A Bayesian nonparametric approach for the two-sample problem

Rafael de Carvalho Ceregatti de Console

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

Rafael de Carvalho Ceregatti de Console

A BAYESIAN NONPARAMETRIC APPROACH FOR THE TWO-SAMPLE PROBLEM

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

Advisor: Prof. Dr. Luis Ernesto Bueno Salasar

São Carlos
June 2019

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

Rafael de Carvalho Ceregatti de Console

UMA ABORDAGEM BAYESIANA NÃO PARAMÉTRICA PARA O PROBLEMA DE DUAS AMOSTRAS

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. Luis Ernesto Bueno Salasar

São Carlos
Junho de 2019

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 Rafael Carvalho Ceregatti de Console, realizada em 19/11/2018:

Prof. Dr. Luis Ernesto Bueno Salasar
UFSCar

Prof. Dr. Danilo Lourenço Lopes
UFSCar

Prof. Dr. Anderson Luiz Ara Souza
UFBA

Certifico que a defesa realizou-se com a participação à distância do(s) membro(s) Anderson Luiz Ara Souza e, depois das arguições e deliberações realizadas, o(s) participante(s) à distância está(ao) de acordo com o conteúdo do parecer da banca examinadora redigido neste relatório de defesa.

Prof. Dr. Luis Ernesto Bueno Salasar

I would like to dedicated this work to my beloved family,
my parents José and Ana and my brothers Felipe and Maria.

ACKNOWLEDGEMENTS

First of all I thank God for all the wonderful things that are happening in my life and for the opportunity to do this work. I want to thanks my dad José, my mom Ana, my brother Felipe and my little sister Maria Eduarda for all their love and support. Without them I would not be able to complete this work.

I also want to thank my advisor Luiz Ernesto Bueno Salasar and my co-advisor Rafael Izbicki. I can not describe how good it is to work with you guys. I want to thank for the many rich discussions, for the advises and specially for friendship.

I would also like to thank my friends from Estatcamp, fon (Afonso), Rafael, Daniel, Hiro (Alexandre), Rubens and in special professor Dorival Leão. Thanks for this opportunity, the feeling of working with you is indescribable.

I could not forget to thanks the friends and residents of the CC, our student resident. It is really nice live with those guys. I also want to thanks my friends Everton (Tom) and Bruno (Kuririn) for the friendship and funny momentos.

I also thank Capes for the financial support and the opportunity to develop this work.

"He who knows others is clever;
He who knows himself has discernment.
He who overcomes others has force;
He who overcomes himself is strong.
He who knows contentment is rich;
He who perseveres is a man of purpose;
He who does not lose his station will endure;
He who lives out his days has had a long life."
(Lao Tzu)

RESUMO

CONSOLE, R. C. C. **Uma abordagem bayesiana não paramétrica para o problema de duas amostras**. 2018. 51 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

Neste trabalho, discutimos o problema conhecido como problema de duas amostras PEARSON; NEYMAN utilizando uma abordagem bayesiana não-paramétrica. Considere $X_1,\ldots,X_n$ and $Y_1,\ldots,Y_m$ duas amostras independentes, geradas por P_1 e P_2, respectivamente, o problema de duas amostras consiste em decidir se P_1 e P_2 são iguais. Assumindo uma priori não-paramétrica, propomos um índice de evidência para a hipótese nula $H_0 : P_1 = P_2$ baseado na distribuição a posteriori da distância $d(P_1,P_2)$ entre P_1 e P_2. O índice de evidência é de fácil implementação, tem uma interpretação intuitiva e também pode ser justificada no contexto da teoria da decisão bayesiana. Além disso, em um estudo de simulação de Monte Carlo, nosso método apresentou bom desempenho quando comparado com o teste de Kolmogorov-Smirnov, com o teste de Wilcoxon e com o método de Holmes. Finalmente, aplicamos nosso método em um conjunto de dados sobre medidas de escala de três grupos diferentes de pacientes submetidos a um questionário para diagnóstico de doença de Alzheimer.

Palavras-chave: Bayesiano Não-paramétrico, Processo de Dirichlet, Teste de Hipótese, Problema de Duas Amostras.

ABSTRACT

CONSOLE, R. C. C. **A Bayesian nonparametric approach for the two-sample problem**. 2018. 51 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2018.

In this work, we discuss the so-called two-sample problem (PEARSON; NEYMAN, 1930) assuming a nonparametric Bayesian approach. Considering $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ two independent i.i.d samples generated from P_1 and P_2, respectively, the two-sample problem consists in deciding if P_1 and P_2 are equal. Assuming a nonparametric prior, we propose an evidence index for the null hypothesis $H_0 : P_1 = P_2$ based on the posterior distribution of the distance $d(P_1, P_2)$ between P_1 and P_2. This evidence index has easy computation, intuitive interpretation and can also be justified in the Bayesian decision-theoretic context. Further, in a Monte Carlo simulation study, our method presented good performance when compared to the well known Kolmogorov-Smirnov test, the Wilcoxon test as well as a recent testing procedure based on Polya tree process proposed by Holmes (HOLMES *et al.*, 2015). Finally, we applied our method to a data set about scale measurements of three different groups of patients submitted to a questionnaire for Alzheimer's disease diagnostic.

Keywords: Bayesian Nonparametrics, Dirichlet Process Prior, Hypothesis Testing, Two-sample problem.

LIST OF FIGURES

Figure 1 – Geometric interpretation of the WIKS index. 28
Figure 2 – Power function comparison for different settings 37
Figure 3 – Power for different settings changing the sample size 39
Figure 4 – Boxplot of CAMCOG scores for the groups MCD, AD and CG. 41

LIST OF ALGORITHMS

Algorithm 1 – WIKS computation 30
Algorithm 2 – Power Function 36
Algorithm 3 – Dirichlet Process Simulation with base probability β and concentration parameter k 47

LIST OF SOURCE CODES

Source code 1 – Functions employed in the simulation 49

LIST OF TABLES

Table 1 – Threshold c values (0.95 sample quantiles of the index) for different support distributions and λ values considering $n = m = 50$. 33
Table 2 – Quantiles 0.05, 0.25, 0.5, 0.75, 0.95 and mean of WIKS index simulated from different populations P (under H_0) with $n = 10$ and $m = 20$ 33
Table 3 – Quantiles 0.05, 0.25, 0.5, 0.75, 0.95 and mean of WIKS index simulated from different populations P (under H_0) with $n = 20$ and $m = 30$ 34
Table 4 – Quantiles 0.05, 0.25, 0.5, 0.75, 0.95 and mean of WIKS index simulated from different populations P (under H_0) with $n = 30$ and $m = 10$ 34

CONTENTS

1 INTRODUCTION 25

2 BAYESIAN NONPARAMETRIC INDEX 27
2.1 The nonparametric Bayesian WIKS index 27
2.1.1 Index Definition *27*
2.1.2 Decision-theoretic formulation *31*
2.2 Prior Specification and Decision Procedure 31
2.3 Invariance of WIKS sampling distribution under H_0 33
2.4 Simulation Study 34
2.4.1 Holmes's method *34*
2.4.2 Power Simulation *35*
2.5 Application 38

3 FINAL REMARKS AND FUTURE WORKS 43

BIBLIOGRAPHY 45

APPENDIX A STICK-BREAKING REPRESENTATION 47

ANNEX A R SCRIPTS 49

CHAPTER

1

INTRODUCTION

One basic interest in statistics is to test the difference between groups, e. g., testing the difference between a group of patients that received a drug and a group that received placebo. Besides the pharmaceutical examples, according to Borgwardt and Ghahramani (2009), the problem of comparing two groups is surprisingly common in practice and appears in several research fields. In the literature this problem is known as the "two-sample problem" Pearson and Neyman (1930) and consists in deciding whether two samples are drawn from the same population.

