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

Marco Henrique de Almeida Inacio

Comparing two populations using Bayesian Fourier series density estimation

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

Orientador: Prof. Dr. Rafael Izbicki

São Carlos
Abril de 2017

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia

Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a defesa de dissertação de mestrado do candidato Marco Henrique de Almeida Inácio realizada em 12/04/2017:

Prof. Dr. Rafael Izbicki
UFSCar

Prof. Dr. Danilo Lourenço lopes
UFSCar

Prof. Dr. Marcos Oliveira Prates
UFMG

Certifico que a sessão de defesa foi realizada com a participação à distância do membro Prof. Dr. Marcos Oliveira Prates e, depois das arguições e deliberações realizadas, o participante à distância está de acordo com o conteúdo do parecer da comissão examinadora redigido no relatório de defesa do(a) aluno(a) Marco Henrique de Almeida Inácio.

Prof. Dr. Rafael Izbicki
Presidente da Comissão Examinadora
UFSCar

This work is dedicated to the One that has made it all possible “`in fieri et in esse`”.

ACKNOWLEDGEMENTS

First of all, I would like to say thanks to my parents Nivaldo and Claúdia and my brother João Pedro.

Then to my advisor Rafael Izbicki and to professor Luis Ernesto Bueno Salasar (especially for the help with the introduction of chapter 2) for their really amazing help and dedication to this work.

And then to professors José Galvão Leite, Rafael Stern, Luís Gustavo Esteves, Gustavo Henrique de Araújo Pereira, Marcio Alves Diniz, Adriano Kamimura Suzuki, Sandro Gallo, Ricardo Sandes Ehlers, the professors of the evaluation board, the DE/UFSCar graduation secretary Izabel Araujo, and the members of Stan development team.

I would also like to give my acknowledgments to the financial support of CAPES and our nation's citizens, and hope that the present work will be useful to achieve common good.

And last but not least, my special thanks to the almighty God Jesus Christ and His immaculate mother.

“If the highest aim of a captain were to preserve his ship, he would keep it in port forever.”

(St. Thomas Aquinas)

RESUMO

INACIO, M. H. DE A. **Comparação de duas populações utilizando estimação bayesiana de densidades por séries de Fourier**. 2017. 56 p. Master dissertation (Master student joint Graduate Program in Statistics DEs-UFSCar/ICMC-USP) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

Dadas duas amostras de duas populações, pode-se questionar o quão parecidas as duas populações são, ou seja, o quão próximas estão suas distribuições de probabilidade. Para distribuições absolutamente contínuas, uma maneira de mensurar a proximidade dessas populações é utilizando uma medida de distância (métrica) entre as funções densidade de probabilidade (as quais são desconhecidas, em virtude de observarmos apenas as amostras). Nesta dissertação, utilizamos a distância quadrática integrada como métrica. Para mensurar a incerteza da distância quadrática integrada, primeiramente modelamos a incerteza sobre cada uma das funções densidade de probabilidade através de uma método bayesiano não paramétrico. O método consiste em estimar a função de densidade de probabilidade f (ou seu logaritmo) usando séries de Fourier $\{\phi_0, \phi_1, ..., \phi_I\}$. Atribuir uma distribuição a priori para f é então equivalente a atribuir uma distribuição a priori aos coeficientes dessa serie. Utilizamos a priori sugerida em Scricciolo (2006) (priori de sieve), a qual não coloca uma priori somente nesses coeficientes, mas também no próprio I, de modo que, na realidade, trabalhamos com uma mistura bayesiana de modelos de dimensão finita. Para obter amostras a posteriori dessas misturas, marginalizamos o parâmetro (discreto) de indexação de modelos, I, e usamos um software estatístico chamado Stan. Concluímos que o método bayesiano de séries de Fourier tem boa performance quando comparado ao de estimativa de densidade kernel, apesar de ambos os métodos frequentemente apresentarem problemas na estimação da função de densidade de probabilidade perto das fronteiras. Por fim, mostramos como a metodologia de series de Fourier pode ser utilizada para mensurar a incerteza a cerca da similaridade de duas amostras. Em particular, aplicamos este método a um conjunto de dados de pacientes com doença de Alzheimer.

Palavras-chave: séries de Fourier, séries ortogonais, estimação de densidade, stan, amostragem discreta.

ABSTRACT

INACIO, M. H. DE A. **Comparing two populations using Bayesian Fourier series density estimation**. 2017. 56 p. Master dissertation (Master student joint Graduate Program in Statistics DEs-UFSCar/ICMC-USP) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

Given two samples from two populations, one could ask how similar the populations are, that is, how close their probability distributions are. For absolutely continuous distributions, one way to measure the proximity of such populations is to use a measure of distance (metric) between the probability density functions (which are unknown given that only samples are observed). In this work, we work with the integrated squared distance as metric. To measure the uncertainty of the squared integrated distance, we first model the uncertainty of each of the probability density functions using a nonparametric Bayesian method. The method consists of estimating the probability density function f (or its logarithm) using Fourier series $\{\phi_0, \phi_1, ..., \phi_I\}$. Assigning a prior distribution to f is then equivalent to assigning a prior distribution to the coefficients of this series. We used the prior suggested by Scricciolo (2006) (sieve prior), which not only places a prior on such coefficients, but also on I itself, so that in reality we work with a Bayesian mixture of finite dimensional models. To obtain posterior samples of such mixture, we marginalize out the discrete model index parameter I and use a statistical software called Stan. We conclude that the Bayesian Fourier series method has good performance when compared to kernel density estimation, although both methods often have problems in the estimation of the probability density function near the boundaries. Lastly, we showed how the methodology of Fourier series can be used to access the uncertainty regarding the similarity of two samples. In particular, we applied this method to dataset of patients with Alzheimer.

Keywords: fourier series, orthogonal series, density estimation, stan, discrete sampling.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

HMC	Halmiltonian Monte Carlo
KDE	kernel density estimation
MAE	mean absolute error
MCMC	Markov Chain Monte Carlo
MSE	mean squared error
NUTS	No-U-Turn Sampler
PDF	probability density function

CONTENTS

CHAPTER 1

INTRODUCTION

Given two samples D_1 and D_2 from two (possibly different) populations, one could question how similar the populations are, that is, how close are the probability distributions from which each sample came from. Soriano (2015) and Holmes *et al.* (2015) among others have addressed that question by developing methods to do Bayesian nonparametric hypothesis tests via Bayes Factor.

In this dissertation, on the other hand, we propose a different approach that is promising in its intuitiveness: for absolutely continuous distributions on $[0,1]$, one way to measure the proximity of such populations, is to use a measure of distance (metric) between the probability density function (PDF) f_1 of the first distribution and the PDF f_2 of the second. In this dissertation, we'll work with what we called integrated squared distance:

$$\mathbb{M}(f_1, f_2) = \int_0^1 (f_1(y) - f_2(y))^2 \mathrm{d}y \tag{1.1}$$

However, to measure the posterior uncertainty of $\mathbb{M}(f_1, f_2)$, we first need to measure the posterior uncertainty of f_1 and f_2, driving us towards a problem of density estimation which is among the most significant problems in Statistics. Thanks to the abundance of data in many applications, today there is a great emphasis on nonparametric methods. In particular, Bayesian nonparametric methods have gained great notoriety lately. Among the most used nonparametric Bayesian density estimation methods, we have finite mixtures or Dirichlet process mixtures (GELMAN *et al.*, 2014), Polya trees (LAVINE, 1992), Bernstein polynomials (PETRONE, 1999; PETRONE; WASSERMAN, 2002) and wavelets (MÜLLER; VIDAKOVIC, 1998). Unfortunately, many of these methods are complex and difficult to interpret, causing technical difficulties like having priors that are hard to elicit.

In this dissertation, we work with an approach that is common in the frequentist literature, but despite being very simple and intuitive, has received little attention in literature as a Bayesian density estimation method, being reduced, most of the times, to theoretical objectives (SCRICCIOLO, 2006; RIVOIRARD; ROUSSEAU *et al.*, 2012). The approach consists of estimating a PDF f (or its logarithm) using Fourier Series $\{\phi_0, \phi_1, ..., \phi_I\}$. Assigning a prior distribution to f is then equivalent to assigning a prior distribution to the coefficients of this series.

We use the prior suggested at Scricciolo (2006) (sieve prior), which not only places a prior on such coefficients, but also on I itself, so that in reality we work with a Bayesian mixture of finite dimensional models. To obtain posterior samples of such mixture we use a statistical software called Stan, but to be able to do so, we first marginalize out the discrete model index parameter I.

Therefore, in this work, we propose first to investigate how this approach performs compared to the kernel density estimation in simulated datasets and then, use it to check how the density estimation method perform to measure the distance.

The rest of this work follows the following structure: in chapter 2 we present the theoretical details of the Fourier series estimation methods. In chapter 3 we provide a short introduction to Stan and provide a description of the method used for discrete sampling with Stan and the sampling algorithms that the software implements. In chapter 4 we present a empirical study to assess the performance of the method of density estimation compared to KDE. In chapter 5 we present another empirical study, but in this case, with the goal of checking the ability of the method to compare two populations and we also present an example of two sample comparison using real data. Chapter 6 concludes the dissertation.

CHAPTER

2

NONPARAMETRIC DENSITY ESTIMATION VIA FOURIER SERIES

In this chapter we first give a brief introduction to Fourier series, and then show how to use it to estimate densities.

Let $L^2([0,1])$ be the linear space of continuous functions $f : [0,1] \to \mathbb{R}$ such that

$$\int_0^1 f(x)\mathrm{d}x \leq \infty$$

The usual inner product is defined by

$$\langle f, g \rangle = \int_0^1 f(x)g(x)\mathrm{d}x$$

This inner product induces the following norm and distance in $L^2([0,1])$:

$$\|f\| = \left(\int_0^1 f^2(x)\mathrm{d}x \right)^{1/2}$$

$$\sqrt{\mathbb{M}(f,g)} = \left(\int_0^1 (f-g)^2\mathrm{d}y \right)^{1/2}$$

where $f, g \in L^2([0,1])$.

The sequence of functions $\{\phi_0, \phi_1, \phi_2, ...\}$ is called orthogonal system when

$$\langle \phi_i, \phi_j \rangle = 0$$

for $i \neq j$ and

$$\|\phi_i\| \neq 0$$

for all i. Furthermore, such a system is called orthogormal basis if for any $f \in L^2([0,1])$ there exists an unique sequence of scalars $\{\alpha_n\}_{n\in\mathbb{N}_+}$ such that

$$\left\| f - \sum_{k=1}^{I} \alpha_k \phi_k \right\| \to 0$$

as $I \to \infty$.

