EFFICIENT CHOICE OF PRIORS FOR BAYESIAN HYPOTHESIS TESTS IN PSYCHOLOGY AND FOR DYNAMIC MODELING OF ZERO-INFLATED DIRECTED NETWORKS

A Thesis

by

SANDIPAN PRAMANIK

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Valen E. Johnson
Co-Chair of Committee,	Yang Ni
Committee Members,	Anirban Bhattacharya
	Debdeep Pati
	Pat Rubio Goldsmith
Head of Department,	Brani Vidakovic

August 2022

Major Subject: Statistics

ABSTRACT

We propose efficient priors for two different statistical problems: (1) designing Bayesian hypothesis tests with reduced costs for detecting the presence or absence of hypothesized effects, and (2) efficient modeling of dynamic zero-inflated directed networks. Our contributions cover computational and methodological aspects and touch upon theoretical aspects in some cases.

Costs of conducting experiments to test hypothesized effects are often directly related to the number of tested items or participants. To address this, in the first part of the thesis we propose cost-efficient Bayesian hypothesis tests. We describe a modified sequential probability ratio test that can be used to reduce the average sample size required to perform statistical hypothesis tests at specified levels of significance and power. Examples are provided for z and t tests, and tests of binomial success probabilities. A description of the software package to implement the test is provided. We compare the sample sizes required in fixed design tests conducted at 5% significance levels to the average sample sizes required in sequential tests conducted at 0.5% significance levels, and we find that the two sample sizes are approximately equal. To generalize this framework, we found the default implementations of Bayesian tests prevent the accumulation of strong evidence in favor of true null hypotheses because associated default alternative hypotheses assign a high probability to data that are most consistent with a null effect. We propose the use of "non-local" alternative hypotheses to resolve this paradox. The resulting class of Bayesian hypothesis tests permits a more rapid accumulation of evidence in favor of both true null hypotheses and alternative hypotheses that are compatible with standardized effect sizes of most interest in psychology.

The second part of the thesis extends the discussion of choosing efficient priors and proposes the Hurdle Network Model for modeling zero-inflated directed networks. We assume node-specific dynamic latent attributes to account for the underlying network structure and *apriori* assume the Dynamic Shrinkage Process on them. We find the model has good predictive performance. Simulation studies and an application on bilateral trade flows from the apparel industry are included to support this.

DEDICATION

To my family

ACKNOWLEDGMENTS

I would like to start by extending my sincere gratitude towards Dr. Valen Johnson. I thank him for welcoming me as one of his students and for being a great advisor. Thank you for always being there a door knock away. I fail to find words in explaining how his enormous knowledge and beautiful insights into the subject amazes me. On the academic side, he has really installed in me the importance of doing serious, rigorous work, and find simple yet powerful solutions to problems. He has always pushed me to hold myself to a high standard. As I keep growing as a researcher in pursuit of unfolding the truth and revealing the mystery of the universe, I will always strive for such excellence. He has really guided me through my Ph.D., but also allowed me the freedom to practice independent thinking. I cannot suppress the EQUAL importance of our informal meetings as well. They have made me comfortable to the point I do not hesitate even for a split second to reveal my ignorance of the subject to him. This has always been essential to my education as I believe the first step in learning about the unknown is admitting that we do not know. So I cannot appreciate this enough. As I move towards that phase of my life where I will try to provide for my family, his duties as the Head of the Department and the Dean of the College of Science have taught me to treat my colleagues with dignity and interact with my subordinates with respect. Thank you for everything.

I would like to extend my sincere appreciation to Dr. Yang Ni for being such a nice mentor. He is surely one of the nicest persons I have met and I am sure all of his current and future students will feel the same way. There were numerous occasions where I kept on repeating the same mistakes. I would reach out to him and he would always discuss them with the same enthusiasm. I really appreciate him for being so patient with me and for always giving me his time whenever I needed any help.

Drs. Anirban Bhattacharya and Debdeep Pati are inspirations to new researchers like me. They lead by example as they strive for research excellence every day. I was fortunate enough to sit in multiple courses taught by them. They are one of the finest teachers I have ever come across.

Thank you for motivating me and guiding me both in research and career. I would like to thank Dr. Pat Rubio Goldsmith for being in my committee and for all the helpful comments.

This would be incomplete without thanking the Department of Statistics for giving me the opportunity to pursue my PhD, for providing a fantastic work environment and many helpful resources. Whatever I have achieved, I cannot imagine experiencing any of it without their helps. I thank all the other faculties and staffs whom I have interacted with over the years. I have just gotten used to running into everyone of you in hallways and seminars among other places, and have refreshing and stimulating conversations. Everyone is so kind and they work so hard to maintain the department as one of the finest research institutions in Statistics. It is hard to describe what a privilege it has been. I thank A&M and College Station for everything, but particularly for giving me a grad life full of numerous memories that I will cherish for the rest of my life.

I would like to end by thanking my family and friends for being there with me. I owe all my achievements to my family.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Drs. Valen Johnson (Chair), Yang Ni (Co-Chair), Anirban Bhattacharya, Debdeep Pati of the Department of Statistics, and Dr. Pat Rubio Goldsmith of the Department of Sociology.

The data analyzed in Chapter 5 was provided by Dr. Raymond Robertson of the Bush School of Government and Public Service.

All works conducted for the dissertation was completed by the student independently.

Funding Sources

Drs. Valen Johnson and Yang Ni offered me Research Assistant opportunities in many occasions. The Department of Statistics offered me the Graduate Assistant Non-Teaching positions many a time. I was also fortunate enough to be accepted for the Graduate Assistant Teaching position to teach a course. I was financially supported in all these occasions. I am thankful to Drs. Valen Johnson and Yang Ni, their funding sources, the Department of Statistics, and the A&M for either directly or indirectly providing financial support for my research.

NOMENCLATURE

SPRT	Sequential Probability Ratio Test
MSPRT	Modified Sequential Probability Ratio Test
UMPBT's	Uniformly Most Powerful Bayesian Tests
ASN	Average Sample Number
IRT	Item Response Theory
OC	Operating Characteristics
GLR	Generalized Likelihood Ratio
MLE	Maximum Likelihood Estimate
CDF	Cumulative Distribution Function
GS	Group Sequential
SSD	Solid State Drive
SBF	Sequential Bayes Factor
NHST	Null Hypothesis Significance Test
NAP	Non-local Alternative Prior Density
$N(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$\phi(a \mid m, c^2)$	Normal density with mean μ and variance σ^2
$NM\left(\beta_0, \tau^2, m\right)$	Normal Moment prior of order m for the null hypothesized value β_0, scale parameter τ
Hurdle-Net	Hurdle Network Model
DSP	Dynamic Shrinkage Process
MSE	Mean Squared Error
MSPE	Mean Squared Prediction Error

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Experimental science relies on controlled experiments that test whether effects predicted by a scientific theory can be produced and measured in laboratory settings. Observational science is based on measuring outcomes as they occur naturally, without experimental intervention. In practice, measured outcomes from both observational studies and experiments are subject to random variation and measurement error. For this reason, hypothesis testing procedures and statistical modeling must be employed to determine whether data support or do not support a hypothesized effect or model. Innovative statistical methods to evaluate the plausibility of scientific theories have attracted increased attention over the last decade. This attention has resulted in renewed interest in Bayesian methods for assessing evidence [e.g., 1], and several novel approaches to sequential testing procedures have recently been proposed [2, 3]. As [2] point out, each of these sequential testing methods can be motivated from a Bayesian perspective towards testing.

Recent technological advancements in diverse areas of science have also made it easier to collect and analyze structured data in the form of networks. Some examples include data observed from social networks, genetic circuits and protein interaction networks. This has increased popularity of statistical analysis of networks over the past few years. Given that a model has been specified, a Bayesian approach for statistical inference requires us to assume priors on model parameters. Although there exists a plethora of methods for modeling such data, there still exists many open questions. One such question is how we can propose smart modeling strategies and propose efficient priors that can incorporate network structure, has interpretable parameters and also lets us draw statistical inference efficiently.

So the common underlying theme in this thesis is choosing efficient priors in these two different real life problems with a common goal of drawing valid statistical inference. Below we discuss some of the key questions that still persists in each of these problems and then briefly outline the way this thesis attempts to solve them.

1.1 Reducing Sample Size to Attain Higher Statistical Significance

In the classical hypothesis testing paradigm, two types of errors are considered to assess whether data support or do not support a hypothesized effect. Type I error occurs when the null hypothesis of "no effect" is rejected when the hypothesized effect does not exist. To limit claims of false discovery, hypothesis testing procedures are commonly designed so that the probability of a Type I error (i.e., α) is limited to be less than a prespecified value, often 0.05. Type II error occurs when we fail to reject the null hypothesis when the hypothesized effect does exist (the probability of a Type II error is denoted by β).

Recent concerns over the replicability of scientific studies have led to calls to move away from p values and significance testing [4, 5, 6, 7]. However, p values and significance testing continue to play critical roles in many areas, including genomics, high-energy physics, and clinical trials. An examination of recent articles in prominent psychology journals also suggests that p values and significance testing continue to play an important role in psychological research [8, 9]. Elsewhere, we have proposed to address the limitations of p values by reducing the significance thresholds required for declaring a positive finding from $\alpha = 0.05$ to $\alpha = 0.005$ [10, 11]. While this change would improve the replicability of scientific claims of discoveries, it would also increase the costs of conducting studies because larger sample sizes would be required if similar controls on Type II error probability were maintained.

In Chapter 2, we describe a modification of the sequential probability ratio test (SPRT) of [12] that reduces the sample sizes required to achieve specified Type I and Type II error probabilities. The modified design can be applied to many studies conducted in the social and natural sciences in which the goal is to establish the existence of a hypothesized effect. Implicit in this goal is the detection of effects that are not arbitrarily close to zero (or the null value of the parameter). In this regard, the proposed design differs from recent developments of sequential procedures designed to estimate various effect sizes, such as standardized mean differences, correlation and regression coefficients, and coefficients of variation [13, 14, 15]. We propose the Modified Sequential Probability Ratio Test (MSPRT) for testing a point null hypothesis against a one or two-sided alternative

hypothesis. In designing these tests, we objectively set alternative hypotheses. The alternative hypotheses we propose are based on uniformly most powerful Bayesian tests (UMPBT's) or approximate UMPBT's [11, 16]. We note that exact UMPBT's are known only for one-parameter exponential family models and tests for the non-centrality parameters of chi-squared statistics [17]. Approximate UMPBT's are known for *t* tests. Thus, a limitation of the MSRPT is that is applicable primarily to z and *t* tests, tests of binomial proportions and Poisson means, and chi-squared tests.

For this class of tests, empirical evidence suggests that MSPRT's require sample sizes that can be less than 50% of the sample size that is required in corresponding fixed designs when the null hypothesis of no effect is true, and sample sizes that can be 20% smaller when alternative hypotheses are true. In general, the sample size savings accrued by the use of the MSPRT depends on the test statistic chosen and the targeted Type I and II error probabilities for the test.Theoretical support for these findings is provided in [18], where approximate formulae for the average sample number (ASN) and operating characteristics for truncated SPRTs are derived. These results approximate discrete time stochastic processes (representing the observed sequential tests) by Brownian motion or Wiener processes, which are continuous time stochastic processes. For sufficiently large sample sizes, these processes provide approximate operating characteristics and ASN's for truncated SPRT's. In the case of one- and two-sample z tests, the underlying assumptions required in deriving those formulae apply to the MSPRT, and the approximate values from these results agree with our empirical findings. Specific details regarding this connection appear in Section 2.5.1. As noted by [18], this theoretical result "leads to appreciable qualitative insight; and quantitatively it does provide a first, crude approximation which can often be used as a basis for subsequent refinement."

1.2 Efficient Alternative for Bayesian Hypothesis Tests

To identify invariances, hypothesis testing procedures must permit accumulation of evidence in support of both null and alternative hypotheses (see also [19, 20, 21]). In this regard, Bayesian testing procedures differ from classical testing procedures, in which one can only fail to reject the null hypothesis [e.g., 22], by allowing researchers to quantify evidence in favor of true null

hypotheses, which can reflect the presence of an invariance or lack of an effect. In the Bayesian paradigm, the posterior odds in favor of an alternative hypothesis H_1, based on data $\mathbf{x}$, can be expressed as the product of the Bayes factor and the prior odds in favor of H_1; that is

$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds}, \tag{1.1}$$

or

$$\frac{\mathbf{P}(H_1 \,|\, \mathbf{x})}{\mathbf{P}(H_0 \,|\, \mathbf{x})} = \frac{m_1(\mathbf{x})}{m_0(\mathbf{x})} \times \frac{\mathbf{P}(H_1)}{\mathbf{P}(H_0)}. \tag{1.2}$$

It is important to note that this equation can be interpreted from both a frequentist and subjective view of probability. From the frequentist perspective, all probabilities can interpreted as the limiting proportion of the occurrence of an event. That is, if the null hypothesis H_0 is repeatedly sampled with probability $\mathbf{P}(H_0)$ (or H_1 with probability $\mathbf{P}(H_1) = 1 - \mathbf{P}(H_0)$), and data $\mathbf{x}$ is generated according to $m_0(\mathbf{x})$ (or $m_1(\mathbf{x})$), then the posterior probability that data was generated under H_1, for a given $\mathbf{x}$, converges in probability to

$$\mathbf{P}(H_1 \,|\, \mathbf{x}) = \frac{\text{BF}_{10}(\mathbf{x})\, \mathbf{P}(H_1)}{\mathbf{P}(H_0) + \text{BF}_{10}(\mathbf{x})\, \mathbf{P}(H_1)}, \tag{1.3}$$

where $\text{BF}_{10}(\mathbf{x}) = m_1(\mathbf{x})/m_0(\mathbf{x})$ is the Bayes factor in favor of H_1.

