2017

ON THE EQUIVALENCE BETWEEN BAYESIAN AND FREQUENTIST NONPARAMETRIC HYPOTHESIS TESTING

Qiuchen Hai
Michigan Technological University, qhai@mtu.edu

ON THE EQUIVALENCE BETWEEN BAYESIAN AND FREQUENTIST NONPARAMETRIC HYPOTHESIS TESTING

By

Qiuchen Hai

A REPORT

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

In Mathematical Sciences

MICHIGAN TECHNOLOGICAL UNIVERSITY

2017

This report has been approved in partial fulfillment of the requirements for the Degree of MASTER OF SCIENCE in Mathematical Sciences.

Department of Mathematical Sciences

Report Advisor:	*Dr. Jianping Dong*
Committee Member:	*Dr. Yu Cai*
Committee Member:	*Dr. Renfang Jiang*
Department Chair:	*Dr. Mark Gockenbach*

Contents

List of Figures

List of Tables

Abstract

Testing of hypotheses about the population parameter is one of the most fundamental tasks in the empirical sciences and is often conducted by using parametric tests (e.g., the t-test and F-test), in which they assume that the samples are from populations that are normally distributed. When the normality assumption is violated, nonparametric tests are employed as alternatives for making statistical inference. In recent years, the Bayesian versions of parametric tests have been well studied in the literature, whereas in contrast, the Bayesian versions of nonparametric tests are quite scant (for exception, [17]) in the literature, mainly due to the lack of sampling distribution of data.

It is well known that like the frequentist counterparts, the Bayesian tests perform well in practical applications, whereas unlike the frequentist ones, they are generally fail to control the Type I error and can even result in different decisions from them. To avoid these issues, we integrate the ideas of [17] and [2] and develop Bayes factor tests for comparing the difference between the means among several populations, which can not only control the Type I error, but also allow researchers to make the identical decisions between frequentists and Bayesians on the basis of the observed data. In addition, they depend on the data only through nonparametric statistics and can thus be easily computed, so long as one has conducted the nonparametric tests. More importantly, they can quantify evidence from empirical data favoring the null hypothesis, and this property is not shared by the frequentist counterparts, which lack the ability to quantify evidence favoring the null hypothesis in the case of failing to reject the null hypothesis.

Chapter 1

Introduction

Testing of hypotheses about a population parameter is one of the most fundamental tasks in virtually most areas of scientific study, as it helps researchers answer practical questions: Did the gasoline price increase by an average of only \$0.10 per gallon last year? Is there a difference in median yields per acre between two fertilizers A and B of growing corn? Do the different types of diets appear to affect the amount of iron present in the livers of white rates after feeding them one of the diets for a certain period of time. There are always two hypotheses involved for these problems at hand: one is the alternative hypothesis (H_1), which represents the statement that researchers would like to support, the other is the null hypothesis (H_0), which is an initial statement that researchers may specify according to their prior knowledge. For instance, the hypotheses correspond to oil price example are: H_0: the oil price has not increased by \$0.10 per gallon and H_1: the oil price was increased by \$0.10 per gallon.

The problem of hypothesis testing is usually covered in most elementary statistics courses, and in particular, we were taught how to implement parametric tests in making statistical inference, such as the t-test and the analysis of variance (ANOVA) for comparing group means. A parametric test often requires certain assumptions of the parameters of the population distribution. For instance, the t-test assumes the samples to be drawn from normal populations, even though this assumption is seldom met in practical applications, especially when the data exhibit heavy-tailed behavior. When the normality assumption is violated and/or outliers are present, the power of parametric tests can drop considerably, and thus nonparametric tests, free of the distribution assumption of the data, are good alternative to parametric ones. For instance, the Wilcoxon signed rank test is generally more powerful than the t-test for comparing the difference between two population means for paired data, when the data are asymmetric while heavy-tailed. More details about the implementation of nonparametric statistics can be found in [1].

No matter whether we adopt the parametric or nonparametric testing procedures, the common decision rule of these tests for rejecting or failing to reject H_0 is based on the p-value from a certain test statistic: we reject H_0 if the p-value $< \alpha$, say $\alpha = 5\%$, a specified significance level. The advantage of this decision rule is its ability to control the Type I error rate, whereas its drawback is that it provides little information about the truth of H_0 if it is not rejected. In addition, the p-value approach has a tendency to overstate the evidence against H_0, leading to instances where it has been banned by [15]. This motivates researchers to consider Bayesian hypothesis testing as an alternative to the p-value test, given that Bayesian procedures to model testing can quantify evidence in favor of both hypotheses.

Bayesian testing procedures are often conducted by comparing the posterior probability of each hypothesis. In this report, let $p(Y \mid \theta_j)$ and $\pi_j(\theta_j)$ be the likelihood function of Y and the prior for θ_j under H_j and let π_j be the prior probability for H_j satisfying $\pi_0 + \pi_1 = 1$ for $j = 1, 2$. In the absence of prior knowledge, the equal prior probabilities can be assigned for both hypotheses (i.e., $\pi_0 = \pi_1 = 1/2$), the so-called assumption of equipoise in this report. By using Bayes theorem, the posterior probability of H_j is given by

$$P(H_j \mid Y) = \frac{\pi_j m_j(Y)}{\pi_0 m_0(Y) + \pi_1 m_1(Y)}, \tag{1.1}$$

where $m_j(Y)$ represents the marginal likelihood of Y given H_j, i.e.,

$$m_j(Y) = \int p(Y \mid \theta_j)\pi_j(\theta_j)\, d\theta_j. \tag{1.2}$$

Note that the posterior probability of H_1 can be rewritten in the form

$$P(H_1 \mid Y) = \frac{\pi_1 \mathrm{BF}_{10}}{\pi_0 + \pi_1 \mathrm{BF}_{10}} = \left[1 + \frac{\pi_0}{\pi_1}\frac{1}{\mathrm{BF}_{10}}\right]^{-1}, \tag{1.3}$$

where BF_{10} is the Bayes factor (BF) between hypotheses H_1 to H_0 given by

$$\mathrm{BF}_{10} = \frac{m_1(Y)}{m_0(Y)}. \tag{1.4}$$

One appearing property of the BF is that it represents the relative plausibility of the observed data under two considered hypotheses. For example, $\mathrm{BF}_{10} = 10$ means that the data are 10 times as more likely to be generated under H_1 than under H_0. We here refer the interested readers to [5] and [7] for a detailed interpretation of the BF. As a Bayes test of decision making, the null hypothesis is rejected if the BF (equivalently, the posterior probability of H_1) exceeds a certain threshold, and in general, we are more likely to choose H_1 (H_0) if $\text{BF}_{10} > 1$ (< 1). The value of 1 results in an optimal action to reject H_0 under the zero-one loss function; see [9]. Here, the optimal decision means that the expected posterior loss of failing to reject H_0 exceeds the expected posterior loss of rejecting H_0.