The first researchers on this subject were Behrens (1929) and Fisher (1935). The method developed by Behrens and Fisher consists in comparing the means of the characteristic of interest that is measured on two samples from different populations assumed normally distributed. Nowadays, the approach based on the normality assumption and considering equal variance to the populations are the well known t-test. Welch (1947) considered how to extend such test for the setting where the variances of the two populations are different, and Cressie and Whitford (1986) discussed the problems of the t-test for two samples when the assumptions of independence, homogeneity and normality do not hold. A detailed study on the t-test was presented by Neyman and Pearson (1967), showing its main particularities and properties.

Kolmogorov (1933) and Smirnov (1939) presented a nonparametric test for deciding whether two samples are generated from the same distribution. Their approach consists in measuring the distance between the two empirical distributions. Wilcoxon (1945), Mann and Whitney (1947) presented an alternative approach to the t-test without the need of such restrictive assumptions such as assuming normal distribution. The Wilcoxon-Mann-Whitney test is a nonparametric test that uses the ranks of the samples to conclude on the equality of the medians. Halperin (1960) generalized the Wilcoxon-Mann-Whitney test for the setting where the samples are censored at a fixed point.

The classic Bayesian parametric formulation to this problem is in terms of the Bayes

factor, introduced by Kass and Raftery (1995). Baldi and Long (2001) proposed a Bayesian two-sample test tailored to microarray data analysis and, to general cases, Gönen *et al.* (2005) presented the Bayesian two-sample t-test. In their paper, they present a prior that provides a closed form for the Bayes factor, which can then be written in terms of the distribution of the t statistics of the two samples on the null and alternative hypothesis, respectively. For a review of several Bayesian approaches to the problem see Bernardo and Smith (2001).

Unfortunately, there are very few attempts of attacking this problem from a Bayesian nonparametric perspective. The only exceptions that we are aware of are Basu and Chib (2003), which use Bayes factors for Dirichlet process-based models; Borgwardt and Ghahramani (2009), which discuss two-sample tests based on Dirichlet process mixture models and derived a formula to compute the Bayes factor in this case; Ma and Wong (2011), which propose to allow the two random distributions under the alternative to randomly couple on different parts of the sample space, thereby achieving transference of information; Labadi, Masuadi and Zarepour (2014), which propose a Bayesian method for comparing two-samples based on the Kolmogorov distance; and Holmes *et al.* (2015), who developed a Bayesian nonparametric procedure for the problem considering a Polya tree process prior.

In this dissertation, we propose a test for the equality of the two populations by means of a nonparametric Bayesian evidence index, which is given by a posterior weighted mean of the distance between P_1 and P_2, the probability distributions associated with each population. We call this evidence index WIKS (weight integrated Kolmogorov-Smirnov). The WIKS, is easy to compute, has intuitive interpretation and can be justified in the Bayesian decision-theoretic framework. We show that WIKS achieves greater power when compared with the Kolmogorov-Smirnov and Wilcoxon in a wide variety of simulation scenarios. When compared with Holmes' method, the WIKS present better performance in some settings. Finally, we apply our method to a data set about scale measurements of three different groups of patients submitted to a questionnaire used to diagnostic Alzheimer's disease.

In Chapter 2 we motivate the WIKS and present a simulation study designed to compare our proposal with the Kolmogorov-Smirnov test, the Wilcoxon-Mann-Whitney test and with the Holmes test and in chapter 3, we present the final considerations.

CHAPTER

2

BAYESIAN NONPARAMETRIC INDEX

This chapter is organized as follows. In Section 2.1 we present our evidence index and a decision theoretical justification for it. In section 2.2 we specify the necessary procedures to the computational study of the index. We investigate the WIKS invariance in section 2.3 and in section 2.4 we present a simulation study designed to compare our proposal with the Kolmogorov-Smirnov test, the Wilcoxon-Mann-Whitney test and the Holmes test. In Section 2.5 we apply our method to a data-set of scale measurements for Alzheimer disease.

2.1 The nonparametric Bayesian WIKS index

Assume that two independent samples $X_1,\cdots,X_n$ and $Y_1,\cdots,Y_m$ are drawn from P_1 and P_2, respectively. Our aim is to test the null hypothesis $H_0 : P_1 = P_2$ against the alternative $H_1 : P_1 \neq P_2$. Assuming a suitable metric d between probability measures[1], we can express the magnitude of the difference between the two populations P_1 and P_2 by $d(P_1,P_2)$. Using this metric, our problem can be reformulated as testing $H_0 : d(P_1,P_2) = 0$ against $H_1 : d(P_1,P_2) > 0$. In the following, we shall assume that the metric d is bounded above.

2.1.1 Index Definition

Considering a given nonparametric prior for (P_1,P_2), let us assume that $\mathbb{P}^{x,y}$ is the posterior distribution for (P_1,P_2) given the observed samples $x = (x_1,\ldots,x_n)$ and $y = (y_1,\ldots,y_m)$. The WIKS index is defined as follows.

Definition 1. The WIKS index against hypothesis H_0 is defined by

$$\text{WIKS}(\mathscr{D}_{n,m}) = \int_0^M w(\varepsilon)\ \mathbb{P}^{x,y}\big(d(P_1,P_2) > \varepsilon\big)d\varepsilon, \tag{2.1}$$

[1] Common choices for this metric are the Kolmogorov-Smirnov metric, the L2 metric and Lévy metric. For a survey of metrics between probability measures see Rachev *et al.* (2013).

where $\mathscr{D}_{n,m} = \{x, y\}$ denotes the two observed samples of sizes n and m, $w : [0, M) \longrightarrow (0, \infty]$ (the weight function) is a probability density function over $[0, M)$ and $M = \sup_{P_1, P_2} d(P_1, P_2)$ is the maximum value (possibly being $+\infty$) of the distance d.

The idea behind this index is to express a discrepancy between the posterior distribution of $d(P_1, P_2)$ and 0 and to facilitate the index understanding, a geometric interpretation of WIKS is displayed in Figure 1.

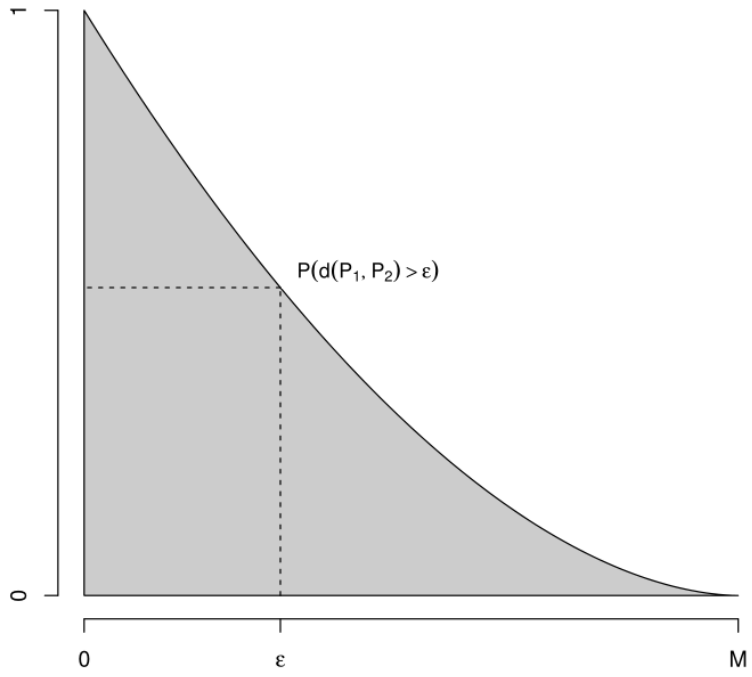

Figure 1 – Geometric interpretation of the WIKS index.