Also as of theorem 3.5.2 from Kreyszig (1989),

$$\alpha_k = \langle f, \phi_i \rangle$$

Thus, f has the following series representation:

$$\sum_{i=0}^{\infty} \langle f, \phi_i \rangle \phi_i$$

In this monograph we shall consider the Fourier basis where $\phi_i : [0,1] \to [-\sqrt{2}, \sqrt{2}]$ and

$$\phi_i(x) = \begin{cases} 1 & \text{if } i = 0 \\ \sqrt{2}\sin(\pi(i+1)x) & \text{if } i \in \{1,3,5,...\} \\ \sqrt{2}\cos(\pi i x) & \text{if } i \in \{2,4,6,...\} \end{cases}$$

Note that there are also many possible smoothness and/or boundary conditions (like twice differentiability) to ensure a somewhat fast convergence rate (see Efromovich (1999) for details). Intuitively, the less smooth the function f is, the larger will be the number of necessary components to get a reasonable approximation. To give the reader some intuition on this, Figure 1 shows the curves for some of the components of Fourier series. It can be seen that higher order components are needed in order to better “explain” less smooth functions.

We now proceed with the exposition of the procedures for statistical inference using frequentist and Bayesian approaches in sections 2.1 and 2.2, respectively.

2.1 Frequentist Inference

Given i.i.d. random variables $Y_1, Y_2, ..., Y_n$ with density function $f : [0,1] \to \mathbb{R} \in L_2[0,1]$, a simple approach to infer f from a frequentist perspective using Fourier series is to use:

$$\hat{f}_I(y) = 1 + \sum_{i=1}^{I} \widehat{\alpha}_i \phi_i(y)$$

where

$$\widehat{\alpha}_i = \frac{1}{n} \sum_{j=1}^{n} \phi_i(Y_j) \approx \int \phi_i(y) f(y) dy = \langle \phi_i, f \rangle$$

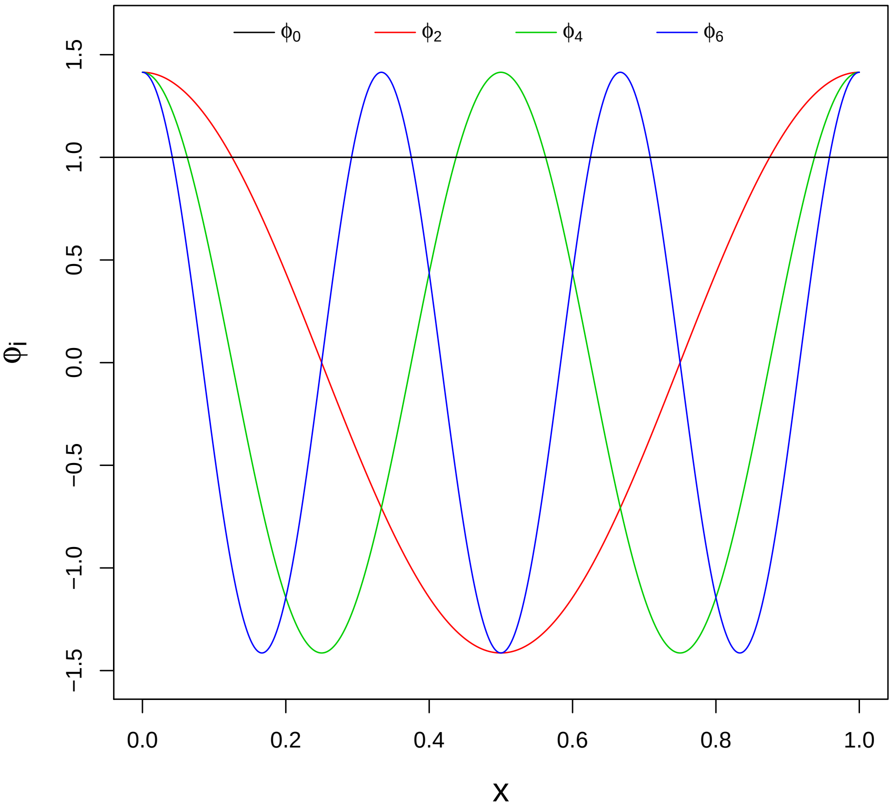

Figure 1 – Plot of some of the components of Fourier series: if a linear combination of such functions is used to approximate a function f, then it can be easily seen that the less smooth f is, the larger will be number of Fourier series components needed in order to better "explain" f.

This estimator is a special case of the modulator estimator which can found in Wasserman (2006), where we can also find the expected value and variance of each $\widehat{\alpha}_i$:

$$E(\widehat{\alpha}_i) = \int_0^1 \phi_i(x) f(x) \mathrm{d}x = \theta_i$$

$$Var(\widehat{\alpha}_i) = \frac{\int_0^1 \phi_i^2(x) f(x) \mathrm{d}x - \theta_i^2}{n}$$

as well as risk of the estimator $\hat{f}_I$:

$$R(\hat{f}_I; f) = E\left[\int_0^1 (\hat{f}_I(x) - f(x))^2 \mathrm{d}x\right] = \sum_{i=1}^{I} Var(\widehat{\alpha}_i) + \sum_{i=I+1}^{\infty} \theta_i^2$$

Therefore the choice of the estimator cutoff parameter I can be seen as bias-variance trade-off problem (in practice, a possible solution is to use cross-validation to choose I).

Finally, we note that the estimate from $\hat{f}_I$ might not respect the constrait $\forall y \in [0,1], f(y) \geq 0$, in which case a "surgery" method is necessary (see Wasserman (2006) and Glad, Hjort and Ushakov (2003)).

2.2 Bayesian Inference

Just as in the previous section, given i.i.d. random variables $Y_1, Y_2, ..., Y_n$ with density function $f : [0,1] \to \mathbb{R} \in L_2[0,1]$, directly proceeding with the Bayesian inference (which is the focus of this work) of f using Fourier series is a somewhat difficult problem since we have to define and work with priors in the constrained space where $f(y) \geq 0$ for all $y \in [0,1]$.

One way to overcome this issue is to use the approach of sieve priors suggested by Scricciolo (2006), which places a prior directly on the coefficient vector β of the Fourier series expansion of $\log(f)$ (instead of f) so that conditionally on the threshold parameter (cutoff parameter) I we have:

$$f(y|I,\beta) = \frac{1}{g(\beta,I)} \exp\left\{\sum_{i=1}^{I} \beta_i \phi_i(y)\right\}$$

where g is a normalizing factor such that

$$g(\beta,I) = \int_0^1 \exp\left\{\sum_{i=1}^{I} \beta_i \phi_i(y)\right\} \mathrm{d}y$$

which is necessary in order to have $\int_0^1 f(y|I,\beta)\mathrm{d}y = 1$. Note that each β_i lives in $\mathbb{R}$, which solves the constrained space problem.

As a drawback, we introduced the difficulty of calculating a normalizing factor (using numerical integration) when evaluating the likelihood function[1].

With such approach (of Scricciolo (2006)) we have

$$\begin{cases} \beta_i \sim \text{Normal}(0, (i+1)^{-2p-1}) & \text{if } i \in \{1,3,5,...\} \\ \beta_i \sim \text{Normal}(0, i^{-2p-1}) & \text{if } i \in \{2,4,6,...\} \end{cases}$$

as prior distributions (with each β_i independent from each other) for the Fourier series coefficients, where p is a strictly positive natural number and the normal distribution is parameterized in terms of mean and variance.

The approach also places a prior distribution on the threshold parameter I, such that for all k strictly positive natural number:

$$P(I=k) = \frac{\exp\{-\gamma k\}}{\sum_{i=1}^{\infty} \exp\{-\gamma i\}}$$

(a geometric distribution) which therefore gives us a Bayesian mixture of Fourier series models. Here γ a is positive hyperparameter.

As we saw in the beginning of this chapter, the less smooth the function is, the greater will number of components needed to get an arbitrary good approximation. Therefore the intuition

[1] This was addressed by adding a numerical integrator to *Stan*, which by time of submission of this work, wasn't yet included in the official distribution of the software.

behind $P(I = k)$ decreasing on k is that we are assuming a prior belief on a somewhat smooth structure for the PDF (see Scricciolo (2006)). This also justifies the use of decreasing variances in priors for β_i (as $i \to \infty$). In principle, one could wonder whether it would be better to not assume this smooth structure and work with more general prior assumptions, but as we saw in section 2.1, we would incur in a problem of lower bias, but greater variance which in general would lead to estimates of f that overfits data.

In this work, primarily because of computational and time restrictions we worked with the series with I up to 10 (that is, $P(I = k) = 0$ if $k > 10$)[2]. We choose to work with somewhat conservative values for those priors with $p = 1$ and $\gamma = 1/10$. As we will see in the next chapters, the results were reasonable for the chosen prior distributions.

[2] There are two technical justifications for this not being problematic: first $P(I = k)$ quickly decreases to zero (a priori) as $k \to \infty$, and also the prior distribution for β_i quickly concentrates on zero (since its variance gets smaller) as as $i \to \infty$.

CHAPTER 3

HMC, NUTS AND STAN

To obtain the posterior samples of the Fourier series models studied in this work, we used Stan (Stan Development Team, 2014), which is a statistical software for obtaining Markov Chain Monte Carlo (MCMC) samples[1] using either the algorithm Halmiltonian Monte Carlo (HMC)[2] or the algorithm No-U-Turn Sampler (NUTS).

The HMC is a MCMC sampling algorithm that was brought from Hamiltonian molecular dynamics to statistics by Duane *et al.* (1987) and Neal (2011). The idea is to solve an Hamiltonian dynamic simulation problem where the variables of original problem (sampling from a target distribution) will be the position variables of Halmitonian system and new artificially introduced variables, typically with Gaussian distribution, will be the "momentum variables" of the Halmitonian system (NEAL, 2011). Among other advantages, the HMC utilizes the gradient of the log-posterior to produce less correlated samples.

The sampler requires, in practice, at least two sampler specific parameters called number of leapfrogs steps and leapfrog stepsize: what happens is that if one could solve the HMC differential equations exactly, it would be possible to sample using only a single sampler specific parameter: the "size", that is, the size of the "jump" we do on the differential equation system for each new sample we want.

However, except for very simple toy models, this is not possible and the differential equation system needs to be approximated using the leapfrog method: approximate the full differential equation jump by many little steps on a difference equations system (similar to Euler method). In any case, HMC is very sensible to these parameters and manually tuning them requires some expertise.

NUTS is a variation of the HMC described at Hoffman and Gelman (2014) that doesn't

[1] Stan can also be used to do optimization and variational inference.

[2] It was originally called Hybrid Monte Carlo by Duane *et al.* (1987), but the term Hamiltonian Monte Carlo used by Neal (2011) seems to be more widely used in literature by now.

require the user to set one the sampler specific parameter: the number of leapfrogs steps. Instead of having an specific number of leapfrog steps, the algorithm will advance in the trajectory up to the point where it starts to retraces its steps (hence the name No-U-Turn Sampler). The authors of the algorithm also propose a method for adapting the other sampler specific parameter left (leapfrog stepsize).

