delta^i \in T_{\gamma(t)}P_+(I)$ and it is the Levi-Civita parallel transport of the vector w along the geodesic curve $\gamma(t)$. ■

Let the parallel transport between two measures be computed along the geodesic curve. The following theorem gives the explicit formula for the transportation.

Theorem V.5

Given two distinct probability measures μ and ν in $\mathcal{P}_+(I)$, a nontrivial tangent vector $w \in T_\mu\mathcal{P}_+(I)$ and the geodesic curve $\alpha : [0, l] \to \mathcal{P}_+(I)$ such that $\alpha(0) = \mu$ and $\alpha(l) = \nu$. The Levi-Civita parallel transport, $\Gamma_{\mu\rightarrowtail\nu} : T_\mu\mathcal{P}_+(I) \to T_\nu\mathcal{P}_+(I)$, that transports a vector w from $T_\mu\mathcal{P}_+(I) = T_{\alpha(0)}\mathcal{P}_+(I)$ to $T_\nu\mathcal{P}_+(I) = T_{\alpha(l)}\mathcal{P}_+(I)$

is given by

$$\Gamma_{\mu\rightarrow\nu}(w) = \sum_{i\in I}\sqrt{\nu_i}\left(-F_0\sqrt{\mu_i}\left(2\sin\frac{l}{2} - 2\frac{\tau_i}{\mu_i}\cos\frac{l}{2}\right) + \frac{w_i}{\sqrt{\mu_i}} - 2F_0\frac{\tau_i}{\sqrt{\mu_i}}\right)\delta^i, \tag{V.87}$$

where $l = 2\arccos\sum_{i\in I}\sqrt{\mu_i\nu_i}$, $F_0 = \frac{1}{2}\mathfrak{g}_\mu(w,\tau)$, *and* τ *is the unit tangent vector*

$$\tau = \frac{1}{\sin\frac{l}{2}}\sum_{i\in I}\left(\sqrt{\frac{d\nu}{d\mu}}(i) - \sum_{j\in I}\sqrt{\frac{d\nu}{d\mu}}(j)\mu(j)\right)\mu_i\delta^i. \tag{V.88}$$

Proof We make use of the geodesic curve $\alpha(t)$ joining two points μ and ν as detailed in Theorem V.3 together with taking $t = l$ in Theorem V.4. ■

We have studied the space $\mathcal{P}_+(I)$ with all the geometrical tools needed. In the next section we consider some statistical models on this space.

V.4 Transfer Learning

In this section, we consider the issue of transporting statistical models on the space of probability measures $\mathcal{P}_+(I)$. Specifically, we illustrate the benefits of exploiting the Riemannian geometry of $\mathcal{P}_+(I)$ and discuss how these tools can be incorporated into transfer learning problems. To address this issue, let $P_L = \{\mu^i\}_{i=1}^L$ and $P_S = \{\nu^i\}_{i=1}^S$ denote two populations in $\mathcal{P}_+(I)$, with P_L is of large size and P_S is of small size. We are interested in learning a statistical model from the dataset on P_S such as covariance matrix, PCA, and regression models while leveraging statistical information from the large dataset on P_L. Firstly, we need to identify the geometric mean of P_L and P_S. Indeed, for a set of data points $\{\mu^i\}_{i=1}^L \in \mathcal{P}_+(I)$, we consider the Karcher mean defined as

$$\mu^\star = \text{argmin}_{\mu\in\mathcal{P}_+(I)}\sum_{i=1}^{L} d^{FR}(\mu,\mu^i)^2 \tag{V.89}$$

where $d^{FR}(\mu,\mu^i)$ denotes the distance on $\mathcal{P}_+(I)$ determined with respect to the Fisher-Rao metric. Now suppose that we have computed the means μ^* and ν^* of the two populations P_L and P_S, we lift data points $P_L = \{\mu^i\}_{i=1}^L$ onto the tangent space $T_{\mu^*}\mathcal{P}_+(I)$ and the data points $P_S = \{\nu^i\}_{i=1}^S$ onto the tangent space $T_{\nu^*}\mathcal{P}_+(I)$ by means of the logarithm map defined on the set of probability measures $\mathcal{P}_+(I)$. The populations

are now represented in the tangent vector space. Indeed, projecting data into a vector space allows us to apply common statistical analysis methods. Let's denote the mapped data respectively by $\{v^i\}_{i=1}^L$ and $\{w^i\}_{i=1}^S$ with $v^i = \log_{\mu^*}(\mu^i), i = 1, ..., L$ and $w^i = \log_{\nu^*}(\nu^i), i = 1, ..., S$. We wish to state steps of our approach for transferring statistical models and we will discuss the performance of the parallel transport defined previously on $\mathcal{P}_+(I)$ with different applications in the next section.

V.4.1 Covariance and PCA from large populations

Since the space $\mathcal{P}_+(I)$ is nonlinear, the usual PCA cannot be applied to the large population P_L. A variant has been introduced in [41] as an extension of PCA for data lying on Riemannian manifolds. This method requires to solve a nonlinear optimization problem that is hard to solve in general [111]. In [24], the authors give an efficient and exact algorithm for manifolds with constant sectional curvature. In this chapter, we consider an approximation on the tangent space at μ^*, called TPCA.

Let $\{v^i\}_{i=1}^L$ in $T_{\mu^*}\mathcal{P}_+(I)$. We define the variance as

$$\sigma^2 = \frac{1}{L-1}\sum_{i=1}^{L} \|v^i\|_{\mu^*}^2 = \frac{1}{L-1}\sum_{i=1}^{L} \mathfrak{g}_{\mu^*}(v^i, v^i). \tag{V.90}$$

By the isometry map Φ, (V.43), maps $\mathcal{P}_+(I)$ into the sphere $\mathbb{S}^+_{(0,2)}(I)$, we have $\mathfrak{g}_{\mu^*}(v^i, v^i) = \langle d\Phi(\mu^*)\left[v^i\right], d\Phi(\mu^*)\left[v^i\right]\rangle$, where $\langle\cdot,\cdot\rangle$ is the usual inner product in the ambient space $\mathbb{R}^{n+1}$. The variance can be rewritten as

$$\sigma^2 = \frac{1}{L-1}\sum_{i=1}^{L} \left\langle d\Phi(\mu^*)\left[v^i\right], d\Phi(\mu^*)\left[v^i\right]\right\rangle. \tag{V.91}$$

We note that σ^2 captures the total variance of data and is equal to the trace of the covariance matrix $C_{\bar{V}}$ given in the following definition.

Definition V.5

Let $\{v^i\}_{i=1}^L$ be a dataset in $T_{\mu^*}\mathcal{P}_+(I)$ and let $\bar{V} = \left[\bar{v}^1, ..., \bar{v}^L\right] \in \mathbb{R}^{(n+1)\times L}$ be a matrix, where the i-th column $\bar{v}^i$ is simply the vector $d\Phi(\mu^*)\left[v^i\right] = v^i/\sqrt{\mu}$, the divide is done component-wise. We define the sample covariance matrix $C_{\bar{V}}$ as

$$C_{\bar{V}} = \frac{1}{L-1}\bar{V}\bar{V}^T. \tag{V.92}$$

We remind that the main goal of TPCA is to find an orthonormal basis $\{e_1, ..., e_n\}$

in $T_{\mu^*}\mathcal{P}_+(I)$ that solves the optimization problems:

$$\begin{cases} e_1 = \text{argmax}_{\|e\|_{\mu^*}=1} \frac{1}{L-1} \sum_{i=1}^{L} \mathfrak{g}_{\mu^*}(e, v^i)^2, \\ e_k = \text{argmax}_{\|e\|_{\mu^*}=1} \frac{1}{L-1} \sum_{i=1}^{L} \mathfrak{g}_{\mu^*}(e, v^{ik})^2, \quad \text{where } v^{ik} = v^i - \sum_{j=1}^{k-1} \mathfrak{g}_{\mu^*}(e_j, v^i) e_j. \end{cases} \tag{V.93}$$

We summarize the solution in the following proposition.

Proposition V.4

Let $\bar{V} = \left[\bar{v}^1, ..., \bar{v}^L\right] \in \mathbb{R}^{(n+1)\times L}$ *with* $\bar{v}^i = d\Phi(\mu^*)\left[v^i\right]$. *Let* $C_{\bar{V}}$ *be the sample covariance matrix. The solution of (V.93) is given by* $\{d\Phi(\mu^*)^{-1} f_1, \ldots, d\Phi(\mu^*)^{-1} f_n\}$, *where* $\{f_1, ..., f_n\}$ *is the first* n *eigenvectors of* $C_{\bar{V}}$.

Proof Using the isometry map Φ, we rewrite (V.93) on $\mathbb{R}^{n+1}$ as

$$\begin{cases} f_1 = \text{argmax}_{\|f\|=1} \frac{1}{L-1} \sum_{i=1}^{L} \left\langle f, \bar{v}^i \right\rangle^2, \\ f_k = \text{argmax}_{\|f\|=1} \frac{1}{L-1} \sum_{i=1}^{L} \left\langle f, \bar{v}^{ik} \right\rangle^2, \quad \text{where } \bar{v}^{ik} = \bar{v}^i - \sum_{j=1}^{k-1} \left\langle f_j, \bar{v}^i \right\rangle f_j, \end{cases} \tag{V.94}$$

with $\bar{v}^i = d\Phi(\mu^*)\left[v^i\right]$. Then, the solution is given by the ordered eigenvectors $\{f_k\}_{k=1}^{n+1}$ of $C_{\bar{V}}$. To show that $\{d\Phi(\mu^*)^{-1} f_1, ..., d\Phi(\mu^*)^{-1} f_n\}$ is an orthonormal basis of $T_{\mu^*}\mathcal{P}_+(I)$ we consider the Hyperplane H perpendicular to $\Phi(\mu^*)$. We remind that $\langle \Phi(\mu^*), \bar{v}^i \rangle = 0$ and we easily check that $\{f_k\}_{k=1}^{n}$ spans H and f_{n+1} is it's normal vector. ■

Definition V.6

Let $(\lambda_i, f_i)_{i=1}^{n+1}$ *be the eigenpairs of the covariance matrix* $C_{\bar{V}}$ *as in the Proposition V.4. We define the corresponding covariance matrix on the tangent space as*

$$C_V = E S^2 E^T, \tag{V.95}$$

where $E = [e_1, ..., e_n], e_i = d\Phi(\mu^*)^{-1}(f_i)$ *and* S *is the diagonal matrix of* $(\sqrt{\lambda_i})_{i=1}^{n}$.

