Flexible Bayesian Inference for Families of Random Densities

By
Bastián Galasso-Díaz

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR IN STATISTICS
AT
PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
SANTIAGO, CHILE
January 30, 2020

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

DEPARTMENT OF STATISTICS

The undersigned hereby certify that they have read and recommend to the Faculty of Mathematics for acceptance of the thesis entitled **"Flexible Bayesian Inference for Families of Random Densities"** by **Bastián Galasso-Díaz** in partial fulfillment of the requirements for the degree of **Doctor in Statistics**.

Dated: January 30, 2020

Internal Supervisor: _____

Jorge González
Pontificia Universidad Católica de Chile

Lead Supervisor: _____

Miguel de Carvalho
The University of Edinburgh, United Kingdom

Examing Committee: _____

Alejandro Jara
Pontificia Universidad Católica de Chile

_____

Fernando Quintana
Pontificia Universidad Católica de Chile

_____

Garritt Page
Brigham Young University, USA

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

Date: **January 30, 2020**

Author:	**Bastián Galasso-Díaz**
Title:	**Flexible Bayesian Inference for Families of Random Densities**
Department:	**Statistics**
Degree:	**Doctor in Statistics**
Convocation:	**November**
Year:	**2019**

_____

Signature of Author

Contents

Acknowledgments

During this long journey, a lot of people have walked along with me and now it comes the opportunity to say thank you very much for your continuous and unconditional support.

First of all, I would to thank my wife, Maria Ignacia; she brings me love and unconditional support during my studies. She always was there when I needed the most, and this dissertation would not have been possible without María Ignacia. I love you so much! My next words of gratitude go to my children, Lucca and Santino; I do everything for you, and you are my "cable to earth" in the tough moments.

I would like to thank my parents (Marcela and Italo) for allowing me to study and discover this beautiful career—I will always be grateful. Also, thanks to my siblings (Cony, Javi, and Gianni) and to my family in law (Maria Eugenia, Francisco, Nico, JP, Pelu, Mona, and Felipe), for always standing by my side.

To my supervisor, Miguel de Carvalho, who guided me in this adventure, always pushing me forward and continuously challenging me beyond expectations so that I could break new ground. This journey was beautiful, and I will be eternally grateful for that. It was a pleasure to work with you, and a privilege to know you!

To my colleague Yoav Zemel, with whom I have worked for hours—I learned a lot from you. Thanks for your collaboration and for sharing your time with me, it was a real honor work with you.

To my friends, Erik, Rodrigo, José, Luz, César, Alvaro, Beto, Cristian, Gabriel and Perla for the good moments and for being there to me. Your friendship is something that I always have in my heart.

To my tutors and masters, Fernando Quintana, Garritt Page, Jorge Gonzalez, Reinaldo Arellano, Alejandro Jara, Mario Ponce, Godofredo Iommi and Manuel Galea, for all their teachings and advice, not only on Statistics but about life—much appreciated. Thanks to

the Faculty of Mathematics of Pontificia Universidad Católica de Chile for their support since the very beginning of my career, I have fond memories of my time here.

Thanks also to the School of Mathematics at the University of Edinburgh for welcoming and receiving me twice. In particular, I would like to thank Vanda Inácio de Carvalho for so kindly welcoming me and for the enjoyable moments over my stay in Edinburgh.

Thanks to Steve MacEachern (ISBA president, 2016), from The Ohio State University, for supporting my trip to BNP11 (Paris).

This thesis was developed under the umbrella of the FCT project PTDC/MAT-STA/28649/2017. Finally, funding from Chilean NSF (CONICYT) via scholarship 21140901, is greatly appreciated. With their financial support, I could attend to national and international conferences, and also conduct research stays at the University of Edinburgh. These activities were vital in training me as a researcher.

I dedicate this thesis to my family.

Bastián Galasso-Díaz

Santiago, Chile

January 30, 2020

Abstract

A main goal of this thesis is to propose and study novel flexible Bayesian models for setups that entail families of random densities. Two specific contexts will be examined: one involves phase-varying point processes, whereas the other involves functional principal component analysis. The common denominator underlying these contexts is the need to model families of random measures to each of which corresponds a different data generating process. On both contexts, prior processes will be used so to devise priors on the target objects of interest.

In more detail, one context entails separating amplitude variation from phase variation in a multiple point process setting. In this framework, I pioneer the development of priors on spaces of warping maps by proposing a novel Bayesian semiparametric approach for modeling registration of multiple point processes. Specifically, I develop induced priors for warp maps via a Bernstein polynomial prior so to learn about the structural measure of the point process and about the phase variation in the process. Theoretical properties of the induced prior, including support and posterior consistency, are established under a fairly mild proviso. Also, numerical experiments are conducted to assess the performance of this new approach; finally, a real data application in climatology illustrates the proposed methodology.

The other context that will be considered in this thesis involves modeling families of random densities using functional principal component analysis through the so-called Karhunen–Loève decomposition. For this, I develop a data-driven prior based on the Karhunen–Loève decomposition which can be used to borrowing strength across samples. The proposed approach defines a prior on the space of families of densities. Theoretical properties are developed to ensure that the trajectories from an infinite mixture belong to L^2 which is a necessary condition for the Karhunen–Loève decomposition to hold. Numerical experiments are conducted to assess the performance of the proposed approach against competing methods, and we offer an illustration by revisiting Galton's height parents dataset.

List of Figures

CHAPTER 1

Introduction and Background

This introductory chapter offers preparations and sketches the problems to be addressed over this thesis. Particularly, this chapter provides an introduction to semiparametric and nonparametric Bayesian inference via prior processes and it reviews key concepts, methods, and ideas related with functional data analysis, decomposition of random functions, and with phase variation.

1.1 Introduction

This thesis develops novel Bayesian models for two specific contexts that will require modeling families of random densities. Prior to giving details on the problems to be addressed, I start with background and preparations.

A main goal of Statistical Science is to learn about random phenomena from data. In this context, parsimonious simplifications of reality are often made to achieve a desirable degree of mathematical tractability, but that aim to preserve intact key aspects of the scientific problem of interest. Of course, in this process not only data plays a primary role but also background knowledge on the scientific problem of interest is of the utmost importance. Formally, we refer data to as a collection $X_1, \ldots, X_n$, where X_i is a random vector corresponding to the ith experimental unit from a sample of size n from a joint probability distribution F.

In the classical frequentist parametric setting, we assume that F belongs to some known class of distribution functions $\mathscr{F} = \{F_\theta \ : \ \theta \in \Theta\}$, where Θ is a finite-dimensional parameter space. The set $\mathscr{F}$ is known as a statistical model and the main goal of the statistician is to learn about θ, such that $F = F_\theta$, from data. In this framework, θ is assumed to be fixed but unknown, and therefore we need to learn about a plausible value of θ from data, which is the so-called problem of point estimation.

The Bayesian take on the setting above, relies on considering the parameter θ as random, and it aims to learn about the (posterior) distribution of θ from data. There are two inputs in a Bayesian analysis: Likelihood and prior. The prior distribution can be specified according to the prior knowledge of the problem and then it is updated using the data, yielding another distribution (posterior) via Bayes theorem.

Both approaches described above require a finite-dimensional parameter space. Further flexibility can be achieve by working with richer parameter spaces, such as spaces of functions. In the latter setting, the class of distribution functions $\mathscr{F} = \{F_\theta \ : \ \theta \in \mathbb{F}\}$,

where $\mathbb{F}$ is an infinite-dimensional space, is called a nonparametric statistical model. The infinite-dimensional space $\mathbb{F}$ is usually a space of functions and a common class of interest is that of probability distributions. Of course, it is not the only class of functions of interest, in some context we can have interest in other types of functions, such as, for instance, the conditional mean function in a nonparametric regression. Also, there exists an intermediate class of models between the parametric and nonparametric models, which are called semiparametric; semiparametric models factorize the parameter space into a part belonging to a finite-dimensional space and another in an infinite-dimensional space, formally $\mathscr{F} = \{(F_\theta, F) \ : \ \theta \in \Theta, F \in \mathbb{F}\}$. Therefore, under the Bayesian (semi and nonparametric) paradigm we need to set a prior distribution over F, i.e., a probability measure over probability measures, which can be made precise using the concept of random probability measure; see Kallenberg (1983) for an introduction to random probability measures.

A key goal of this thesis is to propose novel flexible Bayesian models for setups that involve families of random densities in two specific contexts to be discussed below. In both contexts the setup relies on a K-sample setting where the interest is on learning about a family of continuous distributions $\{F_1, \ldots, F_K\}$. Therefore, for the latter setting we assume that we observe $\{X_{i,k}\}$ with

$$X_{1,k}, \ldots, X_{n_i,k} \mid F_k \sim F_k, \quad k = 1, \ldots, K;$$

if the F_k are absolutely continuous their corresponding density is denoted by f_k.

1.2 Prior Distributions on Probability Measures

The Dirichlet Process

Since the proposed inferences on this thesis are based on prior processes, we start by offering some preparations on these. Below, a prior process is to be understood as a prior over a space of functions. Such processes are the bread and butter of Bayesian semi and nonparametric inferences (see for example Ghosal (2010), Phadia (2015), Ghosal and Van der Vaart (2015), Müller et al. (2015), among others).

Ferguson (1973) states the following two desirable properties that random probability measures should possess so to define a prior process:

(i) Their support should be large.

(ii) Posterior inference should be analytically manageable.

In this line, the Dirichlet process was introduced by Ferguson (1973) as a random probability measure motivated by these two properties. The formal definition is as follows.

Definition 1 (Dirichlet Process)**.** *Let α be a positive real number and let H^* be a probability measure on a given Polish space $\mathscr{X}$. A random measure H on $\mathscr{X}$ is called a Dirichlet process if for every finite measurable partition $\{A_1, \ldots, A_k\}$ of $\mathscr{X}$, the joint distribution of $(H(A_1), \ldots, H(A_k))$ is a k-dimensional Dirichlet distribution with parameters $(\alpha H^*(A_1), \ldots, \alpha H^*(A_k))$.*

In Definition 1, α is the precision parameter, the measure H^* is the centering measure, and the notation for this type of process is $\mathrm{DP}(\alpha, H^*)$. The parameters in the Dirichlet process receive their names because it can be shown that if $H \sim \mathrm{DP}(\alpha, H^*)$, then

$$\mathbb{E}\{H(\cdot)\} = H^*(\cdot), \quad \mathbb{V}\mathrm{ar}\{H(\cdot)\} = \frac{H^*(\cdot)(1 - H^*(\cdot))}{1 + \alpha}. \tag{1.2.1}$$

Therefore, $H(\cdot)$ is a random probability measure whose realizations are centered around the measure $H^*(\cdot)$, whereas α controls the spread around $H^*(\cdot)$. Figure 1.1 shows realizations of $H \sim \text{DP}(\alpha, H^*)$ for different values of α and $H^*(\cdot) = \Phi(\cdot)$, where $\Phi(\cdot)$ stands for the standard normal distribution function. Below the centering distribution will be considered to be a distribution function. Yet it should be mentioned that extensions of the Dirichlet process have been devised where the centering distribution is itself a random probability measure. Particularly, Teh et al. (2006) and Teh and Jordan (2010) propose the so-called hierarchical Dirichlet process, which consists of a family of DPs whose centering distribution follows a common DP.

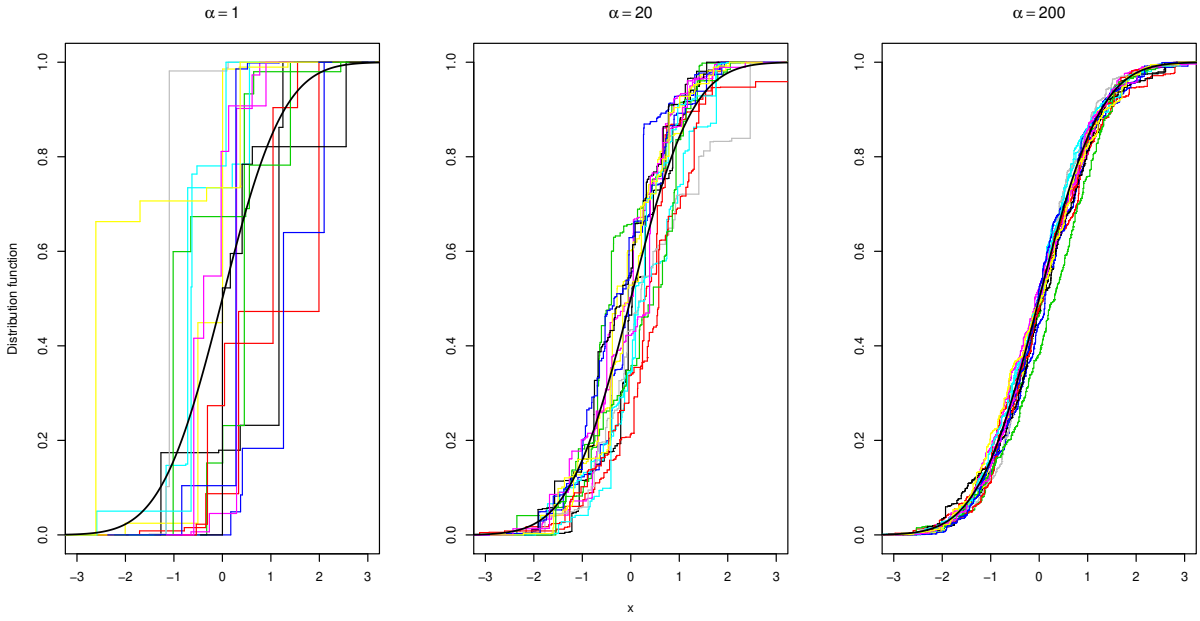


Figure 1.1: Trajectories of realizations of Dirichlet processes; the solid black line represents the centering distribution functions (standard normal)

A fundamental motivation for the Dirichlet process was the simplicity to obtain a posterior update. In fact, if we assume that $X_1, \ldots, X_n \mid H \overset{\text{iid}}{\sim} H$ where $H \sim \text{DP}(\alpha, H^*)$, then the posterior distribution of H given $X_1, \ldots, X_n$ is again a Dirichlet process; that is,

$H \mid X_1, \ldots, X_n \sim \mathrm{DP}(n + \alpha, H^{**})$ with

$$\mathbb{E}\{H \mid X_1, \ldots, X_n\} = \frac{\alpha}{\alpha + n} H^*(\cdot) + \frac{n}{\alpha + n} F_n(\cdot),$$

where $F_n(\cdot)$ denotes the empirical distribution function of $X_1, \ldots, X_n$. This follows from the Multinomial-Dirichlet conjugacy along with an argument based on the martingale convergence theorem (Ghosal 2010, Section 2.2.3).

A key property of the Dirichlet process is its discrete nature. This is a corollary to Sethuraman (1994) stick-breaking representation according to which any $H \sim \mathrm{DP}(\alpha, H^*)$ can be represented in the following form

$$H(\cdot) = \sum_{h=1}^{\infty} \omega_h \delta_{\theta_h}(\cdot), \tag{1.2.2}$$

where

$$\theta_h \overset{\text{iid}}{\sim} H^*, \quad \omega_h = v_h \prod_{1 \leq j < h} (1 - v_j), \quad v_j \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \alpha). \tag{1.2.3}$$

The stick-breaking weight construction mechanism described in (1.2.3) mimics the physical process of sequentially breaking a unit stick in a random way. We start with a unit stick at period $h = 0$ and at period $h = 1$ we break a portion of length $\omega_1 = v_1$, so that we are left with $1 - v_1$. At period $h = 2$ we break a piece of what was left, $\omega_2 = v_2(1 - v_1)$ and so that there remains $(1 - v_2)(1 - v_1)$ left to break; and so on.

There is another representation of the Dirichlet process that is based on the marginal distributions, and which is known as Polya urn representation of Blackwell and MacQueen (1973). The Polya urn scheme can be described as follows. Let $\mathscr{X} = \{1, \ldots, k\}$, and suppose that we have an urn with k different colors; each color has α_i balls with that color, for $i = 1, \ldots, k$. Let X_j be the color of the jth ball; we first take the first ball, whose

probability of being of color i is

$$P[X_1 = i] = \frac{\alpha_i}{\sum_i \alpha_i}.$$

After that, we return into the urn another ball of the same color, and then we draw another ball from the urn—say, X_2. The probability of the event $\{X_2 = j\}$ depends on the value of X_1, and thus we can compute the conditional probability as

$$P[X_2 = j \mid X_1 = i] = \frac{\alpha_j + \delta_j}{1 + \sum_i \alpha_i}, \quad \delta_j = 1 \text{ if } j = i.$$

The process can be repeated so as to obtain a sequence of exchangeable random variables $X_1, X_2, \ldots$, and using the de Finetti's theorem, we can obtain a prior distribution over the law of the X_i. Blackwell and MacQueen (1973) generalize this scheme for a continuum of colors, and showed that the Dirichlet process can also be defined in this way which provides convenient form to obtain random samples from DP.

Finally, another interesting aspect of the Dirichlet process is the behavior of its tail. Indeed, if $H \sim \mathrm{DP}(\alpha, H^*)$ then we know that $\mathbb{E}\{H(\cdot)\} = H^*(\cdot)$ and thus we might think that the tails of H^* and H are equal 'on average'. Yet, this is false as the tails of H are much thinner almost surely. The complete proof of this can be seen in Doss and Selke (1982) but a sketch of the argument is as follows. From Fristedt (1967) we know that if h is a strictly increasing and convex function on $(0, \epsilon)$, for some sufficiently small $\epsilon > 0$, then

$$\limsup_{x \to \infty} \frac{1 - H(x)}{h(1 - H^*(x))} = 0 \text{ a.s.} \quad \text{or} \quad \limsup_{x \to -\infty} \frac{H(x)}{h(H^*(x))} = 0 \text{ a.s.},$$

if and only if

$$\int_0^\epsilon \log h(x) \mathrm{d}x > -\infty.$$

A particular choice of h will yield, for example, that for almost every distribution function H, for all sufficiently large x,

$$1 - H(x) \leq \exp\left(-\frac{1}{(1 - H^*(x))[\log(1 - H^*(x))]^2}\right) < 1 - H^*(x),$$

This last expression is a direct consequence of Fristedt (1967, Th. 1), using the definition of limit for x large enough, and the fact that $x^3 < \exp\{x\}$ for any x. So, taking $x = -\log(1 - H^*(x))$ we have the result, and thus we can conclude that the tail of H is almost surely thinner than the centering measure H^*.

Further properties on the Dirichlet process are discussed by Ferguson (1973), Korwar and Hollander (1973), Antoniak (1974), Diaconis and Kemperman (1996), Cifarelli and Melilli (2000), Ghosal (2010), among others.

Dirichlet Process Mixtures and Extensions

Since the Dirichlet process generates discrete probability measures almost surely, it cannot be directly used for density estimation. This can be fixed by convolving its trajectories with a continuous kernel; in other words, in practice the Dirichlet process is often used as a mixing measure. This approach was introduced by Ferguson (1983), Lo (1984), Escobar (1988, 1994), and Escobar and West (1995).

In a formal specification, let Θ be a finite-dimensional parameter space and let $\mathbb{K}(x \mid \theta)$ be a continuous probability density function for $\theta \in \Theta$. Given a probability distribution function H defined on Θ, a mixture of $\mathbb{K}(x \mid \theta)$ with respect to the mixing measure H has the probability density function

$$f(x) = \int_{\Theta} \mathbb{K}(x \mid \theta) \mathrm{d}H(\theta). \tag{1.2.4}$$

This type of mixture forms a very rich family and a prior on densities may be induced by putting a Dirichlet process prior on the mixing measure H; this model is known as Dirichlet process mixture (DPM). The model in (1.2.4) with a Dirichlet process prior on the mixture measure can be formulated in an equivalent way as a hierarchical model as

$$\begin{aligned} X_i \mid \theta_i &\overset{\text{ind}}{\sim} \mathbb{K}(x \mid \theta_i), \\ \theta_i \mid H &\overset{\text{iid}}{\sim} H, \\ H &\sim \text{DP}(\alpha, H^*). \end{aligned}$$

In the hierarchical model representation of Dirichlet process mixture we introduce new latent variables θ_i, which can be used to induce a probability model on clusters. The discrete nature of the Dirichlet process, implies that ties among the latent variables θ_i will occur with positive probability. If we set θ_j^*, $j = 1, \ldots, k \leq n$ to be the unique values of the latent variables and $S_j = \{i \ : \ \theta_i = \theta_j^*\}$, then, $\rho_n = \{S_1, \ldots, S_k\}$ is a random partition of $\{1, \ldots, n\}$. The model $p(\rho_n)$ is known as the Polya urn.

I close this section with some remarks on extensions. Using Eq. (1.2.4) we can construct different kinds of prior on densities by changing the law of the mixing measure. Formally,

$$f(x) = \int_{\Theta} \mathbb{K}(x \mid \theta) \mathrm{d}P(\theta), \quad P \sim \text{RPM}(\mathbf{\Phi}),$$

where RPM is a random probability measure over Θ with hyperparameters $\mathbf{\Phi}$. Some examples of random probability measures that can be used as mixing measures include: generalized gamma NRMI (normalized random measure with independent increments) (Barrios et al. 2013), normalized inverse Gaussian (Lijoi et al. 2007), N-stable process (Kingman 1975), stick-breaking priors (Ishwaran and James 2001), and probit stick-breaking priors (Rodríguez and Dunson 2011).

Random Bernstein Polynomials

When we need to put a prior measure over all densities which are defined on a closed bounded interval, $[0,1]$ say, we can use the Dirichlet process mixture as in Eq. (1.2.4) with a kernel, $\mathbb{K}(x \mid \theta)$, defined on the mentioned compact set. While natural, the latter approach is far from elegant from a computational viewpoint. The developments to be presented in Chapter 2 will require priors on spaces of densities supported on bounded intervals, and thus we offer below some preparations on prior processes that achieve this. A more elegant approach has been proposed by Petrone (1999a,b), by resorting to a class of functions known as Bernstein polynomials.

To define this type of prior, we will start by defining Bernstein polynomials.

Definition 2 (Bernstein Polynomial)**.** *Let G be a bounded function on $[0,1]$ and let k be a positive integer. The Bernstein polynomial is defined as*

$$B(x \mid k, G) = \sum_{i=0}^{k} G\left(\frac{i}{k}\right) \binom{k}{i} x^i (1-x)^{k-i}. \tag{1.2.5}$$

Some comments on the parameters k and G are in order. The parameter k is tantamount to the number of components in a mixture model, when G is a distribution function. The next theorem shows that B approximates G as $k \to \infty$.

Theorem 1. *For a function G bounded on $[0,1]$, the relation*

$$\lim_{k\to\infty} B(x \mid k, G) = G(x),$$

holds at each point of continuity x of G. Moreover, this relation holds uniformly on $[0,1]$ if G is continuous on this interval.

It can be shown that the Bernstein polynomial described in Eq. (1.2.5) is a distribution function, provided G is also a distribution function. Moreover, if $G(0) = 0$ then we can

obtain the derivative of Eq. (1.2.5), which corresponds to the density

$$b(x \mid k, G) = \sum_{i=1}^{k} w_{i,k} \beta(x \mid i, k - i + 1), \tag{1.2.6}$$

where $w_{i,k} = G(i/k) - G((i-1)/k)$ and $\beta(x \mid a, b)$ is a beta density function with parameters $a, b > 0$. Also, if $G(1) = 1$ it follows that $(w_{1,k}, \ldots, w_{k,k})$ is in the unit simplex

$$S_k = \left\{ (w_1, \ldots, w_k) \in [0,1]^k \ : \ \sum_{i=1}^{k} w_i = 1 \right\}. \tag{1.2.7}$$

Now, if we take G and k as random (distribution and positive integer, respectively), then $B(x \mid k, G)$ is called random Bernstein polynomial and the probability measure induced by B is called Bernstein prior.

The parameters in the Bernstein prior are the distribution function G and the positive integer k, and thus in principle we would need to specify a joint distribution for (k, G). As claimed by Petrone (1999a), it is sufficient to specify a probability function $p(k)$ for the hyperparameter k and a conditional finite-dimensional distribution of G, given k, at the points $(0, 1/k, \ldots, (k-1)/k)$, since $B(x \mid k, G)$ depends on G only through the values $(G(0), G(1/k), \ldots, G((k-1)/k))$. For instance, we can define a Bernstein prior as follows

$$\begin{aligned} F(x) &= B(x \mid k, G), \\ G \mid k &\sim \mathrm{DP}(\alpha_k, G_k^*), \\ k &\sim p(k), \end{aligned}$$

which is known as Bernstein–Dirichlet prior. Figure 1.2 depicts trajectories of random densities simulated using this prior.

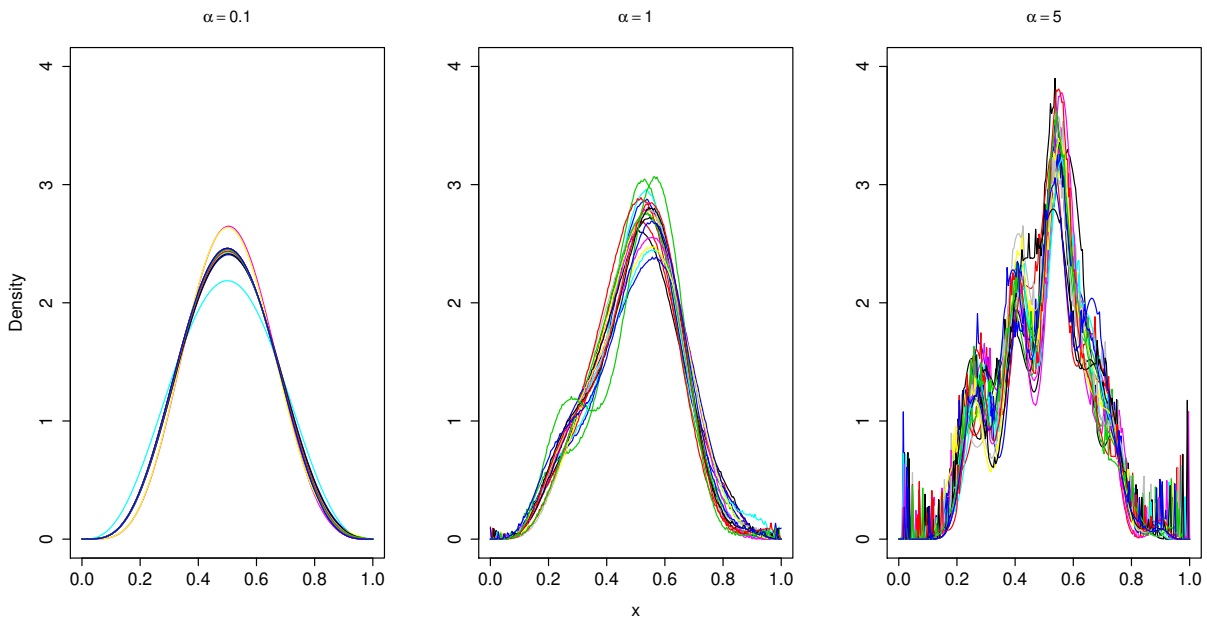


Figure 1.2: Trajectories of random densities sampled from a Bernstein–Dirichlet prior. Here $k \sim \text{Unif}\{1, \ldots, 1000\}$ and $G^*(\cdot) = \text{Beta}(\cdot \mid 1, 1)$

Selected Comments on Regression

A natural extension of the Dirichlet process for a regression context was introduced by MacEachern (2000); the said extension consists in a predictor-dependent version of the Dirichlet process, and it is obtained as a generalization of (1.2.2) by considering

$$H_x(\cdot) = \sum_{h=1}^{\infty} \omega_h(x) \delta_{\theta_h(x)}(\cdot). \tag{1.2.8}$$

Here, x is a covariate and the $\theta_h(x)$ are independent stochastic processes indexed over the covariate space $\mathcal{X} \subset \mathbb{R}^p$; for the weights we use the same construction as Eq. (1.2.3) but instead of $v_h \sim \text{Beta}(1, \alpha)$ now we use $v_h(x) \sim \text{Beta}(1, \alpha_x)$, for all x.

When we refer to dependent Dirichlet process (DDP) it means that both the atoms and weights are indexed by $\mathcal{X}$, i.e. as in Eq. (1.2.8), but other versions of the DDP can be

devised. If only the atoms are indexed by $\mathcal{X}$ then we refer to the process as a single weights DDP, and if only the weights are index by $\mathcal{X}$, then we refer to the process as a single atoms DDP.

Using the support properties in Theorem 4 of Barrientos et al. (2012), we consider a single weights mixing of the type

$$H_x(\cdot) = \sum_{h=1}^{\infty} \omega_h \delta_{\theta_h(x)}(\cdot). \tag{1.2.9}$$

In words, the latter theorem ensures that under some mild assumptions on the kernel, mixture models obtained from (1.2.8) have the same support as those obtained from (1.2.9). The weights $\{\omega_h\}$ in (1.2.9) match those from a standard DP.

Replacing the DP by the (single weights) DDP in (1.2.4) we can obtain a predictor-dependent Dirichlet process mixture that can be regarded as an infinite mixture of regression models as,

$$f(y \mid x) = \int_{\Theta} \mathbb{K}(y \mid \theta) \mathrm{d}H_x(\theta) = \sum_{h=1}^{\infty} \omega_h \mathbb{K}(y \mid \theta_h(x)). \tag{1.2.10}$$

For recent applications of (1.2.10) see for instance Inacio de Carvalho et al. (2016) and de Carvalho et al. (2019a). Other contributions in the regression context can be found in Müller et al. (1996), Poynor and Kottas (2019), to name a few.

I now switch gears and move towards preparations on functional data analysis.

1.3 Preparations on Functional Data Analysis

Context

The main contributions from Chapter 2 and 3 will require methods and concepts from functional data analysis (including amplitude and phase variation, and Karhunen–Loève decomposition), and thus I will review in this section background on these methods.

Functional data analysis (FDA) deals with the analysis and theory of data that are in the form of a function, that is, data that can be seen as a random sample of functions $X_1(t), \ldots, X_n(t)$ for t in some interval I (see, for example, Fig. 1.3). This kind of objects can be viewed as realizations of a one-dimensional stochastic process, which typically belongs to the Hilbert space L^2 and are intrinsically infinite-dimensional, so more flexibility is needed for modeling this type of objects.

Is is important to stress that in some contexts (such as in Chapter 3) the random functions of interest may not be "data" themselves but rather a functional parameter.

