



UNICA

UNIVERSITÀ
DEGLI STUDI
DI CAGLIARI

Ph.D. DEGREE IN
MATHEMATICS AND COMPUTER SCIENCE
Cycle XXXV

TITLE OF THE Ph.D. THESIS

A new evidence measure for Bayesian Inference

Scientific Disciplinary Sector(s)

SECS-S/01

Ph.D. Student: Mara Manca

Supervisor Prof. Monica Musio

Final exam. Academic Year 2021/2022
Thesis defence: April 2023 Session



REGIONE AUTONOMA DELLA SARDEGNA



Mara Manca gratefully acknowledges the Sardinian Regional Government for the financial support of her PhD scholarship (P.O.R. Sardegna F.S.E. - Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2014-2020 - Axis III Education and training, Thematic goal 10, Investment Priority 10ii), Specific goal 10.5”.

Abstract

In the last sixty years there have been several attempts to build a measure of evidence that covers, in a Bayesian context, the role that the p -value has played in the frequentist setting. A prominent example is the decision test based on the Bayes Factor. Worth to mention it is also the e -value, another Bayesian evidence measure on which the Full Bayesian Significance Test procedure is based.

The aim of this thesis is to make a contribution to the Bayesian testing procedure of precise hypotheses for parametric models. To this end we propose a Bayesian measure of evidence, called Bayesian Discrepancy Measure, which gives an absolute evaluation of the suitability of a hypothesis H in light of the prior knowledge about the parameter and the observed data. The starting point is the idea that a hypothesis may be more or less supported by the available evidence contained in the posterior distribution. Since reference is made to a precise hypothesis H and no alternative against this hypothesis is considered, we do not adopt the approach whereby there is no test that can lead to the rejection of a hypothesis except by comparing it with an alternative one (the Bayes factor in the Bayesian perspective and Neyman-Pearson-Wald in the frequentist one). The proposed measure of evidence has the desired properties of invariance under reparametrisations and consistency for large samples.

In this thesis we also show that it is possible to construct a testing procedure, based on the Bayesian Discrepancy Measure, that allows to compare parameter functions from two independent populations. This approach is flexible, as it can be adapted to take into account different distributions and different parameter transformations. Moreover, we address the problem of comparing k parameters, or their transformations, from independent populations. This is not a simple extension of the procedure for two populations, due to the geometry of the hypothesis and the parameter space, and it therefore demands a separate discussion. In conclusion, this methodologies enables us to tackle some problems that are not yet covered in the literature.

Contents

1	Introduction	1
1.1	A bit of history	1
1.1.1	The Bayesian proposals	4
1.2	Some remarks	6
1.3	Outline of the thesis	8
2	A new Bayesian Evidence Measure	11
2.1	The Bayesian Discrepancy Measure for a scalar parameter	12
2.2	The Bayesian Discrepancy Measure in presence of nuisance parameters	16
2.3	Illustrative examples	17
2.3.1	Examples of the univariate parameter case	17
2.3.2	Examples of the more general case	20
2.4	Comparison with the FBST	29
2.4.1	Similarities and differences between the procedures	29
3	The Bayesian Discrepancy Test for the comparison of two independent pop- ulations	37
3.1	The procedure	38
3.2	Illustrative examples	39
3.2.1	Skewness coefficients of two Inverse Gaussian populations	39
3.2.2	Means of two Inverse Gaussian populations	41
3.2.3	Variances of two Gamma populations	42
3.2.4	Coefficients of variation of two independent populations	43
3.2.5	Correlation coefficients of two Normal populations	50
3.2.6	Regression coefficients	52
4	Comparing k independent populations through the Bayesian Discrepancy Measure	57
4.1	Problem setting	57
4.1.1	Prior, posterior and marginals	58
4.2	Comparison of k independent populations	59

4.2.1	Distribution of a fixed contrast	60
4.2.2	The Optimality Criterion	61
4.3	Examples	63
5	The Bayesian Discrepancy Test for Partial Correlations	67
5.1	Problem setting	67
5.1.1	The case of one conditioning variable	68
5.1.2	The case of more conditioning variables	72
5.2	The BDT for the Partial Correlation	74
5.2.1	Illustrative examples	75
6	Conclusions & Discussion	79
A	Short references to some statistical tests	83
A.1	Measures of surprise	83
A.1.1	Surprise indices	84
A.1.2	Bayesian <i>p-value</i> proposals	90
A.2	Bayesian decision-making tests: the Bayes factor	95
A.2.1	Bayes factor and odds	95
A.2.2	The Jeffreys-Lindley's paradox	99
A.3	Full Bayesian Significance Test	102
A.3.1	FBST definition	102
A.3.2	FBST invariance	103
	Bibliography	107

Chapter 1

Introduction

This thesis deals with the description and discussion of a new parametric test referred to as Bayesian Discrepancy Test (BDT). Before giving a detailed description of the test, and its development potential, it is useful to briefly review the history of the testing procedures in order to highlight similarities and differences to the existing tests and thus to determine its proper collocation in the literature.

In this regard it is appropriate to anticipate that the BDT is a test having a evaluative nature, in the sense that it measures the conformity of the parametric hypothesis H of interest by means of an appropriate evidence measure, called Bayesian Discrepancy Measure (BDM), which does not fix or consider any alternative hypotheses to H .

In some way, with the BDT there is a return to the classical idea of testing, which goes back to the pioneering work of K. Pearson, Yule, Gosset (“*Student*”), etc., and to the fundamental contribution of Fisher who outlined the theory of pure significance.

1.1 A bit of history

In 1895, the journal *Biometrika* published K. Pearson’s work on what would become known as the *Chi-Square goodness of fit test* and the *Chi-Square Test of Independence* between two variables expressed in a contingency table. Shortly afterwards, in the period 1907-1911, the same journal accepted some articles by Gosset concerning the heterogeneity or bias test (spatial poissonianity test) and the paired “*Student*” test, a test for comparing the averages of measurements carried out, under two different conditions, on

the same statistical units. It is known that modern statistical test theory has developed from a frequentist perspective.

It was soon realised that the adopted procedures (and especially their logic) could be extended to a much broader range of situations and models. The most capable and ready to grasp its potential was Fisher who, in 1921, succeeded in giving an organic theoretical framework to what is remembered as the *theory of pure significance*. Fisher himself, more than anyone else, proposed and developed an impressive number of tests.

K. Pearson, Gosset and Fisher adopted a measure of evidence called *p-value*, which is defined as the probability of obtaining a result at least as extreme as the one obtained, if the null hypothesis H_0 and all other premises of the model were indeed valid. According to Fisher, if H_0 is true it is unlikely to obtain a low *p-value*. If this happens, either we are in the presence of a rare event or H_0 is false. About an experiment that had produced a very small *p-value*, Fisher's comment (see Fisher, 1956 p. 39) was

“The force with which such a conclusion is supported is logically that of the simple disjunction: *Either* an exceptionally rare chance has occurred, *or* the [...] hypothesis H_0 [...] is not true”.

The expression “rare chance” refers to the (inevitably subjective) concept of the *p-value* threshold. While arguing that the compatibility of observations with respect to other hypotheses should not be excluded, Fisher always claimed that

“it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis”

see Fisher, 1935 p. 16. Thus he argued that the *p-value* was the absolute measure of compatibility, i.e. the “closeness” of the observations with respect to the hypothesis H_0 . Specifically, the *test* is the instrument *pour excellence* for falsifying theories by means of experiments.

The incredible success of the significance test among the scientific community is given by

- (a) theoretical and analytical developments mainly due to Fisher himself;
- (b) the relative simplicity of the *test* application due, certainly, to the possibility of providing it with a paradigmatic form;

- (c) the divulgative skills of the authors of the *test*. Fisher himself was the author of highly appreciated manuals oriented towards researchers;
- (d) the need of researchers to limit the time and cost of research, with the possibility of access to procedures valid for *small samples*;
- (e) the printing of statistical tables, by K. Pearson, Fisher, Yates, Sheppard, F. N. David, Snedecor, etc.

Still in the frequentist field and oppositely to the pure significance theory, Neyman and E. S. Pearson, in subsequent articles (1928-1932), set the foundations of what they called *hypothesis-testing theory*. It should be noted that this proposal started out as an attempt to develop the theory of pure significance and that, even today, Neyman-Pearson's theory adopts notions and terms that originated with the significance test.

After opposing the null hypothesis H_0 with the alternative hypothesis H_1 , Neyman and E. S. Pearson conceived the hypothesis test as a real decision between H_0 and H_1 . It was Neyman himself, in later articles, who named his theory as the theory of *inductive behaviour*.

The test requires to fix *pre-experimentally* the probability α of rejecting H_0 when it is true (*error of the first kind*) and to determine, as a function of α and on the basis of an appropriate optimality criterion, the partition $\{\mathcal{A}_{H_0}^\alpha, \mathcal{R}_{H_0}^\alpha\}$ of the sample space, whose elements are called acceptance and rejection regions of H_0 . If the observations fall in $\mathcal{A}_{H_0}^\alpha$ [in $\mathcal{R}_{H_0}^\alpha$] one accepts [one rejects] H_0 . The above criterion ensures that the probability of accepting H_0 when H_0 is false (*error of the second kind*) is minimal.

The term *pre-experimental*, associated with the hypothesis test, indicates that the regions $\mathcal{A}_{H_0}^\alpha$ and $\mathcal{R}_{H_0}^\alpha$ are established before the experiment is carried out and that, only after the experiment is concluded, the test decides which hypothesis to assume. On the contrary, note that the significance test is *post-experimental*, in the sense that the *p-value* can only be calculated once the experiment is concluded. Note again that, unlike all other tests in the literature (whether frequentist or Bayesian), the hypothesis test does not include any measure of evidence.

A few years later, Neyman and E. S. Pearson were joined by Wald who, by making hypothesis testing part of the more general decision theory under uncertainty, caused a further departure from Fisher's approach.

According to the pioneers of the theory of statistical testing, K. Pearson, Gosset, Yule and Fisher, a *test* had to consider only the hypothesis H_0 , without excluding the compatibility of the observations with other unspecified hypotheses, and without thereby leading to the specification of an alternative hypothesis H_1 . This fact was, in Fisher's opinion, "the serious nonsense" committed by Neyman and E. S. Pearson. Later on, Fisher modified the *theory of pure significance* proposing the *theory of significance* in which the alternative hypothesis H_1 was also included.

The modifications, in practice, did not change the methodological structure of the test in which the distribution of the *test* statistics, subordinate to H_0 , always plays an essential part. This hypothesis now plays the role of a privileged hypothesis, while H_1 intervenes only at the moment of the calculation of the *p-value*. In this way, the test of significance is not given the accentuated *comparative* character that is proper to hypothesis-tests and, above all, is not led to deny the post-experimental and valutative characters of the test. In conclusion, it is worth mentioning that after Fisher's death in 1962, the significance test was also known as *Null Hypothesis Significance Test* (NHST).

1.1.1 The Bayesian proposals

It took more than forty years to see the first attempts to construct a Bayesian test, a test that required the elicitation of prior distributions (or that adopted default prior distributions), that accepted the likelihood principle and that, in short, was fully based on the Bayesian paradigm. In this regard, see the following comment made by Lindley, 1965 (preface p. xi)

"...hypothesis testing looms large in standard statistical practice, yet scarcely appears as such in the Bayesian literature."

Among the Bayesian statisticians who dedicated themselves to this task, it should be mentioned

- Lindley, 1965 who worked on "translating", from a Bayesian perspective, some of the more usual frequentist tests. He also developed the idea that under certain conditions (for example, assuming a suitable non-informative distribution as a prior distribution, etc.) frequentist and Bayesian tests provide the same inferential conclusions.

- Guttman, 1970, Zellner, 1971, Box and Tiao, 1973, who redeveloped numerous multivariate statistics tests from a Bayesian perspective and worked on the development of predictive tests, see the summary due to Geisser, 1993.
- Box, 1980, revisiting Fisher's *p-value* idea in a Bayesian key, proposed a method for analyzing the conformity of the data to a given model through a new Bayesian evidence measure called *prior predictive p-value*. Guttman, 1967 and Rubin, 1984, independently from each other, attempted to eliminate some of the problems by presenting a new measure, *the posterior predictive p-value*. This was followed by some new proposals from Bayarri and Berger, 2000. One of them is called *conditional predictive p-value*.
- Evans, 1997 who proposed a test that has a Bayesian version of the *surprise index* as its measure of evidence. He relied on the work done by Weaver, 1948, 1963, that brought the first frequentist surprise index definition, and Good, 1950 which gave a second frequentist reformulation.
- J. O. Berger, 1985; J. O. Berger and Delampady, 1987; J. O. Berger and Selke, 1987 used the Bayes factor, a measure of evidence introduced by Jeffreys, 1939 to develop a Bayesian test comparing two hypotheses H_0 and H_1 . O'Hagan, 1995 with the *fractional Bayes factor* and J. O. Berger and Pericchi, 1996 with the *intrinsic Bayes factor* gave additional decisive contributions to the problem of comparing nested models.
- Pereira and Stern, 1999 presented the Full Bayesian Evidence Test (FBST). It is based on a Bayesian measure of evidence for precise hypotheses which aims to provide a Bayesian alternative to significance tests. Their proposal involved the definition of a subset of the parameter space called *Highest Posterior Density Set* (HPDS) which, as stated by the authors, is "tangent" to the set defining the null hypothesis H_0 . The measure of evidence in favour of H_0 is the complement of the posterior probability, i.e. its credibility, in the "tangent" region. Notice that this first formulation received some criticism concerning the non-invariance of the measure with respect to alternative parameterizations of the parameter space. Madruga et al., 2003, inspired by Good's concept of relative surprise, proceeded to the introduction of a reference function to overcome this obstacle. In this way,

the evidence measure is no longer based on the integration over the HPDS, but it considers a new integration region called *Highest Relative Surprise Set* (HRSS). For a full exposition of the FBST see Pereira and Stern, 2020.

The Bayesian tests for comparing parameters of independent populations, which are endowed with a certain degree of generality and meet with a fair degree of success in the world of research, are essentially these last two test in the list.

In some of the aforementioned attempts to construct Bayesian tests it is not uncommon to encounter ideas and insights of interest. This is the reason why the Surprise index, the Bayesian *p-value* proposals, the test based on the Bayes factor and the FBST are given in Appendix A in a slightly more analytical form. For anyone wishing to develop a Bayesian test an in-depth critical study of the history of the ideas that gave rise to and developed Bayesian tests is advisable.

1.2 Some remarks

It is well known that, in the research world, the significance test and the hypothesis test are by far the most popular tests. Moreover, it is a fact that prejudices and psychological difficulties connected to the elicitation of prior distributions, or even to the adoption of non-informative a priors, have been and still are a serious obstacle to the affirmation of Bayesian procedures and to the development of Bayesian tests. This circumstance is the real responsible for the use of frequentist tests by the Bayesian statisticians themselves when working with biologists, doctors, psychologists, economists, who only know “classical” tools.

Over the past two decades, especially in the psychology field, following repeated announcements of “new discoveries” or even “significant experimental results” that demolished established theories, succeeded by quick turnarounds, many attempts have been made to address the distortions associated with the often clumsy use of the *p-value*. A warning, addressed to researchers with limited statistical expertise, and not only to them, is to be found in the essay/manifesto, see Benjamin et al., 2018, of only three pages signed by no less than 72 authors. An essay which, in challenging aspects that may have been considered long-established, makes it clear, already in the subtitle, that

“We propose to change the default P-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.”

The trend, now dating back 30 years, to reduce the distances between opposing approaches and viewpoints has also affected the Fisher and Neyman - Pearson schools up to the open proposal by Lehmann, Neyman’s student, for the unification of theories of significance and hypothesis testing (see Lehmann, 1993; Lehmann, 2011). A visible manifestation of this widespread tendency is that in recent and less recent publications and, above all, in ordinary statistical practice, the Fisher and Neyman - Pearson approaches are too often syncretically confused. One reason for this mixture is due to the fact that the Neyman-Pearson approach does not require a measure of evidence; thus the investigator, who often reasons in terms of evidence measures, is inclined to resort to the *p-value*. Such a mixed approach, unable to capture the salient and distinctive methodological aspects of the two theories, fatally ends up with impoverishing and confusing the same statistical analysis.

Prominent examples of this practice, which is now widespread, concern very expensive trials requiring numerous statistical units; a typical one is the testing of new drugs. In the design phase of the clinical experiment, i.e. *ex ante*, it is usual to resort to hypothesis testing to establish the sample size to be adopted, after having preliminarily fixed the first kind error, e.g. $\alpha = 0.01$, and imposed a constraint on the error of the second kind, e.g. $\beta \leq 0.15$ (or, equivalently, to the power of the test $\eta \geq 1 - \beta$). Once the experiment is completed, when the results are evaluated, i.e. *ex post*, the significance test is used (see Spiegelhalter, 2019).

The spirit of reconciliation between these approaches has also made its way among Bayesians, with the imposition of an artificial distinction between Bayesian approaches. The first is called objective Bayesian, which forbids the involvement in the analysis of any information held by the investigator and which, consequently, prescribes the exclusive use of non-informative default priors. The second, called subjective Bayesian, which does not exclude the use of elicited priors.

Among objective Bayesians, there has arisen a large strand of studies with the intention of demonstrating the substantial correspondence between frequentist and Bayesian results. Procedures involving the re-use of the data are not infrequent: in the first phase the data are used to construct the prior, in the second, with the same data, inferences

are made. This is all in the presence of more or less open violations of the likelihood principle. See in this respect J. O. Berger and Sellke, 1987, O'Hagan, 1995, J. O. Berger and Pericchi, 1996, Bayarri and Berger, 2000, as well as the extensive bibliography to be found there.

1.3 Outline of the thesis

In Chapter 2 we define the Bayesian Discrepancy Measure, with the intent to make a contribution to the Bayesian procedure of testing precise hypotheses for parametric models. It is an evidence measure which allows one to evaluate the suitability of a given hypothesis with respect to the available information (prior law and data). The definition of the proposed index is presented for a scalar parameter of interest, both in the absence or presence of nuisance parameters. In order to facilitate its comprehension, different illustrative examples are discussed involving one or two independent populations. In addition, a comparison between the Bayesian Discrepancy Test and the Full Bayesian Significance Test is presented.

In Chapter 3 we extend this testing procedure allowing to compare parameter functions from two independent populations. Some particular scenarios are developed explicitly and the related examples are presented. The proposed approach is flexible, as it can be adapted to take into account different distributions and different parameter transformations. In addition, this methodology enables us to tackle some problems that are not yet covered in the literature.

In Chapter 4 we propose a general procedure, based on the BDT, for comparing k parameters, or their transformations, from independent populations. The presented approach is not a simple extension of the one outlined in the previous chapter since the geometry of the hypothesis and the parameter space require the construction of an entirely new approach. Once again, this methodology allows us to discuss problems not yet addressed in the literature.

In Chapter 5 we deal with the demanding problem of testing partial correlation coefficients when considering a multivariate Gaussian model. This discussion is complemented by several examples.

Finally, Chapter 6 contains conclusions and directions for further research.

This work has led to the publications of the following short papers, see Bertolino, Columbu, and Manca, 2022 and Manca et al., 2022. Further papers are under review, see Bertolino et al., 2021 and Bertolino, Columbu, Manca, and Musio, 2022, 2023.

Chapter 2

A new Bayesian Evidence Measure

Consider a parametric statistical model

$$\mathcal{F} = \{f(x|\boldsymbol{\theta}), x \in \mathcal{X}, \boldsymbol{\theta} \in \Theta\} \quad (2.1)$$

indexed on the parameter $\boldsymbol{\theta}$, where $\Theta \subseteq \mathbb{R}^p$ is the parameter space and $\mathcal{X} \subseteq \mathbb{R}^k$ is the sample space. Let g_0 be a prior distribution over $\Theta \subseteq \mathbb{R}^p$ that summarizes the subject's prior knowledge. Given data $\boldsymbol{x} = (x_1, \dots, x_n)$, consisting of n *iid* observations dependent on the parameter $\boldsymbol{\theta}$, the posterior probability distribution is denoted by

$$g_1(\boldsymbol{\theta}|\boldsymbol{x}) \propto g_0(\boldsymbol{\theta}) \cdot L(\boldsymbol{\theta}|\boldsymbol{x}),$$

where $L(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^n f(x_i|\boldsymbol{\theta})$ is the likelihood function. Assume that

- (a) the model is identifiable;
- (b) $f(x|\boldsymbol{\theta})$ have support not depending on $\boldsymbol{\theta}, \forall \boldsymbol{\theta} \in \Theta$;
- (c) the operations of integration and differentiation with respect to $\boldsymbol{\theta}$ can be exchanged.

We also assume a prior probability density $g_0(\boldsymbol{\theta})$ following Cromwell's Rule which states that “*it is inadvisable to attach probabilities of zero to uncertain events, for if the prior probability is zero so is the posterior, whatever be the data. A probability of one is equally dangerous because then the probability of the complementary event will be zero*” (see Section 6.2 in Lindley, 1991).

First we discuss the case of a scalar parameter. Then we discuss the case of a scalar parameter of interest in the presence of nuisance parameters.

2.1 The Bayesian Discrepancy Measure for a scalar parameter

Assume that $k = p = 1$. Given an *iid* random sample $\mathbf{x} = (x_1, \dots, x_n)$ from \mathcal{P}_θ^X , let $L(\theta|\mathbf{x})$ be the corresponding likelihood function based on data \mathbf{x} and let $g_0(\theta)$ and $g_1(\theta|\mathbf{x})$ be, respectively, the prior and the posterior distributions on $\Theta \subseteq \mathbb{R}$.

Moreover, given the posterior distribution function $G_1(\theta|\mathbf{x})$, the posterior median is any real number m_1 which satisfies the inequalities $G_1(m_1|\mathbf{x}) \geq \frac{1}{2}$ and $G_1^-(m_1|\mathbf{x}) \leq \frac{1}{2}$, where $G_1^-(m_1|\mathbf{x}) = \lim_{\theta \uparrow m_1} G_1(\theta|\mathbf{x})$. In the case in which G_1 is continuous and strictly increasing we have $m_1 = G_1^{-1}(\frac{1}{2}|\mathbf{x})$. We are interested in testing the precise hypothesis

$$H : \theta = \theta_H. \quad (2.2)$$

In order to measure the discrepancy of the hypothesis (2.2) w.r.t. the posterior distribution, in the case $\Theta = \mathbb{R}$, we consider the following two intervals:

1. the *discrepancy interval*

$$I_H = \begin{cases} (m_1, \theta_H) & \text{if } m_1 < \theta_H \\ \{m_1\} & \text{if } m_1 = \theta_H, \\ (\theta_H, m_1) & \text{if } m_1 > \theta_H \end{cases} \quad (2.3)$$

2. the *external interval*

$$I_E = \begin{cases} (\theta_H, +\infty) & \text{if } m_1 < \theta_H \\ (-\infty, \theta_H) & \text{if } \theta_H < m_1. \end{cases} \quad (2.4)$$

When $m_1 = \theta_H$, the external interval I_E can be $(-\infty, m_1)$ or $(m_1, +\infty)$. Note that, by construction, $\mathbb{P}(I_H \cup I_E) = \frac{1}{2}$ (see Figure 2.1). If the support of the posterior is a subset of \mathbb{R} , the intervals I_H and I_E can be defined consequently.

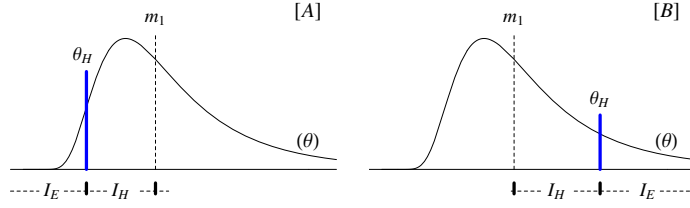


Figure 2.1: Posterior density $g_1(\theta|\mathbf{x})$, the corresponding discrepancy interval I_H and external interval I_E when $\theta_H < m_1$ ([A]) and $\theta_H > m_1$ ([B]).

Definition 2.1. Given the posterior distribution function $G_1(\theta|\mathbf{x})$, we define the Bayesian Discrepancy Measure of the hypothesis H as

$$\delta_H = 2 \mathbb{P}(\theta \in I_H|\mathbf{x}) = 2 \int_{I_H} dG_1(\theta|\mathbf{x}). \quad (2.5)$$

The measure can be also computed by means of the external interval as

$$\delta_H = 1 - 2 \mathbb{P}(\theta \in I_E|\mathbf{x}) = 1 - 2 \int_{I_E} dG_1(\theta|\mathbf{x}), \quad (2.6)$$

which can also be written as

$$\delta_H = 1 - 2 \min\{G_1(\theta_H|\mathbf{x}), 1 - G_1^-(\theta_H|\mathbf{x})\}, \quad (2.7)$$

where $G_1^-(\theta_H|\mathbf{x}) = \lim_{\theta \uparrow \theta_H} G_1(\theta|\mathbf{x})$. In the absolutely continuous case, this simplifies to

$$\delta_H = 1 - 2 \min\{G_1(\theta_H|\mathbf{x}), 1 - G_1(\theta_H|\mathbf{x})\}. \quad (2.8)$$

Formulations (2.7) and (2.8) have the advantage of not involving the posterior median in the integral computation. Furthermore, one can interpret the quantity $\min\{G_1(\theta_H|\mathbf{x}), 1 - G_1(\theta_H|\mathbf{x})\}$ as the posterior probability of a “tail” event concerning only the precise hypothesis H . Doubling this “tail” probability, related to the precise hypothesis H , one gets a posterior probability assessment about how “central” the hypothesis H is, and hence how it is supported by the prior and the data.

It is important to highlight that the hypothesis H induces the following partition

$$\{\Theta_a = (-\infty, \theta_H), \Theta_H = \{\theta_H\}, \Theta_b = (\theta_H, \infty)\} \quad (2.9)$$

of the parameter space Θ . Then formulations (2.7) and (2.8) can be equivalently expressed as

$$\delta_H = 1 - 2 \cdot \min_{a,b} \{\mathbb{P}(\theta \in \Theta_a | \mathbf{x}), \mathbb{P}(\theta \in \Theta_b | \mathbf{x})\}. \quad (2.10)$$

The last formula can be naturally extended to the case where, besides the scalar parameter of interest, nuisance parameters are also present. This issue will be developed in Section 2.2.

As pointed out before, the further θ_H is from the posterior median m_1 of the distribution function $G_1(\theta | \mathbf{x})$, the closer δ_H is to 1. It can then be said that H does *not conform* to $G_1(\theta | \mathbf{x})$. On the contrary, the smaller δ_H the stronger is the evidence in favor of H . Following this idea, we can define a general testing procedure by choosing a certain threshold to establish how large the measure must be, before we can state that H does not conform to the posterior distribution function.

Definition 2.2. The Bayesian Discrepancy Test (BDT) is the procedure based on the Bayesian Discrepancy Measure (BDM) that rejects the hypothesis H when δ_H is higher than some critical value $\omega \in \{0.95, 0.99, 0.995, 0.999, \dots\}$.

As for all measures of evidence (Bayesian or frequentist), the chosen value for ω inevitably has a character of subjectivity. Fisher himself adopted a fixed level and, speaking of the values at which the χ^2 -test statistic is significant at a given level, he justifies his reasoning (see Lehmann, 1993) as follows “... we do not want to know the exact value of P for any observed χ^2 , but, in the first place, whether or not the observed value is open to suspicion. If P is between .1 and .9, there is certainly no reason to suspect the hypothesis tested. If it is below .02, it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05 and consider that higher values of χ^2 indicate a real discrepancy”.

For a more detailed discussion on the threshold choice for the BDT see the remark in Example 2.4.

Proposition 2.3. The following properties apply to the BDM, for a scalar parameter θ :

- (i) δ_H always exists and, by construction, $\delta_H \in [0, 1)$;
- (ii) δ_H is invariant under invertible monotonic transformations of the parameter θ ;
- (iii) if θ^* is the true value of the parameter and $\theta_H = \theta^*$, then $\delta_H \sim Unif(\cdot | 0, 1)$, for all sample sizes n ; if $\theta_H \neq \theta^*$, then $\delta_H \xrightarrow{P} 1$ (consistency property).

Proof. (i) The first property follows immediately from the fact that in (2.5) the posterior probability $\mathbb{P}(\theta \in I_H | \mathbf{x}) \in [0, \frac{1}{2}]$.

- (ii) Let $\lambda = \lambda(\theta)$ be an invertible monotonic transformation of the parameter θ and let K_1 be the cumulative distribution function of the parameter λ . We denote with $\lambda_H = \lambda(\theta_H)$ and we notice that $m'_1 = \lambda(m_1)$ thanks to the monotonic invariance of the median. Suppose, for simplicity, that $\theta_H > m_1$. Then

$$\delta_H = 2 \int_{m_1}^{\theta_H} dG_1(\theta | \mathbf{x}) = 2 \left| \int_{m'_1}^{\lambda_H} dK_1(\lambda | \mathbf{x}) \right|.$$

Therefore, the invariance of the BDM follows immediately from the invariance of the median under invertible monotonic transformations. Notice that if instead of the median m_1 we consider, for example, the posterior mean $E(\theta | \mathbf{x})$, which is not invariant under invertible monotonic reparametrizations, the property will not hold in general. Moreover, $E(\theta | \mathbf{x})$ for some models may not even exist.