Either way, implementation of any of the algorithms requires time and some expertise and one also needs to code the log-posterior function and its gradient for each model. The process is clearly time-consuming and error-prone, but it's done automatically by `Stan` once the user specify the model in `Stan` language which will handle internally the Hoffman and Gelman (2014) algorithm without requiring any configuration by the user.

3.1 Stan and discrete parameters

`Stan` cannot sample models with discrete parameters directly: this is due to the inability of `HMC/NUTS` to sample discrete variables since they are not differentiable.

However this can be accomplished by marginalizing out the discrete parameters. Suppose, for instance, that $H \in \mathbb{H}$ is a discrete parameter and $\theta \in \Theta$ is the vector of all the other (continuous) parameters. Then we can get the (marginal) likelihood of θ by averaging over all possible values of H,

$$P(Y|\theta) = \sum_{h \in \mathbb{H}} P(Y|\theta, H = h) P(H = h|\theta) \tag{3.1}$$

where Y is the vector containing all known data. After obtaining S simulations $(a_1, a_2, ..., a_S)$ from posterior $P(\theta|Y)$ of the "marginalized model", one can proceed directly to get, for example, the posterior predictive distribution of some $\tilde{Y}$,

$$\begin{aligned} P(\tilde{Y}|Y) &= \int_\Theta \sum_{h \in \mathbb{H}} P(\tilde{Y}, \theta, H = h|Y) \mathrm{d}\theta \\ &= \int_\Theta \sum_{h \in \mathbb{H}} P(\tilde{Y}|\theta, H = h, Y) P(H = h|\theta, Y) P(\theta|Y) \mathrm{d}\theta \\ &\approx \frac{1}{S} \sum_{j=1}^{S} \sum_{h \in \mathbb{H}} P(\tilde{Y}|\theta = a_j, H = h, Y) P(H = h|\theta = a_j, Y) \end{aligned}$$

Note that

$$P(H = h|\theta = a_j, Y) = \frac{P(Y|\theta = a_j, H = h) P(H = h|\theta = a_j)}{\sum_{k \in \mathbb{H}} P(Y|\theta = a_j, H = k) P(H = k|\theta = a_j)} \tag{3.2}$$

which are terms that we already calculated (in equation 3.1) for each posterior simulation.

It is easy to get the posterior for H averaging over the probability in equation 3.2,

$$P(H = h|Y) \approx \frac{1}{S} \sum_{j=1}^{S} P(H = h|\theta = a_j, Y)$$

We can also use it to get the (conditional) posterior for $\theta|H$,

$$\begin{aligned} &P(\theta \in B|H = h,Y) \\ &= \frac{P(H = h|\theta \in B,Y)P(\theta \in B|Y)}{\int_\Theta P(H = h|\theta,Y)P(\theta|Y)\mathrm{d}\theta} \\ &\approx \frac{1}{S}\sum_{j=1}^{S} \frac{\mathbb{I}_B(a_j)P(H = h|\theta = a_j,Y)}{P(H = h|\theta = a_j,Y)} \end{aligned}$$

where $\mathbb{I}$ is the indicator function. That is, for each $h \in \mathbb{H}$ we have S weighted posterior simulations where $\theta = a_j|H = h$ has relative weight $P(H = h|\theta = a_j,Y)$. See Stan Development Team (2014) for more details.

The main weakness of this method of marginalization of discrete parameters is that we have to calculate $P(Y,H = h|\theta)$ for every $h \in \mathbb{H}$ every time we want the likelihood $P(Y|\theta)$, making it computationally unfeasible when $\mathbb{H}$ is huge (unless the model has a known shortcut to calculate the sum in equation 3.1). However, this method has an advantage over methods like Metropolis and Gibbs that explore only a single "marginalized model" (sample θ given H) for each simulation: those methods might explore well only some of the models (those with high posterior probability), and might actually get stuck in a single "marginalized model", not being ergodic.

Imagine for instance that, $H \in \{h_1,h_2\}$, and that the region for continuous parameter θ where $P(\theta|H = h_1) > 0$ is disjoint from the region where $P(\theta|H = h_2) > 0$. Then, if we a start a Gibbs sampler at say $H = h_1$, it will get stuck on it forever since $P(H = H_2|\theta)/P(H = H_1|\theta)$ will always evaluate to zero and if we are using Metropolis and the proposal samples the parameters in block, we might still get stuck if the regions are very close from each other and the proposal has difficulty "connecting" each of them (this is specially troublesome for high dimensional θ). This is a extreme case, but something less extreme may also be very detrimental, the regions might not be disjoint, but the union might be a set with very little probability (again, this gets worse with high dimensionality) to the point where they are "probabilistic" disjoint with our finite computational resources.

On the other hand, with this method, Stan samples the whole "full mixture" model (H marginalized out) for each `MCMC` simulation and as a bonus has the nice properties of the `NUTS` algorithm such that the usage of the gradient information for efficient sampling.

CHAPTER

4

A SIMULATION STUDY TO ASSESS THE PERFORMANCE OF THE METHOD

In this chapter we present the results of a simulation study performed to assess the performance of the Fourier series method compared to the kernel density estimation (KDE). We worked with datasets being generated 100 times from 3 different true models with each dataset having 50 observations. We describe those 3 data generating models in Table 1.

Table 1 – Description of the models used to generate data for the simulation study in this chapter and the next one.

Model 1: data from a random variable $U \in [0,1]$ where:	**Model 2**: data from a random variable $Z \in [0,1]$ where:
$U = \text{logit}^{-1}(V)$	$Z_1 \sim \text{Beta}(1.3, 1.3)$
$V_1 \sim \text{Normal}(-2, 1)$	$Z_2 \sim \text{Beta}(1.1, 3.0)$
$V_2 \sim \text{Normal}(0, 1)$	$Z_3 \sim \text{Beta}(5.0, 1.0)$
$V_3 \sim \text{Normal}(1, 1)$	$Z_4 \sim \text{Beta}(1.5, 4.0)$
$V_4 \sim \text{Normal}(2, 1)$	$P(Z = Z_2) = 0.25$
$P(V = V_1) = P(V = V_4) = 0.2$	$P(Z = Z_3) = 0.35$
$P(V = V_2) = P(V = V_3) = 0.3$	$P(Z = Z_1) = P(Z = Z_4) = 0.2$
Model 3: data from a random variable $W \in [0,1]$ where $W \sim Beta(2,5)$	

Using data from these true models, we fitted Fourier Series models conditional on $I = 1, 2, ..., 10$ and the Bayesian mixture of those 10 models (sieve prior on I, as explained in chapter 2). For comparison purposes, we also fitted the same data using KDE method of Sheather (1991).

In Table 2, we presented the estimates of the errors for each true model from Table 1, using the following procedure for $i \in \{1, 2, 3\}$:

1. Generate a new dataset D from true model i.

2. Fit data using Fourier series (calculate $E[f|D]$) and KDE (calculate $\hat{f}$).
3. Calculate the integrated squared error for Fourier series ($\int_0^1 (f(y) - E[f(y)|D])^2 \mathrm{d}y$) and for KDE ($\int_0^1 (f(y) - \hat{f}(y))^2 \mathrm{d}y$).
4. Calculate the integrated absolute error for Fourier series ($\int_0^1 |f(y) - E[f(y)|D]| \,\mathrm{d}y$) and for KDE ($\int_0^1 \left|f(y) - \hat{f}(y)\right| \mathrm{d}y$).
5. Repeat the procedure 100 times and average over the results.

Hence, the table presents the estimates of expected integrated mean squared error (MSE)

$$\begin{aligned}&\int_{\mathbb{D}} \int_0^1 (f(y) - E[f(y)|D])^2 \,\mathrm{d}y \, P(D|f) \,\mathrm{d}D \\ &= E_D \left[\int_0^1 (f(y) - E[f(y)|D])^2 \,\mathrm{d}y \right]\end{aligned}$$

and the expected integrated mean absolute error (MAE)

$$E_D \left[\int_0^1 |f(y) - E[f(y)|D]| \,\mathrm{d}y \right]$$

for each true model $f(y)$.

Table 2 – Estimated mean squared error and mean absolute error for 3 different data generating models (true models, see Table 1) which were used to fit a Fourier series model with up to 10 component functions and a mixture model of those 10 models (a sieve prior). For comparison purposes, data were also used in a KDE method.

Inferential	Mean Squared Error			Mean Absolute Error		
Model	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
I=1	0.173	0.087	0.082	0.349	0.237	0.236
I=2	0.096	0.055	0.110	0.206	0.180	0.259
I=3	0.098	0.056	0.078	0.211	0.183	0.215
I=4	0.103	0.059	0.084	0.218	0.187	0.224
I=5	0.104	0.059	0.082	0.219	0.186	0.221
I=6	0.103	0.060	0.084	0.218	0.188	0.224
I=7	0.103	0.060	0.084	0.219	0.188	0.224
I=8	0.102	0.060	0.084	0.217	0.188	0.224
I=9	0.102	0.060	0.084	0.217	0.187	0.225
I=10	0.101	0.060	0.084	0.216	0.187	0.225
Mixture	0.102	0.058	0.082	0.217	0.185	0.223
KDE	0.122	0.102	0.077	0.261	0.238	0.187

The Bayesian mixture of Fourier series models with sieve prior (and even most of its component models individually) have outperformed the KDE for true models 1 and 2, but haven't done so for true models 3.