Unlike the p-value approach, the BF may fail to control the Type I error and result in different decisions from the parametric and nonparametric tests. To remedy these limitations, [6] followed the idea of a uniformly most powerful test of statistical hypotheses and proposed a uniformly most powerful Bayes test (for short, UMPBT), which was obtained by maximizing the probability that the BF favoring H_1 exceeds a specified threshold. The UMPBT can lead to an identical decision with the frequentist counterpart, whereas it only exists in a few testing scenarios. This motivates [2] to consider a natural extension of the UMPBT, the so-called restricted most powerful Bayes test (RMPBT), which is obtained by restricting the class of priors for the unknown parameters under the alternative hypothesis to a certain parametric class. They have shown that the RMPBT performs well for testing the regression coefficients in linear models. More recently, [16] developed the Bayes t-tests based on the RMPBT for testing the presence of correlations between two continuous random variables. The developed Bayes t-tests depend simply on the t-statistics and can also result in an identical decision with the t-tests on the basis of the observed data.

It is of particular note that the current literature mainly focuses on the implementation of the UMPBT for hypothesis testing in a parametric setting. To the best of our knowledge, it is unclear whether the RMPBT can be generalized to develop efficient Bayesian testing approaches in a nonparametric setting. In this report, we follow the seminal work of [17] and develop the BFs to model testing based on a combined use of nonparametric statistics and the UMPBT. The developed Bayesian tests enjoy various appealing properties: (i) they depend simply on the commonly used nonparametric statistics and their associated quantiles of the nonparametric statistics, (ii) they can be easily computed, so long as researchers are familiar with the nonparametric paradigm, and more importantly, (iii) they can result in an identical conclusion with the associated nonparametric tests, which allows researchers to interpret the results from both Bayesian and frequentist points of view.

This report is organized as follows. In Chapter 2, we briefly overview several commonly used nonparametric test statistics covered in most statistics courses. We derive Bayesian nonparametric tests based on a combined use of nonparametric statistics and the RMPBT, and we then discuss their corresponding properties in Chapter 3. We evaluate the performance of the developed Bayesian tests through using simulations and three real-data examples in Chapter 4. Concluding remarks are provided in Chapter 5 with computer codes written in R [12] and mathematical derivations given in the Appendix.

Chapter 2

Frequentist nonparametric test statistics

In this chapter, we overview several commonly used nonparametric tests, which are treated as alternatives to their parametric analogs when the normality assumption of the data appears to be violated. In Section 2.1, we discuss the Wilcoxon signed rank test for one-sample and/or paired-sample problem. In Section 2.2, we provide the Mann-Whitney-Wilcoxon test for two independent sample problem, and in Section 2.3, we consider the Kruskal-Wallis test for locations in several independent samples.

2.1 The Wilcoxon signed rank test

Let (X_i, Y_i) be a paired observation and $D_i = X_i - Y_i$ for $i = 1, \cdots, n$. In the one-sample problem, D_i's can be viewed as observations in the sample. Suppose that all D_i's are independent and identically distributed with a distribution function $F(D \mid \theta)$, which is assumed to be symmetric with the unknown parameter θ. We wish to test the hypothesis if the unknown parameter is $\theta_0 = 0$, that is,

$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta \neq 0. \tag{2.1}$$

Note that in case of $\theta_0 \neq 0$, we reparameterize each observation by using $\widetilde{D}_i = D_i - \theta_0$. To obtain the Wilcoxon signed rank test, we calculate the absolute values of the differences, denoted by $|D_1|, \cdots, |D_n|$, and sort them in an ascending order. Let R_i be the rank of $|D_i|$ for $i = 1, \cdots, n$. The test statistic T is defined as the sum of the positive signed ranks given by

$$T = \sum_{i=1}^{n} \big(R_i \text{ where } D_i \text{ is positive}\big). \tag{2.2}$$

We reject H_0 at the level of α if T is less than its $\alpha/2$ quantile ($\tau_{\alpha/2}$) or greater than its $1-\alpha/2$ quantile ($n(n+1)/2-\tau_{\alpha/2}$) for the distribution of T under H_0. The value of $\tau_{\alpha/2}$ can be found in Table A12 of [1] or Table A.4 of [3]. It deserves mentioning that T and $\tau_{\alpha/2}$ can also be easily calculated in R by using *wilcox.test*() and *qsignrank*() functions, respectively; see [12].

When the sample size is large ($n \geq 20$), T can be approximately normally distributed with mean $\mathbb{E}[T] = n(n+1)/4$ and variance $\text{var}(T) = n(n+1)(2n+1)/24$; see [3]. The standardized version of T, denoted by T^*, is defined as

$$T^* = \frac{T - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}, \tag{2.3}$$

which has limiting normal distributions under both null and alternative hypotheses. This limiting property plays a key role in deriving the BF for testing the hypotheses in (2.1) using nonparametric statistics studied by [17].

2.2 The Mann-Whitney-Wilcoxon test

Let $X = (x_1, \cdots, x_{n_1})'$ and $Y = (y_1, \cdots, y_{n_2})'$ be two data vectors from two populations: the first sample comes from a control group having a distribution function F, and the second from a treatment group with a distribution function G. Without loss of generality, we assume $n_1 \leq n_2$. Suppose also that a location-shift model for G, such that $G(t) = F(t - \theta)$ for some $\theta \in R$. After an appropriate reparametrization mentioned above, we are interested in testing

$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta \neq 0. \tag{2.4}$$

For the calculation of the Mann-Whitney-Wilcoxon test, we combine samples of X and Y and calculate the rank of X_j in the combined sample, denoted by S_j for

$j = 1, \cdots, n_1$. The test statistic W is given by

$$W = \sum_{j=1}^{n_1} S_j. \tag{2.5}$$

We reject H_0 at the level of α if W is less than its $\alpha/2$ quantile ($\omega_{\alpha/2}$) or greater than its $1-\alpha/2$ quantile ($n_1(n_1+n_2+1)/2-\omega_{\alpha/2}$) of W under H_0. The value of $\omega_{\alpha/2}$ can be found from Table A7 of [1] when $n_1 \leq 20$ and $n_2 \leq 20$ or can be approximated by a standard normal quantile given in Table A1 of [1] for larger sample sizes. Similar to the Wilcoxon signed rank test, W and $\omega_{\alpha/2}$ can be easily calculated in R by using $wilcox.test()$ and $qsignrank()$ functions, respectively.