- (iii) Let us examine the first part of the statement for which $\theta_H = \theta^*$. Suppose that $\theta_H < m_1$. The BDM is defined as

$$\begin{aligned} \delta_H &= 2 \int_{\theta_H}^{m_1} dG_1(\theta | \mathbf{x}) \\ &= 1 - 2 \int_{-\infty}^{\theta_H} dG_1(\theta | \mathbf{x}) \\ &= 1 - 2 G_1(\theta_H | \mathbf{x}). \end{aligned} \tag{2.11}$$

Using the integral transform and the fact that we have supposed $\theta_H < m_1$, we easily find that

$$G_1(\theta_H | \mathbf{x}) = \int_{-\infty}^{\theta_H} dG_1(\theta | \mathbf{x}) = W \sim Unif(\cdot | 0, \frac{1}{2}).$$

Then, since $\delta_H = 1 - 2W$, we find $\delta_H \sim Unif(\cdot|0, 1)$. A similar proof holds for $\theta_H > m_1$. If, instead, $\theta_H \neq \theta^*$ and $n \rightarrow \infty$, under suitable regularity conditions (see for instance Section 5.3.2, p. 287 in Bernardo and Smith, 1994) it is well known that $g_1(\theta|\mathbf{x})$ is concentrated around θ^* . In particular, the posterior median m_1 converges in probability to θ^* . Again, suppose for instance that $\theta_H < \theta^*$, then $\lim_{n \rightarrow \infty} \delta_H = 2 \lim_{n \rightarrow \infty} \int_{\theta_H}^{m_1} dG_1(\theta|\mathbf{x}) = 2 \cdot \frac{1}{2} = 1$.

□

2.2 The Bayesian Discrepancy Measure in presence of nuisance parameters

Suppose that $p \geq 2$ and $k \geq 1$. Let $\varphi = \varphi(\boldsymbol{\theta})$ be a scalar parameter of interest, where $\varphi : \Theta \rightarrow \Phi \subseteq \mathbb{R}$. Let us further consider a bijective reparametrization $\boldsymbol{\theta} \Leftrightarrow (\varphi, \boldsymbol{\zeta})$, where $\boldsymbol{\zeta} \in \mathcal{Z} \subseteq \mathbb{R}^{p-1}$ denotes an arbitrary nuisance parameter, which is determined on the basis of analytical convenience (note that the value of the evidence measure is invariant with respect to the choice of the nuisance parameter). We consider hypotheses that can be expressed in the form

$$H : \varphi = \varphi_H, \quad (2.12)$$

where φ_H is known as it represents the hypothesis that it is of interest to evaluate. The transformation φ must be such that, for all $\boldsymbol{\theta} \in \Theta$ and for all $\varphi_H \in \Phi$, it can always be assessed whether φ is strictly smaller, strictly larger or equal to φ_H (i.e. $\varphi < \varphi_H$ either $\varphi > \varphi_H$, or $\varphi = \varphi_H$). Hypothesis (2.12) and transformation φ univocally identify the partition $\{\Theta_a, \Theta_H, \Theta_b\}$ of the parameter space Θ , with

$$\begin{aligned} \Theta_a &= \{\boldsymbol{\theta} \in \Theta \mid \varphi < \varphi_H\} \\ \Theta_H &= \{\boldsymbol{\theta} \in \Theta \mid \varphi = \varphi_H\}. \\ \Theta_b &= \{\boldsymbol{\theta} \in \Theta \mid \varphi > \varphi_H\} \end{aligned} \quad (2.13)$$

We call any hypothesis of type (2.12), which identify a partition of the form (2.13), a *partitioning hypothesis*. It is easy to verify that many commonly used hypotheses are partitioning. In this thesis we only consider hypotheses of this nature. In this setting, we

express the BDM as

$$\begin{aligned}\delta_H &= 1 - 2 \cdot \min_{a,b} \{ \mathbb{P}(\boldsymbol{\theta} \in \Theta_a | \mathbf{x}), \mathbb{P}(\boldsymbol{\theta} \in \Theta_b | \mathbf{x}) \} \\ &= 1 - 2 \cdot \int_{I_E} g_1(\boldsymbol{\theta} | \mathbf{x}) d\boldsymbol{\theta},\end{aligned}\tag{2.14}$$

where *the external set* is given by

$$I_E = \arg \min_{a,b} \{ \mathbb{P}(\boldsymbol{\theta} \in \Theta_a | \mathbf{x}), \mathbb{P}(\boldsymbol{\theta} \in \Theta_b | \mathbf{x}) \} .\tag{2.15}$$


In the particular scenario where the marginal posterior

$$h_1(\varphi | \mathbf{x}) = \int_{\varphi(\boldsymbol{\theta})=\varphi} g_1(\boldsymbol{\theta} | \mathbf{x}) d\boldsymbol{\theta}, \quad \forall \varphi \in \Phi,$$

of the parameter of interest φ can be computed in a closed form, the hypothesis (2.12) can be easily treated using the methodologies seen in Subsection 2.1, i.e. the BDM is computed by means of formula (2.5) or (2.6) applied to the marginal.

Properties reported in Proposition 2.3 naturally extend to the setting we just presented.

2.3 Illustrative examples

The simplicity of the BDT is highlighted by the following examples, some of which deal with cases not usually considered in the literature. Examples 2.4 and 2.5 focus on a scalar parameter of interest, while Examples 2.6, 2.7, 2.8, 2.9, 2.10 also contain nuisance parameters. All of the  codes for these examples can be found online in Manca, 2022.

In all of the examples we have adopted Jeffrey's priors (see Yang and Berger, 1996 for a catalog of non-informative priors). Of course, other priors (objective or subjective) could be equally used.

2.3.1 Examples of the univariate parameter case

Example 2.4. Exponential distribution

Let $\mathbf{x} = (x_1, \dots, x_n)$ be an *iid* sample of size n from the exponential distribution $X \sim \text{Exp}(x|\theta^{-1})$, with $\theta \in \mathbb{R}^+$. We are interested in the hypothesis $H : \theta = \theta_H$. Assuming

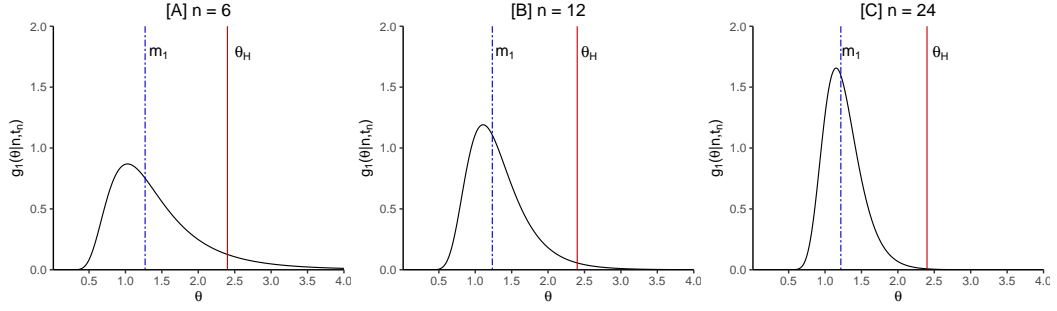


Figure 2.2: Posterior density function $g_1(\theta|n\bar{x})$ and intervals $I_H = (m_1, \theta_H)$ and $I_E = (\theta_H, \infty)$, using data from Example 2.4.

a Jeffreys' prior for θ , i.e. $g_0(\theta) \propto \theta^{-1}$, the posterior distribution is given by $g_1(\theta|\mathbf{x}) \propto \theta^{-n-1} \exp\{-n\bar{x} \cdot \theta^{-1}\}$, with \bar{x} the sample mean.

Figure 2.2 shows the posterior density function as well as the discrepancy and the external intervals for $H : \theta = \theta_H = 2.4$ and the MLE $\bar{x} = 1.2$ for three sample sizes [A] $n = 6$, [B] $n = 12$, [C] $n = 24$. In [A] we have a posterior median $m_1 = 1.27$ and $\delta_H = 0.832$, while in [B] $m_1 = 1.23$ and $\delta_H = 0.960$, in [C] $m_1 = 1.22$ and $\delta_H = 0.997$. In case [A] we do not reject H while in [B] and in [C] we are led to reject the hypothesis.

Note that in all scenarios considered, we find the following relation between δ_H and the p -value,

$$p\text{-value} = 1 - \delta_H \quad (2.16)$$

(in [A] $\delta_H = 0.832$ and $p\text{-value} = 0.168$, in [B] $\delta_H = 0.96$ and $p\text{-value} = 0.04$, while in [C] $\delta_H = 0.997$ and $p\text{-value} = 0.003$). This result depends clearly on the use of the Jeffreys' prior, which is a matching prior for a scalar parameter (see Ruli and Ventura, 2021).

Remark 1. The fact that classical and Bayesian procedures, under certain conditions, produce the same conclusions is well known (see, for instance, Lindley, 1965). The linear relationship (2.16) also occurs in other simple cases. Even if it does not hold for more complicated models and in general for proper priors, it suggests a relationship between the traditional p -value levels of significance $\{0.05, 0.01, 0.005, \dots\}$, and the critical values for the discrepancy measure $\{0.95, 0.99, 0.995, \dots\}$. In this thesis we will not investigate the problem of the choice of the BDM threshold ω . Several aspects about the choice of p -values thresholds have been considered in Benjamin et al., 2018

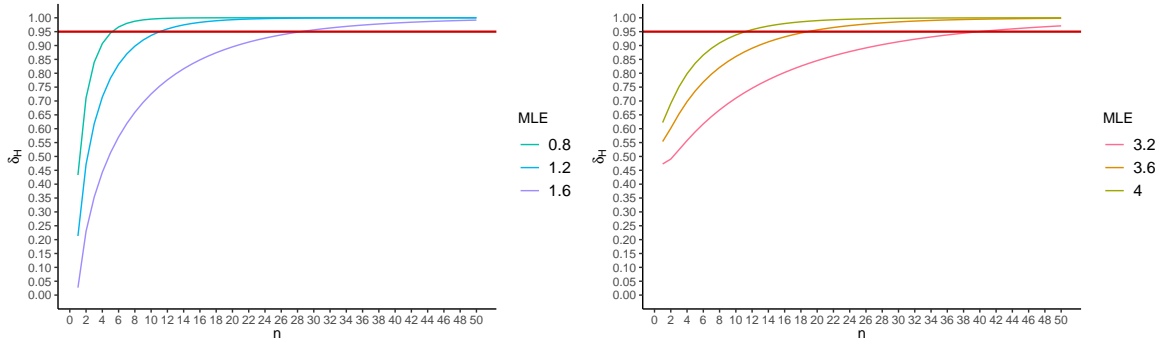


Figure 2.3: BDM for n increasing and for different values of the MLE. Case [A] with MLE = 0.8 (a), 1.2 (b), 1.6 (c) and case [B] with MLE = 3.2 (f), 3.6 (e), 4 (d).

and can be suitably extended to the BDM.

Finally, to conclude Example 2.4, it is useful to show the trend of the BDM when varying $n = 1, 2, \dots, 25$ for six values of the MLE: (a) 0.8, (b) 1.2, (c) 1.6 (case [A]) and (d) 4.0, (e) 3.6, (f) 3.2 (case [B]), see Figure 2.3. In order to explain the difference between the BDM trends in cases [A] and [B], consider that:

- (i) in case [A] the posterior median $m_1 < \theta_H = 2.4$, whereas in case [B] $m_1 > \theta_H = 2.4$;
- (ii) δ_H is monotonically increasing, both with respect to n , and with respect to the distance $|m_1 - \theta_H|$;
- (iii) the posterior g_1 always has a positive asymmetry, which decreases as n increases;
- (iv) the trend difference of the BDM in cases [A] and [B] depends on the fact that the posterior g_1 has ‘small’ tails on the left-hand side of m_1 and ‘large’ tails on the right-hand side.

Moving forward in the discussion, in order to highlight the valutive nature of the BDT, it is worth pointing out that it allows the separate and simultaneous testing of $\ell \geq 2$ hypotheses

$$H_j : \varphi = \varphi_j, \quad j = 1, 2, \dots, \ell, \quad (2.17)$$

as shown in Example 2.5. Remember that with the decisionist approach, among the ℓ competing hypotheses, only one is accepted. On the contrary, under the valutive approach that allows the valuation of a single hypothesis H_j , it may happen that several hypotheses are supported by the data, or even that all hypotheses must be rejected.

Example 2.5. In the 1700s, several hypotheses $H_j : \theta = \theta_j$ were formulated about the birth masculinity rate $\theta = \frac{M}{M+F}$. Among them we consider $\theta_1 = \frac{1}{2}$ (J. Bernoulli), $\theta_2 = \frac{13}{25}$ (J. Arbuthnot), $\theta_3 = \frac{1050}{2050}$ (J. P. Süßmilch), $\theta_4 = \frac{23}{45}$ (P. S. Laplace). We assume that the gender of each newborn is modeled as a $Bin(\cdot|1, \theta)$. Then, using data recorded in 1710 in London (see, for instance, Spiegelhalter, 2019), with 7640 males and 7288 females (the MLE is $\hat{\theta} = 0.512$) and assuming the Jeffreys' prior $Beta(\theta|1, 1)$, we compute δ_{H_j} using the Normal asymptotic approximation

$$\delta_{H_j} \cong 1 - 2 \cdot \int_{I_E^j} \tilde{g}_1(\theta|\hat{\theta}, \frac{1}{n}\hat{\theta}(1-\hat{\theta}))d\theta, \quad j = 1, 2, 3, 4,$$

with \tilde{g}_1 the Normal distribution. Since $\delta_{H_1} = 0.996$, $\delta_{H_2} = 0.955$, $\delta_{H_3} = 0.079$, $\delta_{H_4} = 0.132$, we can conclude that the first two hypotheses has to be rejected, while there is not enough evidence to reject the hypotheses made by Süßmilch and Laplace.

2.3.2 Examples of the more general case

The examples presented hereafter, can be distinguished by tests concerning a parameter or a parametric function of a single population, and tests concerning the comparison of two independent population parameters.

Tests involving a single population

Example 2.6. - Test on the shape parameter, mean and variance of the Gamma distribution

Let $\mathbf{x} = (x_1, \dots, x_n)$ be an iid sample of size n from $X \sim Gamma(x|\alpha, \beta)$, $(\alpha, \beta) \in \mathbb{R}^+ \times \mathbb{R}^+$. We denote by m_g the geometric mean of \mathbf{x} . The likelihood function for (α, β) is given by

$$L(\alpha, \beta|\mathbf{x}) \propto \left(\frac{\beta^\alpha}{\Gamma(\alpha)} \cdot m_g^\alpha \cdot e^{-\bar{x} \cdot \beta} \right)^n.$$

For the fictitious data $\mathbf{x} = (0.8, 1.1, 1.2, 1.4, 1.8, 2, 4, 5, 8)$, we find that the MLEs are $\hat{\alpha} = 1.921$ and $\hat{\beta} = 0.7572$.

We are interested in testing the hypotheses [A] $H_A : \alpha = \alpha_H$, with $\alpha_H = 2.5$, [B] $H_B : \mu = \mu_H$, with $\mu_H = 6$, and [C] $H_C : \sigma^2 = \sigma_H^2$, with $\sigma_H^2 = 2$, where $\mu = \frac{\alpha}{\beta}$ and $\sigma^2 = \frac{\alpha}{\beta^2}$ denote the mean and the variance of X .

Adopting the Jeffreys' prior for (α, β) , i.e.

$$g_0(\alpha, \beta) = g_0^\alpha(\alpha) \cdot g_0^\beta(\beta) \propto \sqrt{\alpha \cdot \psi^{(1)}(\alpha) - 1} \cdot \frac{1}{\beta},$$

where $\psi^{(1)}(\alpha) = \sum_{j=0}^{\infty} (\alpha + j)^{-2}$ denotes the *trigamma* function, the posterior for (α, β) is given by $g_1(\alpha, \beta | \mathbf{x}) = k \cdot g_0^\alpha(\alpha) \cdot g_0^\beta(\beta) \cdot L(\alpha, \beta | \mathbf{x})$, with normalizing constant k .

- Case [A]

The hypothesis H_A identifies the vertical straight line of equation $\alpha = \alpha_H$ and two subsets $\Theta_a = \{(\alpha, \beta) | \alpha < \alpha_H\}$ and $\Theta_b = \{(\alpha, \beta) | \alpha > \alpha_H\}$ (see Figure 2.4 [A]). Then we can compute

$$\begin{aligned} \mathbb{P}((\alpha, \beta) \in \Theta_b | \mathbf{x}) &= \int_{\alpha_H}^{\infty} \int_0^{\infty} g_1(\alpha, \beta | \mathbf{x}) \, d\beta \, d\alpha \\ &= k \cdot \int_{\alpha_H}^{\infty} \int_0^{\infty} \sqrt{\alpha \cdot \psi^{(1)}(\alpha) - 1} \cdot \frac{1}{\beta} \left(\frac{\beta^\alpha}{\Gamma(\alpha)} \cdot m_g^\alpha \cdot e^{-\bar{x} \cdot \beta} \right)^n \, d\beta \, d\alpha \\ &= k \cdot \int_{\alpha_H}^{\infty} \sqrt{\alpha \cdot \psi^{(1)}(\alpha) - 1} \cdot \frac{\Gamma(n\alpha)}{\Gamma(\alpha)^n} \cdot \left(\frac{m_g}{n\bar{x}} \right)^{n\alpha} \, d\alpha = 0.215, \end{aligned}$$

and $\delta_H = 0.570$, a value that does not allows for the rejection of H_A .

- Case [B]

The hypothesis H_B identifies the straight line of equation $\beta = \frac{1}{\mu_H} \alpha$ in the $\alpha\beta$ -plane (see Figure 2.4 [B]) and the two subsets

$$\Theta_c = \{(\alpha, \beta) | \beta > \frac{1}{\mu_H} \alpha\} \quad \text{and} \quad \Theta_d = \{(\alpha, \beta) | \beta < \frac{1}{\mu_H} \alpha\}.$$

We have

$$\mathbb{P}((\alpha, \beta) \in \Theta_d | \mathbf{x}) = \int_{\Theta_d} g_1(\alpha, \beta | \mathbf{x}) \, d\alpha \, d\beta = 0.012,$$

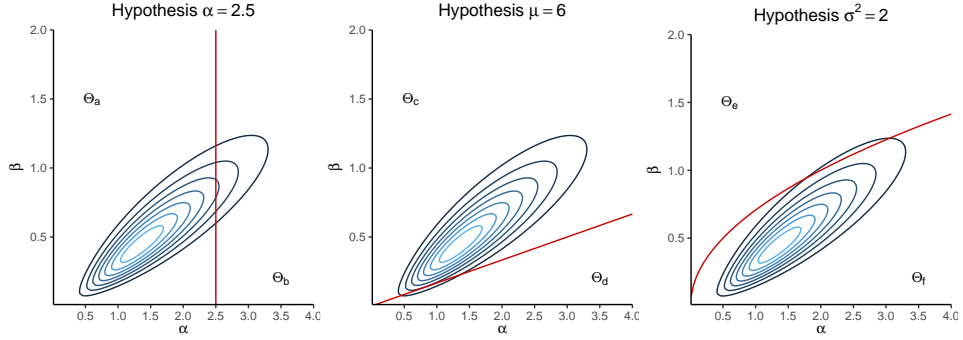


Figure 2.4: Posterior density function $g_1(\alpha, \beta | \mathbf{x})$ from Example 2.6 and corresponding sets of the induced partition in the cases [A], [B] and [C].

and, since $\delta_H = 0.976$, we reject H_B .

- Case [C]

The hypothesis H_C identifies the parabola of equation $\beta = \frac{1}{\sqrt{\sigma_H^2}}\sqrt{\alpha}$, in the $\alpha\beta$ -plane (see Figure 2.4 [C]), and the two subsets

$$\Theta_e = \left\{ (\alpha, \beta) \mid \beta > \frac{1}{\sqrt{\sigma_H^2}}\sqrt{\alpha} \right\} \quad \text{and} \quad \Theta_f = \left\{ (\alpha, \beta) \mid \beta < \frac{1}{\sqrt{\sigma_H^2}}\sqrt{\alpha} \right\}.$$

We have

$$\mathbb{P}((\alpha, \beta) \in \Theta_e | \mathbf{x}) = \int_{\Theta_e} g_1(\alpha, \beta | \mathbf{x}) \, d\alpha \, d\beta = 0.078.$$

Therefore $\delta_H = 0.846$, and so we do not reject H_C .

Example 2.7. - Test on the coefficient of variation for a Normal distribution

Given an iid sample $\mathbf{x} = (x_1, \dots, x_n)$ from $X \sim N(x | \mu, \phi^{-1})$, the parameter of interest is $\psi = \frac{\sqrt{\text{Var}(X)}}{|\mathbb{E}(X)|} = \frac{1}{|\mu| \sqrt{\phi}}$. We are interested in testing the hypothesis

$$H : \psi = \psi_H,$$

with $\psi_H = 0.1$. If we consider the Jeffreys' prior $g_0(\mu, \phi) \propto \phi^{-1} \cdot \mathbf{1}_{\mathbb{R} \times \mathbb{R}^+}$, the posterior distribution is the Normal-Gamma density

$$(\mu, \phi) | \mathbf{x} \sim NG(\mu, \phi | \eta, \nu, \alpha, \beta),$$

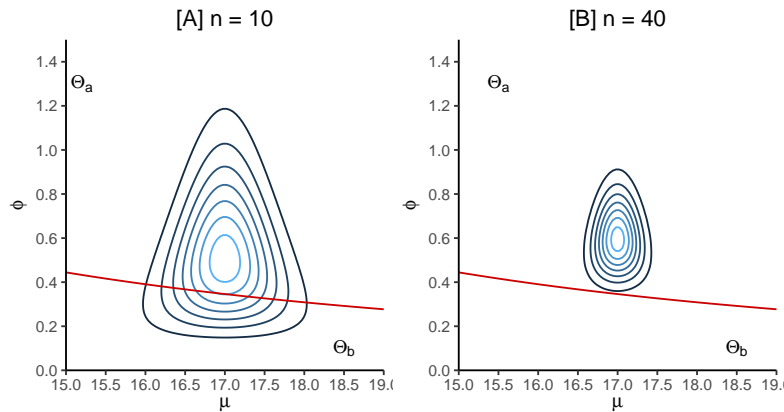


Figure 2.5: Test on the coefficient of variation ψ of a Gaussian population. Data refers to Example 2.7. In the plots, the sets Θ_a , Θ_b and Θ_H are reported for $n = 10$ ([A]) and $n = 40$ ([B]).

with hyperparameters $(\eta, \nu, \alpha, \beta)$, where $\eta = \bar{x}$, $\nu = n$, $\alpha = \frac{1}{2}(n-1)$, $\beta = \frac{1}{2}ns^2$, and density

$$g_1(\mu, \phi \mid \eta, \nu, \alpha, \beta) = \frac{\beta^\alpha \sqrt{\nu}}{\Gamma(\alpha) \sqrt{2\pi}} \phi^{\alpha-1/2} e^{-\frac{\nu\phi}{2}(\mu-\eta)^2} e^{-\beta\phi}.$$

We consider the particular case in which $\bar{x} = 17$ and $s^2 = 1.6$ (so that the MLE is $\hat{\phi} = 0.074$) with two samples of size $n = 10$ (Figure 2.5 [A]) and $n = 40$ (Figure 2.5 [B]). In the $\mu\phi$ -space, the hypothesis H is represented by the curve $\phi = \frac{1}{\psi_H^2} \mu^{-2}$ and determines the subsets Θ_a and Θ_b visualized in Figure 2.5.

In case [A] we have

$$\mathbb{P}((\mu, \phi) \in \Theta_b \mid \mathbf{x}) = \int_{\Theta_b} g_1(\mu, \phi \mid \eta, \nu, \alpha, \beta) d\mu d\phi = 0.215,$$

where $g_1(\mu, \phi \mid \eta, \nu, \alpha, \beta)$ is the Normal-Gamma density, so that $\delta_H = 0.570$ and we do not reject H . In case [B], we have $\mathbb{P}((\mu, \phi) \in \Theta_b \mid \mathbf{x}) = 0.014$ and, since $\delta_H = 0.972$, we reject H . Therefore in such a case, with different sample sizes, the inferential conclusions change.

Example 2.8. - Test on the skewness coefficient of the Inverse Gaussian distribution

Let us consider a Inverse Gaussian random variable X with density

$$f(x | \mu, \nu) = \sqrt{\frac{\nu}{2\pi x^3}} \exp \left\{ -\frac{\nu}{2} \left(\frac{x - \mu}{\mu\sqrt{x}} \right)^2 \right\} \cdot \mathbf{1}_{\mathbb{R}^+}(x),$$

where $(\mu, \nu) \in \mathbb{R}^+ \times \mathbb{R}^+$. The parameter of interest is the skewness coefficient $\gamma = 3\sqrt{\frac{\mu}{\nu}}$ and it is of interest to test the hypothesis $H : \gamma = \gamma_H$, where $\gamma_H = 2$. The Jeffreys' prior is

$$g_0(\mu, \nu) \propto \frac{1}{\sqrt{\mu^3 \nu}} \cdot \mathbf{1}_{\mathbb{R}^+ \times \mathbb{R}^+}(\mu, \nu).$$

Given n observations, the posterior distribution of (μ, ν) is

$$g_1(\mu, \nu | \mathbf{x}) \propto \sqrt{\frac{\nu^{n-1}}{\mu^3}} \cdot \exp \left\{ -\frac{n\nu}{2} \cdot \left(\frac{\bar{x}}{\mu^2} - \frac{2}{\mu} + \frac{1}{a} \right) \right\} \cdot \mathbf{1}_{\mathbb{R}^+ \times \mathbb{R}^+}(\mu, \nu),$$

where \bar{x} and a are the arithmetic and harmonic mean, respectively.

We apply the procedure to the following precipitation data (inches) from Jug Bridge, Maryland, analyzed in Folks and Chhikara, 1978 (p. 272):

1.01	1.11	1.13	1.15	1.16
1.17	1.17	1.20	1.52	1.54
1.54	1.57	1.64	1.73	1.79
2.09	2.09	2.57	2.75	2.93
3.19	3.54	3.57	5.11	5.62.

The hypothesis identifies in the parameter space $\Theta = \mathbb{R}^+ \times \mathbb{R}^+$ the subsets

$$\begin{aligned} \Theta_a &= \left\{ (\mu, \nu) \in \Theta \mid 3\sqrt{\frac{\mu}{\nu}} < \gamma_H \right\}, \\ \Theta_H &= \left\{ (\mu, \nu) \in \Theta \mid 3\sqrt{\frac{\mu}{\nu}} = \gamma_H \right\}, \\ \Theta_b &= \left\{ (\mu, \nu) \in \Theta \mid 3\sqrt{\frac{\mu}{\nu}} > \gamma_H \right\}. \end{aligned}$$

We have that

$$\mathbb{P}((\mu, \nu) \in \Theta_b | \mathbf{x}) = \int_{\Theta_b} g_1(\mu, \nu | \mathbf{x}) \, d\mu \, d\nu = 0.078, \quad (2.18)$$

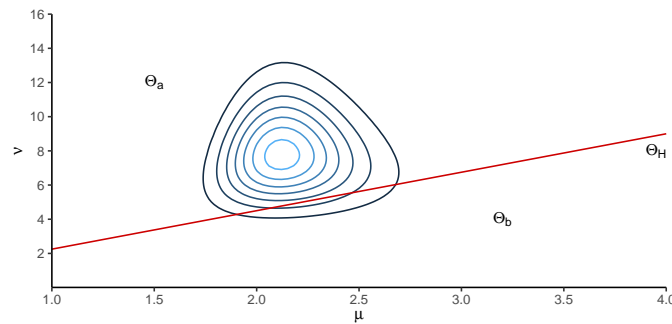


Figure 2.6: Test on the skewness of the Inverse Gaussian distribution with $\gamma_H = 2$. In the plot the sets of the partition induced by H are reported. Data refers to Example 2.8.

see Figure 2.8, then we obtain $\delta_H = 0.844$. This result indicates that we do not have enough evidence to reject the hypothesis H .

Tests involving two independent populations

In this section we consider some examples concerning comparisons between parameters of two independent populations.