We summarize the main steps in Algorithm 3.

Algorithm 3: Covariance matrix/TPCA

Input : Data Populations: $P_L = \{\mu^i\}_{i=1}^L$.

Output: Covariance matrix: $C_V, C_{\bar{V}}$, eigenvectors F, E, eigenvalues $\lambda_1, ..., \lambda_n$.

Compute the Karcher mean μ^* of P_L using (V.89).

Project data points P_L onto tangent space $T_{\mu^*}\mathcal{P}_+(I)$:

$$v^i = \log_{\mu^*}(\mu^i), i = 1, ..., L,$$

$$V = [v^1, ..., v^L].$$

Push forward the data:

$$\bar{v}^i = d\Phi(\mu^*)[v^i], i = 1, ..., L,$$

$$\bar{V} = [\bar{v}^1, ..., \bar{v}^L].$$

Compute the covariance matrix

$$C_{\bar{V}} = \frac{1}{L-1}\bar{V}\bar{V}^T.$$

Find the eigenpairs $(\lambda_i, f_i)_{i=1}^{n+1}$ of $C_{\bar{V}}$.

Pullback the eigenvector $e_i = d\Phi(\mu^*)^{-1}[f_i]$, for $i = 1, ..., n$, and let

$E = [e_1, ..., e_n]$.

The covariance matrix on tangent space is given by

$$C_V = E\text{diag}(\lambda_1, ..., \lambda_n)E^T.$$

V.4.2 Covariance and PCA transport

Given two populations $P_L = \{\mu^i\}_{i=1}^L$ and $P_S = \{\nu^i\}_{i=1}^S$ in $\mathcal{P}_+(I)$, the corresponding covariance matrices are then given by

$$C_{\bar{V}} = \frac{1}{L-1}\bar{V}\bar{V}^T,$$

and

$$C_{\bar{W}} = \frac{1}{S-1}\bar{W}\bar{W}^T,$$

where $\bar{W} = \left[\bar{w}^1, ..., \bar{w}^S\right] \in \mathbb{R}^{(n+1)\times S}$ with the i-th column $\bar{w}^i = d\Phi(\nu^*)\left[w^i\right]$. The corresponding sample covariance matrices on tangent space C_V and C_W are defined as

in Definition V.6. Since P_S is of small size, $C_{\bar{W}}$ may be a poor estimate of the true covariance matrix of P_S. Hence, we would hope to enhance the covariance estimation $C_{\bar{W}}$ by exploiting $C_{\bar{V}}$. To illustrate this idea, we transfer the sample covariance matrix C_V, which is defined on the tangent space $T_{\mu^*}\mathcal{P}_+(I)$ to the tangent space $T_{\nu^*}\mathcal{P}_+(I)$ using the metric parallel transport $\Gamma_{\mu^*\rightarrowtail\nu^*}$ defined on $\mathcal{P}_+(I)$. To accomplish this, our strategy consists in transporting orthonormal basis $\{e_1, ..., e_n\}$ in $T_{\mu^*}\mathcal{P}_+(I)$ to $T_{\nu^*}\mathcal{P}_+(I)$, which in turn has produced the covariance matrix of the transported data $\{v^i\}_{i=1}^L$ in $T_{\nu^*}\mathcal{P}_+(I)$. Thanks to the metric parallel transport, the orthogonality and distance between vectors are preserved.

Improving the covariance matrix $C_{\bar{W}}$ provides an avenue for the application of common dimensionality reduction techniques. Here, we show that TPCA model of $\{\tilde{v}^i\}_{i=1}^L = \Gamma_{\mu^*\rightarrowtail\nu^*}(\{v^i\}_{i=1}^L)$ in $T_{\nu^*}\mathcal{P}_+(I)$ coincides exactly with the TPCA model defined by the transported eigenvectors of $C_{\bar{V}}$. We summarize these results in Proposition V.5.

Proposition V.5. Covariance/PCA transport [46]

Let two populations $P_L = \{\mu^i\}_{i=1}^L$ and $P_S = \{\nu^i\}_{i=1}^S$ in $\mathcal{P}_+(I)$ with the corresponding Karcher means μ^*, ν^*. Let $V = \left[v^1, ..., v^L\right]$ be the matrix where the i-th column v^i is equal to $\log_{\mu^*}(\mu^i)$. Suppose we have an orthonormal basis $\{e_1, ..., e_n\}$ in $T_{\mu^*}\mathcal{P}_+(I)$ and corresponding eigenvalues $\{\lambda_1, ..., \lambda_n\}$, that solves the TPCA problems (V.93). Then

a) the transported vector $\{\tilde{e}_1, ..., \tilde{e}_n\}$, where $\tilde{e}_i = \Gamma_{\mu^*\rightarrowtail\nu^*}(e_i)$, solves the TPCA problems in $T_{\nu^*}\mathcal{P}_+(I)$ of $\tilde{V} = \left[\tilde{v}^1, ..., \tilde{v}^L\right]$, where $\tilde{v}^i = \Gamma_{\mu^*\rightarrowtail\nu^*}(v^i)$.

b) the transported covariance matrix is

$$\tilde{C}_V = \tilde{E}S^2\tilde{E}^T, \tag{V.96}$$

where $\tilde{E} = [\tilde{e}_1, ..., \tilde{e}_n]$ and $S^2 = diag(\lambda_1, ..., \lambda_n)$.

Proof a) Since $\Gamma_{\mu^*\rightarrowtail\nu^*}$ is metric parallel transport, $\{\tilde{e}_1, ..., \tilde{e}_n\}$ is an orthonormal basis in $T_{\nu^*}\mathcal{P}_+(I)$. The TPCA problems for the transported data are given by

$$\begin{cases} u_1 = \text{argmax}_{\|u\|_{\nu^*}=1} \frac{1}{L-1}\sum_{i=1}^L \mathfrak{g}_{\nu^*}(u, \tilde{v}^i)^2, \\ u_k = \text{argmax}_{\|u\|_{\nu^*}=1} \frac{1}{L-1}\sum_{i=1}^L \mathfrak{g}_{\nu^*}(u, \tilde{v}^{ik})^2, \quad \text{where } \tilde{v}^{ik} = \tilde{v}^i - \sum_{j=1}^{k-1}\mathfrak{g}_{\nu^*}(u_j, \tilde{v}^i)u_j. \end{cases} \tag{V.97}$$

We see that the solution of the above problems are $\{\tilde{e}_1, ..., \tilde{e}_n\}$ with the same eigenvectors $\{\lambda_1, ..., \lambda_n\}$.

b) This follows directly from the part a) since $\{\tilde{e}_1, ..., \tilde{e}_n\}$ solves the TPCA problems with corresponding eigenvalues $\{\lambda_1, ..., \lambda_n\}$. ■

By the above Proposition, to transport the TPCA model on μ^* to ν^*, we do not need to transport all the data to solve for the TPCA problems but we only need to transport the orthonormal basis.

Now, the two covariance matrices C_W and $\tilde{C}_V$ lie in the same tangent space $T_{\nu^*}\mathcal{P}_+(I)$, we can use shrinkage estimation method [102] to combine the two covariance matrices as follow,

$$C_\rho = \rho\tilde{C}_L + (1-\rho)C_S, \quad 0 \leq \rho \leq 1. \tag{V.98}$$

It is easily seen that ρ weighs the contribution of $\tilde{C}_L$, and we need to choose the value of ρ depending on each problem. We summarize different steps of transfer learning covariance matrix and PCA model in Algorithm 4.

V.4.3 Linear regression transport

Let $\{v^i\}_{i=1}^L \subset T_{\mu^*}\mathcal{P}_+(I)$. We present similar result as Proposition V.5 for transfer learning of linear regression model. Indeed, let $\{y_i\}_{i=1}^L \in \mathbb{R}$ denote the label of $\{v^i\}_{i=1}^L$. A linear regression model $y : T_{\mu^*}\mathcal{P}_+(I) \to \mathbb{R}, v \to y(v)$ has the following form

$$y(v) = v^T r + r_0 = \left\langle v, G_\mu^{-1} r \right\rangle_\mu + r_0. \tag{V.99}$$

where G_μ denote the Fisher-Rao metric on $\mathcal{P}_+(I)$, $r_0 \in \mathbb{R}$ and r is a tangent vector on $T_{\mu^*}\mathcal{P}_+(I)$. Let us consider $l_i(y)$ the loss function associated with $y_i = y(a_i)$, for example we can take the loss function as a squared error $l_i(y) = (y - y_i)^2$.