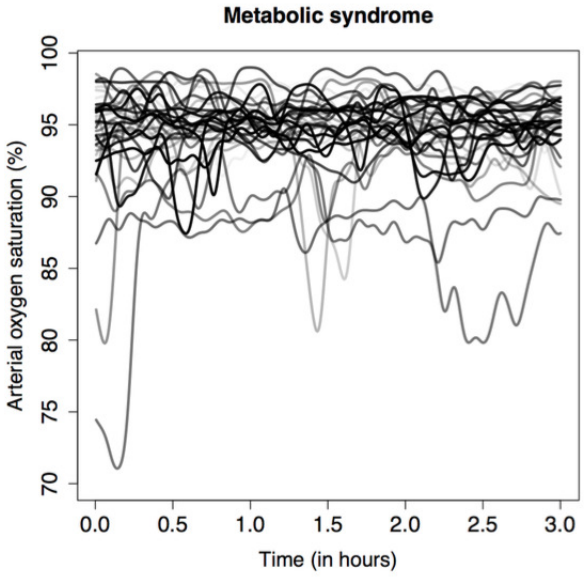


Figure 1.3: Levels of arterial oxygen saturation for women with metabolic syndrome, see Inácio de Carvalho et al. (2016)

The foundations on FDA can be traced back to Grenander (1950) and Rao (1958), but the term functional data analysis originates from Ramsay (1982) and Ramsay and Dalzell (1991). A general introduction to main results in this field can be found in the monographs of Ramsay and Silverman (2002a), Ramsay (2006) or Horváth and Kokoszka (2012).

FPCA and Karhunen–Loève Decomposition

The concepts to be introduced next are key for laying the groundwork of Chapter 3. I will mostly follow Horváth and Kokoszka (2012) over this section.

Principal component analysis is a widely applied dimension-reduction tool for multivariate data; the method has been extended for functional data and termed functional principal component analysis (FPCA). The roots of the idea were given by Grenander (1950), but a more comprehensive framework for statistical inference for FPCA was developed by Dauxois et al. (1982).

First, we introduce the general concept of covariance operator, its relationship with functional principal components, and some decompositions that are derived from those.

Remark 1. *All concepts in this section will be introduced assuming that $X(t)$ is a random function on $L^2(I)$ equipped with the Borel σ-algebra, $\mathbb{E}[X] = 0$ and satisfy that*

$$\mathbb{E}[\|X\|^2] = \mathbb{E}\left[\int_I X^2(t)\mathrm{d}t\right] < \infty. \tag{1.3.1}$$

Definition 3 (Covariance Operator)**.** *Let $X(t)$ be a random function obeying the conditions in Remark 1. The covariance operator of X is defined as*

$$C(Y)(t) = \mathbb{E}\left[\langle X, Y\rangle, X\right] = \int_I \mathbb{E}[X(t)X(s)]Y(s)\mathrm{d}s, \tag{1.3.2}$$

for any $Y \in L^2(I)$.

If $\mathbb{E}[X(t)] = \mu(t) \neq 0$, then the definition of covariance operator in (1.3.2) can be modified as

$$C(Y)(t) = \mathbb{E}\left[\langle X - \mu, Y\rangle(X - \mu)\right] = \int_I \mathbb{E}[(X(t) - \mu(t))(X(s) - \mu(s))]Y(s)\mathrm{d}s, \tag{1.3.3}$$

for $Y \in L^2(I)$.

The covariance operator allows to us introduce the concept of eigenfunction and eigenvalues as follows.

Definition 4 (Eigenvalues and Eigenfunctions)**.** *Let $X(t)$ be a random function obeying the conditions in Remark 1, and let $C(\cdot)(t)$ be its covariance operator. A function $h \in L^2(I)$ is the eigenfunction of $X(t)$, with an associated eigenvalue λ, if the following equation holds*

$$C(h)(t) = \lambda h(t), \tag{1.3.4}$$

for all $t \in I$.

It can be shown that the covariance operator is symmetric and non-negative definite, and so their eigenvalues must be non-negative. Thus, we can construct the spectral decomposition of C as

$$C(h)(t) = \sum_{k=1}^{\infty} \lambda_k \left(\int_I g_k(s)h(s)\mathrm{d}s \right) g_k(t), \tag{1.3.5}$$

where λ_k are the real-valued nonnegative eigenvalues in descending order and $g_k(t)$ are a basis or orthogonal eigenfunctions.

Using (1.3.5), Karhunen and Loève (Karhunen 1946; Loève 1946) independently discovered the functional principal component analysis (FPCA) expansion.

Definition 5 (Karhunen–Loève Decomposition)**.** *Let* $X(t)$ *be a stochastic process in* $L^2(I)$ *and let* $g_k(t)$ *be orthogonal eigenfunctions. The Karhunen–Loève decomposition is defined as*

$$X(t) = \mu(t) + \sum_{k=1}^{\infty} \theta_k g_k(t), \tag{1.3.6}$$

where

$$\theta_k = \langle X(t) - \mu(t), g_k(t) \rangle = \int_I \{X(t) - \mu(t)\} g_k(t)\mathrm{d}t. \tag{1.3.7}$$

If we truncate the series in right-hand-side of the (1.3.6) we have that

$$\sup_{t \in I} \mathbb{E}\left[X(t) - \mu(t) - \sum_{k=1}^{K} \theta_k g_k(t) \right] \to 0, \quad \text{as } K \to \infty.$$

Hence, when K is large enough the truncated version of (1.3.6) provides a good approximation, and thus we can work with

$$X(t) \approx \mu(t) + \sum_{k=1}^{K} \theta_k g_k(t). \tag{1.3.8}$$

The expansion in (1.3.8) can be constructed using another type of bases, such as splines, Fourier bases or wavelets, but what distinguishes the FPCA from other types of basis-expansions, is that FPCA explains most of the variation in X in the L^2 sense.

Now, we discuss the results described above but in the setting when we have a finite sample; specifically, consider $X_1(t), \ldots, X_n(t)$ a sample of random functions identically distributed with X. In this case, we can define the sample covariance operator as

$$\tilde{C}(Y)(t) = \frac{1}{n}\sum_{i=1}^{n}\langle X_i(t) - \mu(t), Y(t)\rangle(X_i(t) - \mu(t)), \tag{1.3.9}$$

where $\mu(t) = n^{-1}\sum_{i=1}^{n} X_i(t)$ and $Y \in L^2(I)$. Using this we can obtain the Karhunen–Loève decomposition as

$$X_i(t) = \mu(t) + \sum_{k=1}^{J}\theta_{i,k}g_k(t), \tag{1.3.10}$$

where $J \leq n$, $n\tilde{C}(g_k)(t) = \lambda_k g_k(t)$ and $\theta_{i,k} = \langle X_i(t) - \mu(t), g_k(t)\rangle$.

The $\theta_{i,k}$ in (1.3.10) are called scores, and the g_k functions are the principal components of X. The $\theta_{i,k}$ are independent across i for a sample of independent trajectories and are uncorrelated across k. Moreover, the scores $\{\theta_{i,k}\}$ obey the following conditions

$$\sum_i \theta_{i,k} = 0, \quad \sum_i \theta_{i,k}\theta_{i,s} \text{ if } k \neq s, \quad \sum_i \theta_{i,k}^2 = \lambda_k. \tag{1.3.11}$$

Karhunen–Loève Decomposition Estimation

In this section, I describe a method to learn about the functional components $\{g_k\}$ and the scores $\{\theta_{i,k}\}$ which involves a $n \times n$ matrix instead the sample covariance operator $\tilde{C}$.

Let $X_1(t), \ldots, X_n(t)$ be a random sample; our objective here is to find an estimator for the functional components $\{g_k\}$ and scores $\{\theta_{i,k}\}$ as described in (1.3.10). To achieve this,

we can construct the $n \times n$ matrix $M = (M_{s,k})$ with entries

$$M_{s,k} = \langle X_s(t) - \mu(t), X_k(t) - \mu(t) \rangle.$$

This matrix has the same nonzero eigenvalues as the sample covariance operator $\tilde{C}$ (Good 1969); and also its eigenvectors have a close relationship with the scores $\{\theta_{i,k}\}$. Indeed, if we consider the eigenvector $p_k = (p_{1k}, \ldots, p_{nk})$ associated to the nonzero eigenvalue λ_k we have that

$$\theta_{i,k} = \sqrt{\lambda_k} p_{ik}.$$

Since $n\tilde{C}(g_k)(t) = \lambda_k g_k(t)$, then

$$g_k(t) = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{n} p_{ik}(X_i(t) - \mu(t)) = \frac{\sum_{i=1}^{n} \theta_{i,k} X_i(t)}{\sum_{i=1}^{n} \theta_{i,k}^2}.$$

Therefore, we can be able to obtain estimators for $\{g_k(t)\}$ and $\{\theta_{i,k}\}$ only using the eigenvalues and eigenvector of matrix M.

Amplitude and Phase Variation in FDA

Chapter 2 will be focus on ideas, concepts, and methods from amplitude and phase variation, and thus I will offer below some background on the subject.

In the context of random functions over some compact domain, the analysis of variation of this function may entail two sources of variation: the first one is about fluctuations around the mean level, that is, the variation in the y-axis direction; this type of variation is known as amplitude variation. This variation is commonly found in multivariate analysis, so it is more natural to understand and therefore, deal with it.

The second source refers to variations in the domain of the random function, that is, continuous deformations or changes in the x-axis direction; this is known as phase variation

and the continuous deformation is known as warp function. Phase variation can be seen as a composition of the stochastic process with a random transformation which is defined in the domain of the process.

In Fig. 1.4 we depict an example of amplitude or/and phase variation for one particular function.

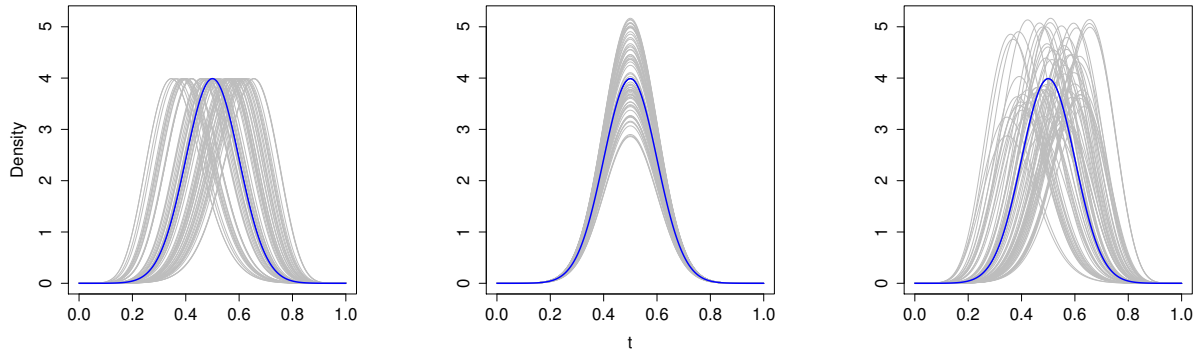


Figure 1.4: Amplitude and phase variation. The blue line corresponds to the original function and the gray lines correspond to functions with phase variation (left), amplitude variation (center) and both types of variation (right).

Definition 6 (Warp Map)**.** *Let T be a function from a compact interval on itself. We say that T is a warp map if it is a strictly increasing homeomorphism.*

Different types of warp maps can be defined (Marron et al. 2015a) but here we will focus on the case where the warp maps correspond to a homeomorphism from a compact interval on itself. Some examples of warp maps obeying Definition 6 can be seen in Fig. 1.5; similar warp maps will be generated in Chapter 2 and thus I will skip here the details on the exact setup used for simulating these. With this concept of warp function, we can define the warped process and thus formalize the concept of phase variation in a random function.

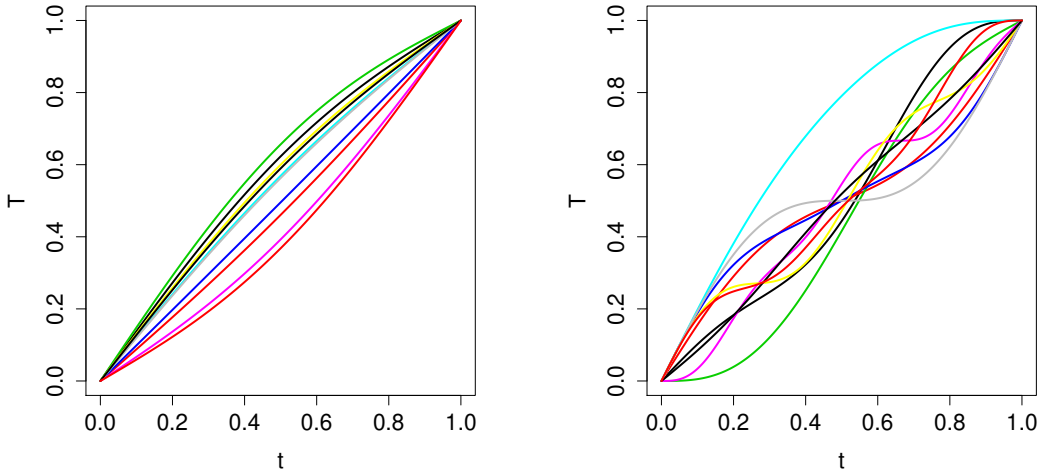


Figure 1.5: Trajectories of two different types of warp maps as in Definition 6.

Definition 7 (Warped Stochastic Processes)**.** *Let $X(t)$ be a random function defined on I and let T be a strictly increasing random homeomorphism from I. The warped stochastic process $\tilde{X}(t)$ is defined as*

$$\tilde{X}(t) = X \circ T(t) = X(T(t)). \tag{1.3.12}$$

Notice that the fact that T is a strictly increasing function means that it generates a distortion of time but preserves its order.

And how do these concepts relate with the Karhunen–Loève decomposition in (1.3.6)? To see the connection, consider a sample of independent realization of the process $X(t)$, say $X_1(t), \ldots, X_n(t)$. A Karhunen–Loève decomposition of each random function yields

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \theta_{i,k} g_k(t). \tag{1.3.13}$$

Eq. (1.3.13) leads to a better understanding of the concept of amplitude variation, as it implies that the X_i are varying around the mean $\mu(t)$ by amplitude oscillations of the functions

$g_k(t)$.

Of course, that is true if we do not have any phase variation present in the $X_i(t)$; if we have, then we need first remove it and this can be achieved as follows. Suppose that every $X_i(t)$ was been warped by an homeomorphism $T_i(t)$; then the unwarped process is

$$\tilde{X}_i(t) = X_i(T_i^{-1}(t)) = \mu(T_i^{-1}(t)) + \sum_{k=1}^{\infty} \theta_{i,k} g_k(T_i^{-1}(t)), \tag{1.3.14}$$

which has the same interpretation as above in the sense of amplitude variation but note that Eq. (1.3.14) is a no longer Karhunen–Loève decomposition of $X_i(t)$, as the functions $g_k(T_i^{-1}(t))$ are no longer eigenfunctions of X_i.

1.4 Selected Comments on Point Processes

I close the preparations for this thesis by recalling basics on point processes, and on how the concept of phase variation can be adapted for point processes, so to facilitate reading Chapter 2. The goal is not provide an encyclopedic account of the topic, but rather to review concepts required for Chapter 2. The theory has its roots in Poisson (1837) where the Poisson process was introduced. Since then a lot of developments have been made, which are revised in the monographs by Daley and Vere-Jones (2003/2008). The definition of point process is as follows.

Definition 8 (Point Process)**.** *Let $S \subset \mathbb{R}^k$ and suppose that for any $A \subset S$, $N(A)$ is a non negative integer-valued random variable. Then $N(\cdot)$ is called a point process (on S) if $N(\phi) = 0$ and $N(A \cup B) = N(A) + N(B)$ for any pair of disjoint sets A and B.*

A celebrated type of point process is the so-called Poisson point process. Formally, a point process $\Pi(\cdot)$ defined on $S \subset \mathbb{R}^k$ is called a homogeneous Poisson process if

(a) For any $A = [a_1, b_1] \times \ldots \times [a_k, b_k] \subset S$, then

$$\Pi(A) \sim \text{Poisson}\left(\lambda \prod_{i=1}^{k} (b_i - a_i)\right), \tag{1.4.1}$$

for some $\lambda > 0$.

(b) For any two disjoint sets A and B, then $\Pi(A) \perp \Pi(B)$.

The parameter λ in (1.4.1) is called intensity parameter of the Poisson process. It is clear that

$$\lambda \prod_{i=1}^{k} (b_i - a_i) = \int_A \lambda \mathrm{d}t.$$

When the intensity parameter becomes a function, say $\lambda(t)$, then the process is called non-homogeneous Poisson process. Therefore

$$\Pi(A) \sim \text{Poisson}\left(\Lambda(A)\right), \quad \Lambda(A) = \int_A \lambda(t)\mathrm{d}t. \tag{1.4.2}$$

Let's now introduce the idea of amplitude and phase variation in a point process context. Amplitude variation, as we mentioned in Section 1.3, refers to fluctuations around the mean level; in a point process setting, this can be understood through the covariance operator

$$\mathcal{C}(A \times B) = \mathbb{E}[\Pi(A)\Pi(B)] \ - \Lambda(A)\Lambda(B). \tag{1.4.3}$$

This covariance operator keeps track of the second order fluctuations of $\Pi(A)$ around its mean $\Lambda(A)$, and also their dependence on the corresponding fluctuations $\Pi(B)$ over $\Lambda(B)$. In this line, we can obtain this Karhunen–Loève decomposition as in (1.3.6) as follows

$$\Pi([0, t]) = \Lambda([0, t]) + \sum_{n=1}^{\infty} \eta_n \psi_n(t).$$

This expression—as in Section 1.3—can be used for appreciating that phase variation in a point process also involves assessing its variations around a mean level. Thus, phase-variation on a point process is equivalent to phase-variation—in the sense described in Section 1.3—over the mean measure $\Lambda(\cdot) = \mathbb{E}[\Pi(\cdot)]$. To see this we will need to introduce some notation. Let Π be a random point process over $[0, 1]$ and let T be a random homeomorphism in the same interval; then we define the warped version of Π, given T, as $\tilde{\Pi}$, where

$$\mathbb{E}[\tilde{\Pi}(\cdot) \mid T] = \Lambda[T^{-1}(\cdot)].$$

Fig. 1.6 depicts an example with realizations of a phase-varying point process along with their corresponding warp maps; similar processes will be generated in Chapter 2 and thus I will skip here details on the exact setup used for simulating these data. This figure depicts realizations of the original point process, their corresponding phase-varying point process along with their corresponding registered versions as obtained using the method proposed in the manuscript; details on the underlying processes can be found in Section 2.3.

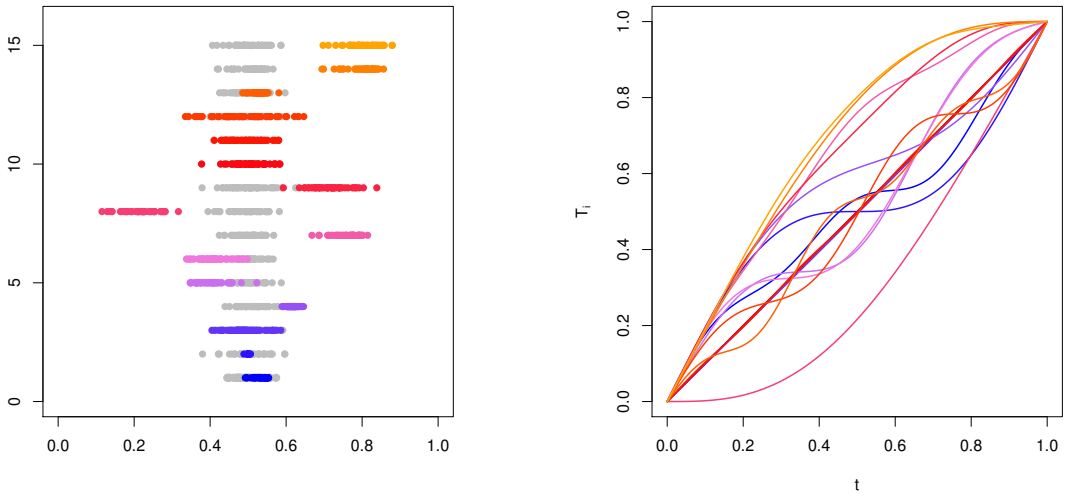


Figure 1.6: Realizations of the original point process (left, gray), this corresponding varying point process (left, colored), along with their corresponding warp maps (right), plotted in the same color palette

The framework of interest in Chapter 2 will require phase-variation over several realization of the point process, which are warped versions of the original one, so we now make that setup precise. The following definition follows from Panaretos and Zemel (2016).

Definition 9 (Phase-Variation on i.i.d. Copies of a Point Process)**.** *Let* Π *be a random point process with* $\mathbb{E}[\Pi(\cdot)] = \Lambda(\cdot)$*. Let* $\Pi_1, \ldots, \Pi_n$ *be* n *realizations of* Π *and let* $T_1, \ldots, T_n$ *be strictly increasing random homeomorphisms satisfying* $\mathbb{E}[T_i(t)] = t$*. Then we define the warped version of* $\{\Pi_1, \ldots, \Pi_n\}$ *by* $\{\tilde{\Pi}_1, \ldots, \tilde{\Pi}_n\}$*, where*

$$\tilde{\Pi}_i(\cdot) = T_i \,_{\#}\Pi_i(\cdot) = \Pi_i(T_i^{-1}(\cdot)).$$

All primitive concepts for understanding the following chapters of this dissertation have now been introduced. To complete this introductory chapter, we now to move on to the

main contributions and the structure of this dissertation.

1.5 Main Contributions of this Thesis

This dissertation will propose and study novel flexible Bayesian models for setups that entail families of random densities. Two specific contexts will be examined here: one involves phase-varying point processes—in the sense introduced in Section 1.4; the other involves functional principal component analysis—as introduced in Section 1.3. The common denominator underlying these contexts is the need to model families of random measures to each of which corresponds a different data generating process. On both contexts, prior processes (Section 1.2) will be used so to devise priors on the target objects of interest. The main contribution of this thesis is documented in Chapters 2 and 3, which aim to deliver the following work packages:

Package 1: Aligning Multiple Phase-Varying Point Processes

In the framework of functional data analysis, several developments have been made on amplitude and phase variation (see for instance Srivastava et al. (2011), Marron et al. (2015a)), but in a point processes framework the field is still in its infancy (Panaretos and Zemel 2016; Xu et al. 2017). Package 1 will focus on modeling phase variation in a multiple point process setting by resorting to the setup in Panaretos and Zemel (2016); see Definition 9. Specifically, I develop induced priors for warp maps via a Bernstein polynomial prior so to learn about the structural measure of the point process and about the phase variation in the process. In this framework, Package 1 pioneers the development of priors on spaces of warping maps by proposing a novel Bayesian approach for modeling registration of multiple point processes. Theoretical properties of the induced prior for the warp maps, including support and posterior consistency, are established under a fairly mild proviso.

Package 2: Borrowing Strength over a Family of Random Densities

Package 2 is instigated by the following question: "How to convert a prior on a space of densities (e.g. DPM) into a prior on a family of densities—that could borrow strength across samples?" Our starting point to answer this question stems from the setup of Kneip and Utikal (2001). Specifically, Package 2 will aim to model families of random densities using functional principal component analysis through the so-called Karhunen–Loève decomposition; see Section 1.3.6. The proposed approach defines a prior on the space of families of densities. Theoretical properties are developed to ensure that the trajectories from an infinite mixture belong to L_2 which is a necessary condition for the Karhunen–Loève decomposition to hold, and also to guarantee the full L_1 support of the proposed prior.

1.6 Structure and Organization

The master-plan for this thesis is as follows. Chapters 2 and 3 contain the proposed solutions for problems described in Packages 1 and 2, respectively. These chapters are self-contained in terms of notations, definitions, and results. For the convenience of the reader, some parts mentioned in the Introduction may be repeated on later chapters. In detail:

- Chapter 2 develops a novel semiparametric model over the space of warp functions and, therefore, registration (aligning) of the multiple warped point processes. The main contribution of this chapter relies on the construction of prior over a space of random warp functions, obtain some key theoretical results such as full support and posterior consistency and a real data application of our model in climatology. Also, numerical experiments and simulation studies were conducted to assess the performance of our method.

- Chapter 3 develops a novel data-driven prior to modeling and borrowing strength across

an entire family of random densities. The main contribution of this chapter is to construct a data-driven prior which satisfies some desirable properties and performs better than other alternatives for families of random densities such as the dependent Dirichlet mixtures. Indeed, simulation studies were conducted to assess the performance and compare with competitors such as DPM and DDPM. Finally, the chapter ends with an illustration using Galton's parents height dataset.

Chapter 4, offers details about computing and implementations as the structure of simulations studies and sampling methods. Also, we comment on the implementations and technical specifications of the instance used in the Google Cloud Platform to achieve the results. We close the thesis in Chapter 5, by putting the main developments in perspective and by discussing directions for future research.

CHAPTER 2

Bayesian Semiparametric Modeling of Phase-Varying Point Processes†

We propose a Bayesian semiparametric approach for modeling registration of multiple point processes. Our approach entails modeling the mean measures of the phase-varying point processes with a Bernstein–Dirichlet prior, which induces a prior on the space of all warp functions. Theoretical results on the support of the induced priors are derived, and posterior consistency is obtained under mild conditions. Numerical experiments suggest a good performance of the proposed methods, and a climatology real-data example is used to showcase how the method can be employed in practice. Appendix A includes supplementary materials

† Joint work with Y. Zemel and M. de Carvalho.

2.1 Introduction

A prototypical characteristic in the analysis of a random function $X(t)$—that distinguishes it from classical multivariate analysis—is that it potentially exhibits two distinct layers of stochastic variability. Amplitude variation is encapsulated in the fluctuations of $X \equiv X(t)$ around its mean function $\mu(t)$, and can be probed by linear tools, perhaps most prominently the covariance operator of X and the subsequent Karhunen–Loève expansion. Phase variation amounts to variability in the argument t, usually modeled by a random warp function T defined on the domain of definition of X, so that one observes realizations (discretized over some grid) from the random function $\widetilde{X}(t) = X(T^{-1}(t))$ instead of $X(t)$. In short, phase variation is randomness in the t-axis, whereas amplitude variation pertains to stochasticity in the X-axis.

Typically, one is interested in inferring properties of the original function X, rather than those of $\widetilde{X}$. In such situations phase variation can be thought of as a nuisance parameter, and failing to account for it may result in a severely distorted statistical analysis: the mean function and Karhunen–Loève expansion of $\tilde{X}$ are smeared and less informative than those of X. Consequently, one needs to undo the warping effect of the phase variation by constructing estimators $\widehat{T}$ for the warp functions, and composing them with the observed realizations from $\widetilde{X}$, a procedure known as registration, or alignment, of the functions. The registered functions $\widetilde{X}_i \circ \widehat{T}_i = X_i \circ T_i^{-1} \circ \widehat{T}_i$ are then treated as distributed approximately as X, allowing for their use in probing the law of X. For a textbook treatment on phase variation, we refer to the books by Ramsay and Silverman (2002b, 2005); one may also consult the review articles Marron et al. (2015b) and Wang et al. (2016).

In this chapter, we propose a Bayesian method for registering phase-varying point processes. Our work is aligned with recent developments focused on modeling phase and amplitude variation of complex objects that are not functional data per se, yet still carry infinite-

dimensional traits. An intriguing example is indeed that of point processes, appearing as spike trains in neural activity (e.g. Wu and Srivastava 2014), where phase variation can be viewed as smearing locations of peaks of activity. See Fig. 2.1 for an example of such phase-varying point processes (and Section 2.3 for more details on the underlying processes). Such data can be transformed into functional data by smoothing and considering density functions (Wu et al. 2013), but can be also be dealt with directly, replacing the ambient space L^2 used for functional data by a space of measures. Indeed, Panaretos and Zemel (2016) formalize the problem and show how the Wasserstein metric of optimal transport arises canonically in the point process version of the problem. Here we aim to devise a Bayesian model that is both flexible and adapted to the warping problem in a point process setting in the sense our priors for the warp functions obey the same classical phase variation assumptions of FDA (functional data analysis). From a conceptual viewpoint, our model can be regarded as a semiparametric Bayesian version of Panaretos and Zemel (2016), but by putting directly a prior on the space of all random measures on the unit interval it allows for straightforward inference from posterior outputs—both in terms of credible bands for warp functions, and credible intervals for registered points. By modeling the mean measure of each phase-varying point process with a random Bernstein polynomial (Petrone 1999a,b), we are able to show that the support of the induced priors for the warping functions and collections of registered points is 'large' in the sense made precise in Section 3.2.4–2.2.4. Posterior consistency is established under a proviso that is asymptotically equivalent to that of Panaretos and Zemel (2016), but our large sample results only require the number of points in each process to increase.

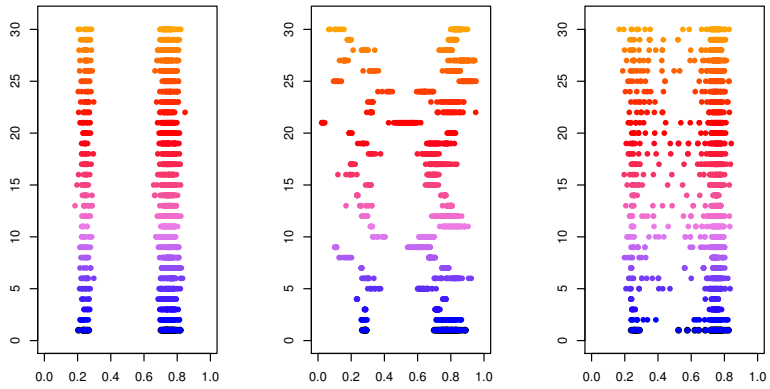


Figure 2.1: Realizations of the original point process (Left), their corresponding phase-varying point process (Middle) along with their corresponding registered versions as obtained using the method proposed in the manuscript (Right); details on the underlying processes can be found in Section 2.3. Appendix A includes supplementary materials.

Section 2.2 develops details of our approach, in Section 2.3 we report numerical experiments, and Section 2.4 includes a climatology real-data example. Concluding remarks are given in Section 2.5. Proofs of results characterizing the prior and limiting posterior can be found in the Section 2.6.