Example 2.9. - Comparison between means and precisions of two independent Normal populations

Let us consider a case study on the dating of the core and periphery of some wooden furniture, found in a Byzantine church, using radiocarbon (see Casella and Berger, 2001, p. 409). The historians wanted to verify if the mean age of the core is the same as the mean age of the periphery, using two samples of sizes $m = 14$ and $n = 9$, respectively, given by

<i>core</i>	1294	1279	1274	1264	1263	<i>periphery</i>	1284	1272	1256
	1254	1251	1251	1248	1240		1254	1242	1274
	1232	1220	1218	1210			1264	1256	1250

We assume that the age of the core X and of the periphery Y are distributed as

$$X \sim N(x|\mu_1, \phi_1^{-1}) \quad \text{and} \quad Y \sim N(y|\mu_2, \phi_2^{-1}),$$

where $Var(X) = \phi_1^{-1}$ and $Var(Y) = \phi_2^{-1}$, and we assume that the data are *iid* conditional on the parameters. We consider for (μ_i, ϕ_i) the Jeffreys' prior

$$g_{0,i}(\mu_i, \phi_i) \propto \phi_i^{-1} \cdot \mathbf{1}_{\mathbb{R} \times \mathbb{R}^+}, \quad i = 1, 2.$$

We obtain $\bar{x} = 1249.86$, $\bar{y} = 1261.33$, $\bar{d} = \bar{x} - \bar{y} = -11.48$, while the MLEs for the sample standard deviations are $s_1 = 23.43$ and $s_2 = 12.51$. The posterior distribution for (μ_i, ϕ_i) is the Normal-Gamma law

$$(\mu_i, \phi_i) \mid \mathbf{x}, \mathbf{y} \sim NG(\mu_i, \phi_i \mid \eta_i, \nu_i, \alpha_i, \beta_i), \quad i = 1, 2,$$

with hyperparameters $\eta_1 = \bar{x}$, $\nu_1 = m$, $\alpha_1 = \frac{1}{2}(m - 1)$, $\beta_1 = \frac{1}{2}ms_1^2$ and $\eta_2 = \bar{y}$, $\nu_2 = n$, $\alpha_2 = \frac{1}{2}(n - 1)$, $\beta_2 = \frac{1}{2}ns_2^2$, and density

$$g_{1,i}(\mu_i, \phi_i \mid \eta_i, \nu_i, \alpha_i, \beta_i) = \frac{\beta_i^{\alpha_i} \sqrt{\nu_i}}{\Gamma(\alpha_i) \sqrt{2\pi}} \phi_i^{\alpha_i - 1/2} e^{-\frac{\nu_i \phi_i}{2} (\mu_i - \eta_i)^2} e^{-\beta_i \phi_i}, \quad i = 1, 2.$$

The hypothesis of interest

$$H_A : \mu_1 - \mu_2 = 0, \quad \forall \phi_1 > 0, \quad \forall \phi_2 > 0,$$

identifies the following subsets in the parameter space

$$\begin{aligned} \Theta_a &= \left\{ \mathbb{R}^2 \times \mathbb{R}_+^2 \mid \mu_1 < \mu_2 \right\}, \\ \Theta_{H_A} &= \left\{ \mathbb{R}^2 \times \mathbb{R}_+^2 \mid \mu_1 = \mu_2 \right\}, \\ \Theta_b &= \left\{ \mathbb{R}^2 \times \mathbb{R}_+^2 \mid \mu_1 > \mu_2 \right\}. \end{aligned}$$

Then we can compute

$$\begin{aligned} &\mathbb{P}((\mu_1, \mu_2, \phi_1, \phi_2) \in \Theta_a \mid \mathbf{x}, \mathbf{y}) \\ &= \int_{\Theta_a} \prod_{i=1}^2 g_{1,i}(\mu_i, \phi_i \mid \eta_i, \nu_i, \alpha_i, \beta_i) d\mu_1 d\mu_2 d\phi_1 d\phi_2 \\ &= \int_{\mu_1 < \mu_2} \prod_{i=1}^2 \frac{\Gamma(\alpha_i + \frac{1}{2})}{\Gamma(\alpha_i)} \left(\frac{\nu_i}{2\pi\beta_i} \right)^{1/2} \left[1 + \frac{\nu_i}{2\beta_i} (\mu_i - \eta_i)^2 \right]^{-(\alpha_i + \frac{1}{2})} d\mu_1 d\mu_2 \\ &= 0.089, \end{aligned}$$

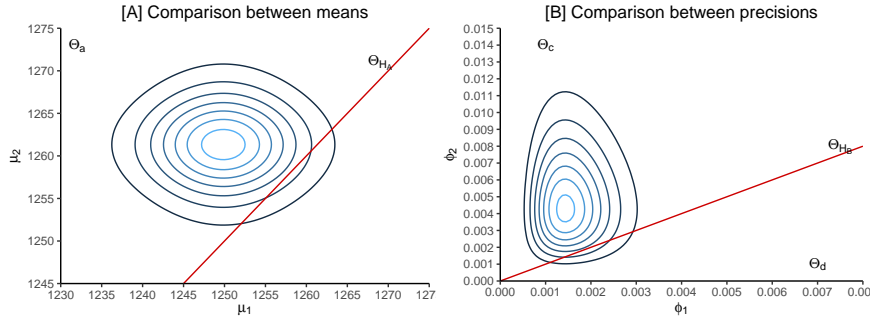


Figure 2.7: Comparisons between means ([A]) and precisions ([B]) of independent normal populations for data in Example 2.9. For both cases we show the contour plots of the marginals of μ_j ([A]) and ϕ_j ([B]), and the partition sets associated with the corresponding hypotheses.

so we have $\delta_H = 0.823$, a value that do not lead to the rejection of the hypothesis. We exploited the fact that the marginal of each μ_i is a Generalized Student's t-distribution (denoted by *StudentG*) with hyperparameters $(\eta_i, \frac{\beta_i}{\nu_i \alpha_i}, 2\alpha_i)$. Figure 2.7 [A] in the space (μ_1, μ_2) shows the contour lines of the distribution

$$StudentG\left(\mu_1 \mid \eta_1, \frac{\beta_1}{\nu_1 \cdot \alpha_1}, 2\alpha_1\right) \cdot StudentG\left(\mu_2 \mid \eta_2, \frac{\beta_2}{\nu_2 \cdot \alpha_2}, 2\alpha_2\right).$$

Note that the homoscedasticity assumption is not necessary. Consider now the hypothesis

$$H_B : \phi_1 - \phi_2 = 0, \quad \forall \mu_1, \mu_2,$$

which determines in the parameter space the subsets

$$\begin{aligned} \Theta_c &= \left\{ \mathbb{R}^2 \times \mathbb{R}_+^2 \mid \phi_1 < \phi_2 \right\}, \\ \Theta_{H_B} &= \left\{ \mathbb{R}^2 \times \mathbb{R}_+^2 \mid \phi_1 = \phi_2 \right\}, \\ \Theta_d &= \left\{ \mathbb{R}^2 \times \mathbb{R}_+^2 \mid \phi_1 > \phi_2 \right\}. \end{aligned}$$

We have

$$\mathbb{P}\left((\mu_1, \mu_2, \phi_1, \phi_2) \in \Theta_c \mid \mathbf{x}, \mathbf{y}\right) = \int_{\phi_1 < \phi_2} \prod_{i=1}^2 \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} \phi_i^{\alpha_i-1} e^{-\phi_i \beta_i} d\phi_1 d\phi_2 = 0.046,$$

from which it follows that $\delta_H = 0.908$ and we reject the hypothesis H . To compute the

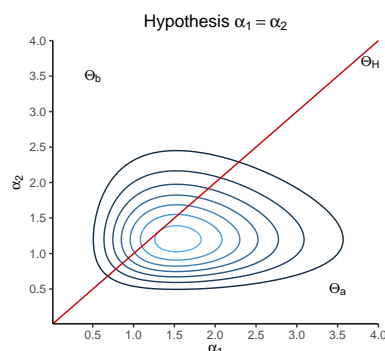


Figure 2.8: Comparison of the shape parameters of two independent Gamma populations, using data of Example 2.10. The sets Θ_a , Θ_b and Θ_H of the partition are reported.

integral we have used the fact that the marginal of each ϕ_i has Gamma distribution with parameters (α_i, β_i) , $i = 1, 2$.

The contour lines of the law $Gamma(\phi_1|\alpha_1, \beta_1) \cdot Gamma(\phi_2|\alpha_2, \beta_2)$, in the space (ϕ_1, ϕ_2) , are reported in Figure 2.7 [B].

Example 2.10. - Comparison between shape parameters of two Gamma distributions

Let us consider two *iid* Gamma populations $X_i \sim Gamma(\alpha_i, \beta_i)$, $(\alpha_i, \beta_i) \in \mathbb{R}^+ \times \mathbb{R}^+$, $i = 1, 2$, and let us consider two samples of sizes $n_1 = 9$ and $n_2 = 12$, respectively, with sample means $\bar{x}_1 = 2.811$ and $\bar{x}_2 = 1.973$, and geometric means $m_{g_1} = 2.116$ and $m_{g_2} = 1.327$.

We are interested in testing $H : \alpha_1 = \alpha_2$. The posterior distribution for $(\alpha_1, \beta_1, \alpha_2, \beta_2)$ is given by

$$g_1(\alpha_1, \beta_1, \alpha_2, \beta_2 | \mathbf{x}_1, \mathbf{x}_2) = g_{1,1}(\alpha_1, \beta_1 | \mathbf{x}_1) \cdot g_{1,2}(\alpha_2, \beta_2 | \mathbf{x}_2),$$

where

$$g_{1,i}(\alpha_i, \beta_i | \mathbf{x}_i) = k_i \cdot g_{0,i}(\alpha_i, \beta_i) \cdot L(\alpha_i, \beta_i | \mathbf{x}_i),$$

with normalizing constant k_i , $i = 1, 2$. Let $\Theta_a = \{(\alpha_1, \alpha_2) \in \mathbb{R}^+ \times \mathbb{R}^+ \mid \alpha_1 > \alpha_2\}$ and $\Theta_b = \{(\alpha_1, \alpha_2) \in \mathbb{R}^+ \times \mathbb{R}^+ \mid \alpha_1 < \alpha_2\}$ (see Figure 2.8). In order to test the hypothesis

H , we compute the probability

$$\begin{aligned} & \mathbb{P}((\alpha_1, \alpha_2) \in \Theta_b \mid \mathbf{x}_1, \mathbf{x}_2) \\ &= \int_{\alpha_1 < \alpha_2} \int_{\mathbb{R}^+ \times \mathbb{R}^+} g_{1,1}(\alpha_1, \beta_1 \mid \mathbf{x}_1) \cdot g_{1,2}(\alpha_2, \beta_2 \mid \mathbf{x}_2) \, d\beta_1 d\beta_2 \, d\alpha_1 \, d\alpha_2 \\ &= \int_{\alpha_1 < \alpha_2} \prod_{i=1}^2 k_i \cdot g_{0,i}(\alpha_i) \cdot \frac{\Gamma(n_i \alpha_i)}{\Gamma(\alpha_i)^{n_i}} \cdot \left(\frac{m_{g_i}}{n_j \bar{x}_i} \right)^{n_i \alpha_i} \, d\alpha_1 d\alpha_2 = 0.311 \end{aligned}$$

and, since $\delta_H = 0.378$, we do not reject H .

2.4 Comparison with the FBST

In this section we present a comparison of the FBST as presented in Pereira and Stern, 2020, which provides an overview of the e -value (see Section A.3.2).

2.4.1 Similarities and differences between the procedures

The most striking similarity between the FBST and the BDT is that both tests, fully accepting the Likelihood Principle¹, see Birnbaum, 1962, and relying on the posterior distribution of the parameter $\theta \in \Theta$, are true Bayesian procedures. Furthermore, since the FBST and the BDT do not demand the adoption of a prior distribution which assigns positive probability for the subset Θ_H , as the *Bayes Factor* does, they do not fall into *Jeffreys-Lindley's paradox*.

Another important similarity is that, asymptotically, both tests lead to the rejection of the hypothesis H when it is false (i.e. when we test $\theta_H \neq \theta^*$ where θ^* is the true value of the parameter). On the contrary, if $\theta_H = \theta^*$ they have a different asymptotic behaviour (see Proposition 2.3 for the BDM and Section A.3.2 for the e -value).

Certainly, the FBST has a more general reach than the BDT. Indeed, it examines the entire class of sharp hypotheses, whereas the extension of the BDT to such hypotheses is not straightforward and, currently, is limited to considering the subclass of the hy-

¹de Finetti, 1979 states that “*the likelihood principle ... simply states that the information available from any set of observations is entirely contained in the corresponding likelihood function. Since this is, in fact, the factor which transforms the prior opinion into the posterior, this is all we require and, indeed, all we can ask for.*”

potheses expressed as $H : \varphi = \varphi_H$ that are able to partition the parameter space Θ as $\{\Theta_a, \Theta_H, \Theta_b\}$. Moreover, notice that while the integration sets Θ_a and Θ_b are determined exclusively by the hypothesis, the tangential set \bar{T} depends on the hypothesis, the posterior density and the choice of the reference function. It is questionable, on the other hand, whether the e -value is as easily computable as the BDM is in cases where the parameter space has dimension higher than 1.

Unlike the BDM, the elimination of nuisance parameters is not recommended when using the e -value. In fact, this measure is not invariant with respect to marginalisations of the nuisance parameter and the use of marginal densities to construct credible sets may produce inconsistency.

It is easy to see that one can create an analogy between the p -value, the e -value and δ_H . Regarding frequentist p -values, the sample space is ordered according to increasing inconsistency with the assumed null hypothesis H . The FBST, instead, orders the parameter space according to increasing inconsistency with the assumed null hypothesis H based, in the first formulation, on the posterior probability density and, in the second one, on the concept of statistical surprise. In the same way, it can be seen that the probability in (2.8) has to do with the posterior probability of exceeding θ_H in a direction in contrast with the data (namely, the side where there is more posterior probability).

Another similarity occurs when considering the reference density $r(\theta)$ as the (possibly improper) uniform density, since the first and second definitions of evidence define the same tangent set, i.e. the HRSS and the HPDS coincide. Then, for a scalar parameter θ , since the BDM is linked to the equi-tailed credible regions while the e -value is linked to the HPDS, we have that if:

- $g_1(\theta|\mathbf{x})$ is symmetric and unimodal, then $\bar{e}v(H) = \delta_H$;
- $g_1(\theta|\mathbf{x})$ is asymmetric and unimodal (for instance with positive skewness) and $m_1 < \theta_H$ [$\theta_H < m_1$], then $\bar{e}v(H) > \delta_H$ [$\bar{e}v(H) < \delta_H$]. When $m_1 = \theta_H$ we have $0 = \delta_H < \bar{e}v(H)$.

Simulation study

In order to determine the resulting false-positive rates of both the FBST and the BDT, we conduct a simulation study for specific sample sizes.

Let $\mathbf{x} = (x_1, \dots, x_n)$ be an *iid* sample of size n from the exponential distribution $X \sim \text{Exp}(x|1/\theta^*)$, with $\theta^* = 1.2$. We are interested in testing the hypothesis $H : \theta_H = \theta^* = 1.2$. Assuming a Jeffreys' prior $g_0(\theta) \propto \theta^{-1}$, the posterior distribution is an *InvGamma* $(\theta|n, \sum x_i)$; see Example 2.4.

Table 2.1 shows the simulation results for three different values of the threshold $\omega = \{0.90, 0.95, 0.99\}$, for $S = 50000$ simulations and $D = 50000$ posterior draws. Across the different sample sizes considered, the false-positive rates are very similar for both tests and, as we expect since we are using objective priors (see Bayarri and Berger, 2004), they are close to the error of the first type $\alpha = \{0.10, 0.05, 0.01\}$, related to ω . Similar results, not reported here, were found adopting a Poisson model.

	$\omega = 0.90$				$\omega = 0.95$				$\omega = 0.99$			
	n				n				n			
	10	100	1000	10000	10	100	1000	10000	10	100	1000	10000
<i>e-value</i> $r(\theta) \propto 1$	0.102	0.100	0.099	0.098	0.052	0.050	0.051	0.049	0.011	0.010	0.011	0.011
<i>e-value</i> $r(\theta) = g_0(\theta)$	0.101	0.102	0.100	0.098	0.051	0.050	0.051	0.049	0.010	0.010	0.011	0.011
δ_H	0.103	0.102	0.101	0.099	0.053	0.049	0.052	0.049	0.010	0.009	0.011	0.011

Table 2.1: False positive rates for different sample sizes n and different thresholds ω .

Some Examples

In order to compare the BDM and the *e-value*, let us consider different situations and then examine the results.

Example 2.11. (Continuation of Example 2.4)

As a first comparative scenario, consider the test performed in Example 2.4 in which $\theta_H = 2.4$ and additionally the case in which $\theta_H = 0.7$. Since the posterior $g_1(\theta|\mathbf{x})$ has a positive skewness and $m_1 < \theta_H = 2.4$ then $\overline{ev}(H) > \delta_H$, on the contrary, for $m_1 > \theta_H = 0.7$ then $\overline{ev}(H) < \delta_H$. Indeed, we find the results reported in Table 2.2.

The differences between the *e-value* and δ_H , which in this example appear to be modest, can actually become meaningful when the posterior has a greater asymmetry

	$\theta_H = 2.4$			$\theta_H = 0.7$		
	<i>e-value</i>		δ_H	<i>e-value</i>		δ_H
	$r(\theta) \propto 1$	$r(\theta) = g_0(\theta)$		$r(\theta) \propto 1$	$r(\theta) = g_0(\theta)$	
[A] $n = 6$	0.909	0.866	0.832	0.646	0.847	0.886
[B] $n = 12$	0.978	0.968	0.960	0.899	0.957	0.968
[C] $n = 24$	0.999	0.998	0.997	0.991	0.997	0.997

Table 2.2: The table shows, for the 3 different cases examined in Example 2.4, the values of δ_H and of the *e-value* considering, as a reference distribution, both a flat reference function and a Jeffreys' prior.

and heavy tails. In such case, comparing different hypotheses, the FBST always leads to favour the hypothesis with higher density. Moreover, the *e-value* may be more or less robust w.r.t. the position of θ_H , as it is highlighted in the example below.

Example 2.12. - Test on the mean of the Inverse Gaussian distribution

Consider a random variable X with Inverse Gaussian distribution $X \sim IG(x|\mu, \nu_0)$, $\mu \in \mathbb{R}^+$ and ν_0 known. Given an *iid* sample \mathbf{x} of size n , the likelihood function for μ is $L(\mu|\mathbf{x}) \propto \exp \left\{ -n\nu_0 \cdot \left(\frac{\bar{x}}{2\mu^2} - \frac{1}{\mu} \right) \right\}$. Adopting the Jeffreys' prior $g_0(\mu) \propto \frac{1}{\sqrt{\mu^3}}$, we obtain the posterior

$$g_1(\mu|\mathbf{x}) \propto \frac{1}{\sqrt{\mu^3}} \cdot \exp \left\{ -n\nu_0 \cdot \left(\frac{\bar{x}}{2\mu^2} - \frac{1}{\mu} \right) \right\}.$$

We are interested in testing the hypothesis $H : \mu = \mu_H$ and we consider a sample of size $n = 8$ for which $\bar{x} = 4.2$ and $m_1 = 4.483$. For $\nu_0 = 5$, we choose to test $H_A : \mu = 2.5$ and $H_B : \mu = 12$. The results of the analysis are displayed in Table 2.3 and Figure 2.9. If we choose $\omega = 0.95$ as a rejection threshold in both cases, and with both references, we are lead to opposite inferential conclusions.

Example 2.13. (Continuation of Examples 2.6, 2.7, 2.8)

Let us now compare the results obtained with the FBST and the BDT for the Examples 2.6, 2.7 and 2.8, when fixing a value of 0.95 as a rejection threshold.

The conclusions reached with the FBST and with the BDT for Example 2.6, which can be seen in Table 2.4, are the same (for both reference functions considered) although, in some cases, there are substantial differences between the values of the evidence mea-

	<i>e-value</i>		δ_H
	$r(\mu) \propto 1$	$r(\mu) = g_0(\mu)$	
$H_A : \mu = 2.5$	0.803	0.848	0.975
$H_B : \mu = 12$	1	1	0.907

Table 2.3: For the two different hypothesis examined in Example 2.12, the table shows δ_H and the *e-value* considering, as a reference distribution, both a flat reference function and a Jeffreys' prior.

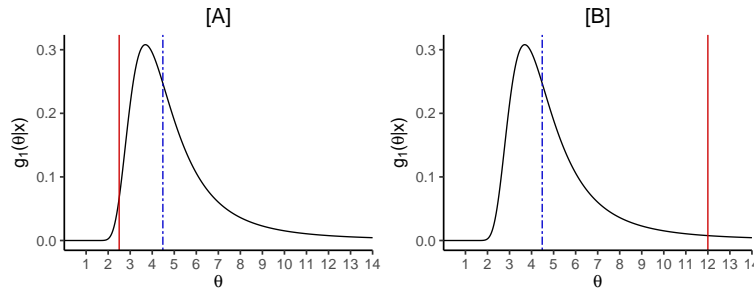


Figure 2.9: Posterior density function $g_1(\mu|\mathbf{x})$ associated to Example 2.12. In [A] we have $\mu_H = 2.5 < m_1$, while in [B] $\mu_H = 12 > m_1$.

sures. To summarise, the hypothesis H_B has to be rejected while not enough evidence is available for the rejection of the hypotheses H_A and H_C .

Moving on to Example 2.7 we can say that the analysis of the findings with the two different tests appears to be more complex than the previous one, see Table 2.5. In case [A], for both BDT and FBST with the flat reference function, there is not enough evidence to reject the hypothesis. On the contrary, if one considers the FBST with the Jeffreys' prior as reference function, one is led to reject this hypothesis. In case [B], by rejecting the hypothesis, the BDT is in agreement with the FBST with the Jeffreys' reference function in contrast to the FBST with the flat reference function for which there is not enough evidence to reject it.

Finally, in the case illustrated in Example 2.8, the conclusion reached with the FBST and with the BDT is the same (for both reference functions considered), i.e. there is not enough evidence to reject the hypothesis (see Table 2.6). It should be noted that, again, there are substantial differences between the values of the evidence measures.

The calculation of the FBST for a scalar parameter of interest without nuisance pa-

	<i>e-value</i>		δ_H
	$r(\boldsymbol{\theta}) \propto 1$	$r(\boldsymbol{\theta}) = g_0(\boldsymbol{\theta})$	
$H_A : \alpha = 2.5$	0.557	0.186	0.570
$H_B : \mu = 6$	0.984	0.963	0.976
$H_C : \sigma^2 = 2$	0.784	0.562	0.846


Table 2.4: The table shows the results of the Example 2.6 on the test on the shape parameter, mean and variance of the Gamma distribution. For the *e-value* we have considered, as a reference distribution, both a flat reference function and a Jeffreys' prior.

	<i>e-value</i>		δ_H
	$r(\boldsymbol{\theta}) \propto 1$	$r(\boldsymbol{\theta}) = g_0(\boldsymbol{\theta})$	
[A] $n = 10$	0.364	0.999	0.570
[B] $n = 40$	0.924	1	0.972

Table 2.5: The table shows the results of the Example 2.7 on the test of the coefficient of variation for a Normal distribution. For the *e-value* we have considered, as a reference distribution, both a flat reference function and a Jeffreys' prior.

	<i>e-value</i>		δ_H
	$r(\boldsymbol{\theta}) \propto 1$	$r(\boldsymbol{\theta}) = g_0(\boldsymbol{\theta})$	
$H : \gamma = 2$	0.650	0.691	0.844

Table 2.6: The table shows the results of the Example 2.8 on the test of the skewness coefficient of the Inverse Gaussian distribution. For the *e-value* we have considered, as a reference distribution, both a flat reference function and a Jeffreys' prior.

rameters, has been carried out through the function defined in the ‘fbst’ package Kelter, 2022 for . Instead, tangential sets \bar{T} and its integrals, for Examples 2.6, 2.7 and 2.8, were determined by means of the *Mathematica* software. Browsing through the code that leads to the calculation of these measures (see Manca, 2022), it is evident that more work is required for the calculation of the integration region related to the FBST. In this sense, the BDT appears to be easier to apply.

Chapter 3

The Bayesian Discrepancy Test for the comparison of two independent populations

In this section we propose a general procedure, based on the BDT, for testing hypotheses on the comparison of two parameters, or their transformations, from two independent populations. This approach is flexible, as it can be adapted to take into account different distributions and different parameter transformations. In addition, this methodology enables us to tackle problems that are not yet covered in the literature.

In what follows, we propose to use the BDM for comparing means, variances, coefficients of variation, skewness and correlation coefficients of two independent populations. The problem of comparing means and variances has been widely discussed in both the frequentist and the Bayesian fields. Whereas the problem of comparing coefficients of variation, although widely addressed in the frequentist paradigm, has had only few contributions in the Bayesian one (see Bertolino, Columbu, Manca, and Musio, 2022 for all the references). Additionally, to the best of our knowledge, the case of the comparison of two skewness indexes treated here is completely original and has not been addressed before.

3.1 The procedure

Generalising the notation of the previous chapter, let us now fix that $k = 1$ and consider two independent parametric models where each one has density $f(x_\ell|\boldsymbol{\theta}_\ell)$, $\ell = 1, 2$. Denoting with $g_{0,\ell}(\boldsymbol{\theta}_\ell)$ a prior density, for a sample $\boldsymbol{x}_\ell = \{x_{\ell_1}, \dots, x_{\ell_{n_\ell}}\}$ of n_ℓ iid observations the corresponding posterior density is

$$g_{1,\ell}(\boldsymbol{\theta}_\ell|\boldsymbol{x}_\ell) \propto g_{0,\ell}(\boldsymbol{\theta}_\ell)L_\ell(\boldsymbol{\theta}_\ell|\boldsymbol{x}_\ell), \quad \ell = 1, 2.$$

Indicate with $\boldsymbol{\theta}$ the joint parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \Theta_1 \times \Theta_2 = \Theta \subseteq \mathbb{R}^p \times \mathbb{R}^p$ of the two populations parameters. We are interested in testing the hypothesis

$$H : \varphi_1 = \varphi_2, \tag{3.1}$$

with $\varphi_\ell = \gamma(\boldsymbol{\theta}_\ell)$, $\ell = 1, 2$, being the parameter of interest, which identify the partition $\{\Theta_a, \Theta_H, \Theta_b\}$ of Θ where

$$\begin{aligned} \Theta_a &= \{\boldsymbol{\theta} \in \Theta \mid \varphi_1 < \varphi_2\}, & \Theta_b &= \{\boldsymbol{\theta} \in \Theta \mid \varphi_1 > \varphi_2\}, \\ \Theta_H &= \{\boldsymbol{\theta} \in \Theta \mid \varphi_1 = \varphi_2\}. \end{aligned}$$

To perform the test we can then express the BDM as

$$\delta_H = 1 - 2 \cdot \min_{a,b} \{\mathbb{P}(\boldsymbol{\theta} \in \Theta_a|\boldsymbol{x}_1, \boldsymbol{x}_2), \mathbb{P}(\boldsymbol{\theta} \in \Theta_b|\boldsymbol{x}_1, \boldsymbol{x}_2)\}, \tag{3.2}$$

where

$$\mathbb{P}(\boldsymbol{\theta} \in \Theta_j \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = \int_{\Theta_j} g_{1,1}(\boldsymbol{\theta}_1|\boldsymbol{x}_1) g_{1,2}(\boldsymbol{\theta}_2|\boldsymbol{x}_2) d\boldsymbol{\theta}_1 d\boldsymbol{\theta}_2, \quad j = a, b. \tag{3.3}$$

The evaluation of these probabilities demands the computation of multidimensional integrals that may not be possible to solve in a closed form. In such cases they can be approximated using the Monte Carlo Integration method.


In the particular case in which the marginal posterior of the parameter of interest $\varphi_\ell \in \Xi_\ell \subseteq \mathbb{R}$ can be computed in a closed form, as we have seen in Section 2.2, then the hypothesis $H : \varphi_1 = \varphi_2$ induces a partition on the marginal parameter space $\Xi =$

$\Xi_1 \times \Xi_2 \subseteq \mathbb{R}^2$, where Ξ_ℓ is the marginal parameter space associated to the ℓ -th population that can be directly used to compute the BDM.

With an appropriate change of notation, this procedure can be naturally extended to the case in which $k > 1$.

3.2 Illustrative examples

We now consider a variety of distributions and parameters transformations and briefly outline the case-by-case tools needed to calculate the BDM. In particular, we specify the posterior distributions and the subsets of the parameter space that are needed to compute the integrals specified in (3.3). In all the cases discussed, a Jeffreys' prior has been adopted.

Most of the  codes for these examples can be found online in https://github.com/maramanca/Chapter3_Example_Thesis.

3.2.1 Skewness coefficients of two Inverse Gaussian populations

Let us consider two independent Inverse Gaussian random variables $X_\ell \sim IG(x_\ell | \mu_\ell, \lambda_\ell)$, $\ell = 1, 2$, i.e.

$$f(x_\ell | \mu_\ell, \lambda_\ell) = \sqrt{\frac{\lambda_\ell}{2\pi x_\ell^3}} \exp \left\{ -\frac{1}{2} \lambda_\ell \left(\frac{x_\ell - \mu_\ell}{\mu_\ell \sqrt{x_\ell}} \right)^2 \right\},$$

where $X_\ell \in \mathbb{R}^+$ and $(\mu_\ell, \lambda_\ell) \in \mathbb{R}^+ \times \mathbb{R}^+$. Given n_ℓ observations $\mathbf{x}_\ell = (x_{\ell,1}, \dots, x_{\ell,n_\ell})$, the posterior distribution of the parameter vector with non-informative prior $g_0(\mu_\ell, \lambda_\ell) \propto \frac{1}{\sqrt{\mu_\ell^3 \lambda_\ell}}$ is

$$g_{1,\ell}(\mu_\ell, \lambda_\ell | \mathbf{x}_\ell) \propto \sqrt{\frac{\lambda_\ell^{n_\ell-1}}{\mu_\ell^3}} \exp \left\{ -\frac{n_\ell \lambda_\ell}{2} \left(\frac{\bar{x}_\ell}{\mu_\ell^2} - \frac{2}{\mu_\ell} + \frac{1}{a_\ell} \right) \right\}, \quad \ell = 1, 2, \quad (3.4)$$

where \bar{x}_ℓ and a_ℓ are the arithmetic and harmonic means respectively. We are interested in comparing the skewness of the two populations. We recall that the skewness of a real-valued random variable is defined as the third standardized moment. In this case it

assumes the expression

$$\varphi_\ell = 3\sqrt{\frac{\mu_\ell}{\lambda_\ell}}.$$

The hypothesis $H : \varphi_1 = \varphi_2$ identifies in the parameter space Θ the subsets

$$\Theta_j = \left\{ (\mu_1, \lambda_1, \mu_2, \lambda_2) \in \mathbb{R}_+^4 \mid \mu_1 \lambda_2 \leq \mu_2 \lambda_1 \right\}, \quad j = a, b.$$

Therefore, the BDM can be computed by means of formula 3.2. It is worth to notice that, for this distribution, the measure of skewness is three times the coefficient of variation. A test on the comparison of two skewness indices is then equivalent to testing the equality of two coefficients of variation.