When the sample sizes are large, T can be approximately normally distributed with mean $\mathbb{E}[W] = n_1(n_1+n_2+1)/2$ and variance $\text{var}(W) = n_1 n_2(n_1+n_2+1)/12$. The standardized version of W, denoted by W^* is defined as

$$W^* = \frac{T - n_1(n_1+n_2+1)/2}{\sqrt{n_1 n_2 (n_1+n_2+1)/12}}, \tag{2.6}$$

which has limiting normal distributions under both null and alternative hypotheses; see [3].

2.3 The Kruskal-Wallis test

The Mann-Whitney-Wilcoxon test is often employed to test difference of two independent samples, and later on, [8] extended this test for analyzing $k(\geq 3)$ independent samples. Suppose that the data consist of k independent random samples of different sample sizes. Let $X_i = (X_{i1}, \cdots, X_{in_i})$ be the ith sample of size n_i drawn from a distribution function $F(x - \theta_i)$, where θ_i represents the median of the ith population for $i = 1, \cdots, k$. We are interested in testing

$$H_0 : \theta_1 = \cdots = \theta_k \quad \text{versus} \quad H_1 : \theta_i \neq \theta_j, \quad \text{for some} \quad 1 \leq i, j \leq k. \tag{2.7}$$

Let $R(X_{ij})$ be the rank of X_{ij} and R_i be the sum of the ranks assigned to the ith sample for $i = 1, \cdots, k$. The Kruskal-Wallis test statistic U is defined as

$$U = \frac{12}{n(n+1)} \sum_{i=1}^{k} n_i \left(\frac{R_i}{n_i} - \frac{n+1}{2} \right)^2, \tag{2.8}$$

where $n = \sum_{i=1}^{k} n_i$ is the total sample size. We reject H_0 at the level of α if U is greater than its $1 - \alpha$ quantile (ν_α) from the null distribution of U. The value of ν_α can be found in [4] and [11]. In the large sample approximation, the approximate quantile can be obtained by the quantile of the central chi-square distribution with $k - 1$ degrees of freedom. This is because when H_0 is true, the test statistic U follows an asymptotic chi-square distribution with $k-1$ degrees of freedom, denoted by χ^2_{k-1}, as $n_i \to \infty$ simultaneously for $i = 1, \cdots, k$. Under H_1, the limiting distribution of U has a non-central χ^2 distribution with $k - 1$ degrees of freedom, denoted by $\chi^2_{k-1}(\rho)$, where

$$\rho = 12 \left\{ \int_{-\infty}^{\infty} f^2(x)\, dx \right\}^2 \sum_{i=1}^{k} a_i (\Delta_i - \overline{\Delta}),$$

where $f(\cdot)$ is the probability density function of F, $a_i = n_i/n$, and $\overline{\Delta} = \sum_{i=1}^{k} a_i \Delta_i$. The limiting distributions of U under both hypotheses paly an important role in developing the Bayes factor test to model hypotheses in (2.7) based on the Kruskal-Wallis test.

Chapter 3

Bayesian hypothesis testing

In this chapter, we overview the Bayes factors using nonparametric statistics of [17] (Section 3.1) and restricted most powerful Bayes test of [2] (Section 3.2). In Section 3.3, we combine ideas of these two procedures and develop alterative Bayesian testing procedures using nonparametric statistics in Chapter 2.

3.1 Bayesian modeling test statistics

Yuan and Johnson [17] developed Bayesian hypothesis tests using nonparametric statistics. In particular, they obtained the BFs based on the sampling distributions of nonparametric statistics, which can be grouped into normal and chi-square distributions, respectively.

For the one- and two-sample testing problems given by (2.1) and (2.4), respectively, [17] adopted the Pitman translation alternative [13] to the alternative hypothesis H_1, leading to the following hypotheses of form

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_0 + \frac{\Delta}{\sqrt{n}}, \tag{3.1}$$

where $\theta_0 = 0$ and Δ is the non-centrality parameter distinguishing the null and alternative hypotheses. Note that the standardized Wilcoxon signed rank test (T^*) and the Mann-Whitney-Wilcoxon test (W^*) have limiting normal distributions under the null and alternative hypotheses, which can be represented as

$$H_0 : \ S^* \sim N\big(0,\ 1\big) \ \text{ and } \ H_1 : \ S^* \sim N\big(c\Delta,\ 1\big), \tag{3.2}$$

where c represents the efficacy of the test S^* (T^* or W^*). Yuan and Johnson [17] specified a normal prior distribution for Δ given by

$$\Delta \sim N(0,\ \kappa^2), \tag{3.3}$$

where κ is a hyperparameter that needs to be prespecified. The Bayes factor in (1.4) under the specified prior can be simplified as

$$\mathrm{BF}_{10} = (1+g)^{-1/2} \exp\left(\frac{T^{*2}}{2} \frac{g}{1+g}\right), \tag{3.4}$$

where $g = (c\kappa)^2$. Yuan and Johnson [17] determined the value of g by finding an upper bound of the Bayes factor in (3.4) over the parameter $g \in (0, \infty)$.

For k-sample testing problem in (2.7), Yuan and Johnson [17] adopted the Pitman translation alternative and obtained the sequence of the local alternatives given by

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0 + \frac{\Delta_i}{\sqrt{n}}, \quad i = 1, \cdots, k, \tag{3.5}$$

where $\theta_0 = 0$, Δ_i is not all equal for $i = 1, \cdots, k$, and $n = \sum_{i=1}^{k} n_i$. The Kruskal-Wallis test U has limiting chi-squared distributions under the null and alternative hypotheses, which are given by

$$H_0 : U \sim \chi^2_{k-1} \ \text{ and } \ H_1 : U \sim \chi^2_{k-1}(\rho).$$

Yuan and Johnson [17] specified a multivariate normal distribution for $\mathbf{\Delta} = (\Delta_1, \cdots, \Delta_k)'$ given by

$$\mathbf{\Delta} \sim N_k\big(\mathbf{0}_k,\ c(\mathbf{R}'\mathbf{R})^{-1}\big),$$

where c is a scaling constant and $\mathbf{R}$ is a nonsingular $k \times k$ matrix satisfying

$$\mathbf{P}'\mathbf{Q}\mathbf{P} = \mathbf{R}' \begin{bmatrix} \mathbf{I}_{k-1} & \mathbf{0} \\ \mathbf{0}' & 0 \end{bmatrix} \mathbf{R}$$

with

$$\mathbf{P} = \mathbf{I}_k - \begin{bmatrix} a_1 & \ldots & a_k \\ \vdots & & \vdots \\ a_1 & \ldots & a_k \end{bmatrix} \ \text{ and } \ \mathbf{Q} = \begin{bmatrix} a_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & a_k \end{bmatrix}.$$

We here refer the interested readers to [17] in detail. The Bayes factor in (1.4) under the specified prior is given by

$$\mathrm{BF}_{10} = (1+g)^{-(k-1)/2} \exp\left\{ \frac{U}{2} \frac{g}{1+g} \right\}, \tag{3.6}$$

where $g = 12c\left(\int f^2(x)\, dx\right)^2$. Again, [17] determined the value of g by finding an upper bound of this Bayes factor over the parameter $g \in (0, \infty)$.

