UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA UFSCar-USP

DIOGO BARBOZA MOREIRA

A NONPARAMETRIC BAYESIAN APPROACH FOR MODELING AND COMPARISON OF FUNCTIONAL DATA

Master dissertation submitted to the Department of Statistics – DEs/UFSCar and the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Interagency Program Graduate in Statistics.

Advisor: Prof. Dr. Luis Ernesto Bueno Salasar

São Carlos
October 2022

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA UFSCar-USP

DIOGO BARBOZA MOREIRA

UMA ABORDAGEM BAYESIANA NÃO PARAMÉTRICA PARA MODELAGEM E COMPARAÇÃO DE DADOS FUNCIONAIS

Tese apresentada ao Departamento de Estatística – DEs/UFSCar e ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Estatística - Programa Interinstitucional de Pós-Graduação em Estatística UFSCar-USP.

Orientador: Prof. Dr. Luis Ernesto Bueno Salasar

São Carlos
Outubro de 2022

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Diogo Barboza Moreira, realizada em 02/09/2022.

Comissão Julgadora:

Prof. Dr. Luis Ernesto Bueno Salasar (UFSCar)

Prof. Dr. José Galvão Leite (USP)

Prof. Dr. Adriano Polpo de Campos (UWA)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

Este trabalho é dedicado à minha família, em especial, ao meu pai, André Luís Moreira.

ACKNOWLEDGEMENTS

Agradeço à Deus, por me conceder a vida e a oportunidade de me tornar um indivíduo melhor, bem como por proporcionar todos os recursos vitais necessários a esse propósito.

Ao meu pai e melhor amigo, André Luís Moreira, pessoa detentora da minha mais nobre confiança e figura inspiradora do meu futuro, sou eternamente grato por me ensinar os principais valores morais da vida e sempre me amparar de todas as maneiras possíveis.

Agradeço à minha família, em especial à minha mãe, Maria Silvia Barbosa dos Santos; à minha boadrasta, Maria Socorro Luciano Moreira; ao meu irmão, Daniel Luciano Moreira e à minha tia, Maria Lucimar Luciano, por constituirem o alicerce e o afago que sustentam todas as minhas conquistas, tanto pessoais quanto profissionais.

Agradeço ao meu orientador, Luis Ernesto Bueno Salasar, pelos esforços em me coordenar do início ao fim do desenvolvimento deste trabalho, pela paciência, pelos conhecimentos compartilhados e pelos valiosos momentos que passamos juntos investigando hipóteses e analisando resultados.

Agradeço aos meus amigos pelos momentos incríveis que compartilhamos durante essa jornada acadêmica.

Agradeço à equipe de docentes da UFSCar e da USP, por compartilhar seus saberes, permitindo meu aperfeiçoamento intelectual para a constituição deste e de outros trabalhos.

Agradeço à equipe de técnicos administrativos da UFSCar e da USP, por manter a infraestrutura e o bom funcionamento das universidades.

Agradeço à minha psicóloga, Maria do Carmo de Sá Borges, pelo suporte emocional dado antes e durante a confecção deste trabalho.

Agradeço a todas as pessoas que contribuíram, de alguma forma, para a realização deste trabalho.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001

ABSTRACT

MOREIRA, D. B. **A nonparametric bayesian approach for modeling and comparison of functional data**. 2022. 79 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

The current advances of technology provides, among other things, several ways of collecting data, which enlarges the possibility of studying new phenomena. Researches focused on studying the functional relation between a variable and some quantity (usually time) produce the called *functional data*. The main feature of this kind of data is that they are registered using devices that can record values almost continuously over time. Suppose two groups of functional data and the interest is to evaluate the similarity of the groups over some range of time. This work proposes a method to compare the groups using predictive samples. The method submit data to a smoothing step using orthonormal functions series and the coefficients of the series are then used to model functional data, due to the bijective relation between the target functions and their respective coefficients. The goal is to estimate the multivariate density associated to the coefficients of each group. Under nonparametric Bayesian context, the densities were estimated using Dirichlet Process Mixture model. Comparison of the functional data groups were performed using a dissimilarity index based on some L_2-distance and estimated using the predcitive samples of the fitted DPM model. The index has a great interpretative appeal and constitute an useful tool for data analysis. Furthermore, it is proposed a bayesian scheme to test the homogeneity of groups of functional data based on the distance between the distributions of the processes for each instant of time. A quick simulation study is presented, as well as preliminary analysis in real functional data set.

Keywords: Functional data, Density estimation, Dirichlet Process Mixtures, Bayesian inference, Nonparametric, Dissimilarity.

RESUMO

MOREIRA, D. B. **Uma abordagem bayesiana não paramétrica para modelagem e comparação de dados funcionais**. 2022. 79 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2022.

Os recentes avanços na tecnologia fornecem, entre outras aspectos, diversas formas de coletar dados, o que aumenta possibilidade de estudar novos fenômenos. Pesquisas cujo foco seja estudar a relação funcional entre uma variável e uma grandeza física (geralmente o tempo) produzem os chamados *dados funcionais*. A principal característica desse tipo de dado é que eles são coletados utilizando dispositivos apropriados para registrar a informação quase que continuamente ao longo do tempo. Suponha dois grupos de dados funcionais e que o interesse é avaliar se os grupos são similares ou não em algum intervalo específico de tempo. Este trabalho apresenta um método para comparação de dois grupos de dados funcionais utilizando amostras preditivas. O método proposto submete os dados originais a uma etapa de suavização utilizando aproximação por as séries de funções ortonormais e os coeficientes da série são utilizados para modelagem dos dados funcionais, devido ao fato de existir uma relação bijetora entre as funções alvo e seus respectivos coeficientes. O objetivo é estimar a densidade multivariada associada aos coeficientes de cada grupos. No contexto de inferência Bayesiana não paramétrica, as densidades foram estimadas através do uso de Misturas de Processos de Dirichlet (DPM). A comparação dos grupos de dados funcionais é então performada atráves de um índice de dissimilaridade baseado em uma medida de distância definida no espaço das funções e estimada através das curvas preditivas amostradas usando o modelo DPM ajustado. O índice possui grande apelo interpretativo e fornece uma ferramenta útil para análise dos dados. É proposto também um esquema bayesiano para testar a homogeneidade das distribuições dos grupos baseado na distância entre as distribuição dos processos para cada instante de tempo. Um rápido estudo de simulação é apresentado, bem como análises preliminares em dados funcionais reais.

Palavras-chave: Dados funcionais, Estimação de densidades, Mistura de Processos de Dirichlet, Inferência Bayesiana, Não paramétrica, Dissimilaridade.

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF TABLES

CONTENTS

CHAPTER

1

INTRODUCTION

Functional Data Analysis (FDA) is a class of statistical methods used to model random processes naturally described as functional in relation to some quantity (usually time), such as meteorology, human growth or human gait studies. The first papers using FDA theory dates on 1990s with the works of James O. Ramsay and Bernard Silverman, who are widely recognized as the founders of this branch of statistics. The technology advance in data collection turns FDA much more feasible, since empirical curves can be recorded closer to the continuous nature proposed by theory. Advances on storage, processing and computational analysis also contribute to efficient application of FDA methods. Moreover, extending the statistical inference tools to more complex spaces expands the range of scientific hypotheses that can be investigated.

Essentially, FDA is indicated when, for each i-th experimental unit, a real valued function is recorded and the set of all curves constitutes the data to be explored, modeled and predicted. The goals of FDA are the same of any other branch of statistics, with the particular feature that FDA deals with whole curves, instead of scalar or vector of real numbers (RAMSAY; SILVERMAN, 1997). Figure 1 shows two examples of functional data studies.

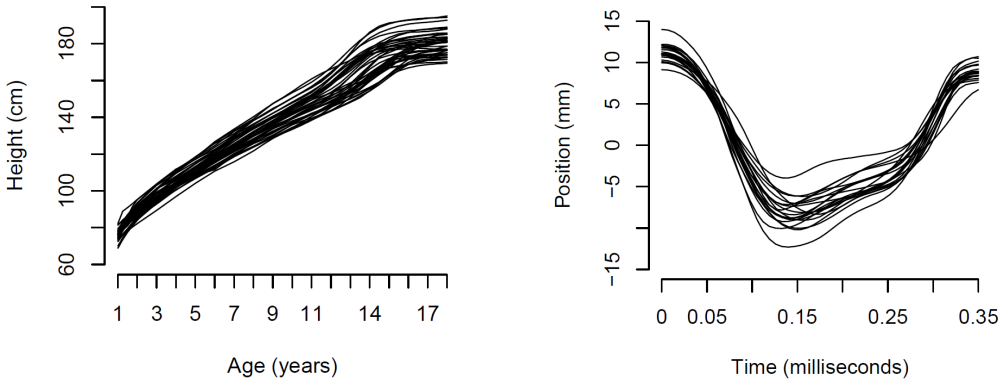
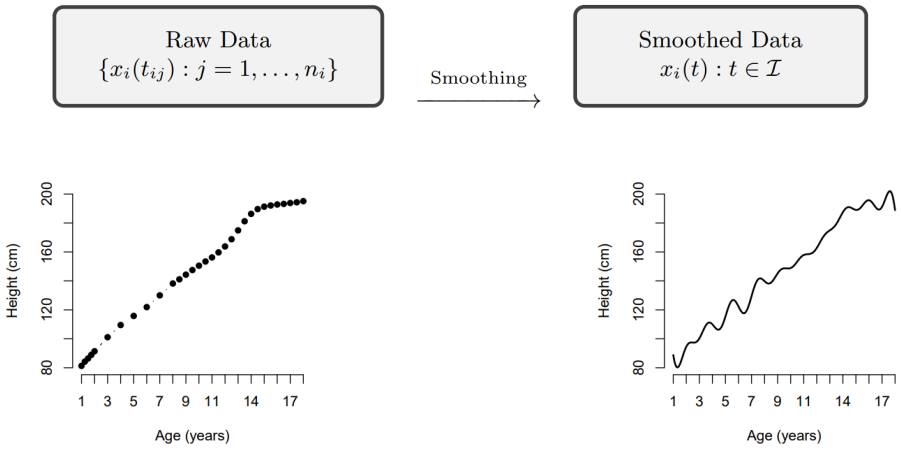


Figure 1 – Berkeley Growth Study Data with heights of 39 boys between the ages 1 and 18 years (left). 20 records of the position of the centre of the lower lip during the uttering of the syllable “bob” (right).

Each curve presented in Figure 1 is the conversion of the recorded values into a function that is computable for any desired argument. If the discrete values are assumed to be errorless, then this process is called *interpolation*, but if they have some observational error that needs removing, then the conversion from discrete data to functions may involve *smoothing*. The scheme below illustrates the idea of smoothing an empirical curve.



We propose to smooth data using **basis function representation**, where each process is represented as a linear combination of orthonormal functions. Essentially, the idea is to model the generating mechanism behind the linear coefficients using nonparametric Bayesian methods, in such a way that we are able to replicate new samples of curves based on predictive distribution. The use of the linear coefficients to represent the functions is feasible due to the strong property that *distinct smoothed processes produces distinct sets of linear coefficients*.

Based on modeling, we tackled the problem of comparing two groups of functional data. Given two independent groups of stochastic processes, we aim: (i) to assess the **dissimilarity of the measurements** of each group for each point of the domain. (ii) to test the **homogeneity of the distributions** of the processes for each point of the domain. The former expresses how likely is to observe two functional measurements close to each other at a certain time and the later is focused on searching for a bayesian criteria to test the following hypotheses:

$$H_{0,t} : X \overset{\mathscr{D}}{=} Y,$$
$$H_{1,t} : X \overset{\mathscr{D}}{\neq} Y.$$

where $\overset{\mathscr{D}}{=}$ means "equal in distribution" and $t \in I$ is the specific instant when the hypotheses are being evaluated. Note that the second goal is different from the first one. While the former focus on provide an intuitive and practical reference of closeness between the processes, the later focus on assess and decide about the equality of the distributions of the processes.

This task has been discussed in several papers and many methods have been proposed. (CUEVAS; FEBRERO; FRAIMAN, 2004) developed a frequentist solution to the problem, comparing the averaged levels of a functional variable using adapted version of the one-way

ANOVA. (ZHANG; PENG; ZHANG, 2010) proposed two statistics to test the equality of two functional group means using Gaussian Process and Bootstrap sampling. (HALL; TAJVIDI, 2002) exhibited a nonparametric approach to test the equality of distributions of two independent functional data groups through permutation tests. Although these tests present great power, they have some inconveniences like assumption of specific parametric forms for the curves or intensive computational needs. (PINI; VANTINI, 2017) provide a tool to test if the average curves of each group are equal over some arbitrary restriction of the domain based on the L^2 distance between the curves and decision is taken using adjusted p-values. The work is structured as following. Chapter 2 discusses the theoretical issues related to smoothing curves using basis function representation. Chapter 3 presents the nonparametric Bayesian models used to model functional data and on Chapter 4, we present the methods to compare two independent groups of functional data. Illustrations with simulated data are presented, as well as an application of the model for the Canadian Data Temperature (RAMSAY; SILVERMAN, 1997).

CHAPTER

2

CONCEPTS OF FUNCTIONAL ANALYSIS AND FUNCTIONAL DATA REPRESENTATION

On this chapter, we define *smoothness*, an important step on functional data analysis that has direct impact on modeling. Here, we review some concepts of *Functional Analysis*, like **basis**, **inner product** and **orthogonality**, which provide the mathematical guarantees needed to represent functions using *basis function representation*. The contents here were almost integrally taken from the book *Introductory Functional Analysis With Applications* (KREYSZIG, 1978).

2.1 A short review of Functional Analysis

Functional analysis (FA) is a branch of mathematics that study abstract spaces endowed with limit-related structures and the linear functions defined on these spaces. Since functional data is approximated by smoothed real continuous curves, it is important to formalize mathematically the space where these curves are defined. For that, three important concepts of FA might be presented: basis, inner product and orthogonality.

Consider an abstract *normed space* $\mathscr{X}$ that contains a sequence $\{e_i\}_{i\geq 1}$ of elements with the property that for every $x \in \mathscr{X}$ there is an **unique** set of scalars $\{\beta_k\}_{k\geq 1}$ such that,

$$||x - (\beta_1 e_1 + \cdots + \beta_n e_n)|| \longrightarrow 0$$

where $||\cdot||$ denotes the norm. The set of $\{e_i\}_{i\geq 1}$ elements is called **Schauder Basis** (or *basis*) for $\mathscr{X}$. Moreover, the series

$$\sum_{k=1}^{\infty} \beta_k e_k$$

is then called *expansion* of x with respect to (e_n) and

$$x = \sum_{k=1}^{\infty} \alpha_k e_k$$

An *inner product* defined on $\mathscr{X}$ is a mapping of $\mathscr{X} \times \mathscr{X}$ into the scalar field K of $\mathscr{X}$; that is, with every pair of vectors $x \in \mathscr{X}$ and $y \in \mathscr{X}$, there is associated a scalar which is written $\langle x, y \rangle$ and is called "the inner product of x and y", such that for all vectors $x, y, z \in \mathscr{X}$ and scalars γ, the following properties hold,

(IP1) $\langle x+y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$

(IP2) $\langle \gamma x, y \rangle = \gamma \langle x, y \rangle$

(IP3) $\langle x, y \rangle = \langle y, x \rangle \qquad \langle x, x \rangle \geq 0$

(IP4) $\langle x, x \rangle = 0 \Longleftrightarrow x = 0$

An **inner product space** is a normed space endowed with inner product, that is, every element of this space satisfies (IP1) to (IP4). If an inner product space $\mathscr{X}$ is **complete**, that is, $\mathscr{X}$ contains the limits of every possible *Cauchy sequence* (see Appendix A for more details) of elements of $\mathscr{X}$, then $\mathscr{X}$ is a **Hilbert Space**. Particularly, an inner product on a Hilbert space defines a norm on it given by,

$$||x|| = \sqrt{\langle x, x \rangle}.$$

Considering $\mathscr{X}$ as a Hilbert space, a measure of distance between distinct elements $x, y \in \mathscr{X}$ is given by

$$d(x, y) = ||x - y|| = \sqrt{\langle x - y, x - y \rangle}$$

Finally, an element $x \in \mathscr{X}$ is said to be *orthogonal* to another element $y \in \mathscr{X}$ if,

$$\langle x, y \rangle = 0$$

that is, if the inner product of x and y is zero. An orthogonal set M of $\mathscr{X}$ is a subset $M \subset \mathscr{X}$ with elements that are pairwise orthogonal. Furthermore, M is said to be an *orthonormal* set if its elements have norm qual to 1, that is, for all $x, y \in M$,

$$\langle x, y \rangle = \begin{cases} 0, \text{ if } x \neq y \\ 1, \text{ if } x = y \end{cases}$$

If an orthogonal or orthonormal set M is countable, we can arrange it in a sequence (x_n) and call it an orthogonal or orthonormal sequence, respectively.

Theorem 2.1. An orthonormal set is linearly independent.

Using Theorem (2.1), we can assume an orthonormal set as a possible choice of basis for $\mathscr{X}$. The great advantage of using orthonormal sequences over arbitrary linearly independent sequences as basis is that if we know that a given x can be represented as a linear combination of some elements of an orthonormal sequence, then the orthonormality makes the actual determination of the coefficients very easy. In fact, if $(e_1, e_2, \dots)$ is an orthonormal sequence in a Hilbert space $\mathscr{X}$ and we have $x \in \text{span}(e_1, \dots, e_n)$ (see Appendix A for definition of **span**), where n is fixed, then

$$x = \sum_{k=1}^{n} \beta_k e_k$$

where

$$\langle x, e_j \rangle = \left\langle \sum \beta_k e_k, e_j \right\rangle = \sum \beta_k \left\langle e_k, e_j \right\rangle = \beta_j.$$

Hence,

$$x = \sum_{k=1}^{n} \langle x, e_k \rangle e_k$$

which shows that the determination of the unknown coefficients in this case is simple.

The above construction can be extended to infinite series and convergence of the corresponding series is discussed below. Given any orthonormal sequence (e_k) in a Hilbert space $\mathscr{X}$, we may consider series of the form,

$$x = \sum_{k=1}^{\infty} \beta_k e_k \tag{2.1}$$

where $\beta_1, \beta_2, \dots$ are any scalars. Such a series converges and has the sum s if there exists an $s \in \mathscr{X}$ such that the sequence (s_n) of partial sums

$$s_n = \beta_1 e_1 + \cdots + \beta_n e_n$$

converges to s. Other important practical advantage of using orthonormal sequences as basis is that if we decide to increase the cutoff in order to improve approximations, it is not necessary to recalculate all the coefficients. In fact, it's only necessary to calculate the new ones. Theorem (2.2) postulates some conditions to guarantee convergence of (2.1).

Theorem 2.2. Let (e_k) be an orthonormal sequence in a Hilbert space $\mathscr{X}$. Then:

(a) The series (2.1) converges (in the norm on $\mathscr{X}$) if, and only if,

$$\sum_{k=1}^{\infty} |\beta_k|^2 < \infty$$

Moreover, the above series corresponds to $||x||^2$.

(b) If (2.1) converges, then the coefficients β_k are the *Fourier coefficients* $\langle x, e_k \rangle$, where x denotes the sum of (2.1); hence in this case, (2.1) can be written

$$x = \sum_{k=1}^{\infty} \langle x, e_k \rangle e_k$$

(c) For any $x \in \mathscr{X}$, the series (2.1) with $\beta_k = \langle x, e_k \rangle$ converges (in the norm of $\mathscr{X}$).

2.1.1 The $L^2[a,b]$ space

Let $\mathscr{C}$ be the vector space of all continuous real-valued functions on $[a,b] \subset \mathbb{R}$. For every $x \in \mathscr{C}$, this space forms a normed space with norm defined by

$$||x|| = \left(\int_a^b x(t)^2 dt \right)^{1/2}$$

This space is not complete, but it can be completed using the fact that every incomplete space has a completion (see Chapter 1 of (KREYSZIG, 1978) for more details). Such completion constitutes the called $L^2[a,b]$ space, which is the **space of all squared Lebesgue-integrable functions**. In such space, the norm can be obtained from the inner product defined by,

$$\langle x, y \rangle = \int_a^b x(t) y(t) dt$$

Since $L^2[a,b]$ is a complete inner product space, then it constitutes a Hilbert space. This means that intuitive notions like distance and angles can be applied to the space of all squared Lebesgue-integrable functions, as well as represent any element through a series of orthonormal basis.

2.2 Functional data representation using Fourier Basis

The orthonormal series approach is the primary mathematical tool for approximation, data compression, and presentation of curves, including functional data. Based on the concepts of the previous section, a function f defined on $L^2[a,b]$ can be fully represented by the following series expansion,

$$f(t) = \sum_{k=1}^{\infty} \beta_k \phi_k(t) \qquad \beta_k = \int_a^b f(t) \phi_k(t) dt \tag{2.2}$$

where $\{\beta_1, \beta_2, \dots\}$ represents the Fourier coefficients defined previously and $\{\phi_1, \phi_2, \dots\}$ is the set of known and fixed *orthonormal* functions which constitutes the *basis* of the $L^2[a,b]$ space.

Since functions $\{\phi_i\}_{i \geq 1}$ are known, the complete representation of f is achieved when $\{\beta_i\}_{i \geq 1}$ is fully determined. However, for practical purposes, the series must be truncated in a

value p, called *cutoff*, that is chosen according to some criteria like cross validation or limited computational resources available.

A desirable characteristic of basis functions is that they have features matching those known to belong to the functions being estimated. A widely used choice for basis functions is the *Fourier Orthogonal System*, whose functions are denoted by,

$$\phi_k(t) = \begin{cases} 1, & \text{if } k = 0, \\ \sqrt{2}\sin(\pi(k+1)t), & \text{if } k = 1,3,5,\ldots, \\ \sqrt{2}\cos(\pi kt), & \text{if } k = 2,4,6,\ldots. \end{cases} \tag{2.3}$$

According to (RAMSAY; SILVERMAN, 1997), the Fast Fourier Transform (FFT) makes it possible to find all the coefficients extremely efficiently. Furthermore, Fourier series is especially useful for extremely stable functions, meaning functions where there are no strong local features and where the curvature tends to be of the same. Ideally, the periodicity of the Fourier series should be reflected to some degree in the data, as is certainly the case for the temperature and gait data.