We present now an application to real data for the comparison of skewness indexes of two Inverse Gaussian populations.

Example 3.1. In the engineering context, river flooding rates have significant economic, social, political implications. The modelling and analysis of such data is an important application of extreme value theory. We consider a study on the Floyd Ever Flood rate data for the years 1935–1954. A more extended version of this dataset was analysed in Mudholkar and Hutson, 1996 where it is shown that the *exponentiated Weibull* model is appropriate for such data, more in general for extreme value analysis. The Inverse Gaussian distribution belongs to this family. With this kind of data it is of interest to compare the asymmetry of the two populations.

In the years 1935–1944 we have that $n_1 = 10$ and with a sample skewness coefficient equal to 2.19 while in the second group, for the years 1945–1954, $n_2 = 10$ and sample skewness coefficient 5.20. As explained in the previous section, the computation of δ_H depends on the evaluation of two four-dimensional integrals (see (3.3)), that can be approximated through the Monte Carlo integration method. Furthermore, since the posterior distribution is known up to a normalizing constant, it is not possible to sample directly from it. We, therefore, applied the random walk Metropolis-Hasting algorithm using a bivariate Normal proposal distribution. In order to obtain a chain that converges to the target distribution, $9 \cdot 10^5$ samples were generated and reduced to $1.2 \cdot 10^5$ after considering a burn-in period and a chain thinning. We, finally, found that the discrepancy measure is $\delta_H \approx 1$, then we can reject the hypothesis of the skewness indexes equality.

3.2.2 Means of two Inverse Gaussian populations

We consider two independent Inverse Gaussian random variables as in section 3.2.1. Now, we are interested in comparing the mean of the two populations. Starting from formula (3.4) we compute the marginal posterior

$$\begin{aligned}
 \mu_\ell \mid \mathbf{x}_\ell &\sim h_{1,\ell}(\mu_\ell \mid \mathbf{x}_\ell) \\
 &= \int g_1(\mu_\ell, \lambda_\ell \mid \mathbf{x}_\ell) d\lambda_\ell \\
 &\propto \sqrt{\frac{1}{\mu_\ell^3}} \cdot \left(\frac{\bar{x}_\ell}{\mu_\ell^2} - \frac{2}{\mu_\ell} + \frac{1}{a_\ell} \right)^{-\frac{n_\ell+1}{2}}.
 \end{aligned} \tag{3.5}$$

In such case, $\varphi_\ell = \mu_\ell$ and the hypothesis $H : \mu_1 = \mu_2$ identifies in the marginal parameter space Ξ the partition

$$\begin{aligned}
 \Xi_a &= \{(\mu_1, \mu_2) \in \mathbb{R}^+ \times \mathbb{R}^+ \mid \mu_1 < \mu_2\}, \\
 \Xi_H &= \{(\mu_1, \mu_2) \in \mathbb{R}^+ \times \mathbb{R}^+ \mid \mu_1 = \mu_2\}, \\
 \Xi_b &= \{(\mu_1, \mu_2) \in \mathbb{R}^+ \times \mathbb{R}^+ \mid \mu_1 > \mu_2\}.
 \end{aligned}$$

Then, after computing

$$\mathbb{P}((\mu_1, \mu_2) \in \Xi_j \mid \mathbf{x}_1, \mathbf{x}_2) = \int_{\Xi_j} \prod_{\ell=1}^2 h_{1,\ell}(\mu_\ell \mid \mathbf{x}_\ell) d\mu_1 d\mu_2, \quad j = a, b,$$

we can evaluate the BDM as

$$\delta_H = 1 - 2 \cdot \min_{a,b} \left\{ \mathbb{P}((\mu_1, \mu_2) \in \Xi_a \mid \mathbf{x}_1, \mathbf{x}_2), \mathbb{P}((\mu_1, \mu_2) \in \Xi_b \mid \mathbf{x}_1, \mathbf{x}_2) \right\}.$$

Example 3.2. We apply this procedure to a dataset collected for a study on the Hodgkin's disease (see Chhikara and Folks, 1989). Two groups of patients with active ($n_1 = 17$) and inactive ($n_2 = 28$) Hodgkin's disease are compared w.r.t. the level of plasma bradykininogen. The outcome variable is measured in micrograms of bradykininogen per milliliter of plasma. The sampled values can be considered as coming from two inverse Gaussian distributions. In the group of patients with active disease we have that $\bar{x}_1 = 4.241$ while in the control group $\bar{x}_2 = 6.791$. The calculation of the BDM for the

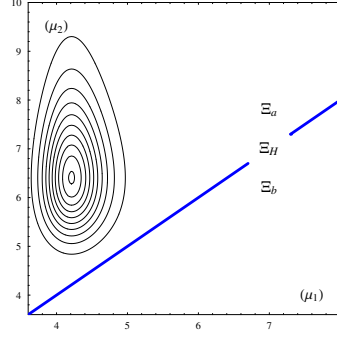


Figure 3.1: Contour plot for the posterior marginal distribution in Example 3.2.

equality of the means in the groups of hill and healthy patients yields $\delta_H = 0.995$, and we can reject the hypothesis. In Figure 3.1 are reported the contour lines of the joint posterior distribution of μ_1 and μ_2 in formula (3.5). The partitioning subsets of the marginal parameter space are also indicated. It is interesting to observe how, when comparing the same samples in terms of their coefficients of variation (see Bertolino, Columbu, Manca, and Musio, 2022), the two samples cannot be considered coming from different populations.

3.2.3 Variances of two Gamma populations

Consider two independent random variables X_ℓ , $\ell = 1, 2$, with distributions

$$X_\ell \sim \text{Gamma}(x_\ell | \alpha_\ell, \beta_\ell) = \frac{\beta_\ell^{\alpha_\ell}}{\Gamma(\alpha_\ell)} x_\ell^{\alpha_\ell - 1} e^{-\beta_\ell x_\ell}, \quad X_\ell \in \mathbb{R}^+, \quad (\alpha_\ell, \beta_\ell) \in \mathbb{R}^+ \times \mathbb{R}^+.$$

We are interested in performing a test on the equality of the variances, i.e. $\text{Var}(X_1) = \text{Var}(X_2)$ where $\text{Var}(X_\ell) = \frac{\alpha_\ell}{\beta_\ell^2} = \varphi_\ell$. The posterior distribution of the parameter vector, after assuming the Jeffreys' prior (see Yang and Berger, 1996), takes the expression

$$\begin{aligned} g_{1,\ell}(\alpha_\ell, \beta_\ell | \mathbf{x}_\ell) &= k_\ell \cdot g(\alpha_\ell, \beta_\ell) \cdot L(\alpha_\ell, \beta_\ell | \mathbf{x}_\ell) \\ &\propto \frac{1}{\beta_\ell} \sqrt{\alpha_\ell \psi^{(1)}(\alpha_\ell) - 1} \left(\frac{\beta_\ell^{\alpha_\ell}}{\Gamma(\alpha_\ell)} g^{\alpha_\ell} e^{-\bar{x}_\ell \beta_\ell} \right)^n \end{aligned}$$

with k_ℓ the normalising constant. The hypothesis $H : \varphi_1 = \varphi_2$ identifies on the parameter space Θ the subsets

$$\Theta_j = \left\{ (\alpha_1, \beta_1, \alpha_2, \beta_2) \in \mathbb{R}_+^2 \times \mathbb{R}_+^2 \mid \alpha_1 \beta_2^2 \leq \alpha_2 \beta_1^2 \right\}, \quad j = a, b.$$

Therefore, the BDM can be computed by means of formula 3.2.

Example 3.3. We consider a study on the hydro-geologic effects of rocks fracturing, associated with ground water yields in southwestern Virginia (see Wright, 1985). It was found that wells in valleys subjected to fractured rocks, produce larger quantities of water than wells in valleys subjected to unfractured rocks. The comparison was made on the water accumulation (expressed in gal/min/ft) from $n_1 = 13$ wells in the first type of valleys and $n_2 = 12$ in the second one. In this application, being aware of the difference in terms of their means, we are interested in studying if the two populations are different also with respect to their variability. In Chang et al., 2011 was observed that the assumption of the Gamma distribution is appropriate for the two populations. In the first group the sample variance was $s_1^2 = 0.098$ while in the second one was $s_2^2 = 0.079$.

Since the posterior distribution is known up to a normalizing constant, we used a MCMC simulation technique with a bivariate Normal proposal distribution. The convergence of the chain was obtained from $2.1 \cdot 10^5$ samples that were reduced to $5 \cdot 10^4$ after considering a burn-in period and a chain thinning. The BDM is $\delta_H = 0.185$ and therefore the hypothesis cannot be rejected.

3.2.4 Coefficients of variation of two independent populations

The case of two Normal populations

Consider two independent Gaussian random variables $X_\ell \sim N(x_\ell | \mu_\ell, \phi_\ell^{-1})$, with $X_\ell \in \mathbb{R}$ and mean and precision $(\mu_\ell, \phi_\ell) \in \mathbb{R} \times \mathbb{R}^+$, for $\ell = 1, 2$. Assuming the non-informative Jeffrey's priors

$$(\mu_\ell, \phi_\ell) \sim g_{0,\ell}(\mu_\ell, \phi_\ell) \propto \phi_\ell^{-1}, \quad \ell = 1, 2,$$

and given n_ℓ observations with sample means \bar{x}_ℓ and sample standard deviations s_ℓ , it is known that the posterior distributions of the parameter vectors are Normal Gamma

$$(\mu_\ell, \phi_\ell) \mid \mathbf{x} \sim NG(\mu_\ell, \phi_\ell \mid \eta_\ell, \nu_\ell, \alpha_\ell, \beta_\ell), \quad \ell = 1, 2,$$

with hyperparameters $\eta_\ell = \bar{x}_\ell$, $\nu_\ell = n_\ell$, $\alpha_\ell = \frac{1}{2}(n_\ell - 1)$, $\beta_\ell = \frac{1}{2}n_\ell s_\ell^2$.

The hypothesis $H : \varphi_1 = \varphi_2$, where $\varphi_\ell = \frac{1}{|\mu_\ell|\sqrt{\phi_\ell}}$, identifies in the parameter space Θ the subsets

$$\Theta_a = \{(\mu_1, \phi_1, \mu_2, \phi_2) \in \mathbb{R}^2 \times \mathbb{R}_+^2 \mid |\mu_1|\sqrt{\phi_1} > |\mu_2|\sqrt{\phi_2}\},$$

$$\Theta_H = \{(\mu_1, \phi_1, \mu_2, \phi_2) \in \mathbb{R}^2 \times \mathbb{R}_+^2 \mid |\mu_1|\sqrt{\phi_1} = |\mu_2|\sqrt{\phi_2}\},$$

$$\Theta_b = \{(\mu_1, \phi_1, \mu_2, \phi_2) \in \mathbb{R}^2 \times \mathbb{R}_+^2 \mid |\mu_1|\sqrt{\phi_1} < |\mu_2|\sqrt{\phi_2}\}.$$

The BDM (as seen in (3.2)) requires the computation of

$$\mathbb{P}\left((\mu_1, \phi_1, \mu_2, \phi_2) \in \Theta_j \mid \mathbf{x}\right) = \int_{\Theta_j} \prod_{\ell=1}^2 g_1^\ell(\mu_\ell, \phi_\ell \mid \eta_\ell, \nu_\ell, \alpha_\ell, \beta_\ell) \, \mathrm{d}\mu_\ell \, \mathrm{d}\phi_\ell,$$

where $j = a, b$ and $g_{1,\ell}$ is the Normal-Gamma density. For this setting we present a small simulation study to assess the false-positive rates of the BDT with varying sample sizes and also an application to real data.

Simulation study

We conducted a simulation study for evaluating the false rejection rate in repeated sampling when comparing CVs of two independent populations with the same distribution assumptions. Let $\mathbf{x}_1 = (x_{1,1}, \dots, x_{1,n_1})$ and $\mathbf{x}_2 = (x_{2,1}, \dots, x_{2,n_2})$ be two *iid* samples of size, respectively, n_1 and n_2 both taken from the Normal distribution $X \sim N(x \mid \mu, \phi^{-1})$. We set $\mu = \mu_1 = \mu_2 = 3$ and $\phi = \phi_1 = \phi_2 = 1$ to have that $\varphi_1 - \varphi_2 = 0$. We are interested in testing the hypothesis of equal CVs under this scenario. In Table 3.2.4 are reported the rate of rejection values under the hypothesis of equal CVs. For three thresholds $\omega = \{0.90, 0.95, 0.99\}$ we have considered $S = 50000$ simulations and computed the BDM by taking $D = 10000$ posterior draws. The false positive rates obtained over S resamplings are in agreement with the nominal accuracy typical of the first type error

	$\omega = 0.90$				$\omega = 0.95$				$\omega = 0.99$			
n_1	10	10	100	1000	10	10	100	1000	10	10	100	1000
n_2	10	50	100	1000	10	50	100	1000	10	50	100	1000
FPR	0.096	0.105	0.1	0.098	0.047	0.054	0.05	0.049	0.009	0.011	0.01	0.01

Table 3.1: False positive rates (FPR) for different sample sizes n_1 and n_2 and different thresholds ω .

$\alpha = \{0.10, 0.05, 0.001\}$ that could be associated to ω . This result is not surprising given the choice of objective priors (see Bayarri and Berger, 2004 and Hartigan, 1966 for other choices of the prior).

Example 3.4. Anthropometric measures in Sardinian population

We consider a set of anthropometric measures concerning the Sardinian population. The sample consists of 280 individuals of both sexes (140 males and 140 females) aged 20–25, that was collected between 1995–1998. We focus on the CVs comparisons among men and women. The same data were presented and analysed in Marini et al., 2005, where the bootstrap test for the difference of CVs developed in Cabras et al., 2006 was applied to evaluate sexual dimorphism. Among the 20 measurements in the dataset, we applied the BDT to a subsample of 10 which can be assumed to be normally distributed. In Table 3.2.4 are reported the principal descriptive statistics together with the values of the discrepancy measure associated to each anthropometric dimension considered. Based on the BDM we conclude that only one hypothesis can be rejected, that is the equality between the coefficients of variation for men's and women's skinfolds triceps. The final conclusions go in the same direction as Marini et al., 2005.

	Men				Women				δ_H
	Mean	SD	n_1	CV	Mean	SD	n_2	CV	
<i>Weight</i>	67.22	8.46	140	0.126	53.71	7.59	140	0.141	0.812
Breadths									
<i>Cephalic</i>	15.10	0.64	141	0.042	14.53	0.58	172	0.040	0.550
<i>Elbow</i>	7.02	0.39	103	0.056	6.01	0.35	117	0.058	0.355
Circumferences									
<i>Midarm relaxed</i>	26.91	2.60	139	0.097	23.47	2.01	134	0.086	0.831
<i>Midarm tensed</i>	30.83	2.74	139	0.089	25.28	2.15	133	0.085	0.388
Skinfolds									
<i>Biceps</i>	4.10	1.79	137	0.437	6.08	2.51	133	0.413	0.420
<i>Triceps</i>	7.76	3.76	140	0.485	13.33	4.78	140	0.359	<u>0.996</u>
<i>Subscapular</i>	10.34	3.78	137	0.366	12.71	4.53	140	0.356	0.213
<i>Suprailiac</i>	9.23	4.34	140	0.470	10.21	4.48	140	0.439	0.507
<i>Abdominal</i>	12.15	6.52	97	0.537	12.77	5.74	111	0.449	0.848

Table 3.2: Dispersion dimorphism in a set of anthropometric dimensions and relative BDM. Weight is expressed in kg; skinfolds in mm; all other measurements in cm.

The case of two Skew Normal populations

Consider two independent independent Skew Normal random variables $X_1, X_2 \in \mathbb{R}$, i.e. $X_\ell \sim SN(x_\ell | \mu_\ell, \sigma_\ell, \lambda_\ell)$, $\ell = 1, 2$ with density

$$\begin{aligned} f(x_\ell | \mu_\ell, \sigma_\ell, \lambda_\ell) &= \frac{2}{\sigma_\ell} \phi\left(\frac{x_\ell - \mu_\ell}{\sigma_\ell}\right) \Phi\left(\lambda_\ell \frac{x_\ell - \mu_\ell}{\sigma_\ell}\right) \\ &= \frac{2}{\sigma_\ell \sqrt{2\pi}} e^{-\frac{(x_\ell - \mu_\ell)^2}{2\sigma_\ell^2}} \int_{-\infty}^{\lambda_\ell \frac{x_\ell - \mu_\ell}{\sigma_\ell}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t_\ell^2}{2}} dt_\ell, \end{aligned}$$

where ϕ is the standard Normal probability density function and Φ is its cumulative distribution function. The location and shape parameter are $\mu_\ell, \lambda_\ell \in \mathbb{R}$ while $\sigma_\ell \in \mathbb{R}^+$ is the scale parameter; hence, $\theta_\ell = (\mu_\ell, \sigma_\ell, \lambda_\ell) \in \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}$ is the ℓ -th parameter vector.

Given $\delta_\ell = \frac{\lambda_\ell}{\sqrt{\lambda_\ell^2 + 1}}$, the expected value is $E(X_\ell) = \mu_\ell + \sigma_\ell \delta_\ell \sqrt{\frac{2}{\pi}}$ and the variance is

$Var(X_\ell) = \sigma_\ell^2 \left(1 - \frac{2\delta_\ell^2}{\pi}\right)$. Therefore, the coefficient of variation can be expressed as

$$CV(X_\ell) = \varphi_\ell = \sqrt{\sigma_\ell^2 \left(1 - \frac{2\delta_\ell^2}{\pi}\right)} / \left| \mu_\ell + \sigma_\ell \delta_\ell \sqrt{\frac{2}{\pi}} \right|.$$

The ℓ -th Jeffreys prior is proportional to

$$\boldsymbol{\theta}_\ell \sim g_{0,\ell}(\boldsymbol{\theta}_\ell) \propto \frac{1}{\sigma_\ell} \tilde{g}_{0,\ell}(\lambda_\ell),$$

where $g_{0,\ell}(\lambda_\ell)$, the prior distribution of the shape parameter, is a generalized t-Student distribution with location parameter $\mu_\ell = 0$, scale parameter $\sigma_\ell = \frac{1}{2}\pi$ and degrees of freedom $\nu_\ell = \frac{1}{2}$ (see Bayes and Branco, 2007), i.e.

$$\tilde{g}_{0,\ell}(\lambda_\ell) = \frac{1}{B\left(\frac{\nu_\ell}{2}, \frac{1}{2}\right)} \sqrt{\frac{\sigma_\ell}{\nu_\ell}} \left(1 + \sigma_\ell \frac{(\lambda_\ell - \mu_\ell)^2}{\nu_\ell}\right)^{-\frac{\nu_\ell+1}{2}}.$$

The posterior distributions of the parameter vectors is then

$$g_{1,\ell}(\boldsymbol{\theta}_\ell | \mathbf{x}_\ell) \propto \frac{1}{\sigma_\ell^{n_\ell+1}} g_{0,\ell}(\lambda_\ell) \prod_{i=1}^{n_\ell} \phi\left(\frac{x_{i\ell} - \mu_\ell}{\sigma_\ell}\right) \Phi\left(\lambda_\ell \frac{x_{i\ell} - \mu_\ell}{\sigma_\ell}\right).$$

The hypothesis $H : \varphi_1 = \varphi_2$ identifies in the parameter space Θ the subsets

$$\begin{aligned} \Theta_a &= \left\{ \boldsymbol{\theta} \in \mathbb{R}^4 \times \mathbb{R}_+^2 \mid \varphi_1 < \varphi_2 \right\}, \\ \Theta_H &= \left\{ \boldsymbol{\theta} \in \mathbb{R}^4 \times \mathbb{R}_+^2 \mid \varphi_1 = \varphi_2 \right\}, \\ \Theta_b &= \left\{ \boldsymbol{\theta} \in \mathbb{R}^4 \times \mathbb{R}_+^2 \mid \varphi_1 > \varphi_2 \right\}, \end{aligned}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is the joint parameter vector.

Example 3.5. We compute the BDM to compare CVs of gene expression levels between two groups of stage 4 high risk Neuroblastoma patients. How shown in Hossain and Beyene, 2015 the possible asymmetry in the distribution of expression levels can justify the assumption of a Skew Normal distribution, which may be a better fit for the expression profiles than the usual considered Normal distribution. The interest is to investigate

	Short survivors			Long survivors			δ_H
	Mean	SD	CV	Mean	SD	CV	
<i>hsa-let-7g</i>	5.10	1.25	0.245	5.04	0.87	0.173	0.733
<i>hsa-miR-24</i>	7.34	1.14	0.156	8.16	1.55	0.190	0.757
<i>hsa-miR-455-3p</i>	2.87	0.89	0.310	3.38	1.16	0.342	0.386
<i>hsa-miR-485-5p</i>	3.55	0.97	0.273	3.89	1.13	0.291	0.282

Table 3.3: Mean, standard deviation, coefficients of variation and δ_H of 4 differentially expressed miRNAs in the Neuroblastoma study (Example 4).

the possible separation between two survivor profiles: short survivors (death within 36 months from diagnosis) and long survivors (alive with an overall survival time > 36 months). The dataset contains the gene profiles of 31 patients of whom 17 were short survivors and 14 long survivors. Data is available online from the public genomics data repository Gene Expression Omnibus (GEO) and accessible through GEO Series accession number GSE16444. In this analysis we consider a small selection of 4 miRNAs among the 319 in the dataset. Also in this case we know the posterior distribution up to a normalizing constant, hence we used a simulation technique through the random walk Metropolis-Hasting algorithm with a bivariate Normal proposal distribution. In order to obtain a chain that converges to the target distribution, $1.04 \cdot 10^6$ samples were employed and reduced to $1.06 \cdot 10^5$ after considering a burn-in period and a chain thinning. In Table 3.3 are reported the sample CVs and relative BDM for their difference. For each of the miRNA analysed we cannot reject the hypotheses of equal CVs.

The case of two Negative Binomial populations

Let us consider two discrete Negative Binomial populations. It is known that, given $X_\ell | \lambda_\ell \sim \text{Pois}(x_\ell | \lambda_\ell)$ with $\lambda_\ell \sim \text{Gamma}(\lambda_\ell | \alpha_\ell, \beta_\ell)$, then the unconditional random variables X_ℓ follow a Negative Binomial distribution

$$X_\ell \sim \text{NB}\left(x_\ell \mid \frac{\beta_\ell}{\beta_\ell + 1}, \alpha_\ell\right), \quad \ell = 1, 2$$

with $(\alpha_\ell, \beta_\ell) \in \mathbb{R}^+ \times \mathbb{R}^+$. Its expected value, variance and coefficient of variation are

$$\begin{aligned} E(X_\ell) &= \frac{\alpha_\ell}{\beta_\ell}, \\ \text{Var}(X_\ell) &= \alpha_\ell \frac{\beta_\ell + 1}{\beta_\ell^2}, \\ \text{CV}(X_\ell) &= \varphi_\ell = \sqrt{\frac{\beta_\ell + 1}{\alpha_\ell}}. \end{aligned}$$

Assuming the Jeffreys prior for λ_ℓ (see Yang and Berger, 1996)

$$g_{0,\ell}(\alpha_\ell, \beta_\ell) \propto \frac{1}{\beta_\ell} \sqrt{\alpha_\ell \psi^{(1)}(\alpha_\ell) - 1}, \quad \ell = 1, 2,$$

where $\psi^{(1)}(\alpha_\ell) = \sum_{j=0}^{\infty} (\alpha_\ell + j)^{-2}$ is the PolyGamma function, the posterior distributions of the parameter vectors take the expressions

$$g_{1,\ell}(\alpha_\ell, \beta_\ell \mid \mathbf{x}_\ell) \propto \frac{\prod_i (x_{i\ell} + \alpha_\ell - 1)!}{[(\alpha_\ell - 1)!]^{n_\ell}} \left[\frac{\beta_\ell^{\alpha_\ell}}{(\beta_\ell + 1)^{\alpha_\ell + \bar{x}_\ell}} \right]^n \frac{1}{\beta_\ell} \sqrt{\alpha_\ell \psi^{(1)}(\alpha_\ell) - 1}.$$

The hypothesis identifies on the parameter space Θ the subsets

$$\begin{aligned} \Theta_a &= \left\{ (\alpha_1, \beta_1, \alpha_2, \beta_2) \in \mathbb{R}_+^2 \times \mathbb{R}_+^2 \mid \alpha_2(\beta_1 + 1) < \alpha_1(\beta_2 + 1) \right\}, \\ \Theta_H &= \left\{ (\alpha_1, \beta_1, \alpha_2, \beta_2) \in \mathbb{R}_+^2 \times \mathbb{R}_+^2 \mid \alpha_2(\beta_1 + 1) = \alpha_1(\beta_2 + 1) \right\}, \\ \Theta_b &= \left\{ (\alpha_1, \beta_1, \alpha_2, \beta_2) \in \mathbb{R}_+^2 \times \mathbb{R}_+^2 \mid \alpha_2(\beta_1 + 1) > \alpha_1(\beta_2 + 1) \right\}, \end{aligned}$$

Example 3.6. In the epidemiology context, when investigating the spreading dynamics of an infectious disease, it is often of interest to model the number of new infections (second cases) generated by an infectious subject. This kind of variable, in cases of individual variability in the transmission patterns, is overdispersed and right skewed (see Lloyd-Smith et al., 2005), therefore the negative binomial distribution is often considered for its analysis. During the Sars-Cov-2 (COVID-19) pandemic many national health systems have collected such kind of information considering contact tracing data to try their best in catching the diffusion pathways of the disease and the behaviours leading

to its increase. In particular, differences in individual contact patterns were a symptom of a superspreading phenomenon as was observed that a small number of infected could cause most of the secondary infections. Looking at the recent literature we used two published data on secondary cases of COVID-19 and compared their CVs through the BDM. The first dataset considered comes from data tracing in two Indian states (see Laxminarayan et al., 2020) from March to July 2020. It is a large dataset containing the offspring distribution for 88,527 cases, with sample mean 0.48 and variance 1.15 resulting in a CV of 2.218. The second one contains 290 cases from Hong Kong (see Adam et al., 2020) recorded from January to April 2020. In this case the observed mean is 0.58 whereas the variance 1.29 leading to a sample CV of 2.217. These two coefficients of variation are extremely close to each other and, as expected, we get to a small BDM of $\delta_H = 0.00972$. We therefore cannot reject the hypothesis on the equality of coefficients of variation between the India and the Hong Kong samples.

Also for this last model, the application of Metropolis-Hasting algorithms was required. In order to obtain a chain that converged to the target distribution, $1.3 \cdot 10^5$ samples were employed and reduced to 10^5 after considering a burn-in period.