It deserves mentioning that the Bayes factors in (3.4) and (3.6) depend on the data only through the associated nonparametric test statistics and can thus be easily implemented, so long as one has performed nonparametric tests. We observe that they depend on the choice of g, in which [17] determined this value by finding an upper bound of the Bayes factor. Even though the Bayes factors based on this choice of g have been shown to perform well in practical applications, they lack of ability to make the identical decisions between Bayesians and frequentists on the basis of the observed data.

In this report, we adopt an alternative way to determine the value of g by matching the rejection regions of both Bayesian and nonparametric testing procedures; see, for example, [6], [2], [16], among others. One appealing property of fixing the value of g in this manner allows researchers to make the identical decisions between two different testing procedures and interpret the results from both Bayesian and frequentist points of view.

3.2 Restricted most powerful Bayes tests

Goddard and Johnson [2] followed the idea of uniformly most powerful Bayes test [6] and developed a restricted most powerful Bayes test (RMPBT), which has been shown to perform well for testing the regression coefficients in the context of normal linear models. They restricted the class of the alternative hypotheses into the form of Zellner's g-prior [18] and obtained the Bayes factor having the same rejection region as the frequentist F-test, provided that its evidence threshold is determined by the significance level of the F-test. They formally defined a RMPBT for hypothesis testing in linear models as follows.

Definition 1 *Let θ be the parameter of interest. A π-restricted most powerful Bayesian test with its evidence threshold $\delta > 0$ in favor of $H_1 : \theta \sim \pi(\theta \mid \psi_1)$ against a fixed null hypothesis H_0 about θ, is a Bayesian hypothesis test, denoted by π-RMPBT(δ), where the Bayes factor for hypothesis testing satisfies*

$$P_{\theta_t}\big[BF_{10} > \delta\big] \geq P_{\theta_t}\big[BF_{20} > \delta\big],$$

for all possible values of the data generating parameter θ_t and all alternative hypotheses $H_2 : \theta \sim \pi(\theta \mid \psi_2)$, where $\pi(\cdot)$ is a density function parameterized by ψ, and $\psi_1, \psi_2 \in \psi$.

The π-RMPBT(δ) is a Bayes test for which the alternative hypothesis is restricted to a class of priors on θ so as to maximize the probability of rejecting H_0, when the Bayes factor exceeds δ over all possible values of ψ and θ_t. An attractive property of this test is that its rejection region can be coincident with that of the frequentist test, provided that δ is determined by the significance level of the associated test.

Goddard and Johnson [2] developed the Bayes factor based on the RMPBT for testing the regression coefficients in linear models. Later on, [16] developed the Bayes factors based on the RMPBT for testing the presence of correlations between two continuous random variables. We observe that these Bayesian tests are developed through a combined use of the RMPBT and parametric testing procedures. To the best of our knowledge, Bayesian tests based on a combined use of the RMPBT and nonparametric tests have not yet been studied in the literature.

3.3 The developed Bayes factors

In this section, we obtain Bayesian tests by determining the value of g through maximizing the probability that the BF exceeds a specified threshold. This is achieved by letting the rejection regions of Bayesian tests and α-sized nonparametric tests be coincident (see Appendix B in detail). In particular, by integrating the ideas from [17] and [2], we obtain the Bayes factors using three nonparametric test statistics in Section 2, which are summarized in the following theorem with proofs given in the Appendix B.

Theorem 1 *(i) For one-sample or paired-samples problem, the Bayes factor based on the Wilcoxon signed rank test is given by*

$$\mathrm{BF}_{10} = \frac{1}{|\tau^*_{\alpha/2}|} \exp\left(\frac{T^{*2}}{2} \frac{\tau^{*2}_{\alpha/2} - 1}{\tau^{*2}_{\alpha/2}} \right), \tag{3.7}$$

where $\tau^*_{\alpha/2}$ the standardized critical value of $\tau_{\alpha/2}$ given by

$$\tau^*_{\alpha/2} = \frac{\tau_{\alpha/2} - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}. \tag{3.8}$$

The corresponding evidence threshold δ_τ is given by

$$\delta_\tau = \frac{1}{|\tau^*_{\alpha/2}|} \exp\left(\frac{\tau^*_{\alpha/2} - 1}{2} \right). \tag{3.9}$$

(ii) For two independent sample problem, the Bayes factor based on the Mann-Whitney-Wilcoxon test is given by

$$\mathrm{BF}_{10} = \frac{1}{|\omega^*_{\alpha/2}|} \exp\left(\frac{W^{*2}}{2} \frac{\omega^{*2}_{\alpha/2} - 1}{\omega^{*2}_{\alpha/2}} \right), \tag{3.10}$$

where $\omega^*_{\alpha/2}$ is the standardized critical value of $\omega_{\alpha/2}$ given by.

$$\omega^*_{\alpha/2} = \frac{\omega_{\alpha/2} - n_1(n_1+n_2+1)/2}{\sqrt{n_1 n_2 (n_1+n_2+1)/12}}. \tag{3.11}$$

The corresponding evidence threshold δ_ω *is given by*

$$\delta_\omega = \frac{1}{|\omega^*_{\alpha/2}|} \exp\left(\frac{\omega^{*2}_{\alpha/2} - 1}{2}\right). \qquad (3.12)$$

(iii) For k (≥ 3) *independent sample problems, the Bayes factor based on the Kruskal-Wallis test is given by*

$$\mathrm{BF}_{10} = \left(\frac{\nu_\alpha}{k-1}\right)^{-(k-1)/2} \exp\left(\frac{U}{2}\frac{\nu_\alpha - (k-1)}{\nu_\alpha}\right), \qquad (3.13)$$

where ν_α *is the* $1-\alpha$ *quantile of the null distribution defined in Section 2.3. The corresponding evidence threshold* δ_ν *is given by*

$$\delta_v = \left(\frac{\nu_\alpha}{k-1}\right)^{-(k-1)/2} \exp\left(\frac{\nu_\alpha - k + 1}{2}\right). \qquad (3.14)$$

This theorem justifies that there is a close relationship between the Bayesian and frequentist nonparametric methods. In addition, we observe that these Bayes factors with their evidence thresholds depend only on nonparametric statistics with their associated critical values and that they can be easily computed by just adding one more step after one has performed the hypothesis testing using nonparametric statistics. For decision making, the Bayes factor greater than its corresponding evidence threshold indicates that H_0 is rejected and its value smaller than the evidence threshold indicates that we fail to reject H_0.