Proposition V.6. Linear regression transport [46]

Let

$$(\beta, \beta_0) = \underset{d \in T_{\mu^*}\mathcal{P}_+(I), d_0 \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{L} l_i\left(\left(v^i\right)^T d + d_0\right). \tag{V.100}$$

be a solution of the linear regression model (V.99) on $T_{\mu^*}\mathcal{P}_+(I)$. Then the tangent vector $\tilde{\beta} = G_{\nu^*}\Gamma_{\mu^* \rightarrowtail \nu^*}(G_{\mu^*}^{-1}\beta)$ is a solution of the linear regression model

Algorithm 4: Transfer Learning of Covariance matrix and PCA model

Input : Data Populations: $P_L = \{\mu^i\}_{i=1}^L$, and $P_S = \{\nu^i\}_{i=1}^S$.
Output: Covariance matrix: C_ρ.

Compute Karcher means μ^* and ν^* of P_L and P_S using (V.89).
Project data points P_L and P_S respectively onto tangent spaces $T_{\mu^*}\mathcal{P}_+(I)$ and $T_{\nu^*}\mathcal{P}_+(I)$:

$$v^i = \log_{\mu^*}(\mu^i), i = 1, ..., L,$$
$$w^i = \log_{\nu^*}(\nu^i), i = 1, ..., S.$$

Push forward the data into the ambient space:

$$\bar{v}^i = d\Phi(\mu^*)[v^i], i = 1, ..., L,$$
$$\bar{w}^i = d\Phi(\nu^*)[w^i], i = 1, ..., S.$$

Apply the Algorithm 3 to find the covariance matrix and TPCA of P_L and P_S: $C_V, C_{\bar{V}}, F_V, E_V, \lambda_i^V$ and $C_W, C_{\bar{W}}, F_W, E_{\bar{W}}, \lambda_i^W$.
Transform the orthonormal basis E_V to $\tilde{E}_V = [\tilde{e}_1^V, ..., \tilde{e}_n^V]$, where $\tilde{e}_i^V = \Gamma_{\mu^* \rightarrowtail \nu^*}(e_i^V)$, and transform F_V to $\tilde{F}_V = [\tilde{f}_1^V, ..., \tilde{f}_n^V]$, where $\tilde{f}_i^V = d\Phi_{\nu^*}(\tilde{e}_i^V)$.
The transformed covariance matrices are

$$\tilde{C}_V = \tilde{E}_V \text{diag}(\lambda_1^V, ..., \lambda_n^V)\tilde{E}_V,$$
$$\tilde{C}_{\bar{V}} = \tilde{F}_V \text{diag}(\lambda_1^V, ..., \lambda_n^V)\tilde{F}_V.$$

Let $C_\rho = \rho\tilde{C}_{\bar{V}} + (1-\rho)C_{\bar{W}}, \quad 0 \leq \rho \leq 1$ be the shrinkage covariance matrix on $\mathbb{R}^{n+1}$.
Compute the eigenvalue decomposition of C_ρ: $C_\rho = V_\rho \text{diag}(\lambda_1^\rho, ..., \lambda_{n+1}^\rho)V_\rho^T$ and the corresponding orthonormal basis E_ρ of $T_{\nu^*}\mathcal{P}_+(I)$.
Fix $k \in \mathbb{N}, k < n$, the k-dimensional TPCA model given by the first k vectors of E_ρ and k eigen-value $(\lambda_1^\rho, ..., \lambda_k^\rho)$.

on $T_{\nu^*}\mathcal{P}_+(I)$,

$$\tilde{\beta} = \textit{argmin}_{c \in T_\nu \mathcal{P}_+(I)} \sum_{i=1}^{L} l_i \left(\left(\tilde{v}^i\right)^T c + \beta_0 \right). \tag{V.101}$$

According to Proposition V.6, there is no need to transport the entire dataset; rather, we only need to transport the tangent vector of the linear model. This significantly reduces computational costs. We summarize the different steps of linear regression

transport in Algorithm 5.

Algorithm 5: Transfer learning of Linear regression model

Input : Data Populations: $P_L = \{\mu^i\}_{i=1}^L$, $P_S = \{\nu^i\}_{i=1}^S$ in $\mathcal{P}_+(I)$ and its labels $\{y_{\mu^i}\}_{i=1}^L$, $\{y_{\nu^i}\}_{i=1}^S$.

Output: γ_ρ: the coefficients of the Linear model.

Compute Karcher means μ^* and ν^* of P_L and P_S using (V.89).

Project data points P_L and P_S respectively onto tangent space $T_{\mu^*}\mathcal{P}_+(I)$ and $T_{\nu^*}\mathcal{P}_+(I)$:

$$v^i = \log_{\mu^*}(\mu^i), i = 1, ..., L,$$

$$w^i = \log_{\nu^*}(\nu^i), i = 1, ..., S.$$

Find the solution (β, β_0) of the Linear regression model model on $T_{\mu^*}\mathcal{P}_+(I)$

$$(\beta, \beta_0) = \text{argmin}_{r \in T_{\mu^*}\mathcal{P}_+(I), r_0 \in \mathbb{R}} \sum_{i=1}^{S} l_i \left(\left(v^i\right)^T r + r_0 \right). \tag{V.102}$$

Split the data $\{\{w^i\}_{i=1}^S, \{y_{\nu^i}\}_{i=1}^S\}$ into the training set B_{train} and test set B_{test}.

Find the solution (θ, θ_0) of the Linear regression model model on $T_{\nu^*}\mathcal{P}_+(I)$ on training set

$$(\theta, \theta_0) = \text{argmin}_{u \in T_{\nu^*}\mathcal{P}_+(I), u_0 \in \mathbb{R}} \sum_{w^i \in B_{train}} l_i \left(\left(w^i\right)^T u + u_0 \right). \tag{V.103}$$

Apply the parallel transport $\Gamma_{\mu^* \to \nu^*}$ defined on $\mathcal{P}_+(I)$ to transport the tangent vector $G_{\mu^*}^{-1}\beta$ to the tangent space $T_{\nu^*}\mathcal{P}_+(I)$.

Set $\tilde{\beta} = G_{\nu^*}\Gamma_{\mu \to \nu}(G_{\mu^*}^{-1}\beta)$, then $(\tilde{\beta}, \beta_0)$ is the solution of the transported linear regression model (V.99) on $T_{\nu^*}\mathcal{P}_+(I)$.

$$\tilde{\beta} = \text{argmin}_{c \in T_{\nu^*}\mathcal{P}_+(I)} \sum_{i=1}^{L} l_i(\Gamma_{\mu^* \to \nu^*}(v^i)^T c + \beta_0). \tag{V.104}$$

Compute the combined solution of the regression model on $T_{\nu^*}\mathcal{P}_+(I)$:

$$\gamma_\rho = \rho(\tilde{\beta}, \beta_0) + (1 - \rho)(\theta, \theta_0). \tag{V.105}$$

Return γ_ρ.

V.5 Experiments

V.5.1 TPCA and TPCA transport

We will apply the TPCA and TPCA transport for two different datasets of histograms.

V.5.1.1 On first datasets

We apply Algorithm 3 to the real dataset Animal-Fclass1 (P_L), which contains 998 elements in $\mathcal{P}_+(I)$, where $|I| = 200$. This dataset represents the histogram of density estimation from the body temperature of animals without disease. The reconstruction is computed by applying the exponential map to the projection of the log of data on the space spanned by principal components (the projection is in the tangent space). Then, the error between the true and the reconstructed data is measured using the geodesic distance. In this experiment, we employ 3 principal components for reconstruction, with an explained variance ratio of 0.9372. The errors are displayed in Figure V.1. We observe that the median of reconstruction errors is approximately 0.1.

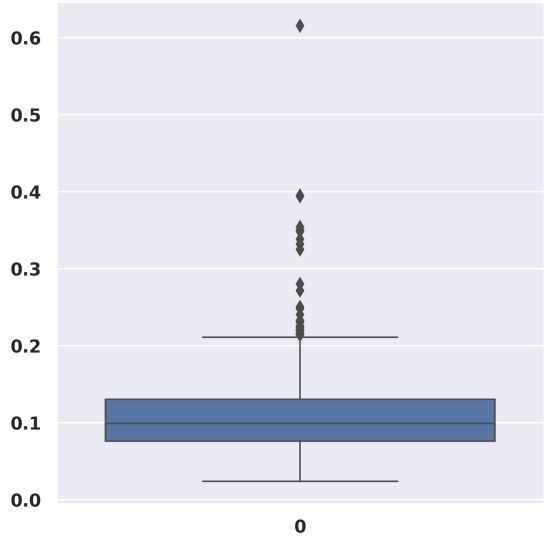


Figure V.1: The boxplot of geodesic distance error of the reconstruction and the true of P_L.

To apply the TPCA transport model developed in this chapter, we consider the second data set, Animal-Fclass2 (P_S), which consists of 350 data points. This dataset represents the histogram of density estimation from the temperature of animals with diseases. Both datasets and their Karcher means are illustrated in Figure V.2.