When Bayesian methods are applied to Null Hypothesis Significance Tests (NHSTs), controversy arises in the "subjective" specification of two quantities in these equations. First, the prior odds in favor of H_1 must be specified. This specification is equivalent to specifying either the prior probability of the alternative hypothesis, $P(H_1)$, or the prior probability of the null hypothesis, $P(H_0)$, since $P(H_0) + P(H_1) = 1$. A simple approach to setting the prior odds is to assume $P(H_0) = P(H_1) = 0.5$, leading to prior odds of 1.0. However, recent evidence gleaned from analyses of replicated experiments suggests that the prior odds in favor of the alternative hypotheses studied in psychology and other social sciences might be closer to $1/9$ [9, 23]. Although it is necessary to set a value of the prior odds in order to calculate the posterior odds, evaluation of

the prior odds is not considered further here. Instead, we encourage researchers to perform their own sensitivity analyses to evaluate how various assumptions regarding the prior odds affect the posterior odds for a given Bayes factor.

The second point of controversy arises in the definition of the marginal density of the data under the alternative hypothesis, given by

$$m_1(\mathbf{x}) = \int_{\Theta} f(\mathbf{x} \,|\, \theta)\pi_1(\theta) d\,\theta. \tag{1.4}$$

Here $\pi_1(\theta)$ represents the prior density for the parameter of interest θ under the alternative hypothesis, i.e., the alternative prior density. (A more detailed description of the Bayesian hypothesis testing framework may be found in, for example, [24] or [19].) In NHSTs, the quantity $m_0(\mathbf{x})$ simply represents the sampling density of the data, say $f(\mathbf{x} \,|\, \theta_0)$, evaluated at the parameter value that defines the null hypothesis, θ_0.

In Chapter 3 we describe a new approach to specifying alternative hypotheses in Bayesian tests of a normal mean or difference between means. Chapter 4 extends this idea and propose a similar approach to specifying alternative hypotheses in Bayesian tests of a proportion or difference between proportions. The approach is based on the use of non-local alternative prior densities (NAPs; [25]). A NAP is a density that exactly equals 0 at parameter values that are consistent with the null hypothesis. Tests specified with NAPs offer several advantages over tests defined with alternative hypotheses based on local priors. The NAPs

- can accumulate stronger evidence for true null hypotheses.
- They achieve comparable or stronger evidence for true alternative hypotheses,
- The sequential tests constructed using NAPs have smaller Average Sample Number (ASN).
- They more accurately reflect the prior belief that under the alternative hypothesis the tested parameter does not equal a value specified under the null hypothesis. This makes the prior specification logically consistent.

1.3 Latent Dynamic Modeling of Networks

Recent technological advancements in diverse areas of studies have made the availability of structured data in the form of networks increasingly popular. A simple dynamic network data is observed from a fixed set of individuals over a time period. Because the dynamics involve the same set of individuals, a key interest in these applications is to take their network structure into account in the modeling. Often the continuous network data that we observe has excessive zeros as observations. This can occur for a variety of reasons. For example, in microbiome data this occurs due to limitations of instruments used for a continuous measurement. Here zeros represent the measuring thresholds in those instruments. Another relevant example is the bilateral trade data where the observed zeros represents absence of trades between country pairs. Other examples include functional connectivity network among widespread brain regions, interactions between people in a social network, email communication networks, citation network among research articles or authors, network of co-purchased products, and bilateral trade flows among countries. This highlights the importance of the observed zeros and suggests taking this into account in the model.

A rapid rise in the network data in many scientific fields have resulted in increased renewed attention to the static and dynamic modeling of networks. Although the existing methods in the literature have been important in setting the premise of network models, they can be improved upon in several aspects. In this research, we motivate ourselves from the bilateral trade flows observed among 29 countries from 1994 to 2013 specific to the apparel industry. The presence or absence of trades and the trade volumes in the presence of trades are observed between each pair of 29 countries. We refer to the network data containing the presence or absence of trades as the binary network. The network data containing the trade volumes in presence of trade occurrences is referred to as the continuous network. Besides the presence of a network structure of the countries, some features of additional interest in the data are: (1) dynamic evolution of the network structure influencing both binary and continuous networks, (2) an abundance of unobserved trades among many country pairs (henceforth refered as the zero-inflated network), (3) available covariates spe-

cific to countries and pairs of countries. Several methods relying on the Gaussian random walk on the latent positions have been proposed to account for the dynamic evolution [26, 27]. But this often restricts the dynamic dependence to a Markov structure. Individual strategies exist in the literature that can separately model a binary or continuous network. In the context of bilateral trade flows, [28] proposed independently modeling binary and continuous networks sequential at each time point. This approach essentially assumes that there is one stochastic process that governs the incidence of trade and another that governs the volume of trade. But this can be inefficient when the proportion of presence and absence of trades become unbalanced. Also, it is counter-intuitive to assume that two independent underlying processes are responsible for the two networks as both networks involve the same set of countries.

In Chapter 5, we propose the *Hurdle Network Model* for zero-inflated network data with two key modifications. First, we assume there is a single stochastic process which governs both the binary and the continuous networks. More precisely, we assume the probability that a trade is present in a binary network is a strictly increasing function of the mean process in a continuous network. This lets us jointly model the two networks. Second, we assume node-specific latent attributes corresponding to each country and we *apriori* assume a dynamic shrinkage process on them independently across countries for modeling [29]. This lets us jointly model their dynamic evolution using continuous scale mixtures of Gaussian distributions in a global-local framework. In latent space, this performs desirable shrinkage as global-local priors, while providing local adaptivity when necessary. This allows for an adaptive way of modeling trend in a time series data.

2. A MODIFIED SEQUENTIAL PROBABILITY RATIO TEST

2.1 Introduction

Experimental science relies on controlled experiments that test whether effects predicted by a scientific theory can be produced and measured in laboratory settings. Observational science is based on measuring outcomes as they occur naturally, without experimental intervention. In practice, measured outcomes from both observational studies and experiments are subject to random variation and measurement error. For this reason, hypothesis testing procedures must be employed to determine whether data support or do not support a hypothesized effect. In the classical hypothesis testing paradigm, two types of errors are considered when making this assessment. Type I error occurs when the null hypothesis of "no effect" is rejected when the hypothesized effect does not exist. To limit claims of false discovery, hypothesis testing procedures are commonly designed so that the probability of a Type I error (i.e., α) is limited to be less than a prespecified value, often 0.05. Type II error occurs when we fail to reject the null hypothesis when the hypothesized effect does exist (the probability of a Type II error is denoted by β).

Recent concerns over the replicability of scientific studies have led to calls to move away from p values and significance testing [4, 5, 6, 7]. However, p values and significance testing continue to play critical roles in many areas, including genomics, high-energy physics, and clinical trials. An examination of recent articles in prominent psychology journals also suggests that p values and significance testing continue to play an important role in psychological research [8, 9]. Elsewhere, we have proposed to address the limitations of p values by reducing the significance thresholds required for declaring a positive finding from $\alpha = 0.05$ to $\alpha = 0.005$ [10, 11]. While this change would improve the replicability of scientific claims of discoveries, it would also increase the costs of conducting studies because larger sample sizes would be required if similar controls on Type II error probability were maintained.

This chapter describes a modification of the sequential probability ratio test (SPRT) of [12]

that reduces the sample sizes required to achieve specified Type I and Type II error probabilities. The modified design can be applied to many studies conducted in the social and natural sciences in which the goal is to establish the existence of a hypothesized effect. Implicit in this goal is the detection of effects that are not arbitrarily close to zero (or the null value of the parameter). In this regard, the proposed design differs from recent developments of sequential procedures designed to estimate various effect sizes, such as standardized mean differences, correlation and regression coefficients, and coefficients of variation [13, 14, 15].

We propose the Modified Sequential Probability Ratio Test (MSPRT) for testing a point null hypothesis against a one or two-sided alternative hypothesis. In designing these tests, we objectively set alternative hypotheses. The alternative hypotheses we propose are based on uniformly most powerful Bayesian tests (UMPBT's) or approximate UMPBT's [11, 16]. Details regarding UMPBT's appear in Section 2.3. We note that exact UMPBT's are known only for one-parameter exponential family models and tests for the non-centrality parameters of chi-squared statistics [17]. Approximate UMPBT's are known for *t* tests. Thus, a limitation of the MSRPT is that is applicable primarily to z and *t* tests, tests of binomial proportions and Poisson means, and chi-squared tests.

For this class of tests, empirical evidence suggests that MSPRT's require sample sizes that can be less than 50% of the sample size that is required in corresponding fixed designs when the null hypothesis of no effect is true, and sample sizes that can be 20% smaller when alternative hypotheses are true. In general, the sample size savings accrued by the use of the MSPRT depends on the test statistic chosen and the targeted Type I and II error probabilities for the test. Empirical studies illustrating such savings are described in Section 2.5. Theoretical support for these findings is provided in [18], where approximate formulae for the average sample number (ASN) and operating characteristics for truncated SPRTs are derived. These results approximate discrete time stochastic processes (representing the observed sequential tests) by Brownian motion or Wiener processes, which are continuous time stochastic processes. For sufficiently large sample sizes, these processes provide approximate operating characteristics and ASN's for truncated SPRT's. In the case of one- and two-sample z tests, the underlying assumptions required in deriving those for-

mulae apply to the MSPRT, and the approximate values from these results agree with our empirical findings. Specific details regarding this connection appear in Section 2.5.1. As noted by [18], this theoretical result "leads to appreciable qualitative insight; and quantitatively it does provide a first, crude approximation which can often be used as a basis for subsequent refinement."

The remainder of this chapter is organized as follows. Section 2.2 reviews sequential hypothesis testing procedures. In Section 2.3, we define MSPRT's, and in Section 2.4 we describe R code that can be used to implement them. In Section 2.5 we present numerical findings from simulation studies, and compare the performance of the MSPRT to group sequential designs and sequential Bayes factors [2]. Section 2.6 complements Section 2.5 by applying the MSPRT to the gambler fallacy study data [30] collected in the Many Labs 1 project [31]. Finally, we summarize our findings in Section 2.7.

2.2 Sequential Testing Procedures

In contrast to fixed sample size designs, sequential testing procedures provide a rule for stopping a study after observing individual participants or groups of participants. A sequential testing procedure specifies a rule that decides, after a group of participants has been measured, whether to (i) continue to collect data, (ii) stop data collection and reject the null hypothesis, or (iii) stop data collection and reject the alternative hypothesis.

Sequential testing procedures have not previously found widespread application in behavioral and social science research. However, the statistical theory for these tests has been developed extensively since their introduction by Wald in the 1940s. For a comprehensive review of statistical theory underlying these procedures, see [18]. Most applications have occurred in item response theory (IRT) and computer adaptive test designs, where sequential tests are often used to terminate IRT-based adaptive classification tests [32, 33]. Other recent applications include an item selection algorithm in a binary IRT model [34] and an extension to Bayesian hypothesis testing, called "Sequential Bayes Factors," that provides an optional stopping rule for multiple testing [2]. From a theoretical point of view, a bound for the expected stopping time (i.e., the test length) was obtained in adaptive mastery tests for dependent data [35].

The SPRT is one of the most widely known sequential testing procedures [12, 36, 37, 38, 39, 40]. This test is based on comparing the likelihood ratio between a simple (i.e., point or precise) null hypothesis and a simple alternative hypothesis, and stopping data collection as soon as the likelihood ratio strongly supports one of the two.

To illustrate this procedure in more detail, suppose that independent data values are collected sequentially. Denote these values by $x_1, x_2, \ldots$. Suppose further that the null hypothesis implies that the probability density function describing a single data value x_i is $f(x_i \,|\, \theta_0)$, and that the alternative hypothesis implies that the probability density function is $f(x_i \,|\, \theta_1)$. Then the likelihood ratio in favor of the alternative hypothesis based on the first n observations is defined as

$$L(\theta_1, \theta_0; n) = \prod_{i=1}^{n} \frac{f(x_i; \theta_1)}{f(x_i; \theta_0)}. \tag{2.1}$$

To simplify notation, we denote $L(\theta_1, \theta_0; n)$ by L_n.