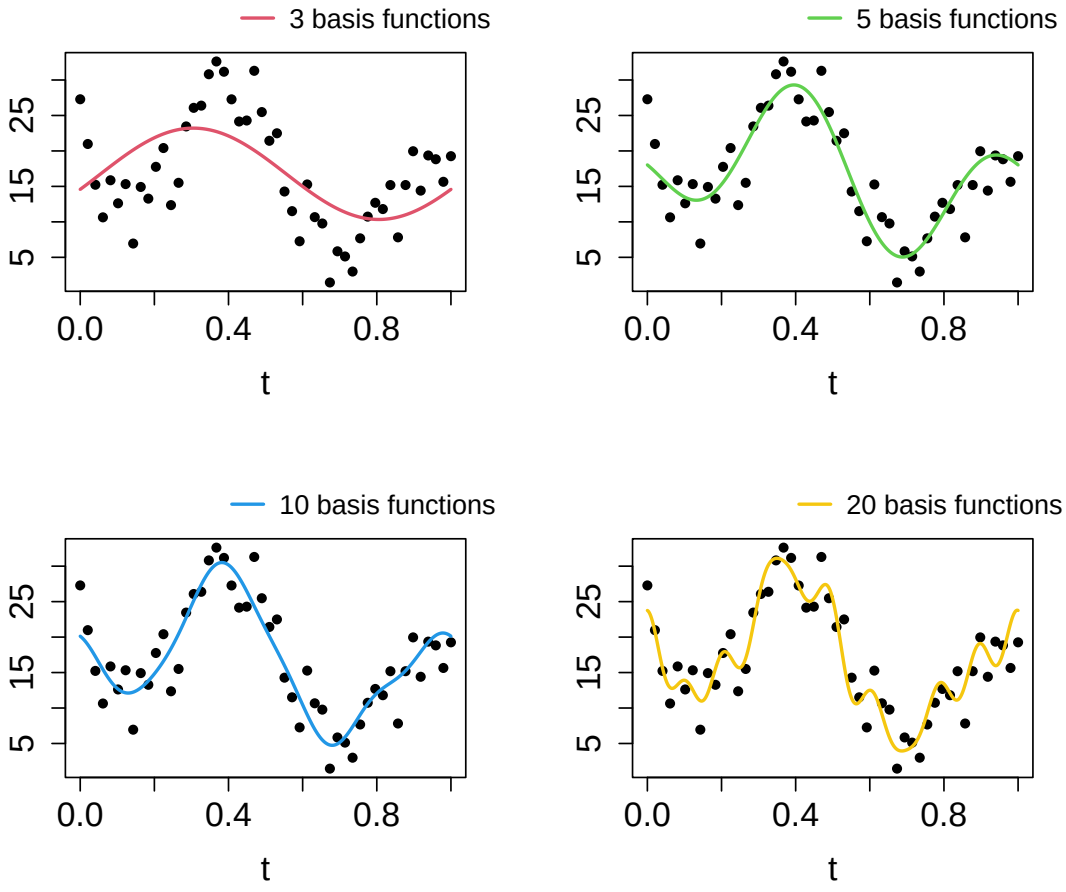


Figure 2 – Examples of smoothness using Fourier basis function.

2.2.1 Estimation of Fourier coefficients using the method of Penalized Least Squares

(GREEN; SILVERMAN, 1993) describes a method to estimate the Fourier coefficients denoted by *Penalized Least Squares*. The procedure is similar to those used in Regression Analysis, but here the approach has some adaptions to the case of functional data smoothing with orthonormal basis.

Let $\{(x_{ij},t_{ij}) : i = 1,\ldots,n; j = 1,\ldots,s\}$ be a sample of n experimental units, where for each one it were recorded s values of some variable of interest over time. For every $i-$th individual, given any twice differentiable function $f_i \in L^2[0,1]$ and a scalar $\lambda > 0$, the following penalized sum of squares is defined,

$$S(f_i) = \sum_{j=1}^{s} \big(x_{ij} - f_i(t_{ij})\big)^2 + \lambda J(f_i). \tag{2.4}$$

The penalized least square estimator $\hat{f}_i$ corresponds to be the minimizes of the functional $S(f_i)$ over all the functions of $L^2[0,1]$. Here, $J(f_i)$ is a quadratic roughness functional of f_i that imposes restrictions to the smoothed curve in order to avoid overfitting problems and λ is a tuning parameter that calibrate the weight of this penalizing and might be estimated.

Assuming the series expansion of $x_i(t)$ discussed previously (with some convenient number of bases p), we can substitute it in (2.4) by,

$$f_i(t) = \sum_{k=0}^{\infty} \beta_{k_i}\phi_k(t) \approx \sum_{k=0}^{p} \beta_{k_i}\phi_k(t) \tag{2.5}$$

which implies that,

$$S(f_i) = \sum_{j=1}^{s} \left(x_{ij} - \sum_{k=0}^{p} \beta_{k_i}\phi_k(t_{ij}) \right)^2 + \lambda J(f_i) \tag{2.6}$$

(GREEN; SILVERMAN, 1993) discusses the use of the following particular roughness penalty $J(f_i)$

$$J(f_i) = \int_0^1 \left\{ \frac{\partial^2 f_i(t)}{\partial t^2} \right\}^2 dt$$

According to the authors, this rule has several convenient mathematical and computational properties. For instance, if two functions differ only by a constant or a linear function, then their roughness should be identical, which can be assessed by the second derivative of the curve under consideration.

Since $J(f_i)$ is quadratic, there will be a $(p+1)\times(p+1)$ matrix such that, for any x_i of the form (2.5),

$$J(f_i) = \boldsymbol{\beta_i}^T \boldsymbol{K} \boldsymbol{\beta_i}$$

where $\boldsymbol{\beta_i} = (\beta_{0_i}, \beta_{1_i}, \ldots, \beta_{p_i})$ are the Fourier coefficients and

$$K_{rs} = \int_0^1 \frac{\partial^2 f_r(t)}{\partial t^2} \frac{\partial^2 f_d(t)}{\partial t^2} dt$$

Using matrix notation, we can rewrite Equation (2.6) as following,

$$S(f_i) = (\boldsymbol{x_i} - \boldsymbol{\Phi\beta_i})^T (\boldsymbol{x_i} - \boldsymbol{\Phi\beta_i}) + \lambda \boldsymbol{\beta_i}^T \boldsymbol{K\beta_i} \tag{2.7}$$

where

$$\boldsymbol{x_i} = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{is} \end{bmatrix}_{s\times 1} \qquad \boldsymbol{\Phi} = \begin{bmatrix} \phi_0(t_1) & \phi_1(t_1) & \ldots & \phi_p(t_1) \\ \phi_0(t_2) & \phi_1(t_2) & \ldots & \phi_p(t_2) \\ \vdots & & & \\ \phi_0(t_s) & \phi_1(t_s) & \ldots & \phi_p(t_s) \end{bmatrix}_{s\times(p+1)}$$

Applying some manipulation in (2.7), we have that

$$S(f_i) = \boldsymbol{x_i}^T \boldsymbol{x_i} - \boldsymbol{x_i}^T \boldsymbol{\Phi\beta_i} - \boldsymbol{\beta_i}^T \boldsymbol{\Phi}^T \boldsymbol{x_i} + \boldsymbol{\beta_i}^T \boldsymbol{\Phi}^T \boldsymbol{\Phi\beta_i} + \lambda \boldsymbol{\beta_i}^T \boldsymbol{K\beta_i}$$

The minimum is achieved when every component of the gradient vector is zero, that is, when

$$\frac{dS(f_i)}{d\boldsymbol{\beta_i}} = \mathbf{0}$$

The gradient of S_i is given by

$$\frac{dS(f_i)}{d\boldsymbol{\beta_i}} = -2\boldsymbol{\Phi}^T \boldsymbol{x_i} + 2\boldsymbol{\Phi}^T \boldsymbol{\Phi\beta_i} + 2\lambda \boldsymbol{K\beta_i}$$

and

$$\frac{dS(f_i)}{d\boldsymbol{\beta_i}} = \mathbf{0} \iff \hat{\boldsymbol{\beta}}_{\boldsymbol{i}} = \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi} + \lambda \boldsymbol{K}\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{x_i} \tag{2.8}$$

To certificate that $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ minimizes (2.7), the second derivative of S_i with respect to $\boldsymbol{\beta_i}$ applied to $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ must be positive. In fact,

$$\frac{d^2 S(f_i)}{d\boldsymbol{\beta_i}^2} = 2\boldsymbol{\Phi}^T \boldsymbol{\Phi} + 2\lambda \boldsymbol{K} > \mathbf{0}$$

whatever the value of $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ is. Hence, $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ is the *Penalized Least Square Estimator* of the Fourier coefficients for i-th individual.

Note that the estimator (2.8) was developed assuming a series of $p+1$ Fourier bases, where p is fixed. On next subsection, we discuss some criteria to choose a good cutoff p, in order to avoid under or overfitting and balance goodness-of-fit versus computational efforts.

2.2.2 A good choice for the cutoff p

The complete representation of the function $f_i(t)$ using series expansion is achieved when the number of bases $p = \infty$. Of couse, this is unmanageable in practice, which means that we must choose p that provides a good representation of $f_i(t)$ and do not generate high computational efforts on the next modeling steps. Here we discuss the choice of the cutoff p based on *k-fold cross validation* (see Chapter 1 of (IZBICKI; SANTOS, 2020) for examples), combined to the penalizing rule discussed previously. Basically, the method is performed as following :

1. For each i-th individual, consider the observed points $(x_{i1}, \ldots, x_{is})$ split in k disjoint groups named by $L_1, \ldots, L_k$;

2. Apply the penalized least square estimator (2.8) with p basis to estimate the Fourier curve $f_i(t)$ using all the observed points **except** those in L_j. This estimated curve is denoted $\hat{f}_{i_{-j}}(t)$;

3. Estimate the Mean Squared Error associated to the estimated curve, named $R(\hat{f}_i)$, using the following measure,

$$\hat{R}(\hat{f}_i) = \frac{1}{n} \sum_{j=1}^{k} \sum_{h \in L_j} \left(f_{ih} - \hat{f}_{i_{-j}}(t_h) \right)^2 \tag{2.9}$$

Reproducing the steps 1 to 3 for other values of p makes it possible to compare the errors associated to the choice of number of basis and then choose a good cutoff p, that is, the the one that returns the minimum error.

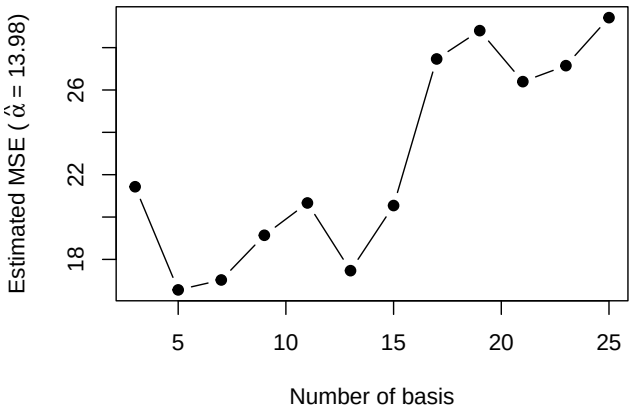


Figure 3 – MSE evaluated using k-fold cross validation (3 lotes) with penalizing rule on Figure 2. The best number of basis was chosen fixing the tuning parameter estimated at $\hat{\alpha} = 13.98$.

CHAPTER

3

NONPARAMETRIC BAYESIAN ESTIMATION OF FUNCTIONAL DATA DISTRIBUTION

The nonparametric Bayesian approach to model Functional Data is presented. The main idea is to regard the Fourier coefficients obtained by some smoothing method as multivariate observations arising from a discrete mixture of multivariate Normal densities. Methods to estimate the distribution of the FD and to predict a new FD observation will be discussed.

Adopting the notation of Chapter 2, let $\{(x_{ij},t_{ij}) : i = 1,\ldots,n, j = 1,\ldots,k_i\}$ be the registered observations associated to the n experimental units. For each experimental unit i, consider

$$\boldsymbol{\beta_i} = (\beta_{0,i},\beta_{1,i},\ldots,\beta_{p,i}) \in \mathbb{R}^{p+1}, \qquad i = 1,\ldots,n,$$

the vectors of smoothed Fourier coefficients. Then, we shall assume that

$$\boldsymbol{\beta_1},\ldots,\boldsymbol{\beta_n} \overset{i.i.d}{\sim} P_{\boldsymbol{\beta}},$$

where $P_{\boldsymbol{\beta}}$ is a probability measure over $\mathbb{R}^{p+1}$. The probability $P_{\boldsymbol{\beta}}$ characterizes the random mechanism that originates the Fourier coefficients, which in turn, define a functional data observation. We assume that $P_{\boldsymbol{\beta}}$ is distributed as a mixture of Dirichlet Process prior (ANTONIAK, 1974), then we discuss methods of estimating the posterior distribution of $P_{\boldsymbol{\beta}}$ and the predictive distribution of a new observation given the observed data (posterior predictive distribution).

The chapter is organized as following. Sections 3.1 and 3.2 defines the Dirichlet Process and Dirichlet Process Mixture models, respectively. Section 3.2 approaches the specific case of DPM model with Normal kernels. Section 3.4 exhibits the analytical distribution of $X(t)$ for each time $t \in [0,1]$. Section 3.5 illustrates the methods fitting DPM model to some simulated data.

3.1 Dirichlet Process

The Dirichlet Process (DP), first introduced by (FERGUSON, 1973), is a nonparametric Bayesian method to model the data generating mechanism as a random probability measure. The DP can be seen as infinite-dimensional generalization of the k-variate Dirichlet distribution, about which we shall make a brief review.

Definition 3.1 (Dirichlet Distribution)**.** The random vector $\theta = (\theta_1, \ldots, \theta_k)$ is said to follow a Dirichlet distribution with parameter $(\alpha_1, \ldots, \alpha_k) \in [0, +\infty)^k$, $\theta \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$, if it can be represented as

$$\theta_j = \frac{U_j}{\sum_{i=1}^{k} U_i}, \qquad j = 1, \ldots, k, \tag{3.1}$$

where $U_1, \ldots, U_k$ are k independent random variables such that $U_j \sim \text{Gamma}(\alpha_j, 1)$, $j = 1, \ldots, k$. The Gamma distribution with parameter (α, β) has support on the non-negative real line $[0, +\infty)$, where $\alpha \geq 0$ is the shape parameter and $\beta > 0$ is the scale parameter. If $\alpha = 0$, the distribution is degenerated at 0, whereas if $\alpha > 0$ the distribution has density

$$f(u|\alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha - 1)} u^{\alpha} e^{-\beta x} I_{[0,+\infty)}(u), \qquad u \in \mathbb{R},$$

where $\Gamma(x)$ is the Gamma function.

From Definition 3.1, it follows that if $\theta = (\theta_1, \ldots, \theta_k) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$, then

(i) $\theta_j \in [0, 1]$ for all $j = 1, \ldots, k$;

(ii) $\sum_{i=1}^{k} \theta_i = 1$.

Since a vector θ that satisfies properties (i) and (ii) characterizes the probability mass function of a discrete random variable X assuming k distinct values, say $1, 2, \ldots, k$, the Dirichlet distribution can be interpreted as a distribution over all probability mass function of that random variable. The Dirichlet Process has an analogous interpretation for a general random variable X.

For $\theta = (\theta_1, \ldots, \theta_k) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$ and all $\alpha_j > 0$, the vector $(\theta_1, \ldots, \theta_{k-1})$ has probability density function given by

$$p(\theta_1, \ldots, \theta_{k-1}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2) \cdots \Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \cdots \theta_{k-1}^{\alpha_{k-1} - 1} \theta_k^{\alpha_k - 1} I_{\Delta_{k-1}}(\theta_1, \ldots, \theta_{k-1}), \tag{3.2}$$

where $\theta_k = 1 - \theta_1 - \cdots - \theta_{k-1}$ and Δ_{k-1} is the $(k-1)$-dimensional simplex defined by

$$\Delta_{k-1} = \left\{ (\theta_1, \ldots, \theta_{k-1}) \in [0, 1]^{k-1} : \sum_{i=1}^{k-1} \theta_i \leq 1 \right\}.$$

The following proposition characterizes the first two moments of the Dirichlet distribution.

Proposition 3.1. If $\theta = (\theta_1, \ldots, \theta_k) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$, then

(a) $E[\theta_j] = \dfrac{\alpha_j}{\sum_{i=1}^{k} \alpha_i}$, for $j = 1, \ldots, k$;

(b) $\text{Var}[\theta_j] = \dfrac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)}$ for $j = 1, \ldots, k$;

(b) $\text{Cov}(\theta_i, \theta_j) = -\dfrac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)}$ if $i \neq j$, $i, j \in \{1, \ldots, k\}$.

Proof. The proof of (a), (b) and (c) follows from the identity

$$\int_{\Delta_{k-1}} \theta_1^{\alpha_1 - 1} \cdots \theta_{k-1}^{\alpha_{k-1} - 1} \theta_k^{\alpha_k - 1} d\theta_1 \ldots d\theta_{k-1} = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2) \cdots \Gamma(\alpha_k)}{\Gamma(\alpha_1 + \cdots + \alpha_k)},$$

which is derived from the fact that (3.2) is a probability density function over Δ_{k-1}. The details are ommited. □

Definition 3.2 (Dirichlet Process)**.** Let $(\Omega, \mathscr{A}, \mathbb{P})$ be a probability space, that is, Ω is the sample space, $\mathscr{A}$ a σ-field of subsets of Ω and $\mathbb{P}$ a probability measure over the measurable space $(\Omega, \mathscr{A})$. A random probability G follows a Dirichlet Process (DP) with baseline probability G_0 and concentration parameter α, $\alpha > 0$, if for each measurable partition $\{A_1, \ldots, A_k\}$ of Ω, $k \geq 1$, we have that

$$(G(A_1), G(A_2), \ldots, G(A_k)) \sim \text{Dirichlet}\,(\alpha G_0(A_1), \alpha G_0(A_2), \ldots, \alpha G_0(A_k)) \tag{3.3}$$

where $\text{Dirichlet}(\cdot|\alpha_1, \ldots, \alpha_k)$ represents the Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_k \geq 0$. The obtained random probability G is denoted by

$$G \sim \text{DP}(\alpha, G_0).$$

A probability measure G sampled from a DP is regarded as a random selection of a probability measure G over Ω.

The existence of a process $\{G(A) : A \in \mathscr{A}\}$ satisfying property (3.3) is proved in (FERGUSON, 1973) by the verification of the consistency conditions of the Kolmogorov Extension Theorem (BILLINGSLEY, 2008).

Proposition 3.2 presents the mean and variance of G for a fixed $A \in \mathscr{A}$.

Proposition 3.2. If $G \sim DP(\alpha, G_0)$ and $A \in \mathscr{A}$, then

$$E[G(A)] = G_0(A), \tag{3.4}$$

$$Var[G(A)] = \frac{G_0(A)[1 - G_0(A)])}{\alpha + 1}. \tag{3.5}$$

Proof. From property (3.3), it follows that $(G(A), G(A^c)) \sim \text{Dirichlet}(\alpha G_0(A), \alpha G_0(A^c))$. Then, from Proposition 3.1 we have that

$$E[G(A)] = \frac{\alpha G_0(A)}{\alpha(G_0(A) + G_0(A^c))} = G_0(A),$$

$$\text{Var}[G(A)] = \frac{\alpha G_0(A)(\alpha - \alpha G_0(A))}{\alpha^2(\alpha+1)} = \frac{G_0(A)(1 - G_0(A))}{\alpha+1}.$$

□

According to Proposition 3.2, G_0 can be interpreted as the prior mean for the random probability G and α as a degree of trust in G_0, in the sense that larger values of α makes the variance of G smaller.

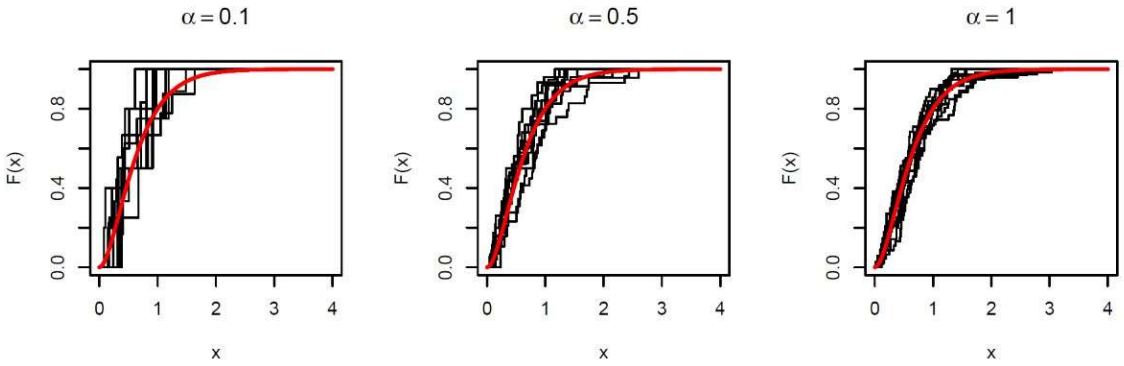


Figure 4 – Illustration of Dirichlet Process. Black curves represents CDF of 10 DP samples of a DP with baseline measure $\Gamma(2,3)$ and concentration parameter α varying. Red line is the real CDF curve.

Proposition 3.3. If $G \sim DP(\alpha, G_0)$, then G has the same support of G_0 and is discrete almost surely.

Proof. For detailed proof, see Section 4 of (FERGUSON, 1973). □

The discreteness nature of DP priors makes it unappealing when we believe that data is continuous. In Section 3.2, we will see a way to overcome this issue using mixtures of Dirichlet Processes.