We apply Algorithm 4 to the model transport with $\rho \in \{0, 0.1, 0.2, ..., 1\}$ (V.98). We also use 3 principal components for reconstruction, and the error is measured by

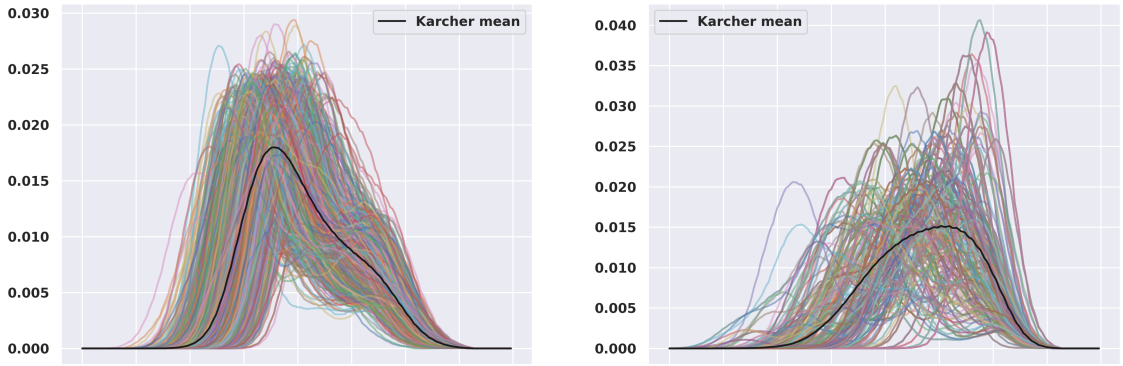


Figure V.2: The data and the Karcher mean of (Left): P_L, (Right): P_S.

geodesic distance. The results are presented in Figure V.3. Notably, we observe that selecting specific ρ values, such as 0.1, 0.2, and 0.3, results in a slight reduction in the reconstruction error.

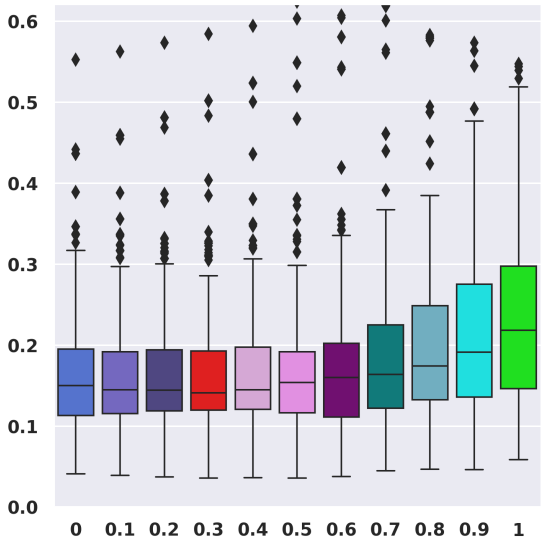


Figure V.3: The boxplots of the reconstruction error for $\rho \in \{0, 0.1, 0.2, ..., 1\}$.

V.5.1.2 On second datasets

Now we apply the same steps as before but to different datasets. The large dataset P_L contains 4000 histograms of images of cat from the training set[1]. The small dataset consists of 300 histogram of cat images from the test set. Figure V.4 shows some of these cat images.

[1] https://www.kaggle.com/datasets/chetankv/dogs-cats-images?select=dataset

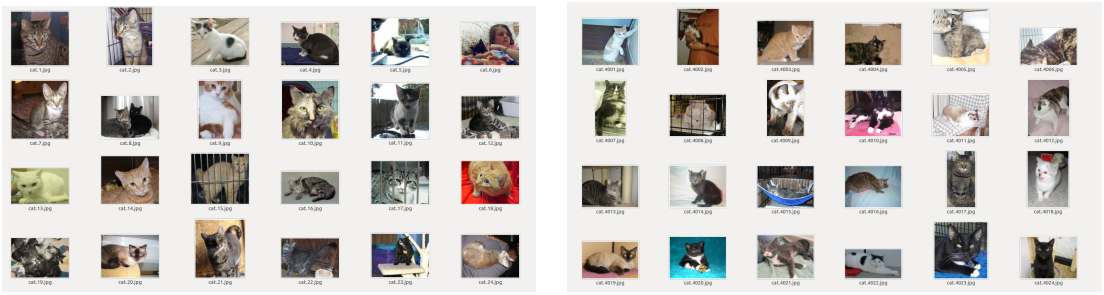


Figure V.4: The images of cats in the large dataset (on the right) and the small dataset (on the left).

The normalized histograms of gray scale images belong to $P_+(256)$ (after adding some small number). Figure V.5 displays Karcher means and some histograms.

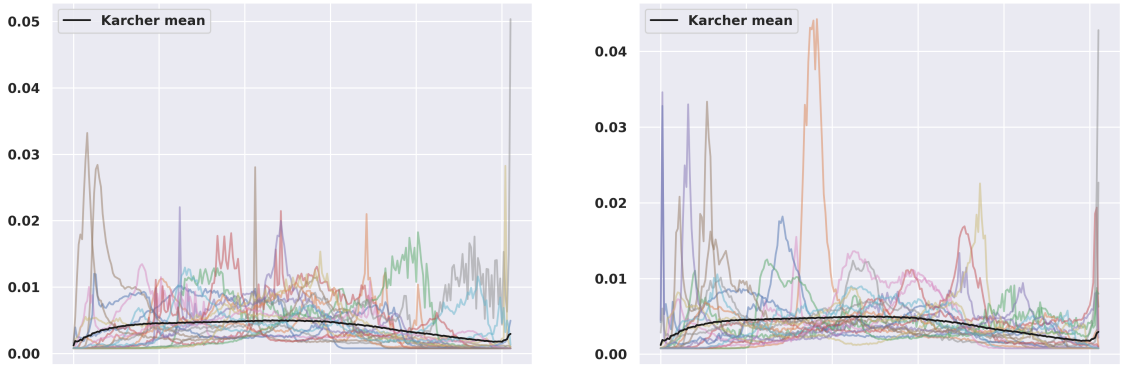


Figure V.5: The images of Karcher mean and some histograms of the large dataset (right) and small dataset (left).

For TPCA of the large dataset P_L, we use 10 principal components for reconstruction, with an explained variance ratio of 0.8516. The errors are displayed in Figure V.6.

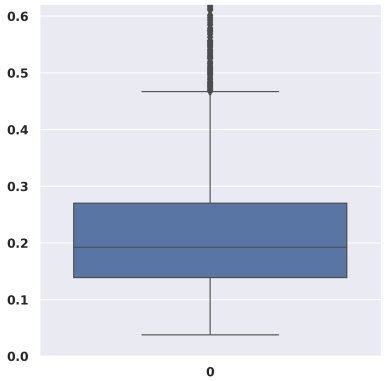


Figure V.6: The boxplot displays the geodesic distance errors on the large dataset of cat image histograms.

For TPCA transport, we apply Algorithm 4 using 10 principal components. The results are presented in Figure V.7.

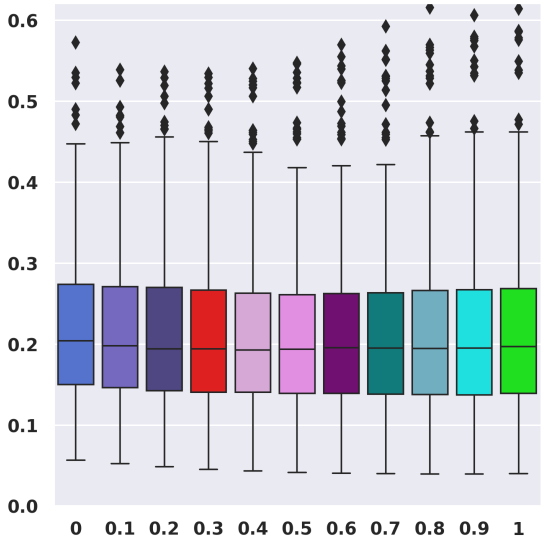


Figure V.7: The boxplots show the reconstruction error for $\rho \in \{0, 0.1, 0.2, ..., 1\}$ for small dataset of cat image histograms.

V.5.2 Linear Regression

As before, we also apply transferring linear regression models to two different datasets.

V.5.2.1 On first datasets

We apply Algorithm 5 to transport a linear regression model. Suppose we have two populations, P_L and P_S, each containing data labeled as 0 or 1. Initially, we train the linear model LM_1 on P_L and then transport its coefficients. We refer to the combined model of the transported one and the model learned from P_S as LM_{com}. Subsequently, we test LM_{com} on the test subset of P_S.

We demonstrate an application on two populations: Human (Male and Female) and Animal (class 1 and class 2). The Human dataset contains the histograms estimated from the heights of males and females, while the Animal dataset comprises histograms estimated from temperatures of animals. Each class in Human and Animal consists of 100 elements. In our application, we randomly split the Animal dataset into a training set and a test set, with the test size set to 0.33.

The first model is leaned on the Human dataset and then transported to the Animal dataset. Then, the second model is trained on the training set of the Animal dataset. Finally, the transported model and the second model are combined and tested on the test set of Animal dataset.

The same method of transported model is also applied for the Heart beats dataset, which contains 500 normal and 1000 abnormal elements, in place of the Human dataset. We conducted the test 50 times with random splits in the Animal dataset to remove the bias. The accuracies of the combined model are depicted in Figure V.8.

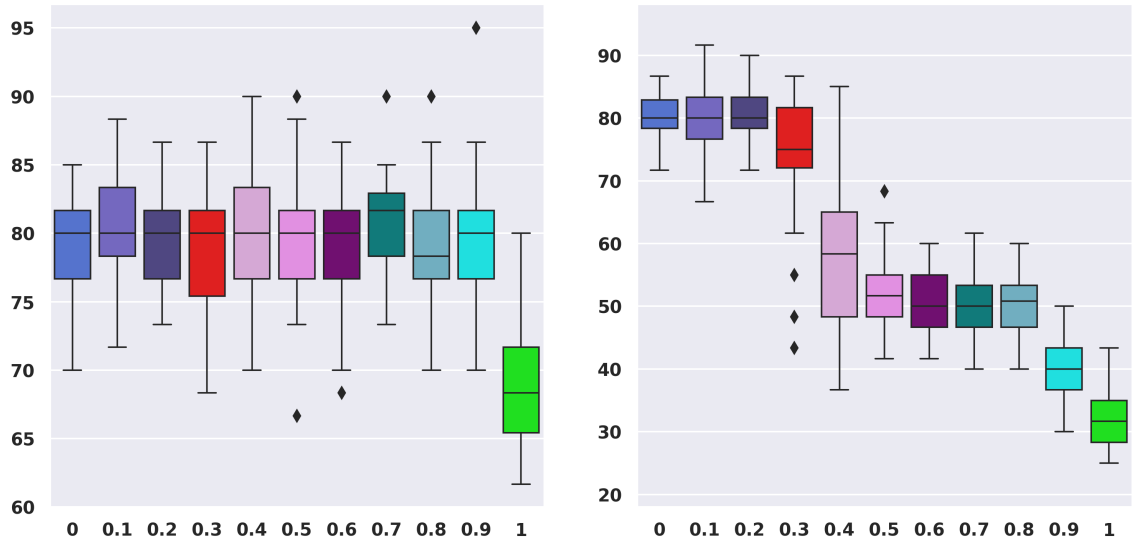


Figure V.8: The Box-plot of the accuracy of the combined models (Left column): the large dataset is Human, (Right column): the large dataset is Heart beats.