Heuristically, the SPRT keeps track of the likelihood ratio L_n as data accumulate, and stops the experiment as soon as the probability assigned to the data under one hypothesis significantly exceeds the probability assigned to the data by the other hypothesis.

More formally, the SPRT proceeds by comparing L_n, $n = 1, 2, \ldots$, to constants A and B, $A > B > 0$, as data from individual study participants are collected. The procedure stops when $L_n \geq A$ or $L_n \leq B$, or equivalently when L_n exits the interval (B, A) for the first time. The quantities A and B are defined as

$$A = \frac{1-\beta}{\alpha} \qquad \text{and} \qquad B = \frac{\beta}{1-\alpha}. \tag{2.2}$$

If $L_n \geq A$, the null hypothesis is rejected; if $L_n \leq B$, the alternative hypothesis is rejected. An important property of the SPRT is that it requires, on average, fewer participants to achieve its specified Type I and Type II error probabilities than any other test whose error probabilities are smaller than or the same as these [41].

A key limitation of the SPRT is that it requires the specification of both a null hypothesis and an

alternative hypothesis. Specifying an alternative hypothesis is not required in classical hypothesis tests when only Type I error probability constraints have been imposed. The proposed MSPRT addresses this limitation by implicitly deriving the alternative hypothesis from the design parameters according to pre-specified criteria. From a user's point of view, this eliminates the need to explicitly specify an alternative hypothesis, even though the procedure does, of course, directly depend on the alternative hypothesis that is used. For this reason, users should carefully consider the magnitude of the effect size implicit in the MSPRT to determine whether it represents a plausible alternative hypothesis. In this regard, the use of the MSPRT mimics classical experimental design procedures in which Type I and II error probabilities, sample size, and targeted effect size are balanced against each other to determine a suitable test design.

Another limitation of the SPRT is that the sample size required to complete a test cannot be determined prior to the start of data collection. In nearly all experimental settings, resources available for testing participants are limited and in observational studies the amount of the data that can be collected from a population is finite. This feature of the SPRT thus complicates the practical design of tests and is resolved by the MSPRT. An earlier modification of the SPRT, known as the truncated SPRT, was proposed by [42] to address this difficulty. However, this modification generally provides less statistical power than our proposed MSPRT. For instance, Tables 3.1 and 3.11 in [18] indicate that for the alternative effect size that provides 80% power in a fixed design test, the truncated SPRT's power is only 74%. By comparison, the MSPRT provides between 78-79% power at the same alternative. Further examples of this difference are provided in Section 2.5.1, where we describe similar differences in power for other effect sizes.

Modifications of the SPRT proposed to handle composite hypotheses are primarily of two types. One is known as the weighted SPRT and was proposed by [12]. This test replaces the likelihood ratio with the ratio of integrated likelihoods, weighted with respect to given weight functions for the respective hypotheses. The weight functions are determined by losses associated with incorrectly accepting various alternative hypotheses. The other type of modification is known as the generalized SPRT, which is based on the ratio of maximized likelihoods under the respective

hypotheses, and is similar to the generalized likelihood ratio (GLR) test [43].

Other extensions of the SPRT, the MaxSPRT and the sequential GLR test, were proposed for drug and vaccine safety surveillance [44, 45]. The goal of the MaxSPRT is to reject the null hypothesis (that a treatment is safe) if there is substantial evidence that a treatment is not safe. Like the SPRT, the MaxSPRT does not impose a bound on the maximum sample size N. In addition, the design does not allow early rejection of the alternative hypothesis. The sequential GLR was proposed to address these issues. Importantly, both tests are based on the GLR, and the alternative hypothesis used by them is the maximum likelihood estimate (MLE) of the parameter being tested. When the null hypothesis is true, the MLE converges to the null value, and as a consequence the tests never terminate a trial in favor of the null hypothesis. Furthermore, for sufficiently large N the tests can, in principle, reject null hypotheses for arbitrarily small effect sizes.

From a practical perspective, there are many hypothesis testing contexts where it is not feasible to implement a SPRT. For instance, a SPRT cannot be applied when data are not collected sequentially. Similarly, it cannot be applied when it is not possible to perform the evaluation as soon as participants are treated. Such is the case in clinical trials of new disease therapies, which are often conducted at multiple treatment centers. Collation of data across centers can be time consuming, and it can be difficult to convene review boards. In addition, patient outcomes are often not known for months or even years after a treatment has been administered. To address these challenges, group sequential designs have been developed to allow for the evaluation of patient outcomes only after groups of patients have been observed or at scheduled interim analysis times [46, 47, 48, 49, 50, 51]. [18, 52] provides detailed discussion of the termination of repeated significance tests for group sequential studies with a maximum sample size.

2.3 The Modified SPRT

To address the limitations of existing sequential tests, we propose a modified SPRT (MSPRT) in which

- the maximum sample size (N) required in a hypothesis test is fixed prior to the start of an

experiment, and

- the effect size defining the alternative hypothesis and used to sequentially compute the likelihood ratio L_n is derived from the size of the test α (Type I error probability), the maximum available sample size N, and the targeted Type II error probability, β.

Thus, N, α, and β are MSPRT design parameters that are fixed at the outset of the study. The effect size defining the alternative hypothesis is determined from these values. Given these values, the MSPRT is defined in a manner similar to Wald's initial proposal.

To objectively set the alternative hypothesis in the MSPRT, we find the uniformly most powerful Bayesian test (UMPBT) or the approximate UMPBT that matches the rejection region of a classical test of size α with a sample size of N [16]. Under fixed designs, UMPBT's are tests that maximize the probability that the Bayes factor in favor of the alternative hypothesis exceeds a specified threshold over the class of all alternative hypotheses. [16] showed that such tests can be obtained by assuming a point alternative and then maximizing the probability mentioned above with respect to such alternatives. The optimum value of the point alternative is defined as the UMPBT alternative. We defer a more detailed description of UMPBTs to Sections A1–A3 of the supplemental materials. The key feature of an UMPBT relevant to our purpose is that it provides an automated procedure for defining an alternative hypothesis against which the null hypothesis is tested. For sampling densities that belong to the class of one-parameter exponential family models (including z tests, tests for proportions, and tests of means of Poisson counts), UMPBTs exist. For other sampling densities, and in particular for t tests, approximate UMPBTs exist. In many cases, the values of the parameter that define the alternative hypotheses in these tests are approximately equal to the maximum likelihood estimate of the parameter obtained from data that lie on the boundary of the rejection region of the test.

To illustrate a simple UMPBT, consider a size α z test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ based on N samples from a normal population with an unknown mean θ and known standard deviation σ. For this problem, the UMPBT alternative hypothesis is $\theta = \theta_0 + z_\alpha \sigma/\sqrt{N}$, where z_α is the $100(1-\alpha)$th quantile of a standard normal distribution.

Table 1 provides the UMPBT alternatives that can be used in some common, one-sided null hypothesis significance tests. In this table, definitions of alternative hypotheses are determined by the maximum sample size N for the z test and test of a binomial proportion. For the t test, it also depends on n, the currently observed sample size. These alternatives are used to compute the likelihood ratio at each step. Thus for a t test, the alternative hypothesis used to compute the likelihood ratio changes after each data point is collected and a new estimate of the observational variance is obtained. For the z and t tests, the UMPBT alternatives are point alternatives. The alternative for the test of a binomial proportion is a mixture distribution of two proportions; a mixture density is used to achieve more accurate Type I error probability control due to the discrete nature of the binomial distribution.

To understand the nature of this mixture distribution, it is necessary to introduce additional notation. Denote the cumulative distribution function (cdf), inverse cdf, and the probability mass function of a binomial distribution with denominator N and success probability θ by $F(\cdot; N, \theta)$, $\bar{F}(\cdot; N, \theta)$, and $f(\cdot; N, \theta)$, respectively. Given N and α for a right one-sided test of the probability θ, define the cut-off point c_0 in a fixed design test by

$$c_0 = \inf \left\{ c = 0, 1, \cdots, N \middle| \bar{F}(c; N, \theta_0) \leq \alpha \right\}.$$

For $\theta \in [0, 1]$ and $\delta > 0$, let

$$h_N(\theta, \delta) = \frac{\log \delta - N \left[\log(1 - \theta) - \log(1 - \theta_0) \right]}{\log \left(\frac{\theta}{1 - \theta} \right) - \log \left(\frac{\theta_0}{1 - \theta_0} \right)},$$

and define $\theta(\delta) = \arg\min_{\theta > \theta_0} h_N(\theta, \delta)$. With these ingredients, we define the UMPBT alternative as the mixture distribution

$$\theta \sim \psi_R \, I_{\theta = \theta_{R,L}} + (1 - \psi_R) \, I_{\theta = \theta_{R,U}},$$

where $I_{a=b}$ is 1 if $a = b$ and 0 otherwise. Also, $\theta_{R,L} = \theta\big(\delta_{R,L}\big)$ and $\theta_{R,U} = \theta\big(\delta_{R,U}\big)$, where $\delta_{R,L}$ and $\delta_{R,U}$ satisfy

$$h_N\Big(\theta\big(\delta_{R,L}\big), \delta_{R,L}\Big) = c_0 - 1, \text{ and } h_N\Big(\theta\big(\delta_{R,U}\big), \delta_{R,U}\Big) = c_0,$$

and $\psi_R = [\alpha - \bar{F}(c_0; N, \theta_0)]/f(c_0; N, \theta_0)$. A similar derivation can be applied to left one-sided tests. Further details are provided in [11, 16] and Section A3 in the supplemental document.

In practice, of course, researchers should examine the design parameters of a MSPRT before the sequential design is initiated. That is, the alternative hypothesis generated by the MSPRT in order to obtain the targeted Type I and Type II error probabilities should be inspected, as should the actual error probabilities achieved by the test design. If the implied effect size is either unreasonably large or substantively unimportant, then investigators should reconsider the maximum sample size and error probability controls that were specified.

In the case of one-sided hypothesis testing, given the alternative hypothesis obtained from the UMPBT or approximate UMPBT, Wald's SPRT is conducted either until the likelihood ratio (z and t tests) or the weighted likelihood ratio (proportion test) exits the interval (B, A) or until N samples (e.g., study participants) have been tested. The values of A and B for the MSPRT are the same as those used in Wald's test and, as noted previously, are given by

$$A = \frac{1-\beta}{\alpha} \qquad \text{and} \qquad B = \frac{\beta}{1-\alpha}.$$

If no decision has been reached after exhausting N samples, a threshold γ is determined numerically so that the Type I error probability of the test equals α for continuous data and is less than or equal to α for discrete data. If $L_N \geq \gamma$, the null hypothesis H_0 is rejected and the experiment is terminated. Otherwise, if $L_N < \gamma$, the alternative hypothesis H_1 is rejected and the experiment is terminated.

The extension of the MSPRT for two-sided tests is accomplished by simultaneously running two one-sided tests of size $\alpha/2$. Before reaching the maximum sample size N, the test terminates

by (a) rejecting H_0 when either of the tests reject H_0, or (b) by not rejecting H_0 if both the tests reject H_1. If the test continues to the maximum sample size N, then a common termination threshold, γ, is determined so as to maintain the desired Type I error probability of the test. The design parameter γ is chosen to be as small as possible while still maintaining the specified size of the test, α. If $L_N \geq \gamma$ for either of the tests, the null hypothesis is rejected. Otherwise, the test rejects the alternative hypothesis.

In practice, it may be useful to examine the value of L_n at the termination of an MSPRT. This value represents the likelihood ratio between hypotheses based on all accumulated data and may be of particular interest when a test terminates after the maximum sample size has been reached. Of course, an advantage of formal hypothesis testing procedures is that they encourage investigators to design experiments that have a reasonably high probability of providing "signficant" evidence in favor of a scientifically important effect. At the end of a parametric hypothesis test, it is usually possible to compute the likelihood ratio in favor of the MLE over the null parameter value. In the particular case of the MSPRT, the investigator is able to go a step further and report the Bayes factor in favor of an alternative hypothesis which was considered scientifically acceptable before the experiment was undertaken. The MSPRT thus encourages the design of tests that will lead to Bayes factors (or likelihood ratios) that differ substantially from 1.0, and they do so with smaller sample sizes than are required in fixed design tests. The values of the likelihood ratio L_n are provided by the **`MSPRT`** software described in Section 2.4.

Figure 1 summarizes the process for conducting a MSPRT for a one-sided test of a normal mean or a population success probability.