3.1.1 Posterior and predictive distributions of a DP prior

The most important and convenient property of Dirichlet Process prior, proved by (FERGUSON, 1973), is its conjugacy under independent and identically distributed sampling.

Proposition 3.4. Let $X_1, \ldots, X_n | G \overset{iid}{\sim} G$ and $G \sim DP(\alpha, G_0)$. Then,

$$G|x_1, \ldots, x_n \sim DP\left(\alpha + n, \frac{\alpha}{\alpha+n} G_0 + \frac{n}{\alpha+n} \sum_{i=1}^{n} \delta_{x_i}\right)$$

where δ_{x_i} is the Dirac measure, that is, for any $A \subseteq \Omega$,

$$\delta_{x_i}(A) = \begin{cases} 1, & \text{if } x_i \in A \\ 0, & \text{if } x_i \notin A \end{cases}$$

Furthermore, DP prior has closed form for posterior predictive distribution.

Proposition 3.5. Let $X_1, \ldots, X_n | G \overset{iid}{\sim} G$ and $G \sim DP(\alpha, G_0)$. Then,

$$X_{n+1} | X_1, \ldots, X_n \sim \frac{\alpha}{\alpha + n} G_0 + \frac{n}{\alpha + n} \widehat{F}_X,$$

where

$$\widehat{F}_X(A) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(A), \qquad A \in \mathscr{A}.$$

We note that the posterior baseline measure is a mixture of G_0 and the empirical distribution function $\hat{F}_X$.

3.1.2 Stick Breaking representation

An alternative representation of the random probability measure G generated by the Dirichlet Process (DP) was formulated by (SETHURAMAN, 1994). This representation makes explicit the discreteness property of the probability measures generated from Dirichlet process, besides providing a simple algorithm to simulate from the Dirichlet Process.

Theorem 3.1. If $G \sim DP(\alpha, G_0)$, then it can be represented, with probability one, by

$$G(A) = \sum_{i=1}^{\infty} \omega_i \delta_{x_i}(A), \qquad A \in \mathscr{A},$$

where δ_x is the Dirac measure over x, $x_1, x_2, \ldots$ are i.i.d observations from G_0, $\omega_1, \omega_2, \ldots$ are weights given by the expressions

$$\omega_1 = \theta_1,$$
$$\omega_i = \theta_i \prod_{j=1}^{i-1} (1 - \theta_j),$$

where $\theta_1, \theta_2, \ldots$ are i.i.d observations from Beta$(1, \alpha)$.

Proof. For a detailed proof, see (SETHURAMAN, 1994). □

This method is known as Stick Breaking construction, alluding to the metaphor that compares the weights to the length of a stick. At the beginning, the stick has length 1. In the first step, we break up the stick at location $\theta_1 \sim$ Beta$(1, \alpha)$. After this step, the remaining stick has

length $1-\theta_1$ and it is broken at point $\theta_2(1-\theta_1)$. The procedure goes on, always breaking up the stick at a fraction θ_i of the length $\prod_{k=1}^{i-1}(1-\theta_k)$ of remaining stick at stage $i = 1,2,\ldots$.

For simulation of Dirichlet Process using Stick Breaking, one can apply the following algorithm.

Algorithm 1 – STICK BREAKING

Input: parameters α, G_0 and error tolerance ε
Output: samples $\boldsymbol{x}$ of G_0; weights $\boldsymbol{\omega}$

Initialize $X_1 \sim G_0$, $\theta_1 \sim Beta(1,\alpha)$, $\omega_i = \theta_i$, i = 1

Repeat until $\sum_{i\geq 1}\omega_i = 1-\varepsilon$
Simulate $X_i \sim G_0$ and $\theta_i \sim Beta(1,\alpha)$
Make $\omega_i = \theta_i\prod_{j<i}(1-\theta_j)$ using (1)
Increment i

3.2 Dirichlet Process Mixtures

One way to mitigate the discreteness limitation of the DP is to add to the discrete distribution G a convolution with a continuous kernel. Let $(X_1, X_2, \ldots, X_n)$ be an i.i.d. sample with unknown distribution. A Dirichlet process mixture prior (DPM) posits that,

$$\begin{aligned} X_1|\theta_1 &\sim f(x|\theta_1), \\ &\vdots \\ X_n|\theta_n &\sim f(x|\theta_n), \\ \theta_1,\ldots,\theta_n|G &\overset{i.i.d}{\sim} G, \\ G &\sim DP(\alpha, G_0), \end{aligned} \tag{3.6}$$

where $f(x_i|\theta)$ is a parametric distribution (often referred to as the kernel of the mixture), which is indexed by a finite dimensional parameter θ. The hierarchical model (3.7) implies on represent the unknown distribution of X as a mixture of distributions with mixing measure given by G. Observe that the above modelling is equivalent to assume that

$$\begin{aligned} X_1,\ldots,X_n|G &\overset{i.i.d}{\sim} f_G(x) = \int f(x|\theta)\,dG(\theta), \\ G|\alpha, G_0 &\sim DP(\alpha, G_0). \end{aligned}$$

Then, conditional on G, the distribution of X_i has a density that is a mixture of the parametric family of densities $\{f(x|\theta) : \theta \in \Omega\}$ using the discrete distribution G over Ω. Since the distribution G is discrete, we may assume the probability of a set $A \subset \Omega$ is

$$G(A) = \sum_{i=1}^{\infty}\omega_i\delta_{\theta_i}(A),$$

where $\theta_i \in \Omega$, $\omega_i \in [0,1]$, $i = 1,2,\ldots$, and $\sum_i \omega_i = 1$. Then, the mixed distribution f_G is given by

$$f_G(x) = \sum_{i=1}^{\infty} \omega_i f(x|\theta_i). \tag{3.7}$$

where

$$\theta_i \sim G_0 \qquad \omega_i = v_i \prod_{k<i}(1-v_k) \qquad v_i \sim Beta(1,\alpha)$$

As argued by (MÜLLER; RODRIGUEZ, 2013), this representation highlights the nature of the DPM model as a discrete mixture. DP mixtures are countable mixtures with an infinite number of components and a specific prior on the weights and the component-specific parameters. Working with an infinite number of components is particularly appealing because it ensures that, for appropriate choices of the kernel $f(x_i|\theta)$, the DPM model has support on a large classes of distributions.

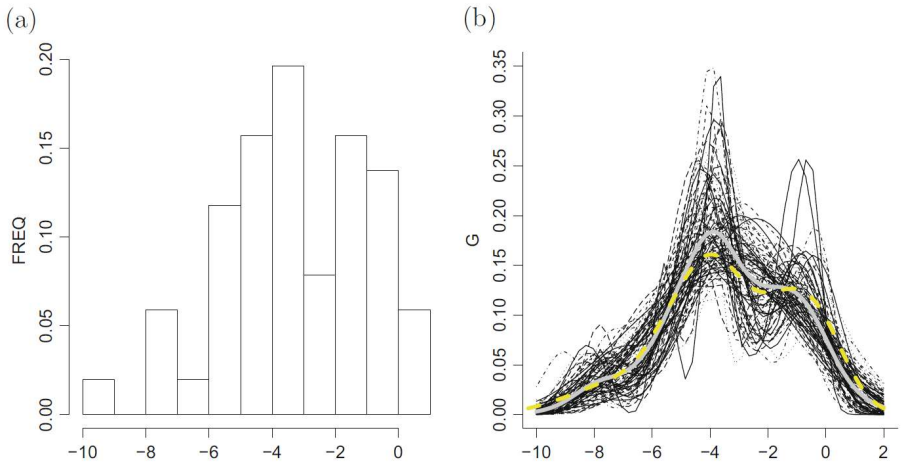


Figure 5 – Data (panel a) and posterior inference for G (panel b). Panel (b) shows 96 posterior draws of the mixture model f_G based on $G \sim p(G|\boldsymbol{x})$ (thin black curves) and the posterior mean $E(f_G|\boldsymbol{y})$ (thick grey curve). For comparison the figure also shows a kernel density estimate (dashed thick yellow line). *Source:* (MÜLLER *et al.*, 2015)

Another consequence of DPM model is that it induces clustering among observations, with α controlling the prior expected number of clusters on the sample. If $\alpha \to 0$, the model reduces to a single component mixture, where all observations are i.i.d. from $f(x|\theta)$ and $\theta \sim G_0$, i.e., a fully parametric model. On the other hand, for $\alpha \to \infty$, each observation is assigned its own singleton cluster and we have $y_i \sim \int f(x_i|\theta) dG_0(\theta)$.

3.3 Modeling functional data using DPM with multivariate normal kernels

Consider $\{X_1(t), X_2(t), \ldots, X_n(t)\}$ to be random curves defined on $t \in [0,1]$ and let $\boldsymbol{B} = \{\boldsymbol{\beta_1}, \ldots, \boldsymbol{\beta_n}\}$ be their respective set of Fourier coefficients obtained after smoothing, where $\boldsymbol{\beta_i} = \{\beta_{0_i}, \ldots, \beta_{p_i}\}$. Assuming a Dirichlet Process Mixture model with normal kernels, we have the following hierarchical structure,

$$\begin{aligned}
\boldsymbol{\beta_1}|\mu_1,\Sigma_1 &\sim N_{p+1}(\mu_1,\Sigma_1)\\
&\vdots\\
\boldsymbol{\beta_n}|\mu_n,\Sigma_n &\sim N_{p+1}(\mu_n,\Sigma_n)\\
(\mu_i,\Sigma_i),\ldots,(\mu_n,\Sigma_n)|G &\overset{iid}{\sim} G\\
G|\alpha,G_0 &\sim DP(\alpha,G_0)\\
(\alpha,G_0)|\psi &\sim \pi_\psi
\end{aligned} \tag{3.8}$$

which can be graphically represent as following,

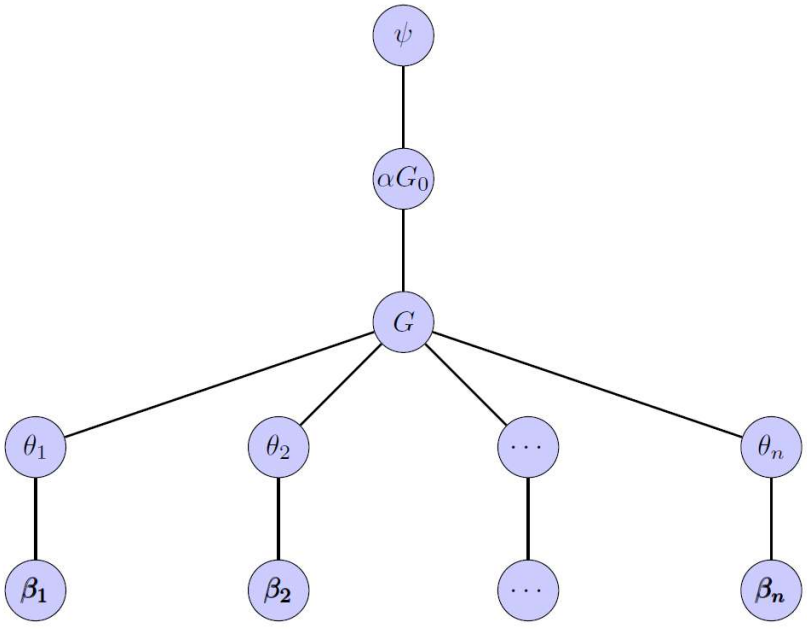


Figure 6 – Illustration of the hierarchical structure of the DPM model.

Using (3.7), we conclude that,

$$\boldsymbol{\beta_i}|G \overset{iid}{\sim} \sum_{h=1}^{\infty} \omega_h \phi_{p+1}(\boldsymbol{\beta_i}|\mu_h,\Sigma_h) \tag{3.9}$$

where ϕ_{p+1} represents the multivariate normal density with dimension $p+1$. The predictive distribution of $\boldsymbol{\beta_i}$ is given by,

$$p(\boldsymbol{\beta_{n+1}}|\boldsymbol{\beta_1},\ldots,\boldsymbol{\beta_n}) = \int p(\boldsymbol{\beta_{n+1}}|\theta_1,\ldots,\theta_n)\ dp(\theta_1,\ldots,\theta_n|\boldsymbol{\beta_1},\ldots,\boldsymbol{\beta_n}) \tag{3.10}$$

where $\theta_j = (\mu_j, \Sigma_j)$. The integrand $p(\theta_1,\ldots,\theta_n|\boldsymbol{\beta_1},\ldots,\boldsymbol{\beta_n})$ is the posterior distribution of normal parameters denoted by (MÜLLER; ERKANLI; WEST, 1996)

$$p(\theta_1,\ldots,\theta_n|\boldsymbol{\beta_1},\ldots,\boldsymbol{\beta_n}) \propto \prod_{i=1}^{n} \phi_{p+1}(\boldsymbol{\beta_i}|\theta_i)\, \frac{\alpha G_0(\theta_i) + \sum_{j<i}\delta_{\theta_i}(\theta_j)}{\alpha + i - 1} \tag{3.11}$$

The other integrand in (3.10) can be rewritten as,

$$p(\boldsymbol{\beta_{n+1}}|\theta_1,\ldots,\theta_n) = \int p(\boldsymbol{\beta_{n+1}}|\theta_{n+1}) p(\theta_{n+1}|\theta_1,\ldots,\theta_n) d\theta_{n+1}$$

Since θ_i has DP prior on its probability measure, the following property is attained,

$$\theta_{n+1}|\theta_1,\ldots,\theta_n \sim \frac{\alpha G_0(\theta_{n+1})}{\alpha+n} + \frac{\sum_{i=1}^{n}\delta_{\theta_i}(\theta_{n+1})}{\alpha+n} \tag{3.12}$$

Moreover, relation (3.12) can be rewritten using the fact that DPM models is also used to cluster data. In this sense, consider k groups of n_j observations per group ($j = 1,\ldots,k$), $\sum_{j=1}^{k} n_j = n$ and let $\theta_j^* = (\mu_j^*, \Sigma_j^*)$ be the normal parameters of group j. Then, (3.12) can be written as,

$$\theta_{n+1}|\theta_1,\ldots,\theta_n \sim \frac{\alpha G_0(\theta_{n+1})}{\alpha+n} + \frac{\sum_{j=1}^{k} n_j \delta_{\theta_j^*}(\theta_{n+1})}{\alpha+n} \tag{3.13}$$

Hence,

$$p(\boldsymbol{\beta_{n+1}}|\theta_1,\ldots,\theta_n) = \frac{\alpha}{\alpha+n}\int f(\boldsymbol{\beta_{n+1}}|\theta) dG_0(\theta) + \frac{1}{\alpha+n}\sum_{j=1}^{k} n_j f(\boldsymbol{\beta_{n+1}}|\theta_j^*) \tag{3.14}$$

and, as argued by (MÜLLER; ERKANLI; WEST, 1996), for practically important cases where α/n is negligible,

$$p(\boldsymbol{\beta_{n+1}}|\theta_1,\ldots,\theta_n) \approx \sum_{j=1}^{k} w_j f(\boldsymbol{\beta_{n+1}}|\theta_j^*)$$

where $w_j = n_j/n$.

Hence, if we were able to evaluate the posterior $p(\theta_j^*|\boldsymbol{\beta_1}\ldots,\boldsymbol{\beta_n})$, then it would be possible to evaluate $p(\boldsymbol{\beta_{n+1}}|\theta_1,\ldots,\theta_n)$ and, consequently, obtain the predictive samples of Fourier coefficients.

3.3.1 Posterior simulation

Keeping the choices made by (MÜLLER; ERKANLI; WEST, 1996), we assumed that the baseline measure G_0 for (μ_i, Σ_i) is given by (assuming independence),

$$\begin{aligned} \mu_i|m,V &\sim N_p(\mu_i; m, V) \\ \Sigma_i^{-1}|s,S &\sim W_p(\Sigma_i^{-1}; s, (sS)^{-1}) \end{aligned}$$

where $W_p(\Sigma_i^{-1}; s, (sS)^{-1})$ denotes the Wishart distribution with s degrees of freedom and expectant matrix $(sS)^{-1}$. The following priors for the hyperparameters m, S, V, δ and α were assumed mutually independent:

$$m \sim N_p(m; a, A) \quad S \sim W_p(S; q, q^{-1}R) \quad V^{-1} \sim W_p(V^{-1}; c, (cC)^{-1}) \quad \alpha \sim \Gamma(\alpha; a_0, b_0) \tag{3.15}$$

(MÜLLER; ERKANLI; WEST, 1996) propose clustering observations using data augmentation. Let $\theta^{(i)} = (\theta_1, \theta_2, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$ denote the main parameter vector with θ_i removed and k_i being the number of distinct values in $\theta^{(i)}$. Denote these distinct values by $\theta_j^{i^*}$ and suppose that $\theta_j^{i^*}$ occurs n_{ij} times. Denoting by $\boldsymbol{B} = \{\boldsymbol{\beta_1}, \ldots, \boldsymbol{\beta_n}\}$, then

$$\theta_i | \theta^{(i)}, \boldsymbol{B}, m, S, V, \alpha \sim q_{i0} G^{(i)}(\theta_i) + \sum_{j=1}^{k} q_{ij} \delta_{\theta_j^{i^*}}(\theta_i)$$

where $G^{(i)}$ is the posterior of θ_i under the prior G_0 updated by the likelihood $f(B_i|\theta_i)$ and the mixing weights q_{ij} are

$$q_{i0} \propto \alpha \int f(\boldsymbol{\beta_i}|\theta_i) dG_0(\theta_i) \qquad q_{ij} \propto n_{ij} f\left(\boldsymbol{\beta_i}|\theta_j^{i^*}\right) \qquad j = (1, \ldots, k_i)$$

Here, q_{i0} is interpreted as the probability of i-th observation belonging to a new cluster and q_{ij} is the probability of it be allocated to cluster j.

Based on the univariate approach developed by (MACEACHERN, 1994), labeling is made introducing a configuration vector $\mathscr{I} = (I_1, \ldots, I_n)$, where each I_i takes values in the set $\{1, 2, \ldots, n\}$ and will be updated at each Gibbs iteration using the posterior distribution of q_{i0} and q_{ij}. For instance, if $I_i = I_{i'} = j$, then observations B_i and $B_{i'}$ share a common parameter $\theta_i = \theta_{i'} = \theta_j^*$. Assuming this structure, the joint posterior distribution of the parameters is denoted by,

$$\begin{aligned} p(\boldsymbol{\theta^*}, m, S, V | \boldsymbol{B}, \mathscr{I}, k, \alpha) &= \left[\prod_{j=1}^{k} \left[\prod_{i:I_i=j} N_p(\boldsymbol{\beta_i}; \mu_j^*, \Sigma_j^*)\right] \cdot N_p(\mu_j^*; m, V) \cdot W_p\left(\Sigma_j^{-1^*}; s, (sS)^{-1}\right)\right] \\ &\times N(m; a, A) \cdot W_p\left(S; q, q^{-1}R\right) \cdot W_p\left(V^{-1}; c, (cC)^{-1}\right) \end{aligned} \tag{3.16}$$

To perform Gibbs sampling using (3.16), we need the full conditional posterior density of each parameter. (MÜLLER; ERKANLI; WEST, 1996) provide these full conditionals and more details about how to achieve these formulas. In Appendix (B), we present the pdf formulas and some properties about the distributions used in this section, as well as the detailed step to obtain the full conditional posterior densities. In (3.16), S, R abd A are covariance matrices, but V^{-1} and C^{-1} are precision matrices. Conditioning on data and all other remaining parameters, we have that the full conditional distributions of the main parameters are

$$\begin{aligned} \mu_j^* | \cdot &\sim N(\mu_j^*; m_j, T_j) \\ \Sigma_j^{-1^*} | \cdot &\sim W_p(\Sigma_j^{-1^*}; s + n_j, \overline{S}_j) \end{aligned}$$

where,

$$\begin{aligned} T_j^{-1} &= V^{-1} + n_j \Sigma_j^{*^{-1}} \\ m_j &= T_j \left(V^{-1} m + n_j \Sigma_j^{-1^*} \overline{b}_j \right) \\ \overline{S}_j^{-1} &= \delta s S + n_j (\mu_j^* - \overline{b_j})(\mu_j^* - \overline{b_j})'. \\ \overline{b}_j &= n_j^{-1} \sum_{i:I_i=j} B_i \end{aligned}$$

The full conditional distributions of the hyperparameters are,

$$\begin{aligned} m|\cdot &\sim N_p(\hat{a}, \hat{A}) \\ S|\cdot &\sim W_p\left(sk+q, \hat{R}\right) \\ V^{-1}|\cdot &\sim W_p(c+k, \hat{C}) \end{aligned} \tag{3.17}$$

where $\mathrm{tr}(\cdot)$ is the trace function. Defining $\bar{\boldsymbol{\mu}}^* = k^{-1} \sum_{j=1}^k \mu_j^*$,

$$\begin{aligned} \hat{A}^{-1} &= A^{-1} + kV^{-1} \\ \hat{a} &= \hat{A}\left(A^{-1}a + kV^{-1}\bar{\boldsymbol{\mu}}^*\right) \\ \hat{R} &= \left(qR^{-1} + s \sum_{j=1}^k \Sigma_j^{*^{-1}} \right)^{-1} \\ \hat{C} &= \left(cC + k(m - \bar{\boldsymbol{\mu}}^*)(m - \bar{\boldsymbol{\mu}}^*)' \right)^{-1} \end{aligned}$$

Full conditional posterior is achieved for α using an auxiliary variable $\eta \in (0,1)$ such that,

$$\begin{aligned} \eta|\cdot &\sim \mathrm{Beta}(\alpha+1, n) \\ \alpha|\cdot &\sim \pi_1 \Gamma(a_0+k, b_0 - \log(\eta)) + \pi_2 \Gamma(a_0+k-1, b_0 - \log(\eta)) \end{aligned}$$

where,

$$\begin{aligned} \pi_1 &= \frac{a_0+k-1}{a_0+k-1+n(b_0-\log(\eta))} \\ \pi_2 &= 1-\pi_1 \end{aligned}$$

We finish presenting the full conditional posterior formulas for configuration vector I. Let $\mathscr{I}^{(i)} = (I_1, \ldots, I_{n-1}, I_{n+1}, \ldots, I_n)$ denote the configuration vector corresponding to the $k^{(i)}$ distinct element $\theta_j^{i^*}$ in $\theta^{(i)} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$. Then, for $j = 1, \ldots, k_i$,

$$\begin{aligned} q_{ij}|\cdot &\propto n_{ij} N(\boldsymbol{\beta_i}; \mu_j^{i^*}, \Sigma_j^{i^*}) \\ q_{i0}|\cdot &\propto \alpha N(\boldsymbol{\beta_i}; m, \Sigma_{n+1}^* + V) \end{aligned}$$

where Σ^*_{n+1} is a new Wishart sample from the baseline measure G_0. The following algorithm summarizes simulation from posterior (3.16) using the full conditional distributions. Due to the high dimension of the parameters, the initial values fixed to the set of hyperparameters (used to start the chain) were obtained using the data. For example, the initial values of covariance matrices S, R and A were set to be the unbiased estimator of the covariance matrix of the data.