In Figure V.8, we observe that the accuracy in the left figure remains relatively stable overall. However, with an appropriate value of ρ, such as $\rho = 0.7$, the results can be improved. In contrast, in the right figure, the accuracies do not exhibit improvement. This suggests that there is little to no relationship between the two datasets, which can be considered an example of negative transfer.

V.5.2.2 On second datasets

In this experiment, we apply the same strategy to transfer a linear regression model to different datasets. The large dataset P_L comprises two classes of histograms from images

of cats and dogs in the training set[2]. Each class consists of 4000 elements. Similar, the small dataset P_S contains two classes of histograms from images of cats and dogs in the test set, with each class contains 300 elements. The results are depicted in Figure V.9.

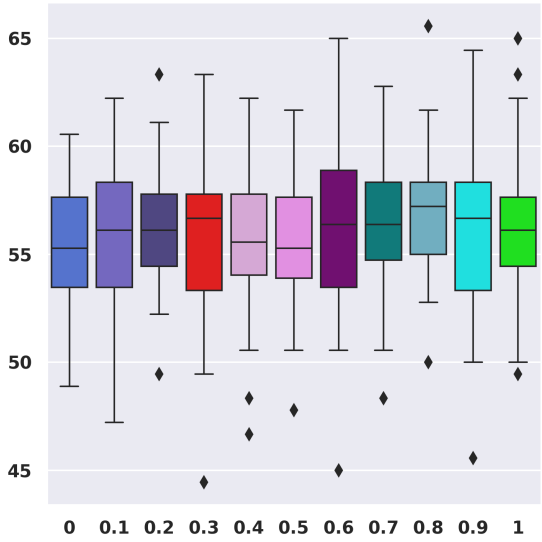


Figure V.9: The Box-plot of accuracy of the combined model for datasets of histograms from images of cats and dogs.

In Figure V.9, we observe that several values of ρ lead to improved results, with the optimal value being $\rho = 0.8$.

V.6 Conclusion

In this chapter, we have proposed an efficient and accurate transfer learning algorithm of statistical models on the space of probability measures $\mathcal{P}_+(I)$. To achieve this goal, we consider a metric parallel transport: Levi-Civita parralel transport. In particular, we have developed newly geodesic operations associated to the metric parallel transport. We have proven that implementing techniques that exploit the underlying geometry of the manifold yield good achievement in transfer learning tasks. Finally, we have applied and discussed the good accuracy of the method and the high efficiency of the proposed algorithm with various and multiple experimental results.

[2] https://www.kaggle.com/datasets/chetankv/dogs-cats-images?select=dataset

Chapter VI: Conclusion and prospects

In this chapter, we conclude the thesis by summarizing our main contributions and results. Additionally, we will list some ongoing works that we aim to finish in the future.

VI.1 Summary of the contributions

Throughout this thesis, we have made contributions by introducing new Gaussian process models, constrained Gaussian processes, and transfer learning on manifolds. We will provide a more detailed summary in the following items:

- We constructed new Gaussian processes based on classical polynomials such as Legendre, Laguerre, Hermite, and Chebyshev. Although these Gaussian processes do not have explicit covariance functions, they do have explicit K-L expansions. By approximating these Gaussian processes through a finite truncation of the K-L expansion, we reduce the computational cost in Gaussian process regression, enabling us to work with large datasets.
- We have incorporated a new type of constraint into Gaussian process models, ensuring that the output function represents a probability distribution. Constrained Gaussian process models pose a challenge because the posterior distribution is typically analytically intractable. As a result, numerical methods are required to approximate its mean and covariance. In this thesis, we introduced the embedded Hamiltonian Monte Carlo (HMC) method on the sphere to simulate and approximate the posterior distribution. Our experiments demonstrate that this proposed framework performs well and, in specific cases, outperforms other methods such as Artificial Neural Networks.
- Our last contribution has two facets: first, a detailed examination of the geometry of finite distribution measures, and second, the introduction of transfer learning on this manifold. In this thesis, we equip the space of finite measures with the Fisher-Rao metric. We provide detailed formulas for the geodesic, parallel transport, exponential map, and log map. These are all the tools necessary for transfer

learning. We also present algorithms for transferring PCA and linear regression. Additionally, we conducted experiments to apply these frameworks.

VI.2 Future work and prospects

In this thesis, we have demonstrated that it is highly beneficial when the K-L expansion of a Gaussian process is known. Unfortunately, discovering the K-L expansion can be challenging. Our future research endeavors will focus on exploring this field further to uncover additional representations of Gaussian processes through series. Indeed, by considering the eigenvalue equations of certain differential and integral operators, we hope to identify more K-L expansions for Gaussian processes.

We also aim to enhance the flexibility of Gaussian process regression in applications by incorporating various types of constraints into the model. Simultaneously, we are exploring the application of alternative simulation methods to improve both cost-effectiveness and accuracy. In our recent work, we have experimented with applying a Neural Network model with constraints to approximate probability density functions. We hope to obtain further results through this approach as well.

Finally, we will delve deeper into Information Geometry to better apply it in Machine Learning. In fact, we will endeavor to define a distance and a covariance function in a more abstract space. From there, we can introduce linear regression models and classification models.

Appendices

Chapter A: Low rank Gaussian processes

A.1 Simulation

In this simulation study, our goal is to predict the true function, denoted as g, which has the form $g(t) = \sum_{j=1}^{P} c_j \varphi_j(t)$. Where the coefficients c_j are generated independently from a normal distribution $\mathcal{N}(0, \lambda_j)$, and (λ_j, φ_j) represent eigenpairs. It's worth noting that the function g exhibits different properties and shapes for different covariance settings. In this section, we keep P fixed at 15.

Figure A.1 illustrates the graphs of g for various covariance settings. To predict the values of g, we utilize 50 observation points that are observed uniformly along the interval I. At each observation point, we introduce a small amount of noise drawn independently from $\mathcal{N}(0, \sigma_n^2 = 10^{-3})$. We employ GP models with $f_M(t) = \sum_{j=1}^{M} a_j \varphi_j(t)$, corresponding to each covariance, with M set to 25. These models are then evaluated at 200 equally spaced points within the interval I. We also employ a standard GP with a Matérn covariance function defined on $\mathbb{R}$, and Sparse GP for the purpose of comparison. The standard GP model is implemented in the scikit-learn library[1], and Sparse GP implemented in GPy[2]. We calculate the errors and present the results in the box-plots in Figure A.2.

[1] https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.GaussianProcessRegressor.html

[2] https://nbviewer.org/github/SheffieldML/notebook/blob/master/GPy/sparse_gp_regression.ipynb

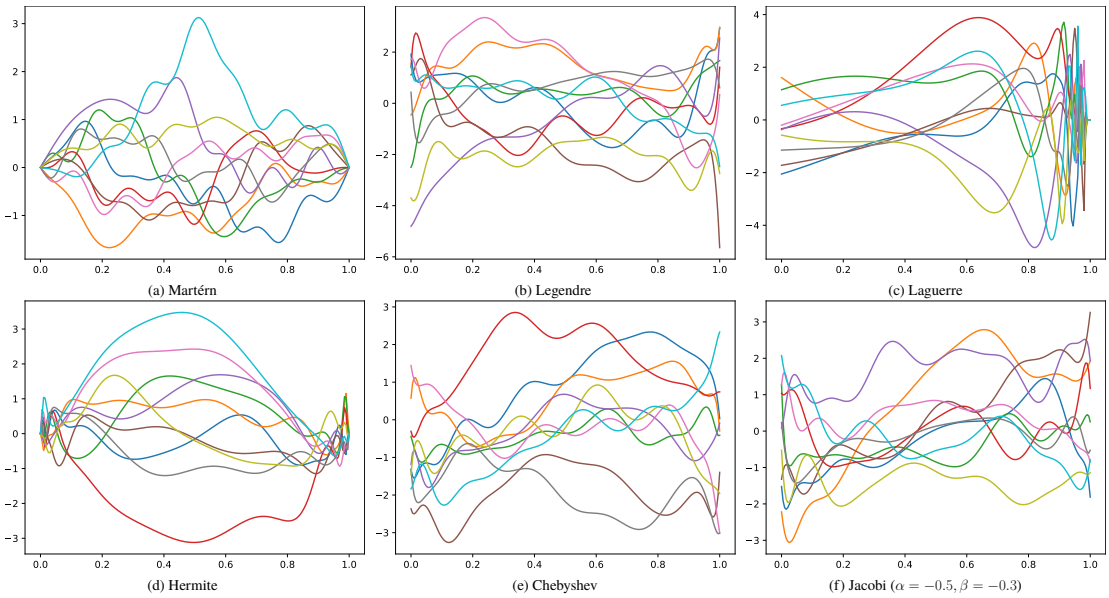


Figure A.1: Realizations of different models.