2.4 Implementation

Software to implement the MSPRT is available from the CRAN R software depository at `https://cran.r-project.org/web/packages/MSPRT/index.html` and on GitHub at `https://github.com/sandy-pramanik/MSPRT`. The software can be used to perform one-sample proportion tests, and one- and two-sample z and t tests. To design and implement a MSPRT, a user must provide a null hypothesis (θ_0), a direction of the alternative hypothesis (right,

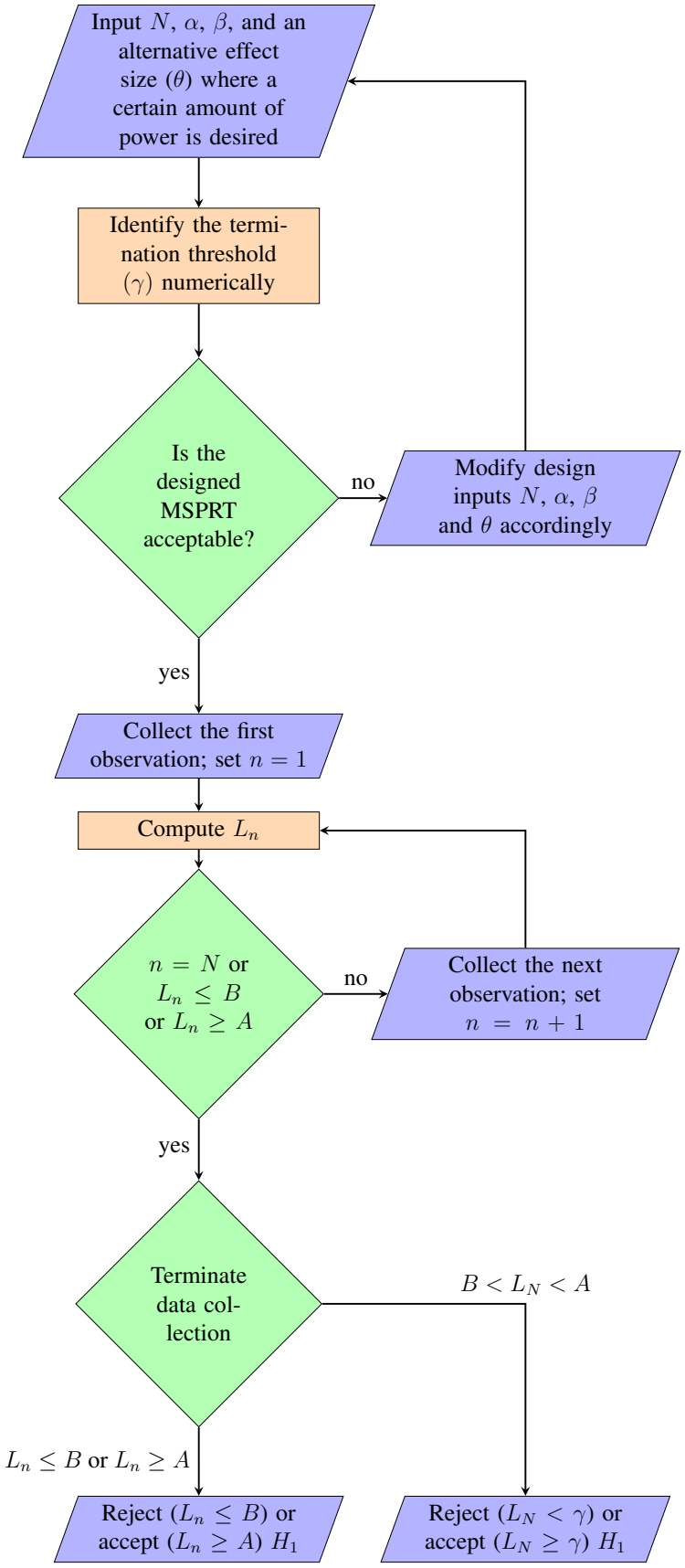


Figure 2.1: A flow chart representing the MSPRT procedure.

Table 2.1: UMPBT alternatives for one-sided tests

Test	H_0	H_1	UMPBT alternative
z test	$\theta = \theta_0$	$\theta > \theta_0$	$\theta = \theta_0 + z_\alpha \frac{\sigma}{\sqrt{N}}$
		$\theta < \theta_0$	$\theta = \theta_0 - z_\alpha \frac{\sigma}{\sqrt{N}}$
t test	$\theta = \theta_0$	$\theta > \theta_0$	$\theta = \theta_0 + t_{\alpha;N-1} \frac{s_n}{\sqrt{N}}$
		$\theta < \theta_0$	$\theta = \theta_0 - t_{\alpha;N-1} \frac{s_n}{\sqrt{N}}$
Test for proportion	$\theta = \theta_0$	$\theta > \theta_0$	$\theta \sim \psi_R \, I_{\theta=\theta_{R,L}} + (1 - \psi_R) \, I_{\theta=\theta_{R,U}}$
		$\theta < \theta_0$	$\theta \sim (1 - \psi_L) \, I_{\theta=\theta_{L,L}} + \psi_L \, I_{\theta=\theta_{L,U}}$

Note. For one-sample z and t tests, UMPBT alternative hypotheses have closed-form expressions. For one-sample tests of proportions, (non-randomized) MSPRT's can be used to more accurately achieve Type I error probability control, but a mixture distribution is required as the alternative in this setting. Details for obtaining explicit values for the alternative using the R package **MSPRT** are described in Section A4.3 of the supplemental materials. The $100(1-\alpha)$th quantiles of a standard normal distribution and central t distribution with $(N-1)$ degrees of freedom are denoted by z_α and $t_{\alpha;N-1}$, respectively, and σ denotes the known population standard deviation in a z test, whereas s_n refers to the sample standard deviation (with divisor $(n-1)$) based on n observations.

left or two-sided), maximum available sample size (N), and pre-specified error probabilities (α and β). Given these design parameters, the **R** package **MSPRT** provides test results based on sequential entry of outcome data. Detailed illustrations are provided in Section A4 of the supplemental materials.

2.5 Simulation Studies

This section analyzes the performance of the MSPRT through simulation studies. For simplicity, we first investigate the performance in one-sample tests for a binomial proportion, z tests, and t tests. Next, we compare the performance of the MSPRT with group sequential (GS) designs. The extension of MSPRT designs to two-sample z and t tests is immediate. We also compare the performance with Sequential Bayes Factors (SBF) [2]. Finally, we discuss the potential benefit that is offered by MSPRT designs when we decrease the p-value threshold for declaring statistical significance from 5% to 0.5%. Throughout the section, 10^6 replications were used to summarize

the performance of the MSPRT.

2.5.1 Performance in one-sample tests

We examine one-sample tests for a binomial proportion, z tests, and t tests of size $\alpha = 0.05$ and $\alpha = 0.005$. For simplicity, we examine one-sided tests with alternative hypotheses of the form $H_1 : \theta > \theta_0$. We also assume that the targeted power of the test is 80% (i.e., $\beta = 0.2$). Two-sided tests, tests of alternative hypotheses of the form $H_1 : \theta < \theta_0$, and tests with different Type I or Type II error probabilities are handled similarly. We compare the resulting MSPRT's to standard fixed design tests having the same α level, sample size N and Type II error probability $\beta = 0.2$. Given N and α for fixed design tests, we define θ_a, the fixed design alternative, as the alternative parameter value that provides the specified β.

We now describe the simulation settings used for analyzing the operating characteristics and ASN's of the MSPRT.

For one-sample z tests, observations are assumed to be independent and identically distributed random samples from a normal distribution with unknown mean θ and known variance σ_0^2. To study the performance of the MSPRT under the null hypothesis, we generate observations from a $N(\theta_0, \sigma^2)$ distribution. For performance under H_1, we use a $N(\theta_a, \sigma_0^2)$ distribution with the fixed design alternative, θ_a, defined in the previous paragraph. We note that it is possible to numerically compute the operating characteristics and ASN of the MSPRT prior to the onset of an experiment. The simulation setup for the proportion test proceeds exactly as above where we simulate the data independently from a Bernoulli(θ) distribution. For our simulations, we use $\theta_0 = 0$ for the z test and $\theta_0 = 0.5$ for the proportion test (Section A4.2 in the supplement provides implementation details). For t-tests, θ_a is interpreted as the standardized effect size θ/σ.

Figure 2.2 illustrates the performance of the MSPRT for a one-sample t test of $H_0 : \theta = 0$ versus $H_1 : \theta > 0$. This plot provides the average proportion of the N samples required by a fixed design test for the MSPRT to achieve nearly equivalent Type I and Type II error probabilities. Type I error probabilities are exactly maintained at targeted levels. Type II error probabilities for the MSPRT's slightly exceed the targeted value of 0.2, but never exceed 0.22.

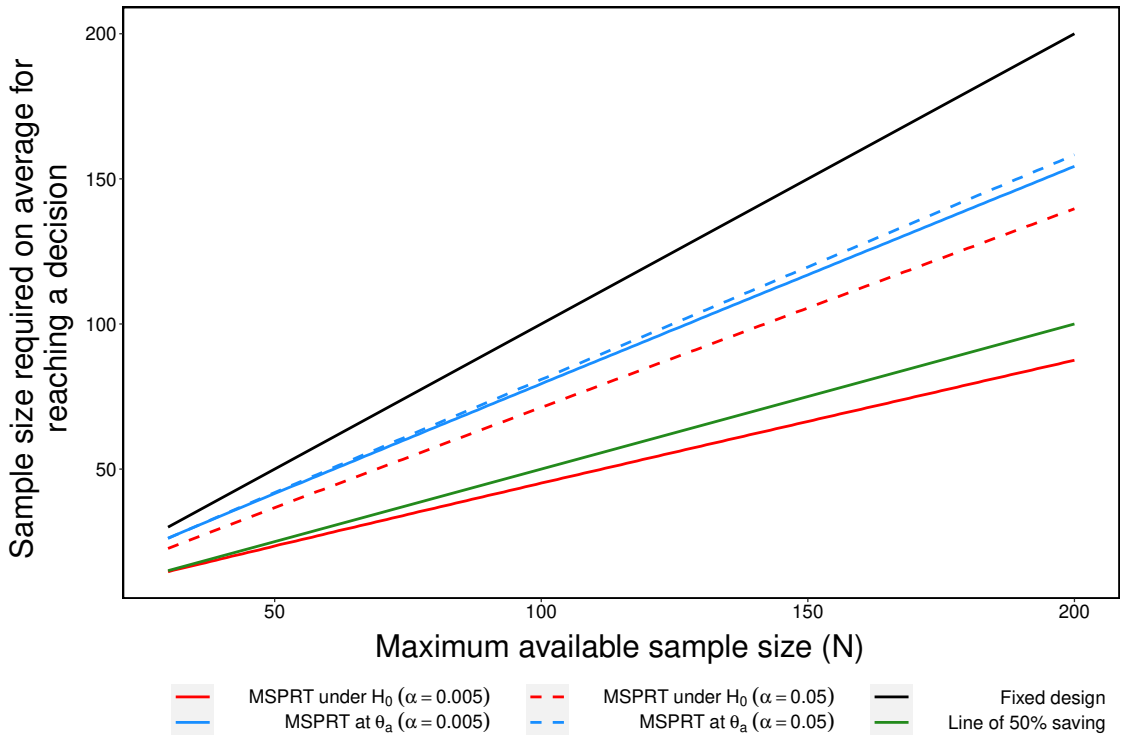


Figure 2.2: **One-sample t test that a population mean is 0.** Hypothesis test of $H_0 : \theta = 0$ vs. $H_1 : \theta > 0$. The population standard deviation is assumed to be unknown. Each curve in the plot represents the average number of samples, out of the maximum sample size (N), used before the MSPRT terminates in favor of the null or alternative hypothesis. The operating characteristics under the alternative are evaluated at the corresponding fixed design point alternatives.

The plot provided in Figure 2.2 for a one-sided t test is nearly indistinguishable from the corresponding plots obtained for one-sample z tests and tests of a binomial proportion; these plots are provided in the supplemental materials.

Two features of these plots are noteworthy. First, for Type I error probabilities of $\alpha = 0.005$, the average sample size required by the MSPRT is less than 50% of the sample size required by the fixed design test when the null hypothesis is true. This finding holds for all three tests. Second, under the alternative hypothesis, the average sample size required for the MSPRT is typically about 80% of the sample size required for the corresponding fixed design test.

To provide a theoretical context for these findings, we note that [18] provides approximate formulae for power and the average sample number function for truncated SPRTs for large N (corollaries 3.45–3.47, page 55–57, Section 6 of Chapter III in [18]). In order to derive these results, a Brownian motion process was used to approximate the operating characteristics and ASN of truncated SPRTs. For $\alpha = 0.005$, the approximations predict that the average sample size required by the MSPRT under the null and the alternative hypothesis is approximately 40% and 70% of the fixed design sample size N, respectively, and the Type II error probability at the fixed design alternative is about 23%. These values match our empirical findings, but are based on approximating the discrete time scale of the MSPRT by the continuous time scale of Brownian motion. They also rely on an assumption that the test statistics are approximately normally distributed. As noted in [18], Brownian motion nonetheless furnishes a convenient way of analyzing properties of SPRTs while avoiding intractable probability calculations.