3.2.5 Correlation coefficients of two Normal populations

We consider two Normal independent bivariate populations $\mathbf{X}_\ell \sim N_2(\mathbf{x}_\ell | \boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell)$ with $\mathbf{X}_\ell = (X_{\ell 1}, X_{\ell 2}) \in \mathbb{R}^2$ and $\ell = 1, 2$. Let $\boldsymbol{\mu}_\ell = E(\mathbf{X}_\ell) \in \mathbb{R}^2$ be the vector of means and let the covariance matrix be

$$\boldsymbol{\Sigma}_\ell = \text{Var}(\mathbf{X}_\ell) = \begin{pmatrix} \sigma_{\ell 1}^2 & \rho_\ell \sigma_{\ell 1} \sigma_{\ell 2} \\ \rho_\ell \sigma_{\ell 1} \sigma_{\ell 2} & \sigma_{\ell 2}^2 \end{pmatrix},$$

with $(\sigma_{\ell 1}^2, \sigma_{\ell 2}^2) \in \mathbb{R}_+^2$, and $\rho_\ell \in (-1, 1)$. We are interested in comparing the correlation coefficients, i.e. $\varphi_\ell = \rho_\ell$ and therefore test the hypothesis

$$H : \rho_1 = \rho_2. \tag{3.6}$$

Given n_ℓ observations for each of the two independent samples we assume the Jeffreys' prior for the parameters $(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell)$, corresponding to

$$g_{1,\ell}(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell) \propto g_{0,\ell}(\mu_{\ell 1}, \mu_{\ell 2}, \sigma_{\ell 1}^2, \sigma_{\ell 2}^2, \rho_\ell) \propto |\boldsymbol{\Sigma}_\ell|^{-1} = \frac{1}{\sigma_{1,\ell}^2 \sigma_{2,\ell}^2 (1 - \rho_\ell^2)},$$

$\ell = 1, 2$, and determine the joint posterior distribution $g_{1,\ell}(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell)$. By marginalizing out the nuisance parameters $(\mu_{1,\ell}, \mu_{2,\ell}, \sigma_{1,\ell}^2, \sigma_{2,\ell}^2)$ we obtain the marginal posterior distributions of the correlation coefficients

$$\begin{aligned} \rho_\ell \mid \hat{\boldsymbol{\mu}}_\ell, \hat{\boldsymbol{\Sigma}}_\ell &= \int g_{1,\ell}(\boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell \mid \hat{\boldsymbol{\mu}}_\ell, \hat{\boldsymbol{\Sigma}}_\ell) d\mu_{1,\ell} d\mu_{2,\ell} d\sigma_{1,\ell}^2 d\sigma_{2,\ell}^2 \\ &\propto \frac{1}{1 - \rho_\ell^2} \int \frac{1}{\sigma_{1,\ell}^2 \sigma_{2,\ell}^2} \cdot L(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) d\mu_{1,\ell} d\mu_{2,\ell} d\sigma_{1,\ell}^2 d\sigma_{2,\ell}^2, \end{aligned}$$

and finally, following Bernardo and Smith, 1994 example 5.27, we find that

$$\begin{aligned} \rho_\ell \mid r_\ell &\sim h_{1,\ell}(\rho_\ell \mid r_\ell) \\ &\propto \frac{(1 - \rho_\ell^2)^{\frac{1}{2}(n_\ell+1)}}{(1 - r_\ell \rho_\ell)^{n_\ell - \frac{1}{2}}} \cdot {}_2F_1\left(\frac{1}{2}, \frac{1}{2}; n_\ell - \frac{1}{2}; \frac{1}{2}(1 + r_\ell \rho_\ell)\right), \end{aligned}$$

which depends only on the sample correlation coefficient r_ℓ . With ${}_2F_1(a \cdot b; c; z)$ we indicate the hypergeometric function.

Hypothesis (3.6) identifies, in the marginal parameter space $\Xi = (-1, +1)^2$, the partition

$$\begin{aligned} \Xi_a &= \{(\rho_1, \rho_2) \in (-1, +1)^2 \mid \rho_1 < \rho_2\}, \\ \Xi_H &= \{(\rho_1, \rho_2) \in (-1, +1)^2 \mid \rho_1 = \rho_2\}, \\ \Xi_b &= \{(\rho_1, \rho_2) \in (-1, +1)^2 \mid \rho_1 > \rho_2\}, \end{aligned}$$

for which we have to compute the probabilities

$$\mathbb{P}((\rho_1, \rho_2) \in \Xi_j \mid \mathbf{x}_{\ell, n_\ell}) = \int_{\Xi_j} \prod_{\ell=1}^2 h_{1,\ell}(\rho_\ell \mid r_\ell) d\rho_1 d\rho_2, \quad j = a, b \quad \ell = 1, 2.$$

and therefore the BDM.

Example 3.7. In a study on animal rights and human behaviours a questionnaire survey

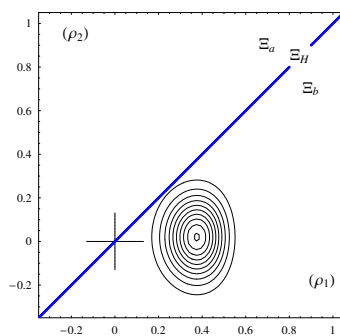


Figure 3.2: Contour plot for the posterior distribution in Example 3.7.

was conducted to measure ethical idealism, misanthropy and attitudes towards animal rights and animal research (Wuensch et al., 2002) in a group of college students. They were classified as being idealistic if their score on the idealism scale was greater than the median score, and non-idealistic otherwise. The relationship between misanthropy and support for animal rights was compared in these two groups. The observed sample correlations between misanthropy and support for animal rights are for the non-idealist $r_1 = 0.3814$ with $n_1 = 63$, and $r_2 = 0.0205$ with $n_2 = 91$ for the idealist.

In Figure 3.2 can be found the contour plots of the product $h_1^1(\rho_1 | r_1) \cdot h_1^2(\rho_2 | r_2)$ of the two posterior marginals. We have that $\delta_H = 0.977$, leading us to reject the hypothesis H of equal correlation in the two groups.

3.2.6 Regression coefficients

Consider a simple linear regression model

$$Y = \alpha + \beta x + \varepsilon, \quad (3.7)$$

with uncorrelated errors $\varepsilon \sim N(\cdot | 0, \phi^{-1})$, $\phi \in \mathbb{R}^+$. Given an *iid* sample of size n , let $\mathbf{y} = (y_1, y_2, \dots, y_n)^\top$ and

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix}^\top$$

be, respectively, the corresponding vector of observations and the regression matrix. The likelihood function is then

$$L(\boldsymbol{\beta}, \phi \mid \mathbf{X}, \mathbf{y}) \propto \phi^{\frac{1}{2}n} \cdot \exp \left\{ -\frac{1}{2}\phi (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\}. \quad (3.8)$$

If we denote with $\boldsymbol{\beta} = (\alpha, \beta)^\top$, and we adopt the Jeffreys prior

$$(\boldsymbol{\beta}, \phi) \sim g_0(\boldsymbol{\beta}, \phi) \propto \phi^{-1}, \quad (3.9)$$

the posterior distribution of $(\boldsymbol{\beta}, \phi)$ is Normal-Gamma, i.e.

$$\boldsymbol{\beta}, \phi \mid \mathbf{X}, \mathbf{y} \sim N_2 \left(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}, \hat{\phi}^{-1} (\mathbf{X}^\top \mathbf{X})^{-1} \right) \cdot \text{Gamma}(\phi \mid \hat{\zeta}, \hat{\lambda}), \quad (3.10)$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$, $\hat{\phi} = n \cdot (\mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X} \hat{\boldsymbol{\beta}})^{-1}$, $\hat{\zeta} = \frac{1}{2}(n-2)$ and $\hat{\lambda} = \frac{n}{2} \hat{\phi}^{-1}$ are the MLEs of the parameters. From (3.10) we find the marginal $h_1(\beta \mid \mathbf{X}, \mathbf{y})$ of the parameter β , i.e.

$$\beta \mid \mathbf{X}, \mathbf{y} \sim \text{StudentG} \left(\beta \mid \hat{\beta}, \hat{\lambda}, n-2 \right), \quad (3.11)$$

with

$$\hat{\beta} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}, \quad \hat{\lambda} = \frac{n}{\sum x_i^2} \cdot \frac{n \sum x_i^2 - (\sum x_i)^2}{\sum y_i^2 - \hat{\alpha} \sum y_i - \hat{\beta} \sum x_i y_i},$$

and $\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$. Suppose, now, to be interested on the hypothesis

$$H : \beta_1 = \beta_2, \quad (3.12)$$

on the correlation coefficients of two simple regression models as (3.7). After considering two *iid* samples

$$\text{data}_\ell = \{ \mathbf{X}_\ell, \mathbf{y}_\ell \}, \quad \ell = 1, 2$$

from the two models, we denote with $\text{data} = (\text{data}_1, \text{data}_2)$ the sample of all the observations. In order to evaluate hypothesis (3.12) we compute the marginal distribution of each regression coefficient β_ℓ , $\ell = 1, 2$ as in (3.11), where $(\beta_1, \beta_2) \in \Xi = \mathbb{R}^2$. The

hypothesis (3.12) identifies a partition $\{\Xi_a, \Xi_H, \Xi_b\}$ of Ξ , where

$$\begin{aligned}\Xi_a &= \{(\beta_1, \beta_2) \in \mathbb{R}^2 \mid \beta_1 < \beta_2\}, \\ \Xi_H &= \{(\beta_1, \beta_2) \in \mathbb{R}^2 \mid \beta_1 = \beta_2\}, \\ \Xi_b &= \{(\beta_1, \beta_2) \in \mathbb{R}^2 \mid \beta_1 > \beta_2\}.\end{aligned}$$

Then, after computing the integrals

$$\mathbb{P}((\beta_1, \beta_2) \in \Xi_j \mid data) = \int_{\Xi_j} \prod_{\ell=1}^2 h_{1,\ell}(\beta_\ell \mid data_\ell) d\beta_1 d\beta_2, \quad j = a, b, \quad (3.13)$$

we can evaluate the hypothesis (3.12) by means of the BDM as

$$\delta_H = 1 - 2 \cdot \min_{a,b} \{\mathbb{P}((\beta_1, \beta_2) \in \Xi_a \mid data), \mathbb{P}((\beta_1, \beta_2) \in \Xi_b \mid data)\}. \quad (3.14)$$

Example 3.8. Consider the following *iid* observations from two independent populations

	<i>sample 1</i>								
(x_i)	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0
(y_i)	2.0	2.3	3.0	3.5	3.8	5.2	6.6	7.0	7.1

	<i>sample 2</i>								
(x_i)	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0
(y_i)	2.0	1.8	3.0	2.0	2.4	3.6	3.3	4.4	3.5

giving $(\hat{\beta}_1, \hat{\lambda}_1, \nu_1) = (0.723, 12.66, 7)$ and $(\hat{\beta}_2, \hat{\lambda}_2, \nu_2) = (0.267, 8.19, 7)$. Thanks to these samples it is possible to write the conjugate marginal distribution as

$$\beta_1, \beta_2 \mid data \sim h_{1,1}(\beta_1 \mid \mathbf{X}_1, \mathbf{y}_1) \cdot h_{1,2}(\beta_2 \mid \mathbf{X}_2, \mathbf{y}_2),$$

whose contour lines are shown in Figure 3.3 on the left hand side.

Starting from (3.13) and (3.14) the hypothesis $H : \beta_1 = \beta_2$ can be evaluated by means of the BDM. Since $\delta_H = 0.640$, there is not enough evidence to reject the hypothesis.

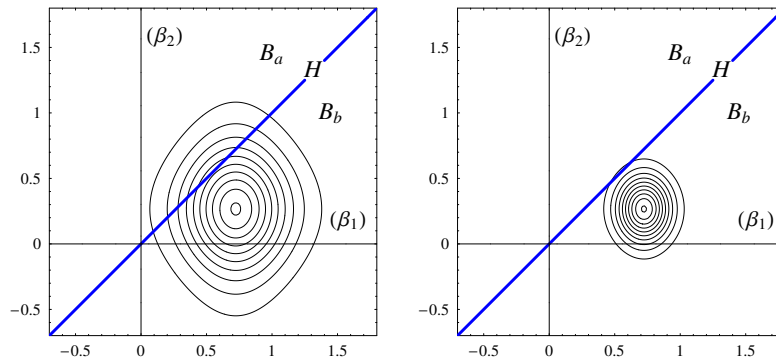


Figure 3.3: Contour plots for Example 3.8.

Now, suppose to have larger samples with $n_A = n_B = 36$ *iid* observations and to have obtained $(\hat{\beta}_1, \hat{\lambda}_1, \nu_1) = (0.723, 50.63, 34)$ and $(\hat{\beta}_2, \hat{\lambda}_2, \nu_2) = (0.267, 32.76, 34)$. The relative contour lines can be seen in Figure 3.3 on the right hand side. In this second case, $\delta_H = 0.951$ leading to the rejection of the hypothesis.

Chapter 4

Comparing k independent populations through the Bayesian Discrepancy Measure

In this section we propose a general procedure, based on the BDT, for comparing k parameters, or their transformations, from independent populations. The presented approach is not a simple extension of the one outlined in the previous chapter and, due to the geometry of the hypothesis and the parameter space, it requires a great number of details and clarifications. Once again, this methodology allows us to discuss problems not yet addressed in the literature.

4.1 Problem setting

Let us consider $k \geq 3$ independent populations $\mathcal{P}_1, \dots, \mathcal{P}_k$, each represented by the parametric model

$$Y_i \sim f(y_i | \theta_i), \quad i = 1, \dots, k, \quad (4.1)$$

where the parameter $\theta_i \in \Theta_i \subseteq \mathbb{R}^p$, $p \geq 1$, and the *iid* random variables Y_1, \dots, Y_k (discrete or continuous) have the same support \mathcal{X} . Let n_i be the number of observations of the *i*-th population and $N = \sum_{i=1}^k n_i$.

Given k populations, the data vector of all observations is $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_k)$, with

$\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})$ and, analogously, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k) \in \Theta = \Theta_1 \times \dots \times \Theta_k$ is the joint vector of all parameters $\boldsymbol{\theta}_i$.

The procedure we propose allows to compare k populations through a comparison of some parameters of interest. In this regard, let us consider a differentiable function

$$\begin{aligned} \phi: \Theta_i &\longrightarrow \Phi_i \subseteq \mathbb{R} \\ \boldsymbol{\theta}_i &\longmapsto \varphi_i = \phi(\boldsymbol{\theta}_i), \quad i = 1, \dots, k \end{aligned}$$

for each population parameter $\boldsymbol{\theta}_i$, which gives the relevant parameter of interest φ_i (where φ_i can be just one of the $\boldsymbol{\theta}_i$ components). Let us further consider a bijective reparametrization

$$\boldsymbol{\theta}_i \mapsto (\varphi_i, \boldsymbol{\zeta}_i), \quad i = 1, \dots, k,$$

for each model (4.1), where $\boldsymbol{\zeta}_i \in \mathcal{Z}_i \subseteq \mathbb{R}^{p-1}$ is a nuisance parameter vector. Then the vector of the parameters of interest is $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_k) \in \Phi = \Phi_1 \times \dots \times \Phi_k \subseteq \mathbb{R}^k$.

4.1.1 Prior, posterior and marginals

Given the i -th sample \mathbf{y}_i , consider the likelihood

$$L^i(\boldsymbol{\theta}_i | \mathbf{y}_i) = \prod_{j=1}^{n_i} f(y_{ij} | \boldsymbol{\theta}_i), \quad i = 1, \dots, k,$$

and a prior distribution $g_0(\boldsymbol{\theta}_i)$ for each parameter vector $\boldsymbol{\theta}_i$. The corresponding i -th posterior distribution is

$$\boldsymbol{\theta}_i | \mathbf{y}_i \sim g_1^i(\boldsymbol{\theta}_i | \mathbf{y}_i) \propto g_0(\boldsymbol{\theta}_i) \cdot L^i(\mathbf{y}_i; \boldsymbol{\theta}_i), \quad i = 1, \dots, k.$$

Since the populations being compared are independent, we have

$$\boldsymbol{\theta} | \mathbf{y} \sim \prod_{i=1}^k g_1^i(\boldsymbol{\theta}_i | \mathbf{y}_i). \quad (4.2)$$

The marginal posterior distribution of the i -th scalar parameter of interest φ_i is then

$$h_1^i(\varphi_i | \mathbf{y}_i) = \int_{\phi(\boldsymbol{\theta}_i) = \varphi_i} g_1^i(\boldsymbol{\theta}_i | \mathbf{y}_i) d\boldsymbol{\theta}_i. \quad (4.3)$$

The independence of φ_i , for all $i = 1, \dots, k$, allows the factorization of the posterior distribution as

$$\boldsymbol{\varphi} | \mathbf{y} \sim \prod_{i=1}^k h_1^i(\varphi_i | \mathbf{y}_i). \quad (4.4)$$

4.2 Comparison of k independent populations

By means of the BDT, we attempt to address and solve the classical problem of comparing k populations. Therefore, we would like to test the hypothesis

$$H : \varphi_1 = \dots = \varphi_k. \quad (4.5)$$

In order to do this, let us consider an appropriate k -pla of known real constants $\mathbf{c}^\top = (c_1, \dots, c_k) \in \mathbb{R}^k$, where $\sum_{i=1}^k c_i = 0$, and a linear combination of interest

$$\psi = \sum_{i=1}^k c_i \varphi_i = \mathbf{c}^\top \boldsymbol{\varphi}, \quad (4.6)$$

said *contrast*, with $\mathcal{C} \doteq \left\{ \mathbf{c} \in \mathbb{R}^k : \sum_{i=1}^k c_i = 0 \right\}$. Testing (4.5) corresponds to test

$$H_{\mathbf{c}} : \psi = 0, \quad (4.7)$$

for all $\mathbf{c} \in \mathcal{C}$ since these two formulations are linked by the relationship

$$\varphi_1 = \dots = \varphi_k \Leftrightarrow \psi = \mathbf{c}^\top \boldsymbol{\varphi} = 0, \forall \mathbf{c} \in \mathcal{C}, \quad (4.8)$$

for a proof see Casella and Berger, 2001 (chap. 11).

Despite the equivalence between the hypotheses, from the geometrical point of view we can notice that H does not partition the space while $H_{\mathbf{c}}$ partitions the space for each $\mathbf{c} \in \mathcal{C}$.

4.2.1 Distribution of a fixed contrast

In the following, we will construct a step of the procedure that allows to test hypothesis (4.7) and, consequently, hypothesis (4.5).

As said before, from a geometrical point of view, the contrast $\psi = \mathbf{c}^\top \boldsymbol{\varphi} = 0$ is, for a fixed $\mathbf{c} \in \mathcal{C}$, one hyperplane of the hyperplanes star \mathcal{S}_k passing through a straight line ℓ_H identified by the hypothesis. Note that for each choice of \mathbf{c} we have a different contrast.

Whatever the vector of contrasts \mathbf{c} chosen, the hypothesis H_c induces the partition $\{\Gamma_1, \Gamma_{H_c}, \Gamma_2\}$ of the space Φ , with

$$\begin{aligned}\Gamma_1 &\doteq \{ \boldsymbol{\varphi} \in \Phi \mid \mathbf{c}^\top \boldsymbol{\varphi} > 0 \} \\ \Gamma_{H_c} &\doteq \{ \boldsymbol{\varphi} \in \Phi \mid \mathbf{c}^\top \boldsymbol{\varphi} = 0 \} . \\ \Gamma_2 &\doteq \{ \boldsymbol{\varphi} \in \Phi \mid \mathbf{c}^\top \boldsymbol{\varphi} < 0 \}\end{aligned}\tag{4.9}$$

Then the BDM is computed as

$$\delta_{H_c}(\mathbf{c}) = 1 - 2 \min \left\{ \int_{\Gamma_1} \prod_{i=1}^k h_1^i(\varphi_i | \mathbf{y}_i) d\varphi_i , \int_{\Gamma_2} \prod_{i=1}^k h_1^i(\varphi_i | \mathbf{y}_i) d\varphi_i \right\} ,$$

which depends from the vector of contrasts \mathbf{c} chosen. The posterior distribution of the contrast, resulting from the selected \mathbf{c} , is

$$\psi | \mathbf{y} \sim p(\psi | \mathbf{y}) = \int_{\mathbf{c}^\top \boldsymbol{\varphi} = \psi} \prod_{i=1}^k h_1^i(\varphi_i | \mathbf{y}_i) d\varphi_i .\tag{4.10}$$

In this way, thanks to the latter, it is possible to test hypothesis (4.7) by computing

$$\delta_H = 1 - 2 \cdot \min \left\{ \int_{-\infty}^0 p(\psi | \mathbf{y}) d\psi , \int_0^{\infty} p(\psi | \mathbf{y}) d\psi \right\} .\tag{4.11}$$

Furthermore, it should be noted that the statistician, depending on his or her needs, may not be interested in hypothesis (4.7) but rather in hypothesis

$$H_c : \psi = \psi_H ,$$

where $\psi_H \neq 0$ is set by the statistician according to his purposes. The proposed proce-

sure can easily deal with functions of interest of the form $\psi = \psi(\boldsymbol{\varphi})$ by computing

$$\delta_H = 1 - 2 \cdot \min \left\{ \int_{-\infty}^{\psi_H} p(\psi | \mathbf{y}) \, d\psi, \int_{\psi_H}^{\infty} p(\psi | \mathbf{y}) \, d\psi \right\}.$$

It must be taken into account that the determination of $p(\psi | \mathbf{y})$, using the (4.10), may also turn out to be laborious. In such a case, it is possible to use a distribution $\tilde{p}(\psi | \mathbf{y})$, obtained by means of well-known simulation techniques. Instead of (4.2.1) we will then have the estimate

$$\tilde{\delta}_H = 1 - 2 \cdot \left\{ \int_{-\infty}^{\psi_H} \tilde{p}(\psi | \mathbf{y}) \, d\psi, \int_{\psi_H}^{\infty} \tilde{p}(\psi | \mathbf{y}) \, d\psi \right\}.$$

4.2.2 The Optimality Criterion

In the space $\Phi \subseteq \mathbb{R}^k$, the hypothesis (4.7) for a fixed \mathbf{c} is a straight line passing through the origin, whose points are equidistant from the co-ordinate axes φ_i . Summarising

- (a) if $k = 2$ then the hypothesis H partition the space of interest Φ and it is possible to compute the BDM knowing distribution (4.4);
- (b) on the contrary, if $k \geq 3$, the hypothesis H does not partition the space Φ . In that case, Φ can only be partitioned through a hyperplane with dimension $k - 1$.

In order to overcome the inconvenience that occurs when $k \geq 3$, we formulated a criterion that consists in taking a partitioning hyperplane of the star \mathcal{S}_k for which

$$\delta_H(\mathbf{c}) = 1 - 2 \cdot \min \left\{ \mathbb{P}(\boldsymbol{\varphi} \in \Gamma_1 | \mathbf{y}), \mathbb{P}(\boldsymbol{\varphi} \in \Gamma_2 | \mathbf{y}) \right\} \quad (4.12)$$

is highest. Hence, we seek to find

$$\mathbf{c}^* = \arg \max_{\mathbf{c} \in \mathcal{C}} \delta_H(\mathbf{c}). \quad (4.13)$$

The vector \mathbf{c}^* , which identifies the optimal contrast and for which the BDM is maximum (optimal), is called *optimal vector*.

Let us now explain the reasoning behind this choice. The fact that to reject $H : \varphi_1 = \dots = \varphi_k$ one must find at least one $\mathbf{c} \in \mathcal{C}$ for which the related contrast is

rejected (see formula (4.8)), implies that with this criterion we are selecting the ‘worst possible’ \mathbf{c} . In this way, if the contrast relating to \mathbf{c}^* is rejected, we have found at least one contrast such that $\psi = \mathbf{c}^\top \boldsymbol{\varphi} = 0$ does not hold and therefore we can reject H . Otherwise, if this is not the case, we cannot reject H since we are sure that all of the other contrasts have a lower $\delta_H(\mathbf{c})$.

In order to simplify the illustration of the procedure for choosing the optimal contrast, consider the situation for which $k = 3$. We write (4.4) as

$$h(\varphi_1, \varphi_2, \varphi_3) = h_1^1(\varphi_1) h_1^2(\varphi_2) h_1^3(\varphi_3), \quad \varphi_1, \varphi_2, \varphi_3 \in \mathbb{R}$$

and express the straight line ℓ_H as

$$\ell_H : \{\varphi_1 = u, \varphi_2 = u, \varphi_3 = u\}, \quad \forall u \in \mathbb{R}.$$

Let $\zeta(u) = h(u, u, u)$ be the profile curve of the hypersurface $h(\varphi_1, \varphi_2, \varphi_3)$ in correspondence to the points of the line ℓ_H . We assume that the marginal distributions h_1^i are unimodal and differentiable *almost everywhere*, then it follows that the contour surfaces of $h(\varphi_1, \varphi_2, \varphi_3)$ are convex. For the same reason, also ζ is differentiable and unimodal.

Inspired by the heuristic intuition, we make a conjecture about which vector \mathbf{c}^* is supposed to maximise $\delta_H(\mathbf{c})$ and therefore maximises the difference

$$|\mathbb{P}(\boldsymbol{\varphi} \in \Gamma_1 | \mathbf{y}) - \mathbb{P}(\boldsymbol{\varphi} \in \Gamma_2 | \mathbf{y})|.$$

Conjecture 4.1. The *optimal vector* \mathbf{c}^* is the vector such that the straight line ℓ_H is tangent with one of the level surfaces.

In fact, the straight line ℓ_H intersects each contour surface in at most two points, but only one of them is intersected in only one point that we indicate with $P^* = (u^*, u^*, u^*)$. It is easy to deduce that u^* is the maximum point of the profile curve ζ . Then, the optimal vector \mathbf{c}^* is (up to a constant) the gradient of the function $h(\varphi_1, \varphi_2, \varphi_3)$ evaluated at the point P^* , i.e.

$$\mathbf{c}^* \propto \vec{\nabla} h(P^*) = \left\{ \frac{\partial h}{\partial \varphi_1}, \frac{\partial h}{\partial \varphi_2}, \frac{\partial h}{\partial \varphi_3} \right\} \Big|_{P^*}.$$


The plane passing through ℓ_H and perpendicular to the gradient, providing the *optimal*

contrast, has equation

$$c_1^* \cdot \varphi_1 + c_2^* \cdot \varphi_2 + c_3^* \cdot \varphi_3 = 0. \quad (4.14)$$

This plane identifies the optimal *division* $\{\Gamma_1^*, \Gamma_{H_c}^*, \Gamma_2^*\}$ which can be expressed similarly to (4.9).

4.3 Examples

We now present some examples that will hopefully clarify the procedure. Again here, we have always adopted Jeffreys' priors. All of the  codes for the following examples can be seen in Manca, 2023.

Example 4.2. (Poisson model)

For this example we consider the study examined by Girón et al., 2022 on potential carcinogenicity of several chemical compounds. From four Poissonian independent populations $Y_i \sim Po(\cdot | \lambda_i)$, $i = 1, \dots, 4$, are extracted $n_1 = 154$, $n_2 = 19$, $n_3 = 20$, $n_4 = 11$ *iid* observations with sums $t_1 = 34$, $t_2 = 4$, $t_3 = 6$, $t_4 = 5$. Let $N = \sum_{i=1}^4 n_i = 204$ and $T = \sum_{i=1}^4 t_i = 49$. These four groups are divided according to different doses of methyl iodide with which some mice were treated with: zero (labeled as $i = 1$), low (corresponds to 0.06 mmoles per kg of mouse and labeled as $i = 2$), medium (0.15 mmoles per kg, $i = 3$) and high (0.31 mmoles per kg, $i = 4$).

The hypothesis to be tested is $H : \lambda_1 = \lambda_2 = \lambda_3 = \lambda_4$. Assuming, the Jeffreys' non-informative prior for the parameter λ_i , the posterior $h_1^i(\varphi_i | \mathbf{y}_i)$ is a Gamma distribution with shape parameter $\alpha_i = \frac{1}{2} + t_i$ and rate parameter $\beta_i = n_i$. Therefore, the posterior distribution of the parameter of interest vector $\boldsymbol{\lambda}$ is

$$\boldsymbol{\lambda} | \mathbf{y} \sim \prod_{i=1}^4 \text{Gamma}(\lambda_i | \alpha_i, \beta_i),$$

and the profile distribution over the hypothesis H is

$$\zeta(\boldsymbol{\lambda}) \propto \lambda^{T-2} \exp\{-N\lambda\},$$

fixing that $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \lambda$ along ℓ_H . The maximum point of the profile curve

ζ is $\lambda^* = \frac{T-2}{N} = 0.23$ and the intersection point between ℓ_H and a certain contour surface is $P^* = (0.23, 0.23, 0.23, 0.23)$. Then, the optimal vector \mathbf{c}^* is

$$\mathbf{c}^* \propto \vec{\nabla} h(P^*) = \left\{ \frac{\partial h}{\partial \lambda_1}, \frac{\partial h}{\partial \lambda_2}, \frac{\partial h}{\partial \lambda_3}, \frac{\partial h}{\partial \lambda_4} \right\} \Big|_{P^*} = (-0.648, -0.287, 0.292, 0.643)$$

and $\delta_H(\mathbf{c}^*) = 0.894$. Therefore the global hypothesis H cannot be strongly rejected. This result is in agreement with the article.

The authors consider also a second example aiming to analyze the patient survival after heart valve replacement operations in a sample of 109 patients along a certain period of observation. The number of months after the operations are, in the four groups, $n_1 = 1259$, $n_2 = 2082$, $n_3 = 1417$, $n_4 = 1647$ while the number of deaths are $t_1 = 4$, $t_2 = 1$, $t_3 = 7$, $t_4 = 9$. The intersection point is $P^* = (0.00297, 0.00297, 0.00297, 0.00297)$, the optimal vector is $\mathbf{c}^* = (-0.033, -0.798, 0.323, 0.508)$ and $\delta_H(\mathbf{c}^*) \approx 1$. Therefore the global hypothesis H can be rejected. Again, this result is in agreement with the one obtained by Girón et al., 2022.

Example 4.3. (Normal model)

From Ventura and Racugno, 2017, we consider a dataset reporting the resistance time (in minutes) measured on professional cyclists in a particular training session, after the ingestion of four different doses of caffeine (0, 5, 9, and 13 mg). The aim of the analysis is to assess whether there are significant differences between the endurance times after caffeine consumption in different doses. It is assumed that $Y_i \sim Normal(y_i | \mu_i, \phi_i)$, $i = 1, \dots, 4$ and that the populations are independent. We conduct two separate analyses:

- first, we perform a test on the means by considering the hypothesis $H : \mu_1 = \mu_2 = \mu_3 = \mu_4$. Assuming, the Jeffreys' non-informative prior for the parameter μ_i , the posterior distribution of the vector parameter of interest is

$$\boldsymbol{\mu} | \mathbf{y} \sim \prod_{i=1}^4 StudentG\left(\mu_i \mid \eta_i, \frac{\beta_i}{\nu_i \alpha_i}, 2\alpha_i\right),$$

using the notation given in Example 2.9.