To give the reader a visual intuition of how Fourier series models look like, Figures 2 and 3 plot of the average estimated density using Fourier series for each value I (except for $I = 4$ and

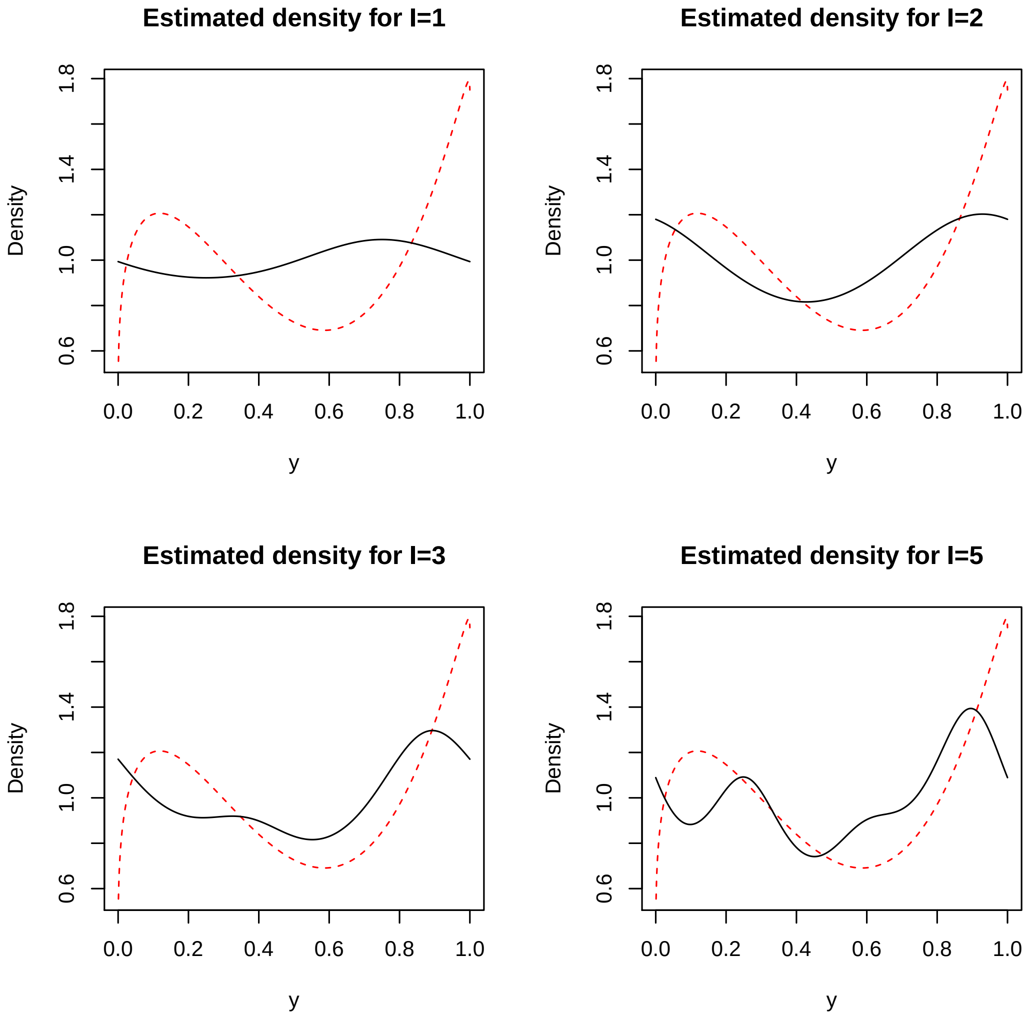

Figure 2 – Plots of the average estimated density using Fourier series for low values of I, data came from (a single dataset of) true model 2. For comparison, the true density is also shown (as a dotted red line).

$I = 7$ for reasons of brevity) using data from a single dataset from true model 2. For comparison, the true density is also shown (as a dotted red line).

Figure 5 shows plots of the curves of the average estimated density from the Bayesian mixture Fourier series model and of the estimate from KDE using the same data (a single dataset from true model 2). Figures 4 and 6 do the same for true models 1 and 3, also using a single dataset from these true models.

These 3 Figures suggest that the Bayesian Fourier series mixture model have performed reasonably well over the whole the domain $[0,1]$ of the density function, but they explored only a single dataset generated from each true model. On the other hand, Table 2 gave us the average errors using 100 different datasets from the true models, but the errors were integrated over $[0,1]$.

Figures 7, 8 and 9 try to combine the advantages of both, that is, to average over 100 different datasets (instead of a single dataset) for the whole domain $[0,1]$ (instead of integrate

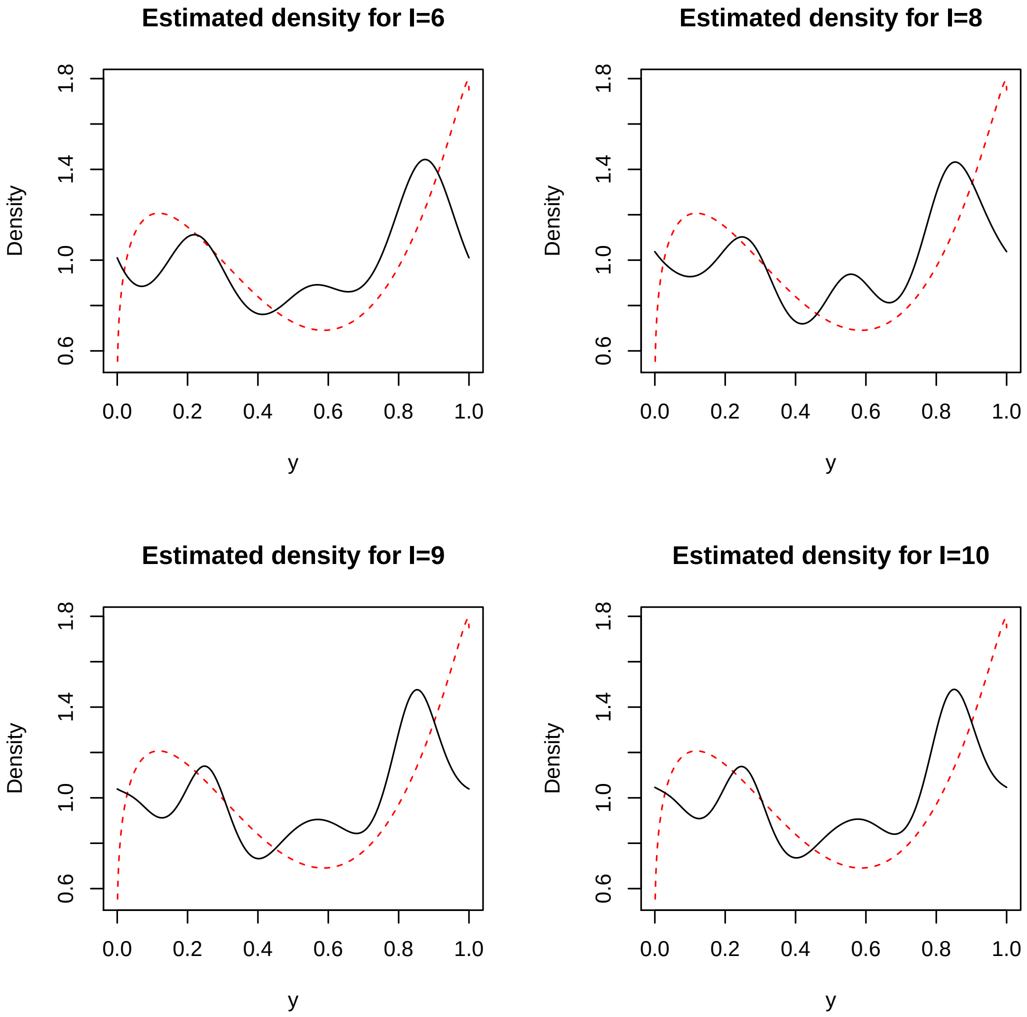

Figure 3 – Plots of the average estimated density using Fourier series for high values of I, data came from (a single dataset of) true model 2. For comparison, the true density is also shown (as a dotted red line).

over $[0,1]$).

Therefore, Figures 7, 8 and 9 show a plot of the estimated MSE of the Bayesian mixture Fourier series model and KDE models for each point in $[0,1]$ for the true models 1, 2 and 3 (respectively) using those 100 different datasets. That is, we estimated and plotted the value of

$$E_D\left[(f(y)-E[f(y)|D])^2\right]$$

for a grid of values $y \in [0,1]$. Therefore, if those plotted curves are integrated against y over $[0,1]$, we'll get the squared errors shown in Table 2.

Figures 7, 8 and 9 indicate that estimating the PDFs near the boundaries with few samples is no easy task and both Fourier series and kernel density methods often had problems (curiously at different places), but as Table 2 has shown, these errors get way less weighty when they are integrated over the domain of the PDF.

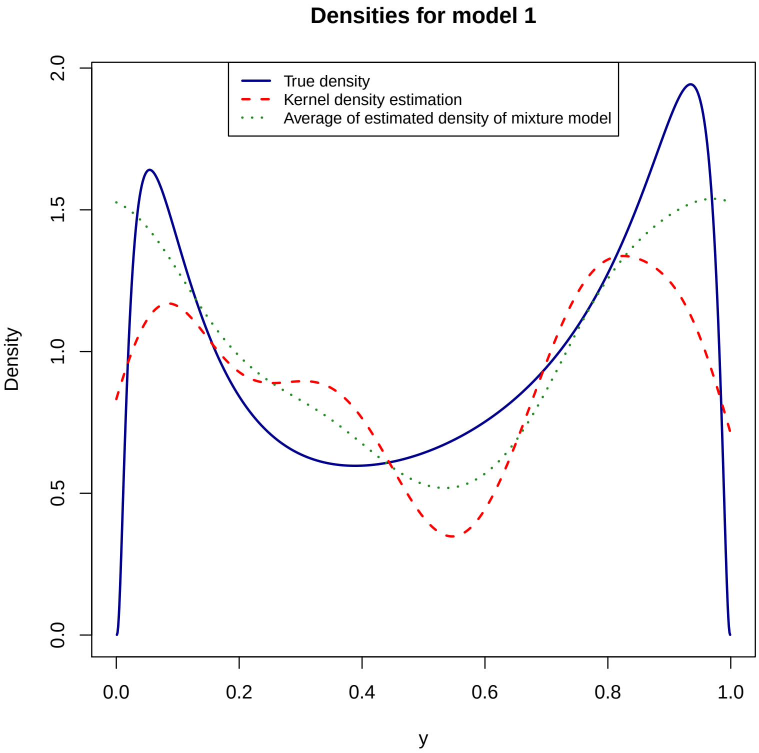

Figure 4 – Average estimated density from the Bayesian mixture of the 10 Fourier series models (sieve prior) against the estimate from KDE, data came from true model 1 (see Table 1) which is also plotted.

We can conclude then that the studied Fourier series estimator is reasonable and therefore we can proceed to the next chapter where it will be used to measure the uncertainty of the distance of the distribution of two population having a sample being observed from each of them.

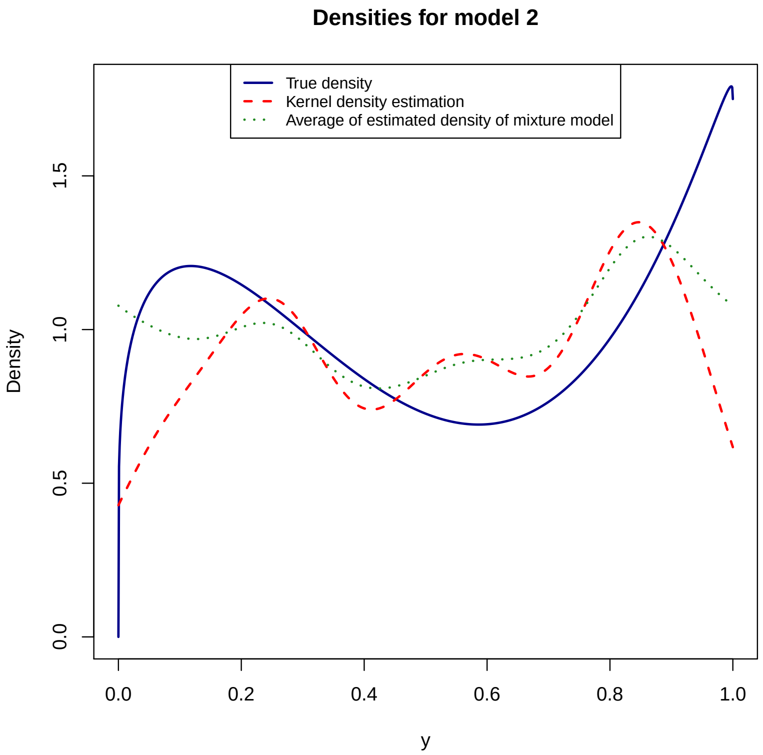

Figure 5 – Average estimated density from the Bayesian mixture of the 10 Fourier series models (sieve prior) against the estimate from `KDE`, data came from true model 2 (see Table 1) which is also plotted.