Figure 2.3 presents the distribution of the number of samples required by the MSPRT to reach a decision in a one-sample t test. As in Figure 2.2, the performance of the MSPRT is compared to the fixed design test having Type I error probability 0.005 and Type II error probability 0.20. The top panel is based on a maximum sample size $N = 30$, and the bottom panel on $N = 100$. From these figures, we see that under H_0 the MSPRT reaches a decision before the maximum sample size is accumulated in about 85% of tests. This proportion slightly increases when N is 100. Under the fixed design alternative, the MSPRT terminates in about 60% of tests before the maximum sample

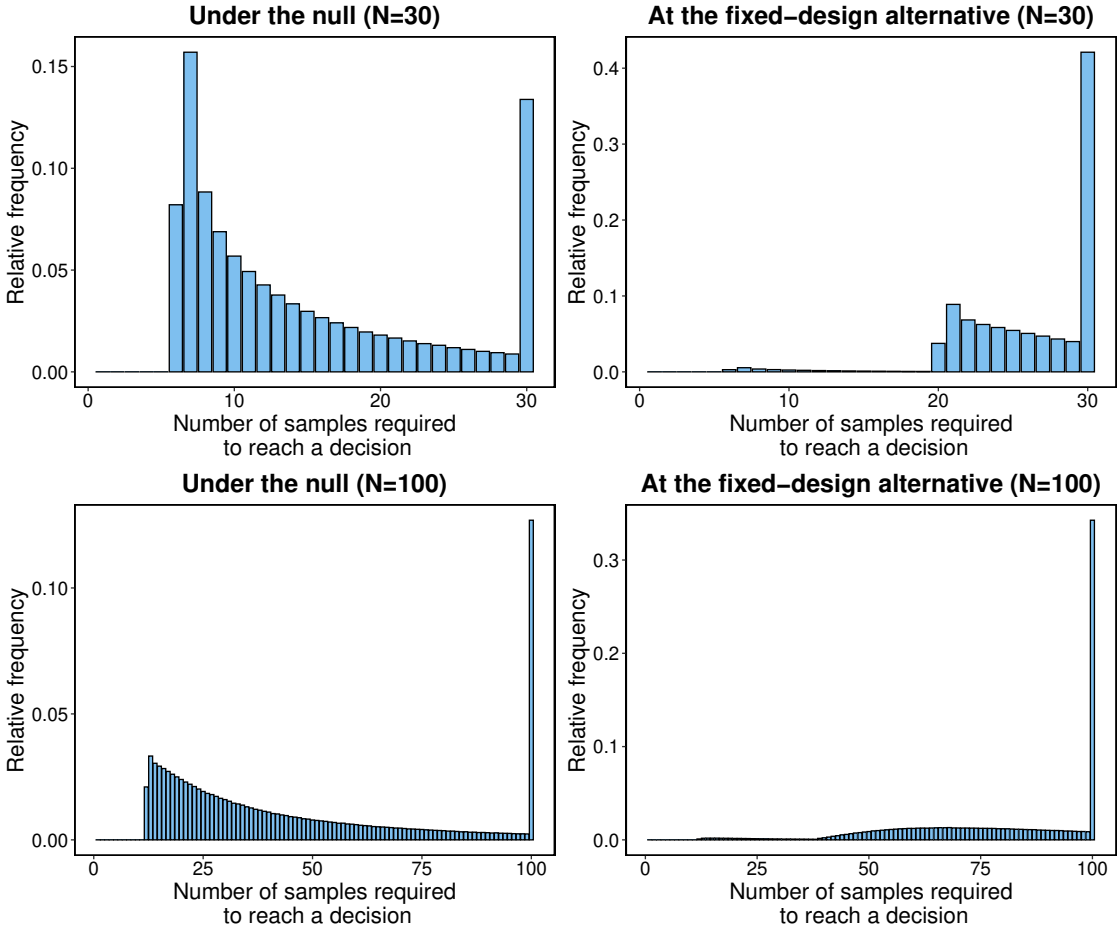


Figure 2.3: **One-sample t test that a population mean is 0.** Hypothesis test of $H_0 : \theta = 0$ vs. $H_1 : \theta > 0$ at $\alpha = 0.005$ and $\beta = 0.2$. The population standard deviation is assumed to be unknown. The barplots represent the distribution of sample size required by the MSPRT for reaching a decision under H_0 and at the corresponding fixed design alternative θ_a. The fixed design alternatives, which provide 20% Type II error probability, are approximately 0.66 for $N = 30$ and 0.35 for $N = 100$.

size is reached when $N = 30$, and in about 65% of tests when $N = 100$. Though not displayed, similar results are obtained for one-sample proportion and z tests.

We also conducted a simulation study to analyze performance of the MSPRT at effect sizes other than the null and fixed-design alternatives θ_a. Specifically, we considered the right-sided one-sample t test with $\alpha = 0.005$ and $\beta = 0.2$. We again set $N = 30$ (left panel in Figure 2.4) and $N = 100$ (right panel). To analyze the operating characteristics and ASN of resulting tests, we generated data using effect sizes that corresponded to fixed design tests having Type II error probabilities $0.05, 0.1, \cdots, 0.9$. The effect sizes range from 0.25 to 0.82 for $N = 30$, and 0.13 to 0.43 for $N = 100$. Figure 2.4 presents the operating characteristics and ASN at these effect sizes. From the top panel we see that the Type II error probabilities of the MSPRT at these effect sizes are almost identical to the corresponding fixed design test (the red line almost coincides with the black line). Thus, the MSPRT achieves almost identical power to the fixed design test at a lower cost. The bottom panel in the same figure displays the ASN of the corresponding MSPRT's at the same effect sizes. The ASN's in this plot are about 70–80$\%$ of N when the Type II error probability of the fixed design tests is 0.05 or smaller. As the effect size decreases and gets closer to the null value, the ASN's increase until they reach a maximum of about 85–90$\%$ of N for Type II error probabilities near 0.5. The ASN then decreases to approximately 40–45$\%$ of N near the null value. The performance of the MSPRT for other tests is similiar to that depicted in Figure 2.4.

2.5.2 Comparison of MSPRT and GS designs

We next compared the MSPRT to GS designs using the **R** software package **gsDesign** [53]. We used the default Hwang–Shih–DeCani error spending function as the sequential stopping criterion. For illustration, we assumed the design had a total of 5 groups/stages (including interim and final analysis) with equal number of subjects entered at each stage. As before, we varied N from 30 to 200, considering two choices of the Type I error probability (0.05 and 0.005), and set the Type II error probability at 0.2.

The **gsDesign** function obtains the critical boundaries of the GS design by assuming a standard normal test statistic. We therefore conducted a right-sided one-sample z test of the form

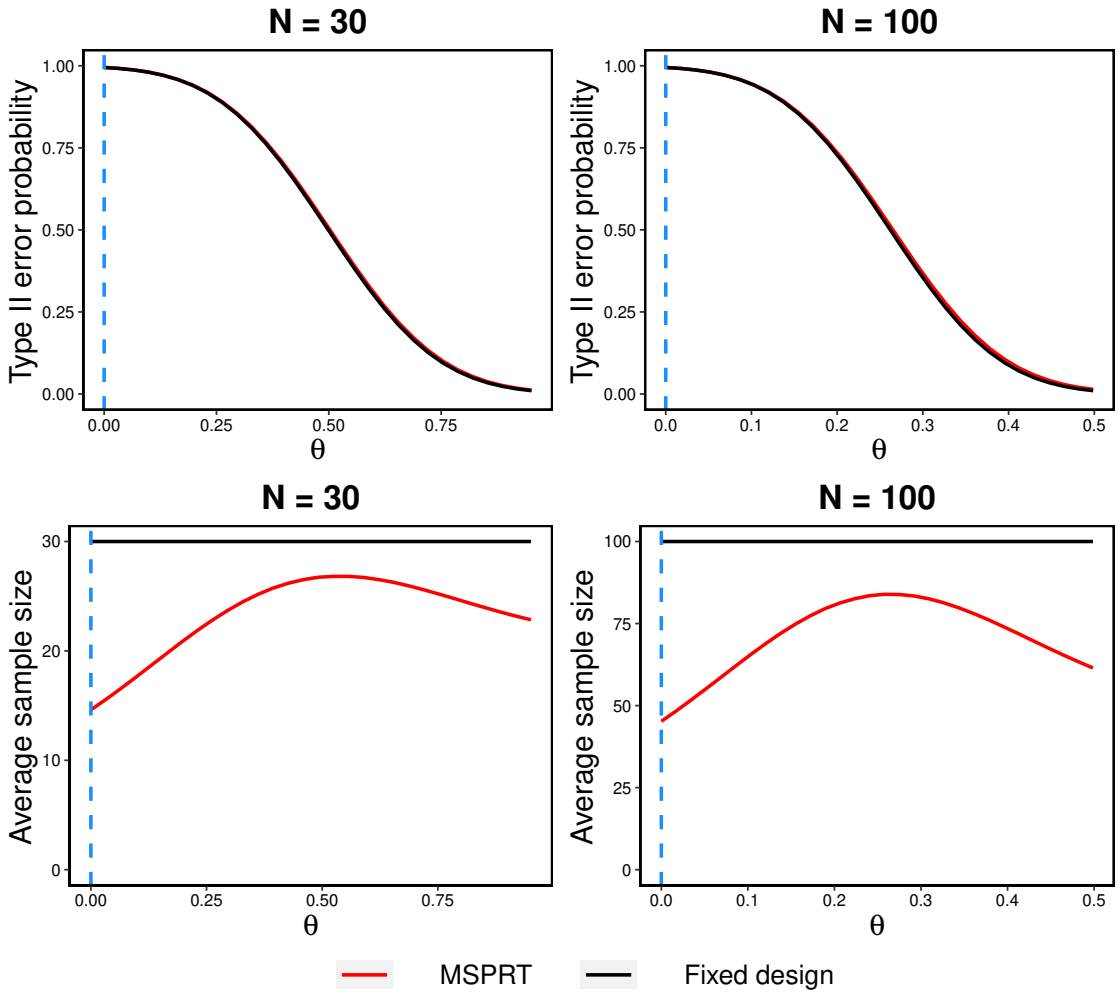


Figure 2.4: **One-sample t test that a population mean is 0.** Hypothesis test of $H_0 : \theta = 0$ vs. $H_1 : \theta > 0$ at $\alpha = 0.005$ and $\beta = 0.2$. The population standard deviation is assumed to be unknown. The above plots compare the Type II error probability and the average sample size of the MSPRT and the fixed design tests for a varied range of alternative effect sizes. The fixed design alternatives, which provide 20% Type II error probability, are approximately 0.66 for $N = 30$ and 0.35 for $N = 100$.

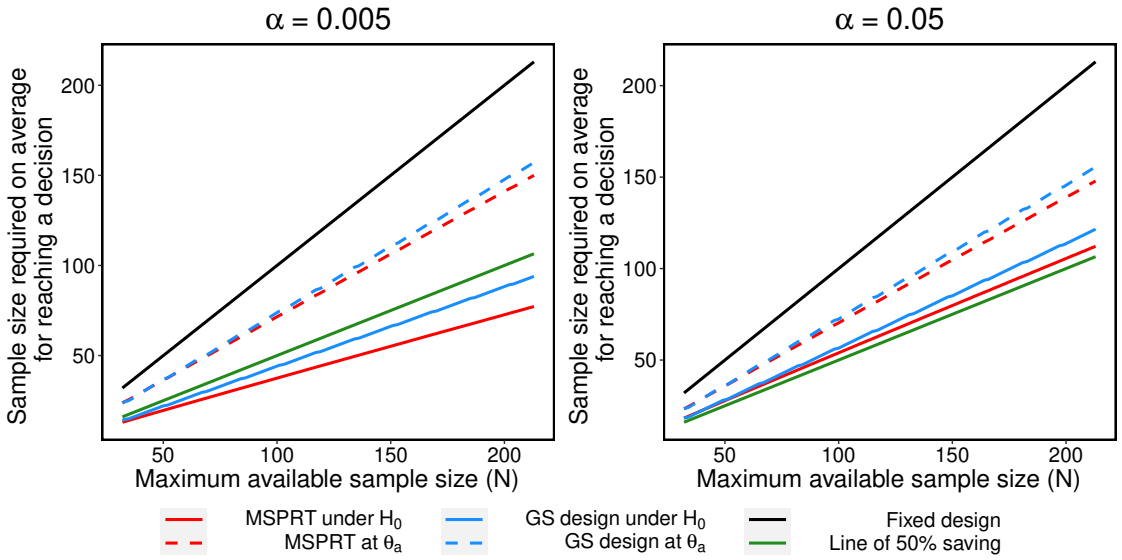


Figure 2.5: **One-sample z test that the population mean is 0.** Hypothesis test of $H_0 : \theta = 0$ vs. $H_1 : \theta > 0$. Each curve in the plot represents the average number of samples, out of the maximum sample size (N), used before the MSPRT or the GS design terminates in favor of the null or alternative hypothesis.