Algorithm 2 – POSTERIOR SIMULATION OF DPM

Input: hyperparameters $a, A, S, R, q, c, C, a_0, b_0$;
data; sample size n;
Output: samples of (μ, Σ)

Initialize the labels with $\mathscr{I} = (1, 2, \ldots, n)$

Repeat N times
Simulate m, S, V, δ and α (1)
Simulate $\theta^* = (\mu^*, \Sigma^*)$(2) using (1)
Simulate new labels $\mathscr{I}$(3) using (2)

3.4 One-dimensional distribution of functional data

As discussed in Chapter 2, the representation of functions using orthonormal basis expansion implies that each random process $X(t)$ is represented as a linear combination of fixed basis functions and a random set of Fourier coefficients (weights),

$$\begin{aligned} X(t) &= \beta_0\phi_0(t) + \beta_1\phi_1(t) + \cdots + \beta_p\phi_p(t) \\ &= \boldsymbol{c_p(t)}^T\boldsymbol{\beta} \end{aligned}$$

where $\boldsymbol{c_p(t)} = \{\phi_0(t), \ldots, \phi_p(t)\}$ is a vector of fixed quantities and $\boldsymbol{\beta} = \{\beta_0, \ldots, \beta_p\}$ is the random component we are interested to model. From DPM model (3.8), we concluded that,

$$\boldsymbol{\beta_i}|G \overset{iid}{\sim} \sum_{h=1}^{\infty} \omega_h N_{p+1}\left(\boldsymbol{\beta_i}|\boldsymbol{\mu_h}, \boldsymbol{\Sigma_h}\right)$$

that is, the distribution of the set of Fourier coefficients is a mixture of multivariate densities.

The following results make it possible to derive the distribution of $X(t)$ for a fixed $t \in [a, b]$.

Result 3.1. The following results are valid for multivariate Normal distributions.

a) If $\boldsymbol{Y}$ is distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then any linear combination of variables $\boldsymbol{a^T Y} = a_1Y_1 + a_2Y_2 + \cdots + a_kY_k$ is distributed as $N_1(\boldsymbol{a^T\mu}, \boldsymbol{a^T\Sigma a})$.

b) If $\boldsymbol{Y}$ is distributed as infinite discrete mixture of multivariate Normal densities, that is, $\boldsymbol{Y}$ has p.d.f given by

$$f_Y(y) = \sum_{h=1}^{\infty} \omega_h \phi_p(y|\mu_h, \Sigma_h),$$

where $\phi_p(\cdot|\mu,\Sigma)$ denotes the p-variate Normal density with mean vector μ and variance-covariance matrix Σ, and the weights $(\omega_h)_{h\geq 1}$ satisfy

(i) $\omega_h > 0$ for all $h \geq 1$,

(ii) $\sum_{h\geq 1} \omega_h = 1$,

then $X = \boldsymbol{a^T Y} = a_1Y_1 + a_2Y_2 + \cdots + a_kY_k$ has p.d.f given by

$$f_X(x) = \sum_{h=1}^{\infty} \omega_h \phi_1(x|\boldsymbol{a^T \mu_h}, \boldsymbol{a^T \Sigma_h a}). \tag{3.18}$$

Proof. a) This is a standard result about the multivariate Normal distribution and can be found, for instance, in (JOHNSON; WICHERN, 2002).

b) Since $\boldsymbol{Y}$ is distributed as infinite discrete mixture of multivariate Normal densities, its distribution can be derived from a two-step procedure:

1. Draw a discrete random variable $H = h$ from the p.m.f given by

$$p_H(h) = \sum_{i=1}^{\infty} \omega_i I_{\{h\}}(i), \qquad h = 1,2,\ldots;$$

2. Draw $Y = y$ from the $\phi_p(\cdot|\mu_h,\Sigma_h)$.

Since $X = \boldsymbol{a^T Y}$, we can conclude from the previous procedure and from part a) that X is obtained from the two-step procedure:

1. Draw a discrete random variable $H = h$ from the p.m.f given by

$$p_H(h) = \sum_{i=1}^{\infty} \omega_i I_{\{h\}}(i), \qquad h = 1,2,\ldots;$$

2. Draw $X = x$ from the $\phi_1(\cdot|\boldsymbol{a^T \mu_h}, \boldsymbol{a^T \Sigma_h a})$.

Thus, from this procedure it follows that X has the p.d.f given in (3.18). □

Using b) of Result 3.1, it follows that,

$$X(t)|G \sim \sum_{h=1}^{\infty} \omega_h N_1\left(x, \boldsymbol{c_p(t)^T \mu_h}, \boldsymbol{c_p(t)^T \Sigma_h c_p(t)}\right) \tag{3.19}$$

represents the analytical distribution of the process $X(t), t \in [a,b]$.

3.5 Illustration using simulated data

In this section we present some illustrations of multivariate density estimation and prediction using Dirichlet Process Mixtures with normal kernels. For that, it was simulated two groups of functional data $\{X_1(t),\ldots,X_{n_1}(t)\}$ and $\{Y_1(t),\ldots,Y_{n_2}(t)\}$ with $n_1 = n_2 = 50$ and,

$$X_i(t) = \sum_{j=1}^{p} \beta_{j_{i_X}} \phi_j(t) \qquad Y_i(t) = \sum_{j=1}^{p} \beta_{j_{i_Y}} \phi_j(t) \qquad i = 1,\ldots,50$$

where $p = 5$ and $\{\phi_0,\ldots,\phi_p\}$ is the set of Fourier orthonormal basis functions defined as in (2.3). It was assumed that,

$$\begin{aligned} \boldsymbol{\beta_{i_X}} = \left(\beta_{0_{i_X}}, \beta_{1_{i_X}}, \ldots, \beta_{p_{i_X}}\right) &\sim N_{p+1}(\mu_X, \Sigma_X) \\ \boldsymbol{\beta_{i_Y}} = \left(\beta_{0_{i_Y}}, \beta_{1_{i_Y}}, \ldots, \beta_{p_{i_Y}}\right) &\sim N_{p+1}(\mu_Y, \Sigma_Y) \end{aligned}$$

and

$$\mu_X = \begin{bmatrix} 20 \\ 3 \\ -1 \\ -1 \\ -0.5 \end{bmatrix} \qquad \Sigma_X = \begin{bmatrix} 0.45 & 0.10 & -0.19 & 0.11 & 0.22 \\ 0.10 & 0.15 & -0.21 & 0.12 & -0.12 \\ -0.19 & -0.21 & 0.72 & -0.37 & -0.07 \\ 0.11 & 0.12 & -0.37 & 0.32 & 0.17 \\ 0.22 & -0.12 & -0.07 & 0.17 & 0.66 \end{bmatrix}$$

$$\mu_Y = \begin{bmatrix} 15.4 \\ 3.4 \\ -1.4 \\ -3.4 \\ -0.4 \end{bmatrix} \qquad \Sigma_Y = \begin{bmatrix} 1.38 & 0.62 & -0.41 & -0.75 & 0.39 \\ 0.62 & 0.46 & -0.24 & -0.20 & 0.17 \\ -0.41 & -0.24 & 0.55 & 0.17 & -0.04 \\ -0.75 & -0.20 & 0.17 & 0.55 & -0.31 \\ 0.39 & 0.17 & -0.04 & -0.31 & 0.58 \end{bmatrix}$$

Figure 7 shows the simulated samples of functional data for each group.

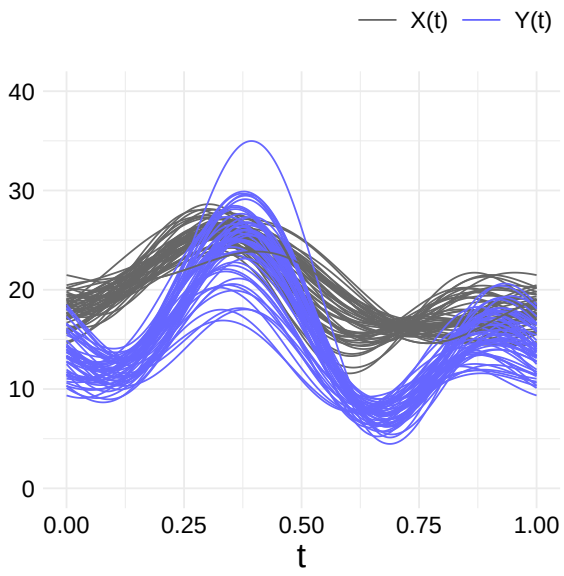


Figure 7 – Functional data simulated using multivariate normal distribution with distinct parameters.

The DPM model was applied to estimate the multivariate distribution of the coefficients of each group. It was simulated $N_0 = 51000$ predicted samples of Fourier coefficients and then applied a *burn* of 1000 from the beginning of the chain (time until convergence) and a *jump* of 4 due to minimize correlation between the predictions. Figure 8 presents the predicted average curve for both groups, as well as the predictive confidence interval with 95% of credibility. Figures 10 and 11 present the projection of raw data and predicted samples from each two-by-two dimensions, for group X and Y, respectively.

The results show that the DPM model using normal kernels is presenting great performance on predicting high dimensional data, even with relatively small sample sizes. In both simulations, we provided initials values for the hyperparameters of (3.15) based on Empirical Bayes methods, due to the fact that as data dimension increases, the more difficult is the chain to converge. The total time spent for each group simulation was about 30 minutes on a device *11th Gen Intel(R) Core(TM) i7-1165G7 16Gb RAM*.

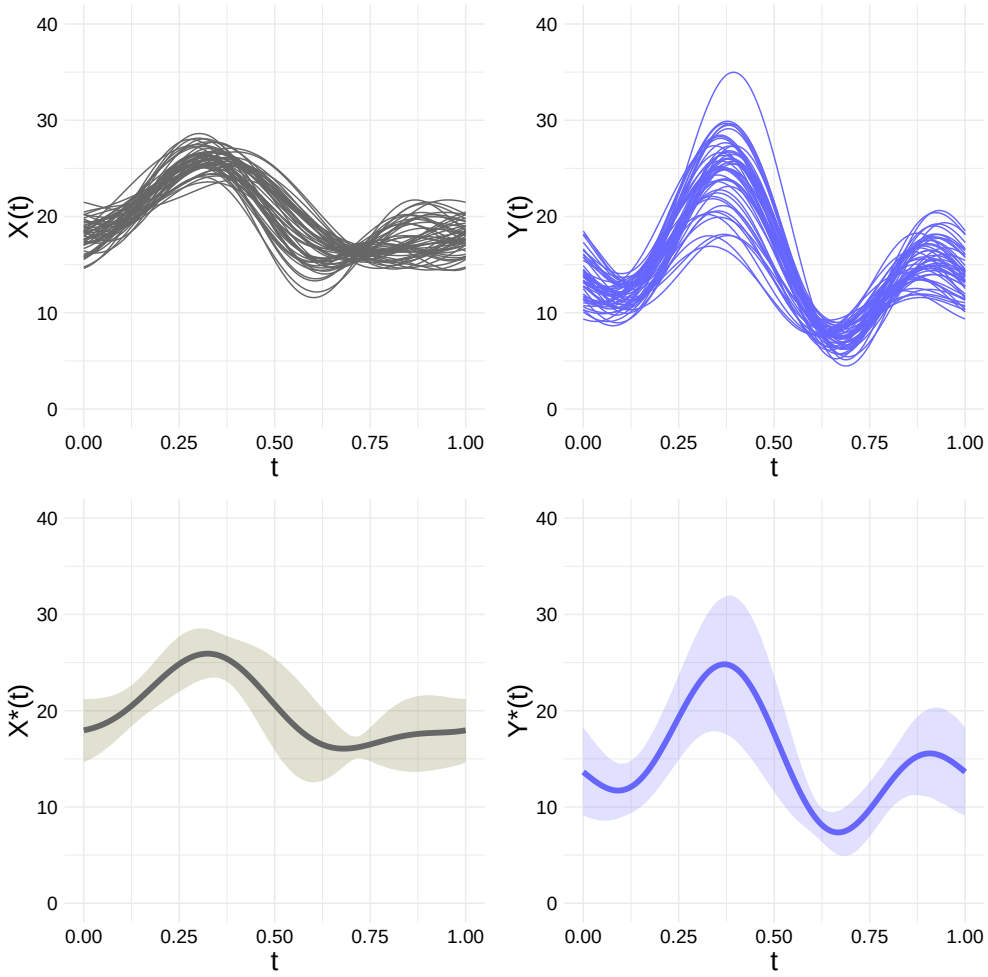


Figure 8 – Comparison between functional data and predicted mean for each group. The shaded areas represent predictive confidence interval with 95% of credibility.

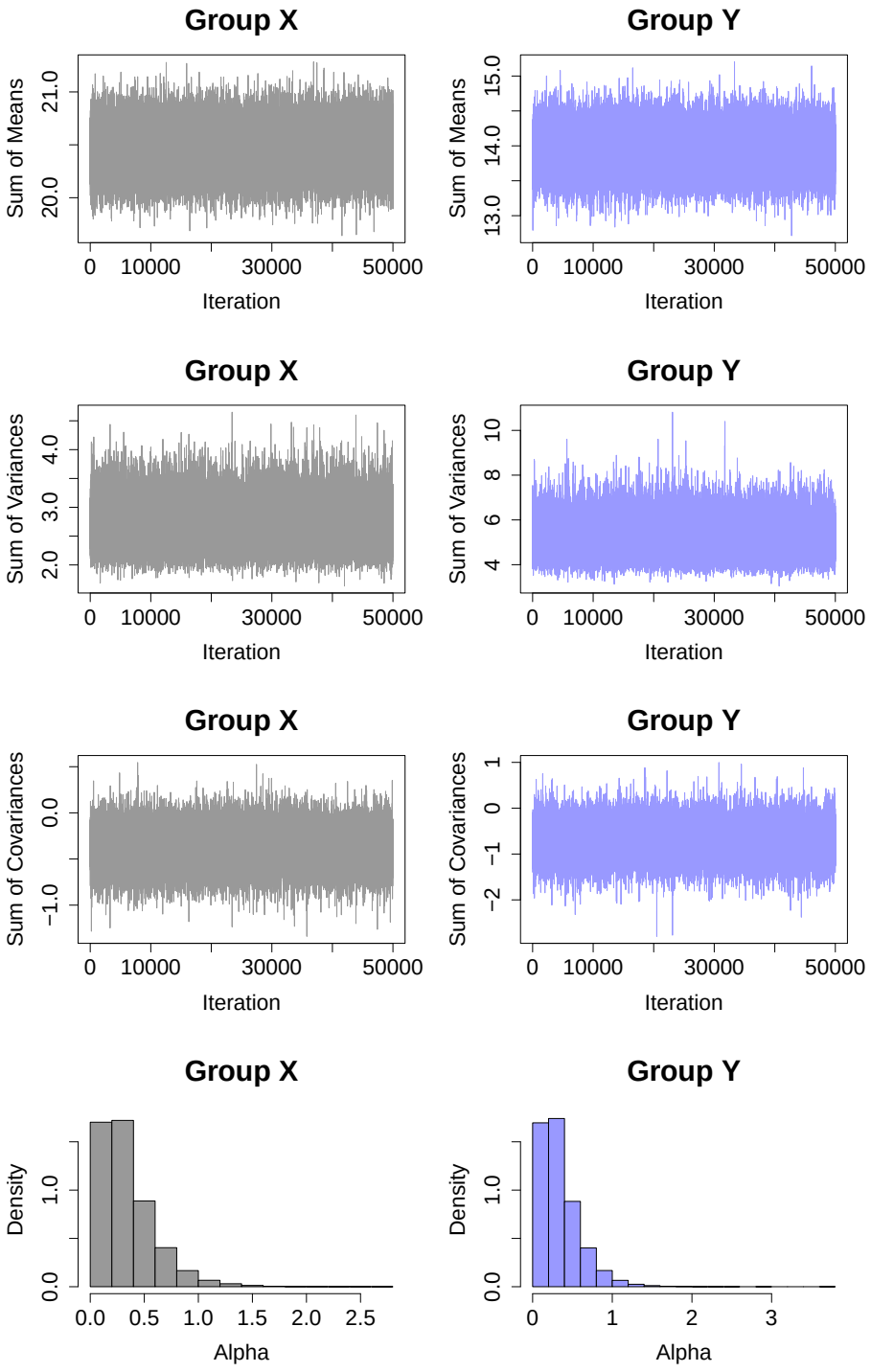


Figure 9 – Convergence diagnostics of predictive samples using MCMC for simulated data.

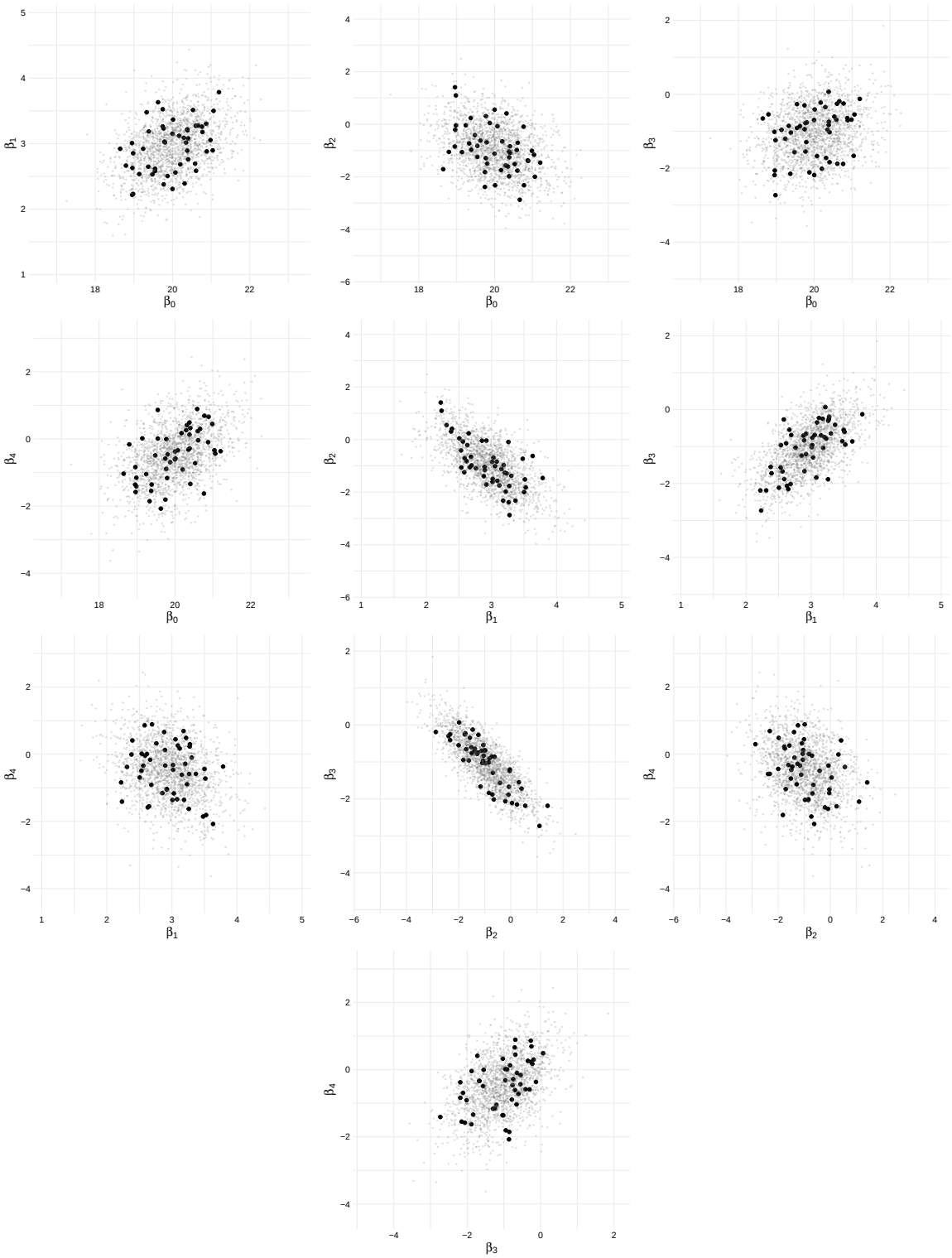


Figure 10 – Comparison between simulated data (large points) and predicted samples (small points) for each pair of coefficients on group X.