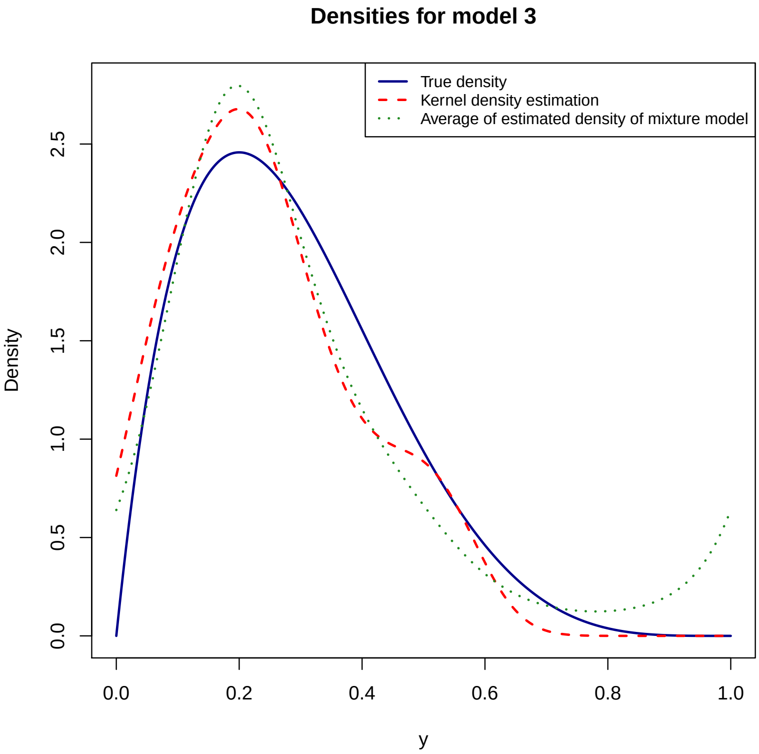

Figure 6 – Average estimated density from the Bayesian mixture of the 10 Fourier series models (sieve prior) against the estimate from KDE, data came from true model 3 (see Table 1) which is also plotted.

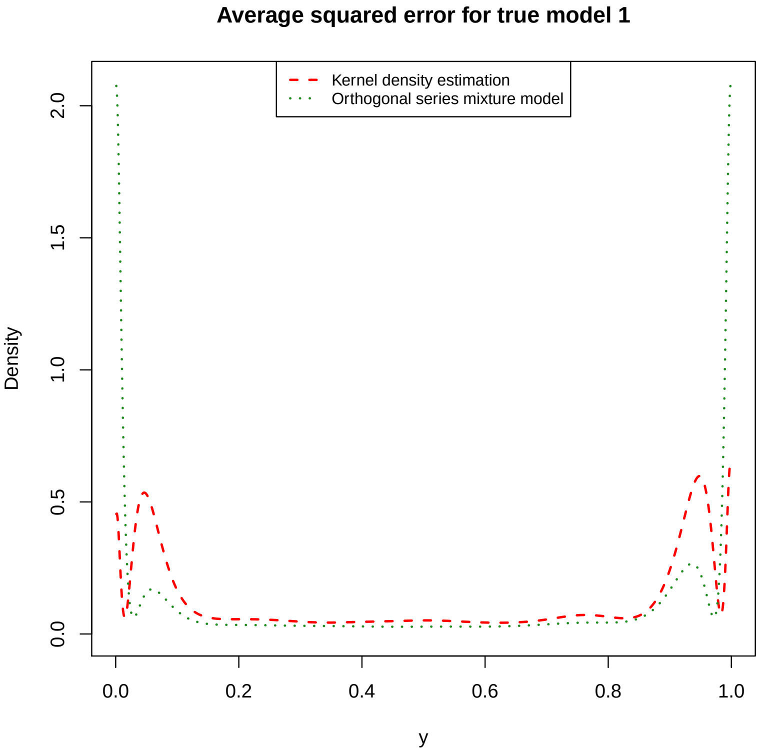

Figure 7 – Estimated mean squared error using data from true model 1 (see Table 1). Data was used to fit a Bayesian mixture of Fourier series model and, for comparison purposes, a KDE model.

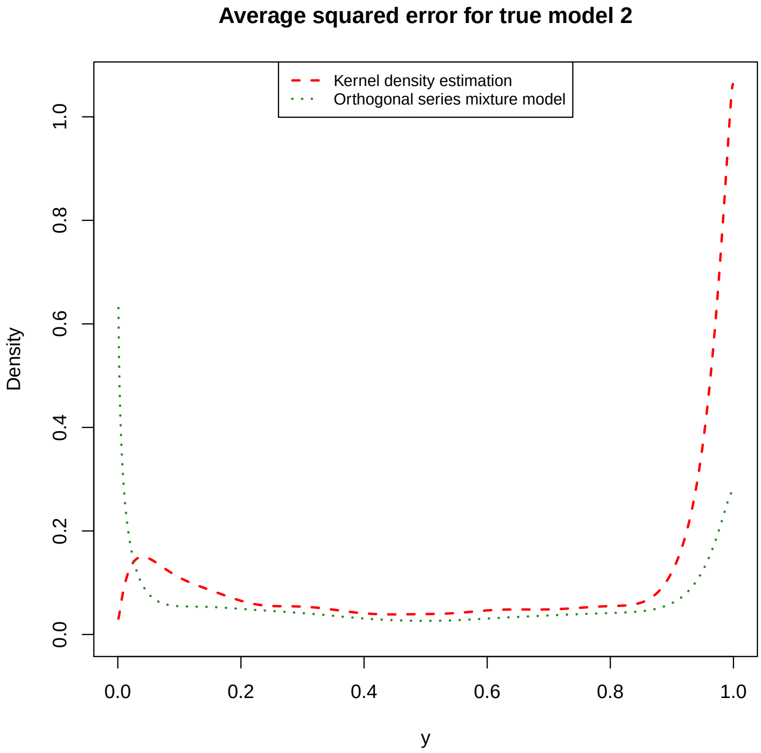

Figure 8 – Estimated mean squared error using data from true model 2 (see Table 1). Data was used to fit a Bayesian mixture of Fourier series model and, for comparison purposes, a `KDE` model.

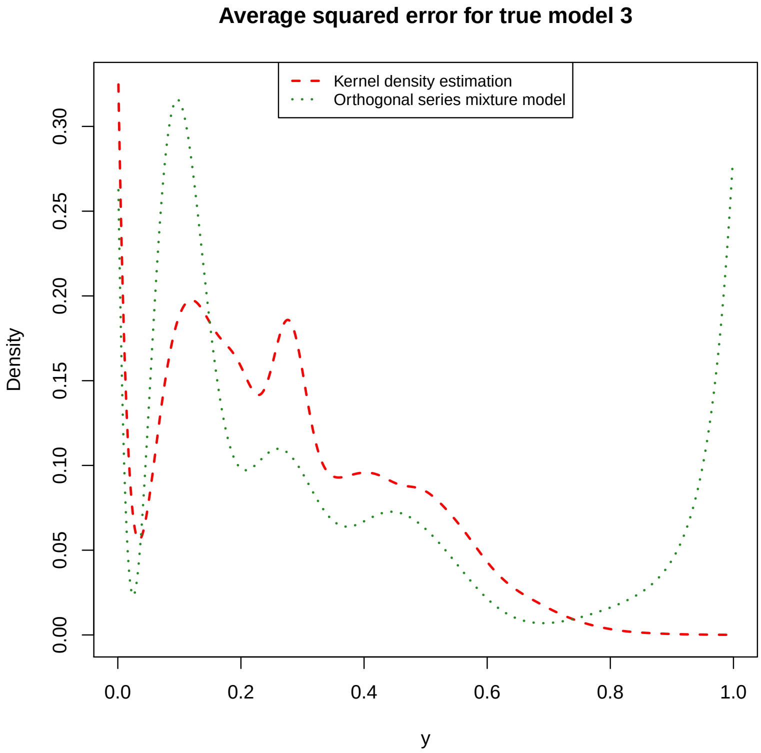

Figure 9 – Estimated mean squared error using data from true model 3 (see Table 1). Data was used to fit a Bayesian mixture of Fourier series model and, for comparison purposes, a `KDE` model.

CHAPTER 5

TWO-SAMPLE SIMILARITY ANALYSIS

In this chapter we deal with the question of how to measure the uncertainty regarding the distance between two populations[1]. That is, given two samples D_1 and D_2 from two (possibly different) populations with PDFs f_1 and f_2, how close (do we expect) the distributions of those two populations are to each other?

To solve this problem we evaluate the posterior probability of some measure of distance (metric), here we work with the integrated squared distance which we denote by $\mathbb{M}$:

$$\mathbb{M}(f_1, f_2) = \int_0^1 (f_1(y) - f_2(y))^2 \mathrm{d}y$$

This distance is itself a parameter, and therefore it has a posterior probability distribution $\mathbb{M}(f_1, f_2)|D_1, D_2$. Having the posterior probability of this distance, we can proceed to calcute $P(\mathbb{M}(f_1, f_2) < \varepsilon|D_1, D_2)$ which gives us a probability that the two populations are close to each other (in the integrated squared distance "sense") up to an arbitrary point ε.

To give the reader some intuition regarding the relationship of $P(\mathbb{M}(f_1, f_2) < \varepsilon|D_1, D_2)$ and ε, in Figure 10 we show a plot of the value of $P(\mathbb{M}(f_1, f_2) < \varepsilon|D_1, D_2)$ against the value of ε with D_1 and D_2 being datasets generated from true model 1 (see Table 1). The Figure indicates that there is a high probability that the distributions are close to each other; compare to Figure 11 where D_1 is generated from true model 1, but D_2 is generated from true model 3.

In this chapter, with the purpose of performing some simulation studies, we work from now on with a fixed value for ε which we'll call ε_0:

$$\varepsilon_0 = \int_{-\infty}^{+\infty} (\varphi_0(y) - \varphi_1(y))^2 \mathrm{d}y \approx 0.125$$

Here φ_0 is the PDF of a standard Gaussian distribution and φ_1 is the PDF of a Gaussian distribution with mean 1 and standard deviation 1.