In Figure 1, we disregard the weight function and we look at only to the posterior survival distribution of the $d(P_1, P_2)$. The area under the curve correspond to the expression:

$$\int_0^M \mathbb{P}^{x,y}\big(d(P_1, P_2) > \varepsilon\big) d\varepsilon. \tag{2.2}$$

For example, assume that $d(P_1, P_2) = M$ almost sure. So the expression 2.2 is equal to M, the highest possible value. Alternatively, if $d(P_1, P_2) = 0$ almost sure (the two populations are equal almost sure), then the expression 2.2 is equal to 0. Then, the bigger the area, greater is the evidence against the null hypothesis. The only difference between the expression 2.2 and the expression 2.1 is the weight function w, thus the same reasoning apply. Therefore, the bigger the value of the WIKS, greater is the evidence against the null hypothesis.

Furthermore, WIKS can be thought of as a compromise between different evidence indexes against the null H_0. More specifically, a naive evidence index against the null is $P^{x,y}(d(P_1,P_2) > \varepsilon)$ for a fixed $\varepsilon > 0$, where larger values indicate greater evidence against the null. Thus, one can decide to reject the null whenever that probability exceeds a given threshold δ (e.g., 0.5)[2]. However, choosing an appropriate ε value is typically not easy, especially in a nonparametric framework. Moreover, it can also lead to inconsistent decisions: for instance, suppose that the actual distance between P_1 and P_2 is ε' in $(0,\varepsilon)$, then $P^{x,y}(d(P_1,P_2) > \varepsilon)$ converges to 0 as the sample sizes increase (since the posterior of $d(P_1,P_2)$ converges to ε') leading one to wrongly accept the null. Instead of fixing an ε value, WIKS combines all the evidences $P^{x,y}(d(P_1,P_2) > \varepsilon)$ for different ε using the weighted average given in (2.1). Notice that, by choosing a constant weight function w, WIKS index (2.1) is proportional to the area below the survival curve of $d(P_1,P_2)$, which is the posterior expected value of $d(P_1,P_2)$. Different choices of the weight function can be considered depending on the specifics of the problem at hand. At last, the naive approach is a particular case when we consider the weight function as a degenerate weight function.

Next we investigate some properties of the index.

Theorem 1. Let $\mathbb{E}^{x,y}$ denote the expectation with respect to $\mathbb{P}^{x,y}$. Then,

$$\text{WIKS}(\mathscr{D}_{n,m}) = \mathbb{E}^{x,y}\big[W(d(P_1,P_2)\big], \tag{2.3}$$

where W is the cumulative distribution function corresponding to the probability density function w.

Proof of Theorem. Let $\mathbb{P}_d$ be the probability distribution of $d(P_1,P_2)$ assuming that (P_1,P_2) is distributed according to $\mathbb{P}^{x,y}$. Thus,

$$WIKS(\mathscr{D}_{n,m}) = \int_0^M w(\varepsilon)\mathbb{P}_d\big((\varepsilon,M]\big) = \int_0^M \int_0^M w(\varepsilon) I_{(\varepsilon,M]}(z) d\mathbb{P}_d(z) d\varepsilon,$$

which implies by the Fubini theorem, that

$$\begin{aligned} WIKS(\mathscr{D}_{n,m}) &= \int_0^M \int_0^M w(\varepsilon) I_{(\varepsilon,M]}(z) d\varepsilon d\mathbb{P}_d(z), \\ &= \int_0^M \int_0^z w(\varepsilon) d\varepsilon d\mathbb{P}_d(z) \\ &= \int_0^M W(z) d\mathbb{P}_d(z) \\ &= E^{x,y}[W(d(P_1,P_2))], \end{aligned}$$

where $I_A(z)$ denotes the indicator function assuming 1 if $z \in A$ and 0 otherwise. □

[2] This approach was suggested by e.g. Swartz (1999) in a Bayesian nonparametric goodness-of-fit context.

Theorem 1 shows that WIKS can be expressed as the expected value of $W(d(P_1,P_2))$ with respect to the posterior distribution. This implies that a Monte Carlo approximation for WIKS is readily available from posterior simulations of (P_1,P_2), what is very useful in practice. A description of such procedure is given in Algorithm 1.

Algorithm 1 – WIKS computation

Require: samples x and y of sizes n and m; posterior distribution $\mathbb{P}^{x,y}(P_1,P_2)$; cumulative weight function W; number of Monte Carlo simulations S
Ensure: WIKS$(\mathscr{D}_{n,m})$

1: Sample $(P_{1,1},P_{2,1}),\ldots,(P_{1,S},P_{2,S})$ independently from the posterior distribution $\mathbb{P}^{x,y}$;
2: Approximate WIKS index by

$$\text{WIKS}(\mathscr{D}_{n,m}) \approx \frac{1}{S}\sum_{s=1}^{S} W(d(P_{1,s},P_{2,s}))$$

In Algorithm 1, we show one way to estimate the WIKS by Monte Carlo.

Straight from the definition, the index has the following properties.

Theorem 2. 1. $0 \leq \text{WIKS}(\mathscr{D}_{n,m}) \leq 1$ for any observed sample $\mathscr{D}_{n,m}$;

2. $\text{WIKS}(\mathscr{D}_{n,m}) = 0$ if, and only if, $d(P_1,P_2) = 0$ almost surely;

3. $\text{WIKS}(\mathscr{D}_{n,m}) = 1$ if, and only if, $d(P_1,P_2) = M$ almost surely;

4. $\text{WIKS}(\mathscr{D}_{n,m})$ is increasing[3] with respect to $d(P_1,P_2)$.

Proof of Theorem. 1. It follows directly from Theorem 1 and the fact that W assumes values in $[0,1]$;

2. Since the random variable $W(d(P_1,P_2))$ is non negative, its expected value is zero if and only if it assumes zero almost surely;

3. The same argument of (2) applied to the non negative random variable $1 - W(d(P_1,P_2))$;

4. Consider D_1 and D_2 two random variables representing two posterior distributions for $d(P_1,P_2)$ such that D_2 is stochastically greater than D_1, i.e., $\mathbb{P}(D_1 \geq x) \leq \mathbb{P}(D_2 \geq x)$ for all $x > 0$. Since $\mathbb{E}[D_i] = \int_0^\infty \mathbb{P}(D_i \geq x)dx$, we have that $\mathbb{E}[D_1] \leq \mathbb{E}[D_2]$.

□

[3] Increasing in the sense of stochasticaly order, for more details see chapter one of Shaked and Shanthikumar (2007)

2.1.2 Decision-theoretic formulation

The WIKS index can also be motivated in the Bayesian decision framework (DeGroot (2005)). Let $\mathbb{D} = \{a, a^c\}$ be the decision space, where a stands for accepting H_0 and a^c for rejecting H_0. Let us consider the following loss function for our decision problem:

$$L\big((P_1,P_2),d\big) = \begin{cases} c_0 W(d(P_1,P_2)), & \text{if } d = a, \\ c_1[1 - W(d(P_1,P_2))], & \text{if } d = a^c, \end{cases} \tag{2.4}$$

where c_0 and c_1 are positive real numbers representing the maximum loss when accepting and rejecting H_0, respectively. Observe that, if we decide to accept H_0, the loss function is zero if $d(P_1,P_2) = 0$ and increases with the value of $d(P_1,P_2)$. On the other hand, if we decide to reject H_0, then the function decreases with the value of the distance $d(P_1,P_2)$ and vanishes if $d(P_1,P_2) = M$.