2.2 Random Bernstein Polynomial-Based Registration of Multiple Point Processes

2.2.1 Random Bernstein Polynomials

Random Bernstein polynomials were introduced by Petrone (1999a,b) and are defined as

$$B(t \mid k, G) = \sum_{i=0}^{k} G\left(\frac{i}{k}\right) \binom{k}{i} t^i (1-t)^{k-i}, \tag{2.2.1}$$

where G is a random function on $[0,1]$ and k is a (positive) integer-valued random variable. It is clear that when G is a distribution function, so is $B(t \mid k, G)$, and if in addition $G(0) = 0$ then $B(t \mid k, G)$ has a density given by

$$b(t \mid k, G) = \sum_{i=1}^{k} w_{i,k} \beta(t \mid i, k-i+1), \tag{2.2.2}$$

where $w_{i,k} = G(i/k) - G((i-1)/k)$ and $\beta(t \mid a, b)$ is a beta density function with parameters $a, b > 0$. Since $G(1) = 1$ it follows that $(w_{1,k}, \ldots, w_{k,k})$ is in the unit simplex $S_k = \{(w_1, \ldots, w_k) \in [0,1]^k : \sum_{i=1}^{k} w_i = 1\}$; if G has a continuous density g, then $b(t \mid k, G)$ approximates g uniformly as $k \to \infty$ (see Lemma 3). Following Petrone (1999a,b) we have the next definition.

Definition 10. *The probability measure π induced by B in (2.2.1), on the set Δ of all continuous distribution functions defined on $[0,1]$, is called Bernstein prior with parameters (k, G). In symbols, $\pi \equiv \pi(k, G)$.*

Further details on random Bernstein polynomials can be found in Ghosal and Van der Vaart (2015, Section 5.5). To avoid unnecessarily burdening notation, measure-theoretical considerations will be kept to a minimum (including the measures with respect to which expected values are defined).

2.2.2 Bayesian Semiparametric Inference for Phase-varying Point Processes

Let Π be a point process in $[0,1]$, with finite second moment: $E\{(\Pi[0,1])^2\} < \infty$ and denote its (finite) mean measure by $\lambda(\cdot) = E\{\Pi(\cdot)\}$. Estimation of λ (say $\widehat{\lambda}$) is straightforward when one has access to multiple realizations $\{\Pi_1, \ldots, \Pi_n\}$ from Π, with $\widehat{\lambda}$ asymptotically normal (Karr 1991, Proposition 4.8). Suppose, however, that one instead observes a sample

$\{\widetilde{\Pi}_1, \ldots, \widetilde{\Pi}_n\}$ with

$$\widetilde{\Pi}_i = T_{i\#}\Pi_i,$$

where $T_{i\#}\Pi_i(\cdot) = \Pi_i\{T_i^{-1}(\cdot)\}$ denotes the push-forward of Π_i through T_i, for all i. In other words, if a given realization of Π_i is the collection of points $\{x_{i,j}\}_{j=1}^{m_i}$, then one observes the deformed collection $\{\widetilde{x}_{i,j}\}_{j=1}^{m_i} \equiv \{T_i(x_{i,j})\}_{j=1}^{m_i}$, for all i. Here, $\{T_1, \ldots, T_n\}$ is a sequence of random warp functions, that is, increasing homeomorphisms on $[0, 1]$. A target of interest will be on learning about the warp functions, so to register the point processes. To achieve this goal we model the (conditional) mean measures of the phase-varying point processes with a Bernstein–Dirichlet prior, which induces a prior on the space of all warp functions. The conditional mean measure of the warped version $\widetilde{\Pi}_i$ given T_i is $\Lambda_i(\cdot) = E\{\widetilde{\Pi}_i(\cdot) \mid T_i\}$, for all i. We impose the rather standard assumptions that $E\{T_i(t)\} = t$ (unbiasedness) for all $t \in [0, 1]$, and that the collection $\{T_1, \ldots, T_n\}$ is independent of $\{\Pi_1, \ldots, \Pi_n\}$; the assumptions of unbiasedness and monotonicity of warp functions are *sine qua non* in the classical FDA phase variation literature, often accompanied with additional conditions (e.g. Tang and Müller 2008; Wang et al. 2016). In words, the assumption is tantamount to requiring that the average time change $E[T(x)]$ to be the identity: on average, the "objective" time-scale should be maintained, so that time is not sped up or slowed down. The assumptions of unbiasedness and monotonicity are key for identifiability. Finally, it follows that the conditional mean measure of the warped version $\widetilde{\Pi}_i$ given T_i is $\Lambda_i(\cdot) = E\{\widetilde{\Pi}_i(\cdot) \mid T_i\}$, for $i = 1, \ldots, n$.

To learn about $F_i(t) = \int_0^t \Lambda_i(\mathrm{d}x)$, for $t \in [0, 1]$, we set the prior

$$F_i(t) = B(t \mid k_i, G_i), \quad t \in [0, 1], \tag{2.2.3}$$

where $\{k_1, \ldots, k_n\}$ is a sequence of independent integer-valued random variables and $\{G_1, \ldots, G_n\}$ is a sequence of independent random measures. In a more concrete specification of (2.2.3), we proceed as follows. Let $\{\widetilde{x}_{i,j}\}_{j=1}^{m_i}$ be the points corresponding to $\widetilde{\Pi}_i$, and for $i = 1, \ldots, n$

we set

$$\begin{aligned} \widetilde{x}_{i,j} \mid F_i \sim F_i, \quad j = 1, \ldots, m_i, \quad F_i(t) = B(t \mid k_i, G_i), \\ G_i \mid \alpha \sim \mathrm{DP}(\alpha, G^*), \quad k_i \sim \rho(k), \end{aligned} \tag{2.2.4}$$

where ρ is a probability function over $\mathbb{N}$. Here 'DP' stands for Dirichlet process (Ferguson 1973), with precision parameter $\alpha > 0$ and centering distribution $G^* = E(G_i)$. To complete the model specification we set $G^* = \mathrm{Beta}(a_0, b_0)$ and $\alpha \sim \mathrm{Gamma}(a_0, b_0)$, for $i = 1, \ldots, n$. More sophisticated versions of (2.2.4) would entail specifying a different precision and centering for the DP per each point process; for simplicity, we will focus on (2.2.4). Below, we assume that the $\{G_i\}$ and $\{k_i\}$ are independent.

Now, $\{F_1, \ldots, F_n\}$, specified as in (2.2.3), can be used to induce a prior F on the mean measure λ of the random point process Π and on the warp maps T_i. The prior F will be centered around the structural mean λ in the Fréchet mean sense that λ is the closest to F in expectation, that is, $E_\lambda\{d^2(\lambda, F)\} \leq E_\lambda\{d^2(\gamma, F)\}$, for all diffuse measures γ on $[0, 1]$. An obvious question that arises is what metric d should one use, but Wasserstein distance (Santambrogio 2015; Panaretos and Zemel 2019) has been shown to be the canonical metric for phase-varying point processes by Panaretos and Zemel (2016, Section 3):

$$d(\mu, \nu) = \inf_{Q \in \Gamma(\mu,\nu)} \sqrt{\int_0^1 \{Q(x) - x\}^2 \mu(\mathrm{d}x)}. \tag{2.2.5}$$

By abuse notation below we identify a measure μ with its distribution function $F_\mu(t) = \mu\{[-\infty, t]\}$. Here $\Gamma(\mu, \nu)$ is the collection of functions $Q : [0, 1] \to [0, 1]$ such that $Q_\# \mu = \nu$. (If μ is not diffuse, then $\Gamma(\mu, \nu)$ may be empty and the definition of d needs to be modified, but we will only have to deal with diffuse measures in the sequel.) Since Fréchet averaging with respect to Wasserstein distance amounts to averaging of quantile functions (Agueh and Carlier 2011), the prior on F is induced from the prior on $\{F_1, \ldots, F_n\}$ as the probability

law of

$$F(t) = \left(\frac{1}{n} \sum_{i=1}^{n} F_i^{-1} \right)^{-1} (t), \quad t \in [0,1]. \tag{2.2.6}$$

The random Bernstein polynomial-induced prior on each T_i defines the optimal transport map of F onto F_i (Santambrogio 2015):

$$T_i = F_i^{-1} \circ F. \tag{2.2.7}$$

Since $F_1, \ldots, F_n$ are independent, identically distributed and increasing distribution functions, it follows that the T_i are homeomorphisms with $E\{T_i(t)\} = t$. Indeed, by construction it can be shown that $T_1(t) + \cdots + T_n(t) = nt$ for every t, $T_1, \ldots, T_n$ are identically distributed given F and so, $E(T_i \mid F) = E(T_{i'} \mid F)$ for every $i \neq i'$, and taking expectation in both sides, we have that $E(T_i) = E(T_{i'})$; therefore,

$$nE\{T_i(t)\} = E\{T_1(t)\} + \cdots + E\{T_n(t)\} = E\{T_1(t) + \cdots + T_n(t)\} = nt, \tag{2.2.8}$$

and thus it follows that $E\{T_i(t)\} = t$, for $i = 1, \ldots, n$.

The random Bernstein polynomial-induced priors on the registered point processes is constructed by pushing them forward through the registration maps

$$\Pi_i = T_{i\#}^{-1} \widetilde{\Pi}_i, \quad i = 1, \ldots, n. \tag{2.2.9}$$

The posterior sampling for the warping maps and registered points is then conducted as follows. Let $F_{i,[1]}, \ldots, F_{i,[M]}$ be posterior samples from F_i, for $i = 1, \ldots, n$, which can be obtained by Gibbs sampling as described in Ghosal and Van der Vaart (2015, Section 5.5); then, for each $j = 1, \ldots, M$ we get $F_{[j]} = (\sum_{i=1}^{n} F_{i,[j]}^{-1}/n)^{-1}$ and so, $T_{i,[j]} = F_{i,[j]}^{-1} \circ F_{[j]}$ and $\Pi_{[j]} = T_{i,[j]}^{-1} {}_{\#} \widetilde{\Pi}_i$. Finally, pointwise estimation for mean measure, warp functions, and

registered points are given by the posterior means,

$$\widehat{F} = \frac{1}{M}\sum_{j=1}^{M} F_{[j]}, \quad \widehat{T}_i = \frac{1}{M}\sum_{j=1}^{M} T_{i,[j]}, \quad \widehat{\Pi}_i = \frac{1}{M}\sum_{i=1}^{M} \Pi_{i,[j]}. \tag{2.2.10}$$

Credible intervals or pointwise credible bands can be also directly obtained from the relevant quantiles of the corresponding posterior outputs.

2.2.3 Kolmogorov–Smirnov, Wasserstein, and Kullback–Leibler supports of induced priors

As it will be shown below, full support of the relevant parameters in our setup holds, under conditions on the support of the law of the k_i and on that of $w_{1,k_i}, \ldots, w_{k_i,k_i} \mid k_i$. Extending the assumptions in Petrone (1999a), we assume that the prior probability function of k_i is positive, that is $p_i(k) > 0$ for $i = 1, \ldots, n$, and that $w_{1,k_i}, \ldots, w_{k_i,k_i} \mid k_i$ has a family of conditional densities $l_i(w_{1,k_i}, \ldots, w_{k_i,k_i} \mid k_i) > 0$, for every $(w_{1,k_i}, \ldots, w_{k_i,k_i}) \in S_{k_i}$ and for every sequence of independent integer valued random variables $\{k_1, \ldots, k_n\}$. Define the supremum norm

$$\|F - H\|_\infty = \sup_{t\in[0,1]} |F(t) - H(t)|.$$

Below, $\mathscr{F} \equiv (F_1, \ldots, F_n)$ denotes the joint Bernstein prior and $N_i \equiv \Pi_i([0,1]) > 0$ is the total number of points in the ith point process, for $i = 1, \ldots, n$.

Theorem 2. *Let $F_1, \ldots, F_n \overset{\text{iid}}{\sim} \pi$ with Fréchet–Wasserstein mean F, and with induced priors T_i and Π_i as defined in (2.2.7) and (2.2.9). For any continuous strictly increasing $\mathbb{F}_1, \ldots, \mathbb{F}_n \in \Delta$, with Fréchet–Wasserstein mean $\mathbb{F}$, transport maps $\mathbb{T}_i = \mathbb{F}_i^{-1} \circ \mathbb{F}$, and registered discrete measures $P_i = \mathbb{T}_{i\#}^{-1}\widetilde{\Pi}_i$, and for any $\varepsilon > 0$ the following events occur with*

positive probability:

$$(a)\ \{\mathscr{F} : \|F_j - \mathbb{F}_j\|_\infty < \varepsilon, j = 1, \dots, n\}, \qquad (b)\ \{\mathscr{F} : \|F - \mathbb{F}\|_\infty < \varepsilon\},$$
$$(c)\ \{\mathscr{F} : \|T_i - \mathbb{T}_i\|_\infty < \varepsilon\}, \qquad (d)\ \{\mathscr{F} : d(\Pi_i/N_i, P_i/N_i) < \varepsilon\},$$

for $i = 1, \dots, n$.

Claims (a), (b), and (c) in the Theorem 2 respectively state that the joint Bernstein prior, the Fréchet–Wasserstein mean, and the warp functions have large Kolmogorov–Smirnov support. Claim (d) states that the registered point processes have large Wasserstein support. The proof actually shows that the intersection of these four events (a)–(d) has positive probability. While the latter properties may not look surprising ex-post, as their proofs show, they are not straightforward facts.

The characterization of the Kullback–Leibler (KL) support is more challenging. By definition, a density f is said to possess the Kullback–Leibler (KL) property relatively to a prior π if for any $\varepsilon > 0$ one has that $\pi\{H : \mathrm{KL}(F, H) < \varepsilon\} > 0$, where

$$\mathrm{KL}(F, H) = \int_0^1 h(t) \log \frac{h(t)}{f(t)} \, \mathrm{d}t,$$

with F and H denoting the distribution functions respectively corresponding to f and h.

Random Bernstein polynomials satisfy the Kullback–Leibler property (Petrone and Wasserman 2002, Theorem 2). The following theorem inspects the permanence of the Kullback–Leibler property on the functionals of interest, and it shows that the property is preserved for Fréchet–Wasserstein mean and the warping functions.

Theorem 3. *Let* $F_1, \dots, F_n \overset{\text{iid}}{\sim} \pi$ *with Fréchet–Wasserstein mean* F *and with transport maps* $T_i = F_i^{-1} \circ F$ *as defined in* (2.2.7). *For any* $\varepsilon > 0$ *and strictly increasing* $\mathbb{F}_1, \dots, \mathbb{F}_n \in \Delta$ *with densities* $\mathbb{f}_i$ *that are continuous on* $(0, 1)$, *Fréchet–Wasserstein mean* $\mathbb{F}$ *and transport maps* $\mathbb{T}_i = \mathbb{F}_i^{-1} \circ \mathbb{F}$, $\mathbb{F}$ *also has a density* $\mathbb{f}$ *and:*

(a) If $\int_0^1 \mathbb{f}(x) \log \mathbb{f}(x) \mathrm{d}x < \infty$ *then* $KL(F, \mathbb{F}) < \varepsilon$ *with positive probability.*

(b) If each $\mathbb{f}_i$ *is strictly positive on* $(0, 1)$*, then with positive probability* $KL(T_i, \mathbb{T}_i) < \varepsilon$ *for all* $i = 1, \ldots, n$*.*

Remark 2. *The densities* $\mathbb{f}_i$ *can be unbounded or approach zero near 0 or 1. The condition* $\int_0^1 \mathbb{f}(x) \log \mathbb{f}(x) \mathrm{d}x < \infty$ *in (a) is very weak and is satisfied when* $\mathbb{f}$ *is a beta density with arbitrary (positive) parameters. This condition is, in fact, necessary; if it fails to hold, then* $KL(F, \mathbb{F}) = \infty$ *almost surely. The assumptions on the densities can be further relaxed to* $\mathbb{f}_i$ *having finitely many discontinuity points on* $[0, 1]$*, and for part (b)* $\mathbb{f}_i$ *may vanish on finitely many points on* $[0, 1]$*. We refrained from this level of generality for the purpose of clarity and because the current version includes the most important case of beta distributions.*

Theorem 3 shows that under mild conditions, the Fréchet–Wasserstein mean and the warping functions possess the Kullback–Leibler property with respect to the prior on F induced from $F_1, \ldots, F_n$ via (2.2.6). We now study the large-sample behavior of the posterior.

2.2.4 Posterior consistency

Contrarily to Panaretos and Zemel (2016, Theorem 1), our asymptotic theory does not require $n \to \infty$; indeed we only require that as $m_i \to \infty$, with $i = 1, \ldots, n$, for any finite n. Yet note that the consequence is that under this assumption one is only able to approximate warping functions of the type $\mathbb{T}_i = \mathbb{F}_i^{-1} \circ \mathbb{F}$, for all i, where $\mathbb{F}$ is the Fréchet–Wasserstein mean of $\mathbb{F}_1, \ldots, \mathbb{F}_n$. This proviso is less and less restrictive as n increases, and it is asymptotically compatible with that of Panaretos and Zemel (2016), as indeed if the $\mathbb{T}_i$ are independent and identically distributed—rather than fixed as assumed in Theorem 3—then it follows that as $n \to \infty$,

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{T}_i(t) \underset{p}{\to} E\{\mathbb{T}_1(t)\} = t.$$

The following result holds.

Theorem 4. *Under the same conditions as in Theorem 3, if $m_i \to \infty$ for $i = 1, \ldots, n$, then the posteriors from the priors induced by (2.2.6) and (2.2.7) are respectively Kolmogorov consistent at $\mathbb{F}$ and $\mathbb{T}_i$, for all i.*

This result closes the large sample properties of our methods; we next focus on assessing their finite-sample properties.

2.3 Numerical Experiments and Computing

2.3.1 Small n, Large m

As our asymptotic theory does not require $n \to \infty$, we start by assessing performance of the proposed methods in a small n, large m setting. We generate random samples $x_{i,1}, \ldots, x_{i,m_i} \mid m_i$, from

$$\lambda(t) = \Phi(t \mid 0.5, (0.15)^2), \quad m_i \sim \text{Poisson}(L),$$

for $i = 1, 2, 3$, with $L = 150$ and $\Phi(t \mid \mu, \sigma^2)$ denoting the normal distribution function. Then the warped data $\tilde{x}_{i,j} = T_i(x_{i,j})$ are obtained using

$$\begin{cases} T_i(t) &= t + \left(a_i - \frac{1}{2}\right) \sin(b_i t \pi)(b_i \pi)^{-1}, \quad i = 1, 2, \\ \\ T_3(t) &= 3t - T_1(t) - T_2(t), \end{cases}$$

where $a_1, a_2 \overset{\text{iid}}{\sim} \text{Unif}([0, 1/4] \cup [3/4, 1])$ and $b_1, b_2 \overset{\text{iid}}{\sim} \text{Unif}\{1, 2\}$. We clearly have that $E(T_i) = t$ since $E(a_j - 1/2) = 0$, for each $i = 1, 2, 3$ and therefore these warp maps are in line with the model assumptions.

A version of the proposed semiparametric approach in Section 2.2 can be implemented

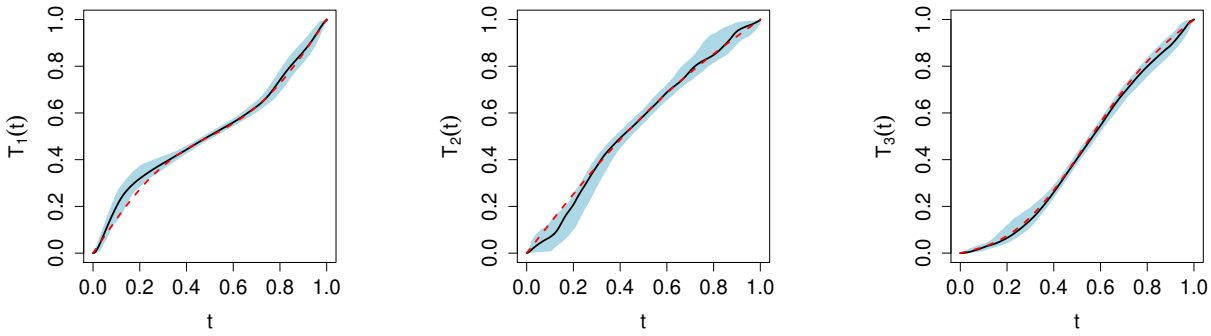


Figure 2.2: True (dashed red) and estimated (solid black) warp functions along with credible bands. The estimators are constructed as the posterior mean of the induced prior as (2.2.7).

with the aid of the R package Rmpp, which implements a version of the algorithm in Petrone (1999a, p. 383). Fig. 2.2 shows the estimators of each of the three warp maps through the posterior mean of the induced prior defined in (2.2.7), along with their credible bands and the true warp maps. From Fig. 2.2 it can be observed that our estimators are reasonably in line with the true warp functions. As a consequence, the method recovers quite well the original point processes, as can be seen when comparing the left and right panels of Fig. 2.3.

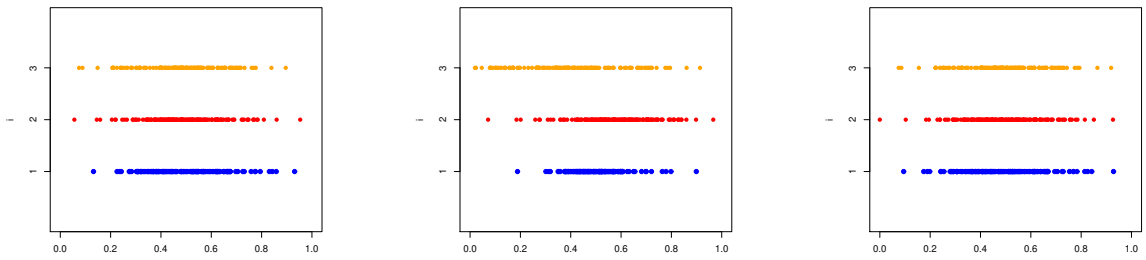


Figure 2.3: Left: Realizations of the original point process from the setup of Section 2.3.1 in the small n, large m regime. Middle: Their corresponding phase-varying point process. Right: Their corresponding registered versions.

A Monte Carlo study was conducted in this setting based on $B = 50$ simulated datasets. We apply our method to each, and then calculate the Monte Carlo L^2-Wasserstein distance

mean (WDM) by

$$\widehat{\mathrm{WDM}} = \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n} d(\widehat{\Pi}_i^{[b]}, \Pi_i^{[b]}), \tag{2.3.1}$$

where the superscript $[b]$ denotes the corresponding object computed from the bth simulated dataset, in order to give a performance of our methods when n is small ($n = 3$) and the m_i's are large. We obtained a value of 0.01274. When taking $L = 75$ instead of 150 the obtained value of $\widehat{\mathrm{WDM}}$ was 0.01697, in accordance with the intuition that this value decreases with L. Boxplots of $d(\widehat{\Pi}_i^{[b]}, \Pi_i^{[b]})$ are given in the Appendix A.2.1, for all i. In the Appendix A.2.2 we also include an additional simulation study suggesting satisfactory performance of the methods under misspecification, with data being warped via biased warp maps (i.e. $E(T) \neq t$).

2.3.2 Large n, Small m

For comparison with Panaretos and Zemel (2016) we now assess performance over a large n setup. We generate random samples $x_{i,1}, \ldots, x_{i,m_i} \mid m_i$, from

$$\lambda(t) = 0.2\,\phi(t \mid 0.25, 0.02^2) + 0.8\,\phi(t \mid 0.75, 0.03^2), \quad m_i \sim \mathrm{Poisson}(L), \quad i = 1, \ldots, n = 30,$$

with $\phi(t \mid \mu, \sigma^2)$ denoting the normal density function and $L = 50$. The warped data $\widetilde{x}_{i,j} = T_i(x_{i,j})$ are obtained using

$$T_i(t) \overset{D}{=} U\,\zeta_{K_1}(t) + (1-U)\,\zeta_{K_2}(t), \qquad \zeta_k(t) = \begin{cases} t, & k = 0, \\ t - \dfrac{\sin(\pi t k)}{|k|\pi}, & \text{otherwise}, \end{cases}$$

where $U \sim \mathrm{Unif}(0,1)$, $K_j \overset{D}{=} V_1 V_2$ with $V_1 \sim \mathrm{Poisson}(3)$ and $P(V_2 = -1) = P(V_2 = 1) = 1/2$.

We start by illustrating our method on this setup on a single run-experiment; a Monte

Carlo study was also conducted this setting along the same lines as in Section 2.3.1 and it will also be reported below. A realization of the original point process can be found in Fig. 2.1. After estimating $F_1, \ldots, F_n$ using random Bernstein polynomials we obtain the posterior Fréchet mean depicted in Fig. 2.4. The posterior mean is quite similar to the kernel-based estimator of Panaretos and Zemel (2016), and both are similar to the true Fréchet mean.

Fig. 2.4 also includes posterior inference for the warp functions. To examine the inference for warp functions in a greater level of detail Fig. 2.5 presents the posterior mean Bernstein polynomial warp function along with credible bands for $i = 5$. As it can be observed from the latter figure, our estimator follows closely that of Panaretos and Zemel (2016), and is reasonably in line with the original warp function; similar evidence holds for the remainder values of i (see Appendix A.2.1). As expected, both estimators have however more difficulty in recovering the true value in the center of unit interval as there tends to be much less data on that region, as can also be seen from Fig. 2.5; it may seem surprising that credible bands are smaller on the center of unit interval, but this is due to an extrapolation issue: Since very few warped points are observed on that region, posterior simulated trajectories overconfidently consider the warp function to be constant there.

While the theoretical claims in Section 2.3.2 extend those of Panaretos and Zemel (2016)—in the sense that under extra conditions they support the use of the methods even under a small n large m setting—numerical experiments in the Appendix A suggest that the pointwise performance of our methods may not dominate that of Panaretos and Zemel (2016).

Fig. 2.5 presents additionally credible intervals for randomly selected registered points for each registered point process. Observe that wider intervals are associated to points falling on the interval separating the two 'clusters' of points.

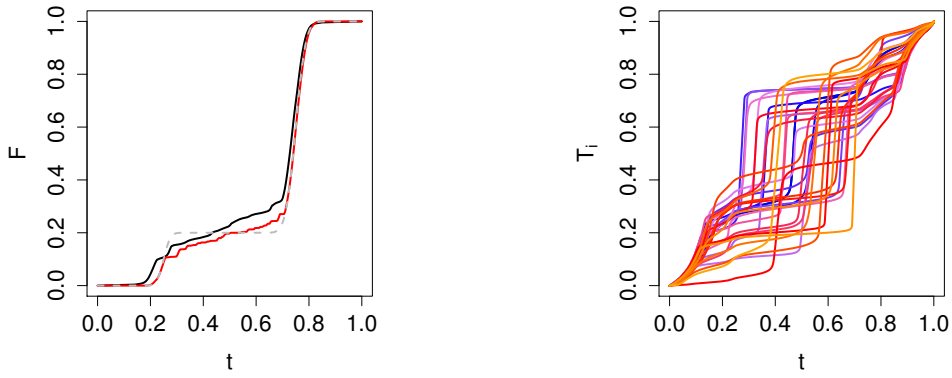


Figure 2.4: Left: Posterior Bernstein polynomial Fréchet mean (solid black), kernel smoothing Fréchet mean (solid red) and original Fréchet mean (grey dashed line). Right: Posterior mean Bernstein polynomial warp functions colored according to the same palette as in Fig. 2.1.

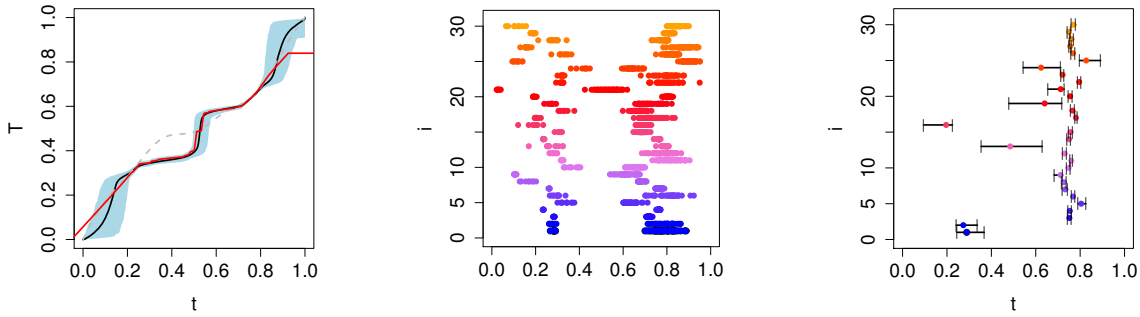


Figure 2.5: Left: Posterior mean Bernstein polynomial warp function (solid black) and corresponding credible band, kernel smoothing warp function estimate (solid red), and original warp function (dashed grey) for $i = 5$. Right: Credible intervals for randomly selected registered points for each registered point process.

2.4 Application: Tracking Phase Variation of Annual Peak Temperatures

We now showcase how our method can be used for tracking the phase variation of annual peak temperatures, that is, temperatures above or below a threshold. Peaks of temperature are related with a variety of hazardous events—including heat-related mortality, destruction of crops, wildfires—and have a direct impact on a wealth of economic decisions—such as demand for fuel and electricity. A better understanding of the variation of the regularity of these peaks is thus of the utmost importance from an applied perspective. A main target of our analysis will be on assessing the variation of the onset of temperature peaks, as well as quantifying how atypical is a certain year's pattern of such peaks. Our analysis has points of contact with the subject of shifts in seasonal cycles (e.g. late start of spring, or growing seasons), which is of wide interest in biology and climatology (e.g. Menzel and Fabian 1999; Schwartz et al. 2006). To illustrate how the method can be used for such purpose we gathered data from "*National Centers for Environmental Information of National Oceanic and Atmospheric Administration (NOAA)*" (https://www.ncdc.noaa.gov/), that consist of average daily air temperatures (in °F, rounded to the nearest integer) of Santiago (Chile) from April, 1990 to March, 2017. Let $\widetilde{x}_{i,j}$ be the temperature on day i, year j. Below, we focus on the point processes of annual peaks over threshold, $\{\widetilde{x}^+_{i,j} \geq u^+_j\}$, and annual peaks below threshold, $\{\widetilde{x}^-_{i,j} \leq u^-_j\}$; in practice we set the thresholds u^+_j and u^-_j using the 95% and 5% quantiles of temperature over year j, and this results in $m^+_1, \ldots, m^+_n$ and $m^-_1, \ldots, m^-_n$ ranging from 19 to 32. The supplementary material includes a sensitivity analysis based on the 97.5% and 2.5% quantiles; the main empirical findings are tantamount to the ones presented here. In Fig. A.6 we present the point processes of interest along with the corresponding warping functions for peaks above the threshold (T^+_j) and peaks below the threshold (T^-_j). For the analysis of annual peaks over threshold, we fully support the warping functions between the

minimum and maximum times corresponding to the pooled exceedances above the threshold; we proceed analogously for the analysis of annual peaks below the threshold.