[1] In this chapter, we use the words population and true model interchangeably

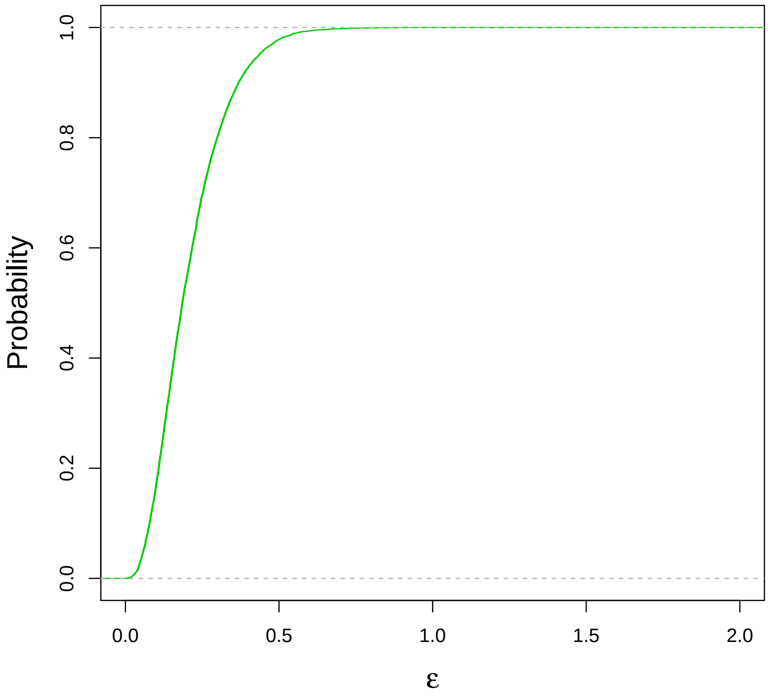

Figure 10 – Plot of $P(\mathbb{M}(f_1, f_2) < \varepsilon | D_1, D_2)$ against ε. Here both D_1 and D_2 are datasets generated from true model 1 (see Table 1).

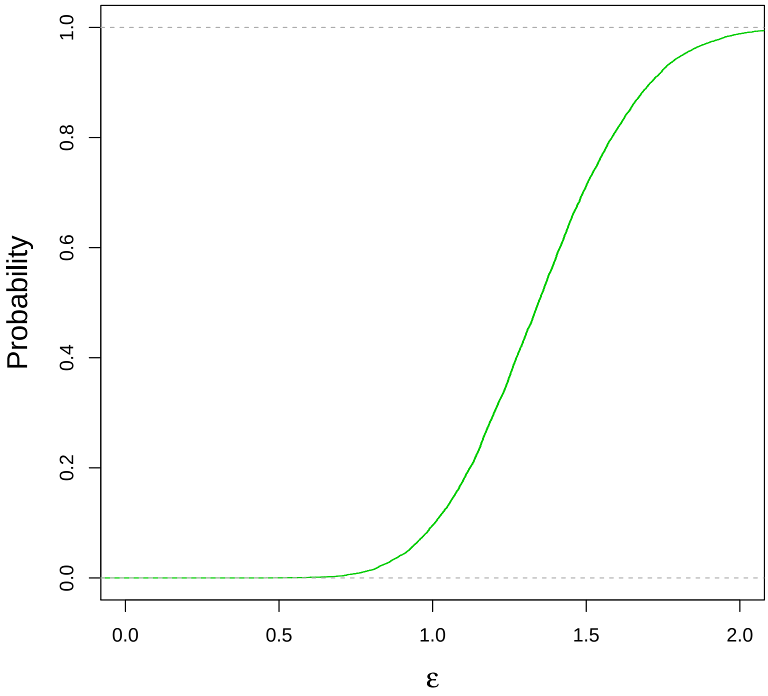

Figure 11 – Plot of $P(\mathbb{M}(f_1, f_2) < \varepsilon | D_1, D_2)$ against ε. Here D_1 is a dataset generated from true model 1 and D_2 is a dataset generated from true model 3 (see Table 1).

After having the theoretic aspects been settled down, we now proceed to some simulation studies with the intention of evaluating the performance of our approach.

5.1 Simulation study for samples from the same population

We start with a simulation study with samples from the same population given the following procedure:

1. Generate 2 samples D_1 and D_2 from true model i.
2. Fit the Fourier series model using each sample separately (get MCMC simulations of $f_1|D_1$ and $f_2|D_2$).
3. Using the simulations from step 2, generate simulations for $\mathbb{M}(f_1, f_2)|D_1, D_2$.
4. Get the proportion of simulations that are below ε_0, therefore obtaining approximately $P(\mathbb{M}(f_1, f_2) < \varepsilon_0|D_1, D_2)$
5. Repeat the procedure 50 times.

The results for such procedure are shown in Figures 12, 13 and 14 for true models 1, 2 and 3 respectively where each of the 50 repetions of the procedure are plotted (points were sorted before being plotted). An horizontal green line in Figures which is the mean of all points and therefore is an approximation to

$$\int_{\mathbb{D}}\int_{\mathbb{D}} P(\mathbb{M}(f_1, f_2) < \varepsilon_0|D_1, D_2)\, P(D_1|f_1)\, \mathrm{d}D_1\, P(D_2|f_2)\, \mathrm{d}D_2$$

Since the “true distance $\mathbb{M}$” from a true model to itself is 0, we should be able to observe a reasonable probability for the event $(\mathbb{M}(f_1, f_2) < \varepsilon_0|D_1, D_2)$, and indeed this is the case for most of the samples as shown in Figures 12, 13 and 14.

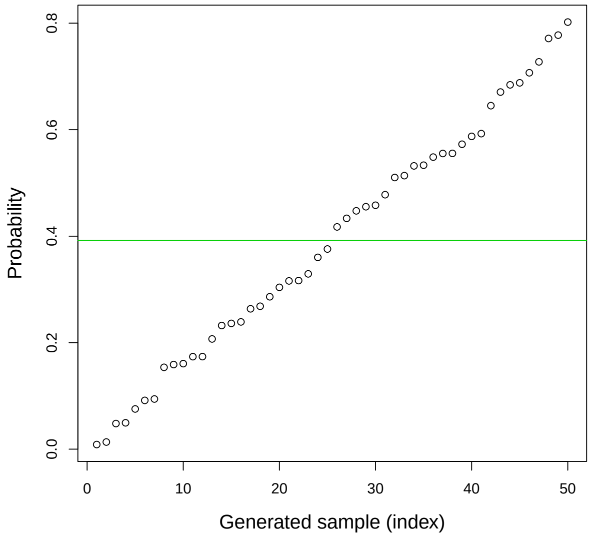

Figure 12 – Each plotted point is the estimated probability that the integrated squared distance between the PDFs of the two samples is less than ε_0 (that is, $P(\int_0^1 (f_1(y) - f_2(y))^2 \mathrm{d}y < \varepsilon_0 | D_1, D_2)$) given samples D_1 and D_2 (for each plotted point, 2 different dataset were generated) **both generated from true model 1** (see Table 1). Points were sorted before plotting. We've choosen to use, as the value of ε_0, the integrated squared distance between the PDF of a standard Gaussian and the PDF of a Gaussian with mean 1 and standard deviation 1 as the value of ε_0. The horizontal green line is the mean of all points.

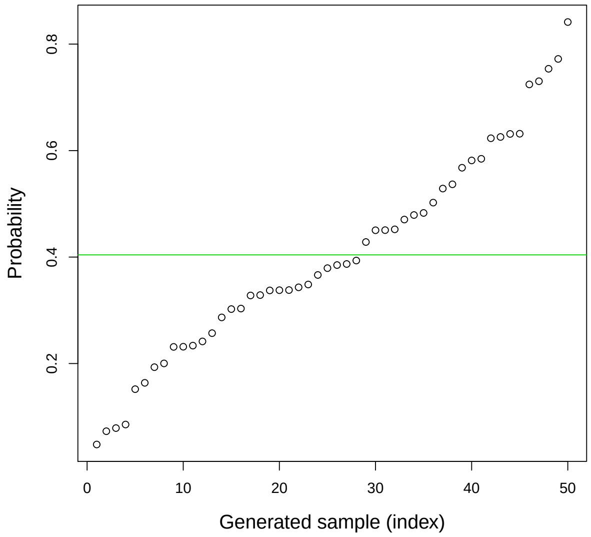

Figure 13 – Each plotted point is the estimated probability that the integrated squared distance between the PDFs of the two samples is less than ε_0 (that is, $P(\int_0^1 (f_1(y) - f_2(y))^2 \mathrm{d}y < \varepsilon_0 | D_1, D_2)$) given samples D_1 and D_2 (for each plotted point, 2 different dataset were generated) **both generated from true model 2** (see Table 1). Points were sorted before plotting. We've choosen to use, as the value of ε_0, the integrated squared distance between the PDF of a standard Gaussian and the PDF of a Gaussian with mean 1 and standard deviation 1 as the value of ε_0. The horizontal green line is the mean of all points.

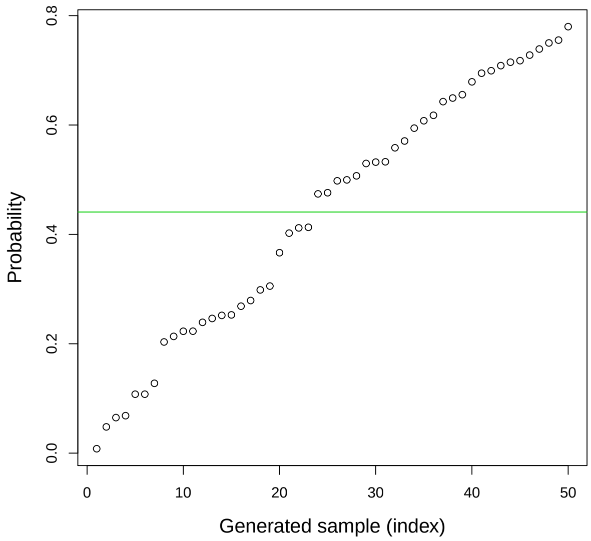

Figure 14 – Each plotted point is the estimated probability that the integrated squared distance between the PDFs of the two samples is less than ε_0 (that is, $P(\int_0^1 (f_1(y) - f_2(y))^2 \mathrm{d}y < \varepsilon_0 | D_1, D_2)$) given samples D_1 and D_2 (for each plotted point, 2 different dataset were generated) **both generated from true model 3** (see Table 1). Points were sorted before plotting. We've choosen to use, as the value of ε_0, the integrated squared distance between the PDF of a standard Gaussian and the PDF of a Gaussian with mean 1 and standard deviation 1 as the value of ε_0. The horizontal green line is the mean of all points.