For a decision $\delta(x,y) \in \mathbb{D}$, the posterior expected loss is given by

$$\mathbb{E}^{x,y}\big[L\big((P_1,P_2),\delta(x,y)\big)\big] = \begin{cases} c_0 \mathbb{E}^{x,y}[W(d(P_1,P_2))], & \text{if } \delta(x,y) = a, \\ c_1\big[1 - \mathbb{E}^{x,y}[W(d(P_1,P_2))]\big], & \text{if } \delta(x,y) = a^c. \end{cases}$$

Theorem 3. The Bayes rule for the loss function (2.4) is given by rejecting H_0 if

$$\text{WIKS}(\mathscr{D}_{n,m}) > c, \tag{2.5}$$

where $c = c_1/(c_1 + c_0)$.

Proof of Theorem. Our decision is a^c, in others words reject H_0, if

$$\begin{aligned} c_0 \mathbb{E}^{x,y}[W(d(P_1,P_2))] &> c_1\big[1 - \mathbb{E}^{x,y}[W(d(P_1,P_2))]\big] \\ &= c_1 - c_1 \mathbb{E}^{x,y}[W(d(P_1,P_2))] \end{aligned}$$

simplifying

$$c_0 \mathbb{E}^{x,y}[W(d(P_1,P_2))] + c_1[\mathbb{E}^{x,y}[W(d(P_1,P_2))] > c_1$$

that is equivalent to

$$\text{WIKS}(\mathscr{D}_{n,m}) > \frac{c_1}{c_1 + c_0} \tag{2.6}$$

□

2.2 Prior Specification and Decision Procedure

Our approach to solve the two-sample problem is fairly general. Basically, if we can draw samples from the posterior, we can compute the index. Thus, the index can be applied to any prior distribution, such as Pólya trees and the Beta processes. In this thesis, however, we focus on one of the most used methods to perform Bayesian nonparametric inference, which is the Dirichlet process prior (Ferguson (1973)).

In order to proceed to test the hypothesis $P_1 = P_2$ using our index, we need to specify the prior distribution for P_1 and P_2, choose a metric d and a weight function w. The prior for P_1 and P_2 is specified as two independent Dirichlet process with the same base probability G and concentration parameter K. The concentration parameter K is set to 1 because the initial idea is to verify the behavior of the index without "changing" the variation of the Dirichlet process, that is, increase variation with $K < 1$ or decrease variation with $K > 1$. The base probability G is chosen accordingly to the known support of the data: for observations taking values on the real line, we choose the standard Gaussian distribution $N(0,1)$; for observations taking values in the nonnegative real line, we choose the standard lognormal distribution $LN(0,1)$ and for observations taking values in the $[0,1]$ interval, we choose the uniform distribution $U(0,1)$. In the Appendix A, we discuss the Sethuraman's approach to sample from the Dirichlet process. The metric d considered is the Kolmogorov metric defined by $d(P_1,P_2) = \sup_x |P_1((-\infty,x]) - P_2((-\infty,x])|$ and, since the maximum of the Kolmogorov distance is 1, the weight function w is taken to be a Beta$(1,\lambda)$ density ($\lambda \geq 1$), which has cumulative weight function $W_\lambda(t) = 1-(1-t)^\lambda$, $t \in [0,1]$. We choose this probability density function to the weight function because they are asymmetric to the left and as we have in mind especially medical and pharmaceutical applications, we want to prioritize small distances due to the critically involved in these fields. For example, for $W_4(t) = 1-(1-t)^4$ approximately 60% of the total weight is attributed up to 0.2 of the distance.

Now, it only remains to decide how to choose the threshold value c for the decision criterion in (2.5). At this point, we follow the philosophical approach suggested in Good (1992) and adopt a Bayesian / non-Bayesian compromise to select the threshold. The idea is to select the value c that controls the type I error, that is, given that the hypothesis H_0 is true, we declare it false with probability less than α, e.g., $\alpha = 0.05$.

We compute the quantiles in the following way. The first step consists in simulate two samples with sizes n and m from the null distribution [4]. The second steps consists in applying the Algorithm 1 with all parameters laid down above to simulate one value of the WIKS. We repeat the latter procedure a thousand times and take c as the 0.95 sample quantile of the index values. Table 1 presents the obtained threshold c considering the three different settings for the population support and $\lambda = 1,2,3,4$. Thus, the H_0 hypothesis should be rejected if the index calculated for a given value of λ exceeds the correspondent c value given in Table 1.

In Table 1 we have the decision procedure for samples with sizes $n = m = 50$. So, if for example we have two samples with sizes 50 taking values in the real line and set $\lambda = 1$, we compare the WIKS index with the threshold $c = 0.2848$. If the WIKS index is greater than 0.2848 we reject the null hypothesis. Otherwise we declare the null hypothesis true.

[4] This null distribution is defined to be N(0,1), LN(0,1) or U(0,1), in the same way as the base measure, accordingly to the known support of data.

Table 1 – Threshold c values (0.95 sample quantiles of the index) for different support distributions and λ values considering $n = m = 50$.

	Distributions		
λ	N(0,1)	LN(0,1)	U([0,1])
1	0.2848	0.2927	0.2826
2	0.4755	0.4875	0.4844
3	0.6218	0.6349	0.6241
4	0.7272	0.7232	0.7215

2.3 Invariance of WIKS sampling distribution under H_0

Our index reminds the Kolmogorov-Smirnov statistics, which is invariant under H_0 (the distribution of the statistics depends only on the sample sizes). This motivates us to find out if the index exhibits similar behavior. In this section we present some results obtained during the invariance investigation. We studied some descriptive statistics of the index, as the quantiles and the mean, changing the distribution. To this study we set as prior two independent Dirichlet process with concentration parameter $K = 1$ and same base probability $N(0,1)$. The distance employed was the Kolmogorov metric $d(P_1,P_2) = \sup_x |P_1((-\infty,x]) - P_2((-\infty,x])|$ and the weight function was $W_4(t) = 1-(1-t)^4$. Below we present the procedure used to describe the sampling distribution of the WIKS index under the null.

1. Simulate two samples x and y independently with respectives sizes n and m under the null hypothesis.
2. Compute the WIKS$(\mathscr{D}_{n,m})$ index by means of Algorithm 1 with $S = 1000$.
3. Repeat the steps 1 and 2 a thousand times.
4. Compute the quantiles and the mean of the simulated values.

In Table 2, we consider 4 scenarios under the null hypothesis: Exponential(1), Beta(5,5), Skew Normal(1.5) and Uniform(50,50) with $n = 10$ and $m = 20$ and for each scenario we calculate the quantiles and the mean following the above algorithm.

Table 2 – Quantiles 0.05, 0.25, 0.5, 0.75, 0.95 and mean of WIKS index simulated from different populations P (under H_0) with $n = 10$ and $m = 20$

	Exponential(1)	**Beta(5,5)**	**Skew_Normal(1,5)**	**Uniform(50,50)**
5%	0,7651	0,7659	0,7626	0,7658
25%	0,7909	0,7922	0,7903	0,7917
50%	0,8187	0,8195	0,8204	0,8230
75%	0,8546	0,8512	0,8574	0,8602
95%	0,9155	0,9096	0,9149	0,9208
mean	0,8269	0,8257	0,8272	0,8298

In Table 3, we consider the scenarios Exponential(10), Beta(10,10), Skew Normal(1) and Uniform(100,100) with sample sizes $n = 20$ and $m = 30$.

Table 3 – Quantiles 0.05, 0.25, 0.5, 0.75, 0.95 and mean of WIKS index simulated from different populations P (under H_0) with $n = 20$ and $m = 30$

	Exponential(10)	**Beta(10,10)**	**Skew_Normal(1)**	**Uniform(100,100)**
5%	0,6747	0,6739	0,6771	0,6773
25%	0,7035	0,7056	0,7046	0,7068
50%	0,7353	0,7360	0,7344	0,7374
75%	0,7727	0,7720	0,7719	0,7755
95%	0,8439	0,8445	0,8452	0,8461
mean	0,7424	0,7432	0,7444	0,7452

In Table 4, we considered Exponential(1), Beta(1,1), Skew Normal(0.5) and Uniform(100,100) with sample sizes $n = 30$ and $m = 10$.