To interpret Fig. A.6 we first focus on annual peaks below the threshold, for which there are at least two patterns of points that readily look unusual to the naked eye: 1991, for which there was an atypical cold weather event almost taking place in the summer; 2010, given that lower temperatures peaked later on a concentrated period. The fact that these patterns of points look unusual agrees with what can be observed from the corresponding warping functions, that are among the ones that further deviate from the identity; cf Fig. 6 and 7 from the supplementary material. In terms of peaks above the threshold, note how the antepenultimate pattern of points started much later than all the remainder, thus meaning that higher temperatures peaked much later than expected.

To assess how atypical is the climatological pattern of onset of peaks, we define the following measures to which we refer as scores of peak irregularity (SPI), and for temperatures above and below a threshold are respectively defined as

$$\text{SPI}^{+} = \int_0^1 |T_j^{+}(t) - t| \, \mathrm{d}t, \quad \text{SPI}^{-} = \int_0^1 |T_j^{-}(t) - t| \, \mathrm{d}t; \tag{2.4.1}$$

to combine peaks over and below a threshold, we also define a global $\text{SPI} = (\text{SPI}^{+} + \text{SPI}^{-})/2$. Fig. A.7 depicts the scores of peak irregularity over time for peaks above and below a threshold. Fig. A.7 is coherent with what was expected given the comments above surrounding Fig. A.6 on the patterns of points that looked immediately atypical, and on the shape of the corresponding warping functions.

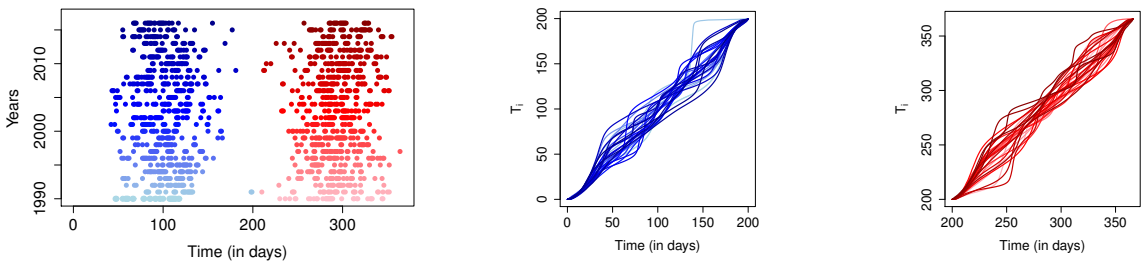


Figure 2.6: Left: Point processes of annual peaks for peaks above (red) and below (blue) the thresholds. Middle and Right: Corresponding posterior mean warp functions in the same palette of colors.

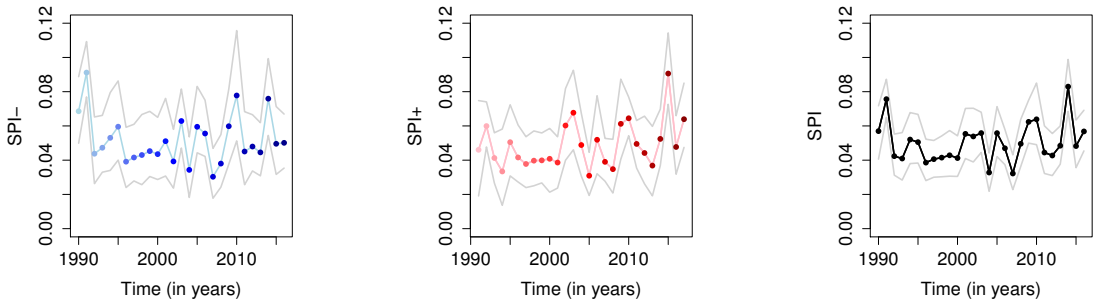


Figure 2.7: Posterior mean SPI (scores of peak irregularity), as defined in (2.4.1), along with credible intervals, for below threshold (Left), above threshold (Middle), and global (Right).

2.5 Closing Remarks

We propose a semiparametric Bayesian approach for the purpose of separating amplitude and phase variation in point process data. This paradigm has the advantage of providing a straightforward construction of credible sets via the posterior distribution, and in particular, we are able to quantify the uncertainty in learning not only the structural mean measure λ, but also the warping functions T_i and the latent point processes Π_i. The Bernstein–Dirichlet prior interweaves elegantly with the Wasserstein geometry of optimal transport. Indeed, its favorable support properties (as established by Petrone and Wasserman 2002) carry over to the induced priors on the structural mean measure λ and all sufficiently regular warping functions, allowing to obtain Bayesian consistency in a genuinely infinite-dimensional setup.

A natural question would be extending this work to the case of spatial point process supported on e.g., $[0,1]^D$ with $D > 1$, as explored by Boissard et al. (2015) and Zemel and Panaretos (2017); a natural extension of our work to this setup would entail modeling the mean measures of the corresponding spatial point processes via multivariate Bernstein polynomials (Zheng et al. 2009). The computation of the empirical Fréchet–Wassertein mean can no longer however be done in closed form, requiring numerical schemes (Peyré and Cuturi 2018). From a statistical viewpoint, another natural avenue for future research would be on modeling the phase variation of point processes conditionally on a covariate, by resorting to predictor-dependent versions of the Bernstein–Dirichlet prior (Barrientos et al. 2017).

2.6 Technical Details

2.6.1 Auxiliary Lemmas

We begin by stating a number of auxiliary lemmas that will be useful to deriving our main results. Lemma 1 is often known as Pólya's theorem (Lehmann and Romano 2006, The-

orem 11.2.9). Lemma 2 states that inversion is continuous in supremum norm (Lehmann and Romano 2006, Lemma 11.2.1). Lemma 3 discusses sufficient conditions for (local) uniform convergence of the Bernstein polynomial density; see Ghosal and Van der Vaart (2015, Lemma E.3) for a related result under further smoothness assumptions on $\mathbb{f}$. Proofs of Lemmas 1–3 are available from Appendix A.1.

Lemma 1. *Let $\mathbb{F}$ be a continuous distribution function and let F_n be a sequence of distribution functions that converge weakly to $\mathbb{F}$. Then $\|F_n - \mathbb{F}\|_\infty \to 0$.*

Lemma 2. *Let $\mathbb{F} : [0,1] \to [0,1]$ be continuous, strictly increasing and with $F(0) = 0$, $F(1) = 1$. Then $\mathbb{F}^{-1}$ is also continuous and strictly increasing, and for any $\varepsilon > 0$ there exists $\delta > 0$ such that for any continuous strictly increasing $H : [0,1] \to [0,1]$:*

(a) If $\|\mathbb{F} - H\|_\infty < \delta$, then $\|\mathbb{F}^{-1} - H^{-1}\|_\infty < \varepsilon$.

(b) If $\|\mathbb{F}^{-1} - H^{-1}\|_\infty < \delta$, then $\|\mathbb{F} - H\|_\infty < \varepsilon$.

Lemma 3. *Let $\mathbb{F} : [0,1] \to \mathbb{R}$ be differentiable with derivative $\mathbb{f}$ that is continuous on $(0,1)$. Then for any $a > 0$, $b(x \mid k, \mathbb{F})$ as defined in (2.2.2) converges to $\mathbb{f}$ uniformly on $[a, 1-a]$. If $\mathbb{f}$ is continuous on $[0,1]$, then $b(x \mid k, \mathbb{F}) \to \mathbb{f}$ uniformly on $[0,1]$.*

As the proof shows, the uniform converges holds on any set bounded away from the discontinuity points of $\mathbb{f}$.

2.6.2 Proofs of Main Results

Proof of Theorem 1.

(a) The proof follows from Theorem 3 in Petrone (1999b), combined with the fact that by assumption $F_1, \ldots, F_n \overset{\text{iid}}{\sim} \pi$. Indeed,

$$\pi^{(n)}\{\mathscr{F} : \|F_j - \mathbb{F}_j\|_\infty < \varepsilon,\ j = 1, \ldots, n\} = \prod_{j=1}^{n} \pi\{F_j : \|F_j - \mathbb{F}_j\|_\infty < \varepsilon\} > 0.$$

(b) From Theorem 2(a) and Lemma 2 it follows that

$$\pi^{(n)}\{\mathscr{F} : \|F_i^{-1} - \mathbb{F}^{-1}\|_\infty < \eta, i = 1, \ldots, n\} > 0, \quad \eta > 0. \tag{2.6.1}$$

Also, note that

$$\|F^{-1} - \mathbb{F}^{-1}\|_\infty = \left\| \frac{1}{n} \sum_{i=1}^{n} F_i^{-1} - \mathbb{F}^{-1} \right\|_\infty \leq \frac{1}{n} \sum_{i=1}^{n} \|F_i^{-1} - \mathbb{F}^{-1}\|_\infty. \tag{2.6.2}$$

From (2.6.2) and Lemma 2, it follows that to have $\|F - \mathbb{F}\|_\infty < \varepsilon$ it would suffice having $\|F_i^{-1} - \mathbb{F}^{-1}\|_\infty < \delta$ for all i, thus implying that

$$\pi^{(n)}\{\mathscr{F} : \|F - \mathbb{F}\|_\infty < \varepsilon\} \geq \pi^{(n)}\{\mathscr{F} : \|F_i^{-1} - \mathbb{F}^{-1}\|_\infty < \delta\} > 0.$$

(c) Lemma 1 and the assumption that the $\mathbb{F}_j$ are (uniformly) continuous on $[0, 1]$ imply that the $\mathbb{F}_i^{-1}$ are also uniformly continuous, for $i = 1, \ldots, n$. Given $\eta > 0$, let $\delta > 0$ such that $|t - s| \leq \delta \Rightarrow |\mathbb{F}_i^{-1}(t) - \mathbb{F}_i^{-1}(s)| \leq \eta$, for $i = 1, \ldots, n$. From Theorem 1 (a) and (b) it respectively follows

$$\pi^{(n)}\{\mathscr{F} : \|F_i - \mathbb{F}_i\|_\infty \leq \eta, \, i = 1, \ldots, n\} > 0, \quad \pi^{(n)}\{\mathscr{F} : \|F - \mathbb{F}\|_\infty \leq \delta\} > 0.$$

Thus, $\pi^{(n)}\{\mathscr{F} : |F(t) - \mathbb{F}(t)| < \delta, \, t \in [0, 1]\} > 0$, and this implies that the event

$$\begin{cases} F_i^{-1}(F(t)) \leq F_i^{-1}(\mathbb{F}(t) + \delta) \leq \mathbb{F}_i^{-1}(\mathbb{F}(t) + \delta) + \eta \leq \mathbb{T}_i(t) + 2\eta, \\ F_i^{-1}(F(t)) \geq F_i^{-1}(\mathbb{F}(t) - \delta) \geq \mathbb{F}_i^{-1}(\mathbb{F}(t) - \delta) - \eta \geq \mathbb{T}_i(t) - 2\eta, \end{cases} \tag{2.6.3}$$

occurs with positive probability, for $i = 1, \ldots, n$. This thus yields that

$$\pi^{(n)}\{\mathscr{F} : \|T_i - \mathbb{T}_i\|_\infty \leq 2\eta\} > 0, \quad i = 1, \ldots, n.$$

(d) The strategy of the proof is similar to that Panaretos and Zemel (2016, p. 798). We start by noting that $(T_i^{-1} \circ \mathbb{T}_i) \in \Gamma(\Pi_i/N_i, P_i/N_i)$ as a consequence of

$$\Pi_i = T_{i\,\#}^{-1}\widetilde{\Pi}_i = (T_i^{-1} \circ \mathbb{T}_i)_{\#}P_i, \quad i = 1, \ldots, n.$$

It thus follows that

$$d^2(\Pi_i/N_i, P_i/N_i) \leq \int_0^1 \{(T_i^{-1} \circ \mathbb{T}_i)(x) - x\}^2 \frac{\Pi_i(\mathrm{d}x)}{N_i} \leq \|\{T_i^{-1} \circ \mathbb{T}_i - x\}^2\|_\infty.$$

To complete the proof just note that Theorem 2(c) implies that for all i

$$\pi^{(n)}\{\mathscr{F} : \|T_i^{-1} \circ \mathbb{T}_i - x\|_\infty < \varepsilon\} = \pi^{(n)}\{\mathscr{F} : \|T_i^{-1} - \mathbb{T}_i\|_\infty < \varepsilon\} > 0,$$

from where the final result follows.

□

Proof of Theorem 3. The derivatives of the induced priors (2.2.6) and (2.2.7) will be required for the proofs, and are respectively

$$f(x) = n\left(\sum_{i=1}^n \frac{1}{f_i(T_i(x))}\right)^{-1}, \quad T_i'(x) = \frac{f(t)}{f_i(T_i(x))}, \quad i = 1, \ldots, n, \qquad f_i = F_i'.$$

(a) Let $\mathbb{f}_i$ be the density corresponding to $\mathbb{F}_i$ and $\mathbb{f}$ that corresponding of $\mathbb{F}$. Then

$$|f(x) - \mathbb{f}(x)| = n\left|\left(\sum_{i=1}^n \frac{1}{f_i(T_i(x))}\right)^{-1} - \left(\sum_{i=1}^n \frac{1}{\mathbb{f}_i(\mathbb{T}_i(x))}\right)^{-1}\right|. \tag{2.6.4}$$

We first assume that $\inf \mathbb{f}_i \geq 2l > 0$ for all i, and consequently $\inf \mathbb{f} > 2l$ as well. For $g : [0,1] \to \mathbb{R}$ and $1/2 > a > 0$ denote $\|g\|_{\infty,a} = \sup_{x\in[a,1-a]} |g(x)|$. We shall show that

the event

$$\Omega_{a,\rho} = \{f_i \geq l \ \& \ \|f_i - \mathbb{f}_i\|_{\infty,a} < \rho, i = 1, \ldots, n\}, \qquad a, \rho > 0$$

has positive probability for all $a, \rho > 0$. Let k_i be large so that $\|b(x \mid k_i, \mathbb{f}_i) - \mathbb{f}_i\|_{\infty,a} < \rho/2$ (using Lemma 3), set $k = \max_i k_i$ and denote $b(x \mid k, \mathbb{f}_i) = \sum_{j=1}^{k} w_{i,j}\beta(x \mid j, k - j + 1)$. The set of polynomials with slightly perturbed coefficients

$$\mathcal{P}_{i,\delta} = \left\{ p = \sum_{j=1}^{k} w'_{i,j}\beta(x \mid j, k - j + 1) : (w'_{i,1}, \ldots, w'_{i,k}) \in S_k \text{ with } |w'_{i,j} - w_{i,j}| < \delta, \text{for all } j \right\}$$

has positive probability under the Bernstein polynomial prior, for all $\delta > 0$, as a consequence of Petrone and Wasserman (2002, p. 84) as the corresponding set where $(w'_{i,1}, \ldots, w'_{i,k})$ lies is open in the unit simplex; in addition, each $p \in \mathcal{P}_{i,\delta}$ satisfies

$$\|p - b(x \mid k, \mathbb{f}_i)\|_\infty \leq \delta k \max_{1 \leq j \leq k} \sup_x \beta(x \mid j, k - j + 1) < \infty$$

because $1 \leq j \leq k$. Thus for small enough δ, $\|p - b(x \mid k, \mathbb{F}_i)\|_{\infty,a} < \rho/2$. Since the F_i's are independent, there is a positive probability that $f_i \in \mathcal{P}_{i,\delta}$ for all i, which implies that $\|f_i - \mathbb{f}_i\|_{\infty,a} < \rho$ for all i. Moreover, as $\mathbb{f}_i \geq 2l$, $w_{i,j} \geq 2l/k$ and if $\delta < l/k$ this yields $w'_{i,j} \geq l/k$ and thus $b(x \mid k, \mathbb{F}_i) \geq l$. Hence $\Omega_{a,\rho}$ has positive probability.

Fix $\epsilon > 0$; we wish to show that $\|F_i - \mathbb{F}_i\|_\infty \leq \epsilon$ holds on $\Omega_{a,\rho}$ for appropriate $a, \rho > 0$. Let $1/2 > a > 0$ such that $\mathbb{F}_i(a) < \epsilon/3$ and $\mathbb{F}_i(1-a) > 1 - \epsilon/3$, and let $\rho < \epsilon/3$. When $\Omega_{a,\rho}$ holds, we have

$$\begin{aligned} 1 \geq F_i(1-a) = F_i(a) + \int_a^{1-a} f_i(x)\mathrm{d}x &\geq F_i(a) + \int_a^{1-a} \mathbb{f}_i(x)\mathrm{d}x - \rho(1-2a) \\ &= F_i(a) + \mathbb{F}_i(1-a) - \mathbb{F}_i(a) - \rho(1-2a). \end{aligned}$$

Thus

$$-\epsilon \leq F_i(a) - \mathbb{F}_i(a) \leq 1 - \mathbb{F}_i(1-a) + \rho(1-2a) < 2\epsilon/3.$$

For $x \leq a$ we have

$$-\epsilon \leq F_i(x) - \mathbb{F}_i(x) \leq F_i(a) - \mathbb{F}_i(x) \leq 1 - \mathbb{F}_i(1-a) + \rho(1-2a) + \mathbb{F}_i(a) - \mathbb{F}_i(x) \leq \epsilon.$$

Thus $|F_i - \mathbb{F}_i| \leq \epsilon$ on $[0, a]$ and by a similar argument the same holds on $[1-a, 1]$. For $x \in [a, 1-a]$ observe that

$$|F_i(x) - \mathbb{F}_i(x)| \leq |F_i(a) - \mathbb{F}_i(a)| + \int_a^x |f_i(y) - \mathbb{f}_i(y)| \mathrm{d}y \leq |F_i(a) - \mathbb{F}_i(a)| + \rho < \epsilon.$$

Conclude that $\|F_i - \mathbb{F}_i\|_\infty \leq \epsilon$. As in the proof of Theorem 2 we have as a consequence that for sufficiently small a and ρ, on $\Omega_{a,\rho}$ $\|F_i^{-1} \circ F - \mathbb{F}_i^{-1} \circ \mathbb{F}\|_\infty < \epsilon$. Fix $a, \rho_2 \in (0, 1/2)$. Let $c_i = \min(\mathbb{F}_i^{-1}(\mathbb{F}(a)), 1 - \mathbb{F}_i^{-1}(\mathbb{F}(1-a)))$ and $a_1 = \min_i c_i/2$. Since $\mathbb{f}_i$ is uniformly continuous on $[a_1, 1-a_1]$, there exists $\delta_2 > 0$ such that $|\mathbb{f}_i(x) - \mathbb{f}_i(y)| \leq \rho_2$ for all $x, y \in [a_1, 1-a_1]$ such that $|x-y| \leq \delta_2$; without loss of generality $\delta_2 \leq a_1$. Choose small $a_1 > a_2, \rho > 0$ such that on $\Omega_{a_2,\rho}$, $\|F_i^{-1} \circ F - \mathbb{F}_i^{-1} \circ \mathbb{F}\|_\infty < \delta_2$. Then on $\Omega_{a_2,\rho}$

$$\|f_i \circ F_i^{-1} \circ F - \mathbb{f}_i \circ F_i^{-1} \circ F\|_{\infty,a} \leq \|f_i - \mathbb{f}_i\|_{\infty,a_1} \leq \rho$$

and

$$\|\mathbb{f}_i \circ F_i^{-1} \circ F - \mathbb{f}_i \circ \mathbb{F}_i^{-1} \circ \mathbb{F}\|_{\infty,a} \leq \sup_{x,y \in [a_1, 1-a_1], |x-y| \leq \delta_2} |\mathbb{f}_i(x) - \mathbb{f}_i(y)| \leq \rho_2.$$

This means that for any $\rho, \rho_2, a > 0$ there is positive probability that for all $i = 1, \ldots, n$

$$\|f_i \circ F_i^{-1} \circ F - \mathbb{f}_i \circ \mathbb{F}_i^{-1} \circ \mathbb{F}\|_{\infty,a} \leq \rho + \rho_2,$$

and since $\Omega_{a_2,\rho}$ implies also that $f_i, \mathbb{f}_i \geq l$, it follows that for all $a, \epsilon > 0$ there is positive probability that $\|f - \mathbb{f}\|_{\infty,a} < \epsilon$. Now write

$$\mathrm{KL}(F, \mathbb{F}) = \int_{x\in[a,1-a]} \mathbb{f}(x) \log \frac{\mathbb{f}(x)}{f(x)} \mathrm{d}x + \int_{x\notin[a,1-a]} \mathbb{f}(x) \log \frac{\mathbb{f}(x)}{f(x)} \mathrm{d}x = \mathrm{KL}_1 + \mathrm{KL}_2.$$

The definition of $\Omega_{a,\rho}$ implies that on this event $f \geq l$. Hence

$$\begin{aligned}\mathrm{KL}_2 &= \int_{x\notin[a,1-a]} \mathbb{f}(x) \log \mathbb{f}(x) \mathrm{d}x - \int_{x\notin[a,1-a]} \mathbb{f}(x) \log f(x) \mathrm{d}x \\ &\leq \int_{x\notin[a,1-a]} \mathbb{f}(x) \log \mathbb{f}(x) \mathrm{d}x - [1 - F(1-a) + F(a)] \log l \to 0, \qquad a \to 0\end{aligned}$$

since $\int \mathbb{f}(x) \log \mathbb{f}(x) \mathrm{d}x < \infty$. Hence we can pick $a > 0$ such that $\mathrm{KL}_2 < \epsilon$. To bound KL_1 notice that when $\epsilon < l = \inf \mathbb{f}$, the maximal value $|\log \frac{\mathbb{f}}{f}|$ can attain is $\log \frac{l}{l-\epsilon}$. Thus, for all $\epsilon > 0$ we have with positive probability

$$\mathrm{KL}(F, \mathbb{F}) \leq \epsilon + \log \frac{l}{l-\epsilon}.$$

As this vanishes when $\epsilon \to 0$, the proof is complete under the assumption that $\inf \mathbb{f}_i > 0$ for all i. This assumption can be relaxed as in Petrone and Wasserman (2002, p. 85)[1]: define $\mathbb{f}^a(x) = \max(\mathbb{f}(x), a)/A$, where $A = \int_0^1 \max(\mathbb{f}(x), a) \mathrm{d}x \in [1, 1+a]$. Applying the theorem to $\mathbb{f}_1 = \cdots = \mathbb{f}_n = \mathbb{f}^a$ we deduce the KL property for $\mathbb{f}^a$. Now, as $\mathbb{f} \leq A\mathbb{f}^a$ we have (Ghosal et al. 1999, Lemma 5.1)

$$\mathrm{KL}\left(\int h(x)\mathrm{d}x, \int \mathbb{f}(x)\mathrm{d}x\right) \leq (A+1)\log A + A\left[\mathrm{KL}\left(\int h(x)\mathrm{d}x, \int \mathbb{f}^a(x)\mathrm{d}x\right)\right.$$

$$\left. + \sqrt{\mathrm{KL}\left(\int h(x)\mathrm{d}x, \int \mathbb{f}^a(x)\mathrm{d}x\right)}\right].$$

[1] beware that they denote $\mathrm{KL}(F, \mathbb{F})$ by $\mathrm{KL}(\mathbb{F}, F)$

As $a \searrow 0$, $A \searrow 1$. If we choose $a > 0$ such that $A < 2$ and $(A+1)\log A < \varepsilon/3$, and then $\delta > 0$ such that $\delta + \sqrt{\delta} < \varepsilon/3$ then

$$\left\{h : \mathrm{KL}\left(\int h(x)\mathrm{d}x, \int \mathbb{f}(x)\mathrm{d}x\right) \leq \varepsilon\right\} \supseteq \left\{h : \mathrm{KL}\left(\int h(x)\mathrm{d}x, \int \mathbb{f}^a(x)\mathrm{d}x\right) \leq \delta\right\},$$

and the latter has positive prior probability. This completes the proof.

(b) Again begin with the assumption that $\inf \mathbb{f}_i > 0$ for all i. Let $T_i'(x) = f(x)/f_i(T_i(x))$ and $\mathbb{T}_i'(x) = \mathbb{f}(x)/\mathbb{f}_i(\mathbb{T}_i(x))$, and note that

$$|T_i'(x) - \mathbb{T}_i'(x)| \leq \frac{|f(x) - \mathbb{f}(x)|}{f_i(T_i(x))} + \mathbb{f}(x)\left|\frac{1}{f_i(T_i(x))} - \frac{1}{\mathbb{f}_i(\mathbb{T}_i(x))}\right|.$$

For all $a, \epsilon > 0$, since $\mathbb{f}$ is bounded on $[a, 1-a]$, the same idea as in part (a) shows that with positive probability $\|T_i' - \mathbb{T}_i'\|_{\infty,a} < \epsilon$. Write again

$$\mathrm{KL}(T_i, \mathbb{T}_i) = \int_{x\in[a,1-a]} \mathbb{T}_i'(x)\log\frac{\mathbb{T}_i'(x)}{T_i'(x)}\mathrm{d}x + \int_{x\notin[a,1-a]} \mathbb{T}_i'(x)\log\frac{\mathbb{T}_i'(x)}{T_i'(x)}\mathrm{d}x = \mathrm{KL}_1 + \mathrm{KL}_2.$$

These two terms can be made small as in part (a) because $\mathbb{T}_i' \leq n$.

To relax the condition $\inf \mathbb{f}_i > 0$ we use a similar idea as for part (a) but the argument is more subtle. Fix $a > 0$ and define

$$A_i = \int_0^1 \max(\mathbb{T}_i'(x), a)\,\mathrm{d}x, \qquad h_i^a(x) = \max(\mathbb{T}_i'(x), a)/A_i, \qquad H_i^a(x) = \int_0^x h_i^a(t)\,\mathrm{d}t.$$

For brevity we omit the dependence of h_i, H_i and A_i on a. Clearly H_i is strictly increasing, differentiable almost surely with derivative bounded below by a/A_i, $H_i(0) = 0$ and $H_i(1) = 1$. Moreover h_i is continuous and strictly positive on $(0,1)$ because so is $\mathbb{T}_i'$. We shall view H_i as transport maps from a Fréchet mean to well-behaved measures; first we need to fix the issue that they do not necessarily average to the identity by

adding another transport map that corrects the discrepancy.

By assumption

$$A_i \leq \int_0^1 (\mathbb{T}_i'(x) + a)\mathrm{d}x = \mathbb{T}_i(1) - \mathbb{T}_i(0) + a = 1 + a$$

and similarly $A_i \geq 1$. Thus we can choose $a > 0$ small such that $(1+a)/A_i \leq 1+1/(2n)$ for all $i = 1, \ldots, n$. Define the correction function

$$H_{n+1}(x) = (n+1)x - \sum_{i=1}^{n} H_i(x).$$

Then H_i, $i = 1, \ldots, n+1$ average to the identity. Since $\mathbb{T}_i$, $i = 1, \ldots, n$ average to the identity, whenever they are differentiable (that it, Lebesgue almost everywhere since they are nondecreasing) we have $\sum_{i=1}^n \mathbb{T}_i'(x) = n$. Hence H_{n+1} is differentiable almost surely with derivative

$$n+1-\sum_{i=1}^{n} h_i(x) \geq n+1-\sum_{i=1}^{n} \frac{T_i'(x)}{A_i} - \frac{na}{A_i} \geq n+1-n\frac{1+a}{A_i} \geq n+1-n(1+\frac{1}{2n}) = \frac{1}{2}.$$

Now consider the distribution functions $\mathbb{G}_i = H_i^{-1}$, $i = 1, \ldots, n+1$ and let $\mathbb{G}$ denote the identity. Then $\mathbb{G}_i$ have Fréchet mean $\mathbb{G}$ with densities bounded above by $\max(2, A_i/a)$ and below by $1/(n+1)$. Therefore, by the previous part of the proof $H_i^a = \mathbb{G}_i^{-1} \circ \mathbb{G}$ is in the KL support of the induced Bernstein polynomial prior. Since $\mathbb{T}_i' \leq A_i h_i^a$ almost surely we have (Ghosal et al. 1999, Lemma 5.1)

$$\begin{aligned}\mathrm{KL}(S, \mathbb{T}_{\mathbb{i}}) &\leq (A_i + 1)\log A_i + A_i[\mathrm{KL}(S, H_i^a) + \sqrt{\mathrm{KL}(S, H_i^a)}] \\ &\leq (a+2)\log(a+1) + (a+1)[\mathrm{KL}(S, H_i^a) + \sqrt{\mathrm{KL}(S, H_i^a)}].\end{aligned}$$

As $(H_i^a)'$ is continuous and strictly positive, $\mathrm{KL}(S, {H_i}^a)$ can be made as small as we

wish with positive probability. The fact that $a > 0$ is arbitrary completes the proof.

□

Proof of Theorem 4. Under the given assumptions the prior on F_i satisfies the Kullback–Leibler property (Petrone and Wasserman 2002, Theorem 2) at $\mathbb{F}_i$ and consequently the sequence of posteriors are weakly consistent for each $\mathbb{F}_i$. The operations

$$(F_1, \ldots, F_n) \mapsto (F_1^{-1}, \ldots, F_n^{-1}) \mapsto \left[F^{-1} = \frac{1}{n} \sum_{i=1}^{n} F_i^{-1} \right] \mapsto F,$$

are continuous in the supremum norm around $(\mathbb{F}_1, \ldots, \mathbb{F}_n)$ by Lemma 2, (2.6.2) and again Lemma 2. Taking into account the equivalence of the supremum norm with weak convergence (Lemma 1), conclude that the operation $(F_1, \ldots, F_n) \mapsto F$ is weakly continuous around $(\mathbb{F}_1, \ldots, \mathbb{F}_n)$. Since each F_i is weakly consistent for $\mathbb{F}_i$, this yields that F is weakly (in fact, Kolmogorov) consistent for $\mathbb{F}$.

Weak (and Kolmogorov) consistency of T_i to $\mathbb{T}_i$ follows in the same way, since in Equation (2.6.3) it has been established that

$$(F_i^{-1}, F) \mapsto F_i^{-1} \circ F$$

is continuous in supremum norm around $(\mathbb{F}_i, \mathbb{F})$. □

CHAPTER 3

Karhunen–Loève Priors for Families of Random Densities†

In this chapter we propose a data-driven prior on the space of families of density functions. The starting point for constructing our prior is to treat density functions as functional data, and to resort to tools and methods from functional data analysis to link all elements in a family densities. The proposed prior resorts to a Karhunen–Loève decomposition so to borrow strength across samples, by suitably distorting a baseline density. Theoretical properties of the proposed prior are discussed, along with conditions for enforcing that each element of the family is in L^2. Computationally, our method can be regarded as a post-processing procedure that could be used as a companion to models such as Dirichlet process mixtures. The simulation studies suggest that our method is competitive against other Bayesian nonparametric approaches. An illustration is given by revisiting Galton's dataset. Appendix B includes supplementary materials.