$H_0 : \theta = 0$ vs $H_0 : \theta > 0$ with a known variance of 1. For a comparison under H_1, we focused at the fixed design alternative (θ_a) corresponding to design parameters N, α and β. After α, β and θ_a are specified in the **`gsDesign`**, the software designs a test by exactly spending α and β (at θ_a) but with a slightly larger maximum sample size (than N). To make a fair comparison, we designed the MSPRT using this larger sample size as the maximum available sample size. We also adjusted the design parameters β and γ so that the designed MSPRT has approximately $1-\beta$ power (within 1%) at θ_a.

In Figure 2.5 we compare the average sample size used in the MSPRT and GS tests. For a varied range of N, the MSPRT achieves a uniformly smaller ASN than the GS design. Their performances are quite similar under both H_0 and θ_a when $\alpha = 0.05$, and at θ_a when $\alpha = 0.005$. A more visible difference can be seen under H_0 when $\alpha = 0.005$. At the higher significance level, the GS design uses about 44% of the maximum available sample size. The MSPRT on an average uses about 3–8% fewer samples for the same Type I and Type II error probabilities. The difference in ASN becomes larger as the maximum available sample size increases.

2.5.3 Performance comparison between MSPRT and SBF in two-sample t tests

In this section we compare the performance of the MSPRT to the sequential Bayes factor (SBF; [2]). At each step of a sequential analysis, a SBF computes the Bayes factor under a Cauchy prior on the standardized effect size. The stopping boundaries are based on verbal labels for grades of evidence [19]. We note that SBF tests, like the SPRT, do not fix maximum sample sizes in advance.

We make this comparison for two-sided two-sample t tests because of their widespread application. Let θ_1 and θ_2 be the population means of two groups of subjects. Under the assumption that the observations from the underlying populations are normally distributed and their common variance is unknown, a two-sided two-sample t test compares the hypothesis $H_0 : \theta_1 - \theta_2 = 0$ against the alternative hypothesis $H_1 : \theta_1 - \theta_2 \neq 0$. To conduct this two-sided test with Type I error probability α, a MSPRT simultaneously performs two separate one-sided tests, each with Type I error probability $\alpha/2$. At each sequential step it (i) rejects H_1 if both the tests reject H_1, (ii) rejects H_0 if either test rejects H_0, or (iii) continues sampling.

To simplify exposition, we assume the maximum number of subjects available in both groups is equal and is denoted by N, and that sequential testing is performed so that one pair of subjects from each group are measured simultaneously. The total sample size for the experiment is thus $2N$.

Figure 2.6 presents a comparison between the two sequential procedures for testing $H_0 : \theta_1 - \theta_2 = 0$ against $H_1 : \theta_1 - \theta_2 \neq 0$. The performance under H_1 is examined at the corresponding fixed-design alternatives (θ_a). Figure 2.6 presents results for the right-sided alternative θ_a, the results for the left-sided alternative being similar. To implement the SBF test, we followed the recommendations of [2] and set the Cauchy scale parameter r equal to 0.707. The null and alternative boundaries for the Bayes factor were fixed at $1/6$ and 6, respectively, and the minimum sample size was set to 20 in each group. We assumed that a maximum of N samples were available for each group, and the sample sizes in the two groups were equal. If the SBF test did not reach a decision after accruing all subjects (i.e., $1/6 <$ SBF < 6), it was assumed that the test failed to reject the null hypothesis. Such outcomes thus decrease the Type I error of the SBF tests.

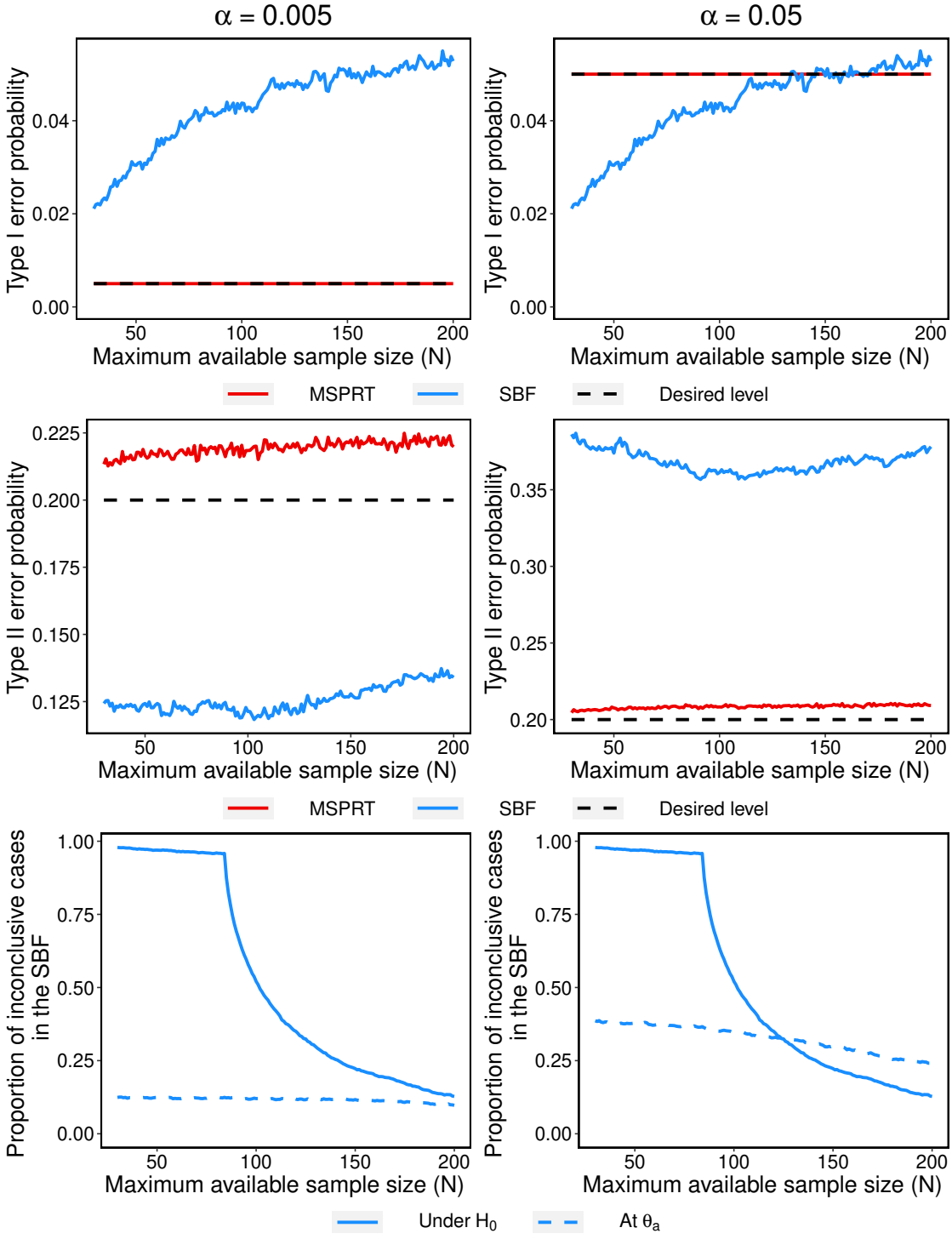


Figure 2.6: **Comparison of error probabilities for SBF and MSPRT tests.** Two choices for the targeted Type I error probabilities of 0.005 (left column) and 0.05 (right column) for the MSPRT are considered. For both the tests we varied the maximum available sample size (N) and compared the Type I (first row) and the Type II (second row) error probabilities achieved. The final column displays the proportion of inconclusive cases at the maximum sample size for the SBF.

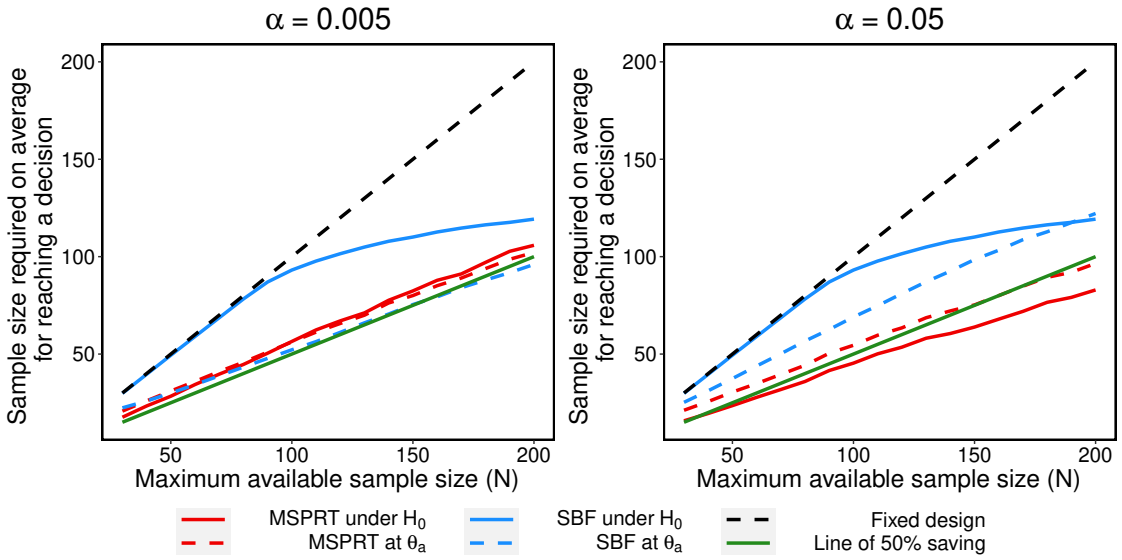


Figure 2.7: **Comparison of ASN for MSPRT and SBF.** This plot displays the proportion of the maximum sample size under various assumptions on null and alternative hypotheses for the MSPRT and SBF tests.

Because the goals and philosophy underlying the SBF and MSPRT are different, choosing the design parameters for the MSPRT to make a comparison to the SBF is difficult. For this reason, we choose two default settings for the MSPRT corresponding to $\alpha = 0.05$ and $\alpha = 0.005$, holding $\beta = 0.2$. In all comparisons, we assumed that pairs of observations were collected sequentially until each test terminated (possibly at the maximum sample size $2N$). We emphasize that the SBF is not intended to control either Type I or Type II error probabilities, so achieving these rates should not be regarded as a basis of comparison.

Figure 6 displays results for this comparison. The first row shows that that the MSPRT achieves its targeted Type I error probability for both tests. The Type I error achieved by the SBF is identical in both plots since the design parameters of that test did not change.

The second row of plots in Figure 6 displays the Type II error of each test, at the alternative targeted by the MSPRT. When $\alpha = 0.005$, the Type II error probability of the MSPRT is higher than the SBF, while it is lower for $\alpha = 0.05$. The Type II error probability for the SBF changes between plots because the alternative being tested has changed. The final row of this plot indicates

the proportion of tests that were inconclusive at the maximum sample size for the SBF test.

Because the SBF does not control error probabilities at pre-assigned values, additional care is needed to compare the ASN needed for each test. To make such a comparison, we therefore implemented the following procedure. At each N, we determined the (positive) value of the alternative hypothesis that provided 80% power in a fixed-design, two-sided t test with Type I error probability of either 0.05 or 0.005, against a null hypothesis of 0. Through simulation, we then determined the Type I and Type II error probabilities of the SBF (using the truncation rule described above). We then constructed the MSPRT with the same (within 1% numerical error) Type I and Type II error probabilities. This procedure allowed us to compare the average sample sizes of the two testing procedures with approximately similar error probabilities. The resulting comparison of the average sample sizes is presented in Figure 7.

For $\alpha = 0.005$, Figure 7 suggests that both tests require approximately the same ASN when the alternative hypothesis is true. However, the MSPRT is substantially more efficient when the null hypothesis is true. In both plots, the solid blue curve represents the ASN for the SBF when the null hypothesis is true, and this curve falls well above the corresponding solid red curve representing the ASN for the MSPRT.

The SBF test's use of a median-zero Cauchy prior to define the alternative hypothesis provides a partial explanation of these findings. This prior assigns significant mass around 0, the hypothesized effect size under the null. The Cauchy prior is a particular example of a local prior, and it is known that the evidence in favor of a true null hypothesis accumulates much slower than it does under a true alternative hypothesis when local priors are used [25]. To fix this asymmetric accumulation of evidence, Johnson and Rossell proposed non-local priors on effect sizes under alternative hypotheses which assign zero prior density to the null value. Since the UMPBT alternatives place all their mass at non-null effect sizes, they are non-local alternative priors and can thus be expected to accumulate evidence more rapidly in favor of true null hypotheses.