Table 4 – Quantiles 0.05, 0.25, 0.5, 0.75, 0.95 and mean of WIKS index simulated from different populations P (under H_0) with $n = 30$ and $m = 10$

	Exponential(1)	**Beta(1,1)**	**Skew_Normal(0.5)**	**Uniform(10,10)**
5%	0,7468	0,7458	0,7455	0,7462
25%	0,7734	0,7739	0,7719	0,7740
50%	0,8018	0,8012	0,7960	0,8028
75%	0,8392	0,8405	0,8354	0,8438
95%	0,8996	0,8989	0,9010	0,9064
mean	0,8100	0,8102	0,8067	0,8123

Note that changing the scenario under the null hypothesis, apparently there is no difference between the quantiles and the mean, providing an evidence that the sampling distribution of the WIKS index is invariant under the null hypothesis.

2.4 Simulation Study

In this section, we present a simulation study to compare our decision criteria with the Kolmogorov-Smirnov test, the Wilcoxon test and the Holmes test. For a detailed explanation of the Wilcoxon and the Kolmogorov-Smirnov test, we cite Wilcoxon (1945), Kolmogorov (1933) and Smirnov (1939). Before we start the simulation study. in the subsection below we describe in few lines the Holmes's method (HOLMES *et al.*, 2015).

2.4.1 Holmes's method

Assume that P_1 and P_2 are two populations and that $x = (x_1, \cdots, x_n)$ and $y = (y_1, \cdots, y_m)$ are samples from P_1 and P_2 respectively. Holmes presented a procedure to test the hypothesis

$$\begin{aligned} H_0 &: P_1 = P_2 \\ H_1 &: P_1 \neq P_2. \end{aligned}$$

The idea of Holmes is to quantify the evidence in favor of the null hypothesis H_0 by means of the probability $\mathbb{P}(H_0|(x,y))$ where (x,y) represent the pooled data.

To proceed with the test, Holmes consider as prior the Pólya tree[5] centered on some distribution G. Under the null, he considered that the pooled date came from a Pólya tree $F^{1,2}$ and under the alternative hypothesis $P_1 \sim F^1$ and $P_2 \sim F^2$ are modelled as independent Pólya tree priors.

As Holmes argues, in such a way the test procedure could be faced as a model comparison problem. From Bayes theorem

$$\mathbb{P}(H_0|(x,y)) = \mathbb{P}((x,y)|H_0)\mathbb{P}(H_0),$$

therefore the posterior odds is equal to

$$\frac{\mathbb{P}(H_0|(x,y))}{\mathbb{P}(H_1|x,y)} = \frac{\mathbb{P}((x,y)|H_0)}{\mathbb{P}(x,y|H_1)}\frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)}. \tag{2.7}$$

Holmes found an explicit expression to the Bayes factor

$$\frac{\mathbb{P}((x,y)|H_0)}{\mathbb{P}(x,y|H_1)}$$

2.4.2 Power Simulation

To compare the power functions, just as Holmes presented in his article (HOLMES *et al.*, 2015), we consider 8 scenarios representing different departures from the null

1. Beta symmetry: $\mathbf{X} \sim Beta(1,1)$ and $\mathbf{Y} \sim Beta(\theta,\theta)$, $\theta = [1,6]$
2. Gamma shape: $\mathbf{X} \sim Gamma(3,2)$ and $\mathbf{Y} \sim Gamma(\theta,2)$, $\theta = [3,6]$
3. Lognormal mean shift: $\log\mathbf{X} \sim N(0,1)$ and $\log\mathbf{Y} \sim N(\theta,1)$, $\theta = [0,3]$
4. Lognormal variance shift: $\log\mathbf{X} \sim N(0,1)$ and $\log\mathbf{Y} \sim N(0,\theta)$, $\theta = [1,5]$
5. Normal mean shift: $\mathbf{X} \sim N(0,1)$ and $\mathbf{Y} \sim N(\theta,1)$, $\theta = [0,3]$
6. Normal mixtures: $\mathbf{X} \sim N(0,1)$ and $\mathbf{Y} \sim \frac{1}{2}N(-\theta,1) + \frac{1}{2}N(\theta,1)$, $\theta = [0,3]$
7. Fat tails: $\mathbf{X} \sim N(0,1)$ and $\mathbf{Y} \sim t(\theta^{-1})$, $\theta = [10^{-3},10]$.
8. Normal variance shift: $\mathbf{X} \sim N(0,1)$ and $\mathbf{Y} \sim N(0,\theta)$, $\theta = [1,4]$

The comparison is made in terms of the "power to detect the alternative". That is, we compare the power functions of the different decision rules at level 5%. The power function is calculated in according to Algorithm 2.

[5] The Pólya tree is a nonparametric prior proposed by Lavine *et al.* (1992)

Algorithm 2 – Power Function

1: Fix a value for θ on the grid
2: Draw $\mathbf{X}$ and $\mathbf{Y}$ with sample sizes $n = m = 50$ from the respective distribution determined by θ
3: Decide to reject or not the null hypothesis
4: Repeat this procedure 1000 times
5: Compute the proportion of times that the null hypothesis was rejected
6: Repeat steps 1 to 5 for a new value of θ

Figure 2 presents the power function for the different methods considered under the above scenarios.