† Joint work with M. de Carvalho.

3.1 Introduction

Families of random densities are a natural model for a variety of contexts of applied interest, such as the K-sample setting and repeated-cross sectional surveys. In this chapter, we propose a prior for inference for families of random densities by combining ideas from two fast-evolving fields, namely: nonparametric Bayes and functional data analysis. One of the goals of the chapter is on proposing Bayesian inference for families of random densities using a post-processing procedure through infinite mixture models and functional principal component analysis. To further be able to describe the contributions of the chapter, we will briefly discuss what these fields are all about.

A key ingredient to the increasing popularity of Bayesian nonparametrics is the fact that it offers modeling flexibility and safeguard against misspecification (Müller et al. 2015). Bayesian nonparametric methods rely on parametric approaches as baseline models, while allowing for deviations from these when data provide evidence for it. Some Bayesian nonparametric approaches—such as Polya trees (e.g. Hanson 2006) and Dirichlet processes (e.g. Ghosal 2010)—can be understood as extensions of standard parametric methods in the sense that they are centered a priori around a parametric model, but assign positive mass to a variety of alternatives. Thus, a recurring theme in much of the Bayesian nonparametric literature is to regard a parametric approach as a reference, while allowing for other alternative models to 'take over' when data suggests that the parametric model is inappropriate. To do Bayesian nonparametrics—in the sense described above—we need prior probability models for probability distributions, and for this the natural probabilistic concept is that of a random probability measure. Other Bayesian nonparametric approaches—such as Gaussian process priors—consist of probability models over spaces of functions, and for these the natural probabilistic concept is that of a random function. The latter priors are particularly tailored for modeling, for example, mean functions (e.g. Rodriguez and Martinez 2014) and

even densities (e.g. Adams et al. 2009). See Hjort et al. (2010), Phadia (2015), Müller et al. (2015) for some recent surveys of prior processes, along with the review articles by Hjort (2003), Quintana and Müller (2004), and Müller and Mitra (2013).

In functional data analysis, data are random curves (stochastic processes). For example, due to recent advances in technology, medical diagnosis data are becoming increasingly complex and, nowadays, applications where measurements are curves are ubiquitous (Inácio de Carvalho et al. 2016). For an introduction to functional data analysis see, for instance, the monographs of Ferraty and Vieu (2006), Ramsay (2006), and Horváth and Kokoszka (2012); for a recent review of functional data analysis see Wang et al. (2015).

The starting point for our prior is the functional principal component analysis model of Kneip and Utikal (2001). A particularly interesting aspect of the Kneip–Utikal model is that the curves of interest are themselves densities, and a Karhunen–Loève decomposition is used to link all members of a family of densities $\{f_k\}_{k=1}^K$ through principal components. Such principal components, along with their corresponding 'scores' (dynamic strength components), can be used to assess how a baseline density ($f_\mu = K^{-1}\sum_{k=1}^K f_k$) would need to be 'tilted' so that each element of $\{f_k\}_{k=1}^K$ could be obtained; for recent applications of the Kneip–Utikal setup see Huynh et al. (2011) and references therein.

The proposed prior uses the Kneip–Utikal setup so to link all elements in a family of densities via a Karhunen–Loève decomposition. Different specifications of our prior can be constructed by setting up different models for each f_k individually. The scores of the decomposition then act tilting the baseline density so to link it with the f_k. Computationally, the proposed method can be understood as a post-processing procedure that could be used as a companion to model such as Dirichlet process mixtures.

In the next section we discuss the proposed method, in Section 3.3 we report numerical experiments, and in Section 3.4 we offer an illustration. Concluding remarks are given in Section 3.5.

3.2 The Karhunen–Loève Prior

3.2.1 Definition and Properties

The data configuration of interest is essentially a K-sample setting, which we denote by $\boldsymbol{x} = \{x_{i,k}\}$. Specifically, we assume that the data consist of random samples from K random distributions,

$$x_{i,k} \mid F_k \overset{\text{iid}}{\sim} F_k, \quad i = 1, \ldots, n_k, \quad k = 1, \ldots, K, \tag{3.2.1}$$

where $F_k(x) = P(x_{i,k} < x)$ denotes the random distribution function of the kth population throughout, we assume that each F_k is absolutely continuous, a.s., and define its corresponding density as $f_k = \mathrm{d}F_k/\mathrm{d}x$. Let L_1^2 denote the space of square integrable random densities, i.e. $f \in L_1^2$ if and only if, $\int f(x)\,\mathrm{d}x = 1$ and $\int \{f(x)\}^2\,\mathrm{d}x < \infty$, a.s, with inner product $\langle f, g\rangle = \int f(x)g(x)\,\mathrm{d}x$, for $f, g \in L_1^2$. Throughout, we assume that $f_k \in L_1^2$, for $k = 1, \ldots, K$.

The setting above assumes an underlying family of random densities $\{f_k^*\}_{k=1}^K$ and the purpose of the following construction is to link-up all elements via a common link (to be termed below, baseline density). To do that, given a family of random densities we can define the common mean $f_\mu = K^{-1}\sum_{k=1}^K f_k^*$, and then the question would be how to distort f_μ so to recover the remainder elements in the family. The Karhunen–Loève decomposition offers a natural way to conduct such inquiry. In this setting, it would be given by

$$f_k^* = f_\mu + \sum_{j=1}^{J} \theta_{k,j} g_j, \quad k = 1, \ldots, K, \tag{3.2.2}$$

where $J \leq K$, and $\{\theta_{k,j}\}$ and $\{g_j(x)\}$ obey the following equations

$$\boldsymbol{\nu} g_j = \lambda_j g_i \quad \text{and} \quad \theta_{k,j} = \langle f_j^* - f_\mu, g_j\rangle, \tag{3.2.3}$$

where $\boldsymbol{\nu}$ denote the empirical covariance operator. Then we can write any member of the family f_j^* as a distortion of the common mean, which can be conducted using the Karhunen–Loève decomposition as we define in (3.2.2). Therefore, putting this together we can introduce the data-driven Karhunen–Loève prior—which we call the KL prior—as the probability measure associated to the random density.

Definition 11 (Karhunen–Loève prior)**.** *A family of random density functions* $\{f_k\}_{k=1}^K$ *is said to follow a Karhunen–Loève prior with hyperparameters* $\{f_k^*\}_{k=1}^K$*, with* $f_k^* \in L_1^2$*, if*

$$f_k = f_\mu + \sum_{j=1}^{J} \mathbb{E}[\theta_{k,j} \mid \boldsymbol{x}]\, \mathbb{E}[g_j \mid \boldsymbol{x}]. \tag{3.2.4}$$

Here, $f_\mu = K^{-1} \sum_{k=1}^K f_k^*$*, whereas* $\theta_{k,j}$ *and* g_j *respectively denote the scores and principal components of* $\{f_k^*\}_{k=1}^K$ *resulting from the Karhunen–Loève decomposition* (3.2.2)*.*

Remark 3. *Note that the expectations* $\mathbb{E}[\theta_{k,j} \mid \boldsymbol{x}]$ *and* $\mathbb{E}[g_j \mid \boldsymbol{x}]$ *can be calculate using the induce measure by* $\{f_k^*\}$*, which follows from the Eq.* (3.2.3)*.*

Remark 4. *Note that the expectations used in the definition of our prior are of the type* $\mathbb{E}[\cdot \mid \boldsymbol{x}] \equiv \mathbb{E}[\cdot \mid \boldsymbol{X} = \boldsymbol{x}]$ *rather than* $\mathbb{E}[\cdot \mid \boldsymbol{X}]$*.*

As it can be seen from (3.2.4) the Karhunen–Loève (KL) prior takes the baseline density f_μ as a starting point, but it allows the data $\boldsymbol{x}$ to take over and to tilt the baseline in a way that borrows strength across samples. It is evident from Definition 11 that KL prior is a data-driven prior; data-driven priors are common in Bayesian statistics including, for example, empirical Bayes (Lehmann and Casella 1998, Section 4.6), Hartigan's maximum likelihood prior (Hartigan 1998), and max-compatible priors (de Carvalho et al. 2019b). Below we use the notation $\{f_k\} \sim \mathrm{KL}\{f_k^*\}$ to denote that the family $\{f_k\}$ follows a Karhunen–Loève prior with hyperparameters $\{f_k^*\}$; natural specifications for the hyperparameters are discussed in Section 3.2.2.

The following proposition summarizes some properties of the KL prior:

Proposition 1. *Suppose* $\{f_k\} \sim KL\{f_k^*\}$ *with* f_i^* *being independent of* f_j^**, for* $i \neq j$*. Then:*

1) $\sum_{k=1}^K var[f_k] \leq \sum_{k=1}^K var[f_k^*]$,

2) $cov(f_i, f_j) = K^{-2} \sum_{k=1}^K var[f_k^*] > 0$*, for* $i \neq j$,

a priori.

Proof. See Appendix. □

This proposition warrants some remarks. We refer to 1) and 2) as the *variance-reducing* and *generating-dependence* properties of the KL prior. The variance reducing property implies that the total variance of the prior (i.e., $\sum_{k=1}^K \text{var}[f_k]$) is smaller than of the corresponding hyperparamters. From the generating-dependence property is follows that, although the hyperparameters are a priori independent, the random density functions $\{f_k\}$ are a priori dependent; a consequence of the generating-dependence property is thus that the KL prior can be seen as a device to build dependent family $\{f_k\}$ from standard nonparametric Bayesian (hyper)priors for $\{f_k^*\}$, such as Dirichlet process mixtures; we explore this aspect of KL priors in further details in Section 3.2.2.

Similarly to Kneip and Utikal (2001) our prior will put positive mass on possibly negative f_k. While this may seem to be a shortcoming, our numerical experiments in Section 3.3 show that in line with Kneip and Utikal (2001) this is not as problematic as it may look ex-ante. The possibility for negative f_k comes from the fact that $\mathbb{E}[\theta_{k,j} \mid \boldsymbol{x}]$ and $\mathbb{E}[g_j \mid \boldsymbol{x}]$ can be negative for some $x \in \mathscr{X}$, and therefore so can the f_k be. We note below that all members in the family $\{f_k\}$ will integrate to one.

Proposition 2. *The family* $\{f_k\}$ *defined in Definition 11 obeys*

$$\int_{\mathscr{X}} f_k(x)\mathrm{d}x = 1, \quad k = 1, \ldots, K.$$

Proof. See Appendix. □

The next section explores a specific version of the setup above, where the hyperparameters consist of an infinite mixture model.

3.2.2 Karhunen–Loève–Dirichlet Prior

The construction of KL prior depends on the choice of the hyperparameters $\{f_k^*\}$. Since our data configuration of interest is a K-sample setting as defined in (3.2.1), we suggest considering infinite mixture models of the type

$$f_k^*(x) = \int_\Theta \mathbb{K}(x \mid \boldsymbol{\theta}) \, \mathrm{d}H_k(\boldsymbol{\theta}), \tag{3.2.5}$$

where $\mathbb{K}$ is a kernel and H_k is a random mixing measure. A popular approach is to consider H_k as a Dirichlet process (Ferguson 1973) (i.e., $H_k \sim \mathrm{DP}(\alpha_k, H_{k,0})$), and a normal kernel in (3.2.5), $\mathbb{K}(x \mid \boldsymbol{\theta}) = \phi(x \mid \mu, \sigma)$, in which case one obtains the so-called Dirichlet process mixtures (DPM) of normal kernels, and denote by $\{f_k^*\} \sim \mathrm{DPM}(\mathbb{K}, \{\alpha_k\}, \{H_{k,0}\})$; here and below, the f_k^* are assumed to be independent of $f_{k'}^*$, for $k \neq k'$. The instance of a KL prior with DPM hyperparameters will be of particular interest, and thus it will deserve a name of its own.

Definition 12 (Karhunen–Loève–Dirichlet prior)**.** *A family of random densities $\{f_k\}_{k=1}^K$ is said to follow a Karhunen–Loève–Dirichlet prior (to be denote as $\{f_k\} \sim KLD\{f_k^*\}$) if*

1) $\{f_k\} \sim KL\{f_k^*\}$,

2) $\{f_k^*\} \sim DPM(\mathbb{K}, \{\alpha_k\}, \{H_{k,0}\})$.

Different specifications of the Karhunen–Loève prior can be obtained by replacing H with some random distributions. For example, Barrios et al. (2013) consider infinite mixture

models of the type (3.2.5), but where the mixing distribution H is a generalized gamma NRMI (normalized random measure with independent increments), to be denoted by $H \sim \text{NGG}(\alpha, \kappa, \gamma, H^*)$. The parameters $\alpha > 0$ and H^* play analogous roles as in the DP, and the parameters $\kappa \geq 0$ and $\gamma \in [0, 1)$ can be used to specify members of this class, including the DP ($\kappa = 1, \gamma = 0$), the normalized inverse Gaussian (Lijoi et al. 2007) ($\gamma = 1/2$), and the N-stable process ($\alpha = 1, \kappa = 0$) (Kingman 1975). Still, other natural candidates for modeling H have been proposed, including stick-breaking priors (Ishwaran and James 2001), and probit stick-breaking priors (Rodríguez and Dunson 2011).

3.2.3 Computing and Implementation

Some comments in terms of fitting our model are in order. There exist quite a few algorithms for fitting (3.2.5) (e.g. Neal 2000; Ishwaran and James 2002). In the case of Dirichlet process mixtures in (3.2.5), posterior sampling can be conducted through a blocked–Gibbs sampler (Ishwaran and James 2001), which relies on a truncated DP prior; this is our choice in practice as it is a manageable approximation to the DP prior, and a major difference in comparison to standard parametric mixture models is that here we fix a maximum on the number of clusters, and not the number of clusters itself (Dunson 2010, pp. 232–233). There exist algorithms which do not rely on truncated versions of the DP, such as the collapsed Gibbs sampler (MacEachern 1994), and reversible jump Markov chain Monte Carlo approaches (Jain and Neal 2004). Futher details on computing and implementations will be given in the Section 4.1 and 4.2.

As mentioned above the proposed method can be regarded as a post-processing method. Given $f_k^{*[1]}, \ldots, f_k^{*[T]}$ posterior samples from f_k^* we can get posterior samples for f_k just plugging-in the $f_k^{*[b]}$ in f_k for each k. Also, we can estimate the conditional expectations of $\theta_{k,j}$ and g_j using the posterior sample scores, $\theta_{k,j}^{[t]}$, and posterior sample components, $g_j^{[t]}$, which are computed from the matrix $M = (M_{k,k'})$ where $M_{k,k'}^{[t]} = \langle f_k^{[t]} - f_\mu^{[t]}, f_{k'}^{[t]} - f_\mu^{[t]} \rangle$. The

last statement is based in the relationship between eigenvalues of the empirical covariance operator and the matrix M defined as before (see Good 1969). The implementation of the proposed methods is available from the function `dKLD` of our `R` package, `ROCstudio` (see Appendix C).

3.2.4 Theoretical Properties

Let $\mathscr{P}_{(\Theta,\mathbb{B}_\Theta)}$ be the space of all probability measures that can be defined on $(\Theta, \mathbb{B}_\Theta)$, with $\mathbb{B}_\Theta$ denoting the Borel sigma algebra over Θ. Also, let $\mathscr{X} = \{x \in \mathbb{R} : \mathbb{K}(x \mid \boldsymbol{\theta}) > 0\}$. To ensure that each f_k—as defined in (3.2.5)—is a random element of the space of square integrable random densities L_1^2, we assume the following conditions:

(A1) $\int_{\mathscr{X}} \mathbb{K}(x \mid \boldsymbol{\theta})\, \mathrm{d}x = 1$, for $\boldsymbol{\theta} \in \Theta$.

(A2) $\boldsymbol{\theta} \mapsto \mathbb{K}(x \mid \boldsymbol{\theta})$ is uniformly bounded, and $\mathbb{B}_\Theta$-measurable for $x \in \mathscr{X}$.

Indeed, under such regularity conditions on the kernel, the following result holds.

Proposition 3. *Under A1 and A2, each f_k as defined in (3.2.5) belongs to L_1^2.*

Proof. See Appendix. □

To connect theory with practice, let's consider (3.2.5) in which case the kernel corresponds to the normal density function, that is, $\mathbb{K}(x \mid \boldsymbol{\theta}) = \phi(x \mid \mu, \sigma^2)$, with $\boldsymbol{\theta} = (\mu, \sigma^2)$. Thus, if we suppose that $\sigma \in [\underline{\sigma}, \infty)$ for some fixed $\underline{\sigma} > 0$, then it holds that

$$\phi(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \leq \frac{1}{\underline{\sigma}\sqrt{2\pi}}.$$

Therefore, in theory—provided we support the prior of σ away from 0—the normal kernel satisfies that regularity condition A1 and A2, and thus Proposition 3 implies that a density constructed via (3.2.5) corresponds to a random element of L_1^2. In practice, in our numerical

experiments we found no need to support the prior of σ away from 0 as a plain vanilla DPM yielded satisfactory results overall.

3.3 Simulation Study

3.3.1 Simulation Scenarios and One-Shot Experiments

Below we compare the performance of our KLD prior against the DPM model and the B-splines DDP mixture model. A simulation study will be reported in Section 3.3.2; for now we concentrate on illustrating the methods on a single-run experiment, and on describing the design underlying the numerical study. We consider three scenarios. For Scenario I we sample from $y_{ik} \mid f_k \overset{\text{iid}}{\sim} f_k$ for $i = 1, \ldots, n$, with

$$f_k(y) = \omega_{1,k}\phi(y \mid -2, 1) + \omega_{2,k}\phi(y \mid 0, 0.25^2) + \omega_{3,k}\phi(y \mid 4, 1), \quad k = 1, \ldots, 10,$$

where $\omega_{1,k} = 0.32 + 0.02 \times (k-1)$, $\omega_{2,k} = 0.2 + 0.01 \times (k-1)$ and $\omega_{3,k} = 1 - \omega_{1,k} - \omega_{2,k}$. For scenario II we sample from $y_{ik} \mid f_k \overset{\text{iid}}{\sim} f_k$ for $i = 1, \ldots, n$, with

$$f_k(y) = \omega_k\phi(y \mid -2, k^{-\frac{1}{2}}) + (1 - \omega_k)\phi(y \mid 2, k^{\frac{1}{4}}), \quad k = 1, \ldots, 10,$$

with $y > 0$ and weights $\omega_k = 0.5 + 0.01(k-1)$, for $k = 1, \ldots, 10$. And for scenario III–the most complex one–we sample from $y_{ik} \mid f_k \overset{\text{iid}}{\sim} f_k$ for $i = 1, \ldots, n$, with

$$f_k(y) = \omega_{1,k}\phi(y \mid k/10, 1) + \omega_{2,k}\phi(y \mid -1, 0.25) + \omega_{3,k}\phi(y \mid 4, 1/\log(k+1)), \quad k = 1, \ldots, 10,$$

where

$$\omega_{1,k} = \begin{cases} 0.52 + 0.02 \times (k-1) & \text{if } k \in \{1, \ldots, 5\} \\ 0.52 + 0.02 \times (k-1) & \text{if } k \in \{6, \ldots, 10\} \end{cases},$$

$\omega_{2,k} = 0.22 + 0.02 \times (k-1)$ and $\omega_{3,k} = 1 - \omega_{1,k} - \omega_{2,k}$. In all cases, $\phi(\cdot \mid \mu, \sigma^2)$ stands for a normal density with mean μ and variance σ^2.

Above we set $n = 500$; the Appendix B.2 contain the outcomes with for $n = 1000$.

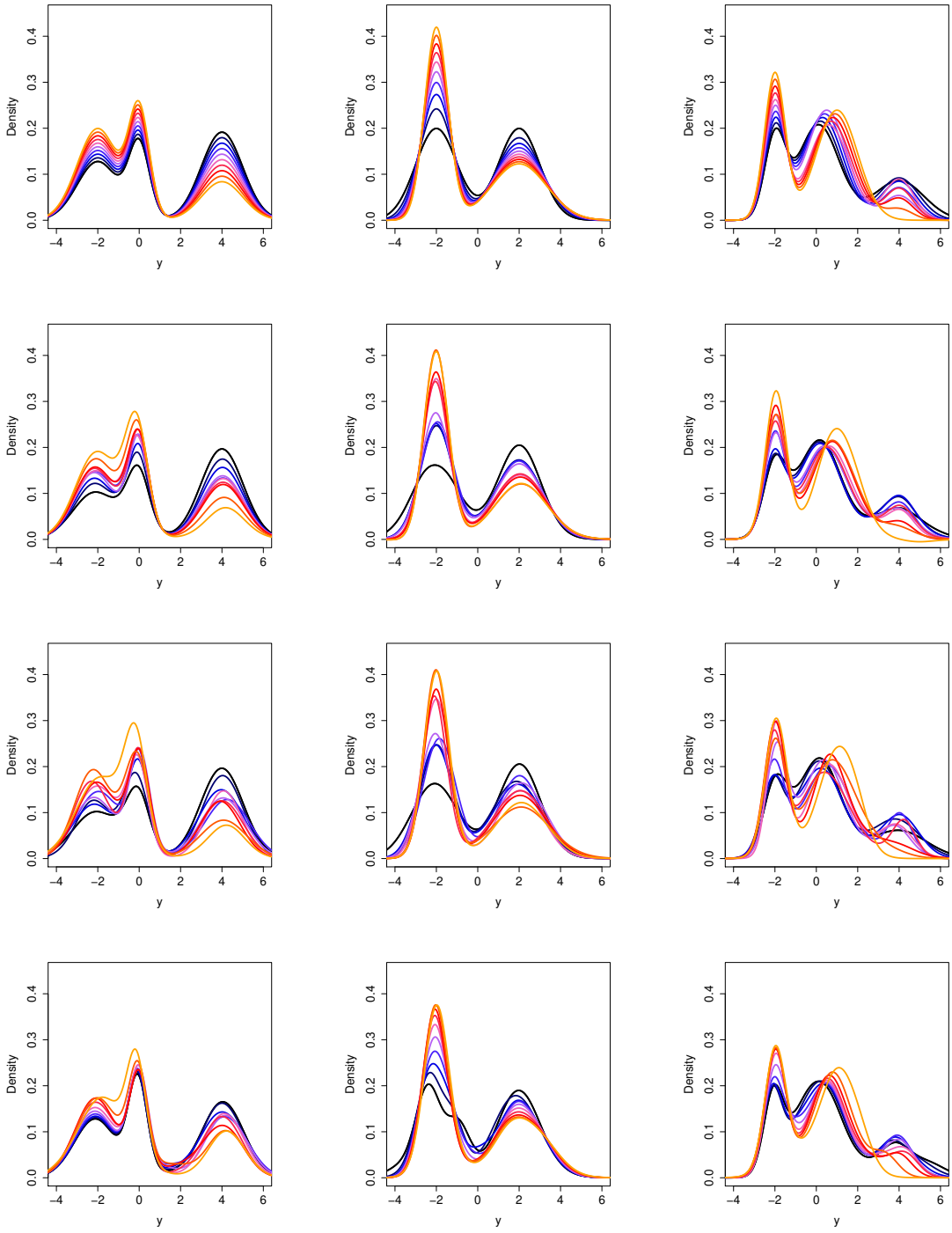


Figure 3.1: Single-run experiment illustrating KLD against DPM and DDP over three scenarios ($n = 500$). The true densities underlying all scenarios are depicted in the first row.

The single-run experiment in Figure 3.1 allows us to anticipate strengths and limitations with the methods under discussion. As it can be seen from Figure 3.1, KLD-based density

estimates tend to look closer to the real densities than the corresponding estimates by DPM and DDP models. For now, these conclusions should be regarded as tentative—since they summarize the outcome of a single run-experiment—and the goal of the next section is to assess how robust these conclusions are over other runs of simulated data. For each model, we generate 5 000 trajectories using a blocked Gibbs sampler (Ishwaran and James 2001), where we set a truncation value—for DPM and DDP models—equal to 20, and burn the first 500. Also we standardize the data prior to fitting the DPM model and set $\alpha_k = 1$ for both, DPM and DDP. In terms of prior information and kernel: For all the scenarios, we use $\mu_{k,h} \overset{\text{iid}}{\sim} N(0, 100)$ and $\sigma^2_{k,h} \overset{\text{iid}}{\sim} \text{IG}(0.1, 0.1)$ as hyper-priors of DPM model with normal kernel; In the DDP, we fit a B-splines DDP mixture with $Q = 4$, for normal-gamma we set $\boldsymbol{\mu} = 0_Q, \boldsymbol{S} = 100\boldsymbol{I}_Q, \nu = Q + 2, \boldsymbol{\psi}^{-1} = \boldsymbol{I}_Q$, $a = b = 0.1$ and normal kernel.

3.3.2 Monte Carlo Study

A Monte Carlo study was performed by simulating $B = 1000$ datasets for Scenarios I, II and III, as described in Section 3.3.1. We compare the performance of competing approaches using MISE (Mean Integrated Squared Error),

$$\text{MISE}_k = E\left[\int_{\mathscr{Y}} \{\hat{f}_k(y) - f_k(y)\}^2 \text{d}y\right] \approx \frac{1}{B}\sum_{b=1}^{B}\int_{\mathscr{Y}} \{\hat{f}_k^{(b)}(y) - f_k(y)\}^2 \,\text{d}y, \tag{3.3.1}$$

where $\hat{f}_k^{(b)}$ is the estimate underlying the bth simulated data set.The MISE resulting from this Monte Carlo study is summarized in Figure 3.2.

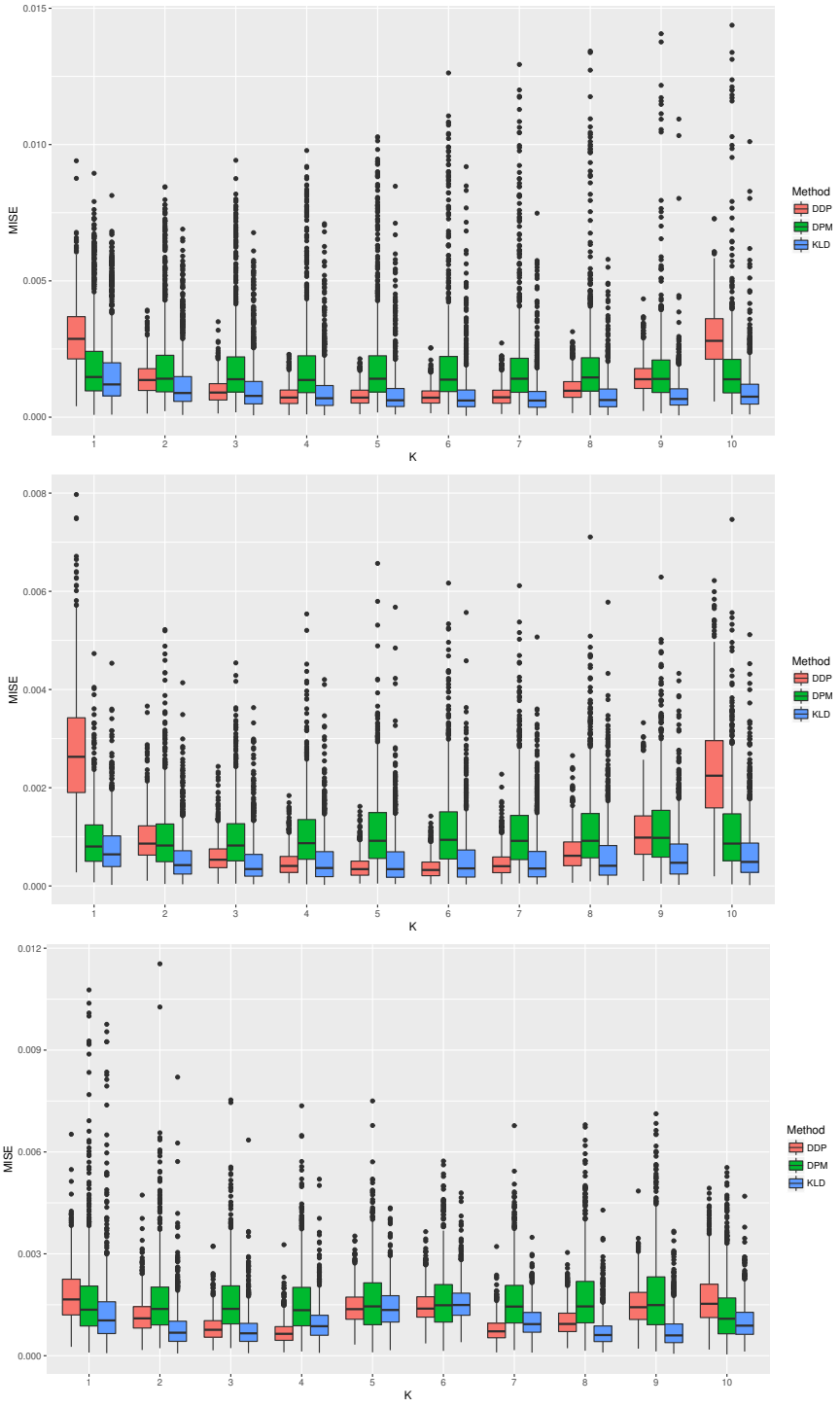


Figure 3.2: Boxplot of MISE for KLD, DPM and DDP estimates resulting from Monte Carlo study, for each density. The plots are Scenario I–III from top to bottom.

Figure 3.2 suggests that the KLD estimator tends to achieve a lower or competitive MISE over the corresponding DPM and DDP estimates, for different scenarios.