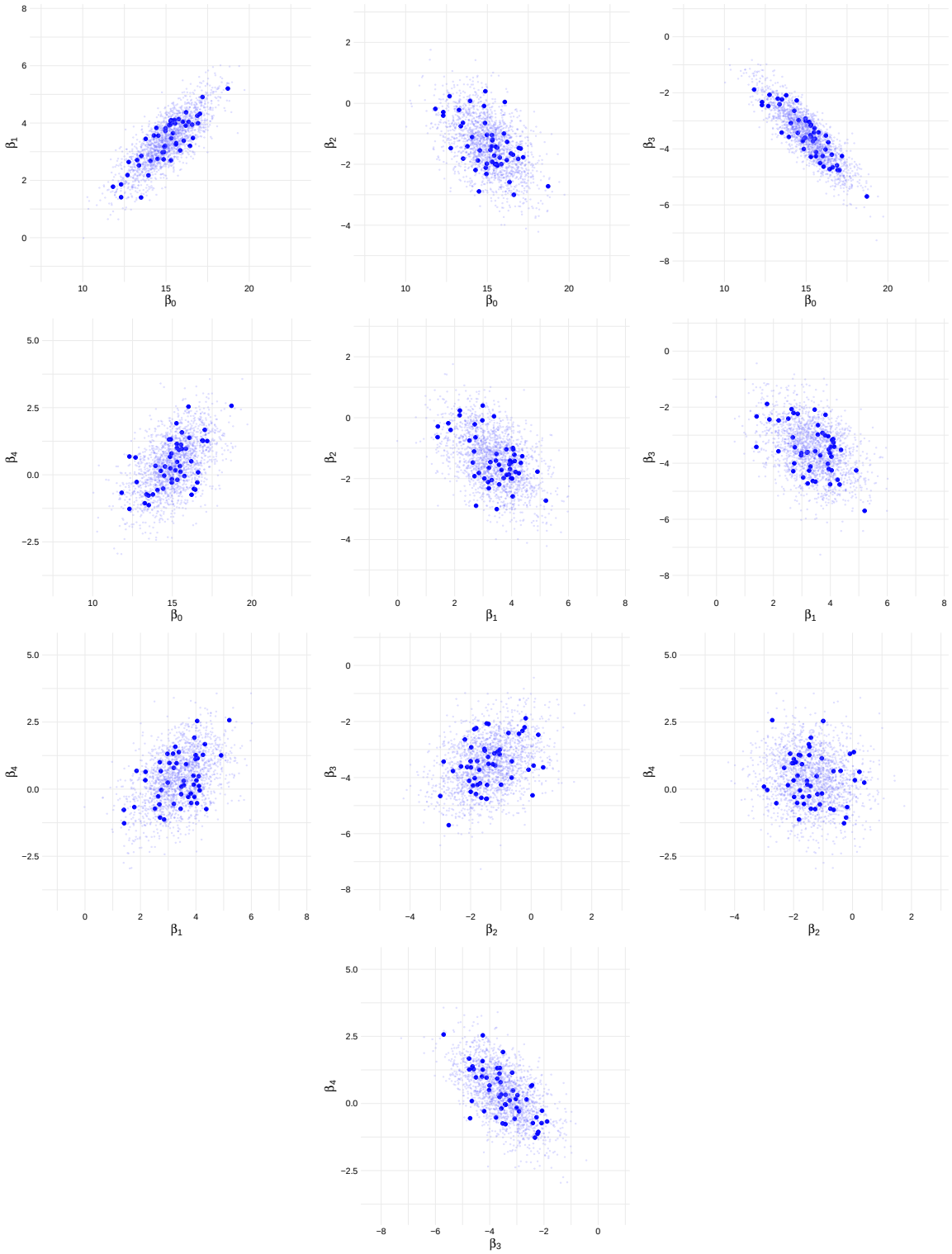


Figure 11 – Comparison between simulated data (large points) and predicted samples (small points) for each pair of coefficients on group Y.

CHAPTER

4

COMPARISON OF TWO INDEPENDENT FUNCTIONAL DATA GROUPS

In this chapter, we present methods to compare two independent groups of functional data based on two different notions of closeness: (1) assessment of the dissimilarity between the groups and (2) evaluation of the homogeneity between the generating mechanisms of the curves of each group. For the first goal, we proposed a **Predictive Dissimilarity Index** that measures the distance between predicted curves of each group and has a strong interpretative appeal. For the second goal, we used the result discussed on Section 3.4 to test the following hypotheses:

$$H_{0,t} : X_t \overset{\mathscr{D}}{=} Y_t, \text{ for } t \in [a,b]$$

where $\overset{\mathscr{D}}{=}$ denotes "equals in distribution" and $[a,b] \subseteq [0,1]$.

4.1 Predictive Dissimilarity Index

Consider two independent stochastic processes $X = \{X(t) : t \in [0,1]\}$, $Y = \{Y(t) : t \in [0,1]\}$ inside a fixed sub-interval $[a,b] \subseteq [0,1]$ of the domain. Consider a metric d between two real functions $\tilde{X}, \tilde{Y} \in L^2([a,b])$. The dissimilarity between the processes can be assessed inside $[a,b]$ using the following measure

$$D_\varepsilon(X,Y;[a,b]) = \mathbb{P}\big(d(\tilde{X},\tilde{Y}) > \varepsilon\big)$$

where $\tilde{X}, \tilde{Y}$ are the restrictions of X and Y to the sub-interval $[a,b]$, and ε is fixed positive value representing a practical significance level for the difference between the two measurements. Hence, the dissimilarity $D_\varepsilon(X,Y;[a,b])$ is the probability of observing a significant difference between X and Y inside the interval $[a,b]$. The choice of the ε value is a subjective one and should take into account the practical consequences associated to the problem at hand.

In case this choice is not direct, the consideration of the following dissimilarity index can be helpful

$$\begin{aligned} PDI_\alpha(X,Y;[a,b]) &= \sup\Big\{\varepsilon > 0 : D_\varepsilon(X,Y;[a,b]) \geq \alpha\Big\} \\ &= \sup\Big\{\varepsilon > 0 : \mathbb{P}(d(\tilde{X},\tilde{Y}) > \varepsilon) \geq \alpha\Big\}, \end{aligned} \tag{4.1}$$

where $\alpha \in [0,1]$ is a large probability indicating the degree of confidence desired. The index $PDI_\alpha(X,Y;[a,b])$ represents the largest possible difference ε that we can state, with probability α, that a practical significance difference exists. In other words, the *PDI* index is the $(1-\alpha)$ quantile of the distribution of $d(\tilde{X},\tilde{Y})$.

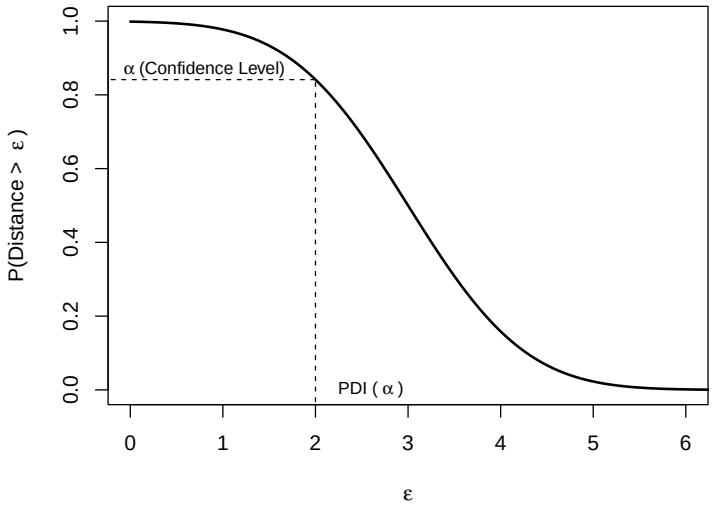


Figure 12 – Illustration of PDI.

The greater the value of $PDI_\alpha(X,Y;[a,b])$, the greater is the difference between the observed values of X and Y inside the sub-interval $[a,b]$. For instance, if $PDI_{0.90}(X,Y;[a,b]) = 5$, than there is a probability greater than 0.9 that the processes are 5 units apart. If one considers that a difference up to $\varepsilon = 1$ is not relevant in practice, then he can conclude that a practical difference exists, since the probability of $[d(\tilde{X},\tilde{Y}) > \varepsilon]$ is greater than 0.9.

We used two possibilities for the metric d in $L^2([a,b])$:

$$\begin{aligned} d_1(\tilde{X},\tilde{Y}) &= \sup_{t\in[a,b]} |\tilde{X}(t) - \tilde{Y}(t)|, \\ d_2(\tilde{X},\tilde{Y}) &= \frac{1}{b-a}\int_a^b |\tilde{X}(t) - \tilde{Y}(t)|dt. \end{aligned} \tag{4.2}$$

The metric d_1 is the largest absolute difference between $\tilde{X}$ and $\tilde{Y}$ and the metric d_2 is the mean value of the absolute difference between $\tilde{X}$ and $\tilde{Y}$, known as L_1 norm. Considering these two

metrics, we define

$$PDI_1(\alpha) = \sup\Big\{\varepsilon > 0 : \mathbb{P}(d_1(\tilde{X},\tilde{Y}) > \varepsilon) \geq \alpha\Big\}, \tag{4.3}$$

$$PDI_2(\alpha) = \sup\Big\{\varepsilon > 0 : \mathbb{P}(d_2(\tilde{X},\tilde{Y}) > \varepsilon) \geq \alpha\Big\}. \tag{4.4}$$

Estimates of *PDI* is achieved as following. Assuming two independent observed samples of functional data $\{X_1(t),\ldots,X_{n_1}(t)\}$ and $\{Y_1(t),\ldots,Y_{n_2}(t)\}$ and its respective Fourier coefficients $\boldsymbol{B_X} = \{\boldsymbol{\beta_{1X}},\ldots,\boldsymbol{\beta_{n_1X}}\}$ and $\boldsymbol{B_Y} = \{\boldsymbol{\beta_{1Y}},\ldots,\boldsymbol{\beta_{n_2Y}}\}$, where $\boldsymbol{\beta_{i\cdot}} = (\beta_{0_{i\cdot}},\ldots,\beta_{p_{i\cdot}})$, with p being the number of basis (*cutoff*) used in the smoothing step, we can evaluate *PDI* using predictive samples $\{X_1^*(t),X_2^*(t),\ldots,X_N^*(t)\}$ and $\{Y_1^*(t),Y_2^*(t),\ldots,Y_N^*(t)\}$ obtained adjusting DPM models.

In summary, we can use the observed samples of coefficients to estimate the probability density and predictive distributions of each group, which enables us to get predictive samples $\left\{\boldsymbol{\beta_{(n_1+1)X}},\ldots,\boldsymbol{\beta_{(n_1+N)X}}\right\}$ and $\left\{\boldsymbol{\beta_{(n_2+1)Y}},\ldots,\boldsymbol{\beta_{(n_2+N)Y}}\right\}$, where N is some arbitrary number of predictive samples desired. Through these new samples of fourier coefficients, we can construct new curves that is expected to have the same probabilistic behavior as the original functions, which allows us to evaluate *PDI* in $[a,b]$, according to some confidence level α.

Algorithm 3 – PREDICTIVE DISSIMILARITY INDEX

Input: predictive samples of Fourier coefficients for group 1 and 2; confidence level α, distance metric, limits $[a,b]$ of domain
Output: predicitive dissimilarity index $PDI(\alpha)$

For each $t \in [a,b]$, do
  Repeat N times
    Evaluate $X^*(t)$ and $Y^*(t)$ using the predictive samples of Fourier coefficients
    Calculate $d(X^*(t),Y^*(t))$ using the input distance metric
  Evaluate the quantile $(1-\alpha)$ of the $Nestimateddistances$

4.2 Hypothesis Test for Distribution Equality

Here we discuss the second approach to evaluate the closeness of two functional data groups: hypotheses test. Assuming two independent stochastic processes $X = \{X(t) : t \in [0,1]\}$, $Y = \{Y(t) : t \in [0,1]\}$ inside a fixed sub-interval $[a,b] \subset [0,1]$ of the domain, the hypothesis of interest is,

$$\begin{aligned} H_{0,t} &: X_t \overset{\mathscr{D}}{=} Y_t, \\ H_{1,t} &: X_t \overset{\mathscr{D}}{\neq} Y_t, \end{aligned} \tag{4.5}$$

for a fixed $t \in [a,b]$, where $\overset{\mathscr{D}}{=}$ means "equals in distribution". In words, H_0 suggests that the processes are generated according to the same probabilistic mechanism over the period $[a,b]$.

To evaluate the plausibility of H_0 using Bayesian principles, we might calculate the posterior probability of H_0, that is

$$P(H_{0,t}|\mathscr{B}) \quad = \quad P(F_{X_t} = F_{Y_t}|\mathscr{B}) \tag{4.6}$$

where $F_{X(t)}$ and $F_{Y(t)}$ are the distribution functions of X and Y at time t, respectively, and $\mathscr{B}$ represents the data. Since the processes are continuous, the value of (4.6) is always zero. For practical purposes, it is reasonable to substitute the original hypothesis of equality with an enlarged one that means the same. (COSCRATO *et al.*, 2019) calls this enlargement as **Pragmatic Hypothesis** . Essentially, a pragmatic hypothesis is *an imprecise hypothesis that is sufficiently good from the practical purpose of an end-user of the theories, using an appropriate precision level* (ESTEVES *et al.*, 2019).

- **Definition (pragmatic hypothesis for a singleton):** let $H_0 : \theta = \theta_0$, $d_{\mathbf{Z}}$ be a predictive dissimilarity function and $\gamma > 0$. The pragmatic hypothesis for H_0, denoted by $Pg(\{\theta_0\}, d_{\mathbf{Z}}, \gamma)$ is,

$$Pg(\{\theta_0\}, d_{\mathbf{Z}}, \gamma) = \{\theta^* \in \Theta : d_{\mathbf{Z}}(\theta_0, \theta^*) \leq \gamma\} \tag{4.7}$$

Adapting definition (4.7) to the hypotheses postulated in (4.5), we have that,

$$Pg(H_{0,t}, d_{\mathbf{Z}}, \gamma) = \left\{ \left(F^*_{X_t}, F^*_{Y_t}\right) : d_{\mathbf{Z}} \left(F^*_{X_t}, F^*_{Y_t}\right) \leq \gamma \right\} \tag{4.8}$$

which means that using an appropriate precision level $0 \leq \gamma \leq 1$ and a convenient dissimilarity measure between distributions $d_{\mathbf{Z}}$, we can calculate the probability of the pragmatic version of the hypothesis of interest. Here we used the Kolmogorv-Smirnov (KS) distance as a candidate to $d_{\mathbf{Z}}$, that is,

$$d_{\mathbf{Z}} \left(F_{X_t}, F_{Y_t}\right) = \sup_{x \in \mathbb{R}} |F_{X_t}(x) - F_{Y_t}(x)|$$

For the choice of the precision level γ, we suggests simulation studies. An alternative way is given by (SWARTZ, 1999), that suggests to consider an initial measurement precision γ_0 of the variable of interest whose difference has little impact in decisions in practice (this answer can be obtained using prior knowledge of the researcher). For instance, if the variable is measured in feet and $\gamma_0 = 0.5$, then we are stating that the values are rounded to the nearest foot and, in practice, this rounding doesn't cause significant impact in decisions. Having specified γ_0, we might evaluate the impact of this rounding over the distribution of the data. It is important to highlight that the choice of this distribution might not consider the observed data, because the definition of the scientific hypotheses is a previous step.

Hence, the researcher can provide an idea of scale σ of the variable of interest and measure the impact of γ_0 using the KS distance between the original distribution and the

distribution offset by γ_0. The distribution M might be in a location-scale family (Normal, Logistic, etc), but the procedure is invariant by location. Then, a candidate for the precision level γ is given by

$$\gamma^* = \sup_{x\in\mathbb{R}} \left| M\left(x + \frac{\gamma_0}{\sigma}\right) - M(x) \right| \tag{4.9}$$

Hence, with the precision level γ^* at hand, we can rewrite the pragmatic hypotheses of interest as following:

$$H_{0,t}: \quad X_t \overset{\mathscr{D}}{=} Y_t, \text{ for } t \in [a,b] \qquad \Longrightarrow \qquad Pg(H_{0,t};\gamma^*): \quad \sup_{x\in\mathbb{R}} |F_{X_t}(x) - F_{Y_t}(x)| \leq \gamma^*$$

where F_{X_t} and F_{Y_t} are the distribution of the process X and Y at time t, respectively.

Algorithm 4 – PROBABILITY OF PRAGMATIC NULL HYPOTHESIS

Input: posterior simulations of (μ,Σ); precision level γ^*
Output: probability of pragmatic hypothesis $\mathbb{P}(P(H_{0,t};\gamma^*)|\mathscr{B})$
For each $t \in [a,b]$, do
Repeat N times
Estimate $\hat{F}_{X_t}$ and $\hat{F}_{Y_t}$ using the posterior simulations of (μ,Σ)
Calculate $d_{KS}(\hat{F}_{X_t},\hat{F}_{Y_t})$
Estimate $P(H_{0,t}) = (1/n)\sum_{i=1}^{N} I(d_{KS}(\hat{F}_{X_t},\hat{F}_{Y_t}) \leq \gamma^*)$

The decision criteria used to decide about the acceptance of the pragmatic null hypothesis at $t \in [a,b]$ is,

$$\mathbb{P}\left(P(H_{0,t},\gamma^*)|\mathscr{B}\right) > \frac{1}{2} \tag{4.10}$$

4.3 Illustration using simulated data

In this section, we evaluate the PDI index and KS distance over time using the same simulations of Section (3.5). Figure 13 exhibits the results when the groups has the same generating mechanism (H_0 is true) and when they has distinct generating mechanisms (H_0 is false).

The results show that according to PDI_1, *there is a confidence level greater than* 95% *that the maximum difference between the curves is greater than 2 units (over* $[0,1]$*) when the groups are equally distributed and greater than 8 when the groups are not equally distributed*. Similarly, according to PDI_2, there is a probability greater than 95% that the mean difference of the curves is about 1 when the groups are equally distributed and about 4 when the groups are not equally distributed (over the same domain).

Note that when the groups doesn't have the same distribution over $[0,1]$, PDI index shows a high confidence level that the dissimilarity between the curves is relatively "far" from

zero, with doesn't happen when the groups have the same distribution, indicating that this index has a good potential to detect dissimilarities on the processes over any subset $[a,b] \subseteq [0,1]$ and it is easy to interpret.

The KS average kept very close to the real KS when the groups are not equally distributed. However, when H_0 is true, the KS average kept constant and close to 0.1, which can be a first guess of the precision level γ used on pragmatic hypotheses discussed on Section 4.2. A more extended simulation study is recommended to evaluate the behavior of this threshold varying the sample sizes and distributions of the groups.

Another illustration was made using polynomial basis instead of Fourier basis (Figure 14), because the periodic outline of Fourier basis does not always fit well to some studies, making it desirable that the methods performs well for other sets of basis functions, like polynomials, splines, etc. In this scheme, we have processes $\{P_1(t),\ldots,P_{n_1}(t)\}$ and $\{Q_1(t),\ldots,Q_{n_2}(t)\}$ with $n_1 = n_2 = 50$ and,

$$P_i(t) = \sum_{j=1}^{p} \tilde{\beta}_{j_{i_X}} \psi_j(t) \qquad Q_i(t) = \sum_{j=1}^{p} \tilde{\beta}_{j_{i_Y}} \psi_j(t) \qquad i = 1,\ldots,50$$

where $p = 5$, $\{\psi_0,\ldots,\psi_p\}$ is the set of polynomial basis functions, that is, $\psi_i(t) = t^i$, $\tilde{\boldsymbol{\beta}}_{\boldsymbol{i_P}} \sim N_{p+1}(\mu_P,\Sigma_P)$ and $\tilde{\boldsymbol{\beta}}_{\boldsymbol{i_Q}} \sim N_{p+1}(\mu_Q,\Sigma_Q)$, with

$$\mu_P = \begin{bmatrix} 4.7 \\ 10.4 \\ 5.1 \\ 2.7 \\ -0.3 \end{bmatrix} \qquad \Sigma_P = \begin{bmatrix} 1.62 & 1.46 & 1.23 & 0.48 & -0.04 \\ 1.46 & 2.42 & 1.61 & -0.01 & 0.59 \\ 1.23 & 1.61 & 2.10 & -0.60 & 1.21 \\ 0.48 & -0.01 & -0.60 & 2.66 & -0.99 \\ -0.04 & 0.59 & 1.21 & -0.99 & 1.36 \end{bmatrix}$$

$$\mu_Q = \begin{bmatrix} 12.0 \\ 8.9 \\ 1.2 \\ 10.3 \\ 9.5 \end{bmatrix} \qquad \Sigma_Q = \begin{bmatrix} 0.65 & -0.26 & 1.04 & -0.81 & 1.51 \\ -0.26 & 2.08 & 1.18 & 0.61 & 0.04 \\ 1.04 & 1.18 & 3.81 & -1.55 & 3.32 \\ -0.81 & 0.61 & -1.55 & 1.81 & -1.86 \\ 1.51 & 0.04 & 3.32 & -1.86 & 5.73 \end{bmatrix}$$

The results show that the PDI index is performing well on estimate the expected and maximum difference between the functional data smoothed using polynomial basis function. The averaged KS Distance is closed to the real, but it seems to be relatively overestimated for some regions of the domain, even when H_0 is false.