5.2 Simulation study for samples from the different populations

In this section we have a similar procedure, but with samples D_1 and D_2 coming from different true models:

1. Generate sample D_1 from true model i and sample D_2 from true model j.
2. Fit the Fourier series model using each sample separately (get MCMC simulations of $f_1|D_1$ and $f_2|D_2$).
3. Using the simulations from step 2, generate simulations for $\mathbb{M}(f_1, f_2)|D_1, D_2$.
4. Get the proportion of simulations that are below ε_0, thefore obtaining approximately $P(\mathbb{M}(f_1, f_2) < \varepsilon_0 | D_1, D_2)$
5. Repeat the procedure 100 times.

The results for such procedure are shown in Figure 15 with data from true model 1 against data from true model 2, in Figure 16 with data from true model 1 against data from true model 3 and in Figure 17 with data from true model 2 against data from true model 3. In a similar

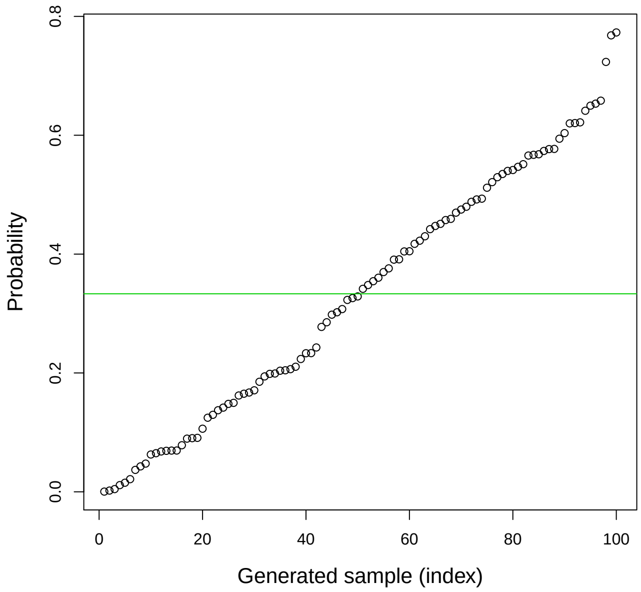

Figure 15 – Each plotted point is the estimated probability that the integrated squared distance between the PDFs of the two samples is less than ε_0 (that is, $P(\int_0^1(f_1(y)-f_2(y))^2\mathrm{d}y < \varepsilon_0|D_1,D_2)$) given samples D_1 and D_2 (for each plotted point, 2 different dataset were generated) where D_1 **was generated from true model 1, and D_2 was generated from true model 2** (see Table 1). Points were sorted before plotting. We've choosen to use, as the value of ε_0, the integrated squared distance between the PDF of a standard Gaussian and the PDF of a Gaussian with mean 1 and standard deviation 1. The horizontal green line is the mean of all points.

manner, each of the 100 repetitions of the procedure are plotted (points were sorted before being plotted). We used the same value for ε_0 and the plots also include an horizontal green line in plot which is the mean of all points.

Note that the "true distance $\mathbb{M}$":

- Between true model 1 and true model 2 is approximately 0.118.
- Between true model 1 and true model 3 is approximately 1.280.
- Between true model 2 and true model 3 is approximately 0.881.

and, as we have already seen, $\varepsilon_0 \approx 0.125$, so it does make sense that, as we can see on Figure 15, samples from true model 1 and 2 have a reasonable probability for the event $(\mathbb{M}(f_1,f_2) < \varepsilon_0|D_1,D_2)$. On the other hand, this probability should be low if samples come from true model 1 and 3, or from true models 2 and 3, and this is what one can clearly see on Figures 16 and 17.

Therefore, we conclude that the proposed method is reasonable to access the uncertainty regarding the similarity of two samples.

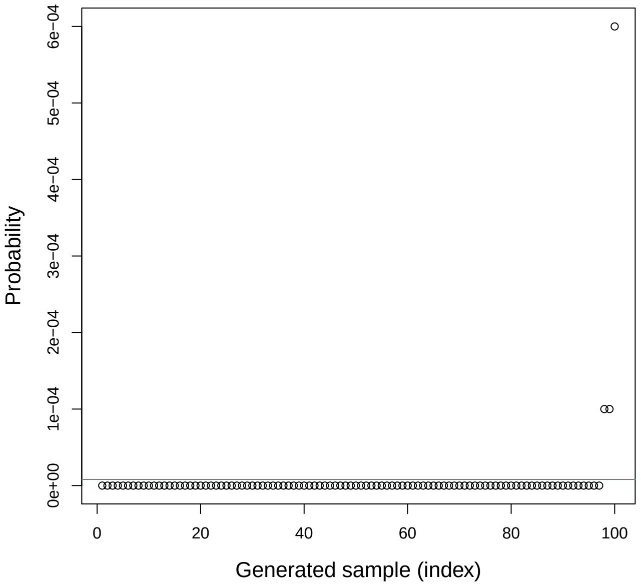

Figure 16 – Each plotted point is the estimated probability that the integrated squared distance between the PDFs of the two samples is less than ε_0 (that is, $P(\int_0^1 (f_1(y) - f_2(y))^2 \mathrm{d}y < \varepsilon_0 | D_1, D_2))$ given samples D_1 and D_2 (for each plotted point, 2 different dataset were generated) where D_1 **was generated from true model 1, and D_2 was generated from true model 3** (see Table 1). Points were sorted before plotting. We've choosen to use, as the value of ε_0, the integrated squared distance between the PDF of a standard Gaussian and the PDF of a Gaussian with mean 1 and standard deviation 1. The horizontal green line is the mean of all points.

5.3 An example of two-sample similarity analysis with real data

We now present an example of two-sample similarity analysis using real data. We used data from the Montreal cognitive assessment used in Cecato *et al.* (2016).

Here we have 3 groups patients: **DA** represents the group of patients diagnosed with Alzheimer, **CCL** represents the group of patients diagnosed with a light cognitive decay and **GC** represents the control group (the response measures the performance of each patient on the Montreal cognitive assessment).

Data was transformed to $[0,1]$ using $S = (R - 50)/(107 - 50)$. Although the maximum values and minimum possible values for this test are 0 and 107, we decided a priori, specially for simplicity, to assume the minimum value to be 50, since it would be very unlikely to a person with such a score to even be placed to take such test.

For simplicity, we have also eliminated two missing data points from datasets. After transformation and removal of missing data, we had 45, 52 and 39 observations for **CCL**, **DA** and **GC** groups, respectively and the descriptive statistics presented in Table 3.

We then estimated the densities for each group independently. Figure 18 shows the

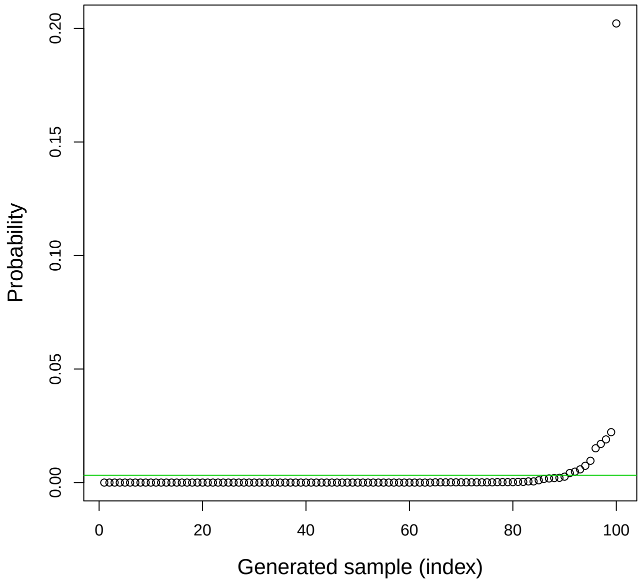

Figure 17 – Each plotted point is the estimated probability that the integrated squared distance between the PDFs of the two samples is less than ε_0 (that is, $P(\int_0^1 (f_1(y) - f_2(y))^2 \mathrm{d}y < \varepsilon_0 | D_1, D_2)$) given samples D_1 and D_2 (for each plotted point, 2 different dataset were generated) where D_1 **was generated from true model 2, and D_2 was generated from true model 3** (see Table 1). Points were sorted before plotting. We've choosen to use, as the value of ε_0, the integrated squared distance between the PDF of a standard Gaussian and the PDF of a Gaussian with mean 1 and standard deviation 1. The horizontal green line is the mean of all points.

Table 3 – Descriptive statistics of data used in this section.

	Obs	Min	1st Qu	Median	Mean	3rd Qu	Max	SD
CCL	45	0.4035	0.6140	0.6842	0.6737	0.7719	0.8772	0.1163
DA	52	0.0351	0.3421	0.5263	0.4716	0.6272	0.8070	0.2042
GC	39	0.5614	0.7632	0.8596	0.8376	0.9123	1.0000	0.0941

average estimated densities of each group using the Fourier series method. Figure 19 shows a plot of $P(\mathbb{M}(f_1, f_2) < \varepsilon | D_1, D_2)$ against ε comparing each of groups against each other (therefore, totaling 3 comparisons).

What's interesting to note is that the curves in Figure 19 allow us to visually compare the groups (by comparing the three curves) without the need to choose a specific value for ε. As it can be seen, the groups diagnosed with Alzheimer and with a light cognitive decay are more alike to each other than to control group.