While the MISE in (3.3.1) provides us with a score for each element in the family, it is also natural wondering how one can obtain a score which ranks the performance of an estimation method for an entire family. With the latter purpose in mind, we propose using the global MISE,

$$\text{GMISE} = \frac{1}{K}\sum_{k=1}^{K}\text{MISE}_k \approx \frac{1}{KB}\sum_{k=1}^{K}\sum_{b=1}^{B}\int_{\mathscr{Y}}\{\hat{f}_k^{(b)}(y) - f_k(y)\}^2\,\mathrm{d}y.$$

The global MISE from this Monte Carlo study is reported in Figure 3.3

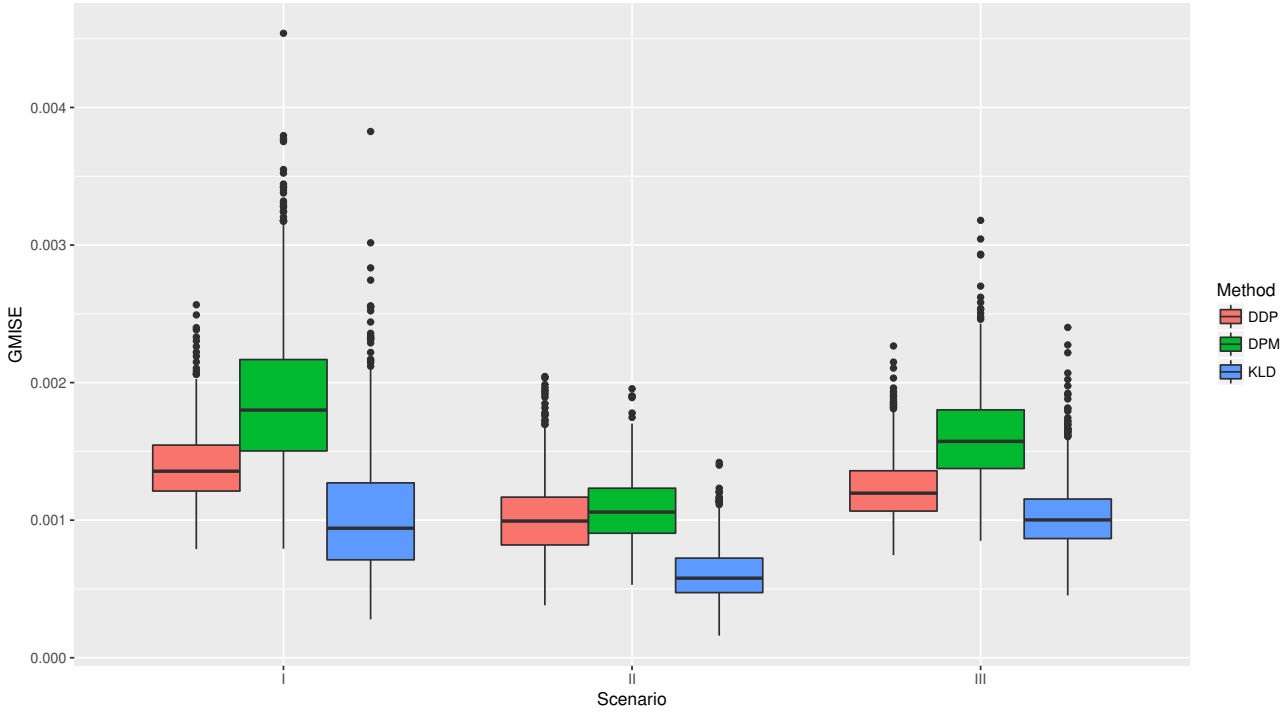


Figure 3.3: Boxplot of global MISE for KLD, DPM and DDP estimates resulting from Monte Carlo study for Scenario I–III.

As it can be noticed from Figure 3.3, the KLD estimator seems had the lower global MISE in each scenario but clearly it is the best in Scenario II, and and very competitive with DDP (Scenario I) and DPM (Scenario III).

To wrap up, the findings above show that KLD presents a competitive performance against popular models as DPM and DDP, and in some cases will actually outperform the latter models.

3.4 Revisiting Galton's Data

To illustrate the methods, we revisit Galton's dataset on the heights of parents and their children (Galton 1886). This data set is quite popular as it played a key role on the early developments of regression methods (Stigler 1986, Chapter 8), and it is readily available from R in the `HistData` package. In Galton's own words: "When Mid-Parents are taller than mediocrity, their Children tend to be shorter than they" (cf Galton 1886). These data have been examined widely (see Wachsmuth et al. 2003; Hanley 2004, and references therein) and in many analysis a small amount of jittering is typically introduced.

Score for Parents Height	Principal components			
	PC 1	PC 2	PC 3	PC 4
64.0 (f_1)	0.1488	0.1014	0.0336	0.0164
64.5 (f_2)	0.1878	0.0815	-0.0135	-0.0148
65.5 (f_3)	0.0946	-0.0219	-0.0054	-0.0011
66.5 (f_4)	0.0959	-0.0566	-0.0151	0.0005
67.5 (f_5)	0.0403	-0.0732	-0.0056	0.0015
68.5 (f_6)	0.0020	-0.0717	0.0003	0.0008
69.5 (f_7)	-0.0495	-0.0410	0.0030	-0.0005
70.5 (f_8)	-0.1257	-0.0400	0.0087	-0.0005
71.5 (f_9)	-0.1756	-0.0195	0.0147	-0.0019
72.5 (f_{10})	-0.2187	0.1411	-0.0207	-0.0004

Table 3.1: Scores ($\theta_{k,j}$) associated to $K = 10$ levels of parents height [in inches (in)], corresponding to $\{f_k\}_{k=1}^{10}$, for the first four principal components (g_j) for Galton's data.

In Figure 3.5(a), we plot the version of the data to be used; a light amount of jittering was added to artificially smooth the child's height. A height-adjusted palette is used to aid

in the interpretation of the results below. KLD density estimates for child height are shown in Figure 3.4; the same prior information as in Section 3.3 was used here.

To facilitate interpretations we set the baseline (f_1) as the density of child heights for which parents height equals the sample minimum, i.e., 64in; the corresponding data from which the baseline is estimated are represented in black in Figure 3.5(a). To put it differently, the black line in Figure 3.5(b) represents the baseline KLD estimate which corresponds to parents whose height equals 64in, and thus coincides with one of density estimates in Figure 3.4.

The first principal component explains around 71.12% of the variance, with a 95% credible interval of $(59.85\%, 81.96\%)$ whereas the first two components are responsible for about 96.79% of the variance; see Table 3.2.

	Principal components			
	PC 1	PC 2	PC 3	PC 4
Explained variability	71.12%	25.67%	2.55%	0.55%
	(59.85%,81.96%)	(16.09%,35.69%)	(0.74%,5.11%)	(0.09%,1.61%)

Table 3.2: Explained variability and their 95% credible intervals (in parenthesis) for the first four principal components for Galton's data.

A natural question from an applied viewpoint is now the following: "What is the role played by the first principal component in terms of tilting the baseline density?" To answer this question, in Figure 3.5(b) we plot the data along with the first principal component deformations of f_1, i.e. $(\theta_{k,1} - \theta_{1,1})g_1$. To aid the interpretation, we color the first principal component deformations of f_1 using the same palette as the one used to represent the data as in 3.5(a); the scores associated to the first four principal components can be found in Table 3.1. The first principal component deformations in Figure 3.5(b) have a straightforward interpretation: more mass is increasingly assigned to higher child heights, for parents whose height is also higher.

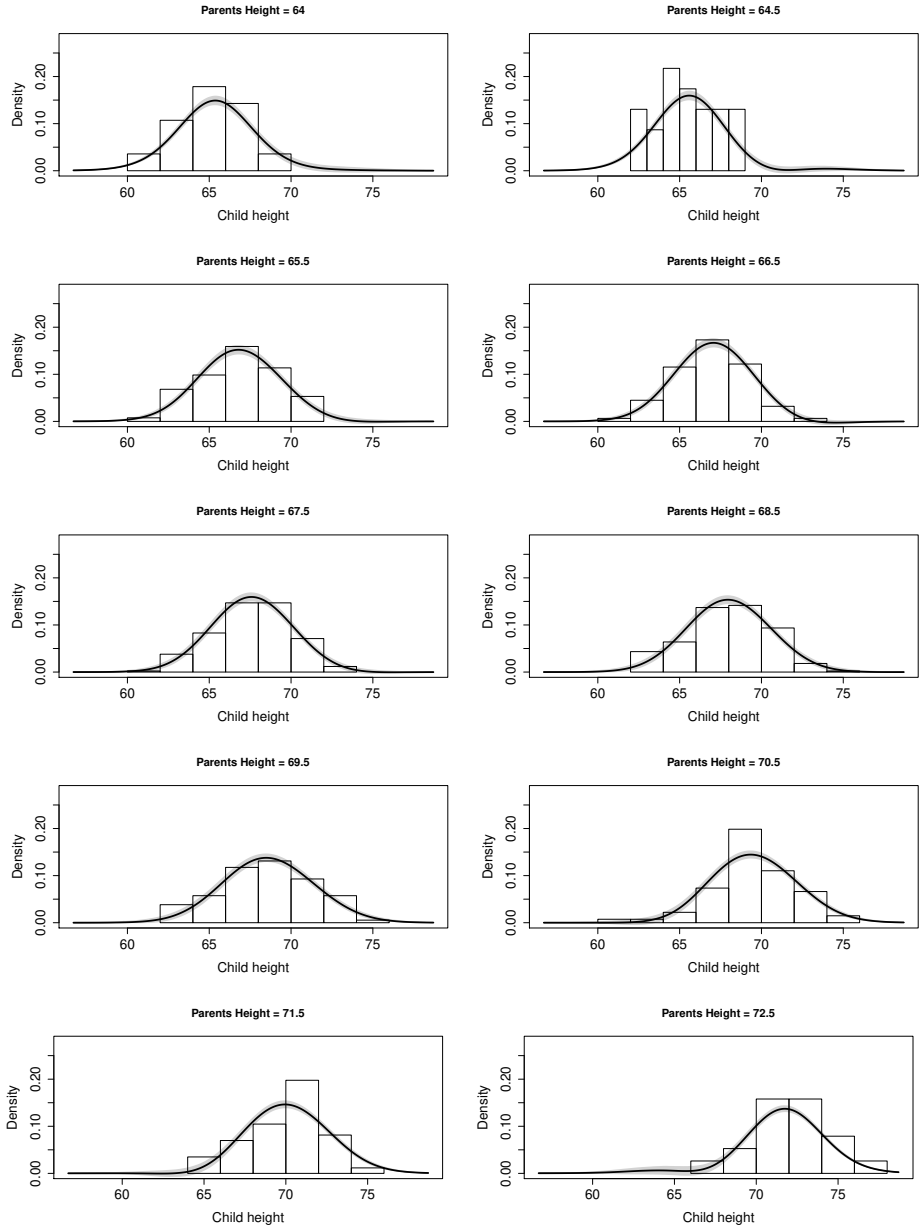
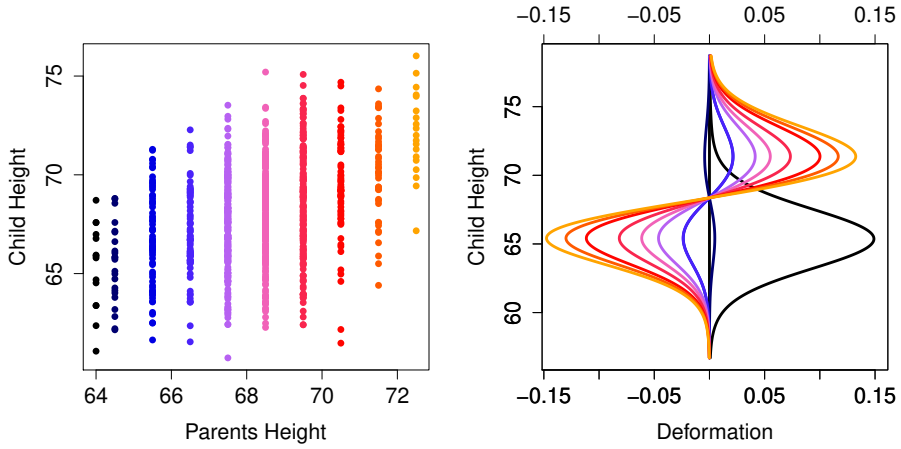


Figure 3.4: KLD density estimates (solid black line) for child height [in inches (in)], corresponding credible bands (grey), and histograms, with parents height ranging from 64in to 72.5in.



(a) (b)

Figure 3.5: (a) Galton's regression towards mediocrity data; (b) The solid black represents the baseline density estimate (f_1) through from a Karhunen–Loève–Dirichlet model; the remainder curves represent first component deformations of f_1, i.e., $(\theta_{k,1}-\theta_{1,1})g_1$, represented with the same palette as in (a). All heights are in inches.

Indeed, as it can be observed from Figure 3.5(b) the higher the parents height the further the need of transferring mass from the bulk of the baseline into the right tail.

3.5 Closing Remarks

We propose a data-driven Bayesian prior approach for modeling families of random densities, based in the Karhunen–Loève decomposition. This decomposition allowed us to induce dependence a priori between the members of the family, writing each one as a tilted version of the common mean f_μ. To satisfy the assumptions of Karhunen–Loève decomposition, theoretical properties are discussed to ensure that every member of the family belongs to L^2. Numerical experiment and simulation studies were conducted to assess the performance of our model against other Bayesian nonparametric models; such simulations suggest that the proposed method may overperform these competitors in three different simulation scenarios and different sample sizes. Finally, we illustrated our method using Galton's dataset on the

heights of parents.

3.6 Technical Details

3.6.1 Proofs of Main Results

Proof of Proposition 1. From Definition 11, if $\{f_k\} \sim \text{KL}\{f_k^*\}$ then

$$f_k = \frac{1}{K}\sum_{k=1}^{K} f_k^* + \sum_{j=1}^{J} \mathbb{E}[\theta_{j,k} \mid \boldsymbol{x}]\,\mathbb{E}[g_j \mid \boldsymbol{x}],$$

and thus

$$\begin{aligned}
\text{var}[f_k] &= \text{var}\left[\frac{1}{K}\sum_{k=1}^{K} f_k^* + \sum_{j=1}^{J} \mathbb{E}[\theta_{j,k} \mid \boldsymbol{x}]\,\mathbb{E}[g_j \mid \boldsymbol{x}]\right] \\
&= \text{var}\left[\frac{1}{K}\sum_{k=1}^{K} f_k^*\right] \\
&\overset{\text{ind}}{=} \frac{1}{K^2}\sum_{k=1}^{K} \text{var}[f_k^*].
\end{aligned}$$

Therefore,

$$\sum_{k=1}^{K} \text{var}[f_k] = \frac{1}{K}\sum_{k=1}^{K} \text{var}[f_k^*] < \sum_{k=1}^{K} \text{var}[f_k^*],$$

which completes the proof of part 1). The part 2) is based on the covariance properties, which implies that

$$\text{cov}(f_i, f_j) = \text{cov}(f_\mu, f_\mu) = \text{var}[f_\mu] = \frac{1}{K^2}\sum_{k=1}^{K} \text{var}[f_k^*] > 0,$$

which completes the proof of proposition. □

Proof of Proposition 2. Remember that

$$\mathbb{E}[g_j \mid \boldsymbol{x}] = \sum_{k=1}^{K} \omega_{k,j}\, \mathbb{E}[f_k^* \mid \boldsymbol{x}],$$

and thus,

$$\int_{\mathscr{X}} \mathbb{E}[g_j \mid \boldsymbol{x}]\mathrm{d}x = \sum_{k=1}^{K} \omega_{k,j} \left\{ \int_{\mathscr{X}} E[f_k^* \mid \boldsymbol{x}]\mathrm{d}x \right\} = \sum_{k=1}^{K} \omega_{k,j}.$$

But

$$\sum_{k=1}^{K} \omega_{k,j} = \frac{1}{\sum_{k=1}^{K} \theta_{k,j}^2} \sum_{k=1}^{K} \theta_{k,j} = 0,$$

by definition. Therefore,

$$\begin{aligned}
\int_{\mathscr{X}} f_k(x)\mathrm{d}x &= \int_{\mathscr{X}} \left\{ \frac{1}{K} \sum_{j=1}^{K} f_j^*(x) + \sum_{j=1}^{J} \mathbb{E}[\theta_{k,j} \mid \boldsymbol{x}] \cdot \mathbb{E}[g_j \mid \boldsymbol{x}] \right\} \mathrm{d}x \\
&= \frac{1}{K} \sum_{j=1}^{K} \underbrace{\left\{ \int_{\mathscr{X}} f_j^*(x)\mathrm{d}x \right\}}_{=1} + \sum_{j=1}^{J} \mathbb{E}[\theta_{k,j} \mid \boldsymbol{x}] \underbrace{\left\{ \int_{\mathscr{X}} \mathbb{E}[g_j \mid \boldsymbol{x}]\mathrm{d}x \right\}}_{=0} \\
&= 1,
\end{aligned}$$

for each $k = 1, \ldots, K$, then the result is follows. □

Proof of Proposition 3. Let $\mathscr{X} = \{x \in \mathbb{R} : \mathbb{K}(x \mid \boldsymbol{\theta}) > 0\}$. Note that

$$f_k^2(x) = \left\{ \int_{\Theta} \mathbb{K}(x \mid \boldsymbol{\theta}) H(\mathrm{d}\boldsymbol{\theta}) \right\}^2 \leq \int_{\Theta} \mathbb{K}^2(x \mid \boldsymbol{\theta}) H(\mathrm{d}\boldsymbol{\theta}), \tag{3.6.1}$$

by Jensen's inequality, thus

$$\int_{\mathscr{X}} f_k^2(x)\mathrm{d}x \leq \int_{\mathscr{X}} \int_{\Theta} \mathbb{K}^2(x \mid \boldsymbol{\theta}) H(\mathrm{d}\boldsymbol{\theta})\, \mathrm{d}x. \tag{3.6.2}$$

As $\mathbb{K}^2(x \mid \boldsymbol{\theta})$ is non–negative, Fubini's theorem (Durrett 2010, Theorem 1.7.2) can be applied

to the right hand-side expression on (3.6.2), yielding

$$\int_{\mathscr{X}} f_k^2(x)\mathrm{d}x \leq \int_{\Theta}\int_{\mathscr{X}} \mathbb{K}^2(x \mid \boldsymbol{\theta})\mathrm{d}x\, H(\mathrm{d}\boldsymbol{\theta}), \tag{3.6.3}$$

but, by A2, $\mathbb{K}(x \mid \boldsymbol{\theta}) \leq M$ for all $(x, \boldsymbol{\theta})$ in $\mathscr{X} \times \Theta$, and by A1,

$$\int_{\mathscr{X}} \mathbb{K}^2(x \mid \boldsymbol{\theta})\mathrm{d}x \leq M \int_{\mathscr{X}} \mathbb{K}(x \mid \boldsymbol{\theta})\mathrm{d}x = M. \tag{3.6.4}$$

Combining (3.6.3) and (3.6.4) yields

$$\int_{\mathscr{X}} f_k^2(x)\mathrm{d}x \leq M \int_{\Theta} H(\mathrm{d}\boldsymbol{\theta}) = M,$$

and thus $f_k \in L_1^2$. □

CHAPTER 4

Computing and Implementations

This chapter offers some details about computational aspects, including posterior sampling algorithms, selected comments on implementations, and specifications of the virtual machine instance from the Google Cloud Platform used to run the Monte Carlo simulations. Appendix C includes supplementary materials.

4.1 Selected Comments on Posterior Sampling

Fitting Phase-Varying Point Processes

For fitting the Bernstein–Dirichlet model from Chapter 2, we use the hybrid Gibbs sampler by Petrone (1999b), which is implemented in the `R` package `DPpackage` (Jara et al. 2011). In Algorithm 1 we present the Gibbs sampler used to conduct posterior sampling for each conditional mean measure Λ_i; the algorithm is based in the following hierarchical structure

- $k \sim p(k)$,
- $G \sim \mathrm{DP}(\alpha, G_0)$,
- $z_1, \ldots, z_n \mid k, G \overset{\text{i.i.d.}}{\sim} F$,
- $x_1, \ldots, x_n \mid k, G, z_1, \ldots, z_n \overset{\text{i.i.d.}}{\sim} \prod_{i=1}^{n} \sum_{j=1}^{k} \beta(x_i \mid j, k-j+1) \mathbb{1}_{((j-1)/k, j/k)}(z_i)$.

Algorithm 1: Gibbs sampler

---

1. *sampling* k:

$$k^{(s)} \mid \text{else} \sim p(k) \prod_{i=1}^{n} \beta(x_i \mid \eta(z_i^{(s-1)}, k^{(s-1)}), k^{(s-1)} - \eta(z_i^{(s-1)}, k^{(s-1)}) + 1),$$

   where $\eta(z, k) = i$ if $(i-1)/k < z \leq i/k$, for $i = 1, \ldots, n$.

2. *sampling* Z: We use $b(x \mid k, G)$ as defined in (1.2.6).

   - $z_i^{(s)} \mid \text{else} \sim g_0(z_i^{(s-1)}) \beta(x_i \mid \eta(z_i^{(s-1)}, k^{(s-1)}), k^{(s-1)} - \eta(z_i^{(s-1)}, k^{(s-1)}) + 1)$, with probability

$$q_{i,0} \propto \alpha\, b(x_i \mid k, G_0),$$

   - $z_i^{(s)} \mid \text{else} = z_j^{(s-1)}$, with probability

$$q_{i,j} \propto \beta(x_i \mid \eta(z_i^{(s-1)}, k^{(s-1)}), k^{(s-1)} - \eta(z_i^{(s-1)}, k^{(s-1)}) + 1).$$

Fitting Karhunen–Loève Prior-Based Model

For fitting densities through the DPM model from Chapter 3, we use a finite approximation of the Dirichlet process as the mixing measure. This allows us to use the blocked Gibbs sampling method proposed by Ishwaran and James (2002), which is implemented in the `R` package `ROCstudio`. In Algorithm 2 we present the blocked Gibbs sampler used to conduct posterior sampling; in the algorithm, N denotes the truncation level and S_i is the cluster indicator.

Algorithm 2: Blocked–Gibbs sampler

---

1. *Multinomial sampling*: Allocate observations to component mixtures with

$$P(S_i = h \mid \text{else}) = \frac{\pi_h \exp\{-1/2(Y_{i,k} - \mu_h)^2/\sigma_h^2\}}{\sum_{j=1}^{N} \pi_j \exp\{\exp\{-1/2(Y_{i,k} - \mu_h)^2/\sigma_h^2\}},$$

   and compute

$$n_h = \sum_{i=1}^{n_k} I(S_i = h), \quad n_h^+ = \sum_{i=1}^{n_k} I(S_i > h).$$

2. *Beta sampling*: Update stick-breaking weights

$$V_h \mid \text{else} \overset{\text{ind}}{\sim} \text{Beta}(1 + n_h, \alpha + n_h^+).$$

3. Update parameters of component mixtures:

$$\mu_h \mid \text{else} \overset{\text{ind}}{\sim} N\left(\frac{n_h}{(n_h\sigma_h^2 + \sigma^{-1})}\left(\frac{m}{s} + \sum_{\{i:S_i=h\}} \frac{Y_i}{\sigma_h^2}\right), \frac{1}{n_h/\sigma_h^2 + \sigma^{-1}}\right),$$

$$\sigma_h^{-1} \mid \text{else} \overset{\text{ind}}{\sim} \text{Gamma}\left(a + \frac{n_h}{2}, b + \frac{1}{2}\sum_{\{i:S_i=h\}} (Y_{i,k} - \mu_h)^2\right),$$

   where m, s, a and b are hyperparameters.

4.2 Selected Comments on Implementations

Fitting Phase-Varying Point Processes

Now, we show how to implement in `R` the methodology proposed in Chapter 2. I have wrote the package `Rmpp` to implement the proposed methods; part of the code uses a function earlier available from the `R` package `DPpackage` (namely, `BDPdensity`).

We start with the code for simulating a phase-varying process, following the scenario described in Section 2.3.1:

```
## Generating raw data
require(Rmpp)


set.seed(6789)
N <- 3 ; points <- 150 ; lgrid <- 2^8
grid <- seq(0,1,length.out = 2^8)
PARALLEL <- TRUE ; CORES <- parallel::detectCores() - 1


Npoints <- rpois(N,points); Y <- matrix(NA,nrow=max(Npoints),ncol=N)
for(i in 1:N)
  Y[1:Npoints[i],i] <- rnorm(Npoints[i],mean = 1/2,sd = .15)
```

Then, we can sample random warping functions T_i described in Section 2.3.1, and thus we can obtain the warped point processes with the following code:

```
## Generating warps maps and warped data
ar <- c(); part <- sample(c(1,2),2,replace = T)
for(k in 1:2){
  if(part[k] == 1)
    ar[k] <- runif(1,3/4,1)
  else
    ar[k] <- runif(1,0,1/4)
```

```
}
br <- sample(c(1,2),2,replace = TRUE)
warping <- function(x,k){
  if(k < N)
    a <- x - (ar[k]-1/2)*sin(x*pi*br[k])/(pi*br[k])
  else
    a <- x + (ar[1]-1/2)*sin(x*pi*br[1])/(pi*br[1]) +
      (ar[2]-1/2)*sin(x*pi*br[2])/(pi*br[2])
  return(a)
}
functions <- matrix(NA,nrow=lgrid,ncol=N); data <- matrix(NA,nrow=max(Npoints),ncol=N)
for(k in 1:N){
  data[1:Npoints[k],k] <- warping(na.omit(Y[,k]),k)
  functions[,k] <- warping(xx,k)
}
```

The code above generates simulated warped data. We now fit the proposed method by using the function `BAlignment` in `Rmpp` package as follows:

```
## Fitting Phase-Vaying Point Processes
mcmc <- list(nburn=500,nsave=4500,nskip=0,ndisplay=100)
prior <- list(aa0=2,ab0=2,kmax=1000,a0=1,b0=1)

fit <- BAlignment(data,prior = prior,mcmc = mcmc,grid = seq(0,1, length = 2^8),
                  parallel = TRUE, objective = 1)
```

Our package uses a modification of the `BDPdensity` function from the `R` package `DPpackage`, which allow us to extract the posterior trajectories rather than only the posterior mean.

The output of this function depends on the arguments (see Appendix C), but in the case that we show above the result is a 3D-array containing posterior simulated trajectories of the registered point process. Taking other values for arguments, we can recover the trajectories

of the warp maps or the Frechét mean.

A charts similar to Figure 2.1 can now reproduced using the following code:

```
plot(Y,fit, objective = 1, k = 0)
```

Fitting Karhunen–Loève Prior-Based Model

In this section we explain how to fit in R the model proposed in Chapter 3. The example replicates the fit from Section 3.3.1. I have produced a function which will be available from the R package ROCstudio. Prior to fitting the model I start by loading packages and by defining the simulation scenario.

The data generating mechanism is described in detail in Section 3.3.1, so we shown our code to implement that mechanism.

```
## Generating raw data
require(ROCstudio)


Ndata <- 500; Nfun <- 10; Nmix <- 3
w_1 <- seq(0.32,0.5,.02); w_2 <- seq(.2,.29,.01)
W <- cbind(w_1,w_2,1-w_1-w_2)
mu <- c(-2,0,4); sigma <- sqrt(c(1,.25,1))
lgrid <- 2^8; grid <- seq(-6,8,length.out = lgrid)
T <- 4500


set.seed(8)
Y=matrix(NA,nrow=Ndata,ncol=Nfun)
  for(i in 1:Nfun){
    Sam <- sample(Nmix,Ndata,replace = TRUE,prob = c(W[i,]))
    Y[,i] <- rnorm(Ndata,mu[Sam],sigma[Sam])
  }
```

Now we present the fit of proposed model in Chapter 3. The implementation requires the algorithm presented in Section 3.2.3, which combined with the Karhunen–Loève decomposition allows us to sample from our prior. To fit the model from Chapter 3, I have designed the `dKLD` function which is available from `ROCstudio`.

```
### Fitting KLD model
prior <- list(alpha = 1,mu = 0,sigma = 100,a = .1,b = .1)
fit <- dKLD(y = Y,prior = prior,kernel = 'gaussian',N = 20,T = 5000,
            burn = 500,grid = grid)
```

The outputs of our function are documented in Appendix C, and include for example the posterior trajectories and posterior mean for: density estimation, scores, and functional principal components. Also we can recover the posterior trajectories and posterior mean for DPM model.

A chart similar to Figure 3.1 can be produced using the following code:

```
plot(fit)
```

Details on Monte Carlo Studies

The examples from previous sections correspond to one shot experiments. Although such experiments do not require a high-performing machine, when we repeat this process several times, we need more computing power. The simulation studies conducted in Chapters 2 and 3 were executed using the Google Cloud Platform (GCP). Google Cloud Platform is a cloud service provided by Google LLC which has a lot of services such as storage, virtual machine instances, AI tools, big data tools, and so on. GCP is one of the big three cloud platforms in the industry (AWS and Azure, the other two) and provide a limited one-year free subscription (limited by US$300 in the year). For our purposes, we use the virtual machine (VM) instance service to conduct the simulation studies simultaneously to reduce the execution time. We used two setups of instances:

- In Chapter 2, we used three VM instances with Linux SO, 30 Gb SSD, 4 vCPU and 16 Gb RAM each.

- In Chapter 3, we used four VM instances with Linux SO, 25 Gb SSD, 8 vCPU and 32 Gb RAM each.

The simulation studies were conducted using these setups and running the corresponding code through the Linux shell, meanwhile the one-shot experiment was conducted using the (web) Rstudio server installed in the VM instances mentioned above. For more information about GCP, see https://cloud.Google.com.

CHAPTER 5

Discussion

This chapter puts the main contributions into perspective, it discusses the implications of this research, and it formulates new questions to be addressed in future work.

5.1 Final Comments

This dissertation addresses problems that involve inferences for families of random densities. In this line, in Chapter 2 we propose a Bayesian semiparametric approach for modeling registration of multiple point processes, where we use the Bernstein–Dirichlet prior to induce a prior over the space of all warp maps. For this induced prior we derived theoretical properties on the support and (weak) posterior consistency under mild conditions. Also, numerical experiments and simulation studies was conducted to assess the performance of our model and comparing with the kernel-based approach by Panaretos and Zemel (2016). Finally, a real data application in climatology showcased our model in practice.

For Chapter 3, we propose a data-driven prior based in the Karhunen–Loève decomposition for modeling a family of random densities. Our approach uses the Karhunen–Loève decomposition to induce dependence between members of the family. Our prior borrows strength across the elements of the family, as every member of the family can be regarded as a tilted version of the mean (f_μ). In order to satisfy the Karhunen–Loève decomposition assumptions, we discuss some theoretical properties to ensure that each element in the family belongs to L^2; other theoretical properties of the proposed prior–such as a priori variance–are discussed. Numerical experiments and simulation studies were conducted to assess the performance of the proposed method against other Bayesian nonparametric options, such as the Dirichlet process mixture model and the dependent Dirichlet process mixture model; such simulation studies suggest that our model may overperform these competitors. Finally, we offer an illustration of the proposed prior using Galton's dataset on the heights of parents and their children.