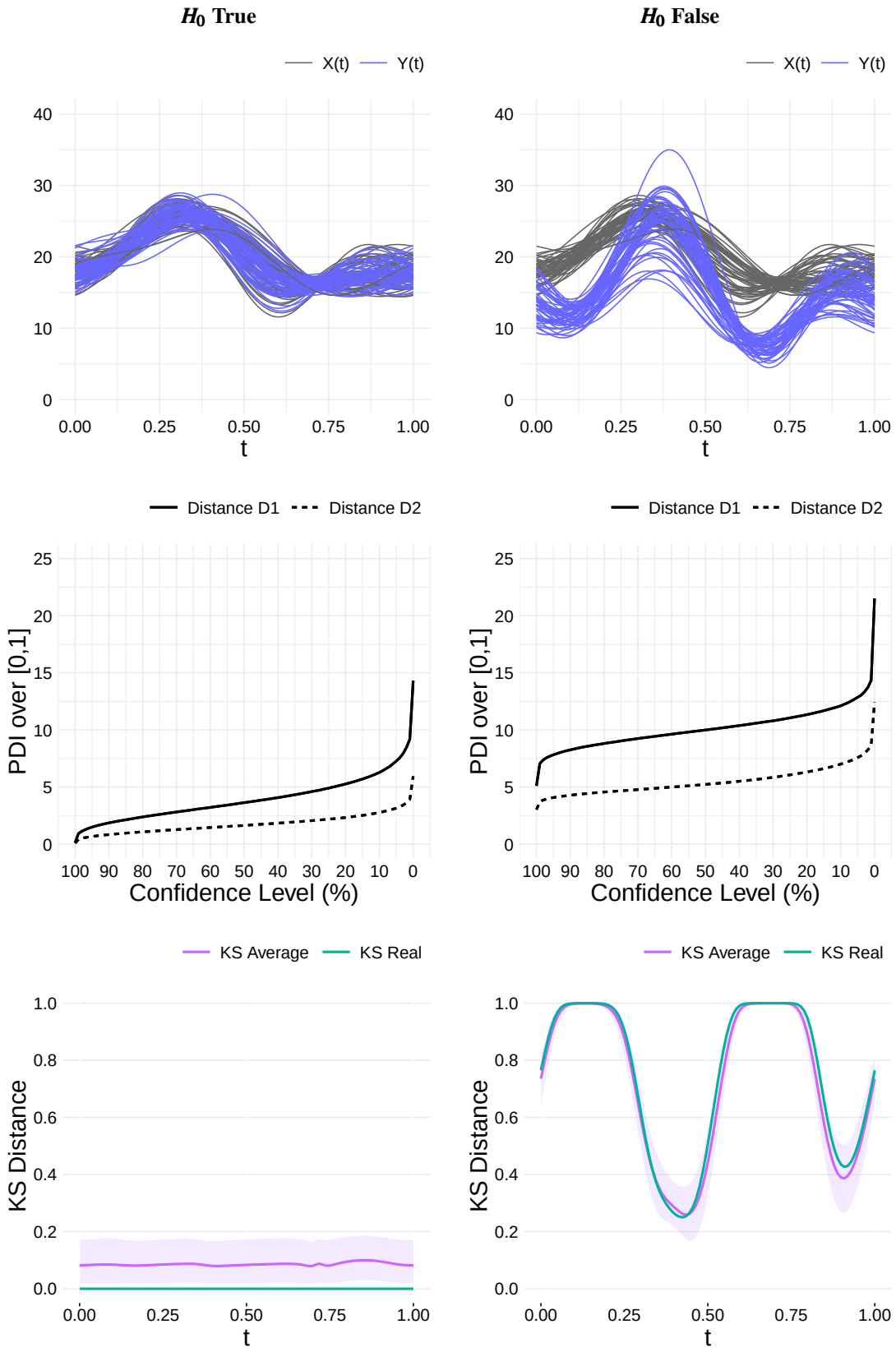


Figure 13 – Comparison of functional data groups using simulated data. On left, the PDI curve using both d_1 and d_2 distances and KS average over time when H_0 is true. On right, the same information when H_0 is false. The shaded pink areas corresponds to pointwise predictive bands with 90% of credibility.

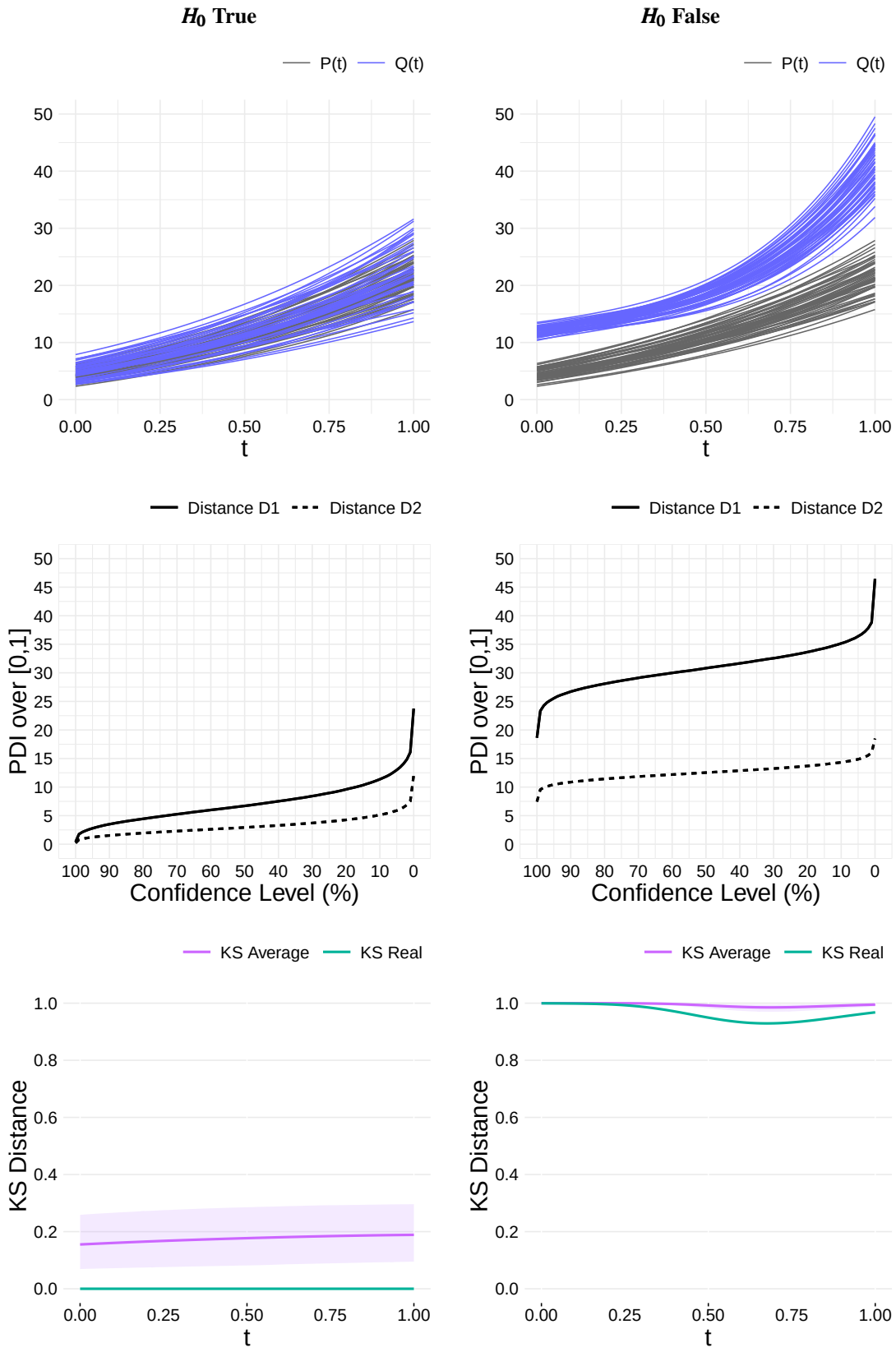


Figure 14 – Comparison of functional data groups using simulated data adjusted using polynomial basis. On left, the PDI curve using both d_1 and d_2 distances and KS average over time when H_0 is true. On right, the same information when H_0 is false. The shaded pink areas corresponds to a pointwise predictive bands with 90% of credibility.

CHAPTER

5

APPLICATION: CANADIAN WEATHER DATA

In this section, we present an application of the methods for a real functional data set. (RAMSAY; SILVERMAN, 1997) introduce the famous *Canadian Weather Data*, where temperature of 35 provinces across Canada were recorded daily by weather stations and averaged over 34 years (1960 to 1994). The provinces are divided in four geographic climate regions: Atlantic, Continental, Pacific and Arctic. Figure 15 shows the locations of the provinces grouped by region and tables 1 and 2 show some details of the provinces and how data is structured, respectively.

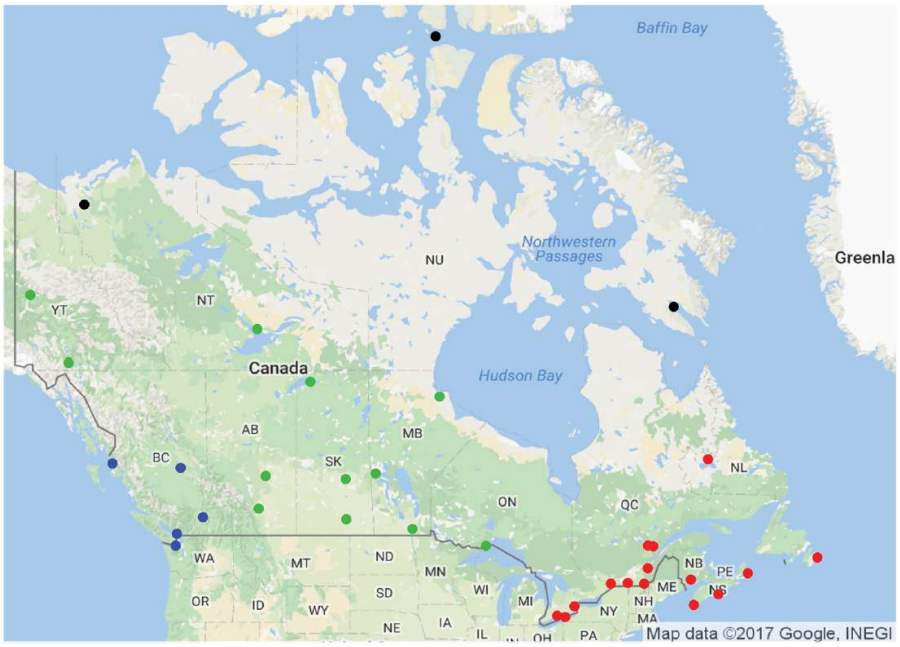


Figure 15 – Location of the weather stations across Canada, grouped by regions: Atlantic (red), Continental (green), Pacific (blue) and Arctic (black). *Source:* (PINI; VANTINI, 2017)

Table 1 – Number of provinces and province names by region.

Region	Number of provinces	Province names	
Atlantic	15	St. Johns Halifax Sydney Yarmouth Charlottvl Fredericton Scheffervll Arvida	Bagottville Quebec Sherbrooke Montreal Ottawa Toronto London
Continental	12	Thunderbay Winnipeg The Pas Churchill Regina Pr. Albert	Uranium Cty Edmonton Calgary Whitehorse Dawson Yellowknife
Pacific	5	Kamloops Vancouver Victoria	Pr. George Pr. Rupert
Arctic	3	Iqaluit Inuvik	Resolute
TOTAL	**35**		

Table 2 – Structure of the Canadian Temperature dataset.

Date	**Averaged Temperature (1961 - 1994)**											
	Atlantic			**Continental**			**Pacific**			**Arctic**		
	Province 1	…	**Province 15**	**Province 1**	…	**Province 12**	**Province 1**	…	**Province 5**	**Province 1**	…	**Province 3**
01/jan	-3.6	…	-5.2	-14.0	…	-24.5	-5.3	…	0.4	-23.3	…	-30.7
02/jan	-3.1	…	-5.7	-14.0	…	-25.3	-5.6	…	0.5	-24.0	…	-30.6
03/jan	-3.4	…	-5.3	-13.5	…	-26.1	-6.5	…	-0.2	-24.4	…	-31.4
04/jan	-4.4	…	-5.8	-13.9	…	-27.7	-6.8	…	-0.6	-24.7	…	-31.9
05/jan	-2.9	…	-6.5	-14.4	…	-27.8	-7.3	…	-1.0	-25.3	…	-31.5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
27/dez	-12.4	…	-5.6	-13.0	…	-26.1	-4.3	…	0.8	-24.1	…	-29.8
28/dez	-10.1	…	-4.1	-12.8	…	-25.4	-6.2	…	0.1	-23.1	…	-30.1
29/dez	-9.1	…	-4.7	-13.8	…	-24.5	-7.0	…	-0.1	-23.5	…	-29.0
30/dez	-11.3	…	-5.7	-14.0	…	-25.8	-6.4	…	-0.1	-23.9	…	-29.4
31/dez	-10.7	…	-4.9	-13.9	…	-25.1	-6.3	…	0.0	-24.5	…	-30.5

The data were smoothed using Fourier orthonormal basis functions of the form (2.3). The number of basis functions were chosen using k-fold cross validation with penalizing rule as presented in (2.6). Table 3 shows the best number of basis for each province as well as their respective estimated MSE. Due to reduce computational efforts, we choose to assume 50 basis functions for all the provinces (the computational costs using $p = 95$ were worse than the gains of MSE involved). Figure 16 shows the raw recorded data and smoothed curves for each region using $p = 50$.

The Fourier coefficients obtained from smoothing were modeled and predicted using Dirichlet Process Mixtures with normal kernels as presented in (3.7). The initial values used on MCMC were based on Empirical Bayes due to the high dimension of the data. For each region, it was simulated $N = 100,000$ samples, with *burn* = 1000 and *jump* = 4 , due to guarantee convergence of the chain and avoid high correlation between the samples. Figure 17 presents some diagnostics that justify the choice of these quantities. Figure 18 shows the predicted average for the functional processes of each region.

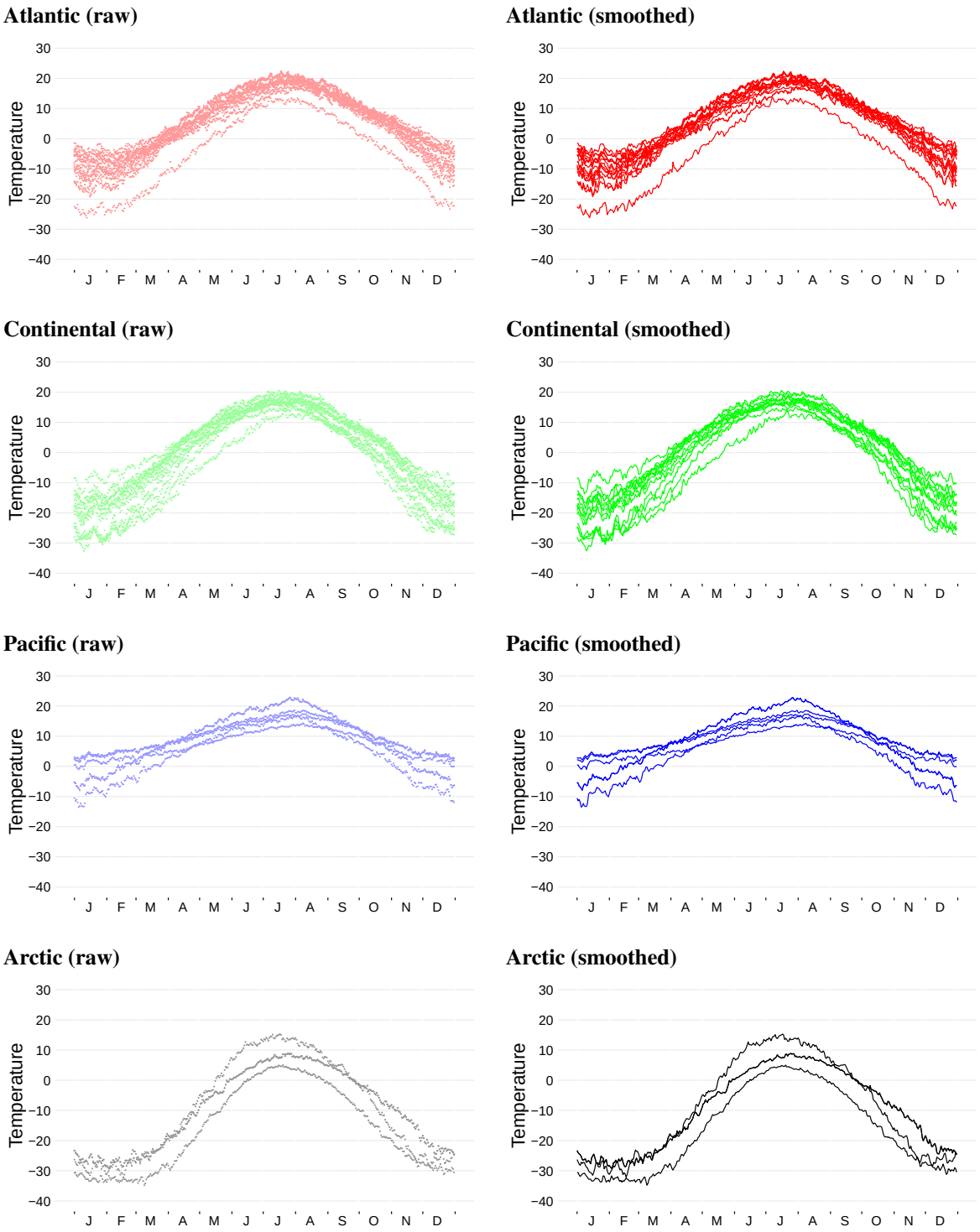


Figure 16 – Raw recorded points (left) and smoothed curves (right) over time for the four regions of Canada. Smoothing was made using $p = 50$ Fourier basis function.

Table 3 – Best number of basis and estimated MSE for each province.

Province	Number of basis with smaller MSE	Estimated MSE	Province	Number of basis with smaller MSE	Estimated MSE
St. Johns	75	0.3650	**Churchill**	75	0.3442
Halifax	55	0.3980	**Regina**	90	0.3832
Sydney	90	0.3419	**Pr. Albert**	65	0.3960
Yarmouth	75	0.3895	**Uranium Cty**	55	0.4066
Charlottvl	75	0.3966	**Edmonton**	95	0.3447
Fredericto	75	0.3958	**Calgary**	90	0.3335
Scheffervl	75	0.3413	**Kamloops**	65	0.3988
Arvida	75	0.3663	**Vancouver**	70	0.3950
Bagottvill	75	0.3799	**Victoria**	90	0.3427
Quebec	80	0.3540	**Pr. George**	65	0.3646
Sherbrooke	50	0.4184	**Pr. Rupert**	75	0.3551
Montreal	70	0.3936	**Whitehorse**	75	0.3928
Ottawa	90	0.3428	**Dawson**	75	0.3263
Toronto	80	0.3857	**Yellowknife**	90	0.3626
London	85	0.3993	**Iqaluit**	90	0.3758
Thunderbay	90	0.3583	**Inuvik**	55	0.4171
Winnipeg	65	0.3986	**Resolute**	80	0.3955
The Pas	75	0.3639			

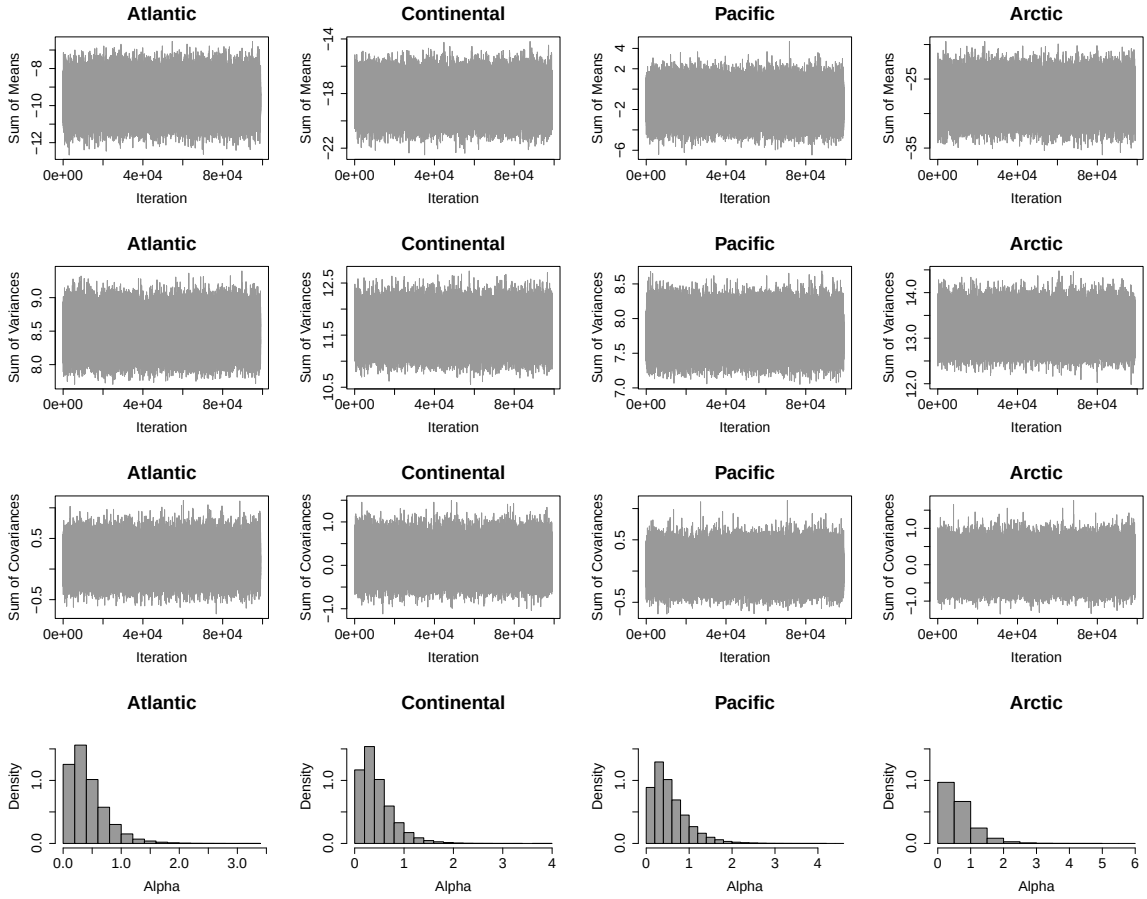


Figure 17 – Convergence diagnostics for predictive samples using MCMC for Canadian Weather Data.