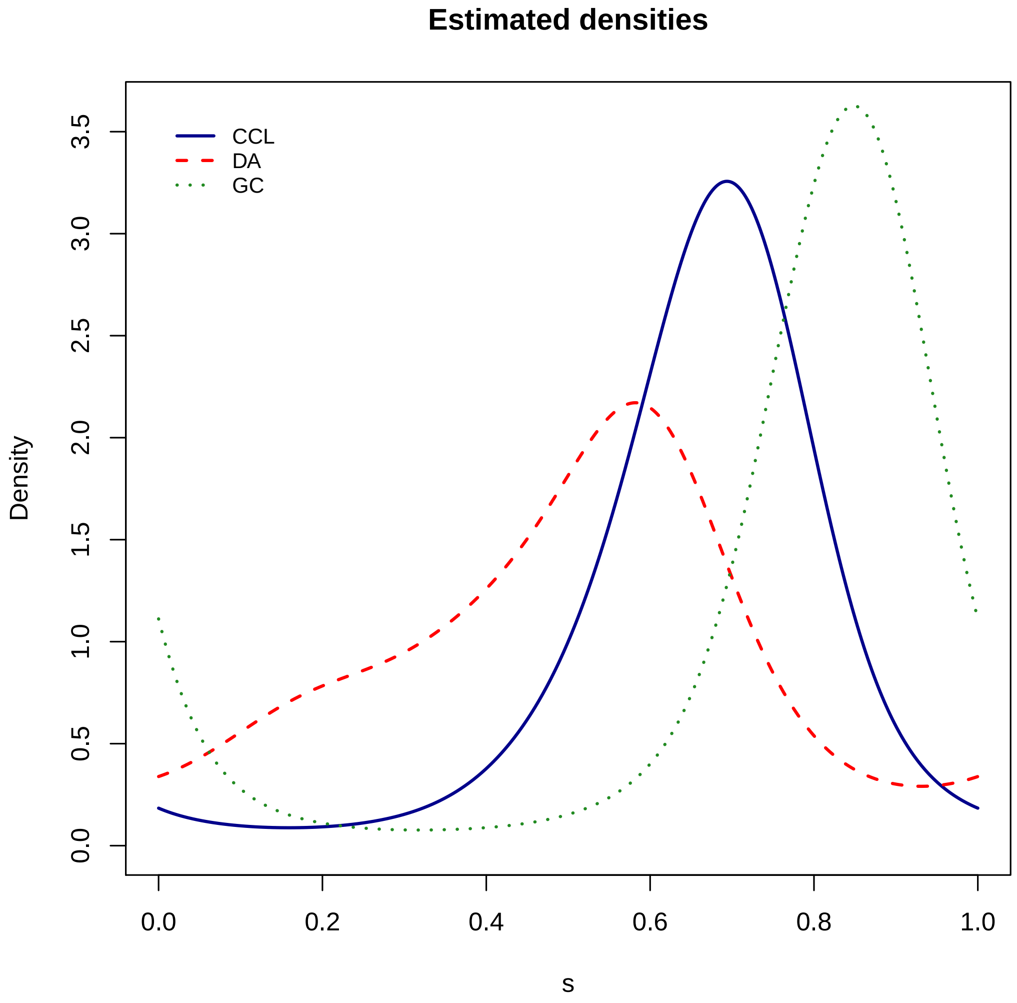

Figure 18 – Plots of the average estimated density using Fourier series (using real data). Here we have 3 groups patients: **DA** represents the group of patients diagnosed with Alzheimer, **CCL** represents the group of patients diagnosed with a light cognitive decay and **GC** represents the control group.

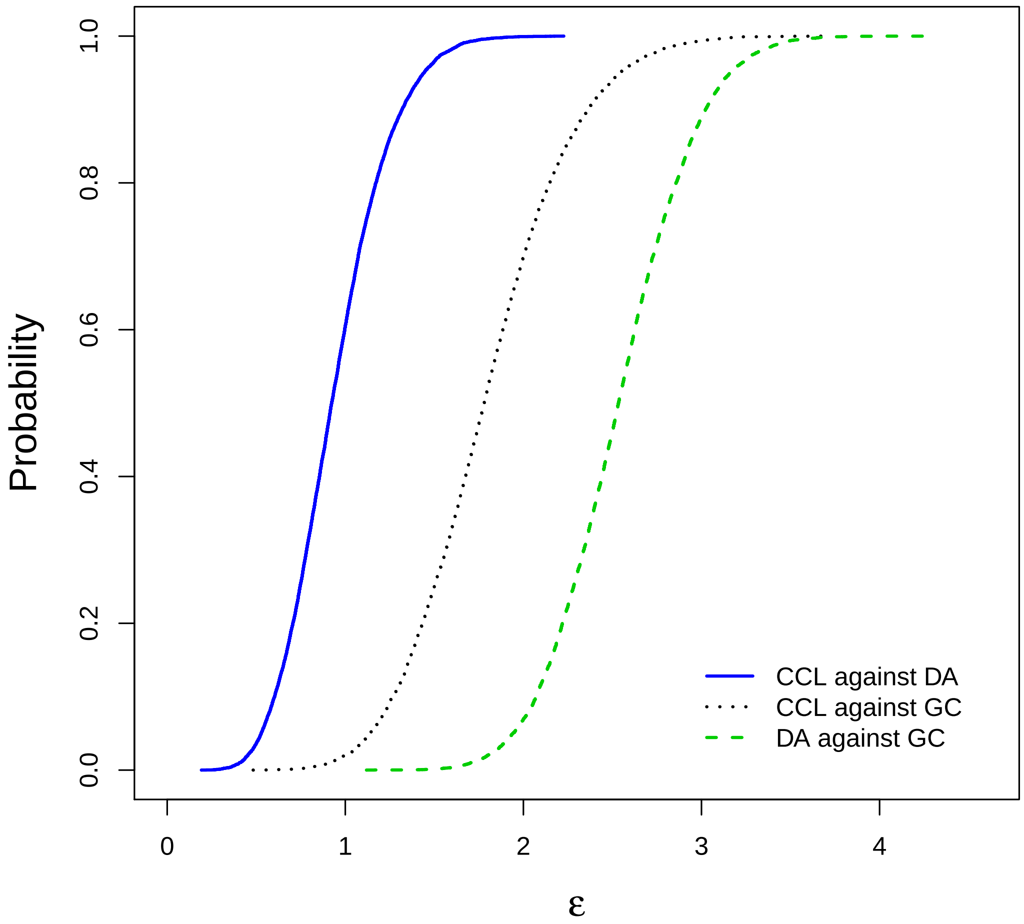

Figure 19 – Plot of $P(\mathbb{M}(f_1, f_2) < \varepsilon | D_1, D_2)$ against ε (using real data). Here both D_1 and D_2 are datasets for the observations groups **CCL**, **DA** and **GC** (being compared against each other). Here we have 3 groups patients: **DA** represents the group of patients diagnosed with Alzheimer, **CCL** represents the group of patients diagnosed with a light cognitive decay and **GC** represents the control group. As it can be seen, the groups diagnosed with Alzheimer and with a light cognitive decay are more alike to each other than to control group.

CHAPTER 6

CONCLUSION

In this work, we proposed a comparison study between Bayesian Fourier series and the kernel density estimation. We concluded that Fourier series method has reasonable goodness of fit compared to the kernel density estimation method and that both often had problems to estimate the PDF near the boundaries for reasonable samples sizes.

We also proposed a method to verify the ability of the Fourier series method to measure the uncertainty regarding the similarity between two populations. We conclude that the proposed method is reasonable for such objective. In particular, it yielded sensible results in a real application. We notice that other Bayesian density estimators could be used; our framework for comparing two populations is very general with that respect.

Possible extensions of this work are:

- A rerun the simulation studies using different true models and/or different samples sizes.
- Usage of other orthogonal series other than Fourier series.
- A comparison of the proposed Bayesian density estimation method with other established methods other than KDE.
- Use different sieve prior parameters, possibly using cross validation to choose them.
- A case study (using real data) of the two sample comparison method.
- Proposal of a method to obtain ε, possibly justifying that using a loss function and Decision Theory techniques.
- Study of the properties of the estimation methods near the boundaries.

BIBLIOGRAPHY

CECATO, J. F.; MARTINELLI, J. E.; IZBICKI, R.; YASSUDA, M. S.; APRAHAMIAN, I. A subtest analysis of the montreal cognitive assessment (moca): which subtests can best discriminate between healthy controls, mild cognitive impairment and alzheimer's disease? **International Psychogeriatrics**, Cambridge University Press, Cambridge, UK, v. 28, n. 5, p. 825–832, 005 2016. Citation on page 49.

DUANE, S.; KENNEDY, A.; PENDLETON, B. J.; ROWETH, D. Hybrid monte carlo. **Physics Letters B**, v. 195, n. 2, p. 216 – 222, 1987. ISSN 0370-2693. Citation on page 31.

EFROMOVICH, S. **Nonparametric curve estimation: methods, theory and applications**. New York: Springer, 1999. ISBN 0-387-98740-1. Citation on page 26.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; DUNSON, D. B.; VEHTARI, A.; RUBIN, D. B. **Bayesian data analysis**. Third. [S.l.]: CRC Press, 2014. ISBN 978-143984095-5. Citation on page 23.

GLAD, I. K.; HJORT, N. L.; USHAKOV, N. G. Correction of density estimators that are not densities. **Scand J Stat**, Wiley-Blackwell, v. 30, n. 2, p. 415–427, jun 2003. Citation on page 27.

HOFFMAN, M. D.; GELMAN, A. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. **Journal of Machine Learning Research**, v. 15, p. 1593–1623, 2014. Citations on pages 31 e 32.

HOLMES, C. C.; CARON, F.; GRIFFIN, J. E.; STEPHENS, D. A. Two-sample Bayesian nonparametric hypothesis testing. **Bayesian Anal.**, International Society for Bayesian Analysis, v. 10, n. 2, p. 297–320, 06 2015. Citation on page 23.

KREYSZIG, E. **Introductory Functional Analysis with Applications**. [S.l.]: Wiley, 1989. ISBN 0471504599. Citation on page 26.

LAVINE, M. Some aspects of polya tree distributions for statistical modelling. **Ann. Statist.**, The Institute of Mathematical Statistics, v. 20, n. 3, p. 1222–1235, 09 1992. Citation on page 23.

MÜLLER, P.; VIDAKOVIC, B. Bayesian inference with wavelets: Density estimation. **Journal of Computational and Graphical Statistics**, Taylor & Francis Group, v. 7, n. 4, p. 456–468, 1998. Citation on page 23.

NEAL, R. M. Mcmc using hamiltonian dynamics. In: BROOKS ANDREW GELMAN, G. L. J. S.; MENG, X.-L. (Ed.). **Handbook of Markov chain Monte Carlo**. Boca Raton, USA: CRC PressTaylor & Francis, 2011. ISBN 1420079417. Citation on page 31.

PETRONE, S. Bayesian density estimation using bernstein polynomials. **Canadian Journal of Statistics**, Wiley Online Library, v. 27, n. 1, p. 105–126, 1999. Citation on page 23.

PETRONE, S.; WASSERMAN, L. Consistency of bernstein polynomial posteriors. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 64, n. 1, p. 79–100, 2002. Citation on page 23.

RIVOIRARD, V.; ROUSSEAU, J. *et al.* Posterior concentration rates for infinite dimensional exponential families. **Bayesian Analysis**, International Society for Bayesian Analysis, v. 7, n. 2, p. 311–334, 2012. Citation on page 23.

SCRICCIOLO, C. Convergence rates for Bayesian density estimation of infinite-dimensional exponential families. **The Annals of Statistics**, v. 34, n. 6, p. 2897–2920, 2006. Citations on pages 11, 13, 23, 24, 28 e 29.

SHEATHER, M. C. J. S. J. A reliable data-based bandwidth selection method for kernel density estimation. **Journal of the Royal Statistical Society. Series B (Methodological)**, [Royal Statistical Society, Wiley], v. 53, n. 3, p. 683–690, 1991. ISSN 00359246. Citation on page 35.

SORIANO, J. **Bayesian Methods for Two-Sample Comparison**. Phd Thesis (PhD Thesis) — Duke University, 2015. Citation on page 23.

Stan Development Team. **Stan Modeling Language Users Guide and Reference Manual, Version 2.8.0**. [S.l.], 2014. Citations on pages 31 e 33.

WASSERMAN, L. **All of nonparametric statistics**. New York London: Springer, 2006. ISBN 0-387-25145-6. Citation on page 27.