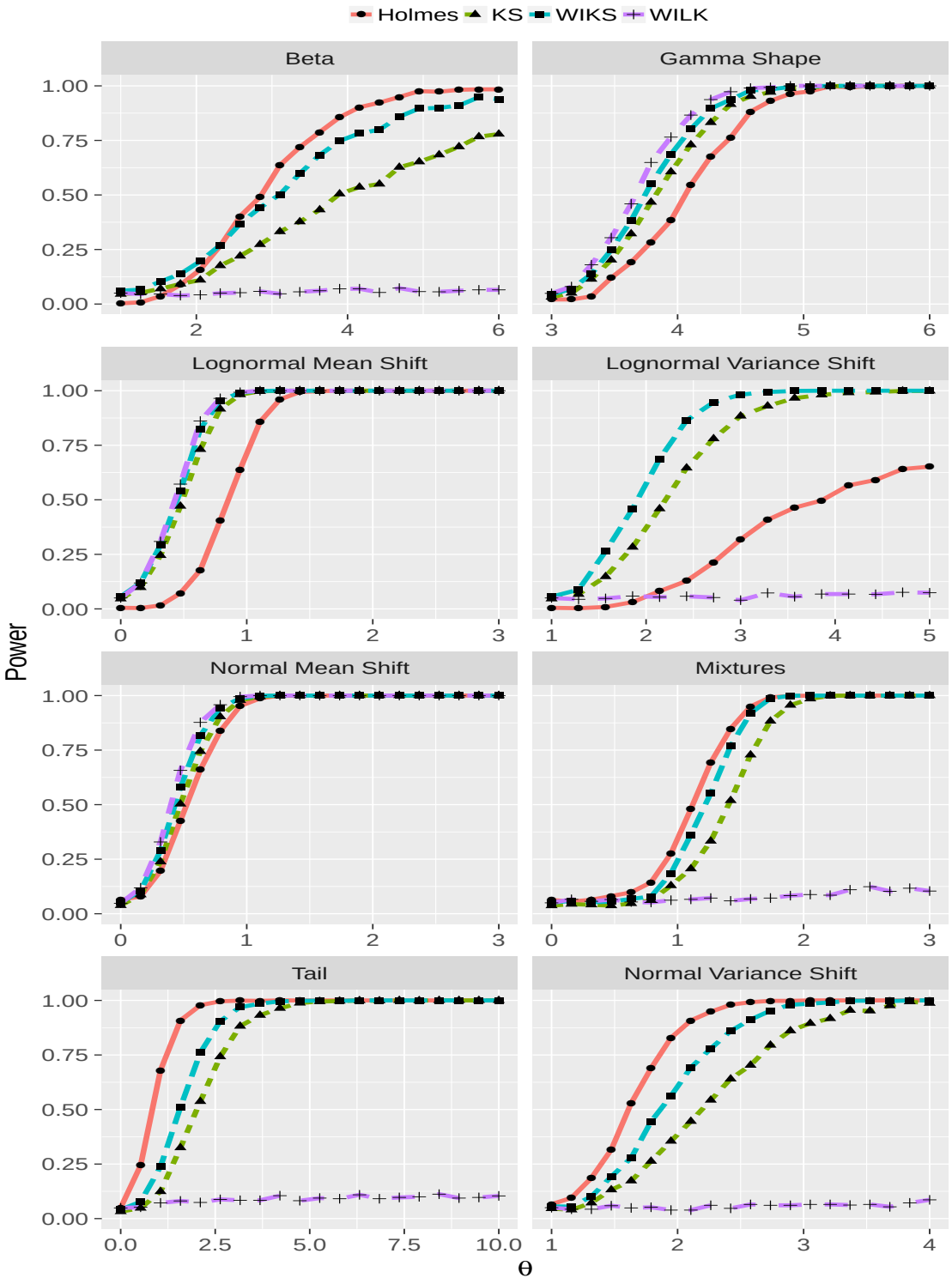

Figure 2 – Power function comparison for different settings

Note that the Wilcoxon test is able to detect changes in the location parameter and gamma shape parameter (scenarios (2), (3) and (5)), but shows extremely low power in detecting the alternative for all other scenarios. The Kolmogorov-Smirnov test presents a medium power performance over all scenarios. The Holmes test present low power in detecting the variance shift in lognormal distribution. The WIKS overperfomed its competitor in 1 scenarios ((4)).

Compared with the another Bayesian nonparametric test (Holmes method), our index presentd better performance in the location parameter, the gamma shape parameter and the lognormal variance shift (scenarios (2), (3), (4) and (5)). Besides, the Kolmogorov metric does not work well with differences in distribution tails, so changing the metric in this scenarios could improve the results of our approach.

Another interesting point in Figure 2, is that the Holmes method does not control the type one error. While we study the Holmes's routines, that is worth mentioning was in Octave, we observe that for each point under the null hypothesis he established a criterion, that is, for each scenario he determined the cutoff point to proceed with the test and this is not clear in his article. Therefore, for Holmes test, for each distribution and sample size a cut-off point must be specified. In the power function for Holmes method, we defined the cutoff point for all scenarios as the one with the maximum 5% of error of type one.

Additionally, we also investigate the consistency of the proposed method, that is, we study the power function under the alternative for increasing sample sizes. In order to do so, we consider the scenarious:

1. Beta symmetry: $\mathbf{X} \sim Beta(1,1)$ and $\mathbf{Y} \sim Beta(3,3)$
2. Gamma shape: $\mathbf{X} \sim Gamma(3,2)$ and $\mathbf{Y} \sim Gamma(4,2)$
3. Lognormal mean shift: $\log \mathbf{X} \sim N(0,1)$ and $\log \mathbf{Y} \sim N(1,1)$
4. Lognormal variance shift: $\log \mathbf{X} \sim N(0,1)$ and $\log \mathbf{Y} \sim N(0,3)$
5. Normal mean shift: $\mathbf{X} \sim N(0,1)$ and $\mathbf{Y} \sim N(1,1)$
6. Normal mixtures: $\mathbf{X} \sim N(0,1)$ and $\mathbf{Y} \sim \frac{1}{2}N(-1,1) + \frac{1}{2}N(1,1)$
7. Fat tails: $\mathbf{X} \sim N(0,1)$ and $\mathbf{Y} \sim t(1.25^{-1})$
8. Normal variance shift: $\mathbf{X} \sim N(0,1)$ and $\mathbf{Y} \sim N(0,3)$

The results are reported in Figure 3 for $n = m = 10, 20, 30 \cdots, 100$. In all scenarios we can observe the expected behavior, increasing the sample size increases the power. Again, the WIKS overperfomed the Holmes method in the scenarios ((2), (3), (4) and (5)).

2.5 Application

We apply our methods to a data set of three groups of patients (CG: the control group, MCD: with mild cognitive decline and AD: with Alzheimer's disease) submitted to a questionnaire for Alzheimer's disease diagnostic (CAMCOG) with respective sample sizes 39, 45 and

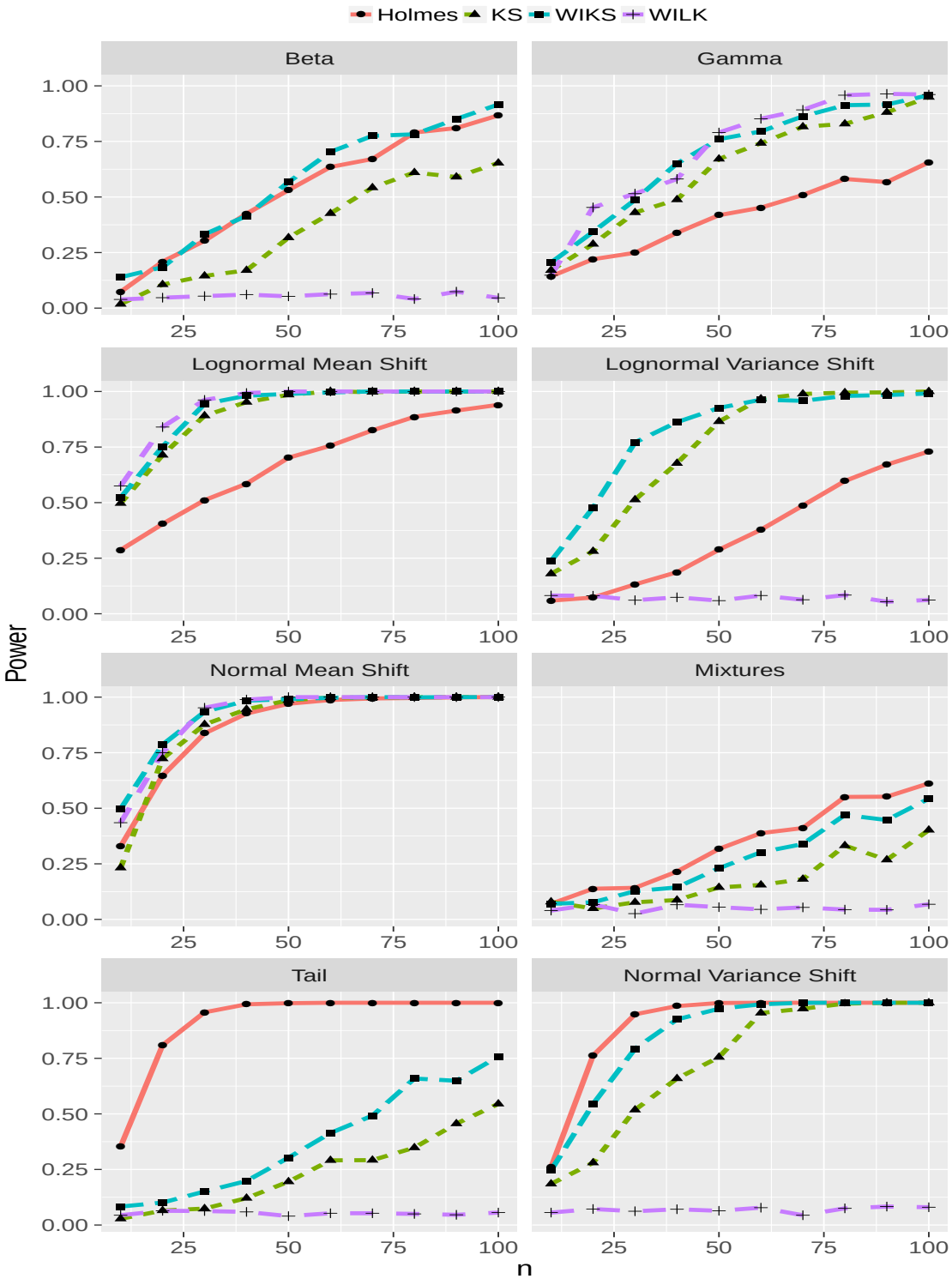

Figure 3 – Power for different settings changing the sample size

50. More details on this dataset can be obtained in (Cecato *et al.* (2016)). The main idea is to quantify the differences between the groups using our methods.