5.2 Directions for Future Research

Package 1: Aligning Multiple Phase-Varying Point Processes

A natural question would be extending the research from Chapter 2 to the case of spatial point process supported on e.g., $[0,1]^D$ with $D > 1$, as explored by Boissard et al. (2015) and Zemel and Panaretos (2017); a natural extension of our paper to this setup would entail modeling the mean measures of the corresponding spatial point processes via multivariate Bernstein polynomials (Zheng et al. 2009). The computation of the empirical Fréchet–Wasserstein mean can no longer however be done in closed form, requiring numerical schemes (Peyré and Cuturi 2018). From a statistical viewpoint, another natural avenue for future research would be on modeling the phase-variation of point processes conditionally on a covariate, by resorting to predictor-dependent versions of the Bernstein–Dirichlet prior (Barrientos et al. 2017).

Package 2: Borrowing Strength over a Family of Random Densities

Future extensions of the methods developed here, could take advantage of the square-root representation of Bhattacharyya (1946), $\sqrt{f_k}$, as this would lead to densities on the so-called Hilbert unit sphere, i.e. $\|\sqrt{f_k}\| = \int_{\mathscr{Y}} f_k(y)\,\mathrm{d}y = 1$, for $k = 1, \ldots, K$. The corresponding Karhunen–Loève decomposition in this case would be

$$\sqrt{f_k(y)} = \sqrt{f_\mu(y)} + \sum_{j=1}^{J} \vartheta_{k,j} g_j(y), \tag{5.2.1}$$

with $\vartheta_{k,j}$ and g_j representing the scores and principal components underlying (5.2.1). While (5.2.1) is not confined to densities in L_1^2, the corresponding principal component, $\{\vartheta_{k,j} g_j(y)\}_{k=1}^K$, are operated on the space of square-root of densities, and thus their interpretation is not as straightforward as that of $\{\theta_{k,j} g_j(y)\}_{k=1}^K$.

In addition, the proposed approach extends naturally to a regression setting. Indeed, for contexts where the interest is on a family of conditional densities, say $\{f_k(y \mid x)\}_{k=1}^K$, a conditional version of (3.2.4) could be written as

$$f_k(y \mid x) = f_\mu(y \mid x) + \sum_{j=1}^{J} \theta_{k,j}(x) g_j(y \mid x),$$

with the obvious notation. A Karhunen–Loève prior for the regression context could then be constructed by resorting to dependent Dirichlet process MacEachern (2000), rather than via Dirichlet process mixtures as in Chapter 3.

APPENDIX A

Supplementary Material for Chapter 2

A.1 Proofs of Auxiliary lemmas

Proof of Lemma 1. Since $\mathbb{F}$ is continuous $F_n \to \mathbb{F}$ pointwise. Let $\epsilon > 0$ and let $x < y$ such that $\mathbb{F}(x) \leq \epsilon$ and $\mathbb{F}(y) \geq 1 - \epsilon$. Since $\mathbb{F}$ is uniformly continuous on $[x, y]$ there exists a finite grid $x = x_1 < \cdots < x_k = y$ with $\mathbb{F}(x_i) \geq \mathbb{F}(x_{i+1}) - \epsilon$ for all $i \leq k - 1$. For n large $|F_n(x_i) - \mathbb{F}(x_i)| \leq \epsilon$ for all i so that

$$\sup_{z \in [x_i, x_{i+1}]} F_n(z) - \mathbb{F}(z) \leq F_n(x_{i+1}) - \mathbb{F}(x_i) \leq |F_n(x_{i+1}) - \mathbb{F}(x_{i+1})| + |\mathbb{F}(x_{i+1}) - \mathbb{F}(x_i)| \leq 2\epsilon.$$

In the same way

$$\sup_{z \notin [x,y]} |F_n(z) - \mathbb{F}(z)| \leq 2\epsilon, \qquad \sup_{z \in [x_i, x_{i+1}]} \mathbb{F}(z) - F_n(z) \leq \mathbb{F}(x_{i+1}) - F_n(x_i) \leq 2\epsilon,$$

and we conclude that $\|F_n - \mathbb{F}\|_\infty \leq 2\epsilon$ for n sufficiently large. □

Proof of Lemma 2. Since $\mathbb{F}$ is bijective, it has an inverse $\mathbb{F}^{-1}$. The latter is nondecreasing and, being a bijection, must also be continuous and with $\mathbb{F}^{-1}(0) = 0$, $\mathbb{F}^{-1}(1) = 1$. Let $p \in (0, 1)$, and let $x \in (0, 1)$ such that $\mathbb{F}(x) = p$. For $\epsilon \in (0, 1 - p)$ we have $F_n(x + \epsilon) \to$

$\mathbb{F}(x+\epsilon) > p$, which means that $x+\epsilon \geq F_n^{-1}(p)$ for n large. Similarly, $x-\epsilon \leq F_n^{-1}(p)$ for any $\epsilon \in (0,p)$ and all n large. This implies that $F_n^{-1}(p) \to x = \mathbb{F}^{-1}(p)$ for all $p \in (0,1)$. Since

$$0 \leq F_n^{-1}(0) \leq F_n^{-1}(p) \overset{n\to\infty}{\to} \mathbb{F}^{-1}(p) \overset{p\to 0}{\to} \mathbb{F}^{-1}(0) = 0,$$

it also follows that $F_n^{-1}(0) \to \mathbb{F}^{-1}(0)$. Similarly $F_n^{-1}(1) \to \mathbb{F}^{-1}(1)$ and we conclude that $F_n^{-1} \to \mathbb{F}^{-1}$ pointwise on $[0,1]$. By Lemma 1 the convergence is uniform. Convergence of sequences is equivalent to the statement of the lemma because the supremum norm defines a metric space.

Part b) is shown in the same way, since $(\mathbb{F}^{-1})^{-1} = \mathbb{F}$. There is a slight complication though because F_n^{-1} is only defined on $[F_n(0), F_n(1)]$ which may be a strict subinterval of $[0,1]$. Let $x = \mathbb{F}^{-1}(p)$ for $x, p \in (0,1)$. Then $F_n^{-1}(p-\epsilon) \to \mathbb{F}^{-1}(p-\epsilon) > x$ for $\epsilon > 0$ small, which means in particular that $F_n(0) \leq p-\epsilon$ and $F_n^{-1}(p-\epsilon)$ is defined, and also that $F_n(x) \leq p-\epsilon$ for n large. The inequality $F_n(x) \geq p+\epsilon$ is shown in the same way and we conclude that $F_n \to \mathbb{F}$ pointwise, and hence uniformly on $[0,1]$ by Lemma 1. □

Proof of Lemma 3. Since $\mathbb{F}$ is differentiable, there exists $x_j^* \in [j, j+1]/k$ such that

$$b(x \mid k, \mathbb{F}) = B(x \mid k-1, \mathbb{f}) + \sum_{j=0}^{k-1} \left[\mathbb{f}\left(x_j^*\right) - \mathbb{f}\left(\frac{j}{k-1}\right) \right] \binom{k-1}{j} x^j (1-x)^{k-1-j}.$$

Notice that $|x_j^* - j/(k-1)| \leq j/(k-1) - j/k \leq 1/(k-1) \to 0$ uniformly in j and as $\mathbb{f}$ is uniformly continuous on $[a-1/k, 1-a+1/k]$ for all $k > 1/a$, the sum at the right-hand side vanishes uniformly in $x \in [a, 1-a]$ as $k \to \infty$. If $\mathbb{f}$ is continuous on $[0,1]$ then it is uniformly continuous there and the sum at the right-hand side vanishes uniformly in $x \in [0,1]$. Since $B(x \mid k-1, \mathbb{f})$ converge to $\mathbb{f}$ uniformly, this completes the proof. □

A.2 Further Numerical Experiments

A.2.1 Supporting Outputs

In this section we present some figures which are derived from the simulation studies conducted in Section 3 in the paper. In details, Fig. A.1 refers to results in the simulation study in Section 3.1 (paper), Fig. A.2 refers to the comparison conducted in Section 3.2 (paper) and Fig. A.3 correspond to Fig. 5 (paper, left) but for all warp maps.

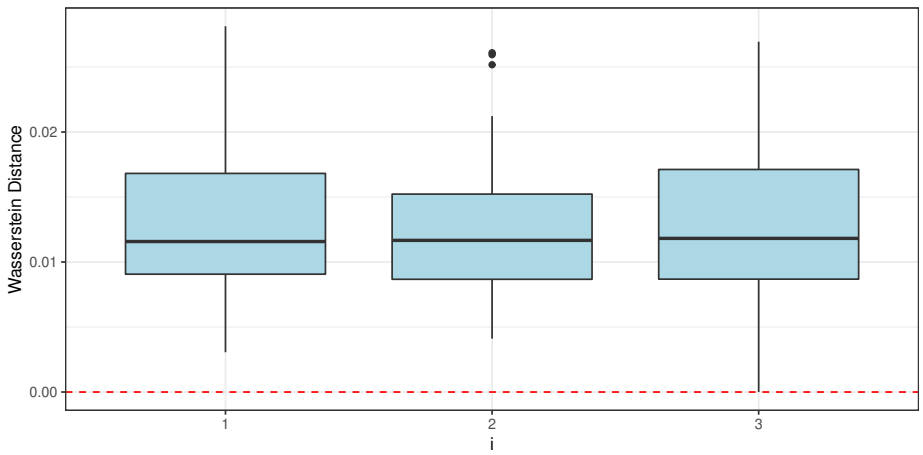


Figure A.1: Boxplots of the L^2-Wasserstein distance between the original processes $\Pi_i^{[b]}$ and the registered ones $\widehat{\Pi}_i^{[b]}$. Here b ranges from 1 to $B = 50$ and $i = 1, 2, 3$ correspond to the three panels.

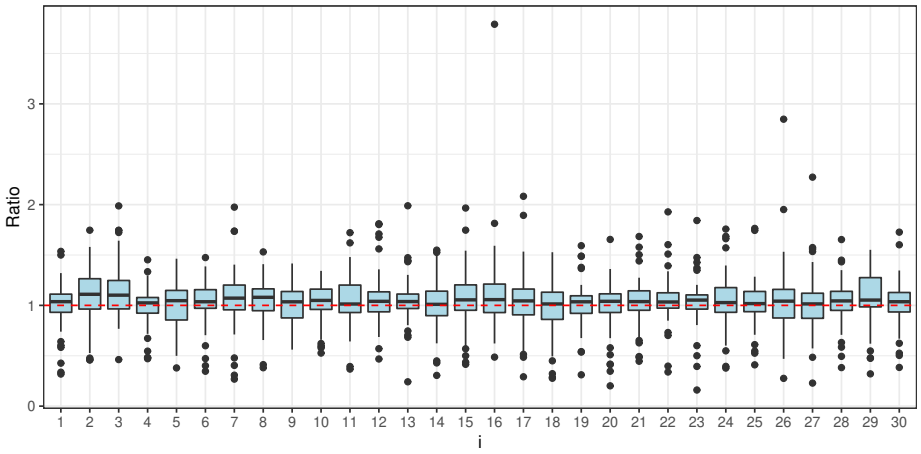


Figure A.2: Comparison of our Bayesian registration with the kernel-based registration of Panaretos and Zemel (2016). Each boxplot contains the ratio $d(\widehat{\Pi}_i^{[b,\text{Bayes}]}, \Pi_i^{[b]})/d(\widehat{\Pi}_i^{[b,\text{Kernel}]}, \Pi_i^{[b]})$ for all $i \in \{1, \ldots, 30\}$.

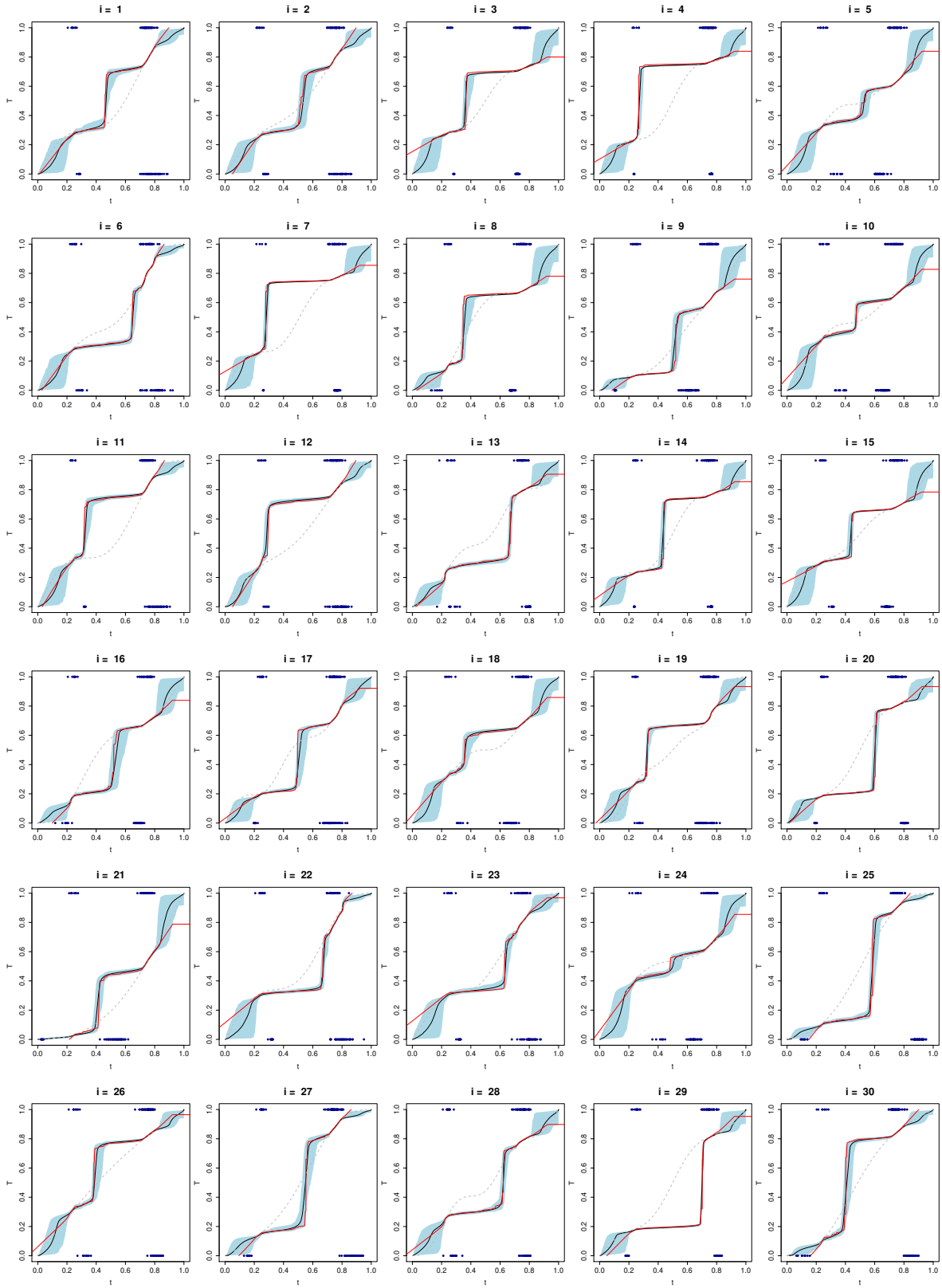


Figure A.3: 30 posterior mean Bernstein polynomial warp functions (solid black) and corresponding credible bands, with their kernel-based counterparts (solid red) and the original warp functions (dashed grey). Warp and original data are in the bottom and top, respectively.

A.2.2 Simulation Study under Misspecification

In this section, we analyze a simulation scenario similar to that we shown in Section 3.1 (paper) but this time using warp maps T_i which not satisfy $\mathbb{E}[T_i(t)] = t$, i.e., we try to assess the performance of our method under misspecification.

For this scenario, we generate random samples $x_{i,1}, \ldots, x_{i,m_i} \mid m_i$, from

$$f(t) = 0.45\,\{\phi(t \mid 0.25, 0.02^2) + \phi(t \mid 0.75, 0.03^2)\} + 0.1\,\beta(t \mid 1.5, 1.5), \quad m_i \sim \text{Poisson}(L), \quad i = 1, 2, 3,$$

with $L = 150$, $\phi(t \mid \mu, \sigma^2)$ denoting the normal density function and $\beta(t \mid a, b)$ denoting the beta density. Then the warped data $\tilde{x}_{i,j} = T_i(x_{i,j})$ are obtained using

$$T_i(t) = \int_0^t \beta(y \mid a, b)\,\mathrm{d}y, \quad i = 1, 2, \quad T_3(t) = 3t - T_1(t) - T_2(t), \quad a, b \overset{\text{iid}}{\sim} \text{Unif}[1, 3].$$

Here, it is direct that $\mathbb{E}[T_i(t)] \neq t$ for each $i = 1, 2, 3$. Fig. A.4 shows the estimators of each of the three warp maps through the posterior mean of the induced prior defined in Section 2.2 (paper), along with their credible bands and the true warp maps.

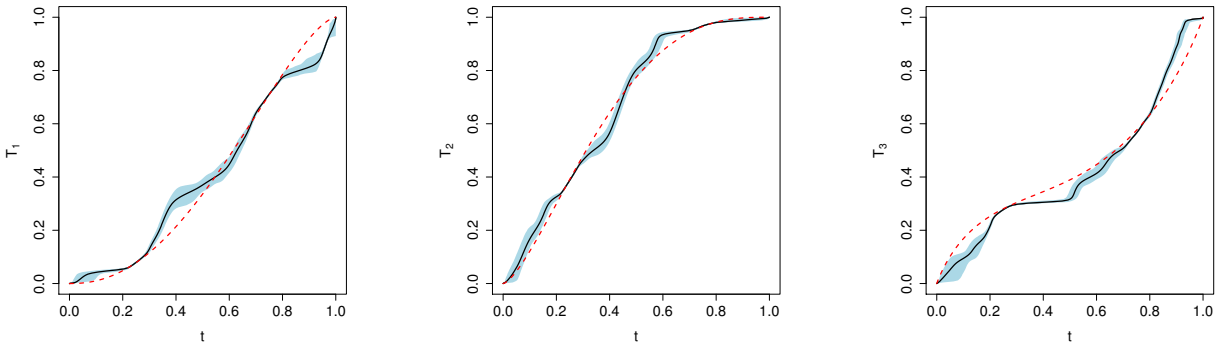


Figure A.4: True (dashed red) and estimated (solid black) warp functions along with credible bands. The estimators are constructed as the posterior mean of the induced prior.

From Fig. A.4 it can be observed that our estimators are reasonably in line with the true

warp functions even under misspecification, and as a consequence, the method recovers quite well the original point processes, as can be seen when comparing the left and right panels of Fig. A.5.

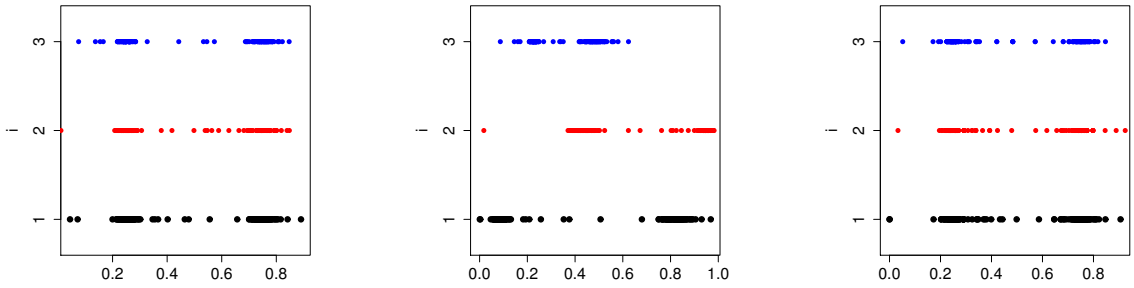


Figure A.5: Left: Realizations of the original point process from the setup of Section 1 (paper) in the small n, large m regime. Middle: Their corresponding phase-varying point process. Right: Their corresponding registered versions.

Also, a Monte Carlo study was conducted in this setting based on $B = 50$ simulated datasets, and we calculate the L^2-Wasserstein distance defined in Eq. 11 (paper) where we obtain that where the superscript $[b]$ denotes the corresponding object computed from the $\widehat{\text{WDM}} \approx 0.041677$ which is close to 0 and close to the value obtain in Section 3.1 (paper) in the well specify setting.

A.3 Additional outputs from Application

As in §4, we analyze the annual peaks over threshold, $\{\widetilde{x}_{i,j}^{+} \geq u_j^{+}\}$, and annual peaks below threshold, $\{\widetilde{x}_{i,j}^{-} \leq u_j^{-}\}$; we set the thresholds u_j^{+} and u_j^{-} using the 97.5% and 2.5% quantiles of temperature over year j, and this results in $m_1^{+}, \ldots, m_n^{+}$ ranging from 10 to 18 and $m_1^{-}, \ldots, m_n^{-}$ ranging from 10 to 20.

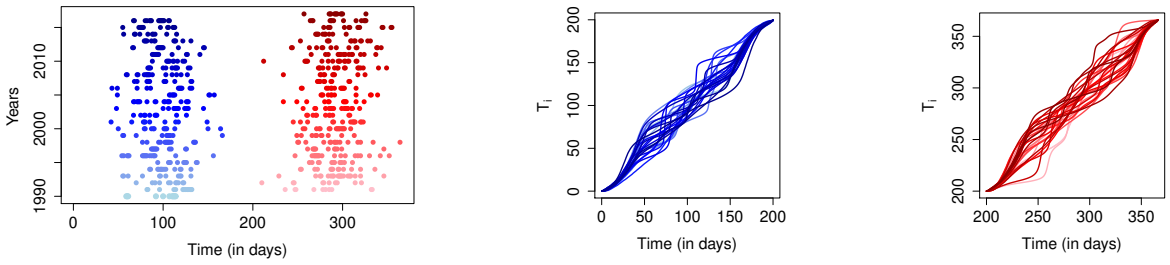


Figure A.6: Left: Point processes of annual peaks for peaks above (red) and below (blue) the thresholds. Middle and Right: Corresponding posterior mean warp functions in the same palette of colors for the 2.5% and 97.5% quantiles data.

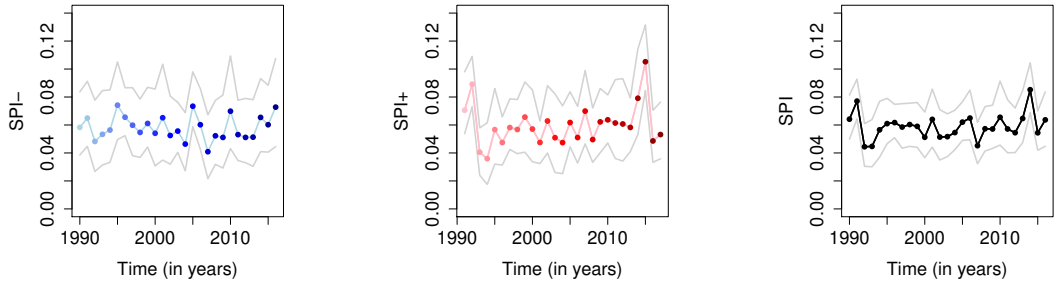


Figure A.7: Posterior mean SPI (scores of peak irregularity), as defined in (11), along with credible intervals, for below threshold (Left), above threshold (Middle), and global (Right), for the 2.5% and 97.5% quantiles data.

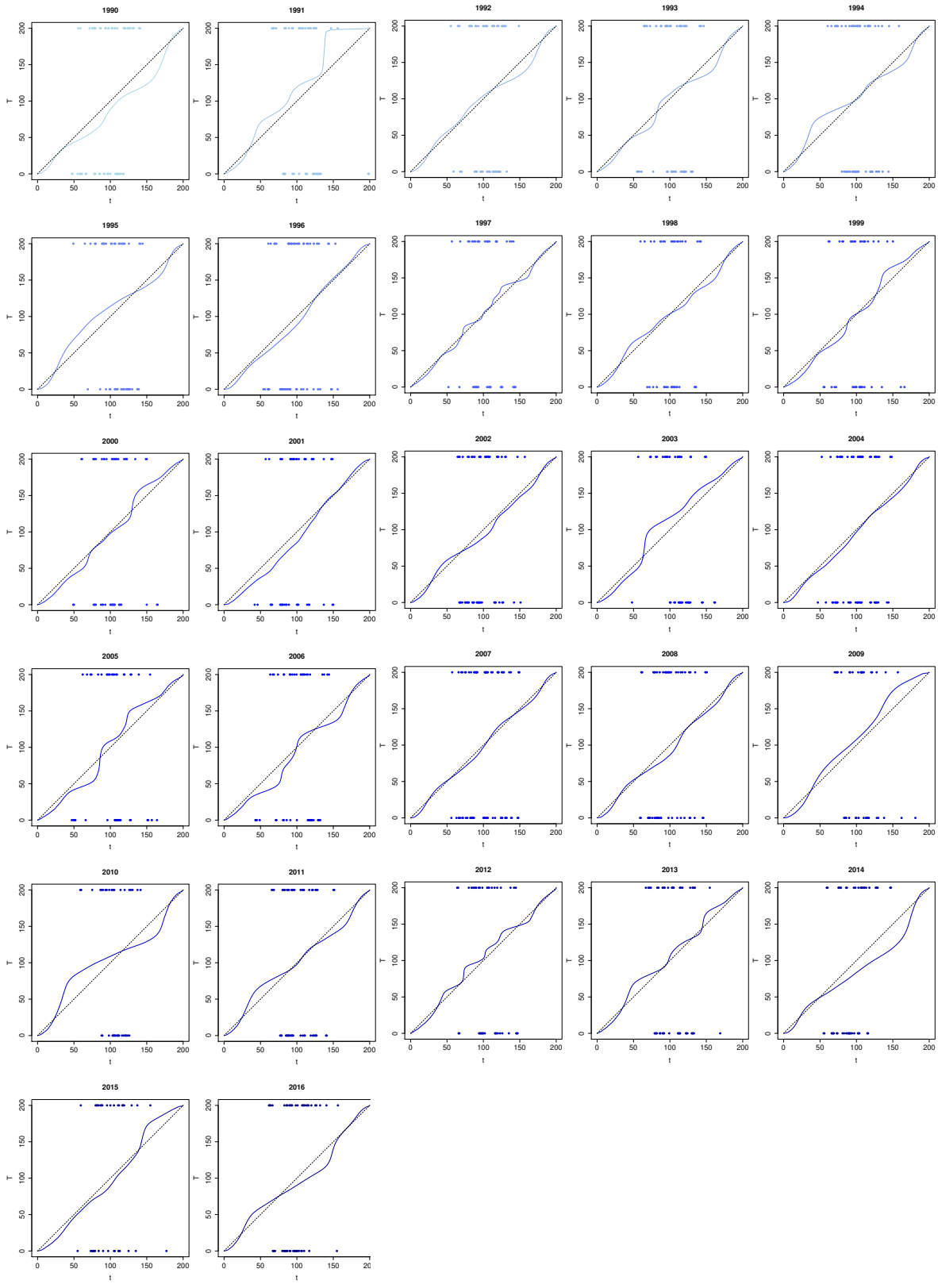


Figure A.8: Yearly posterior mean Bernstein polynomial warp functions of low-temperatures in the same color palette as data, plotted with raw data (bottom), registered points (top) and the identity function (dashed black).

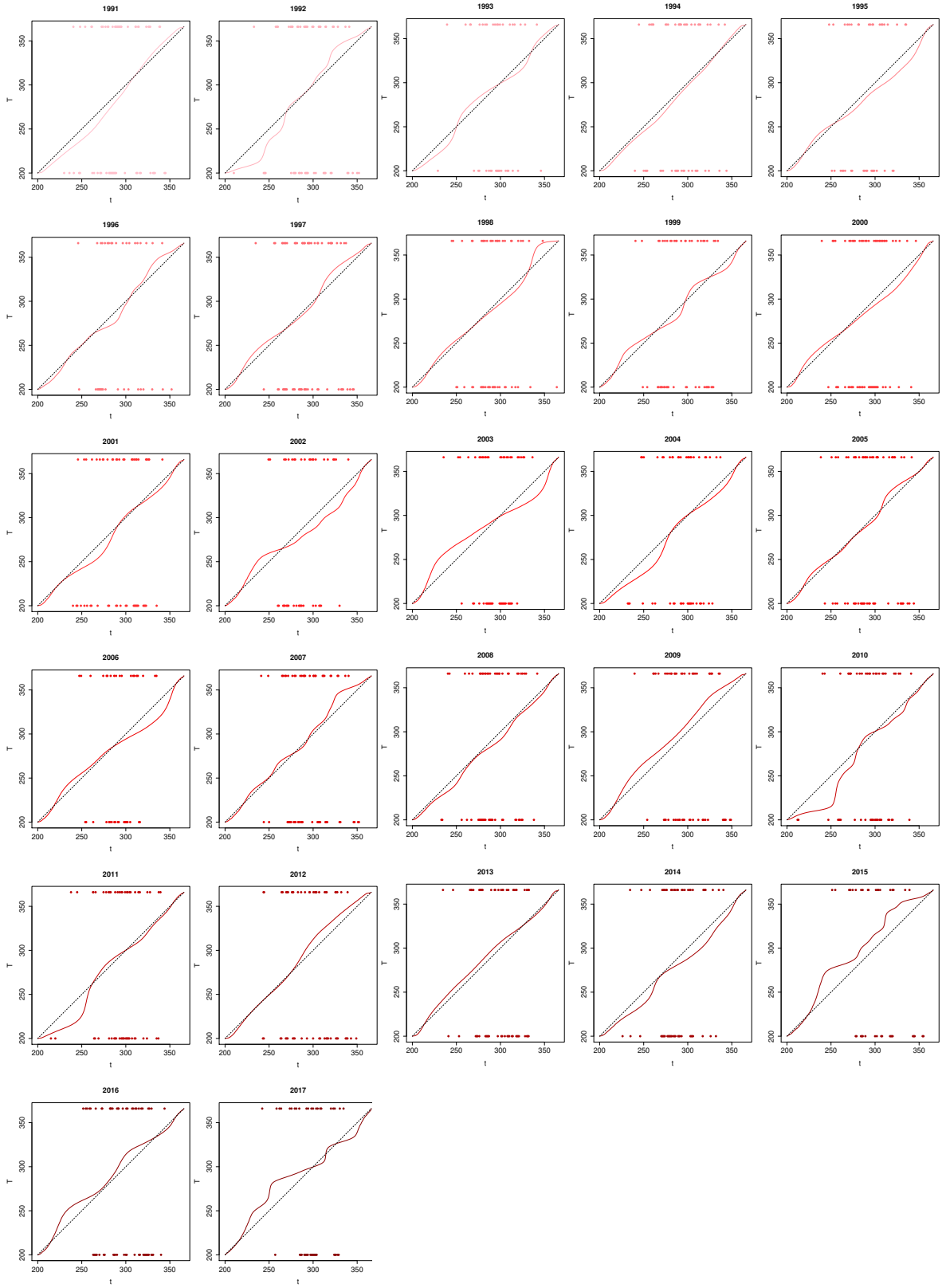


Figure A.9: Yearly posterior mean Bernstein polynomial warp functions of high-temperatures in the same color palette as data, plotted with raw data (bottom), registered points (top), and the identity function (dashed black). Here the year refers to that of onset of summer.