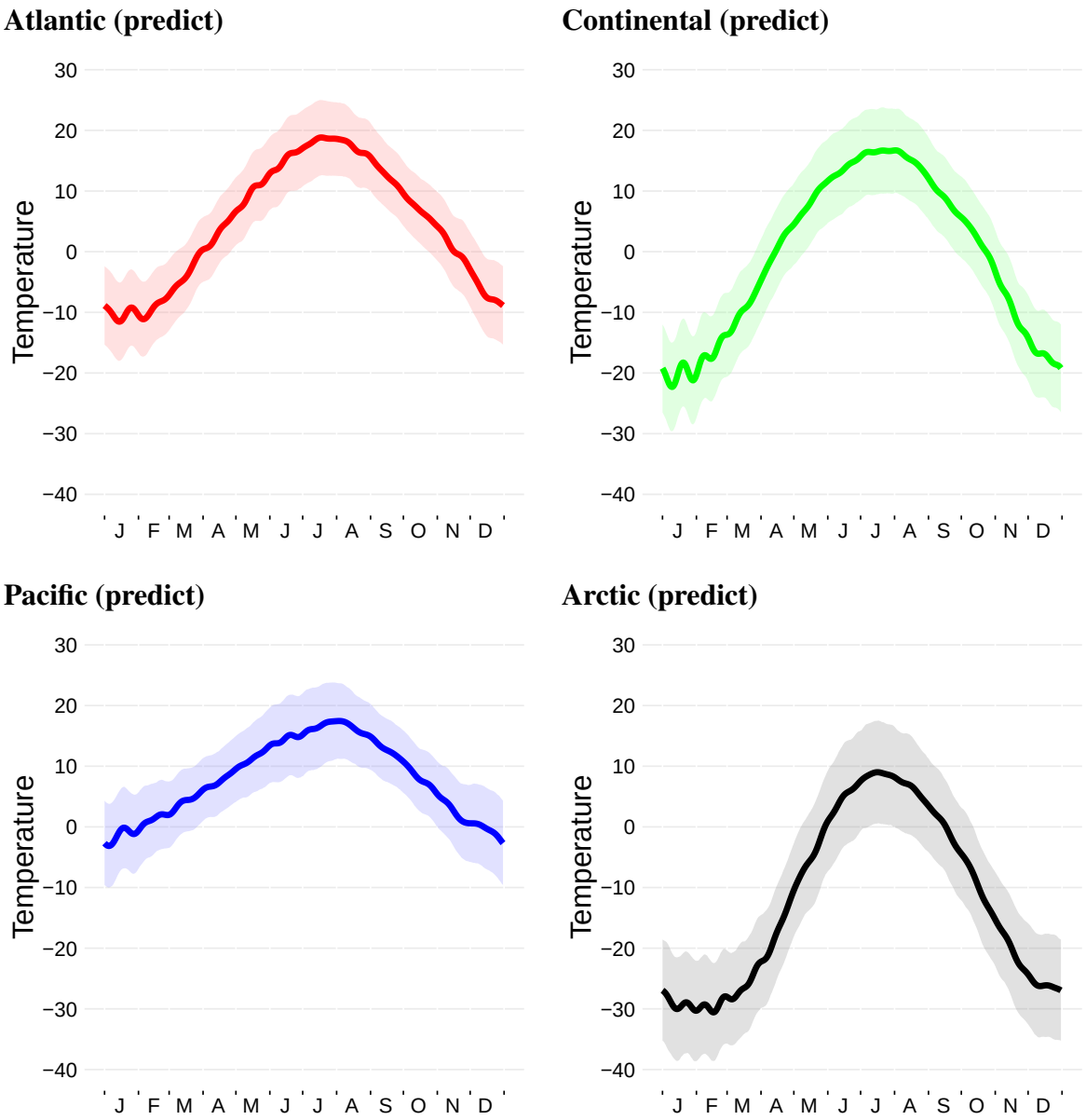


Figure 18 – Predicted temperature (^oC) over time for the four regions of Canada. Thicker lines represent the predicted average and shaded areas represent pointwise predictive bands with 95% of credibility.

The graphs presented in Figure 19 are related to the estimation of the Predictive Dissimilarity Index over different periods (full year, summer and winter), using the two distance metrics (4.2). The results show that considering the full year, the regions *Pacific x Arctic* have the *largest degree of dissimilarity* when compared with other combinations two-by-two. For these regions, there is a confidence level greater than 95% that the average difference between them is expected to be greater than 13 ^oC and the maximum difference is expected to be greater than 30 ^oC. the regions *Atlantic x Pacific* have the *smallest degree of dissimilarity* when compared with others. For these regions, there is a confidence level greater than 95% that the average difference between them is expected to be greater than 3 ^oC and the maximum difference is expected to be

greater than 9.3 ^oC. Table 4 highlights important values of PDI based on the graphs.

Looking at the PDI index for summer and winter periods, we observe that, in general, the regions tend to have minor temperature dissimilarities on summer than on winter. Furthermore, the temperature differences between Continental, Atlantic and Pacific are similar on summer, and the Arctic appears to have higher but similar differences from the other regions during this period. On winter, the average temperature difference tends to be distinct for each two-by-two regions comparison. During the summer, Continental temperature seems to be more similar to Pacific than the Arctic, bu on winter, Continental temperature seems to be more similar to Arctic than to Pacific. This probably happened because there are four provinces classified on Continental region (Churchill, Uranium Cty, Whitehorse and Dawson) that has latitudes closed to latitudes of Arctic provinces.

The homogeneity of the distributions for every $t \in [a,b]$ was evaluated through the KS distance between the distributions of the processes (result (3.19)) using the posterior DPM parameters simulated from the scheme presented in Section 3.3. Figure 20 shows the results of homogeneity tests. The highlighted threshold $\gamma^* = 0.2$ is a suggestion of precision level based on the results of simulated data (Section 4.3) and it was used to calculate the pragmatic hypothesis $P(H_{0,t};\gamma^*) = d(F_X,F_Y) \leq \gamma^*$.

The graphs have the same display of the ones presented in (PINI; VANTINI, 2017), in order to make it easy some comparisons between the results. Comparing Arctic to other regions, the null hypothesis of homogeneity was rejected over the full year. Looking at Continental *versus* Atlantic and Continental *versus* Pacific, we observe that the null hypothesis is rejected for autumn and winter. Finally, when comparing Atlantic *versus* Pacific, we conclude that the null hypothesis is rejected only for winter. The conclusions are similar to that presented in (PINI; VANTINI, 2017), but on their work, the acceptance regions were lightly larger than the ones presented here.

Table 4 – Predictive Dissimilarity Index varying periods of the year and confidence levels.

Confidence Level	Regions	Predictive Dissimilarity Index					
		Average Temperature Difference			Maximum Temperature Difference		
		Full Year	Summer	Winter	Full Year	Summer	Winter
99%	Pacific - Arctic	13.08	7.53	13.47	27.19	22.46	19.07
	Atlantic - Arctic	9.70	6.39	10.04	19.91	16.80	15.67
	Continental - Arctic	5.32	4.79	3.46	14.58	12.99	7.98
	Continental - Pacific	5.00	2.44	4.69	13.87	7.52	9.80
	Atlantic - Continental	2.70	1.92	2.25	8.39	5.64	6.28
	Atlantic - Pacific	2.60	1.84	1.73	7.90	5.34	4.26
95%	Pacific - Arctic	15.00	9.13	16.44	29.80	25.35	22.14
	Atlantic - Arctic	11.53	8.16	12.73	22.41	19.48	18.50
	Continental - Arctic	6.73	6.22	5.15	17.15	15.82	10.52
	Continental - Pacific	6.17	2.95	6.76	16.58	9.22	12.73
	Atlantic - Continental	3.36	2.27	3.22	10.27	6.71	8.36
	Atlantic - Pacific	3.12	2.23	2.26	9.32	6.49	5.36
90%	Pacific - Arctic	16.05	10.19	18.07	31.24	26.84	23.79
	Atlantic - Arctic	12.62	9.31	14.10	23.77	21.03	19.99
	Continental - Arctic	7.68	7.17	6.27	18.53	17.29	12.03
	Continental - Pacific	6.83	3.27	8.04	18.05	10.25	14.38
	Atlantic - Continental	3.84	2.50	4.00	11.43	7.39	9.71
	Atlantic - Pacific	3.46	2.47	2.66	10.24	7.21	6.05

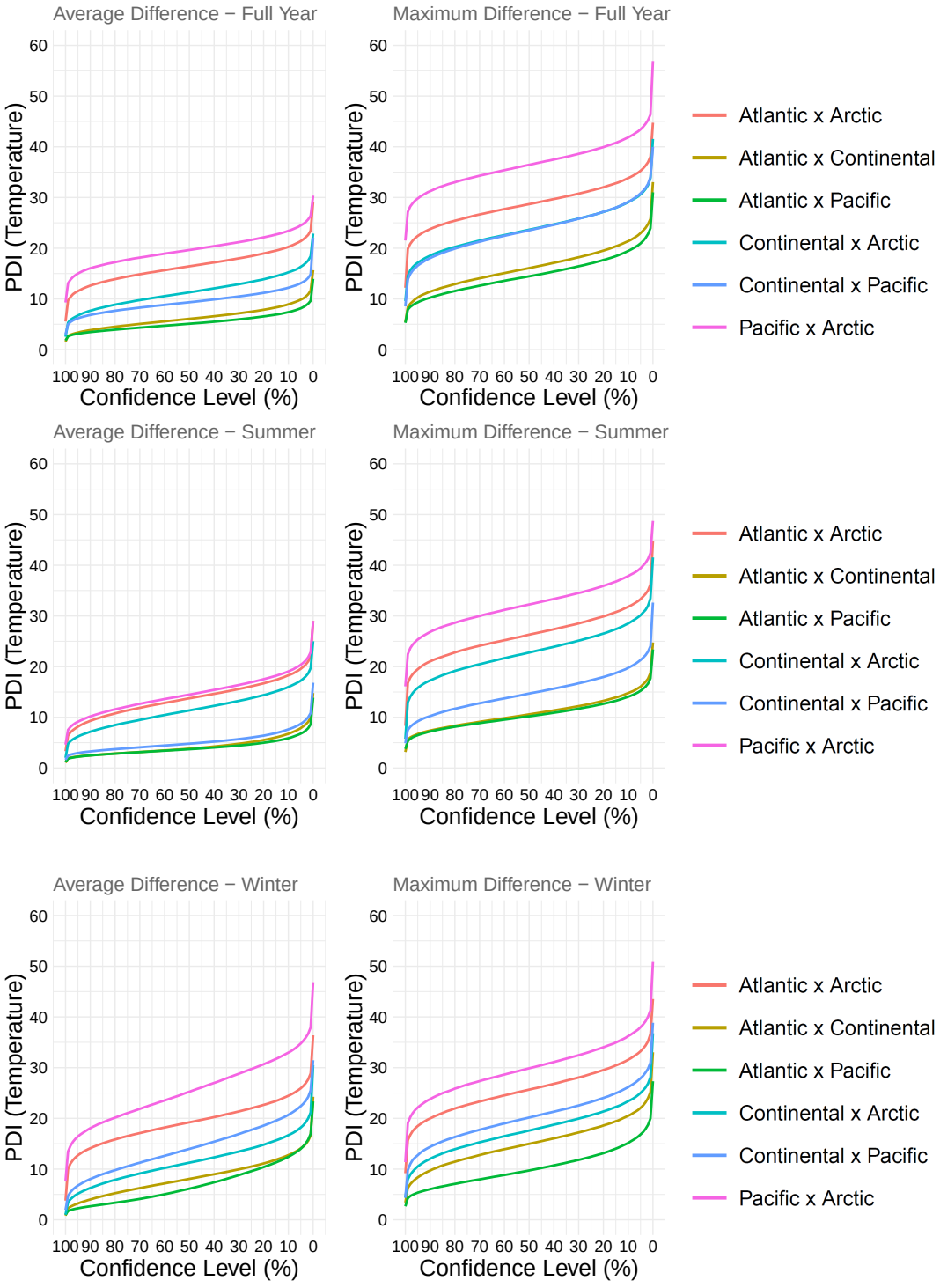


Figure 19 – Predictive Dissimilarity Index for regions two-by-two using distance metrics d_1 (maximum difference) and d_2 (average difference) over full year, summer and winter.

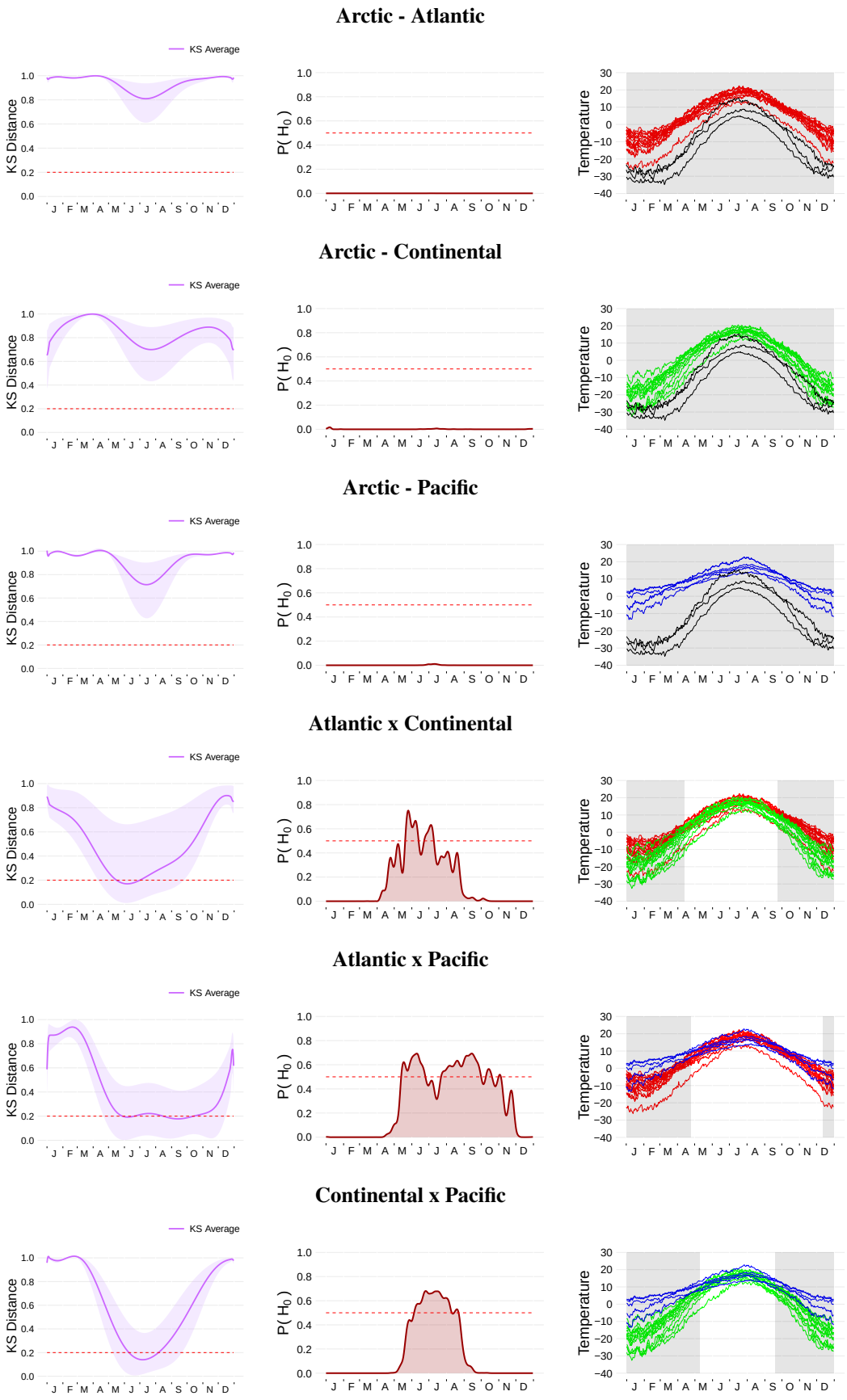


Figure 20 – On left: estimated KS Distance between the distributions of the curves for each groups two-by-two over time (shaded areas represent pointwise predictive bands with 95% of credibility.). On center: estimated probability of null hypothesis of homogeneity between the curves over time. On right: overlap of observed curves of the compared groups (gray areas represent periods where the hypothesis of homogeneity is rejected).

CHAPTER

6

CONCLUSIONS

The focus of this work was to present Bayesian methods to model and compare two groups of functional data. For that, we divided the problem into three main parts. The first is *functional data representation*, which is a pre-processing step necessary to link the empirical measurements collected to the functional nature presented in theory. This representation was made using series expansion with orthonormal basis functions due to the bijective relation between the curves and their respective coefficients of the series. This property enables us to perform analyses only with coefficients, which makes it easier to apply statistical models.

The second step was *modeling functional data*. Considering that each individual is associated to a vector of real coefficients, the challenge is resumed on modeling multivariate data and the interest is to model the multivariate density associated to the coefficients. Under nonparametric Bayesian methods, the densities were estimated using Dirichlet Process Mixture model with normal kernels, where the fourier coefficients is distributed as a mixture of parametric distributions (multivariate normals) and a Dirichlet Process prior is assumed to the mixing measure. With the posterior simulations and estimated densities at hand, we were able to get predictive samples of coefficients and build new predicted curves. This step demanded intensive computational programming skills due to the complexity of the model and the dimension of data involved. Simulations showed that up to now the model is predicting multivariate data in a very effective way with relatively low time, even for high dimensions.

The third step was to propose methods *to compare two groups of functional data*. For that, we suggested an index that measures the dissimilarity between the groups in some fixed interval $[a,b]$ of the domain, using predictive samples of the fitted model. Thie index is a great tool to compare functional data groups global and locally and has a strong interpretative appeal, providing the dissimilarity notion at the same scale of the interest variable. We also proposed a bayesian approach to assess the homogeneity of the groups through the measure of distance between the distributions of the processes for every point of time. The evaluation of probability of null hypothesis is a topic to be explored with more details in future works. As we could see,

the use of pragmatic hypothesis requires the choice of a precision level that can be difficult to be found in practice and might be adjusted to the demands of the referred research. A detailed simulation study can be done to specify sets of distance that, in practice, can be considered references of how much the pragmatic hypothesis can be enlarged under equally distributed data.

In general, both modeling and comparison of functional data were successfully achieved in this work. The next steps include studies of other decision rules to test the homogeneity of two or more groups of functional data and compare the method to other tests proposed in literature, like (PINI; VANTINI, 2017) and (KIM; LEE; LEI, 2019).

We finish acknowledging all the researchers and students who collaborate to the growth and development of (R Core Team, 2021), the software where the analysis were integrally performed.

BIBLIOGRAPHY

ANTONIAK, C. E. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. **The annals of statistics**, JSTOR, p. 1152–1174, 1974. Citation on page 33.

BILLINGSLEY, P. **Probability and measure**. [S.l.]: John Wiley & Sons, 2008. Citation on page 35.

COSCRATO, V.; ESTEVES, L. G.; IZBICKI, R.; STERN, R. B. Interpretable hypothesis tests. **arXiv preprint arXiv:1904.06605**, 2019. Citation on page 54.

CUEVAS, A.; FEBRERO, M.; FRAIMAN, R. An anova test for functional data. **Computational statistics & data analysis**, Elsevier, v. 47, n. 1, p. 111–122, 2004. Citation on page 22.

ESTEVES, L. G.; IZBICKI, R.; STERN, J. M.; STERN, R. B. Pragmatic hypotheses in the evolution of science. **Entropy**, Licensee MDPI, v. 21, n. 9, 2019. Citation on page 54.

FERGUSON, T. S. A bayesian analysis of some nonparametric problems. **The annals of statistics**, JSTOR, p. 209–230, 1973. Citations on pages 34, 35, and 36.

GREEN, P. J.; SILVERMAN, B. W. **Nonparametric regression and generalized linear models: a roughness penalty approach**. [S.l.]: Crc Press, 1993. Citation on page 30.

HALL, P.; TAJVIDI, N. Permutation tests for equality of distributions in high-dimensional settings. **Biometrika**, Oxford University Press, v. 89, n. 2, p. 359–374, 2002. Citation on page 23.

IZBICKI, R.; SANTOS, T. M. dos. **Aprendizado de máquina: uma abordagem estatística**. [S.l.: s.n.], 2020. ISBN 978-65-00-02410-4. Citation on page 32.

JOHNSON, R.; WICHERN, D. **Applied multivariate statistical analysis**. 5. ed. ed. Upper Saddle River, NJ: Prentice Hall, 2002. Citation on page 45.

KIM, I.; LEE, A. B.; LEI, J. Global and local two-sample tests via regression. **Electronic Journal of Statistics**, Institute of Mathematical Statistics and Bernoulli Society, v. 13, n. 2, p. 5253–5305, 2019. Citation on page 68.

KREYSZIG, E. **Introductory functional analysis with applications**. [S.l.]: wiley New York, 1978. Citations on pages 25 and 28.

MACEACHERN, S. N. Estimating normal means with a conjugate style dirichlet process prior. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 23, n. 3, p. 727–741, 1994. Citation on page 42.

MÜLLER, P.; ERKANLI, A.; WEST, M. Bayesian curve fitting using multivariate normal mixtures. **Biometrika**, Oxford University Press, v. 83, n. 1, p. 67–79, 1996. Citations on pages 41, 42, and 75.

MÜLLER, P.; QUINTANA, F. A.; JARA, A.; HANSON, T. **Bayesian nonparametric data analysis**. [S.l.]: Springer, 2015. Citations on pages 13 and 39.

MÜLLER, P.; RODRIGUEZ, A. **Nonparametric bayesian inference**. [S.l.]: Institute of Mathematical Statistics; American Statistical Association, 2013. Citation on page 39.

PINI, A.; VANTINI, S. Interval-wise testing for functional data. **Journal of Nonparametric Statistics**, Taylor & Francis, v. 29, n. 2, p. 407–424, 2017. Citations on pages 14, 23, 59, 64, and 68.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2021. Available: <https://www.R-project.org/>. Citation on page 68.

RAMSAY, J.; SILVERMAN, B. W. **Functional data analysis**. [S.l.]: (pringer, 1997. Citations on pages 21, 23, 29, and 59.

SETHURAMAN, J. A constructive definition of dirichlet priors. **Statistica sinica**, JSTOR, p. 639–650, 1994. Citation on page 37.

SWARTZ, T. Nonparametric goodness-of-fit. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 28, n. 12, p. 2821–2841, 1999. Citation on page 54.

ZHANG, C.; PENG, H.; ZHANG, J.-T. Two samples tests for functional data. **Communications in Statistics—Theory and Methods**, Taylor & Francis, v. 39, n. 4, p. 559–578, 2010. Citation on page 23.