Figure 4 shows that all groups present different behavior with respect the score obtained from CAMCOG. The group with Alzheimer's disease (AD) has the lowest CAMCOG scores and the control group (CG) the highest ones. The group with mild cognitive decline (MCD) has score

values in-between the other two groups. Thus, it is expected that the WIKS index will be greater when comparing AD and CG groups than for the other comparisons. In fact, for AD vs CG, CG vs MCD and MCD vs AD the WIKS index are $0.9993, 0.9629, 0.9312$ with respective thresholds 0.7558, 0.7681 and 0.7314, leading to the rejection of null for all pairwise comparisons. The index was computed with concentration parameter $K = 1$, base probability $G \sim N(0,1)$, weight function $W(t) = 1-(1-t)^4$ and the Kolmogorov metric. From this analysis, we conclude that CAMCOG is an useful tool for initial diagnostic of Alzheimer disease, being able to properly distinguish between the three groups.

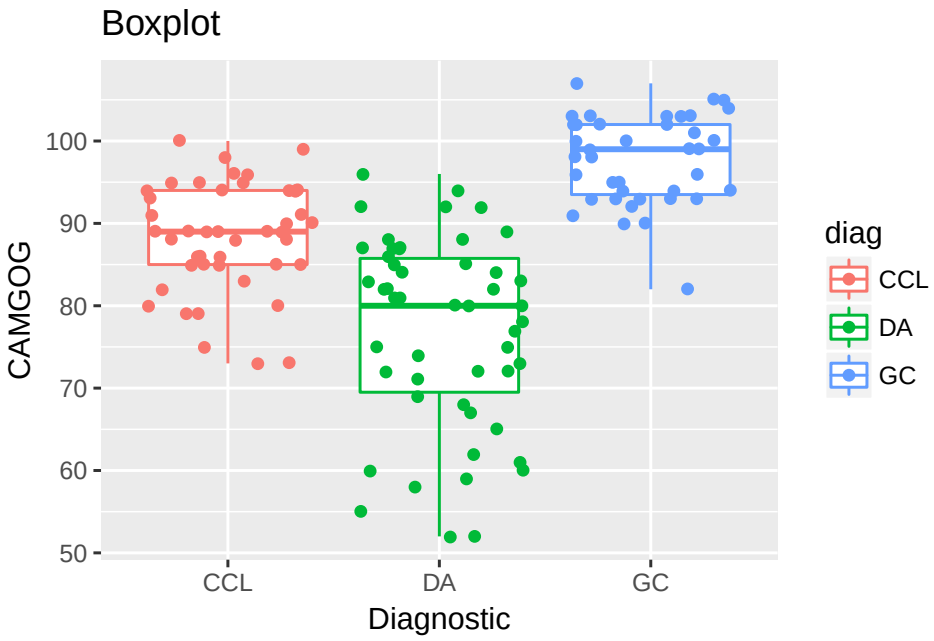

Figure 4 – Boxplot of CAMCOG scores for the groups MCD, AD and CG.

CHAPTER

3

FINAL REMARKS AND FUTURE WORKS

We propose a method to compare two populations P_1 and P_2 that relies on a Bayesian nonparametric discrepancy index (WIKS) defined as a weighted average of the posterior survival function of the Kolmogorov distance $d(P_1, P_2)$. The WIKS index can also be expressed as the posterior expectation in terms of $d(P_1, P_2)$, which makes it easier to compute its value using samples of the posterior distribution. The WIKS definition can be seen as an aggregated evidence against the null and the proposed decision procedure is the Bayes rule under a suitable loss function.

In a power function simulation study, WIKS presents better performance than the well-established Wilcoxon and Kolmogorov-Smirnov tests. When compared to the method proposed by Holmes *et al.* (2015), WIKS shows similar performance in many settings and is superior when the support of data are restricted to the positive real numbers or the unitary interval. For a data-set on questionaire scores used for Alzheimer diagnose applied to 3 groups, WIKS could correctly indentify the difference between the groups.

We conclude that WIKS is a powerful and flexible method to compare populations with low computational cost. Even thought we have chosen the Dirichlet Process as our prior, any other nonparametric (e.g, the Polya tree or the Beta processes) or even parametric prior could be used without the need of adjustments: WIKS computation only requires a sampling algorithm for posterior simulation. Further investigation is needed to assess the effect of the choices of the metric d and the weight function w on the performance of the method. Future research directions are extending the methods presented here to goodness-of-fit problems and investigating the performance in high-dimensional settings.

BIBLIOGRAPHY

BALDI, P.; LONG, A. D. A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. **Bioinformatics**, Oxford Univ Press, v. 17, n. 6, p. 509–519, 2001. Citation on page 26.

BASU, S.; CHIB, S. Marginal likelihood and bayes factors for dirichlet process mixture models. **Journal of the American Statistical Association**, Taylor & Francis, v. 98, n. 461, p. 224–235, 2003. Citation on page 26.

BEHRENS, W. Ein beitrag zur fehlerberechnung bei wenigen beobachtungen. landw. jb. 68, 807-37. **Behrens80768Landw. Jb**, 1929. Citation on page 25.

BERNARDO, J. M.; SMITH, A. F. **Bayesian theory**. [S.l.]: IOP Publishing, 2001. Citation on page 26.

BORGWARDT, K. M.; GHAHRAMANI, Z. Bayesian two-sample tests. **arXiv preprint arXiv:0906.4032**, 2009. Citations on pages 25 and 26.

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, v. 28, n. 5, p. 825–832, 2016. Citation on page 39.

CRESSIE, N.; WHITFORD, H. How to use the two sample t-test. **Biometrical Journal**, Wiley Online Library, v. 28, n. 2, p. 131–148, 1986. Citation on page 25.

DEGROOT, M. H. **Optimal statistical decisions**. [S.l.]: John Wiley & Sons, 2005. Citation on page 31.

FERGUSON, T. S. A bayesian analysis of some nonparametric problems. **The annals of statistics**, JSTOR, p. 209–230, 1973. Citation on page 31.

FISHER, R. A. The fiducial argument in statistical inference. **Annals of eugenics**, Wiley Online Library, v. 6, n. 4, p. 391–398, 1935. Citation on page 25.

GÖNEN, M.; JOHNSON, W. O.; LU, Y.; WESTFALL, P. H. The bayesian two-sample t test. **The American Statistician**, Taylor & Francis, v. 59, n. 3, p. 252–257, 2005. Citation on page 26.

GOOD, I. J. The bayes/non-bayes compromise: A brief review. **Journal of the American Statistical Association**, Taylor & Francis, v. 87, n. 419, p. 597–606, 1992. Citation on page 32.

HALPERIN, M. Extension of the wilcoxon-mann-whitney test to samples censored at the same fixed point. **Journal of the American Statistical Association**, Taylor & Francis, v. 55, n. 289, p. 125–138, 1960. Citation on page 25.

HOLMES, C. C.; CARON, F.; GRIFFIN, J. E.; STEPHENS, D. A. *et al.* Two-sample bayesian nonparametric hypothesis testing. **Bayesian Analysis**, International Society for Bayesian Analysis, v. 10, n. 2, p. 297–320, 2015. Citations on pages 13, 26, 34, 35, and 43.