For tests based on fixed sample sizes, we note that UMPBTs (when they exist) are, by definition, the tests that provide the highest probability of exceeding a specified Bayes factor threshold.

2.5.4 Higher significance with similar sample sizes

We next examine the potential benefit that the MSPRT offers in offsetting the increase in the sample size that would be required if the bar for declaring a result "statistically significant" were moved from $p < 0.05$ to $p < 0.005$. Specifically, we compare the sample size required in standard fixed design tests at the 5% level to the average sample size required by the MSPRT at the 0.5% level.

If the null hypothesis is true, this comparison is straightforward. If not, care must be taken to make sure that the same alternative hypotheses are compared at both levels of significance in the fixed design and MSPRT design scenarios. To make this comparison, we determine the θ^* that achieves the targeted Type II error probability in a fixed design test of size $\alpha = 0.05$. For that θ^*, we next determine the N^* needed to achieve the same Type II error probability in a fixed design test of size $\alpha = 0.005$. We then set that N^* as the maximum sample size for the MSPRT.

Because the average sample size used in the MSPRT depends on whether the null or alternative hypothesis is true, and because we are interested in the long-run effect of implementing the MSPRT over many experiments, it is useful to examine the effect on the total sample size as the proportion of true null hypotheses is varied. Recent research suggests that this proportion is likely to be in the range 0.80–0.95 [9, 23].

In the case of a one-sample t test, Figure 2.8 displays the average multiple of the fixed 5% test's sample size N that is required to perform the MSPRT with size 0.5% as the proportion of tested null hypotheses π_0 is decreased from 1 (the dashed red line at the bottom) to 0.6 (the light blue line). Also displayed is the multiple of N that is required to achieve a Type I error probability of 0.005 in a fixed design test (the solid black line at the top). The latter multiple tends to fall between 1.89 and 2.14. Similar plots are obtained for one-sample z tests and tests of a binomial proportion; these plots are provided in the supplemental materials.

The key finding from Figure 2.8 is that MSPRTs for $\alpha = 0.005$ require, on average, essentially the same sample sizes that are required to conduct one-sided, fixed design tests for $\alpha = 0.05$ for tests designed to have Type II error probabilities of 0.2. We emphasize that such gains may not

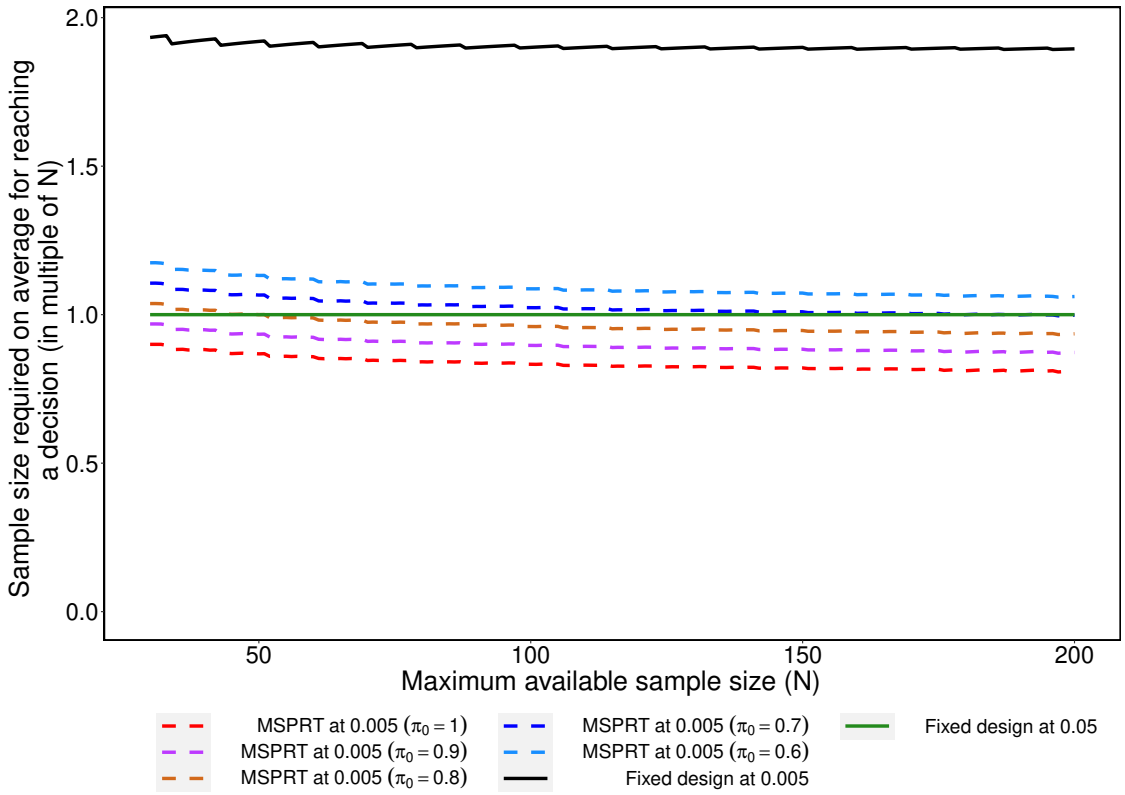


Figure 2.8: **One-sample t test that a population mean is 0.** Curves in this plot represent the average multiple of the sample size in a fixed design test of size $\alpha = 0.05$ required to perform the MSPRT of size $\alpha = 0.005$ of approximately the same power. Average sample sizes are dependent on the proportion of tested null hypotheses that are true. The MSPRT maintains a Type I error probability of 0.005, and its power at θ^* always exceeds 0.77 for the indicated proportion of N^* (the sample size of the corresponding fixed design test).

be achieved at tests implemented with more stringent Type II error probabilities or in two-sided test, and it is important to study the operating characteristics of any particular design before its implementation. In the case of one-sided z, t and proportion tests, however, this finding holds because N^* is roughly two times that of N, but at $\alpha = 0.005$ the MSPRT saves more than 50% of the maximum available sample size when the null hypothesis is true and the test is powered to achieve a Type II error of 0.2. For such tests, "raising the significance bar" from 0.05 to 0.005 could be accomplished without significantly increasing sample sizes if MSPRTs were used in place of fixed design tests.

2.6 An Application to the retrospective gambler's fallacy study

In this section we illustrate the use of the MSPRT to the replication data of the retrospective gambler's fallacy study, one of many studies available from the first "Many Labs" project [31]. The data is openly accessible from the Open Science Framework (`https://osf.io/wx7ck/`). In the original study, [30] investigated the influence of observing a rare, independent, chance event on individuals' perception of preceding events. For the experiment, the participants imagined that they saw someone rolling dice in a casino and then witnessed one of the following three outcomes (or conditions). In one condition, the participants imagined that they observed three dice being rolled and all came up "6" (the "three6 condition"). In the second condition, two dice came up "6" and one was "3" (the "two6 condition"). Finally, in a third condition, two dice were rolled and both came up "6." All participants then estimated the number of times the dice were rolled before they observed the outcomes. The results from the study support a theoretical prediction that participants perceive unlikely outcomes to have arisen from longer sequences than more common outcomes.

In the Many Labs project, the same study was replicated with only the first and second conditions. In that study, there were a total of 5942 participants out of whom 2680 participants witnessed the three6 condition and 3262 witnessed the two6 condition. To keep the illustration simple we consider a sequential MSPRT with equal number of participants from each group. Thus, we ran-

domly selected 2680 participants from the two6 group and the full set of 2680 responses from the three6 condition as our data for the sequential MSPRT. Furthermore, following [30] and [31], we took the square-root of the subjects' estimated number of dice rolls prior to their imagined outcome as the response variable. (The square-root transformation of Poisson counts is approximately variance stabilizing).

To test the hypothesis of a mean difference, we assumed that the underlying population of the transformed responses corresponding to the three6 and two6 conditions were independently and normally distributed with means θ_3 and θ_2 with an unknown common variance σ^2. We then applied a right-sided two-sample t-test of the form

$$H_0 : \theta_3 - \theta_2 = 0 \qquad \text{vs.} \qquad \theta_3 - \theta_2 > 0 \tag{2.3}$$

with the Type I and the Type II error probabilities constrained by α and β, respectively.

Approximately 90 subjects, on average, were assigned to each group in the Many Labs project, so we arbitrarily set $N = 90$ in this study. We then varied $\alpha = \{0.005, 0.05\}$ and $\beta = \{0.05, 0.2\}$ and examined the operating characteristics of the MSPRT by repeatedly sampling 90 subjects from the two treatment groups.

For each (α, β) combination, we designed the MSPRT using the **`design.MSPRT()`** function in the **`R`** package **`MSPRT`**. For example, when $\alpha = 0.005$ and $\beta = 0.05$, the **`R`** command to obtain the MSPRT design parameter is as follows:

```
# design the MSPRT
>out = design.MSPRT(test.type = 'twoT',
                    Type1.target = 0.005,
                    Type2.target = 0.05,
                    N1.max = 90, N2.max = 90)

# display termination threshold
```

```
>out$termination.threshold

# display simulation estimate of Type I error
# probability obtained by the MSPRT
>out$Type1.attained

# display simulation estimate of Type II error
# probability obtained by the MSPRT at
# the fixed-design alternative
>out$Type2.attained

# display the ASN of the MSPRT under the null
>out$EN$H0

# display the ASN of the MSPRT at the fixed-design
# alterative
>out$EN$H1
```

We next applied the MSPRT to actual data sets by randomly selecting 90 (sequential) observations from the available 2680 observations in each treatment group. For $\alpha = 0.005$ and $\beta = 0.05$, applying the MSPRT to the first sequence of sampled outcomes led to rejection of the null hypothesis at the 0.005 level of significance after 60 observations were observed from each group. The MSPRT was implemented using the **`implement.MSPRT()`** function as follows:

```
>implement.MSPRT(obs1 = three6.resp.MSPRT,
                 obs2 = two6.resp.MSPRT,
                 design.MSPRT.object = out)
```

Here, `three6.resp.MSPRT` and `two6.resp.MSPRT` are numeric vectors containing the sequential observations of the three6 and two6 responses, and `out` is the object storing the MSPRT

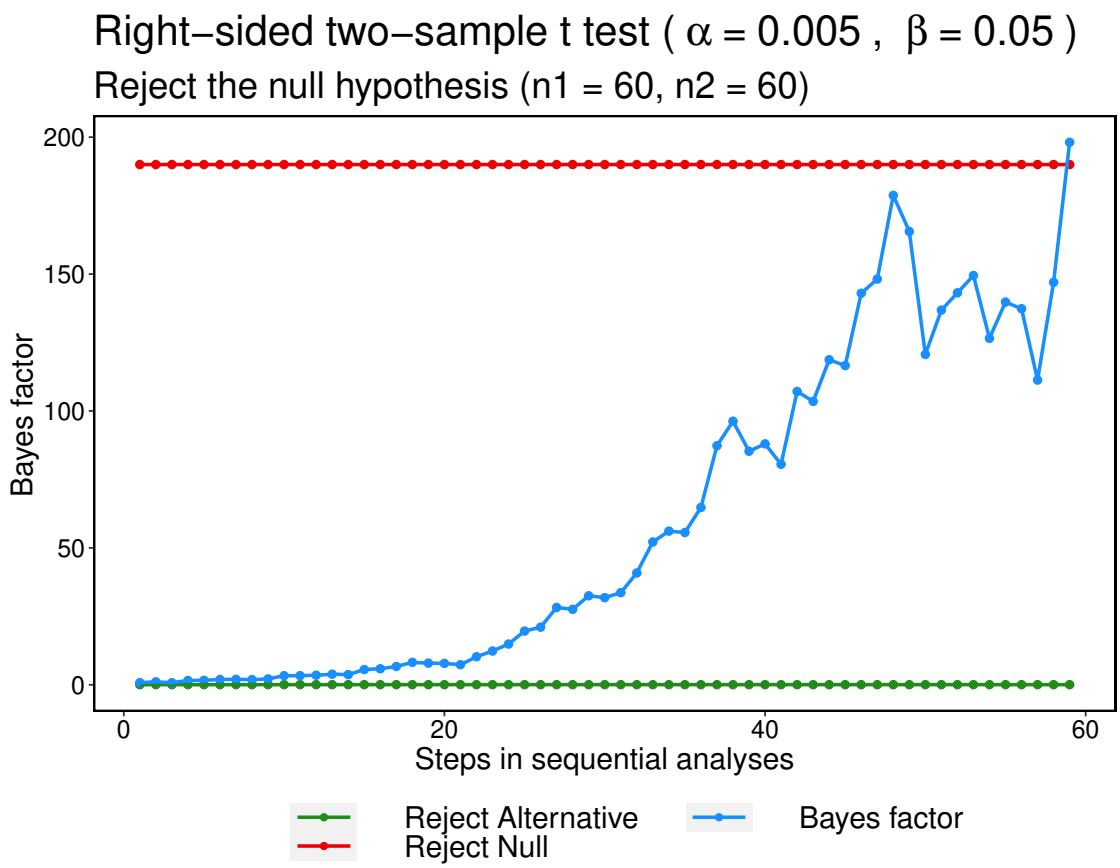


Figure 2.9: Application of the MSPRT at $\alpha = 0.005$ and $\beta = 0.05$ to a specific simulated sequence of observations from each group available from the retrospective gambler's fallacy study.

output. The **`implement.MSPRT()`** command can be executed sequentially after responses are observed and the response variables have been updated. The Bayes factor obtained at the MSPRT alternative for this sequence of observations is displayed in Fig. 9.