The Fig. A.6 and A.7 are the corresponding to the Fig. 6 and 7 in the paper, respectively; and Fig. A.8 and A.9 are the fit of each warp maps for the setting in the paper.

APPENDIX B

Supplementary Material for Chapter 3

B.1 Supporting Outputs

The following figures are analogous to Fig. 3.3 and 3.2 but considering $n = 1000$ instead $n = 500$.

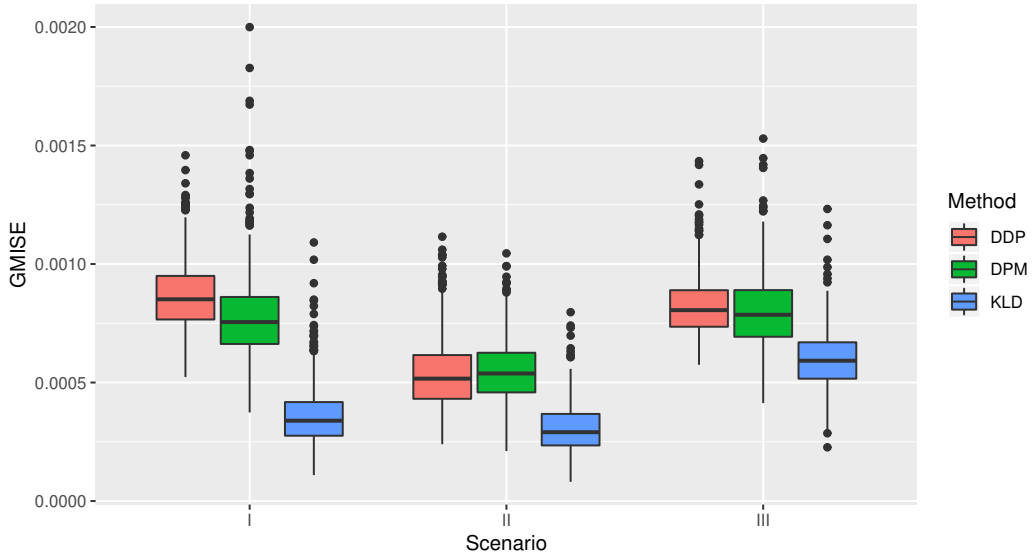


Figure B.1: Boxplot of global MISE for KLD, DPM and DDP estimates resulting from Monte Carlo study for Scenario I–III and $n = 1000$.

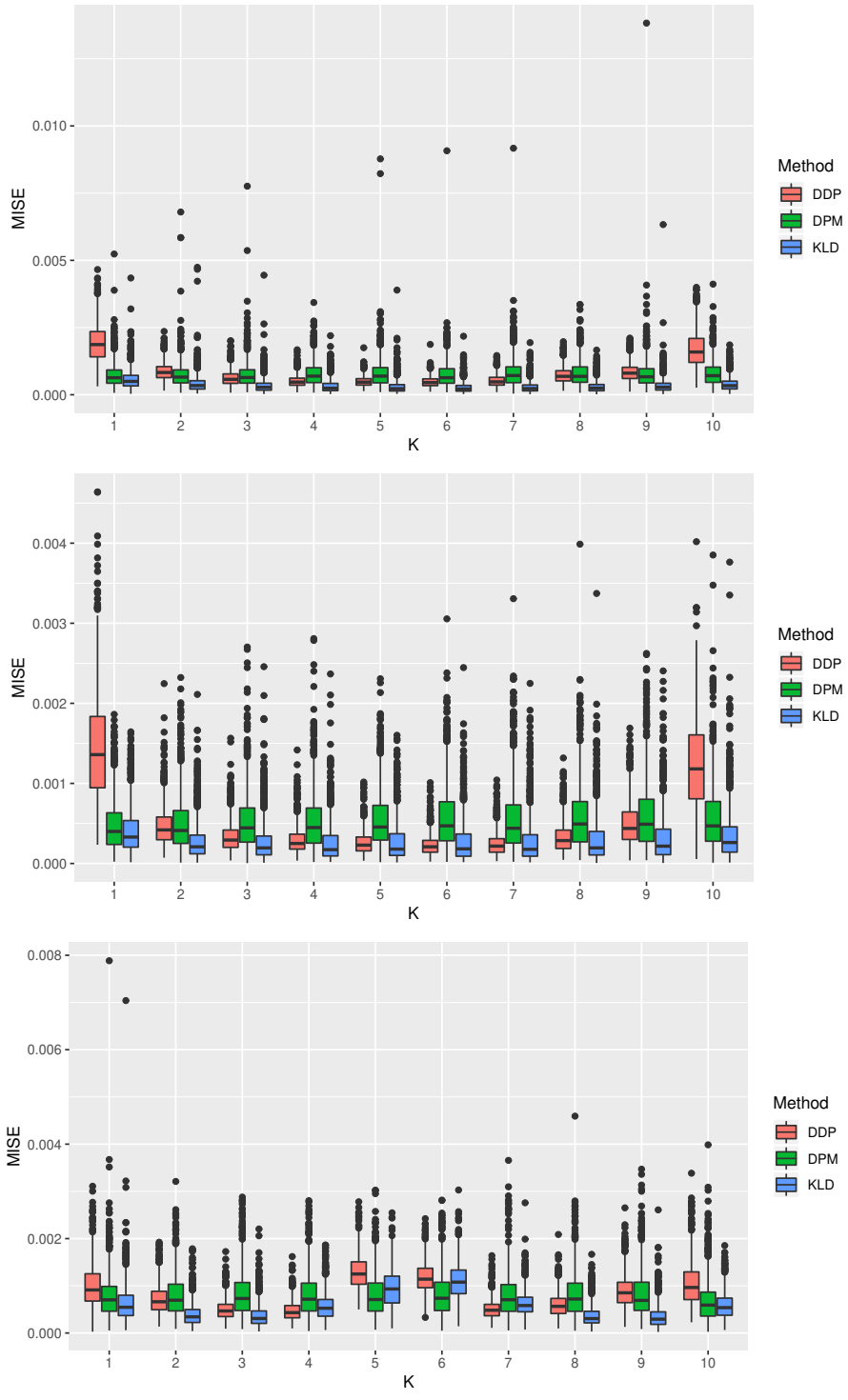


Figure B.2: Boxplot of MISE for KLD, DPM and DDP estimates resulting from Monte Carlo study, for each density. The plots are Scenario I–III from top to bottom with $n = 1000$.

B.2 Additional of Numerical Experiments

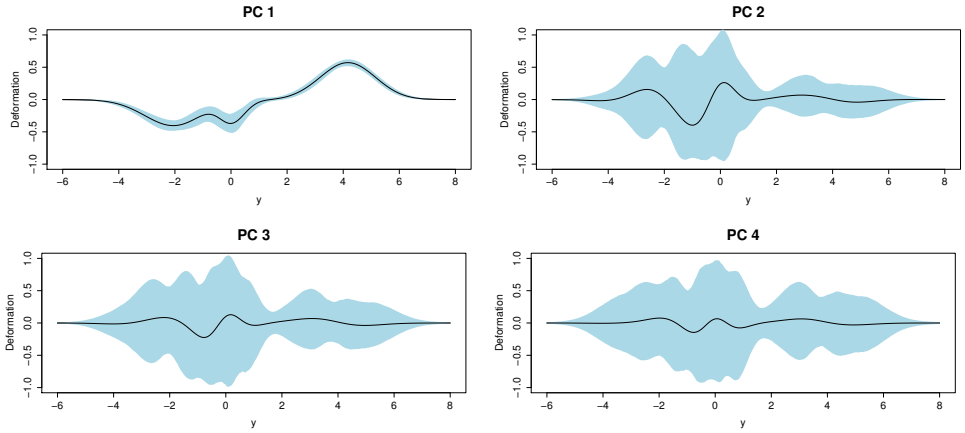


Figure B.3: First four principal components with their corresponding credible bands, for Scenario II as defined in Section 3.1 of the paper.

APPENDIX C

Supplementary material for Chapter 4

BAlignment {Rmpp} R Documentation

Bayesian phase variation alignment process

Description

Bayesian semi-parametric method for align several point processes with phase-variation.

Usage

```
## Default S3 method:
BAlignment(y, prior, mcmc, grid = seq(0,1, length = 2^8),
           parallel = TRUE, objective = 1)
```

Arguments

`y`	A matrix with data from which the alignment is to be computed. Different realizations in each column.
`prior`	a list with prior information for the Dirichlet-Bernstein polynomial prior. The list includes the following parameter: `aa0` and `ab0` giving the hyperparameters for prior distribution of the precision parameter of the Dirichlet process prior, `alpha` giving the value of the precision parameter (it must be specified if `aa0` is missing, see details below), `a0` and `b0` giving the parameters of the beta centering distribution of the DP prior, and `kmax` giving the maximum value of the discrete uniform prior for the degree of the Bernstein polynomial.
`mcmc`	a list giving the MCMC parameters. The list must include the following integers: `nburn` giving the number of burn-in scans, `nskip` giving the thinning interval, `nsave` giving the total number of scans to be saved, and `ndisplay` giving the number of saved scans to be displayed on screen (the function reports on the screen when every `ndisplay` iterations have been carried out).
`grid`	grid on which the posterior object of interest is to be evaluated; by default `grid = seq(0, 1, length = 2^8)`.
`parallel`	logical; if `TRUE`, the alignment process will be conducted using n - 1 cores, where n is the total number of (virtual) cores.
`objective`	a number indicating the process to be conducted. If `objective = 1` then alignment of point process will be conducted, if `objective = 2` then warp maps will be obtained and if `objective = 3` then Frechet mean will be obtained.

Details

This function fits a model for alignment point processes with phase variation, as described in Galasso, Zemel and de Carvalho (2019). The Balignment function use a fit of Dirichlet-Bernstein polynomial prior created by A. Jara (Jara et al, 2011)).

Value

`Align.traj`	3D-array containing posterior simulated trajectories of the registered point process (for `objective = 1`).
`Warp.traj`	3D-array containing posterior simulated trajectories of the warp maps (for `objective = 2`).
`Frechet.traj`	3D-array containing posterior simulated trajectories of the Frechet mean (for `objective = 3`).

Author(s)

Bastian Galasso-Diaz, Yoav Zemel and Miguel de Carvalho

References

Galasso, B., Zemel, Y. and de Carvalho, M. (2019). *Bayesian semiparametric modelling of phase-varying point processes*. Work in progress

Jara, A., Hanson, T., Quintana, F., Muller, P. and Rosner, G. (2011) DPpackage: Bayesian Semi- and Nonparametric Modeling in R. *Journal of Statistical Software*, **40**.

Examples

```
## Example 1: experiments on simulated data
# Initial values
```

[Package *Rmpp* version 1.0]

KLD {ROCstudio} R Documentation

KLD-Based Inference

Description

Karhunen–Loeve–Dirichlet-based inference for a family of density functions.

Usage

```
## dKLD(y, prior, ...)

## Default S3 method:
dKLD(y, prior, kernel = "gaussian", N = 20, T = 5000, burn = 500,
     grid = seq(min(y), max(y), length = 2^8))

pKLD(y, prior, kernel = "gaussian", N = 20, T = 5000, burn = 500,
     grid = seq(min(y), max(y), length = 2^8))
```

Arguments

`y`
: A matrix with data from which the estimate is to be computed. Different densities for each column.

`prior`
: prior information: (`alpha`, `mu`, `sigma`, `a`, `b`); see details.

`kernel`
: character string giving the smoothing kernel to be used; this must be one of `"gaussian"` or `"log-gaussian"`; by default `kernel = "gaussian"`.

`N`
: truncation value for the maximum number of mixture components (Ishwaran and James, 2001, 2002); by default, `N = 20`.

`T`
: number of MCMC iterations; by default: `T = 5000`.

`burn`
: number of burn in iterations; by default: `burn = 500`.

`grid`
: grid on which the posterior object of interest is to be evaluated; by default `grid = seq(min(y), max(y), length = 2^8)`.

Details

This function fits a model based on the Karhunen–Loeve–Dirichlet (KLD) prior as described in de Carvalho and Galasso (2019). The KLD model requires fitting a density for each level of a factor (here fitted using a DPM model with a blocked Gibbs sampler (Ishwaran and James, 2001, 2002)).

Value

`traj`
: 3D-array containing posterior simulated trajectories of densities or distribution function (`dKLD`) computed over `grid`.

`trajhat`
: matrix containing the mean trajectory of the estimated object of interest evaluated at grid.

`trajDPM`
: 3D-array containing posterior simulated trajectories of the density and distribution function computed over `grid`, for the DPM model.

`DPMhat`
: matrix containing the mean trajectory of the estimated object of interest evaluated at grid, for the DPM model.

`grid`
: grid on which the posterior object of interest has been evaluated.

`scores`
: matrix containing the estimated scores for the K-L decomposition.

`vartraj`
: matrix with the trajectories of the percentage of variability explained for each principal component.

`PC`
: matrix containing the estimated functional principal component for the K-L decomposition.

The `plot` method depicts the estimated objects [densities (`dKLD`) or distribution functions (`pKLD`) all in the same plot. If `k = j` then the method depicts the estimated densities along with its 95% credible bands. If `def = TRUE` then the method depicts a side-by-side plot with the data on the left and the baseline density (k=1) and all corresponding (1st) deformations on the right. The `summary` method depicts the variability explained for each one of the first j components along with their 95% credible intervals.

Author(s)

Miguel de Carvalho and Bastian Galasso-Diaz

References

de Carvalho, M. and Galasso, B. (2019). *Karhunen–Loeve Priors for families of random densities*. Work in progress

Ishwaran, H. and James, L. F. (2001) Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**, 161–173.

Ishwaran, H. and James, L. F. (2002) Approximate Dirichlet process computing in finite normal mixtures. *Journal of Computational and Graphical Statistics*, **11**, 508–532.

Examples

```
## Example 1: experiments on simulated data
# Initial values
n <- 500
Nf = 10
N <- 3
w_1 <-seq(0.32,0.5,.02)
w_2 <- seq(.2,.29,.01)
W=cbind(w_1,w_2,1-w_1-w_2)
mu <- c(-2,0,4)
sigma <- sqrt(c(1,.25,1))
lgrid = 2^8
T <- 4500
grid = seq(-6,8,length.out = lgrid)
# Simulating data
Y=matrix(NA,nrow=n,ncol=Nf)
  for(i in 1:Nf){
    Sam <- sample(N, n, replace = TRUE, prob = c(W[i,]))
    Y[,i] <- rnorm(n, mu[Sam], sigma[Sam])
  }
# prior specification and fir the KLD-model
prior <- list(alpha = 1, mu = 0, sigma = 100, a = .1, b = .1)
fit <- dKLD(y = Y, prior = prior,
            kernel = 'gaussian',
            N = 20, T = 5000,
            burn = 500,
            grid = grid)
# Plot of complete family
plot(fit)
# One element of the family along its 95% credible bands
plot(fit,k=1)
# Variability explained for the first four principal components
summary(fit,k=4)

## Example 2: Illustration with Galton's data
data(Galton)
attach(Galton)
lgrid <- 2^8
K <- nrow(table(Galton)) - 1
nk <- rowSums(table(Galton))
fit <- y <- y.grid <- list()
grid = seq(min(Galton$child) - 5, max(Galton$child) + 5,length=lgrid)
```

```
x <- c(64, 64.5, 65.5, 66.5, 67.5, 68.5, 69.5, 70.5, 71.5, 72.5)
set.seed(1)
for (k in 1:K)
  y[[k]] <- child[which(parent == x[k])] + rnorm(sum(parent == x[k]))

L <- unlist(lapply(y,FUN = length))
N <- max(L)
G <- matrix(NA,nrow = N,ncol=K)
for(k in 1:K)
  G[1:L[k],k] <- y[[k]]

prior <- list(alpha = 1, mu = 0, sigma = 100, a = .1, b = .1)
fit <- dKLD(y = G, prior = prior,
            kernel = 'gaussian',
            N = 20, T = 5000,
            burn = 500,
            grid = grid)
plot(fit)
plot(x = fit,y = G, cnames = x, def = TRUE,
  labs = c("Child Height","Parents Height",
    "Deformation","Parents Height"))
```

[Package *ROCstudio* version 1.0 Index]

Bibliography

Adams, R. P., Murray, I., and Mackay, D. (2009), "The Gaussian process density sampler," *Advances in Neural Information Processing Systems*, 9–16.

Agueh, M. and Carlier, G. (2011), "Barycenters in the Wasserstein space," *Soc. Ind. Appl. Math.*, 43, 904–924.

Antoniak, C. E. (1974), "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems," *Annals of Statistics*, 2, 1152–1174.

Barrientos, A. F., Jara, A., and Quintana, F. A. (2012), "On the support of MacEachern's dependent dirichlet processes and extensions," *Bayesian Analysis*, 7, 277–310.

— (2017), "Fully nonparametric regression for bounded data using dependent Bernstein polynomials," *Journal of the American Statistical Association*, 112, 806–825.

Barrios, E., Lijoi, A., Nieto-Barajas, L., and Prünster, I. (2013), "Modeling with Normalized Random Measure Mixture Models," *Statistical Science*, 28, 313–334.

Bhattacharyya, A. (1946), "On a measure of divergence between two multinomial populations," *Sankhyā: the Indian Journal of Statistics*, 401–406.

Blackwell, D. and MacQueen, J. (1973), "Ferguson distributions vio Pólya urn schemes," *The Annals of Statistics*, 1, 353–355.

Boissard, E., Le Gouic, T., and Loubes, J.-M. (2015), "Distribution's template estimate with Wasserstein metrics," *Bernoulli*, 21, 740–759.

Cifarelli, D. and Melilli, E. (2000), "Some new results for Dirichlet priors," *Annals of Statistics*, 28, 1390–1413.

Daley, D. and Vere-Jones, D. (2003/2008), *An introduction to the Theory of Point Processes. Volume I: Elementary Theory and Methods, Volumen II: General Theory and Structure*, Springer New Yorkç.

Dauxois, J., Pousse, A., and Romain, Y. (1982), "Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inferences," *Journal of Multivariate Analysis*, 12, 136–154.

de Carvalho, M., Barney, B., and Page, G. (2019a), "Affinity-Based Measures of Biomarker Performance Evaluation," *Statistical Methods in Medical Research*.

de Carvalho, M., Page, G. L., and Barney, J. B. (2019b), "On the geometry of Bayesian Inference," *Bayesian Analysis*, 14, 1013–1036.

Diaconis, P. and Kemperman, J. (1996), "Some new tools for Dirichlet priors," in *Bayesian Statistics*, eds. Bernardo, J., Berger, J., Dawid, P., and Smith, A., Oxford University Press, vol. 5, pp. 97–106.

Doss, H. and Selke, T. (1982), "The tails of probabilities chosen from a Dirichlet prior," *Annals of Statistics*, 4, 1302–1305.

Dunson, D. B. (2010), "Nonparametric Bayes applications to biostatistics," in *Bayesian Nonparametrics*, eds. Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G., Cambridge, UK: Cambridge University Press, pp. 223–273.

Durrett, R. (2010), *Probability Theory and Examples*, Cambridge University Press.

Escobar, M. (1988), "Estimating the means of several normal populations by nonparametric estimation of the distributions of the means," *Unpublished Doctoral Thesis, Department of Statistics, Yale University*.

— (1994), "Estimating normal means with a Dirichlet process prior," *Journal of the American Statistical Association*, 89, 268–277.

Escobar, M. and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 209–230.

— (1983), "Bayesian density estimation by mixtures of normal distribution," *Recent advances in statistics*, 287–302.

Ferraty, F. and Vieu, P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice*, New York: Springer.

Fristedt, B. (1967), "Sample function behavior of increasing processes with stationary independent increments," *Pacific Journal of Mathematics*, 21, 21–33.

Galton, F. (1886), "Regression towards mediocrity in hereditary stature," *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246–263.

Ghosal, S. (2010), "The Dirichlet process, related priors and posterior asymptotics," Cambridge, UK: Cambridge University Press, vol. 2, pp. 36–83.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. (1999), "Consistent semiparametric Bayesian inference about a location parameter," *Journal of Statistical Planning and Inference*, 77, 181–193.

Ghosal, S. and Van der Vaart, A. W. (2015), *Fundamentals of Nonparametric Bayesian Inference*, Cambridge University Press, Cambridge.

Good, I. (1969), "Some Applications of Singular Value Decomposition of a Matrix," *Technometrics*, 11, 823–831.

Grenander, U. (1950), "Stochastic processes and statistical inferences," *Arkiv för Matematik*, 1, 195–277.

Hanley, J. A. (2004), "«Transmuting» women into men: Galton's family data on human stature," *The American Statistician*, 58, 237–243.

Hanson, T. E. (2006), "Inference for mixtures of finite Polya tree models," *Journal of the American Statistical Association*, 101, 1548–1565.

Hartigan, J. (1998), "The maximum likelihood prior," *The Annals of Statistics*, 26, 2083–2103.

Hjort, N. L. (2003), "Topics in nonparametric Bayesian statistics," in *Highly Structured Stochastic Systems*, eds. Green, P., Hjort, N. L., and Richardson, S., pp. 455–487.

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010), *Bayesian Nonparametrics*, Cambridge, UK: Cambridge University Press.

Horváth, L. and Kokoszka, P. (2012), *Inference for Functional Data with Applications*, New York: Springer.

Huynh, K. P., Jacho-Chávez, D. T., Petrunia, R. J., and Voia, M. (2011), "Functional principal component analysis of density families with categorical and continuous data on Canadian entrant manufacturing firms," *Journal of the American Statistical Association*, 106, 858–878.

Inácio de Carvalho, V., de Carvalho, M., Alonzo, T. A., and González-Manteiga, W. (2016), "Functional covariate-adjusted partial area under the specificity-ROC curve with an application to metabolic syndrome diagnosis," *Annals of Applied Statistics*, 10, 1472–1495.

Inacio de Carvalho, V., de Carvalho, M., and Branscum, A. J. (2016), "Nonparametric Bayesian Regression Analysis of the Youden Index," .

Ishwaran, H. and James, L. (2001), "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*.

Ishwaran, H. and James, L. F. (2002), "Approximate Dirichlet process computing in finite normal mixtures: Smoothing and prior information," *Journal of Computational and Graphical Statistics*, 11, 508–532.

Jain, S. and Neal, R. M. (2004), "A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model," *Journal of Computational and Graphical Statistics*, 13, 158–182.

Jara, A., Hanson, T. E., Quintana, F. A., Müller, P., and Rosner, G. L. (2011), "DPpackage: Bayesian Semi-and Nonparametric Modeling in R," *Journal of Statistical Software*, 40, 1.

Kallenberg, O. (1983), *Random Measures*, Academic Press.

Karhunen, K. (1946), "Zur spektraltheorie stochastischer prozesse," *Annales Academiae Scientiarum Fennicae, Series A. I, Mathematica*.

Karr, A. F. (1991), *Point Processes and Their Statistical Inference*, Probability: Pure and Applied, New York: Dekker, 2nd ed.

Kingman, J. F. C. (1975), "Random discrete distributions," *Journal of the Royal Statistical Society, Ser. B*, 37, 1–22.

Kneip, A. and Utikal, K. J. (2001), "Inference for density families using functional principal component analysis (with Discussion)," *Journal of the American Statistical Association*, 96, 519–542.

Korwar, R. and Hollander, M. (1973), "Contributions to the theory of Dirichlet processes," *Annals of Probability*, 1, 705–711.

Lehmann, E. L. and Casella, G. (1998), *Theory of Point Estimation*, Springer texts in statistics, New York: Springer, 2nd ed.

Lehmann, E. L. and Romano, J. P. (2006), *Testing statistical hypotheses*, New York: Springer.

Lijoi, A., Mena, R. H., and Prünster, I. (2007), "Controlling the reinforcement in Bayesian non-parametric mixture models," *Journal of the Royal Statistical Society: Ser. B*, 69, 715–740.

Lo, A. (1984), "On a class of Bayesian nonparametric estimates I: Density estimates," *The Annals of Statistics*, 12, 351–357.

Loève, M. (1946), "Fonctions aléatoires à décomposition orthogonale exponentielle," *La Revue Scientifique*, 84, 159–162.

MacEachern, S. N. (1994), "Estimating normal means with a conjugate style Dirichlet process prior," *Communications in Statistics—Simulation and Computation*, 23, 727–741.

— (2000), "Dependent Dirichlet processes," *Unpublished manuscript, Department of Statistics, The Ohio State University*, 1–40.

Marron, J., Ramsay, J., Sangalli, L., and Srivastava, A. (2015a), "Functional data analysis of amplitude and phase variation," *Statistical Science*, 30, 468–484.

Marron, J. S., Ramsay, J. O., Sangalli, L. M., and Srivastava, A. (2015b), "Functional data analysis of

amplitude and phase variation," *Statistical Science*, 30, 468–484.

Menzel, A. and Fabian, P. (1999), "Growing season extended in Europe," *Nature*, 397, 659.

Müller, P., Erkanli, A., and West, M. (1996), "Bayesian curve fitting using multivariate normal mixtures," *Biometrika*, 83, 67–79.

Müller, P. and Mitra, R. (2013), "Bayesian nonparametric inference—why and how," *Bayesian Analysis*, 8, 269–302.

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015), *Bayesian Nonparametric Data Analysis*, New York: Springer.

Neal, R. M. (2000), "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, 9, 249.

Panaretos, V. M. and Zemel, Y. (2016), "Amplitude and Phase Variation of Point Processes," *Annals of Statistics*, 44, 771–812.

— (2019), "Statistical Aspects of Wasserstein Distances," *Annu. Rev. Stat. Appl.*, 6, 405–431.

Petrone, S. (1999a), "Bayesian density estimation using Bernstein polynomials," *The Canadian Journal of Statistics*, 27, 105–126.

— (1999b), "Random Bernstein Polynomials," *Scandinavian Journal of Statistics*, 26, 373–393.

Petrone, S. and Wasserman, L. (2002), "Consistency of Bernstein Polynomial Posteriors," *Journal of the Royal Statistical Society, Ser. B*, 64, 79–100.

Peyré, G. and Cuturi, M. (2018), *Computational Optimal Transport*, arXiv:1803.00567.

Phadia, E. G. (2015), *Prior Processes and their Applications: Nonparametric Bayesian Estimation*, New York: Springer.

Poisson, S. D. (1837), "Recherches sur la Probabilité des Judgements en Matière Criminelle et en Matière Civile, Précédées des Règles Générales du Calcul des Probabilités," *Barchelier, Paris.*

Poynor, V. and Kottas, A. (2019), "Nonparametric Bayesian inference for mean residual life functions in survival analysis," *Biostatistics*, 20, 240–255.

Quintana, F. A. and Müller, P. (2004), "Nonparametric Bayesian data analysis," *Statistical Science*, 19, 95–110.

Ramsay, J. O. (1982), "When the data are functions," *Psychometrika*, 47, 379–396.

— (2006), *Functional Data Analysis*, New York: Wiley.

Ramsay, J. O. and Dalzell, C. (1991), "Some tools for functional data analysis," *Journal of the Royal Statistical Society. Series B (Methodological)*, 539–572.

Ramsay, J. O. and Silverman, B. W. (2002a), *Applied Functional Data Analysis: Methods and Case Studies*, vol. 77, Citeseer.

— (2002b), *Applied Functional Data Analysis: Methods and Case Studies*, vol. 77, New York: Springer.

— (2005), *Functional Data Analysis*, New York: Springer, 2nd ed.

Rao, C. (1958), "Some statistical methods for comparison of growth curves," *Biometrics*, 14, 1–17.

Rodríguez, A. and Dunson, D. B. (2011), "Nonparametric Bayesian models through probit stick-breaking processes," *Bayesian Analysis*, 6, 145–177.

Rodriguez, A. and Martinez, J. C. (2014), "Bayesian semiparametric estimation of covariate-dependent ROC curves," *Biostatistics*, 15, 353–369.

Santambrogio, F. (2015), *Optimal Transport for Applied Mathematicians*, Basel: Birkäuser.

Schwartz, M. D., Ahas, R., and Aasa, A. (2006), "Onset of spring starting earlier across the Northern Hemisphere," *Global change biology*, 12, 343–351.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

Srivastava, A., Wu, W., Kurtek, S., Klassen, E., and Marron, J. (2011), "Registration of Functiona Data Using Fisher–Rao Metric," *arXiv:1103.3817*.

Stigler, S. M. (1986), *The History of Statistics: The Measurement of Uncertainty Before 1900*, Cambridge, MA: Harvard University Press.

Tang, R. and Müller, H.-G. (2008), "Pairwise curve synchronization for functional data," *Biometrika*, 95, 875–889.

Teh, Y. and Jordan, M. (2010), "Hierarchical Bayesian nonparametric models with applications," *Bayesian nonparametrics*, 1, 158–207.

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006), "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, 101, 1566–1581.

Wachsmuth, A., Wilkinson, L., and Dallal, G. E. (2003), "Galton's bend: A previously undiscovered nonlinearity in Galton's family stature regression data," *The American Statistician*, 57, 190–192.

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2015), "Review of functional data analysis," *arXiv:1507.05135*.

— (2016), "Functional data analysis," *Annu. Rev. Stat. Appl.*, 3, 257–295.

Wu, S., Müller, H., and Zhang, Z. (2013), "Functional data analysis for point processes with rare events," *Statistica Sinica*, 23, 1–23.

Wu, W. and Srivastava, A. (2014), "Analysis of spike train data: Alignment and comparisons using the extended Fisher–Rao metric," *Electron. J. Stat.*, 8, 1776–1785.

Xu, H., Carin, L., and Zha, H. (2017), "Learning Registered Point Processes from Idiosyncratic Observations," *arXiv:1710.01410*.

Zemel, Y. and Panaretos, V. M. (2017), "Fréchet Means and Procrustes Analysis in Wasserstein Space," *Bernoulli*, in press.

Zheng, Y., Zhu, J., and Roy, A. (2009), "Nonparametric Bayesian inference for the spectral density function of a random field," *Biometrika*, 97, 238–245.