The intersection point is $P^* = (55.18, 55.18, 55.18, 55.18)$ while the optimal vector is $\mathbf{c}^* = (-0.862, 0.218, 0.358, 0.286)$ and $\delta_H(\mathbf{c}^*) = 0.939$. Therefore the

global hypothesis H can be rejected. This result is in agreement with the p -value that is equal to 0.00359.

- Second, we perform a test on the precisions by considering the hypothesis $H : \phi_1 = \phi_2 = \phi_3 = \phi_4$. Assuming, the Jeffreys' non-informative prior for the parameter ϕ_i , the posterior distribution of the vector parameter of interest is

$$\boldsymbol{\phi} \mid \mathbf{y} \sim \prod_{i=1}^4 \text{Gamma}(\phi_i \mid \alpha_i, \beta_i),$$

using again the notation given in Example 2.9. The intersection point is $P^* = (2781.503, 2781.503, 2781.503, 2781.503)$, while the optimal vector is

$$\mathbf{c}^* = (-0.824, 0.501, 0.065, 0.258)$$

and $\delta_H(\mathbf{c}^*) = 0.05812$. Therefore the global hypothesis H cannot be rejected.

Example 4.4. (Binomial model)

Finally, we consider an example from Nashimoto et al., 2013 that, with a frequentist test, compares the proportions in five different groups concerning the effects of five training methods (A , B , C , D , and E) on test scores. It is assumed that $Y_i \sim \text{Binomial}(\cdot \mid \theta_i)$, $i = 1, \dots, 5$ and that the populations are independent. For each group we have that k_i is the number of successes (the proportion of high scores) and n_i is the number of the observations.

	groups				
	A	B	C	D	E
k	12	10	36	36	37
n	60	40	80	60	50

The hypothesis to be tested is $H : \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5$. Assuming, the Jeffreys' non-informative prior for the parameter θ_i , the posterior distribution of the parameter of interest vector $\boldsymbol{\theta}$ is

$$\boldsymbol{\theta} \mid \mathbf{y} \sim \prod_{i=1}^5 \text{Beta}(\theta_i \mid \alpha_i, \beta_i),$$

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>B</i>	0.999995	-	-	-
<i>C</i>	0.922	0.998	-	-
<i>D</i>	0.9995	0.449	0.969	-
<i>E</i>	0.880	1	0.999	0.999998

Table 4.1: Values of δ_H for *Example 4.4*.

with $\alpha_i = k_i + 1/2$ and $\beta_i = n_i - k_i + 1/2$. The intersection point is

$$P^* = (0.45, 0.45, 0.45, 0.45, 0.45)$$

while the optimal vector is $\mathbf{c}^* = (-0.627, -0.336, -0.005, 0.369, 0.598)$ and $\delta_H(\mathbf{c}^*) \approx 1$. Therefore the global hypothesis H can be rejected. This result is in agreement with the one obtained by the authors of the paper.

As done in the paper under consideration, it may be of interest make a pairwise comparison between groups. This means that one has to test the hypothesis $H : \theta_i = \theta_j$ with $i \neq j$. The results are shown in Table 4.4, allowing us to conclude that we reject the hypothesis H for groups $B-D$ and $A-E$. Less strongly, we can also reject the hypothesis for groups $A-C$. These results are somewhat similar to those found by them, of course it all depends on the threshold one chooses.

Chapter 5

The Bayesian Discrepancy Test for Partial Correlations

Partial correlations coefficients express the relationships between two variables while taking into account other confounding variables. In the context of the multivariate normal model we will focus on, a null partial correlation corresponds to a conditional independence between two given variables.

In the frequentist framework expressing the evidence for or against conditional independence using a partial correlation is still in development and, in the Bayesian one, methods for partial correlation estimation and testing are not completely developed for the practical researcher's use. One proposal that aims to overcome this problem is that of Kucharskỳ, 2018, which constructs an analytical version of the default Bayes factor for partial correlations.

5.1 Problem setting

The aim of this chapter is to use the BDT to perform a test on the partial correlation coefficient. This is based on the promising work conducted by Kucharskỳ, 2018, where the marginal distribution of this coefficient is derived starting from a particular parameterisation of the model and a certain prior distribution.

In this section, in order to provide a solid background to our application, we summarise the work done by Kucharskỳ, 2018. The underlying idea is that partial correla-

tion can be viewed from a perspective of linear regression, since the absence of a partial correlation in multivariate normal data means conditional independence between some variables given all the others. First, the case where only one conditioning variable is considered, afterwards, the discussion is extended to the inference on the partial correlation in the situation where several conditioning variables are present.

5.1.1 The case of one conditioning variable

Let (X, Y, Z) be a multivariate normal random vector with mean $\boldsymbol{\mu} = (\mu_x, \mu_y, \mu_z)$ and a variance-covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_x \sigma_y \rho_{xy} & \sigma_x \sigma_z \rho_{xz} \\ \sigma_x \sigma_y \rho_{xy} & \sigma_y^2 & \sigma_y \sigma_z \rho_{yz} \\ \sigma_x \sigma_z \rho_{xz} & \sigma_y \sigma_z \rho_{yz} & \sigma_z^2 \end{pmatrix}.$$

The partial correlation between the variables X and Y taking into account the variable Z can be expressed as

$$\rho_{xy.z} = \frac{\rho_{xy} - \rho_{xz}\rho_{yz}}{\sqrt{(1 - \rho_{xz}^2)(1 - \rho_{yz}^2)}}.$$

Given n observations of the multivariate random vector (X, Y, Z) , they can be summarized by $data = (n, \bar{\mathbf{m}}, S)$, where $\bar{\mathbf{m}} = (\bar{x}, \bar{y}, \bar{z})$ is a vector of sample means and

$$S = \begin{pmatrix} s_x^2 & s_x s_y r_{xy} & s_x s_z r_{xz} \\ s_x s_y r_{xy} & s_y^2 & s_y s_z r_{yz} \\ s_x s_z r_{xz} & s_y s_z r_{yz} & s_z^2 \end{pmatrix}.$$

is an average sum of squares and cross-products matrix. The sample partial correlation is then

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}.$$

A test on the partial correlation $\rho_{xy.z}$ of X and Y conditioned on Z can be seen as a

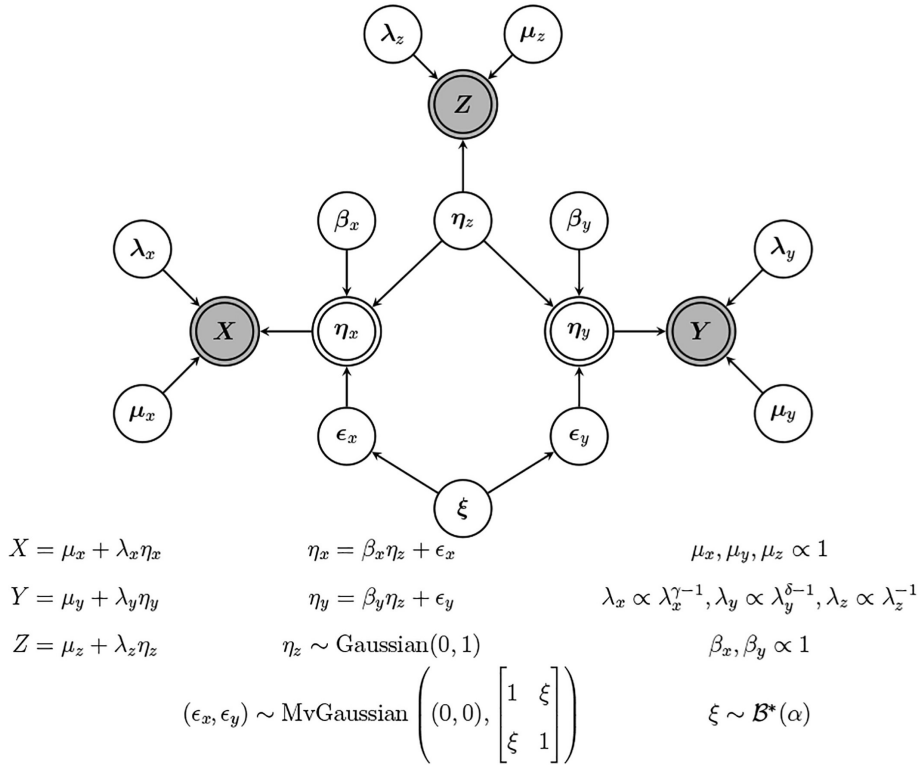


Figure 5.1: Bayesian graphical representation of the reparametrized model for the partial correlation.

comparison of two linear models

$$M_0 : X = \alpha + \beta_1 Z + \epsilon$$

$$M_1 : X = \alpha + \beta_1 Z + \beta_2 Y + \epsilon,$$

where one tests whether the addition of a predictor Y improves the fit.

Kucharský, 2018 (p. 11) states that it is necessary to specify the models M_0 and M_1 such that the Bayes factor depends only on the sample partial correlation by integrating out the nuisance parameters. The aim is to express Σ such that the partial correlation parameter $\rho_{xy.z}$ appears in the matrix entries. This is done specifying the partial correlation as a covariance between the residuals of X and Y after regressing them individually on Z . The reparametrized model results in the Bayesian graphical model shown in Figure 5.1, which implies that now the variance-covariance matrix takes the form

$$\Sigma = \begin{pmatrix} \lambda_x^2(1 + \beta_x^2) & \lambda_x\lambda_y(\xi + \beta_x\beta_y) & \lambda_x\sigma_z\beta_x \\ \lambda_x\lambda_y(\xi + \beta_x\beta_y) & \lambda_y^2(1 + \beta_y^2) & \lambda_y\sigma_z\beta_y \\ \lambda_x\sigma_z\beta_x & \lambda_y\sigma_z\beta_y & \sigma_z^2 \end{pmatrix},$$

where

$$\begin{aligned} \xi &= \rho_{xy.z}, \\ \mu_x &= \mathbb{E}[X], \quad \mu_y = \mathbb{E}[Y], \quad \mu_z = \mathbb{E}[Z], \quad \sigma_z^2 = \text{Var}(Z), \\ \lambda_x^2 &= \sigma_x^2(1 - \rho_{xz}^2), \quad \lambda_y^2 = \sigma_y^2(1 - \rho_{yz}^2), \quad \beta_x = \frac{\rho_{xz}}{\sqrt{1 - \rho_{xz}^2}}, \quad \beta_y = \frac{\rho_{yz}}{\sqrt{1 - \rho_{yz}^2}}. \end{aligned}$$

Prior specification

After this reparametrisation, the new model M_1 is parametrized by a vector

$$\boldsymbol{\theta} = (\mu_x, \mu_y, \mu_z, \lambda_x, \lambda_y, \sigma_z, \beta_x, \beta_y, \xi)$$

of 9 unknown parameters, while in the null model M_0 the partial correlation ξ is set to 0, implying that

$$\boldsymbol{\theta}_0 = (\mu_x, \mu_y, \mu_z, \lambda_x, \lambda_y, \sigma_z, \beta_x, \beta_y).$$

The prior for the alternative model M_1 is

$$g_0(\boldsymbol{\theta}|M_1) = \underbrace{B(1/2, \alpha)^{-1}(1 - \xi^2)^{\alpha-1}}_{g_0(\xi)} \cdot g_0(\boldsymbol{\theta}_0|M_0) \quad (5.1)$$

where $g_0(\boldsymbol{\theta}_0|M_0) = \underbrace{1^3}_{g_0(\boldsymbol{\mu})} \cdot \lambda_x^{\gamma-1} \cdot \lambda_y^{\delta-1} \cdot \sigma_z^{-1} \cdot \underbrace{1^2}_{g_0(\boldsymbol{\beta})}$ is the prior on the 8 parameters for the null model. The prior on the partial correlation ξ is a symmetric beta distribution stretched in the interval $(-1, 1)$ which is controlled by the hyperparameter $\alpha > 0$ (fixed by the researcher) and therefore is denoted as $\mathcal{B}^*(\alpha)$.

Bayes factor

Since we are interested to compute the Bayes factor, we need to have the marginal likelihood of the alternative model M_1 , for details of the challenging calculations see

Kucharský, 2018, p.17 – 24). It factors out as

$$f(\text{data}|M_1) = f(\text{data}|M_0) \cdot \int g_0(\xi) f_{\gamma,\delta}(n, r_{xy.z} | \xi) d\xi, \quad (5.2)$$

where

$$f(\text{data}|M_0) \cdot f_{\gamma,\delta}(n, r_{xy.z} | \xi)$$

is the reduced likelihood after integrating out all the parameters besides ξ , with $f_{\gamma,\delta}(n, r_{xy.z} | \xi) = (1 - \xi^2)^{1/2} h_{\gamma,\delta}(n, r_{xy.z} | \xi)$. The function

$$h_{\gamma,\delta}(r_{xy.z}, n | \xi) = A(n, r_{xy.z} | \xi) + B(n, r_{xy.z} | \xi), \quad (5.3)$$

i.e. it is a sum of an even function

$$A(n, r_{xy.z} | \xi) = (1 - \xi^2)^{(n-\gamma-\delta-1)/2} {}_2F_1\left(\frac{n-\gamma-1}{2}, \frac{n-\delta-1}{2}; \frac{1}{2}; r_{xy.z}^2 \xi^2\right),$$

where ${}_2F_1(a, b; c; z)$ is the Gaussian hypergeometric function, and an odd function

$$B(n, r_{xy.z} | \xi) = 2r_{xy.z} \xi (1 - \xi^2)^{(n-\gamma-\delta-1)/2} W_{\gamma,\delta}(n) \\ \times {}_2F_1\left(\frac{n-\gamma}{2}, \frac{n-\delta}{2}; \frac{3}{2}; r_{xy.z}^2 \xi^2\right),$$

where $W_{\gamma,\delta}(n) = [\Gamma(\frac{n-\gamma}{2}) \Gamma(\frac{n-\delta}{2})] / [\Gamma(\frac{n-\gamma-1}{2}) \Gamma(\frac{n-\delta-1}{2})]$. In short, $h_{\gamma,\delta}(n, r_{xy.z} | \xi)$ is of the form of the “reduced” likelihood of a Pearson’s correlation (see Ly et al., 2018, p. 6) and it is the part of the reduced likelihood depending on the partial correlation ξ given only the sample correlation $r_{xy.z}$ and the sample size n . In conclusion, Kucharský, 2018 (p. 16-17) proves that the Bayes factor testing the nullity of a partial correlation ξ can be expressed as

$$BF_{10} = \frac{f(\text{data}|M_1)}{f(\text{data}|M_0)} = \int_{-1}^1 g_0(\xi) (1 - \xi^2)^{1/2} h_{\gamma,\delta}(n, r_{xy.z} | \xi) d\xi \\ = B\left(\frac{1}{2}, \alpha + \frac{n-\gamma-\delta}{2}\right) / B\left(\frac{1}{2}, \alpha\right) \\ \times {}_2F_1\left(\frac{n-\gamma-1}{2}, \frac{n-\delta-1}{2}; \alpha + \frac{n-\gamma-\delta+1}{2}; r_{xy.z}^2\right), \quad (5.4)$$

where $n > 3$, and $\gamma, \delta, \alpha > 0$.

5.1.2 The case of more conditioning variables

Now, we focus on the case in which the controlling variable is a vector \mathbf{Z} , i.e. one wants to compare two linear models

$$\begin{aligned} M_0 : X &= \alpha + \beta_1 \mathbf{Z} + \epsilon \\ M_1 : X &= \alpha + \beta_1 \mathbf{Z} + \beta_2 Y + \epsilon, \end{aligned}$$

where $\mathbf{Z} = (Z_1, \dots, Z_k)$ and $\beta_1 = (\beta_{11}, \dots, \beta_{1k})$, in order to test whether the addition of the predictor improves the fit. Again, this can be seen as a test of the partial correlation ξ of X and Y conditioned on \mathbf{Z} .

The extension of the partial correlations model based on more controlling variables can be done by including in the model more latent variables (for the details see Kucharský, 2018, p. 24).

Bayes factor for $k = 2$

It can be proved that, for $k = 2$ conditioning variables, the marginal likelihood of the alternative model M_1 factors out as

$$f(\text{data}|M_1) = f(\text{data}|M_0) \cdot \int_{-1}^1 g_0(\xi)(1 - \xi^2)^{2/2} h_{\gamma, \delta}(n, r_{xy.z} | \xi) d\xi. \quad (5.5)$$

Therefore, the Bayes factor testing the nullity of a partial correlation ξ can be expressed as

$$\begin{aligned} BF_{10} &= \frac{f(\text{data}|M_1)}{f(\text{data}|M_0)} = \int_{-1}^1 g_0(\xi)(1 - \xi^2)^{2/2} h_{\gamma, \delta}(n, r_{xy.z} | \xi) d\xi \\ &= B\left(\frac{1}{2}, \alpha + \frac{n - \gamma - \delta + 1}{2}\right) \\ &\quad \times {}_2F_1\left(\frac{n - \gamma - 1}{2}, \frac{n - \delta - 1}{2}; \alpha + \frac{n - \gamma - \delta + 2}{2}; r_{xy.z}^2\right) / B\left(\frac{1}{2}, \alpha\right), \end{aligned} \quad (5.6)$$

where $n > 3$ and $\gamma, \delta, \alpha > 0$.

Bayes factor generalization for $k > 2$

Kucharský, 2018, based on the results of the partial correlation with one and two conditioning variables, conjectures that the marginal likelihood of the alternative model factors out as

$$f(\text{data}|M_1) = f(\text{data}|M_0) \cdot \int g_0(\xi) f_{\gamma,\delta}(n, r_{xy.z} | \xi) d\xi, \quad (5.7)$$

with $f_{\gamma,\delta}(n, r_{xy.z} | \xi) = (1 - \xi^2)^{k/2} h_{\gamma,\delta}(n, r_{xy.z} | \xi)$ the reduced likelihood, where k is the number of conditioning variables Z_1, \dots, Z_k and $h_{\gamma,\delta}(r, n | \xi)$ is of the form (5.3). Consequently, the Bayes factor for the partial correlation given k conditioning variables is

$$\begin{aligned} BF_{10} &= \int_{-1}^1 p(\xi) (1 - \xi^2)^{k/2} h_{\gamma,\delta}(r, n | \xi) d\xi \\ &= B\left(\frac{1}{2}, \mu\right) {}_2F_1\left(\frac{n - \gamma - 1}{2}, \frac{n - \delta - 1}{2}; \mu + \frac{1}{2}; r_{xy.z}^2\right) / B\left(\frac{1}{2}, \alpha\right). \end{aligned} \quad (5.8)$$

where $\mu = \alpha + \frac{n+k-\gamma-\delta-1}{2}$. The conjecture is supported by a simulation given in Kucharský, 2018 at pp. 33–38. He also proved that the Bayes factor appears to be invariant with respect to the correlation structure between the conditioning variables.

Marginal posterior of the Partial Correlation

The novelty produced by the work outlined until here, and that is of interest to us, is the analytical form for the marginal posterior distribution of the partial correlation. With a stretched beta prior on ξ symmetric around zero, the analytic marginal posterior of ξ is expressed as

$$\begin{aligned} g_1(\xi | n, r_{xy.z}) &= \frac{g_0(\xi) f_{\gamma,\delta}(n, k, r_{xy.z} | \xi)}{\int g_0(\xi) f_{\gamma,\delta}(n, k, r_{xy.z} | \xi) d\xi} \\ &= \frac{(1 - \xi^2)^{(2\alpha+k-2)/2}}{BF_{10}(r_{xy.z}, n, k, \alpha) B(1/2, \alpha)} h_{\gamma,\delta}(r_{xy.z}, n | \xi), \end{aligned} \quad (5.9)$$

given the conjecture on the Bayes factor for partial correlation for k conditioning variables, see formula (5.8).

Kucharský, 2018 proves that setting $\gamma = \delta = k$ gives a predictive matching Bayes factor (i.e such that $p(\text{data}|M_1) = p(\text{data}|M_0)$, see Jeffreys, 1939; Ly et al., 2016) following Jeffrey's philosophy for constructing Bayes factors. This choice gives the marginal posterior

$$\begin{aligned}
 g_1(\xi \mid n, r_{xy.z}) &= \frac{(1 - \xi^2)^{(2\alpha+n-k-3)/2}}{B\left(\frac{1}{2}, \alpha + \frac{n-k-1}{2}\right) {}_2F_1\left(\frac{n-k-1}{2}, \frac{n-k-1}{2}; \alpha + \frac{n-k}{2}; r_{xy.z}^2\right)} \\
 &\cdot \times \left[{}_2F_1\left(\frac{n-k-1}{2}, \frac{n-k-1}{2}; \frac{1}{2}; r_{xy.z}^2 \xi^2\right) \right. \\
 &\quad \left. + 2 r_{xy.z} \xi W_{k,k}(n) {}_2F_1\left(\frac{n-k}{2}, \frac{n-k}{2}; \frac{3}{2}; r_{xy.z}^2 \xi^2\right) \right].
 \end{aligned} \tag{5.10}$$

5.2 The BDT for the Partial Correlation

Given n observations of (X, Y, \mathbf{Z}) and considering the same problem setting as in the previous section, it is possible to outline a simple procedure for testing the partial correlation coefficient using the BDM. Taking advantage of the results obtained there, in particular the analytical expression of the marginal posterior of the partial correlation, we want to test the hypothesis $H : \rho_{xy.z} = 0$, which corresponds to testing

$$H : \xi = 0,$$

after reparametrizing the model. The hypothesis H identifies in the marginal parameter space $\Xi = (-1, 1)$ the partition $\{\Xi_a, \Xi_H, \Xi_b\}$, with

$$\begin{aligned}
 \Xi_a &= \{\xi \in \Xi \mid \xi < 0\}, \\
 \Xi_H &= \{\xi \in \Xi \mid \xi = 0\}, \\
 \Xi_b &= \{\xi \in \Xi \mid \xi > 0\}.
 \end{aligned}$$

Then, after computing

$$\mathbb{P}(\xi \in \Xi_j \mid n, r_{xy.z}) = \int_{\Xi_j} g_1(\xi \mid n, r_{xy.z}) \, d\xi, \quad j = a, b,$$

we can evaluate the BDM as

$$\delta_H = 1 - 2 \cdot \min_{a,b} \left\{ \mathbb{P}(\xi \in \Xi_a \mid n, r_{xy.z}), \mathbb{P}(\xi \in \Xi_b \mid n, r_{xy.z}) \right\} .$$

5.2.1 Illustrative examples

In this section we present three examples in order to show the simplicity of the procedure based on the BDM. These are all given in Kucharskỳ, 2018 (pp. 40 – 54).

Example 5.1. The first motivating example is presented to explain the functioning of the Default Bayesian Test for Partial Correlations in the event that the original data are not available. It provides a reanalysis of the results from Lleras et al., 2011 who tested the hypothesis $H : \xi = 0$, where ξ is the partial correlation of a search time and rapid resumption, conditioned on the age of the participants. Since the conditioning variable is only one, then $k = 1$. For the analysis has been chosen a uniform prior, on the partial correlation, between -1 and 1 which corresponds to set $\alpha = 1$. Starting from a sample partial correlation $r_{xy.z} = 0.01$ and a sample size $n = 40$, the Default Bayes Factor is $BF_{10} = 0.200$ indicating that the data are five times more likely to occur under the null model M_0 than under the alternative model M_1 . On the other hand, the *p-value* associated with the partial correlation is $p\text{-value} = 0.952$, informing us that the hypothesis H cannot be rejected at any reasonable α level. In agreement with this, we find that $\delta_H = 0.047$, suggesting that we do not have enough evidence to reject H .

Example 5.2. In a study conducted by Gottlieb and Lombrozo, 2018, the researchers considered $n = 317$ participants who were asked to rate how some psychological topics could be explained by scientific psychology (*scientific possibility*). They were also asked whether they feel comfortable with science explaining the phenomenon (*scientific discomfort*) and how much the phenomenon involves subjective experience (*introspection-phenomenology*). The result of the survey was that the more a phenomenon requires subjective experience, the smaller was the belief that it could be explained by science. Since people could feel more discomfort with science explaining subjective phenomena, the researchers realized that it might happen that the correlation between *introspection-phenomenology* and *scientific possibility* could be influenced by their relationship with *scientific discomfort*. Verifying if this intuition is true, can be done by the partial correlation of the two variables by taking *scientific discomfort* as conditioning variable Z .

Again, $k = 1$ and the resulting sample partial correlation is $r_{xy.z} = 0.027$. For the analysis has been again chosen a uniform prior on the partial correlation ($\alpha = 1$) and the Default Bayes Factor is $BF_{10} = 0.189$, indicating that the data are $1/0.189 = 5.3$ times more likely to occur under the null model M_0 than under the alternative model M_1 . The measure $\delta_H = 0.189$, suggesting that we do not have enough evidence to reject H .

Example 5.3. Finally, we analyse a subset of a dataset, investigated in Soyuyılmaz et al., 2017, containing the responses of $n = 43$ first year psychology students about their scientific thinking. The scores on three questionnaires are the 5 variables that are analyzed as a network

- *NFC*, the Need For Cognition short scale which assess the tendency to enjoy engagement in effortful thinking,
- *SSE*, the Scientific Self-Efficacy scale which has been developed to assess the confidence in engaging in scientific inquiry,

and three subscales of the Epistemic and Ontological Cognition Questionnaire (EOCQ)

- *EOCQa*, regarding the certainty of knowledge,
- *EOCQb*, concerning the justification of knowledge by authority,
- *EOCQc*, regarding the personal justification of knowledge.

In this case $k = 3$ and for the analysis has been again chosen a uniform prior on the partial correlation ($\alpha = 1$). The results of the analysis are reported in Figure 5.2.1, showing that the when the Bayes factor chooses the model M_1 then the BDM rejects the hypothesis H . This happens for the partial correlation on the couple of variables ($EOCQ_c, SSE$) and ($EOCQ_c, EOCQ_a$) given all the others.

		<i>NFC</i>	<i>SSE</i>	<i>EOCQa</i>	<i>EOCQb</i>
<i>SSE</i>	$r_{xy.z}$	0.121	-	-	-
	BF_{10}	0.257	-	-	-
	δ_H	0.533	-	-	-
<i>EOCQa</i>	$r_{xy.z}$	0.130	0.081	-	-
	BF_{10}	0.269	0.222	-	-
	δ_H	0.566	0.373	-	-
<i>EOCQb</i>	$r_{xy.z}$	0.163	0.286	0.917	-
	BF_{10}	0.321	0.925	0.405	-
	δ_H	0.675	0.919	0.767	-
<i>EOCQc</i>	$r_{xy.z}$	0.097	0.416	-0.304	-0.001
	BF_{10}	0.233	6.156	1.136	0.197
	δ_H	0.438	0.991	0.937	0.004

Table 5.1: Bayes factor and BDM for *Example 5.3*.

Chapter 6

Conclusions & Discussion

Several attempts have been made, over the last sixty years, to construct a measure of evidence that covers, in the Bayesian context, the role that the *p-value* has played in the frequentist setting. The goal of this thesis was to contribute to the Bayesian testing procedure for precise hypotheses by defining a new measure of evidence on which a testing procedure that is general is based, in contrast with the frequentist field where the testing procedures are ad hoc as they depend on the particular case study under consideration.

In this regard, this thesis proposed and investigated the use of a new evidence measure, known as BDM, to test precise hypotheses for parametric models. It should be kept in mind that the BDM has an exclusively evaluative nature of the null hypothesis, unlike most of the measures of evidence present in the literature which have a comparative nature as the null hypothesis is placed against an alternative one.

The outline of the thesis is as follows. Chapter 2 dealt with its definition in the univariate case in absence or presence of nuisance parameters. In Chapter 3 and 4 was, respectively, presented the application on the comparison between two and k parameters, or their functions, from independent populations. Last, Chapter 5 discussed the use of the BDT for testing partial correlations when considering a multivariate Gaussian model.

Overall, it has been shown that the BDT represent a flexible procedure that allows hypotheses to be tested intuitively and rather easily in different contexts. Furthermore, it became clear that this methodology allows us to address some problems not yet explored in the literature, as the test on the comparison of two skewness indexes. As we have seen, the Bayesian test that most closely resembles the BDT is the FBST defined in Pereira and

Stern, 2020. It must be emphasised that the BDM is much easier and faster to calculate than the *e-value*, the evidence measure on which the FBST is based, when the parameter space has dimension greater than 1. Furthermore, the BDM is invariant with respect to the marginalisations of the nuisance parameter, unlike the *e-value* for which the use of marginal densities to construct credible sets can produce inconsistency.

On the other hand, it should be highlighted that the extension of the BDM definition from the univariate case to the multi-parameter case is not straightforward due to the non-trivial problem of defining the median in multiple dimensions. This prevents an elegant and unified definition of the measure as the one provided by the FBST that examines the entire class of sharp hypotheses, not only considering the subclass of the hypotheses expressed as that are able to partition the parameter space as the BDT does. A similar problem arises with regard to the impossibility of the direct extension of the procedure for comparing two populations to the comparison of k . This is due to the fact that, in the latter case, the hypotheses do not partition the parameter space leading to the need for a conjecture to find the best partitioning hyperplane.