In Figure A.2, we observe that, in general, all models can predict the true function very effectively with small errors. The model employing the Legendre, Chebyshev and Jacobi covariance functions outperforms the others, as it exhibits the smallest Integrated Squared Error (ISE) and negative log-likelihood. These models slightly better than standard GP. However, the Matérn Hermite and sparse GP models do not perform as well as the others.

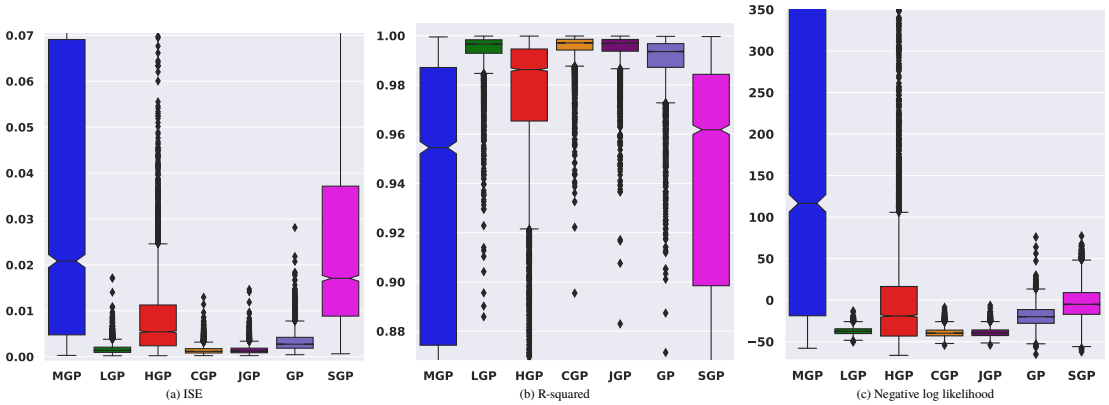


Figure A.2: Box-plots of the results of different Gaussian process models in predicting the true function $g(t) = \sum_{j=1}^{P} c_j \varphi_j(t)$, where c_j generated independently from $\mathcal{N}(0, \lambda_j)$. Where MGP: Matérn, LGP: Legendre, HGP: Hermite, CGP: Chebyshev, GP: standard GP, SGP: Sparse GP.

A.2 Real data

In this part, we give details results for the Experiment part with real data. Table A.1, Table A.2 and Table A.3 display the prediction results using real data, in which Legendre

performs sightly better than the others methods. These results suggest that Legendre is more adaptable in capturing a wide range of patterns and structures within real data. The time required for the Chebyshev model is the smallest, even though its complexity is the same as that of other proposed models. This is due to the faster computation time for the corresponding eigenfunctions. The standard GP (std GP) has the highest complexity, while the sparse GP has a smaller complexity compared to the proposed methods. However, both the standard GP and sparse GP require more computation for solving the best parameters of the covariance functions and inducing points. The polynomial regression has the smallest complexity but we need to spend time preprocessing the data. In all the experiments, we take the number of truncation $M = 25$ in the proposed models, number of inducing points $Z = 10$ for sparse GP and the degree of polynomial regression is two, $R = 3$. The programs were run on the computer Dell Precision with 125 GiB memory and CPU Xeon(R) W 2275 @ 3.30GHz

Table A.1: Results of different methods on CalCOFI data ($N = 500$).

Models	MSE	R2	NLML	Time(s)	$O(.)$
Matérn	1.5186	0.8588	4635.63	0.0048	NM^2
Legendre	1.4988	0.8607	3755.24	0.0033	NM^2
Hermite	1.5095	0.8597	4430.93	0.0059	NM^2
Chebyshev	1.4991	0.8606	3744.62	0.0022	NM^2
Jacobi	1.4992	0.8606	3758.00	0.0171	NM^2
std GP	1.4997	0.8606	822.77	0.2259	N^3
Sparse GP	1.5635	0.8547	830.7	0.1489	NZ^2
Polynomial regression	1.8685	0.8263	N/A	0.0034	NR^2

Table A.2: Results of different methods on MCP non-smoker data ($N = 532$).

Models	MSE	R2	NLML	Time(s)	$O(.)$
Matérn	0.2556	-0.0029	504.54	0.0042	NM^2
Legendre	0.2568	-0.0076	510.09	0.0046	NM^2
Hermite	0.2578	-0.0117	500.47	0.0045	NM^2
Chebyshev	0.2572	-0.0091	508.23	0.0023	NM^2
Jacobi	0.2570	-0.0086	508.91	0.0175	NM^2
std GP	0.2570	-0.0086	368.45	0.5880	N^3
Sparse GP	0.2569	-0.0083	368.0	0.1803	NZ^2
Polynomial regression	0.2575	-0.0103	N/A	0.0033	NR^2

Table A.3: Results of different methods on MCP smoker data ($N = 137$).

Models	MSE	R2	NLML	Time(s)	Complexity (O)
Matérn	0.2152	0.7611	205.38	0.0018	NM^2
Legendre	0.1809	0.7992	149.06	0.0018	NM^2
Hermite	0.2518	0.7205	178.91	0.0020	NM^2
Chebyshev	0.1822	0.7977	147.44	0.0011	NM^2
Jacobi	0.1836	0.7962	150.45	0.0139	NM^2
std GP	0.2018	0.7760	111.99	0.0288	N^3
Sparse GP	0.2227	0.7528	116.27	0.1696	NZ^2
Polynomial regression	0.3058	0.6606	N/A	0.0023	NR^2

Chapter B: Transfer learning

Proof of Claim

Claim: For all $\mu \in \mathcal{P}_+(I)$ and $0 < l < \pi$. Let $\tau, \tilde{\tau} \in T_\mu \mathcal{P}_+$ such that

$$\cos\frac{l}{2} + \frac{d\tau}{d\mu}(i)\sin\frac{l}{2} = \pm\left(\cos\frac{l}{2} + \frac{d\tilde{\tau}}{d\mu}(i)\sin\frac{l}{2}\right), \forall i \in I.$$

Let

$$I_+ = \{i \in I | \tau_i = \tilde{\tau}_i\}, \tag{B.1}$$

$$I_- = \left\{i \in I | \tau_i + \tilde{\tau}_i = -2\mu_i \cot\frac{l}{2}\right\}, \tag{B.2}$$

then $I_- = \emptyset$.

Proof We proof the Claim by induction on the degree of I. If $|I|$ is one or two the Claim is true since I_+ is not empty. Suppose the Claim is trus for $|I| = n$. We go to prove the Claim for $|I| = n + 1$. Let μ ,τ, $\tilde{\tau}$ and l like in the Claim. Suppose $I_- \neq \emptyset$ then $|I_-| \geq 2$. Let g, h be two distinct index in I_-, this means $\tau_g + \tilde{\tau}_g = -2\mu_g \cot\frac{l}{2}$ and $\tau_h + \tilde{\tau}_h = -2\mu_h \cot\frac{l}{2}$. Now let $k \in I_+$ and define three measures $\tau', \tilde{\tau}', \mu'$ on $I \setminus \{k\}$ as follow

$$\tau' = \sum_{i\in I, i\neq k,h,g} \tau_i\delta^i + \tau_g\delta^g + (\tau_h + \tau_k)\delta^h, \tag{B.3}$$

$$\tilde{\tau}' = \sum_{i\in I, i\neq k,h,g} \tilde{\tau}_i\delta^i + (\tilde{\tau}_g + 2\tilde{\tau}_k)\delta^g + (\tilde{\tau}_h - \tilde{\tau}_k)\delta^h, \tag{B.4}$$

$$\mu' = \sum_{i\in I, i\neq k,h,g} \mu_i\delta^i + (\mu_g + \mu_k)\delta^g + \mu_h\delta^h. \tag{B.5}$$

We have $\tau', \tilde{\tau}' \in T_{\mu'}\mathcal{P}_+(I \setminus \{k\})$, and $h \in I_- \neq \emptyset$. This contradicts to the hypothesis. This shows the Claim for $|I| = n + 1$. ■

Bibliography

[1] Abrahamsen, P. (1997). *A Review of Gaussian Random Fields and Correlation Functions*. Norsk Regnesentral/Norwegian Computing Center.

[2] Akhiezer, N. I. and Glazman, I. M. (2013). *Theory of linear operators in Hilbert space*. Dover Books on Mathematics. Dover Publications, New York, USA.

[3] Alvarado, P. A., Alvarez, M. A., and Stowell, D. (2019). Sparse Gaussian process audio source separation using spectrum priors in the time-domain. In *2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pages 995–999, Brighton, UK.

[4] Amari, S. and Nagaoka, H. (2000). *Methods of Information Geometry*, volume 191.

[5] Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D. W., and O'Neil, M. (2016). Fast direct methods for gaussian processes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38:252–265.

[6] Aristidi, E. (2016). Representation of signals as series of orthogonal functions. In *EAS Publications Series*, pages 99–126, Nice, France. Mathematical Tools for Instrumentation & Signal Processing in Astronomy.

[7] Atkinson, C. and Mitchell, A. F. S. (1981). Rao's distance measure. *Sankhy: The Indian Journal of Statistics*, pages 345–365.

[8] Ay, N., Jost, J., Le, H., and Schwachhofer, L. (2017). *Information geometry*. Springer.

[9] Bachoc, F. (2020). Lecture notes gaussian processes and sensitivity analysis for computer experiments.

[10] Bachoc, F., Gamboa, F., Loubes, J.-M., and Venet, N. (2018). A Gaussian process regression model

for distribution inputs. *IEEE Transactions on Information Theory*, 64:6620–6637.