This theorem also shows that the Bayesian and frequentist nonparametric testing procedures can result in an identical decision when we match their rejection regions. This property allows researchers to simultaneously report the conclusions from both Bayesian and frequentist points of view, and more importantly they can quantify evidence in favor of both H_0 and H_1 in terms of the Bayes factors and the posterior probability in (1.3).

We observe that like the nonparametric test, the evidence threshold of the Bayesian approach depends on the specified significance level α. As an illustration, we consider the evidence threshold δ_τ in (3.9), since other evidence thresholds behave similarly. It can be seen from Figure 3.1 that δ is a decreasing convex function of α and that we need to choose δ_τ to be larger than 1 to achieve its agreement with the nonparametric test at the α (≤ 0.10) level of significance.

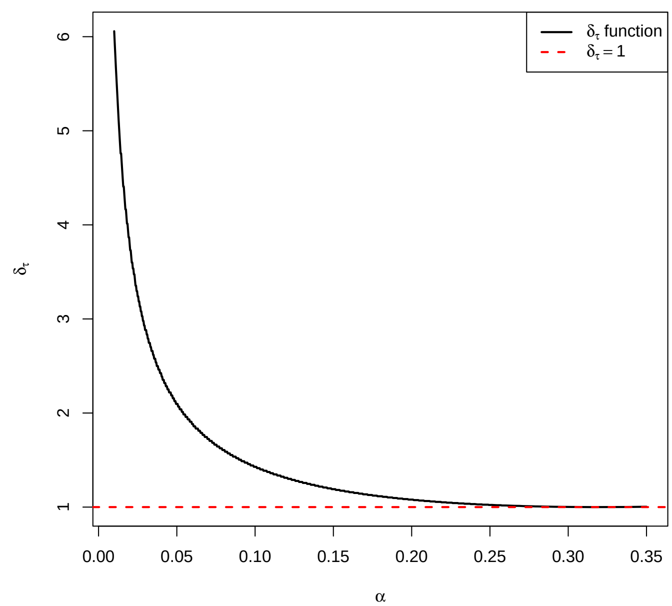

Figure 3.1: The relationship between the evidence threshold δ_τ and the significance level α when $n = 50$.

Chapter 4

Numerical studies

In this chapter, we examine the performance of both Bayesian and frequentist nonparametric methods using simulated data in Section 4.1 and three real-data examples in Section 4.2.

4.1 Simulation studies

We employ simulated data to assess the agreement between the Bayesian and frequentist nonparametric methods. For illustrative purposes, we here only illustrate the agreement between the Bayes factor in (3.7) and the Wilcoxon signed rank test in (2.2), since similar conclusions can be achieved for other two Bayes factors in Theorem 1 and are thus omitted for simplicity.

First, n random variables are generated from the normal distribution with mean μ and standard deviation $\sigma = 1$, where μ ranges from -4 to 4 in increments of 0.01. For each value of μ, we generate 10,000 simulated datasets with $n = 10$ (small) and $n = 100$ (moderate), respectively. The decision criterion is that we select H_1 if the Bayes factor is larger than its evidence threshold δ, and H_0, otherwise. We consider two different choices of δ: (i) $\delta = 1$ from [7] and (ii) $\delta = \delta_\tau$ determined by (3.9), which can control the Type I error at a specified significant level, say $\alpha = 5\%$. The relative frequencies of rejecting H_0 are depicted in Figure 4.1 for two different choices of δ .

Rather than providing exhaustive conclusions from Figure 4.1, we only highlight some most important findings. (i) Like its nonparametric counterpart, the Bayes factor in (3.7) with δ_τ in (3.9) is able to control the Type I error for the given value of α. For instance, when $n = 10$, the frequency of rejecting H_0 is 0.05 when we choose $\alpha = 0.05$, leading to $\delta_\tau = 2.529$; (ii) as the sample size increases, the Type I error rate of the Bayes factor in (3.7) still remains a constant, mainly because we fix the Type I error of the test to be $\alpha = 5\%$, and (iii) when the sample size is large, the Bayes factor with δ_τ in (3.9) behave similarly to the one with $\delta_\tau = 1$. This behavior occurs the value of δ decreases to its limit 1 shown in Figure 3.1, as the sample size increases.

4.2 Real data applications

We here apply the developed Bayes factors in Theorem 1 to three real-data examples. The first is about the paired-sample problem, the second about two independent samples, and the third about three or more independent samples.