Nevertheless, this method has a considerable potential for development. Some of the main extensions concerning the BDT based approach include:

- tests on the regression coefficients of linear models and generalized linear models;
- goodness-of-fit tests for model validation and selection;
- predictive tests;
- parametric tests for non-regular models (Pareto, Cauchy, Laplace, etc.).

Another interesting extension of the work done in this thesis, and in particular Chapter 5, is the use of the BDM as a sequential method for Gaussian graphical model selection. Let's see in detail what it is.

Graphical models are a way of representing conditional independence relationships between variables. A graphical model consists of a vertex set $V = \{1, \dots, p\}$ (which correspond to the variables X_1, \dots, X_p), and an edge set E which contains edges of the form (i, j) where $i, j \in V$.

Let now be $\mathbf{X} = (X_1, \dots, X_p) \sim N_p(\mathbf{x} \mid \mu, \Sigma)$ a p dimensional normally distributed random vector. Let also consider the inverse covariance matrix or precision matrix $\Theta =$

Σ^{-1} , whose elements are indicated by θ_{ij} with $i, j = 1, \dots, p$. The lack of an edge in the graph, i.e. $(i, j) \notin E$, specifies a conditional independence relationship between the variables X_i and X_j . In fact, there exists a close link between graphical models and the precision matrix in the case of Gaussian random variables. This can be summarized by saying that

$$\theta_{ij} = 0 \iff X_i \perp X_j \mid X_{-(i,j)}, \quad (6.1)$$

that is, zero entries in the precision matrix correspond to conditional independence of some variables given all the others. We may then write this expression in terms of partial correlations, i.e.

$$\theta_{ij} = 0 \iff \rho_{X_i X_j \cdot X_{-(i,j)}} = 0. \quad (6.2)$$

Then main goal of graphical models is to identify the graph that best represents the variables based on certain observations. The space of possible graphical models is too large to compare every model in a fully Bayesian way. For this reason, have been introduced some search algorithms which choose a subset of the models among all the models to be compared, and hopefully this subset contains some good or the ‘best’ model (however one defines ‘best’). One Bayesian way to do this is sequentially removing the edge which maximises the Bayes factor score. A possible alternative to the usage of the Bayes factor score in model search algorithms, can be the usage of the BDM for Gaussian graphical model selection in the $p > 2$ case. One can start by using the BDM to test the hypothesis $H : \theta_{ij} \neq 0$ or, equivalently, $H : \rho_{X_i X_j \cdot X_{-(i,j)}} = 0$, thanks to (6.2). From these one would obtain a δ_{ij} , for each θ_{ij} , which can be used to make a model search algorithm.

Appendix A

Short references to some statistical tests

The appendix provides a concise exposition of some statistical tests expressing ideas, which are more or less shareable, that inspired the definition of the measure of evidence under study in this work. The terminology adopted in this appendix will be the same as the one used in the thesis. Therefore, with evident meaning of the symbols, it will be referred to

$$\mathcal{F} = \{f(x|\boldsymbol{\theta}), x \in \mathcal{X} \subseteq \mathbb{R}^k, \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\} \quad (\text{A.1})$$

as the family of distributions indexed by the parameter $\boldsymbol{\theta}$. The prior distribution will be denoted by $g_0(\boldsymbol{\theta})$ and the posterior distribution, given the vector $\boldsymbol{x} = (x_1, \dots, x_n)$ of *iid* observations dependent on $\boldsymbol{\theta}$, will be $g_1(\boldsymbol{\theta}) = \frac{g_0(\boldsymbol{\theta}) \prod_{i=1}^n f(x_i|\boldsymbol{\theta})}{m(\boldsymbol{x})}$, where $m(\boldsymbol{x}) = \int_{\Theta} \prod_{i=1}^n f(x_i|\boldsymbol{\theta}) g_0(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is the marginal or *prior predictive distribution*. The correspondent distribution function will be denoted by $G_1(\boldsymbol{\theta}|\boldsymbol{x})$.

A.1 Measures of surprise

The degree of incompatibility between the data and a certain hypothesised model H is quantified by means of surprise measures, without any reference to alternative models (see Bayarri and Berger, 1997). From a frequentist perspective, a traditional measure of surprise is the *p-value*.

A.1.1 Surprise indices

The first frequentist surprise index definition, other than the usual *p-values*, was formulated by Weaver, 1948, 1963; this was followed by a second frequentist reformulation due to Good, 1956, 1981, 1983, 1985, 1989. Later on Evans, 1997 introduced a Bayesian version of this index, through the concept of relative surprise. The latter formulation is central in this discussion because it will be seen to be the foundation of the Full Bayesian Significance Test (see Section A.3). For the sake of expositional clarity, we present Weaver's and Good's ideas that are necessary to understand Evans' thinking.

The word "surprising" is used to denote the feeling one gets observing the data once a model or hypothesis, reflecting the knowledge previously possessed, is formulated. According to Good, the main function of a feeling of surprise could be to make a person reconsider the validity of previous assumptions they have; it could lead them to change their subjective (personal) probabilities of various assumptions and possibly leading to the consideration of assumptions that were not previously considered.

Weaver's Surprise Index

Weaver, 1948 first clarifies the concepts meanings of *interesting*, *rare* and *surprising* events. A *rare* event, due to the frequentist definition of probability, is an improbable event. An improbable event is not always an *interesting* event. In fact, a *rare* event is *interesting* depending on whether the subject consider it interesting or not (it often depends on the experiment type). Furthermore, he observes that a *surprising* event occurs when its probability is not small in an absolute sense, but is small relative to the probability of any other alternative event. He points out that since the concept of probability coincides with that of degree of rarity, then a *surprising* event is a *rare* event but the vice versa does not apply.

In order to ease the comprehension of these concepts, he considers two examples. The first one involves a single 13-card bridge hand. There may appear

$$\binom{52}{13} = 635,013,559,600$$

different hands, all with equal probability $1/635,013,559,600$. Although each hand has small probability, thus each one is a *rare* event, it doesn't necessarily mean that each is

interesting. An *interesting* result would be represented by 13 ♠, but there is no reason to determine that the result is *surprising*. One could, instead, lump uninteresting events together and treat them as a single alternative event. In this way, the probability of the *interesting* event is much smaller than the alternative event. The result is that when an improbable event is so *interesting* that all its alternatives are grouped together as “dull” events as to be indistinguishable, then the *interesting* event may be a *surprising* event. The second example involves the tossing of a metal disc that can land in the following three positions: head, cross or edge. Assume that the thickness of the coin has been adjusted in such a way that it lands on the edge with a probability of one over a billion. This result is then *improbable*, *rare*, *interesting* and *surprising*. The surprise occurs because not only the probability of the event is low, but it is actually small compared to the other outcomes.

In order to make the concept of *surprising* event more precise, Weaver introduced the definition of *Surprise Index* of an event. This is an index that aims to measure how surprised the subject should be by the occurrence of a particular event. He considered an experiment having a discrete finite set of possible outcomes A_1, \dots, A_n having probabilities p_1, \dots, p_n , where p_i can be seen as the observed value of a random variable P (which takes values p_1, \dots, p_n with probabilities p_1, \dots, p_n). The surprise index, associated with the i -th occurring outcome A_i , is

$$(S.I.)_i = \frac{\mathbb{E}(P)}{p_i} = \frac{\sum_{j=1}^n p_j^2}{p_i},$$

where $\mathbb{E}(P)$ “is the average amount of probability one can expect to realize per trial of the experiment in question”⁽¹⁾. In short, the surprise index compares the observed value of the random variable P to its expectation. If the ratio is large, then one has the right to be surprised.

Good, 1956 argues that the surprise index is open to two criticisms. The first criticism involves the fact that the index changes when, in the discrete case, the results of an experiment are grouped together in a new manner and, in the continuous case, there is a change of the independent variable (then the surprise indices and the observed surprise will generally change due to the Jacobian factor). The second criticism concerns the fact

¹⁾Weaver, 1948.

that the numerator of the index is in some way arbitrary. Regarding the first criticism he considers 20 flips of an unbiased coin which would lead to 2^{20} different results. The two results

$$HTHTHTHTHTHTHTHTHTHTHT$$

and

$$HTTHTHHHTTTHHHHTTTHTH$$

have the same surprise index although it would seem that the former is more surprising. The author points out that this perception is partially due to the fact that the first result is “simpler”. In fact, if we would separate the simpler outcomes from the less simple ones into distinct groups, the first one would have a higher index of surprise than the second as expected. The vagueness of the definition of this index, according to Good, is due to the difficulty of measuring “the simplicity”. However, he tries to defend the connection between surprise and simplicity by stating that maybe the main function of surprise is to lead us to reconsider the validity of some hypotheses we had previously accepted; in the sense that we tend to be surprised when we find evidence against that accepted hypothesis. Formally, he means that we tend to be surprised if E occurs when the likelihood ratio $\mathbb{P}(E|H')/\mathbb{P}(E|H)$ is large for the hypothesis H previously considered valid and H' very simple. Although he attempts to find these arguments, in conclusion, he stresses that since a satisfactory measure of simplicity has not been found, it is unlikely that a satisfactory measure of surprise can be defined.

In order to overcome the second critical issue, as will be seen in the next section, he generalizes the definition of surprise index.

Good’s Surprise Index

Good (1982) underlines that the Weaver’s index for the i -th outcome can also be written as $\frac{\sum_j \mathbb{E}_j(P_j)}{p_i}$ where \mathbb{E}_j denotes an expectation over the j -th random variable. The author points up that this index coincides with the ratio between the Gini’s index of homogeneity $\rho = \sum_j p_j^2$ (the “repeat rate” in Turing’s terminology) and p_i . He also states that for continuous distributions one works with probability densities instead of probabilities, but the surprise index is then invariant only under linear transformations of the independent variable. Formally, if the experiments consists in the measurement

of a continuous vector or scalar variable with a differentiable distribution Good, 1956 defines the surprise index as

$$\frac{\mathbb{E}(P^*|H)}{p},$$

where H is a simple statistical hypothesis, P^* is the random variable that is the probability density of the original random variable and p is a realization of P^* .

Good described and improved his proposal in several articles, see Good (1956, 1982, 1983), generalizing the definition due to Weaver, 1948 and suggesting a “continuum”⁽²⁾ of indices of surprise. He considered u experiments combined into one, where a finite (or enumerable) set of mutually exclusive and exhaustive possible outcomes A_1, A_2, A_3, \dots has probability $p_i = \mathbb{P}(A_i|H)$, $i = 1, 2, \dots$ and H is a simple statistical hypothesis. The surprise index associated with the i -th occurring outcome A_i is defined as

$$\lambda_u = \frac{(\mathbb{E}[P^u])^{1/u}}{p_i} = \frac{(\sum_j p_j^{u+1})^{1/u}}{p_i}, \quad (u > 0).$$

In the continuous case,

$$\lambda_u = \frac{(\mathbb{E}[P^{*u}])^{1/u}}{p^*} = \frac{G.E.(P^*)}{p^*}, \quad (u > 0),$$

where $G.E.$ indicates the “geometric expectation”. Note that Weaver’s index is obtained for $u = 1$.

The limit for $u \rightarrow 0$ of the surprise index just defined, gives

$$\lambda_u \xrightarrow{u \rightarrow 0} \lambda_0 = p_i^{-1} \exp \left\{ \sum_j \mathbb{E}_j [\log(P_j)] \right\},$$

and the limit for $u \rightarrow \infty$ gives

$$\lambda_u \xrightarrow{u \rightarrow \infty} \lambda_\infty = p_i^{-1} \max p_j.$$

If the results of several statistically independent experiments are combined into a single

²⁾Good, 1983.

experiment, it can be seen that λ_u is multiplicative. The logarithmic transformation of λ_u gives the *Logarithmic Surprise Index* $\Lambda_u = \log \lambda_u$ which is additive. Good reminds that the expression $\Lambda_u + \log p_i$ is sometimes called *Rényi's generalized entropy* due to Rényi et al., 1961 which did not mention surprise indexes because he was unaware of Good, 1956. For $u \rightarrow 0$ the Logarithmic Surprise Index tends to

$$\begin{aligned}\Lambda_0 &= -\log p_i + \sum_j \mathbb{E}_j[\log(P_j)] \\ &= -\log p_i + \sum_j p_j \log(p_j)\end{aligned}$$

where the first term is the information provided by the event A_i and the second is the entropy of the experiment (expected amount of information from the experiment).

The author concludes that thanks to this interpretation Λ_0 and λ_0 are, respectively, the most natural additive and multiplicative surprise indexes. Furthermore, this consideration is supported by the fact that $\mathbb{E}(\Lambda_0) = 0$ and it is reasonable to demand that the expected log-surprise is equal to 0 before the experiment is done.

Good, 1983 suggested the possibility, after an observation is made, of selecting one hypothesis among a group of them (or estimating a parameter, which is logically the same thing), by a *principle of least surprise*. The surprise of a certain hypothesis H_k is denoted by

$$\lambda_u(H_k) = \frac{\left(\sum_j \mathbb{P}(E_j|H_k)^{u+1}\right)^{1/u}}{\mathbb{P}(E_i|H_k)}, \quad (u > 0).$$

Then, he underlined that the expression $1/\lambda_u(H_k)$ provides a generalization of the Barndorff-Nielsen, 1976 plausibility function $\frac{\mathbb{P}(E_i|H_k)}{\sup_j \mathbb{P}(E_j|H_k)}$ since the latter coincides with $1/\lambda_\infty(H_k)$. Taking advantage of this relation he proposed to select the hypothesis that maximizes the plausibility function by minimizing a surprise index, which corresponds to a generalization to Barndorff-Nielsen procedure. Finally, he claimed that a continuum of procedures for estimation or hypothesis choice could be obtained in this way, calling the core idea as *principle of least surprise*. When $u = 0$ this principle consists of *choosing the hypothesis* H_k that maximizes

$$\log \mathbb{P}(E_i|H_k) - \sum_j \mathbb{P}(E_j|H_k) \log(\mathbb{P}(E_j|H_k)).$$

Evans' relative surprise

Following Good's work on the concept of surprise, Evans considers the issue of determining Bayesian inference methods using the notion of *relative surprise*, see Evans, 1997. This paper's main argument is that, when applied to a Bayesian model, a modified version of surprise provides a conceptual basis for the development of a variety of estimating, hypothesis testing, model checking, and model selection techniques.

Assume to observe data \mathbf{x} from a statistical model and choose a prior distribution on the model parameter. Following Good's principle of least surprise, Evans proposed to arrange the elements of the parameter space Θ in the following order: if the relative increase in belief for θ_1 is, from a priori to a posteriori, greater than the corresponding increase for θ_2 , then θ_1 is strictly preferred over θ_2 . Basically, inference is made by using a preference ordering that favours θ_1 over θ_2 when

$$\frac{g_1(\theta_1|\mathbf{x})}{g_0(\theta_1)} > \frac{g_1(\theta_2|\mathbf{x})}{g_0(\theta_2)}.$$

Maximizing the ratio $\frac{g_1(\theta|\mathbf{x})}{g_0(\theta)}$ as a function of θ yields the *least relative surprise* estimate, i.e. a value $\theta^* \in \Theta$ that has the greatest relative increase in belief from a priori to a posteriori.

Suppose we want to test a precise hypothesis $H : \theta = \theta_0$ with $\theta_0 \in \Theta$, this means that we want to assess H using the evidence provided by the data. The preference ordering that we have just introduced, leads to compare the relative increase in belief for θ_0 , from a priori to a posteriori, with the increase for each of the other possible values in Θ . Then, if the increase for θ_0 is small compared to the others, then the data suggests that θ_0 is surprising and we have evidence against the hypothesis. Evans proposed to make this comparison by computing the *observed relative surprise* at θ_0 , which is the posterior probability of obtaining a relative increase larger than that observed for θ_0

$$\mathbb{P} \left(\frac{g_1(\theta|\mathbf{x})}{g_0(\theta)} > \frac{g_1(\theta_0|\mathbf{x})}{g_0(\theta_0)} \mid \mathbf{x} \right).$$

If this value is high, then the data suggests that θ_0 is surprising and we have evidence against the hypothesis.

Reminding to Good's principle of least surprise, we can select a certain hypothesis $H : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ which minimizes the *observed relative surprise* at $\boldsymbol{\theta}_0 \in \Theta$. Hence, minimizing the probability

$$\mathbb{P} \left(\frac{g_1(\boldsymbol{\theta}|\mathbf{x})}{g_0(\boldsymbol{\theta})} > \frac{g_1(\boldsymbol{\theta}_0|\mathbf{x})}{g_0(\boldsymbol{\theta}_0)} \mid \mathbf{x} \right),$$

we find the *least relative surprise* estimate $\boldsymbol{\theta}_0$, that is the value most supported by the evidence and therefore the least surprising.

In his paper, Evans introduced also the definition of an α -relative surprise region

$$C_\alpha(\mathbf{x}) = \left\{ \boldsymbol{\theta}_0 \in \Theta \mid \mathbb{P} \left(\frac{g_1(\boldsymbol{\theta}|\mathbf{x})}{g_0(\boldsymbol{\theta})} > \frac{g_1(\boldsymbol{\theta}_0|\mathbf{x})}{g_0(\boldsymbol{\theta}_0)} \mid \mathbf{x} \right) \leq \alpha \right\},$$

which corresponds to the set of all the values belonging to Θ for which the observed relative surprise is no greater than α .

A.1.2 Bayesian *p-value* proposals

Box, 1980, revisiting Fisher's *p-value* idea in a Bayesian key, proposed a method for analyzing the conformity of the data to a given model through a new Bayesian evidence measure called *prior predictive p-value*. This approach has been called *model criticism*, where the criticism of the model occurs by contrasting data and prior information. Although this is an attractive method, it presents some difficulties in numerical processing. Guttman, 1967 and Rubin, 1984, independently from each other, attempted to eliminate some of the problems by presenting a new measure, *the posterior predictive p-value*, which did not come without criticism. In addition, this was followed by some new proposals from Bayarri and Berger, 2000. One of them is called *conditional predictive p-value* and is presented in the following discussion. As mentioned at the beginning, note that tests based on surprise measures does not involve any alternative statistical model that is, therefore, not specified and that does not enter into the analysis.

Prior predictive p -value

Consider a parametric statistical model (A.1). Suppose we want to test the hypothesis

$$H : X \sim f(x|\theta_0), \text{ with } \theta_0 \in \Theta, \quad (\text{A.2})$$

then the *prior predictive distribution* can be seen as a natural tool to quantify the surprise. In fact, given the observed data \mathbf{x}_0 under the null model, small values of $m(\mathbf{x}_0)$, the probability of observing \mathbf{x}_0 , indicates a surprising result.

In order to clarify the analogy between Box measure and Fisher's p -value, it is useful to remind the definition of the latter. It measures the plausibility of a model or an hypothesis (A.2) and it is defined by

$$p = \mathbb{P}^{f(\cdot|\theta_0)} \left(t(\mathbf{X}) \geq t(\mathbf{x}_0) \right), \quad (\text{A.3})$$

where $T = t(\mathbf{X})$, with $t : \mathcal{X} \rightarrow \mathbb{R}^+$, is a statistic that constructs an ordering on \mathcal{X} . The inequality $t(\mathbf{x}') > t(\mathbf{x}'')$ means that \mathbf{x}' is further from H than \mathbf{x}'' . In this way, the statistic T checks the compatibility of the model with the observed data.

Box proposed that the plausibility of a model or an hypothesis of the type (A.2), would be measured by the index

$$p_{prior} = \mathbb{P}^{m(\cdot)} \left(m(\mathbf{X}) \leq m(\mathbf{x}_0) \right), \quad (\text{A.4})$$

where $m(\cdot)$ is the prior predictive distribution. It is easy to see an analogy with Fisher's p -value. Whereas in (A.3) the sample distribution conditional on the hypothesis H is employed, in (A.4) the probability is calculated by using the prior predictive distribution. Furthermore, it can be noticed that p_{prior} is calculated assuming the statistic $T = 1/m(\mathbf{X})$, hence in (A.4) there is a flip of the inequality direction compared to that in (A.3).

The approach proposed by Box is to interpret a small value of p_{prior} as experimental evidence against the assumptions made, as it emphasizes a surprising result based on the observations made. In this way, $1 - p_{prior}$ and $1/p_{prior}$ can be used as a measures of surprise. In fact, note that $1 - p_{prior}$ is for Evans, 1997 the observed surprise.

This index has several weaknesses, three of which relate to model checking and are

presented below:

- (i) as pointed out by Piccinato, 2009, the first one arises because one considers events that are different from those actually observed. In fact while $\mathbf{X} = \mathbf{x}_0$ is observed, the integration is performed in the sample space, i.e. over the set $\{x : m(\mathbf{X}) \leq m(\mathbf{x}_0)\}$;
- (ii) the second one, highlighted by Bayarri and Berger, 2000, is “*The main weakness of p_{prior} for pure model checking is its dependence on the prior $\pi(\theta)$; in essence, $m(x)$ measures the likelihood of x relative to both the model and the prior, and an excellent model could come under suspicion if a poor prior distribution were used*”;
- (iii) to solve problem (ii), Bayarri and Berger, 2000 pointed out that one might consider noninformative priors, but this would cause a third problem. In fact, since noninformative priors are typically improper, it is not guaranteed that the marginal $m(\mathbf{x})$ would be proper, leading to the impossibility of calculating p_{prior} .

Posterior predictive p -value

In order to overcome difficulty (iii) mentioned in the previous section, many Bayesian statisticians proposed to consider the *posterior predictive distribution*

$$m_{post}(\mathbf{x}|\mathbf{x}_0) = \int_{\Theta} f(\mathbf{x}|\boldsymbol{\theta}) dG_1(\boldsymbol{\theta}|\mathbf{x}_0)$$

in place of the prior distribution function, where $f(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^n f(x_i|\boldsymbol{\theta})$. The first ones to use this method were Guttman, 1967 and Rubin, 1984.

Using this trick, the fact that $g_0(\boldsymbol{\theta})$ is improper does not imply that $m_{post}(\mathbf{x}|\mathbf{x}_0)$ is improper too. This approach led to the definition of the *posterior predictive p -value* as

$$p_{post} = \mathbb{P}^{m_{post}(\cdot|\mathbf{x}_0)}(m(\mathbf{X}) \leq m(\mathbf{x}_0)). \quad (\text{A.5})$$

In this case, however, there is another problem related to the double use of the data. In fact, they are first used to convert the prior $g_0(\boldsymbol{\theta})$ into a proper distribution $g_1(\boldsymbol{\theta}|\mathbf{x}_0)$, in

order to determine the posterior predictive distribution $m_{post}(\mathbf{x}|\mathbf{x}_0)$, and then to compute the probability $\mathbb{P}^{m_{post}(\cdot|\mathbf{x}_0)}(m(\mathbf{X}) \leq m(\mathbf{x}_0))$.

Conditional predictive *p*-value

Another variant, which avoids the suspicion of double use of the data, was proposed by Bayarri and Berger, 2000. They work on the *u*-conditional predictive *p*-value, whose underlying idea is to operate a conditioning not with respect to the whole result $\mathbf{X} = \mathbf{x}_0$, but with respect to $u(\mathbf{X}) = u_0$, where $u(\mathbf{X})$ is an appropriate statistic. In this way,

$$p_{cpred(u)} = \mathbb{P}^{m(\cdot|u_0)}(t(\mathbf{X}) \geq t(\mathbf{x}_0)), \quad (\text{A.6})$$

where

$$m(t|u) = \int_{\Theta} f(t|u; \boldsymbol{\theta}) \, dG_1(\boldsymbol{\theta}|u)$$

assuming that

$$g_1(\boldsymbol{\theta}|u) = \frac{f(u; \boldsymbol{\theta}) \cdot g_0(\boldsymbol{\theta})}{\int_{\Theta} f(u; \boldsymbol{\theta}) \, dG_0(\boldsymbol{\theta})}$$

is proper and where, with an abuse of notation, $f(t|u; \boldsymbol{\theta})$ and $f(u; \boldsymbol{\theta})$ indicate the conditional and marginal distributions of T and U under H .

The idea they suggest, for the continuous data, is to choose the conditional statistic as the conditional MLE for θ

$$\hat{\theta}_{cMLE}(\mathbf{x}) = \arg \max_{\boldsymbol{\theta}} f(\mathbf{x}|t, \boldsymbol{\theta}) = \arg \max_{\boldsymbol{\theta}} \frac{f(\mathbf{x}; \boldsymbol{\theta})}{f(t; \boldsymbol{\theta})}.$$

The *conditional predictive p-value* is therefore

$$p_{cpred} = p_{cpred(\hat{\theta}_{cMLE})}.$$

This proposal has also been subject to some criticisms see, among others, Piccinato, 2000:

“One standard criticism to checking a given model is that models can only be compared and cannot be assessed singly”

and also

“The difficulty lies in defining a convincing standard for comparisons across different situations for judging when the compatibility is too low to try to improve the model”

and

“In any case, I think that integrating over the sample space after knowing the data will always introduce too much noise; after all, *p-values* are still *p-values*”.

A.2 Bayesian decision-making tests: the Bayes factor

The most widely used Bayesian test is, without any doubt, the one that has the Bayes factor as its measure of evidence. This test, which involves a system consisting of at least two point or interval hypotheses, is decisionist in its nature. In contrast with the significance testing, Bayes factors support the evaluation of evidence in favor of a null hypothesis, rather than only allowing the null to be rejected or not rejected. In particular, it is a method for testing hypotheses by evaluating decisions through selected loss functions.

A.2.1 Bayes factor and odds

Consider a parametric statistical model of the type (A.1). Suppose one wishes to test two parametric hypothesis

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \quad \text{vs} \quad H_1 : \boldsymbol{\theta} \in \Theta_1, \quad (\text{A.7})$$

with H_0 the null hypothesis, H_1 the alternative one, and where $\Theta_0 \cup \Theta_1 = \Theta \subseteq \mathbb{R}$ and $\Theta_0 \cap \Theta_1 = \emptyset$. The purpose of the test is to establish whether the true value of the parameter θ is an element of the subset Θ_0 or Θ_1 .

In the Bayesian framework, for an observed sample $\boldsymbol{x} \in \mathcal{X}$ it can be seen that

$$\begin{aligned} \mathbb{P}(H_0|\boldsymbol{x}) &= \frac{\mathbb{P}(H_0) \cdot \mathbb{P}(\boldsymbol{x}|H_0)}{\mathbb{P}(H_0) \cdot \mathbb{P}(\boldsymbol{x}|H_0) + \mathbb{P}(H_1) \cdot \mathbb{P}(\boldsymbol{x}|H_1)} \\ &= \left(1 + \frac{1}{\mathcal{O}(H_0)} \cdot \frac{1}{B_{0,1}(\boldsymbol{x})}\right)^{-1}, \end{aligned} \quad (\text{A.8})$$

where $\mathcal{O}(H_0) = \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)}$ is the prior odd and the ratio $B_{0,1}(\boldsymbol{x}) = \frac{\mathbb{P}(\boldsymbol{x}|H_0)}{\mathbb{P}(\boldsymbol{x}|H_1)}$ is the so called *Bayes factor*. It is a measure of the experimental evidence in favour of the hypothesis H_0 or against the hypothesis H_1 given data \boldsymbol{x} . Given the posterior odd

$$\mathcal{O}(H_0|\boldsymbol{x}) = \frac{\mathbb{P}(H_0|\boldsymbol{x})}{\mathbb{P}(H_1|\boldsymbol{x})},$$

the relation

$$\mathcal{O}(H_0|\mathbf{x}) = B_{0,1}(\mathbf{x}) \cdot \mathcal{O}(H_0) \quad (\text{A.9})$$

holds. In this perspective, the Bayes factor is the multiplicative factor that transforms the prior odds into the posterior odds.

Equation (A.9) has a clear decision-making meaning that can be briefly summarised. The idea behind the test is that if $B_{0,1}(\mathbf{x}) > 1$ [< 1] then

$$\mathcal{O}(H_0|\mathbf{x}) > \mathcal{O}(H_0) \quad [\mathcal{O}(H_0|\mathbf{x}) < \mathcal{O}(H_0)],$$

and the experiment giving \mathbf{x} is said to increase our preferences in favour of hypothesis H_0 [H_1].

For simplicity, let us assume that $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ with $p = 1$ and that the hypothesis to be tested is of the type *simple hypothesis vs composite hypothesis*, i.e.

$$H_0 : \theta = \theta_0 \quad vs \quad H_1 : \theta \neq \theta_0. \quad (\text{A.10})$$

Concerning other systems of hypotheses from the literature, such as

(i) *composite hypothesis vs composite hypothesis*

$$H_0 : \theta \in (-\infty, a] \quad vs \quad H_1 : \theta \in (a, \infty)$$

or

$$H_0 : \theta \in [a, b] \quad vs \quad H_1 : \theta \notin [a, b],$$

(ii) *simple hypothesis vs simple hypothesis*

$$H_0 : \theta = \theta_0 \quad vs \quad H_1 : \theta = \theta_1,$$

where $\theta_0 \neq \theta_1$,

we will limit ourselves to a few remarks. The case of more than two hypotheses will not be considered as, for instance,

$$H_0 : \theta = \theta_0, \quad H_1 : \theta < \theta_0, \quad H_2 : \theta > \theta_0.$$

The approach based on Bayes factors is fully justified in the context of decision theory. Limiting to a comparison of only two hypotheses $\{H_0 \text{ vs } H_1\}$, the decision-making scheme involves: (a) the decisions $\{d_0, d_1\}$, where the former consists of accepting H_0 and the latter H_1 ; (b) the states of nature $\{\Theta_0, \Theta_1\}$, i.e. it is true that $\theta \in \Theta_0$ or it is true that $\theta \in \Theta_1$; (c) the loss system for which

- making the decision d_j when $\theta \in \Theta_{1-j}$ is true, correspond to a loss $l_j(\theta) > 0$, $j = 0, 1$. (For the sake of simplicity, assume that losses do not depend on θ);
- making the decision d_j when $\theta \in \Theta_j$ is true, does not cause any loss.