[11] Bali, J. L., Boente, G., and Wang, J. L. (2011). Robust functional principal components: A projection-pursuit approach. *The Annals of Statistics*, 39:2852–2882.

[12] Barzilai, J. and Borwein, J. M. (1988). Two-point step size gradient methods. *IMA Journal of Numerical Analysis*, 8:141–148.

[13] Beskos, A., Girolami, M., Lan, S., Farrell, P. E., and Stuart, A. M. (2017). Geometric MCMC for infinite-dimensional inverse problems. *Journal of Computational Physics*, 335:327–351.

[14] Besse, P. and Ramsay, J. O. (1986). Principal components analysis of sampled functions. *Psychometrika*, 51:285–311.

[15] Bishop, C. M. (2006). *Pattern recognition and machine learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

[16] Borovitskiy, V., Terenin, A., Mostowsky, P., and Deisenroth, M. P. (2020). Matérn gaussian processes on riemannian manifolds. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20.

[17] Brodzicki, A., Piekarski, M., Kucharski, D., Korjakowska, J. J., and Gorgon, M. (2020). Transfer learning methods as a new approach in computer vision tasks with small datasets. *Foundations of Computing and Decision Sciences*, 45.

[18] Buhmann, M. D. (2003). *Radial basis functions: Theory and implementations*. Cambridge University Press, Cambridge.

[19] Burt, D. R., Rasmussen, C. E., and Wilk, M. V. D. (2020). Variational orthonal features.

[20] Byra, M., Wu, M., Zhang, X., Jang, H., Ma, Y. J., Chang, E. Y., Shah, S., and Du, J. (2020). Knee menisci segmentation and relaxometry of 3d ultrashort echo time cones mr imaging using attention u-net with transfer learning. *Magn Reson Med*, 83.

[21] Byrne, S. and Girolami, M. (2013). Geodesic Monte Carlo on embedded manifolds. *Scandinavian Journal of Statistics*, 40:825–845.

[22] Calin, O. and Udriste, C. (2014). *Geometric Modeling in Probability and Statistics*. Springer-Verlag.

[23] Cavoretto, R., Fasshauer, G., and McCourt, M. (2015). An introduction to the Hilbert-Schmidt SVD using iterated Brownian bridge kernels. *Numerical Algorithms*, 68:393–422.

[24] Chakraborty, R., Seo, D., and Vemuri, B. (2016). An efficient exact-pga algorithm for constant curvature manifolds. pages 3976–3984.

[25] Chihara, T. S. (1978). *An introduction to orthogonal polynomials*. Ellis Horwood series in mathematics and its applications. Gordon and Breach, New York, USA.

[26] Da Veiga, S. and Marrel, A. (2012). Gaussian process modeling with inequality constraints. *Annales de la Faculté des sciences de Toulouse: Mathématique*, 21:529–555.

[27] Damianou, A., Titsias, M., and Lawrence, N. (2011). Variational Gaussian process dynamical systems. In *Advances in Neural Information Processing Systems*, pages 2510–2518, Granada, Spain. Curran Associates, Inc.

[28] Daumé III, H. (2007). Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic. Association for Computational Linguistics.

[29] Day, O. and Khoshgoftaar, T. (2017). A survey on heterogeneous transfer learning. *Journal of Big Data*, 4.

[30] Deheuvels, P. and Martynov, G. V. (2008). A Karhunen-Loève decomposition of a Gaussian process generated by independent pairs of exponential random variables. *Journal of Functional Analysis*, 255:2363–2394.

[31] Delicado, P. (2011). Dimensionality reduction when data are density functions. *Computational*

Statistics & Data Analysis, 55:401–420.

[32] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 View*. John Wiley & Sons.

[33] Doman, B. (2016). *The Classical Orthogonal Polynomials*. World Scientific.

[34] Duan, L., Xu, D., and Tsang, I. W. (2012). Learning with augmented features for heterogeneous domain adaptation. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 667–674. Omnipress.

[35] Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern classification*. Wiley, New York, 2 edition.

[36] Dudley, R. M. (1989). Real analysis and probability. Cambridge University Press.

[37] Eaton, E., desJardins, M., and Lane, T. (2008). Modeling transfer relationships between learning tasks for improved inductive transfer. In *Machine Learning and Knowledge Discovery in Databases*, pages 317–332. Springer Berlin Heidelberg.

[38] Einsiedler, M. and Ward, T. (2017). *Functional Analysis, Spectral Theory, and Applications*. Springer.

[39] Fasshauer, G. (2012). Green's functions: taking another look at kernel approximation, radial basis functions, and splines. In *Approximation Theory XIII*, pages 37–63, New York. Springer.

[40] Fasshauer, G. E. (2007). *Meshfree approximation methods with MATLAB*. World Scientific Publishing Company, Singapore.

[41] Fletcher, P., Lu, C., Pizer, S., and Joshi, S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging*, 23:995–1005.

[42] Foreman-Mackey, D., Agol, E., Angus, R., and Ambikasaran, S. (2017). Fast and scalable gaussian process modeling with applications to astronomical time series. *The Astronomical Journal*, 154.

[43] Fradi, A., Feunteun, Y., Samir, C., Baklouti, M., Bachoc, F., and Loubes, J.-M. (2021). Bayesian

regression and classification using Gaussian process priors indexed by probability density functions. *Information Sciences*, 548:56–68.

[44] Fradi, A. and Samir, C. (2020a). Bayesian cluster analysis for registration and clustering homogeneous subgroups in multidimensional functional data. *Communications in Statistics - Theory and Methods*, 49:1–17.

[45] Fradi, A. and Samir, C. (2020b). Bayesian cluster analysis for registration and clustering homogeneous subgroups in multidimensional functional data. *Communications in Statistics - Theory and Methods*.

[46] Freifeld, O., Hauberg, S., and Black, M. J. (2014). Model transport: Towards scalable transfer learning on manifolds. *CVPR*, pages 1378–1385.

[47] Fritz, J., Nowak, W., and Neuweiler, I. (2009). Application of FFT-based algorithms for large-scale universal kriging problems. *Mathematical Geosciences*, 51:199–221.

[48] Ge, L., Gao, J., Ngo, H., Li, K., and Zhang, A. (2014). On handling negative transfer and imbalanced distributions in multiple source transfer learning. *Stat. Anal. Data Min.*, page 254–271.

[49] Ghanem, R. G. and Spanos, P. D. (1991). *Stochastic finite elements: A spectral approach*. Springer-Verlag, Berlin, Heidelberg.

[50] Gnedenko, B. V. (1962). *The theory of Probability*. Chelsea Publishing Co.

[51] Golub, G. H. and Van Loan, C. F. (1996). *Matrix computations*. The Johns Hopkins University Press, Baltimore, MD, third edition.

[52] Greengard, P. and O'Neil, M. (2022). Efficient reduced-rank methods for Gaussian processes with eigenfunction expansions.

[53] Griffiths, D. J. and Schroeter, D. F. (2018). *Introduction to quantum mechanics*. Cambridge University Press, Cambridge, 3 edition.

[54] Hana, D., Liu, Q., and Fan, W. (2018). A new image classification method using CNN transfer learning and web data augmentation. *Expert Systems with Applications*, 95.

[55] Harel, M. and Mannor, S. (2011). Learning from multiple outlooks. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 401–408.

[56] Hartley, R. I., Trumpf, J., Dai, Y., and Li, H. (2013). Rotation averaging. *International Journal of Computer Vision*, 103:267 – 305.

[57] Helgason, S. (2001). *Differential Geometry, Lie Groups, and Symmetric Spaces*. Academic Press.

[58] Hensman, J., Durrande, N., and Solin, A. (2018). Variational fourier features for gaussian processes. *Journal of Machine Learning Research*.

[59] Hoff, P. D. (2009). Simulation of the matrix Bingham–von Mises-Fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18:438–456.

[60] Holbrook, A., Lan, S., Streets, J., and Shahbaba, B. (2020). Nonparametric Fisher geometry with application to density estimation. *Proceedings of Machine Learning Research*, 124:101–110.

[61] Itoh, M. and Satoh, H. (2015). Geometry of fisher information metric and the barycenter map. *Entropy*, pages 1814–1849.

[62] Jensen, B. S., Nielsen, J. B., and Larsen, J. (2013). Bounded gaussian process regression. In *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*.

[63] Jin, S. (2014). *Gaussian Processes: Karhunen-Loève expansion, smallball estimates and applications in time series models*. PhD dissertation.

[64] Jonathan, W. and Francis, B. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620 – 2648.

[65] Jorgensen, P. and Tian, F. (2019a). Decomposition of Gaussian processes, and factorization of

positive definite kernels. *Opuscula Math*, 39:497–541.

[66] Jorgensen, P. and Tian, F. (2019b). Decomposition of Gaussian processes, and fractorization of positive definite kernels. *Opuscula Mathematica*, 39:497–541.

[67] Jost, J. (2011). *Riemannian geometry and geometric analysis*, chapter 3, pages 89–131. Springer-Verlag, Berlin, Heidelberg.

[68] Julio, B., Joaquin, F., Gonzalo, R., and Felipe, T. (2018). Bayesian learning with wasserstein barycenters. *arXiv e-prints,*.

[69] Jupp, P. E. and Mardia, K. V. (2009). *Directional statistics*. Wiley, London, 1st edition.

[70] Jérémie, B., Raúl, G., Thierry, K., and Alfredo, L. (2017). Geodesic PCA in the wasserstein space by convex pca. *Annales de l'Institut Henri Poincaré*, 53 (1):1 – 26.

[71] Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35.