APPENDIX

A

FURTHER TOPICS IN FUNCTIONAL ANALYSIS

A.1 Convergence and completeness of a metric space

Sequences plays an important role when studying convergence of real or complex numbers and the same is applied for arbitrary metric spaces. The following definitions and theorems explain these two important features of metric spaces, which shall be important on later definitions.

Definition: A sequence $\{x_n\}$ in a metric space (X,d) is said to converge if there is an $x \in X$ such that,

$$\lim_{n\to\infty} d(x_n,x) = 0$$

So, x is called *limit* of $\{x_n\}$ and

$$\lim_{n\to\infty} x_n = x.$$

Otherwise, $\{x_n\}$ is said to be *divergent*.

Here, d yields the sequence of real numbers $a_n = d(x_n,x)$ whose convergence defines that of $\{x_n\}$. Hence if $x_n \to x$, an ε being given, there is an $N = N(\varepsilon)$ such that all x_n with $n > N$ lie in the $\varepsilon-$neighborhood $B(x;\varepsilon)$ of x. Finally, it's notable that the limit of a convergent sequence must be a point of the space X.

Theorem: Let (X,d) be a metric space. Then,

(a) A convergent sequence in X is bounded and its limit is unique.

(b) if $x_n \to x$ and $y_n \to y$, then $d(x_n,y_n) \to d(x,y)$.

Definition: A sequence $\{x_n\}$ in a metric space (X,d) is said to be *Cauchy* if for every $\varepsilon > 0$ there is an $N = N(\varepsilon)$ such that,

$$d(x_m, x_n) < \varepsilon$$

for every $m,n > N$. The space (X,d) is said to be *complete* if every Cauchy sequence in X converges.

Theorem: Every convergent sequence in a metric space is a Cauchy sequence.

Theorem: Let M be a nonempty subset of a metric space (X,d) and $\overline{M}$ its closure. Then:

(a) $x \in \overline{M}$ if and only if there is a sequence (x_n) in M such that $x_n \longrightarrow x$.

(b) M is closed if and only if the situation $x_n \in M$, $x_n \longrightarrow x$ implies that $x \in M$.

Theorem: A subspace M of a complete metric space X is itself complete if and only if the set M is closed in X.

The set X of the Riemann integrable and continuous functions in $[a,b]$ is denoted by $L^1([a,b])$. It can be proved that the function space $C[a,b]$, where $a,b \in \mathbb{R}, a < b$, is complete . However, for the space $L^1([a,b])$ and the metric

$$d(x,y) = \int_a^b |x(t) - y(t)|dt,$$

the $L^1([a,b])$ is not complete. Fortunately, this problem can be corrected by completing the space.

A.2 Completion of metric spaces

Definition: Let $X = (X,d)$ and $\tilde{X} = (\tilde{X},\tilde{d})$ be arbitrary metric spaces. Then,

(a) A mapping T of X into $\tilde{X}$ is said to be *isometric* or an *isometry* if T preserves distances, that is, if for all $x,y \in X$,

$$\tilde{d}(Tx,Ty) = d(x,y)$$

where Tx and Ty are the images of x and y, respectively.

(b) The space X is said to be isometric with the space $\tilde{X}$ if there exists a bijective isometry of X onto $\tilde{X}$. The spaces X and $\tilde{X}$ are then called isometric spaces.

Theorem: For a metric space $X = (X,d)$ there exists a complete metric space $\hat{X} = (\hat{X},\hat{d})$ which has a subspace W that is isometric with X and is dense in $\hat{X}$. This space $\hat{X}$ is unique except for isometries, that is, if $\tilde{X}$ is any complete metric space having a dense subspace $\tilde{W}$ isometric with X, then $\tilde{X}$ and $\hat{X}$ are isometric.

This result allows us to complete the function space $L^1([a,b])$. This first approach introduced the concept of completeness and how the choice of d affects this important property of a metric space. On next subsections we will see that important metric spaces, including $L^1([a,b])$, are obtained if we take a vector space and define on it a metric by means of a *norm*.

A.3 Vector spaces

Definition: A *vector space* (or linear space) over a field K is a nonempty set X of elements $x, y, \ldots$ (called vectors) together with two algebraic operations. These operations are called *vector addition* and *multiplication of vectors by scalars*, that is, by elements of K.

Definition: A *subspace* of a vector space X is a nonempty subset Y of X such that for all $y_1, y_2 \in Y$ and all scalars α, β, we have $\alpha y_1 + \beta y_2 \in Y$.

Definition: A *linear combination* of vectors $x_1, \ldots, x_m$ of a vector space X is an expression of the form,

$$\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_m x_m$$

where $\{\alpha_i\}_{1 \leq i \leq m}$ are any set of scalars.

Definition: For any nonempty subset $M \in X$, the set of all linear combinations of vectors of M is called the *span* of M, written $\text{span}(M)$.

Definition: For a given set M of vectors $x_1, \ldots, x_r (r \geq 1)$ in a vector space X such that

$$\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_r x_r = 0$$

where $\{\alpha_i\}_{1 \leq i \leq r}$ are any set of scalars, if the only $r-$tuple of scalar for which the above equation holds is $\alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$, then M is said to be *linearly independent*. Otherwise, M is said to be *linearly dependent*.

Definition: A vector space X is said to be finite dimensional if there is a positive integer n such that X contains a linearly independent set of n vectors whereas any set of $n+1$ or more vectors of X is linearly dependent. n is called the dimension of X, written $n = \dim(X)$. By definition, $X = \{0\}$ is finite dimensional and $\dim(X) = 0$. If X is not finite dimensional, it is said to be infinite dimensional.

Definition: If $\dim(X) = n$, a linearly independent $n-$tuple of vectors of X is called a *basis* for X (or a basis in X). If $\{e_1, ..., e_n\}$ is a basis for X, every $x \in X$ has a unique representation as a linear combination of the basis vectors:

$$x = \alpha_1 e_1 + \alpha_2 e_2 + \cdots + \alpha_n e_n$$

More generally, if X is any vector space, not necessarily finite dimensional, and B is a linearly independent subset of X which spans X, then B is called a basis for X. Hence if B is a basis for X, then every nonzero $x \in X$ has a **unique** representation as a linear combination of elements of B with nonzero scalars as coefficients.

APPENDIX

B

DETAILS ABOUT FULL CONDITIONAL DISTRIBUTIONS OF POSTERIOR DPM WITH NORMAL KERNELS

Here we present the pdf formulas about the multivariate distributions used on Chapter 3. Furthermore, we proof the full conditional distributions of DPM with normal kernels presented by (MÜLLER; ERKANLI; WEST, 1996). Equation (B.1) represents the complete posterior distribution which the full conditionals are derived.

B.1 Probability distribution functions

B.1.1 Multivariate Normal distribution

Definition: a random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ has multivariate normal distribution if,

$$f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^p|\boldsymbol{\Sigma}|}}$$

where μ is the mean vector and Σ is the covariance matrix. The inverse of the covariance matrix, Σ^{-1} is known as precision matrix. Notation: $\boldsymbol{X} \sim N_p(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$.

Property: if $\boldsymbol{X} \sim N_p(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ then,

$$\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu}$$
$$\mathbb{V}[\boldsymbol{X}] = \boldsymbol{\Sigma}$$

B.1.2 Wishart distribution

The Wishart distribution is a distribution over matrices elements. It is a generalization to multiple dimensions of the gamma distribution.

Definition: let $\boldsymbol{\Sigma}$ be a $p \times p$ symmetric matrix of random variables that is positive semi-definite. Then, if $n \geq p$, $\boldsymbol{\Sigma}$ has a Wishart distribution with n degrees of freedom if it has the following probability density function,

$$f_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}|n,\boldsymbol{V}) = \frac{1}{2^{np/2}|\boldsymbol{V}|^{n/2}\Gamma_p\left(\frac{n}{2}\right)}|\boldsymbol{\Sigma}|^{(n-p-1)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{V}^{-1}\boldsymbol{\Sigma}\right)\right\}$$

where $\boldsymbol{V}$ is a symmetric positive definite matrix parameter of size $p \times p$, Γ_p is the multivariate gamma function and $\mathrm{tr}(\cdot)$ is the trace function. Notation: $\boldsymbol{\Sigma} \sim W_p(\boldsymbol{\Sigma};n,\boldsymbol{V})$.

Property 1: if $\boldsymbol{\Sigma} \sim W_p(\boldsymbol{\Sigma};n,\boldsymbol{V})$ then,

$$\begin{aligned}\mathbb{E}[\boldsymbol{\Sigma}] &= n\boldsymbol{V} \\ \mathbb{V}[\boldsymbol{\Sigma}_{ij}] &= (nv_{ij}^2 + v_{ii}v_{jj})\end{aligned}$$

Property 2: in Bayesian context, if $\boldsymbol{X} = (X_1,\ldots,X_m)$ is a random multivariate normal samples of size m, where $X_i \sim N_p(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ and a Wishart prior distribution is assigned to the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, that is, $\boldsymbol{\Omega} \sim W_p(\boldsymbol{\Omega};p,\boldsymbol{V^{-1}})$, then,

$$\boldsymbol{\Omega}|\boldsymbol{X} \sim W_p\left(\boldsymbol{\Omega};n+m,(\boldsymbol{X}\boldsymbol{X}^T)^{-1}+\boldsymbol{V}^{-1}\right)$$

B.2 Full Conditional Posterior densities

B.2.1 Complete Posterior Distribution

$$\begin{aligned}p\left(\boldsymbol{\mu^*},\boldsymbol{\Sigma^{*-1}},m,S,V^{-1}\middle|\boldsymbol{B},\mathscr{I},s,k,\alpha\right) \propto & \prod_{j=1}^{k}\left[\left|\Sigma_j^{*-1}\right|^{\frac{s+n_j}{2}-\frac{1}{2}}\right] \\ & \times \exp\left\{-\frac{1}{2}\sum_{j=1}^{k}\sum_{i:\mathscr{I}_i=j}(B_i-\mu_j^*)^T\Sigma_j^{*-1}(B_i-\mu_j^*)\right\} \\ & \times \exp\left\{-\frac{1}{2}\sum_{j=1}^{k}(\mu_j^*-m)^TV^{-1}(\mu_j^*-m)\right\} \qquad \text{(B.1)} \\ & \times |S|^{\frac{sk}{2}}\exp\left\{-\frac{1}{2}\sum_{j=1}^{k}\mathrm{tr}\left(sS\Sigma_j^{*-1}\right)\right\} \\ & \times \exp\left\{-\left[\frac{1}{2}(m-a)^TA^{-1}(m-a)+1\right]\right\} \\ & \times |S|^{(q-p-1)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left(qR^{-1}S\right)\right\} \\ & \times \left|V^{-1}\right|^{(c+k-p-1)/2}\exp\left\{-\frac{1}{2}\mathrm{tr}\left(cCV^{-1}\right)\right\}\end{aligned}$$

B.2.2 Full conditional for μ^*

$$
\begin{aligned}
p\left(\boldsymbol{\mu^*}\middle|\cdot\right) &\propto \exp\left\{-\frac{1}{2}\sum_{j=1}^{k}\left[(\mu_j^*-m)^T V^{-1}(\mu_j^*-m)+\sum_{i:\mathscr{I}_i=j}(z_i-\mu_j^*)^T\Sigma_j^{*^{-1}}(z_i-\mu_j^*)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\sum_{j=1}^{k}\left[(\mu_j^*-m)^T V^{-1}(\mu_j^*-m)+\sum_{i:\mathscr{I}_i=j}(\mu_j^*-z_i)^T\Sigma_j^{*^{-1}}(\mu_j^*-z_i)\right]\right\}
\end{aligned}
$$

Property 1: Let $Q_i(x)=(x-a_i)^T A_i(x-a_i)$ be a quadratic form on x. So, $\sum_i Q_i(x)=(x-\boldsymbol{a})^T\boldsymbol{A}(x-\boldsymbol{a})$, where,

$$
\begin{aligned}
\boldsymbol{A} &= \sum_i A_i \\
\boldsymbol{a} &= \left[\sum_i\left(A_i+A_i^T\right)\right]^{-1}\cdot\left[\sum_i\left(A_i+A_i^T\right)a_i\right]
\end{aligned}
$$

Then,

$$
\begin{aligned}
p\left(\boldsymbol{\mu^*}\middle|\cdot\right) &\propto \exp\left\{-\frac{1}{2}\sum_{j=1}^{k}\left[(\mu_j^*-m)^T V^{-1}(\mu_j^*-m)+(\mu_j^*-\bar{z_j})^T\left(n_j\Sigma_j^{*^{-1}}\right)(\mu_j^*-\bar{z_j})\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\sum_{j=1}^{k}\left[(\mu_j^*-m_j)^T T_j^{-1}(\mu_j^*-m_j)\right]\right\}
\end{aligned}
$$

where, using Property 1,

$$
\begin{aligned}
T_j^{-1} &= V^{-1}+n_j\Sigma_j^{*^{-1}} \\
m_j &= T_j\left(V^{-1}m+n_j\Sigma_j^{-1}\bar{z_j}\right)
\end{aligned}
$$

Then,

$$
\begin{aligned}
p\left(\boldsymbol{\mu^*}\middle|\cdot\right) &\propto \exp\left\{-\frac{1}{2}\sum_{j=1}^{k}\left[(\mu_j^*-m_j)^T T_j^{-1}(\mu_j^*-m_j)\right]\right\} \\
&= \prod_{j=1}^{k}\exp\left\{-\frac{1}{2}\left[(\mu_j^*-m_j)^T T_j^{-1}(\mu_j^*-m_j)\right]\right\} \\
&\propto \prod_{j=1}^{k} N_p(\mu_j^*;m_j,T_j)
\end{aligned}
$$

Assuming $\boldsymbol{\mu^*}=(\mu_1,\ldots,\mu_k)$ independents, the full conditional posterior density for μ_j^* is,

$$
\boxed{\left(\mu_j^*\middle|\,\Sigma_j^{*^{-1}},m,S,V^{-1},\boldsymbol{B},\mathscr{I},s,k,\alpha\right)\sim N_p(m_j,T_j)} \tag{B.2}
$$

B.2.3 Full conditional for $\boldsymbol{\Sigma^{*^{-1}}}$

$$
\begin{aligned}
p\left(\boldsymbol{\Sigma^{*^{-1}}}\middle|\cdot\right) &\propto \prod_{j=1}^{k}\left[\left|\Sigma_j^{*^{-1}}\right|^{\frac{s+n_j}{2}-\frac{1}{2}}\right]\exp\left\{-\frac{1}{2}\left[\sum_{j=1}^{k}\operatorname{tr}\left(sS\Sigma_j^{*^{-1}}\right)+\sum_{j=1}^{k}\sum_{i:\mathscr{I}_i=j}(z_i-\mu_j^*)^T\Sigma_j^{*^{-1}}(z_i-\mu_j^*)\right]\right\}\\
&\propto \prod_{j=1}^{k}\left[\left|\Sigma_j^{*^{-1}}\right|^{\frac{s+n_j}{2}-\frac{1}{2}}\right]\exp\left\{-\frac{1}{2}\left[s\operatorname{tr}\left(S\sum_{j=1}^{k}\Sigma_j^{*^{-1}}\right)+\sum_{j=1}^{k}\left[(\mu_j^*-\bar{z_j})^T\left(n_j\Sigma_j^{*^{-1}}\right)(\mu_j^*-\bar{z_j})\right]\right]\right\}\\
&\propto \prod_{j=1}^{k}\left[\left|\Sigma_j^{*^{-1}}\right|^{\frac{s+n_j}{2}-\frac{1}{2}}\exp\left\{-\frac{1}{2}\operatorname{tr}\left[\left(sS+n_j(\mu_j^*-\bar{z_j})(\mu_j^*-\bar{z_j})^T\right)\Sigma_j^{*^{-1}}\right]\right\}\right]\\
&\propto \prod_{j=1}^{k}\left[\left|\Sigma_j^{*^{-1}}\right|^{\frac{s+n_j}{2}-\frac{1}{2}}\exp\left\{-\frac{1}{2}\operatorname{tr}\left(S_j^{-1}\Sigma_j^{*^{-1}}\right)\right\}\right]\\
&\propto \prod_{j=1}^{k}W_p\left(\Sigma_j^{-1};s+n_j,S_j\right)
\end{aligned}
$$

where,

$$S_j^{-1}=sS+n_j(\mu_j^*-\bar{z_j})(\mu_j^*-\bar{z_j})^T.$$

Assuming $\boldsymbol{\Sigma^*}=(\Sigma_1,\ldots,\Sigma_k)$ independents, the full conditional posterior density of $\Sigma_j^{*^{-1}}$ is,

$$\boxed{\left(\Sigma_j^{*^{-1}}\middle|\,\mu_j,m,S,V^{-1},\boldsymbol{B},\mathscr{I},s,k,\alpha\right)\sim W_p(s+n_j,S_j)} \tag{B.3}$$

B.2.4 Full conditional for m

$$
\begin{aligned}
p\left(m\middle|\cdot\right) &\propto \exp\left\{-\frac{1}{2}\left[(m-a)^TA^{-1}(m-a)+\sum_{j=1}^{k}(m-\mu_j^*)^TV^{-1}(m-\mu_j^*)\right]\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left[(m-a)^TA^{-1}(m-a)+(m-\bar{\boldsymbol{\mu}}^*)^T\left(kV^{-1}\right)(m-\bar{\boldsymbol{\mu}}^*)\right]\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left[(m-\hat{a})^T\hat{A}^{-1}(m-\hat{a})\right]\right\}\\
&\propto N(m;\hat{a},\hat{A})
\end{aligned}
$$

where,

$$
\begin{aligned}
\hat{A}^{-1} &= A^{-1}+kV^{-1}\\
\hat{a} &= \hat{A}\left(A^{-1}a+kV^{-1}\bar{\boldsymbol{\mu}}^*\right)
\end{aligned}
$$

The full conditional posterior density for m is,

$$\boxed{\left(m\middle|\,\boldsymbol{\mu^*},\boldsymbol{\Sigma^{*^{-1}}},S,V^{-1},\boldsymbol{B},\mathscr{I},s,k,\alpha\right)\sim N_p(\hat{a},\hat{A})} \tag{B.4}$$

B.2.5 Full conditional for S

$$\begin{aligned}
p\left(S\left|\cdot\right.\right) &\propto |S|^{\frac{sk}{2}}|S|^{\frac{q}{2}-\frac{p}{2}-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(qR^{-1}S\right)+\sum_{j=1}^{k}\operatorname{tr}\left(sS\Sigma_j^{*^{-1}}\right)\right]\right\}\\
&\propto |S|^{\frac{sk+q}{2}-\frac{p}{2}-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(qR^{-1}S\right)+\sum_{j=1}^{k}\operatorname{tr}\left(sS\Sigma_j^{*^{-1}}\right)\right]\right\}\\
&\propto |S|^{\frac{sk+q}{2}-\frac{p}{2}-\frac{1}{2}}\exp\left\{-\frac{1}{2}\operatorname{tr}\left[\left(qR^{-1}+s\sum_{j=1}^{k}\Sigma_j^{*^{-1}}\right)S\right]\right\}\\
&\propto |S|^{\frac{sk+q}{2}-\frac{p}{2}-\frac{1}{2}}\exp\left\{-\frac{1}{2}\operatorname{tr}\left(\hat{R}^{-1}S\right)\right\}\\
&\propto W_p\left(S;sk+q,\hat{R}\right)
\end{aligned}$$

where,

$$\hat{R}=\left(qR^{-1}+s\sum_{j=1}^{k}\Sigma_j^{*^{-1}}\right)^{-1}$$

The full conditional posterior density for S is,

$$\boxed{\left(S\left|\,\boldsymbol{\mu}^{*},\boldsymbol{\Sigma}^{*^{-1}},m,V^{-1},\boldsymbol{B},\mathscr{I},s,k,\alpha\right.\right)\sim W_p\left(sk+q,\hat{R}\right)} \tag{B.5}$$

B.2.6 Full conditional for V^{-1}

$$\begin{aligned}
p\left(V^{-1}\left|\cdot\right.\right) &\propto \left|V^{-1}\right|^{\frac{c+k}{2}-\frac{p}{2}-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(cCB^{-1}\right)+\sum_{j=1}^{k}(m-\mu_j^{*})^{T}V^{-1}(m-\mu_j^{*})\right]\right\}\\
&\propto \left|V^{-1}\right|^{\frac{c+k}{2}-\frac{p}{2}-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left[\operatorname{tr}\left(cCB^{-1}\right)+(m-\bar{\boldsymbol{\mu}}^{*})^{T}(kV^{-1})(m-\bar{\boldsymbol{\mu}}^{*})\right]\right\}\\
&\propto \left|V^{-1}\right|^{\frac{c+k}{2}-\frac{p}{2}-\frac{1}{2}}\exp\left\{-\frac{1}{2}\operatorname{tr}\left[\left(cC+k(m-\bar{\boldsymbol{\mu}}^{*})(m-\bar{\boldsymbol{\mu}}^{*})^{T}\right)V^{-1}\right]\right\}\\
&\propto \left|V^{-1}\right|^{\frac{c+k}{2}-\frac{p}{2}-\frac{1}{2}}\exp\left\{-\frac{1}{2}\operatorname{tr}\left[\hat{C}^{-1}V^{-1}\right]\right\}\\
&\propto W_p(V^{-1};c+k,\hat{C}).
\end{aligned}$$

where,

$$\hat{C}=\left(cC+k(m-\bar{\boldsymbol{\mu}}^{*})(m-\bar{\boldsymbol{\mu}}^{*})^{T}\right)^{-1}$$

The full conditional posterior density for B^{-1} is,

$$\boxed{\left(V^{-1}\left|\,\boldsymbol{\mu}^{*},\boldsymbol{\Sigma}^{*^{-1}},m,S,\boldsymbol{B},\mathscr{I},s,k,\alpha\right.\right)\sim W_p(c+k,\hat{C})} \tag{B.6}$$