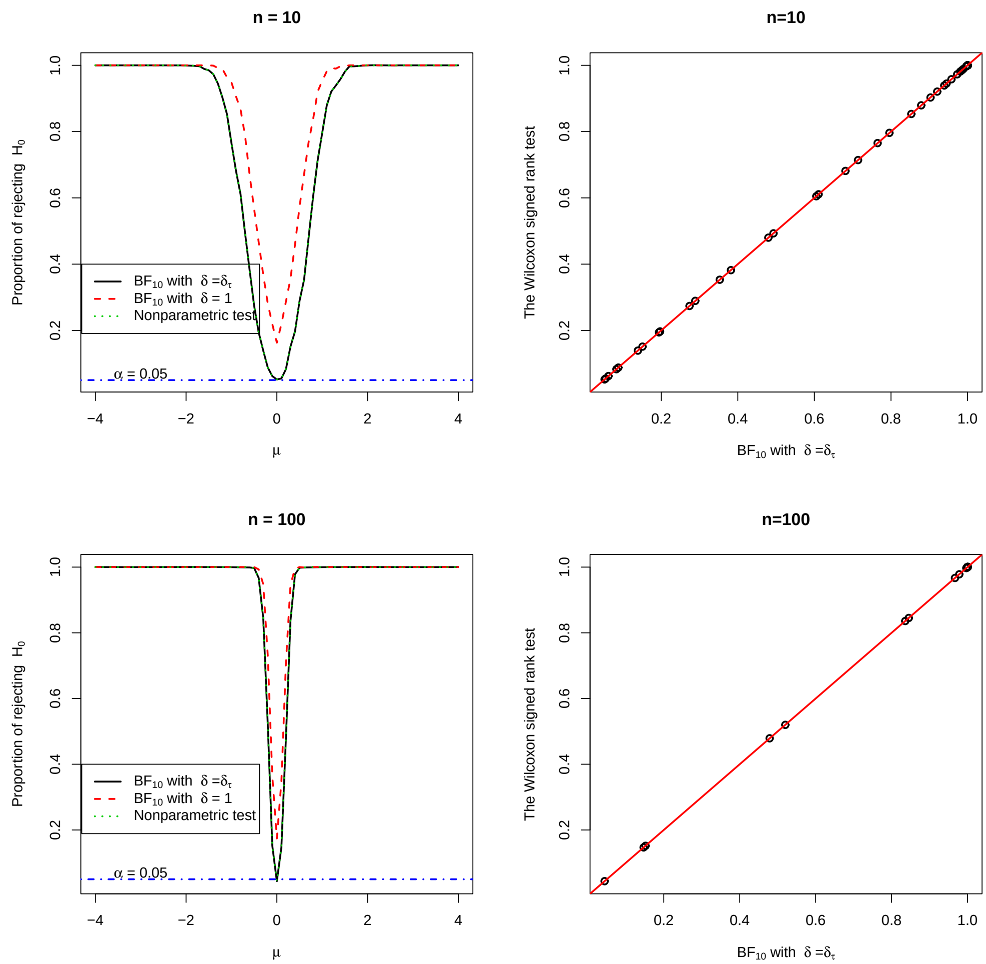

Figure 4.1: The relative frequency of rejection of H_0 under the Bayes factors with two different choices of δ and the Wilcoxon signed rank test (left figures); the relationship between the Bayesian and nonparametric tests (right figures) when $\alpha = 0.05$.

Example 1 The depression data set from [10] is available from the R $ACSWR$ package, created by [14]. The purpose of this data is to investigate changes to Hamilton depression scale Factor IV measurements. The data consist of nine patients with anxiety or depression before and after tranquilizer therapy. We are interested in testing if there exists a treatment effect for reducing symptoms of depression. The Wilcoxon signed rank test is $T = 5$ with the two-sided p-value of 0.03906, and thus we reject the null hypothesis of no treatment effect at a $\alpha = 5\%$ significance level.

We calculate the Bayes factor BF_{10} in (3.7) and the evidence threshold δ_τ in (3.9) corresponding to the Wilcoxon signed rank test. By using equation in (3.8) with $n = 9$ and $T = 5$, we have $T^* = -2.073221$. Thus, by plugging $T^* = -2.073221$ and $\tau^*_{\alpha/2} = -1.954751$ into two functions, we obtain that $\text{BF}_{10} = 2.500316$ and $\delta_\tau = 2.096488$, indicating that the data are about 2.50 times more likely to be generated under H_1 than under H_0 and H_0 should be rejected, since $\text{BF}_{10} > \delta_\tau$. In addition, by using equation in (1.3) with the assumption of equipoise, the posterior probability of H_1 is 0.7143115, or equivalently, the posterior probability of H_0 is $1 - 0.7143115 = 0.2856885$.

Example 2 The blood pressure data consist of the blood pressure measurements for 21 African-American men: ten of the men took calcium supplements and 11 took placebos. The data can be found via the link *http://lib.stat.cmu.edu/DASL/Datafiles/Calcium.html*. In this study, the researchers are interested in testing if blood pressure can be reduced by increasing calcium intake. The Mann-Whitney-Wilcoxon test is $W = 124.5$ with the two-sided p-value of 0.3228, and thus, we fail to reject H_0 of no treatment effect at the $\alpha = 5\%$ significance level.

By plugging $W^* = 1.021059$ and $\omega^*_{\alpha/2} = -1.971701$ into the Bayes factor BF_{10} in (3.10) and the evidence threshold δ_ω in (3.12), we obtain that $\text{BF}_{10} = 0.74699$ and $\delta_\omega = 2.148791$, indicating that the data are about 0.75 times more likely to be generated under H_1 than under H_0 and that we fail to reject H_0 since $\text{BF}_{10} < \delta_\omega$, corresponding to the 5% Mann-Whitney-Wilcoxon test. Under the assumption of equipoise, the posterior probability of H_1 is 0.42759, or equivalently, the posterior

Method			
1	2	3	4
83	91	101	78
91	90	100	82
94	81	91	81
89	83	93	77
89	84	96	79
96	83	95	81
91	88	94	80
92	91		81
90	89		
	84		

Table 4.1
The growing corn data set from Example 4.3.1 of [1].

probability of H_0 is $1 - 0.42759 = 0.5724131$.

Example 3 The growing corn data set from [1] is to investigate whether there exists a difference in yields per acre among four different methods of growing corn. The data is given in Table 4.1. The value of the Kruskal-Wallis test in (2.8) for this dataset is $U = 25.464$ with the two-sided p-value of $1.141e - 05$, which clearly leads to the rejection of the null hypothesis at the 5% significance level. We may thus conclude that some methods of growing corn tend to furnish higher yields than others.

By plugging $U = 25.62884$, $\nu_\alpha = 7.548731$ from [11] into the Bayes factor in (3.13) and the evidence threshold δ_ν in (3.12), we obtain that $\text{BF}_{10} = 565.4317$ and $\delta_\nu = 2.435653$, indicating that the data are about 565 times more likely to be generated under H_1 than under H_0 and that we choose H_1 since $\text{BF}_{10} > \delta_\nu$, corresponding to the 5% Kruskal-Wallis test. Under the assumption of equipoise, the posterior probability of H_1 is 0.9982346, and the posterior probability of H_0 is $1 - 0.9981448 = 0.0017654$.

Chapter 5

Concluding remarks

Based on a combined use of the testing procedures from [17] and [2], we obtained the Bayes factor tests for comparing the difference between the means among two or more populations. The proposed Bayes factors will not only have closed-form expressions in terms of the associated nonparametric statistics and their associated critical values under the null hypothesis, but also justify that there exists a close relationship between the Bayesian and frequentist nonparametric testing procedures. From a practical point of view, they can be easily calculated by one step further, so long as one has performed the corresponding nonparametric tests for the testing problem at hand. In addition, like the nonparametric counterparts, they are able to control the Type I error and also allow researchers to make the identical decisions between frequentists and Bayesians. More importantly, they can quantify evidence from empirical data in favor of H_0, and this property is not shared by the frequentist counterparts, which lack this ability when we fail to reject H_0.