[72] Karakida, R., Okada, M., and Amari, S. (2016). Dynamical analysis of contrastive divergence learning: Restricted boltzmann machines with gaussian visible units. *Neural Networks*, 79.

[73] Karakida, R., Okada, M., and Amari, S. (2018). Universal statistics of fisher information in deep neural networks: mean field approach. *Journal of Statistical Mechanics: Theory and Experiment*, 2020.

[74] Karimi, A., Ripani, L., and Georgiou, T. T. (2021). Statistical learning in wasserstein space. *IEEE Control Systems Letters,*, 5(3):899–904.

[75] Kent, J. T., Ganeiber, A. M., and Mardia, K. V. (2013). A new method to simulate the Bingham and related distributions in directional data analysis with applications.

[76] Kirby, M. and Sirovich, L. (1990). Application of the Karhunen-Loève procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:103–108.

[77] Kneip, A. and Utikal, K. J. (2001). Inference for density families using functional principal component analysis. *Journal of the American Statistical Association*, 96:519–532.

[78] Konig, H. (1986). *Eigenvalues Distribution of Compact Operators*. Birkhäuser.

[79] Koralov, L. B. and Sinai, Y. G. (2007). *Theory of Probability and Random Processes*. Springer, 2st edition.

[80] Kulis, B., Saenko, K., and Darrell, T. (2011). What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, page 1785–1792, USA. IEEE Computer Society.

[81] Lan, S., Zhou, B., and Shahbaba, B. (2014). Spherical Hamiltonian Monte Carlo for constrained target distributions. *JMLR Workshop Conf Proc*, 32:629–637.

[82] Li, X., Grandvalet, Y., Davoine, F., Cheng, J., Cui, Y., Zhang, H., Belongie, S., Tsai, Y. H., and Yang, M. H. (2020). Transfer learning in computer vision tasks: Remember where you come from. *Image and Vision Computing*, 93.

[83] López-Lopera, A. F. (2019). *Gaussian process modelling under inequality constraints*. PhD dissertation.

[84] López-Lopera, A. F., Bachoc, F., Durrande, N., and Roustant, O. (2018). Finite-dimensional Gaussian approximation with linear inequality constraints. *SIAM/ASA Journal on Uncertainty Quantification*, 6:1224–1255.

[85] Maatouk, H. and Bay, X. (2017). Gaussian process emulators for computer experiments with inequality constraints. *Mathematical Geosciences*, 49:557–582.

[86] Melkumyan, A. and Ramos, F. (2009). A sparse covariance function for exact Gaussian process inference in large datasets. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*, page 1936–1942, Pasadena, California, USA. Morgan Kaufmann Publishers Inc.

[87] Moakher, M. and Zerai, M. (2011). The riemannian geometry of the space of positive-definite matrices and its application to the regularization of positive-definite matrix-valued data. *Journal of Mathematical Imaging and Vision*, 40:171–187.

[88] Nourdin, I. (2012). *Normal approximations with Malliavin calculus: From Stein's method to universality*. Cambridge University Press, Cambridge.

[89] Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[90] Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, pages 1345–1359.

[91] Pennec, X., Fillard, P., and Ayache, N. (2006). A riemannian framework for tensor computing. *International Journal of Computer Vision*, 66:41 – 66.

[92] Petersen, A. and Müller, H. G. (2016). Functional data analysis for density functions by transformation to a Hilbert space. *The Annals of Statistics*, 44:183–218.

[93] Pinele, J., Strapasson, J. E., and Costa, S. (2020). The Fisher-Rao distance between multivariate normal distributions: Special cases, boundsand applications. *Entropy*, 22.

[94] Porat, B. and Friedlander, B. (1986). Computation of the exact information matrix of Gaussian time series with stationary random components. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34:118–130.

[95] Prettenhofer, P. and Stein, B. (2010). Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1127.

[96] Quinonero-Candela, J. and Rasmussen, C. E. (2005). *Analysis of some methods for reduced rank*

Gaussian process regression. Springer-verlag Berlin.

[97] Rachev, S. T., Klebanov, L. B., Stoyanov, S., and Fabozzi, F. J. (2013). *The Methods of Distances in the Theory of Probability and Statistics*. Springer New York.

[98] Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, London.

[99] Rosenstein, M. T., Marx, Z., Kaelbling, L. P., and Dietterich, T. G. (2005). To transfer or not to transfer.

[100] Ruder, S., Peters, M. E., Swayamdipta, S., and Wolf, T. (2019). Transfer learning in natural language processing. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, 93.

[101] Samir, C., Loubes, J.-M., Yao, A.-F., and Bachoc, F. (2019). Learning a gaussian process model on the riemannian manifold of non-decreasing distribution functions. *Pacific Rim International Conference on Artificial Intelligence*.

[102] Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat Appl Genet Mol Biol.*

[103] Seah, C. W., Ong, Y. S., and Tsang, I. W. (2013). Combating negative transfer from predictive distribution differences. *IEEE Transactions on Cybernetics,*, 43:1153–1165.

[104] Sharma, C. and Parikh, S. (2022). Transfer learning and its application in computer vision: A review. *Conference: Transfer Learning and its application in Computer VisionAt: Waterloo, Canada*.

[105] Sinho, C., Tyler, M., Philippe, R., and Austin, J. S. (2020). Gradient descent algorithms for bures-wasserstein barycenters. *In Jacob Abernethy and Shivani Agarwal, editors, Proceedings of Thirty Third Conference on Learning Theory*, 125:1276–1304.

[106] Skiadopoulos, G., Hodges, S., and Clewlow, L. (2000). *The dynamics of implied volatility surfaces*,

pages 197–211. Springer US, Boston, MA.

[107] Skovgaard, L. T. (1984). A riemannian geometry of the multivariate normal model. *Scandinavian Journal of Statistics*, 11:211–223.

[108] Snelson, E. and Ghahramani, Z. (2005). Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, page 1257–1264, Cambridge, MA, USA. MIT Press.

[109] 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*, pages 904–912, Reykjavik, Iceland. Proceedings of Machine Learning Research (PMLR).

[110] Solin, A. and Särkkä, S. (2020). Hilbert space methods for reduced-rank Gaussian process regression. *Statistics and Computing*, 30:419–446.

[111] Sommer, S., Lauze, F., Hauberg, S., and Nielsen, M. (2010). Manifold valued statistics, exact principal geodesic analysis and the effect of linear approximations. volume 6316, pages 43–56.

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

[113] Stednick, Z. (2017). Machine-Learning-with-R-datasets. `https://github.com/stedy/Machine-Learning-with-R-datasets`.

[114] Stein, M. L. (1999). *Interpolation of spatial data*. Springer Series in Statistics. Springer-Verlag New York.

[115] Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer Publishing Company, Incorporated, New York, 1st edition.

[116] Stocker, J. J. (1989). *Differential Geometry*. Wiley-Interscience.

[117] Swiler, L. P., Gulian, M., Frankel, A. L., Safta, C., and Jakeman, J. D. (2020). A survey of constrained Gaussian process regression: Approaches and implementation challenges. *Journal of Machine Learning for Modeling and Computing*, 1:119–156.

[118] Takhanov, R. (2023). On the speed of uniform convergence in Mercer's theorem. *Journal of Mathematical Analysis and Applications*, 518:126718.

[119] Tallis, G. M. (1961). The moment generating function of the truncated multi-normal distribution. *Journal of the royal statistical society series b-methodological*, 23:223–229.

[120] Tao, J. and Fang, X. (2020). Toward multi-label sentiment analysis: a transfer learning based approach. *Journal of Big Data*, 7.

[121] Titsias, M. K. (2009). Variational learning of inducing variables in sparse gaussian processes. In *International Conference on Artificial Intelligence and Statistics*.

[122] Trentin, E. (2018). Soft-constrained Neural Networks for nonparametric density estimation. *Neural Process. Lett.*, 48:915–932.

[123] Trentin, E., Lusnig, L., and Cavalli, F. (2018). Parzen Neural Networks: Fundamentals, properties, and an application to forensic anthropology. *Neural networks : the official journal of the International Neural Network Society*, 97:137–151.

[124] Tsybakov, A. B. (2008). *Introduction to nonparametric estimation*. Springer Publishing Company, Incorporated, New York, 1st edition.

[125] Ulrich, K. R., Carlson, D. E., Dzirasa, K., and Carin, L. (2015). GP kernels for cross-spectrum analysis. In *Advances in Neural Information Processing Systems*, page 1999–2007, Montreal, Canada. MIT Press.

[126] Villani, C. (2009). *Optimal Transport: Old and New*. Springer Science & Business Media.

[127] Wahba, G. (1990). *Spline models for observational data*. Society for Industrial and Applied

Mathematics, Philadelphia.

[128] Wang, C. and Mahadevan, S. (2011). Heterogeneous domain adaptation using manifold alignment. IJCAI'11, page 1541–1546. AAAI Press.

[129] Wang, D. and Zheng, T. F. (2015). Transfer learning for speech and language processing. *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1225–1237.

[130] Wang, L. (2008). *Karhunen-Loeve expansions and their applications*. PhD thesis.

[131] Weiss, K., Khoshgoftaar, T., and Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, 3.

[132] Wendland, H. (2005). *Scattered data approximation*. Cambridge University Press, Cambridge.

[133] Whittle, P. (1963). Stochastic-processes in several dimensions. *Bulletin of the International Statistical Institute*, 40:974–994.

[134] Williams, C. K. I. and Seeger, M. (2000). Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, volume 13, pages 585–591. MIT Press.

[135] Xie, Q., Kurtek, S., Le, H., and Srivastava, A. (2013). Parallel transport of deformations in shape space of elastic surfaces. *Image and Vision Computing*.

[136] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and Q. He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109:43–76.