KASS, R. E.; RAFTERY, A. E. Bayes factors. **Journal of the american statistical association**, Taylor & Francis Group, v. 90, n. 430, p. 773–795, 1995. Citation on page 26.

KOLMOGOROV, A. N. **Sulla determinazione empirica di una legge di distribuzione**. [S.l.]: na, 1933. Citations on pages 25 and 34.

LABADI, L. A.; MASUADI, E.; ZAREPOUR, M. Two-sample bayesian nonparametric goodness-of-fit test. **arXiv preprint arXiv:1411.3427**, 2014. Citation on page 26.

LAVINE, M. *et al.* Some aspects of polya tree distributions for statistical modelling. **The annals of statistics**, Institute of Mathematical Statistics, v. 20, n. 3, p. 1222–1235, 1992. Citation on page 35.

MA, L.; WONG, W. H. Coupling optional pólya trees and the two sample problem. **Journal of the American Statistical Association**, Taylor & Francis, v. 106, n. 496, p. 1553–1565, 2011. Citation on page 26.

MANN, H. B.; WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. **The annals of mathematical statistics**, JSTOR, p. 50–60, 1947. Citation on page 25.

NEYMAN, J.; PEARSON, E. On the problem of the most efficient tests of statistical hypotheses. **J. Neyman and ES Pearson**, p. 140–85, 1967. Citation on page 25.

PEARSON, E. S.; NEYMAN, J. **On the problem of two samples**. [S.l.]: Imprimerie de l'university, 1930. Citations on pages 11, 13, and 25.

RACHEV, S. T.; KLEBANOV, L.; STOYANOV, S. V.; FABOZZI, F. **The methods of distances in the theory of probability and statistics**. [S.l.]: Springer Science & Business Media, 2013. Citation on page 27.

SETHURAMAN, J. A constructive definition of dirichlet priors. **Statistica sinica**, JSTOR, p. 639–650, 1994. Citation on page 47.

SHAKED, M.; SHANTHIKUMAR, J. G. **Stochastic orders**. [S.l.]: Springer Science & Business Media, 2007. Citation on page 30.

SMIRNOV, N. V. Estimate of deviation between empirical distribution functions in two independent samples. **Bulletin Moscow University**, v. 2, n. 2, p. 3–16, 1939. Citations on pages 25 and 34.

SWARTZ, T. Nonparametric goodness-of-fit. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 28, n. 12, p. 2821–2841, 1999. Citation on page 29.

WELCH, B. L. The generalization ofstudent's' problem when several different population variances are involved. **Biometrika**, JSTOR, v. 34, n. 1/2, p. 28–35, 1947. Citation on page 25.

WILCOXON, F. Individual comparisons by ranking methods. **Biometrics bulletin**, JSTOR, v. 1, n. 6, p. 80–83, 1945. Citations on pages 25 and 34.

APPENDIX

A

STICK-BREAKING REPRESENTATION

Sethuraman (1994) presented a simple constructive definition of the Dirichlet process. Its representation simplify the way to approximate a sample from the Dirichlet process.

Let $Y_1, Y_2, \cdots$ independent and identically random variables with distribution β. Let $p_1, p_2, \cdots$ probabilities from a discrete distribution over the integers with discrete failure rate $\theta_1, \theta_2 \cdots$ which are independent and identically distributed with a Beta distribution $B(1, K)$, with K a constant. Define

$$\mathscr{P}(\theta, Y, B) = \mathscr{P}(B) = \sum_{n=1}^{\infty} p_n \delta_{Y_n}(B) \tag{A.1}$$

$\mathscr{P}$ is the random probability that put weights p_n at the degenerated measure δ_{Y_n}. Next we present the algorithm to approximate a sample from de Dirichlet process.

Algorithm 3 – Dirichlet Process Simulation with base probability β and concentration parameter k

procedure SAMPLE PATH OF DIRICHLET PROCESS(β, k)
    Sample $y_1, y_2, \cdots, y_n$ from the base probability
    Sample $b_1, b_2, \cdots, b_n$ from the beta distribution $B(1, k)$
    $p_1 \leftarrow b_1$
    $i \leftarrow 1$
    while $i \leq n-1$ **do**
        $p_{i+1} \leftarrow b_{i+1} * \prod_{j=1}^{i-1}(1 - b_j)$
        $i \leftarrow i + 1$
    end while
    For each sample y_i assign the weight p_i
    return $\{(y_1, p_1), (y_2, p_2), \cdots, (y_n, p_n)\}$
end procedure

We use the algorithm 3 to sample paths from the Dirichlet Process in all simulations in this dissertation.

ANNEX

A

R SCRIPTS

Source code 1 – Functions employed in the simulation

```
 1:
 2: # rdirichlet: returns a sample from Dirichlet distribution
 3: rdirichlet <- function(alpha) {
 4:   z <- rgamma(length(alpha), alpha)
 5:   return(z/sum(z))
 6: } # end-function
 7:
 8: # Returns one realization from the posterior DP
 9: post.dp <- function(data, alpha0, r.G0, K = 500)
10: {
11:   # data: vector of data values
12:   # alpha0: concentration parameter (prior)
13:   # r.G0: function that draws from the base probability (prior)
14:   # K: truncation number of simulations for the stick-breaking
       process
15:
16:   # Draw values from the base probability G (posterior) ->
       theta
17:   n <- length(data)
18:   theta <- rbinom(K, 1, alpha0/(alpha0 + n)) # indicator values
        to simulate from G0
19:   k <- sum(theta)
20:   theta.prior <- r.G0(k)
21:   theta.data <- sample(data, K - k, replace = TRUE)
22:   theta[theta == 1] <- theta.prior
23:   theta[theta == 0] <- theta.data
```

```
24:
25:   # Draw values for the weights associated to the theta values
26:   betas <- rbeta(K , 1, alpha0 + n)
27:   w <- c(1, cumprod(1 - betas))[1:K]*betas
28:
29:   # Remove ties from theta
30:   w.new <- tapply(w, theta, sum)
31:   theta.new <- sort(unique(theta))
32:   return(list(theta = theta.new, w = w.new))
33: } # end-function
34:
35: # dist: returns the Kolmogorov distance between 2 discrete
      distribution functions
36: dist <- function(x, y, f = rep(1/length(x), length(x)), g = rep
      (1/length(y), length(y))) {
37:   # (x, f): values and weights for sample 1 (x ordered
      increasingly)
38:   # (y, g): values and weights for sample 2 (y ordered
      increasingly)
39:   Fx <- ewcdf(x, f)
40:   Gy <- ewcdf(y, g)
41:   z <- c(x, y)
42:   max(abs(Fx(z) - Gy(z)))
43: } # end-function
44:
45: # weight: returns the weight function for a given lambda
      parameter
46: weight <- function(x, lambda) {1-(1-x)^lambda}
47:
48: # index: returns a Monte Carlo aproximation for the index
49: index <- function(lambda, data, alpha0, r.G0, K = 500, S =
      1000){
50:   # lambda: parameter of the weight function
51:   # data: list with the 2 data sets
52:   # alpha0: concentration parameter for the DP (prior)
53:   # r.G0: function that draws from the base probability (prior)
54:   # K: truncation number for the stick-breaking process
55:   # S: number of Monte Carlo simulations
56:   d <- numeric(S)
57:   for (i in 1:S){
58:     v1 <- post.dp(data[[1]], alpha0, r.G0, K)
59:     v2 <- post.dp(data[[2]], alpha0, r.G0, K)
```

```
60:     d[i] <- dist(v1[[1]], v2[[1]], v1[[2]], v2[[2]])
61:   } # end-for
62:   mean(weight(d, lambda))
63: } # end-function
```