Provided we have the prior distribution function $G_0(\theta)$, the observed sample \mathbf{x} and thus the posterior distribution function $G_1(\theta|\mathbf{x})$, it is possible to express the posterior risk, associated with decision d_j , as

$$\rho_1(d_j|\mathbf{x}) = E_\theta(l_j|\mathbf{x}) = l_j \cdot \int_{\Theta_{1-j}} dG_1(\theta|\mathbf{x}) = l_j \cdot \mathbb{P}(H_{1-j}|\mathbf{x}) \quad j = 0, 1. \quad (\text{A.11})$$

Set up in this way, the optimal decision problem is brought back to the Bayesian posterior risk comparison. Thus, the following procedure is adopted

$$\frac{\rho_1(d_0|\mathbf{x})}{\rho_1(d_1|\mathbf{x})} \begin{cases} < 1 & \text{take decision } d_0 \\ \geq 1 & \text{take decision } d_1 \end{cases} \quad (\text{A.12})$$

or, equivalently,

$$\mathcal{O}(H_0|x) = \frac{\mathbb{P}(H_0|\mathbf{x})}{\mathbb{P}(H_1|\mathbf{x})} \begin{cases} \geq \frac{l_0}{l_1} & \text{take decision } d_0 \\ < \frac{l_0}{l_1} & \text{take decision } d_1. \end{cases} \quad (\text{A.13})$$

For a detailed discussion of the decision-making procedure underlying the Bayes factor test, see Bernardo and Smith, 1994, J. O. Berger, 1985 and Piccinato, 2009.

In the practice of Bayes factor analysis, the system of hypotheses that is most considered and most challenging is (A.10). The fact that H_0 is a point hypothesis and H_1 is a diffuse hypothesis makes it appropriate to elicitate the parameter in two stages. In the

first phase, the prior probabilities of the hypotheses are elicited

$$\pi_0 = \mathbb{P}(H_0), \quad 1 - \pi_0 = \mathbb{P}(H_1).$$

In the second step the elicitation of the prior density $g_0(\theta)$ subordinate to H_1 , i.e. on the space $\Theta \setminus \{\theta_0\}$, is provided with $\int_{\Theta} g_0(\theta) d\theta = 1$.

The prior distribution function of the parameter θ is then

$$G_0(\theta) = \pi_0 \cdot \mathbf{1}_{(\theta_0, \infty)}(\theta) + (1 - \pi_0) \cdot \int_{-\infty}^{\theta} g_0(t) dt, \quad (\text{A.14})$$

with $\pi_0 \in (0, 1)$. Therefore, the posterior distribution function can be written as

$$G_1(\theta) = \pi_1 \cdot \mathbf{1}_{(\theta_0, \infty)}(\theta) + (1 - \pi_1) \cdot \int_{-\infty}^{\theta} g_1(t|x) dt. \quad (\text{A.15})$$

where $g_1(\theta|x) \propto g_0(\theta) \cdot f(\mathbf{x}|\theta)$ and the posterior probabilities of the hypothesis are

$$\pi_1 = \mathbb{P}(H_0|\mathbf{x}) = \frac{\pi_0 \cdot f(\mathbf{x}|\theta_0)}{\pi_0 \cdot f(\mathbf{x}|\theta_0) + (1 - \pi_0) \cdot m_{g_0}(\mathbf{x})}, \quad 1 - \pi_1 = \mathbb{P}(H_1|\mathbf{x}).$$

Hence, it can be seen that

$$\frac{\mathbb{P}(H_0|x)}{\mathbb{P}(H_1|x)} = \frac{\pi_1}{1 - \pi_1} = \frac{\pi_0}{1 - \pi_0} \cdot B_{0,1}(x),$$

where $B_{0,1}(x) = \frac{f(\mathbf{x}|\theta_0)}{\int_{\Theta} f(\mathbf{x}|\theta) dG_0(t)} = \frac{f(\mathbf{x}|\theta_0)}{m_{g_0}(\mathbf{x})}$. In this way, decision (A.13) corresponds to

$$B_{0,1}(x) = \frac{f(\mathbf{x}|\theta_0)}{m_{g_0}(\mathbf{x})} \begin{cases} \geq \frac{l_1}{l_0} \cdot \frac{1-\pi_0}{\pi_0} & \text{take decision } d_0 \\ < \frac{l_1}{l_0} \cdot \frac{1-\pi_0}{\pi_0} & \text{take decision } d_1. \end{cases} \quad (\text{A.16})$$

Notice that hypotheses of kind (i)-(ii) do not require a two-step elicitation of this type.

The main shortcoming of the test based on Bayes factors is that it does not allow the use of improper prior distributions as this would lead to the Jeffreys-Lindley's paradox described in the next section. The need to not give up improper priors gave rise to the development of the *fractional Bayes factor* by O'Hagan, 1995 and the *intrinsic Bayes*

factor by J. O. Berger and Pericchi, 1996, both of which are marked by the more or less open violation of the likelihood principle. The weaknesses of the aforementioned approaches underlie our preference for the BDT, which does not suffer from the described drawbacks.

A.2.2 The Jeffreys-Lindley's paradox

The Jeffreys-Lindley's paradox, which has generated an interesting and fierce theoretical debate, has a dual interpretation (see Robert, 2014). On one hand, it results in the differentiation between frequentist and Bayesian statistics; on the other hand, it indicates the difficulty of using improper priors during testing.

The first interpretation focuses on the fact that the Jeffreys-Lindley's paradox is actually a counterintuitive situation of the statistical hypothesis testing problem, where the Bayesian and frequentist approaches give different results for certain choices of the prior distribution. The problem of this disagreement, already discussed by Jeffreys himself in his book Jeffreys, 1939, became known as the Jeffreys-Lindley's paradox after Lindley referred to it as a paradox, in Lindley, 1957. The context of the paradox, as set out by Lindley, is to consider a sample of size n from a normal distribution $N(x|\theta, \sigma^2)$ with known variance σ^2 , and to perform a statistical hypothesis testing on the mean, of the sort seen before

$$H_0 : \theta = \theta_0 \quad vs \quad H_1 : \theta \neq \theta_0. \quad (\text{A.17})$$

Summarizing the dataset into the sample mean (a sufficient statistic)

$$\bar{X}_n \sim N\left(\bar{x}_n \mid \theta, \frac{\sigma^2}{n}\right),$$

under the null hypothesis, leads to the t -statistic

$$T_n = \frac{\bar{X}_n - \theta_0}{\sigma/\sqrt{n}} \sim N(t_n \mid 0, 1).$$

Therefore, in the frequentist framework, the test is conducted by evaluating

$$p\text{-value} = \mathbb{P}(|T_n| > |t_n|) = 1 - 2\Phi(|t_n|),$$

while the Bayesian approach to the hypothesis-testing problem relies on the Bayes factor. If we choose a normal prior $\theta \sim N(\theta|\theta_0, \sigma^2)$, then the Bayes factor is

$$B_{0,1}(t_n) = \sqrt{1+n} \exp \left\{ -\frac{t_n^2}{2} \frac{n}{1+n} \right\},$$

giving the measure of the experimental evidence in favour of the hypothesis H_0 or against the hypothesis H_1 , given the data. The paradox occurs when, for a fixed t_n and for almost any choice of the prior distribution, one makes the sample size n tend to infinity. While the *p-value* is constant in n , the Bayes factor tends to infinity. This implies that, if we are in a certain experimental situation, the frequentist approach could lead to rejecting H_0 , while the Bayesian approach always leads to accepting H_0 , regardless of the particular experiment. The differing results do not represent a real disagreement between the two methods, since they answer fundamentally to different questions. On one hand we check if the result t_n is "significant" by a frequentist test of H_0 , indicating sufficient evidence to reject H_0 at a certain α -level; on the other hand, through the Bayes factor, we check if the posterior probability of H_0 given t_n is high, indicating strong evidence that H_0 is in better agreement with t_n than H_1 .

The second problem occurs when using improper distributions in Bayesian hypothesis testing procedures, particularly when the two hypotheses have different dimensions as in (A.17), i.e. for a point-null hypothesis tested against a composite alternative hypothesis. A drawback arises when the variance of a conjugate prior goes to infinity, i.e. if one considers a flat prior distribution. Such an improper prior can be expressed as

$$g_0(\theta) \propto k \cdot h(\theta)$$

where $h(\theta)$ is a positive non-integrable function on Θ , the constant k cannot be determined and the Bayes factor

$$B_{0,1}(x) = \frac{L(\theta_0|x)}{k \cdot \int_{\theta \neq \theta_0} L(\theta|x) \cdot h(\theta) d\theta}$$

depends on it. Moreover, the posterior probability of the null hypothesis $\mathbb{P}(H_0|x)$ cannot be calculated since it depends on the Bayes factor, as seen in equation (A.8). This problem can be seen as the focus shifting from the asymptotic trend of the sample size

to the variance of the prior distribution. Let us return to the normal framework and now consider only one observation from $N(x | \theta, \sigma^2)$ with a prior $\theta \sim N(\theta | \theta_0, n\sigma^2)$. Then,

$$B_{0,1}(x) = \sqrt{1+n} \exp \left\{ -\frac{(x-\theta_0)^2}{2\sigma^2} \frac{n}{1+n} \right\},$$

which again tends to infinity as n grows, no matter what the value of the observation x is. Similarly to the previous case, as highlighted in Robert, 2014, the phenomenon exhibited therein is not paradoxical in the least and can be explained by noticing that the prior gets more and more diffuse with increasing n , leading to the inconvenient result that the only relevant prior information becomes that θ could be equal to θ_0 . In conclusion, as Robert, 2014 underlines that

“There is therefore a deep coherence in the selection of the null hypothesis H_0 in this case: being completely indecisive about the alternative hypothesis means we could and should not choose this alternative. It is not possible to pick the alternative hypothesis of an undefined value of θ when opposed to the very special value θ_0 if we want to be “completely non-informative” about θ under H_1 . This analysis of the Jeffreys- Lindley paradox is justifying (further) the prohibition of the use of improper priors for testing point-null hypotheses”.

A.3 Full Bayesian Significance Test

The Full Bayesian Significance Test (*FBST*) has been introduced by Pereira and Stern, 1999 as a Bayesian measure of evidence for precise null hypotheses. The test is presented as a “Bayesian alternative approach to significance tests or, equivalently, to *p-values*”. The authors modify the definition in progress with the publication of several articles, see Madruga et al., 2003, Pereira et al., 2008, Pereira and Stern, 2020.

Among the Bayesian methods currently present in literature, the *FBST* test is the one that is most similar to the test that this thesis aims to present and discuss.

A.3.1 FBST definition

The *FBST* is presented as a Bayesian alternative to significance test of precise null hypothesis. Let us consider a parametric statistical model (A.1) under appropriate regularity conditions. In order to define the *FBST* procedure for a precise hypothesis $H : \boldsymbol{\theta} \in \Theta_H$, consider

$$g^* = \sup_{\Theta_H} g_1(\boldsymbol{\theta}|x) \quad \text{and} \quad T = \{\boldsymbol{\theta} \in \Theta \mid g_1(\boldsymbol{\theta}|x) > g^*\},$$

where T is called *tangential set* or *highest posterior density set* (HPDS) as it is the “Highest Posterior Density Region” that is “tangent” to the set that defines the null hypothesis.

The *Bayesian evidence value against H*, is defined by the authors as the posterior probability of the tangential set T , i.e.,

$$\bar{ev} = Pr(\boldsymbol{\theta} \in T|x) = \int_T g_1(\boldsymbol{\theta}|x) d\boldsymbol{\theta}, \quad (\text{A.18})$$

while the *Bayesian evidence value supporting H* is $ev = 1 - \bar{ev}$. The *FBST* procedure rejects H whenever ev is small.

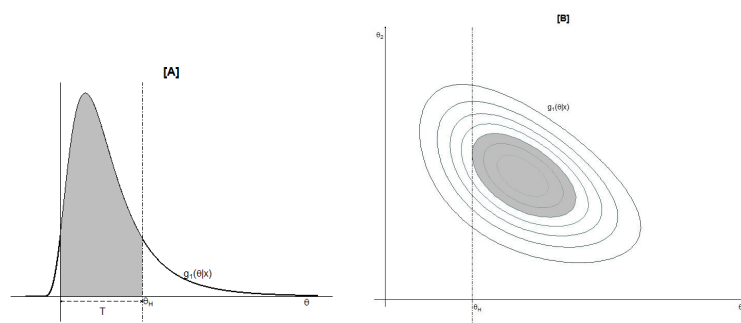


Figure A.1: Tangential set when [A] $p = 1$ and [B] $p = 2$.

The computation of ev needs numerical optimization and integration. Since it is not always possible to calculate some of these integrals, it is often necessary to employ simulation methods to approximate them. Pereira and Stern, 1999 underline that “*This methodology takes into account only the location of the maximum likelihood under the null hypothesis, making it consistent with the Benefit of the Doubt juridical principle*”, also known as the Onus Probandi juridical principle³. Accordingly, a working hypothesis is not rejected if there is not sufficient evidence against it.

Concerning the use of an alternative hypothesis in the procedure mentioned above, it is underlined that the alternative hypothesis is as important as the null hypothesis unlike the classical p -value that does not consider it. This claim is supported by the fact that “as the full parameter space is used in the computation of ev , the alternative hypothesis is always intrinsically considered”, see Pereira et al., 2008.

They also suggest an equivalent formulation of ev where the normalized likelihood, if available, can be used in place of the posterior distribution. This can be useful when it is assumed that the consistency between the data and the null hypothesis should not involve prior opinion about the parameter.

A.3.2 FBST invariance

The authors underline that, in statistical procedures, two types of invariance are required: invariance with respect to the null hypothesis parameterization and invariance with re-

³It states that “*There is no liability as long as there is a reasonable basis for belief, effectively placing the burden of proof (Onus Probandi) on the plaintiff, who, in a lawsuit, must prove false a defendant’s misstatement, without making any assumption not explicitly stated by the defendant, or tacitly implied by existing law or regulation*”, see Madruga et al., 2003

spect to the parameter space parameterization. The *FBST* procedure, in its original formulation outlined in A.3.1, fulfills the first requirement but not the second.

In order to make the procedure invariant with respect to alternative parameterizations of the parameter space, Madruga et al., 2003 introduced a second definition. They set a reference function $r(\boldsymbol{\theta})$ on the parameter space Θ where the prior was defined. Then, they consider the “surprise function” relative to a suitable reference function $r(\boldsymbol{\theta})$ to be chosen, i.e. $s(\boldsymbol{\theta}) = \frac{g_1(\boldsymbol{\theta}|\mathbf{x})}{r(\boldsymbol{\theta})}$ and require that

$$\boldsymbol{\theta}^* = \arg \max_{\boldsymbol{\theta} \in \Theta_H} s(\boldsymbol{\theta}|\mathbf{x}) = \arg \max_{\boldsymbol{\theta} \in \Theta_H} \frac{g_1(\boldsymbol{\theta}|\mathbf{x})}{r(\boldsymbol{\theta})}, \quad s^* = s(\boldsymbol{\theta}^*) = \sup_{\boldsymbol{\theta} \in \Theta_H} s(\boldsymbol{\theta}).$$

Essentially, they introduce an ordering of the parameter space Θ following the procedure proposed by Evans (see Section A.1.1), which is based on Good’s principle of least surprise. In fact, $\boldsymbol{\theta}^*$ is the least relative surprise estimate, i.e. the value in Θ_H that is least supported by the data. The evidence against H provided by the sample x is calculated using the *highest relative surprise set* (HRSS)

$$\bar{T}(s^*) = \{\boldsymbol{\theta} \in \Theta | s(\boldsymbol{\theta}) > s^*\},$$

which includes all parameter values $\boldsymbol{\theta}$ that attain a larger surprise function value than the supremum s^* of the null set. Finally, the e -value, that represents the Bayesian evidence against H , is defined as

$$e\text{-value} = \bar{e}v(H) = \bar{W}(s^*) = \int_{\bar{T}(s^*)} g_1(\boldsymbol{\theta}|\mathbf{x}) d\boldsymbol{\theta}.$$

On the contrary, the e -value in support of H is $ev(H) = 1 - \bar{e}v(H)$, which is evaluated by means of the set $T(s^*) = \Theta \setminus \bar{T}(s^*)$ and the cumulative surprise function $W(s^*) = 1 - \bar{W}(s^*)$. In conclusion, the *FBST* is the procedure that rejects H whenever $\bar{e}v(H)$ is large.

As pointed out in Pereira and Stern, 2020 (Section 3.2) “*the role of the reference density is to make $\bar{e}v(H)$ explicitly invariant under suitable transformations of the coordinate system*”. Notice that the first non-invariant definition of this measure corresponds to the use of a flat reference function $r(\boldsymbol{\theta}) \propto 1$. Some of the suggested choices for the

reference function are the use of uninformative priors such as “*the uniform, maximum entropy densities, or Jeffreys’ invariant prior*” (see Pereira and Stern, 2020, Section 3.2).

Asymptotic properties

Consider the cumulative distribution of the evidence value against the hypothesis $\bar{e}v(H)$, that is $\bar{V}(c) = \Pr(\bar{e}v(H) \leq c)$. Under appropriate regularity conditions, for increasing sample size, Pereira and Stern, 2020 provide a proof for the following statements.

Given θ^* the true value of the parameter:

- if $\theta_H \neq \theta^*$ then $\bar{e}v(H) \xrightarrow{p} 1$, that is, $\bar{V}(0 \leq c < 1) \rightarrow 0$;
- if $\theta_H = \theta^*$ then

$$\bar{V}(c) \approx F_{t-h}[F_t^{-1}[c]],$$

where $t = \dim(\Theta)$, $h = \dim(\Theta_H)$ and F_ν is the cumulative χ^2 -distribution with ν degrees of freedom.

Bibliography

- Adam, D. C., Wu, P., Wong, J. Y., Lau, T. K., E. H. Y. and Tsang, Cauchemez, S., Leung, G. M., & Cowling, B. J. (2020). Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nature Medicine*, 26(11), 1714–1719.
- Barndorff-Nielsen, O. (1976). Plausibility inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 38(2), 103–123.
- Bayarri, M. J., & Berger, J. O. (1997). *Measures of surprise in Bayesian analysis*. Cite-seer.
- Bayarri, M. J., & Berger, J. O. (2000). P-values for composite null models (with discussion). *Journal of the American Statistical Association*, 95(452), 1127–1142.
- Bayarri, M. J., & Berger, J. O. (2004). The Interplay of Bayesian and Frequentist analysis. *Statistical Science*, 19(1), 58–80.
- Bayes, C. L., & Branco, M. D. (2007). Bayesian inference for the skewness parameter of the scalar Skew-Normal distribution. *Brazilian Journal of Probability and Statistics*, 141–163.
- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., et al. (2018). Redefine statistical significance. *Nature human behaviour*, 2(1), 6–10.
- Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis* (2nd ed.). Springer Series in Statistics.
- Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses (with discussion). *Statistical Science*, 317–352.
- Berger, J. O., & Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433), 109–122.

- Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of P values and evidence (with discussion). *Journal of the American Statistical Association*, 82(397), 112–139.
- Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. John Wiley & Sons Inc.
- Bertolino, F., Columbu, S., & Manca, M. (2022). A contribution to the L. J. Savage problem. *SIS 2022 – 51st Scientific Meeting of the Italian Statistical Society. Book of Short Papers*, 1215–1220.
- Bertolino, F., Columbu, S., Manca, M., & Musio, M. (2022). Testing two coefficients of variation: a new Bayesian approach. *arXiv:2204.10147v1*.
- Bertolino, F., Columbu, S., Manca, M., & Musio, M. (2023). Comparing two independent populations through the Bayesian Discrepancy Test. [*Manuscript submitted for publication*].
- Bertolino, F., Manca, M., Musio, M., Racugno, W., & Ventura, L. (2021). A new Bayesian Discrepancy Measure. *arXiv:2105.13716v3*.
- Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298), 269–306.
- Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A (General)*, 143(4), 383–430.
- Box, G. E. P., & Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Wiley.
- Cabras, S., Mostallino, G., & Racugno, W. (2006). A nonparametric bootstrap test for the equality of coefficients of variation. *Communications in Statistics-Simulation and Computation*, 35(3), 715–726.
- Casella, G., & Berger, R. L. (2001). *Statistical inference* (2nd ed.). Duxbury.
- Chang, C. H., Lin, J. J., & N., P. (2011). Testing the equality of several gamma means: A parametric bootstrap method with applications. *Computational Statistics*, 26, 55–76.
- Chhikara, R., & Folks, J. (1989). The Inverse Gaussian distribution. *Inc., New York*.
- de Finetti, B. (1979). *Theory of Probability: A Critical Introductory Treatment*. Wiley.
- Evans, M. (1997). Bayesian inference procedures derived via the concept of relative surprise. *Communications in Statistics-theory and Methods*, 26(5), 1125–1143.
- Fisher, R. A. (1935). *The design of experiments*. Oliver & Boyd.

- Fisher, R. A. (1956). *Statistical methods and scientific inference*. Oliver & Boyd.
- Folks, J. L., & Chhikara, R. S. (1978). The Inverse Gaussian Distribution: Theory Methodology and Applications – A Review. *Journal of Royal Statistical Society Series B*, 40(3), 263–289.
- Geisser, S. (1993). *Predictive Inference: An Introduction*. Chapman & Hall.
- Girón, F. J., Martel–Escobar, M., & Vázquez–Polo, F.-.-J. (2022). A Bayesian homogeneity test for comparing Poisson populations. *Applied Stochastic Models in Business and Industry*.
- Good, I. J. (1950). *Probability and the weighing of evidence*. Charles Griffin & Company.
- Good, I. J. (1956). The Surprise Index for the Multivariate Normal Distribution. *The Annals of Mathematical Statistics*, 27(4), 1130–1135.
- Good, I. J. (1981). Some logic and history of hypothesis testing. In J. C. Pitt (Ed.), *Philosophy in Economics: Papers Deriving from and Related to a Workshop on Testability and Explanation in Economics held at Virginia Polytechnic Institute and State University, 1979* (pp. 149–174). Springer Netherlands.
- Good, I. J. (1983). C169. Barndorff-Nielsen’s plausibility function and a principle of least surprise. *Journal of Statistical Computation and Simulation*, 18(2-3), 215–218.
- Good, I. J. (1985). C226. a new measure of surprise. *Journal of Statistical Computation and Simulation*, 21(1), 88–89.
- Good, I. J. (1989). C332. surprise indexes and p-values. *Journal of Statistical Computation and Simulation*, 32(1–2), 90–92.
- Gottlieb, S., & Lombrozo, T. (2018). Can science explain the human mind? Intuitive judgments about the limits of science. *Psychological science*, 29(1), 121–130.
- Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 29(1), 83–100.
- Guttman, I. (1970). *Statistical Tolerance Regions: Classical and Bayesian*. Griffin.
- Hartigan, J. (1966). Note on the Confidence-Prior of Welch and Peers. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1), 55–56.

- Hossain, A., & Beyene, J. (2015). Application of skew-normal distribution for detecting differential expression to microRNA data. *Journal of Applied Statistics*, 42(3), 477–491.
- Jeffreys, H. (1939). *Theory of probability*. Oxford University Press.
- Kelter, R. (2022). fbst: An R package for the Full Bayesian Significance Test for testing a sharp null hypothesis against its alternative via the e-value. *Behavior Research Methods*, 54(3), 1114–1130.
- Kucharský, Š. (2018). *A Default Bayesian Test for Partial Correlations*. [Unpublished doctoral dissertation]. University of Amsterdam.
- Laxminarayan, R., Wahl, B., Dudala, S. R., Gopal, K., C., M. B., Neelima, S., Jawahar Reddy, K. S., Radhakrishnan, J., & Lewnard, J. A. (2020). Epidemiology and transmission dynamics of COVID-19 in two Indian states. *Science*, 370(6517), 691–697.
- Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, 88(424), 1242–1249.
- Lehmann, E. L. (2011). *Fisher, Neyman, and the creation of classical statistics*. Springer Science & Business Media.
- Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44, 187–192.
- Lindley, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint*. Cambridge University Press.
- Lindley, D. V. (1991). *Making decisions* (2nd ed.). Wiley.
- Lleras, A., Porporino, M., Burack, J. A., & Enns, J. T. (2011). Rapid resumption of interrupted search is independent of age-related improvements in visual search. *Journal of Experimental Child Psychology*, 109(1), 58–72.
- Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., & Getz, W. M. (2005). Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066), 355–359.
- Ly, A., Marsman, M., & Wagenmakers, E.-J. (2018). Analytic posteriors for Pearson's correlation coefficient. *Statistica Neerlandica*, 72(1), 4–13.

- Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, 72, 19–32.
- Madruga, M. R., Pereira, C. d. B., & Stern, J. M. (2003). Bayesian evidence test for precise hypotheses. *Journal of Statistical Planning and Inference*, 117(2), 185–198.
- Manca, M. (2022). Maramanca/a_new_bayesian_discrepancy_measure: A new bayesian discrepancy measure. <https://doi.org/10.5281/zenodo.7317122>
- Manca, M. (2023). maramanca/Examples_Chapter4: Comparing k independent populations through the Bayesian Discrepancy Measure. <https://doi.org/10.5281/zenodo.7539019>
- Manca, M., Columbu, S., & Musio, M. (2022). A Bayesian Test for the comparison of two independent populations. *SIS 2022 – 51st Scientific Meeting of the Italian Statistical Society. Book of Short Papers*, 1209–1214.
- Marini, E., Rebato, E., Racugno, W., Buffa, R., Salces, I., & Borgognini Tarli, S. M. (2005). Dispersion dimorphism in human populations. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, 127(3), 342–350.
- Mudholkar, G. S., & Hutson, A. D. (1996). The exponentiated Weibull family: Some properties and a flood data application. *Communications in Statistics–Theory and Methods*, 25(12), 3059–3083.
- Nashimoto, K., Haldeman, K. M., & Tait, C. M. (2013). Multiple comparisons of k binomial proportions. *Computational Statistics & Data Analysis*, 68, 202–212.
- O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 99–118.
- Pereira, C., & Stern, J. M. (2020). The e -value: A fully Bayesian significance measure for precise statistical hypotheses and its research program. *Sao Paulo J. Math. Sci.*
- Pereira, C., & Stern, J. M. (1999). Evidence and Credibility: Full Bayesian Significance Test for Precise Hypotheses. *Entropy*, 1(4), 99–110.
- Pereira, C., Stern, J. M., Wechsler, S., et al. (2008). Can a significance test be genuinely Bayesian? *Bayesian Analysis*, 3(1), 79–100.

- Piccinato, L. (2000). Asymptotic distribution of p values in composite null models: Comment. *Journal of the American Statistical Association*, 95(452), 1166–1167.
- Piccinato, L. (2009). *Metodi per le decisioni statistiche*. Springer Science & Business Media.
- Rényi, A., et al. (1961). On measures of entropy and information. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*.
- Robert, C. P. (2014). On the Jeffreys-Lindley paradox. *Philosophy of Science*, 81(2), 216–232.
- Rubin, D. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 1151–1172.
- Ruli, E., & Ventura, L. (2021). Can Bayesian, confidence distribution and frequentist inference agree? *Statistical Methods & Applications*, 30(1), 359–373.
- Soyyılmaz, D., Griffin, L. M., Martin, M. H., Kucharský, Š., Peycheva, E. D., Vaupotič, N., & Edelsbrunner, P. A. (2017). Formal and informal learning and first-year psychology students' development of scientific thinking: A Two-Wave Panel Study. *Frontiers in psychology*, 8, 133.
- Spiegelhalter, D. (2019). *The Art of Statistics: Learning from data*. Penguin Books.
- Ventura, L., & Racugno, W. (2017). *Biostatistica. Casi di studio in R*. Egea.
- Weaver, W. (1948). Probability, rarity, interest, and surprise. *The Scientific Monthly*, 67(6), 390–392.
- Weaver, W. (1963). *Lady Luck, The Theory of Probability*. Penguin.
- Wright, W. G. (1985). Effects of fracturing on well yields in the coal field areas of Wise and Dickenson counties, southwestern Virginia. *US geological survey water resources investigations report*, 1–20.
- Wuensch, K. L., Jenkins, K. W., & Poteat, G. M. (2002). Misanthropy, idealism, and attitudes towards animals. *Anthrozoös*, 15, 139–149.
- Yang, R., & Berger, J. O. (1996). *A catalog of noninformative priors*. Institute of Statistics; Decision Sciences, Duke University.
- Zellner, A. (1971). *An Introduction to Bayesian Inference Econometrics*. Wiley.