It is noteworthy that this report mainly focuses on the agreement between Bayesian and nonparametric testing procedures for the location parameters. Given that the Pearson correlation coefficient is a commonly used criterion to measure the strength of a linear relationship between two quantitative variables, [16] recently studied the agreement between Bayesian and freqeuent t-test procedures for the presence of correlations and partial correlations. In an ongoing project, we study the relationship between Bayesian and nonparametric testing (e.g., Kendall's τ) procedures for testing the dependence of two variables, which are currently under investigation and will be reported elsewhere.

References

[1] Conover, W. J. (1999). *Practical Nonparametric Statistics*. Wiley Series in Probability & Mathematical Statistics, 3rd edition.

[2] Goddard, S. D. and Johnson, V. E. (2016). Restricted most powerful Bayesian tests for linear models. *Scandinavian Journal of Statistics* **43**, 1162–1177.

[3] Hollander, M. and Wolfe, D. A. (1999). *Nonparametric Statistical Methods*. Wiley Series in Probability and Statistics: Texts and References Section. John Wiley & Sons, Inc., New York, 2nd ed. A Wiley-Interscience Publication.

[4] Iman, R. L., Quade, D. and Alexander, D. A. (1975). *Exact probability levels for the Kruskal-Wallis test* **III**, 329–384.

[5] Jeffreys, H. (1961). *Theory of Probability*. Statistics and Computing, 3rd edn. London: Oxford University Press.

[6] Johnson, V. E. (2013). Uniformly most powerful Bayesian tests. *The Annals of Statistics* **41**, 1716–1741.

[7] Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

[8] Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* **47**, 583–621.

[9] Li, Y. and Yu, J. (2012). Bayesian hypothesis testing in latent variable models. *Journal of Econometrics* **166**, 237–246.

[10] MacLean, K. A., Ferrer, E., Aichele, S. R., Bridwell, D. A., Zanesco, A. P., Jacobs, T. L., King, B. G., Rosenberg, E. L., Sahdra, B. K., Shaver, P. R., Wallace, B. A., Mangun, G. R. and Saron, C. D. (2010). Intensive meditation training improves perceptual discrimination and sustained attention. *Psychological Science* **21**, 829–839.

[11] Meyer, J. P. and Seaman, M. A. (2013). A comparison of the exact Kruskal-Wallis distribution to asymptotic approximations for all sample sizes up to 105. *The Journal of Experimental Education* **81**, 139–156.

[12] R Core Team (2014). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

[13] Randles, R. H. and Wolfe, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics.* John Wiley & Sons, New York-Chichester-Brisbane. Wiley Series in Probability and Mathematical Statistics.

[14] Tattar, P. (2015). *ACSWR: A Companion Package for the Book "A Course in Statistics with R".* R package version 1.0.

[15] Trafimow, D. and Marks, M. (2017). Editorial. *Basic and Applied Social Psychology* **37**, 1–2.

[16] Wang, M. and Dong, J. (2017). The Bayesian t-tests for correlations and partial correlations. *Submitted for publication*.

[17] Yuan, Y. and Johnson, V. E. (2008). Bayesian hypothesis tests using nonparametric statistics. *Statistica Sinica* **18**, 1185–1200.

[18] Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian inference and decision techniques*, vol. 6 of *Stud. Bayesian Econometrics Statist.* Amsterdam: North-Holland, pp. 233–243.

Appendix A

Sample Code

A.1 SimulationCode.R

```
# R codes for Figure 4.1 when n = 10
n = 10
alpha = 0.05
iter = 10000
mu = seq(-4, 4, by = 0.1)
Rep = length(mu)
Avg_rej = matrix(0, ncol = 3, nrow = Rep)
result = matrix(0, ncol = 3, nrow = iter)
for (j in 1:Rep) {
    for (i in 1:iter) {
        x = rnorm(n, mu[j], 1)
        test = wilcox.test(x, exact = T)
        # calculate T^ast
```

```
        Tast = (test$statistic - n * (n + 1)/4)/sqrt(n *↩
            (n + 1) * (2 *
            n + 1)/24)

        # calculate tau
        tau_1_orig = qsignrank(alpha/2, n, lower.tail = ↩
           TRUE, log.p = FALSE)
        tau_2_orig = qsignrank(1 - alpha/2, n, lower.↩
           tail = TRUE, log.p = FALSE)
        tau_orig = tau_1_orig * (Tast <= 0) + tau_2_orig↩
            * (Tast > 0)
        tau = (tau_orig - n * (n + 1)/4)/sqrt(n * (n + ↩
           1) * (2 * n + 1)/24)
        # The proposed Bayes factor
        BF = 1/abs(tau) * exp(Tast^2/2 * (tau^2 - 1)/tau↩
           ^2)
        delta = 1/abs(tau) * exp((tau^2 - 1)/2)
        # The previous Bayes factor
        BF_tilde = 1/abs(Tast) * exp(-(1 - Tast^2)/2)
        result[i, ] = c(1 * (BF > delta), 1 * (BF > 1), ↩
           1 * (test$p.value <
            0.05))
    }
    Avg_rej[j, ] = c(colMeans(result))
}

# Upper left of Figure 1
plot(mu, Avg_rej[, 1], lwd = 2, xlab = expression(mu), ↩
   ylab = expression(paste("Proportion of rejecting  ",
    H[0])), type = "l", main = "n = 10")
lines(mu, Avg_rej[, 2], lwd = 2, col = 2, lty = 2)
lines(mu, Avg_rej[, 3], lwd = 2, col = 3, lty = 3)
abline(h = 0.05, lty = 4, col = 4, lwd = 2)
```

```
legend(-4.3, 0.4, c(expression(paste(BF[10], " with  ", ↩
   delta, " =", delta[tau])),
     expression(paste(BF[10], " with  ", delta, " = 1")),↩
         "Nonparametric test"),
     col = 1:3, lwd = 2, lty = 1:3)
text(-3, 0.07, label = expression(paste(alpha, " = 0.05"↩
   )))

# Upper right of Figure 1
plot(Avg_rej[, 1], Avg_rej[, 3], lwd = 2, xlab = ↩
   expression(paste(BF[10],
     " with  ", delta, " =", delta[tau])), ylab = "The ↩
        Wilcoxon signed rank test",
     main = "n=10")
abline(c(0, 1), lwd = 2, col = 2)
```

A.2 Example1Code.R

```
# Example 1
# Attache the depression data
library(ACSWR)
attach(depression)
data(depression)

X = depression$X
Y = depression$Y
n = length(X)

# The Wilcoxon signed rank test
(test = wilcox.test(Y, X, paired = TRUE, exact = T))
```

```
T_ast = (test$statistic - n * (n + 1)/4)/sqrt(n * (n + ↩
   1) * (2 * n + 1)/24)
alpha = 0.05
tau = qsignrank(alpha/2, n, lower.tail = TRUE, log.p = ↩
   FALSE)
tau_ast = (tau - n * (n + 1)/4)/sqrt(n * (n + 1) * (2 * ↩
   n + 1)/24)

# Calculate BF and delta_tau
BF = 1/abs(tau_ast) * exp(T_ast^2/2 * (tau_ast^2 - 1)/↩
   tau_ast^2)
delta = 1/abs(tau_ast) * exp((tau_ast^2 - 1)/2)
list(BF = BF, delta = delta)
```

A.3 Example2Code.R

```
# Example 2
x = c(7, -4, 18, 17, -3, -5, 1, 10, 11, -2)
y = c(-1, 12, -1, -3, 3, -5, 5, 2, -11, -1, -3)

n1 = length(x)
n2 = length(y)

# The Mann-Whitney-Wilcoxon test
W = sum((rank(c(x, y))[1:n1]))

# calculate T^ast
```

```
W_ast = (W - n1 * (n2 + n1 + 1)/2)/sqrt(n2 * n1 * (n2 + ↩
   n1 + 1)/12)
# The value of 82 was from Table A12 of Conover (1999)
omega_ast = (82 - n1 * (n2 + n1 + 1)/2)/sqrt(n2 * n1 * (↩
   n2 + n1 + 1)/12)

# Calculate BF and delta_omega
BF = 1/abs(omega_ast) * exp(W_ast^2/2 * (omega_ast^2 - ↩
   1)/omega_ast^2)
delta = 1/abs(omega_ast) * exp((omega_ast^2 - 1)/2)
list(BF = BF, delta = delta)
```

A.4 Example3Code.R

```
# Example 3
m1 = c(83, 91, 94, 89, 89, 96, 91, 92, 90)
m2 = c(91, 90, 81, 83, 84, 83, 88, 91, 89, 84)
m3 = c(101, 100, 91, 93, 96, 95, 94)
m4 = c(78, 82, 81, 77, 79, 81, 80, 81)

# The Kruskal-Wallis test
type <- c(rep(1, times = 9), rep(2, times = 10), rep(3, ↩
   times = 7), rep(4,
    times = 8))
gross = c(m1, m2, m3, m4)
test = kruskal.test(gross ~ type)
U = test$statistic

k = 4
nu_alpha = 7.548731  #from Meyer and Seaman (2013)
```

```
# we can use chi-square approximation alpha = 0.05 nu_↩
   alpha =
# qchisq(1-alpha, df=k-1) Calculate BF and nu_tau
BF = exp(-(k - 1)/2 * log(nu_alpha/(k - 1)) + U/2 * (nu_↩
   alpha - k + 1)/nu_alpha)
delta = exp(-(k - 1)/2 * log(nu_alpha/(k - 1)) + (nu_↩
   alpha - k + 1)/2)
list(BF = BF, delta = delta)
```

Appendix B

Appendix B: Deviations of $\mathbf{BF}_{10}$ and $\delta_{\mathcal{T}}$ given by (3.7) and (3.9)

In this Appendix, we only provide the proof for part (i), since the proofs for others are exactly the same and are thus omitted here for simplicity. It can be seen from [17] that under the limiting distributions of T in (3.2) and the proposed prior in (3.3), the Bayes factor for comparing two competing models in (3.1) is given by

$$\mathrm{BF}_{10} = (1+g)^{-1/2} \exp\left\{ \frac{T^{*2}}{2} \frac{g}{1+g} \right\}. \tag{B.1}$$

Simple algebra shows that the probability of the Bayes factor in (B.1) exceeding the evidence threshold δ can be rewritten as

$$P_\theta\big(\mathrm{BF}_{10} > \delta_\tau\big) = P_\theta\bigg(T^{*2} > 2\frac{1+g}{g}\log\big[\delta_\tau(1+g)^{1/2}\big]\bigg).$$

The maximum of this probability can be achieved by minimizing the term $\frac{1+g}{g}\log\big[\delta_\tau(1+g)^{1/2}\big]$ with respect to $g \in (0,\ \infty)$. By taking derivative of this term with respect to g and then setting it equal to 0, we obtain that

$$\delta_\tau = (1+g)^{-1/2}\exp(g/2). \tag{B.2}$$

In order to match the rejection regions between Bayesian and frequentist nonparametric testing procedures, we reexpress the inequality of $\mathrm{BF}_{10} > \delta_\tau$ as

$$\begin{aligned} &(1+g)^{-1/2}\exp\bigg\{\frac{T^{*2}}{2}\frac{g}{1+g}\bigg\} > \delta_\tau \\ \Rightarrow\ &T^{*2} > 2\frac{1+g}{g}\log\big[\delta_\tau(1+g)^{1/2}\big]. \end{aligned}$$

This shows that by using the Bayes factor for decision making, we reject the null hypothesis if $T^* > \kappa$ or $T^* < -\kappa$, where

$$\kappa = \bigg\{2\frac{1+g}{g}\log\big[\delta(1+g)^{1/2}\big]\bigg\}^{1/2}.$$

Also, the rejection region of the standardized nonparametric statistic T^* is defined as

$$\big\{Y : T^* < -|\tau^*_{\alpha/2}| \quad \text{or} \quad T^* > |\tau^*_{\alpha/2}|\big\},$$

where $\tau^*_{\alpha/2}$ is defined in Section 2.1. Note that the Bayesian and frequentist tests can make an identical decision if we match their rejection regions by

$$\tau^*_{\alpha/2} = -\kappa \quad \text{or} \quad \frac{n(n+1)}{2} - \tau^*_{\alpha/2} = \kappa. \tag{B.3}$$

By solving two equations in (B.2) and (B.3) with respect to δ and g, we obtain that $\hat{g} = \tau^{*2}_{\alpha/2} - 1$. We also obtain that $\delta_\tau = |\tau^*_{\alpha/2}|^{-1} \exp\big[(\tau^{*2}_{\alpha/2} - 1)/2\big]$. By replacing g in (B.1) with $\tau^2_{\alpha/2} - 1$, we have

$$\mathrm{BF}_{10} = \frac{1}{|\tau_{\alpha/2}|} \exp\left\{ \frac{T^{*2}}{2} \frac{\tau^2_{\alpha/2} - 1}{\tau^2_{\alpha/2}} \right\}.$$

This completed the proof of Theorem 1 (i).