We can also find the operating characteristics of the MSPRT at specified effect sizes using the **`OCandASN.MSPRT()`** function. For example, the **`R`** commands to calculate the operating characteristics of this MSPRT, at the estimated standardized effect of 0.69 cited in [30] are as follows:

```
# obtain the OC at theta = 0.69
>oc.out = OCandASN.MSPRT(theta = 0.69,
                         design.MSPRT.object = out)
```

Table 2.2: Operating characteristics and ASN's of the designed MSPRT's for the retrospective gambler's fallacy study

α	β	$\triangle = 0$		$\triangle = \theta_a$		$\triangle = 0.69$	
		Type I	$\mathbb{E}(n)$	Type II	$\mathbb{E}(n)$	Type II	$\mathbb{E}(n)$
0.005	0.05	0.005	63.46	0.0513	63.31	0.023	58.4
	0.2	0.005	39.83	0.2129	69.78	0.029	56.06
0.05	0.05	0.05	84.3	0.0504	63.74	0.001	46.77
	0.2	0.05	63.28	0.204	71.52	0.002	44.13

Note. Type I and Type II indicates the corresponding error probabilities. $\triangle = (\theta_3 - \theta_2)/\sigma$ denotes the standardized effect size. $\mathbb{E}(n)$ denotes the ASN for each group at the corresponding effect size. Effect sizes at the null value $\triangle = 0$, fixed-design alternative θ_a (i.e., the fixed N design providing the specified (α, β)), and the standardized effect size 0.69 estimated from the original study are provided.

```
# display simulation estimate of Type II error probability
# at theta = 0.69
>oc.out$acceptH0.prob


# display ASN from Group-1 at theta = 0.69
>oc.out$EN1
```

The output from these commands, `oc.out`, is a data frame in which rows correspond to effect sizes, and columns refer to the probability of rejecting H_1 and the ASN's from Group 1 and Group 2 (in the case of equal sample sizes in both groups, these values are the same). For reference, these values are displayed in Table 2.2.

For each pair of (α, β), we also evaluated the operating characteristics of the MSPRT when applied to 10^6 sampled sequences. Specifically, we calculated (a) the number of samples required on average by the MSPRT to reach a decision, and (b) the proportion of times the MSPRT rejected the null hypothesis. These results are presented in Figure 2.10.

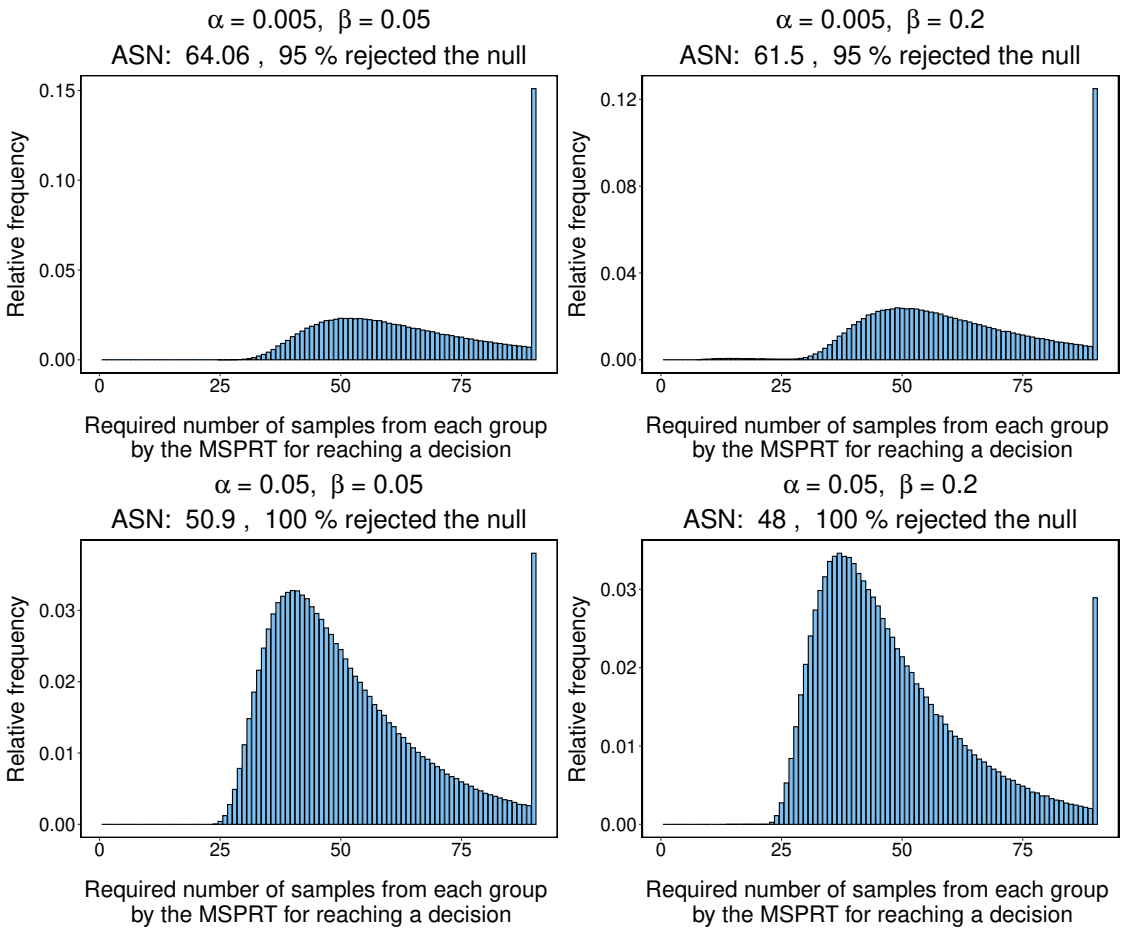


Figure 2.10: Histogram of the required number of samples from each group (condition) by the MSPRT for reaching a decision in 10^6 random permutations of the gambler's fallacy study responses.

2.7 Discussion

The costs of conducting experiments to test hypothesized effects is often related directly to the number of tested items or participants. When the study data can be collected sequentially, the use of sequential testing procedures can dramatically reduce these costs. When tests are designed to identify hypothesized effects that do not exist (i.e., the null hypothesis is true), the use of the MSPRT can reduce the sample sizes to reach a decision. In z and t tests with type II error probabilities targeted at 20%, the reduction in sample sizes can be as much as 20% to 30% in 5% tests, and as much as 50% in 0.5% tests.

Much of this chapter has focused on one-sample z, t, and proportion tests. Mathematically, two-sample z and t tests are similar to one-sample tests, and so our findings extend to two-sample z and t tests. Table S1 of [11] provides a list of the UMPBT alternatives and the likelihood ratios (or Bayes factors) for two-sample z and t tests. Section A4 of the supplemental materials provides a user guide for two-sample tests.

A potential drawback in the implementation of MSPRTs is the firm requirement to specify the outcome variable and test statistic prior to the start of the experiment. Of course, in principle the same requirement applies to fixed design experiments, but failure to ensure that these quantities are clearly identified a priori could lead to additional opportunities for p hacking and other unethical practices in sequential designs. For instance, researchers might apply MSPRTs to several outcome variables simultaneously, which would negatively affect the control of Type I errors. In addition, the conduct of MSPRTs requires that investigators perform statistical analyses after the acquisition of each participant's data, which in some settings may not be feasible. However, for studies in which a high threshold for significance is desired, this technique may offer researchers a method of testing hypotheses while maintaining required sample sizes at a manageable level.

2.8 Supplementary Materials

Supplementary materials, which are available online, contain a detailed discussion of the UMPBT alternative and a comprehensive user guide for the **`MSPRT`** package. Section A2 highlights the gen-

eral MSPRT for testing a simple null against a compositive alternative hypothesis. In Section A3, the UMPBT alternatives are discussed in detail for one-sample z, t and proportion tests. Finally, Section A4 presents an instructional user guide for the R package. Designing and implementing a MSPRT, calculating the UMPBT alternative for different tests and obtaining N^* (as in Section 2.5.4) are reviewed in respective subsections there. Additional simulation results for one-sample z and proportion tests, with similar conclusions as to the one-sample t test, are presented in Section A4.2.

3. EFFICIENT ALTERNATIVES FOR BAYESIAN HYPOTHESIS TESTS IN PSYCHOLOGY

3.1 Introduction

Innovative statistical methods to evaluate the plausibility of scientific theories have attracted increased attention over the last decade. This attention has resulted in renewed interest in Bayesian methods for assessing evidence [e.g., 1], and several novel approaches to sequential testing procedures have recently been proposed [2, 3, 54]. As [2] point out, each of these sequential testing methods can be motivated from a Bayesian perspective towards testing.

[1] provide a useful summary of Bayesian methodology and, through a series of examples, argue that "advances in science often come from identifying *invariances*," or "statements of equality, sameness, or lack of association." As examples, they cite interest in determining "whether cognitive skills vary with gender"; whether subliminal priming occurs; whether detectability of a "briefly flashed stimulus" is invariant to the ratio of the intensity of the stimulus to background, as predicted by the Weber-Fechner law [55]; and whether the exponent in the power function of intensity used to predict sensation is constant for a given intensity variable [56, 57]. To identify invariances, hypothesis testing procedures must permit accumulation of evidence in support of both null and alternative hypotheses (see also [19, 20, 21]). In this regard, Bayesian testing procedures differ from classical testing procedures, in which one can only fail to reject the null hypothesis [e.g., 22], by allowing researchers to quantify evidence in favor of true null hypotheses, which can reflect the presence of an invariance or lack of an effect.

In the Bayesian paradigm, the posterior odds in favor of an alternative hypothesis H_1, based on data $\mathbf{x}$, can be expressed as the product of the Bayes factor and the prior odds in favor of H_1; that is

$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds}, \tag{3.1}$$

or

$$\frac{\mathbf{P}(H_1 \mid \mathbf{x})}{\mathbf{P}(H_0 \mid \mathbf{x})} = \frac{m_1(\mathbf{x})}{m_0(\mathbf{x})} \times \frac{\mathbf{P}(H_1)}{\mathbf{P}(H_0)}. \tag{3.2}$$

It is important to note that this equation can be interpreted from both a frequentist and subjective view of probability. From the frequentist perspective, all probabilities can interpreted as the limiting proportion of the occurrence of an event. That is, if the null hypothesis H_0 is repeatedly sampled with probability $\mathbf{P}(H_0)$ (or H_1 with probability $\mathbf{P}(H_1) = 1 - \mathbf{P}(H_0)$), and data $\mathbf{x}$ is generated according to $m_0(\mathbf{x})$ (or $m_1(\mathbf{x})$), then the posterior probability that data was generated under H_1, for a given $\mathbf{x}$, converges in probability to

$$\mathbf{P}(H_1 \mid \mathbf{x}) = \frac{\mathrm{BF}_{10}(\mathbf{x})\,\mathbf{P}(H_1)}{\mathbf{P}(H_0) + \mathrm{BF}_{10}(\mathbf{x})\,\mathbf{P}(H_1)}, \tag{3.3}$$

where $\mathrm{BF}_{10}(\mathbf{x}) = m_1(\mathbf{x})/m_0(\mathbf{x})$ is the Bayes factor in favor of H_1.

When Bayesian methods are applied to Null Hypothesis Significance Tests (NHSTs), controversy arises in the "subjective" specification of two quantities in these equations. First, the prior odds in favor of H_1 must be specified. This specification is equivalent to specifying either the prior probability of the alternative hypothesis, $P(H_1)$, or the prior probability of the null hypothesis, $P(H_0)$, since $P(H_0) + P(H_1) = 1$. A simple approach to setting the prior odds is to assume $P(H_0) = P(H_1) = 0.5$, leading to prior odds of 1.0. However, recent evidence gleaned from analyses of replicated experiments suggests that the prior odds in favor of the alternative hypotheses studied in psychology and other social sciences might be closer to $1/9$ [9, 23]. Although it is necessary to set a value of the prior odds in order to calculate the posterior odds, evaluation of the prior odds is not considered further here. Instead, we encourage researchers to perform their own sensitivity analyses to evaluate how various assumptions regarding the prior odds affect the posterior odds for a given Bayes factor.

The second point of controversy arises in the definition of the marginal density of the data

under the alternative hypothesis, given by

$$m_1(\mathbf{x}) = \int_{\Theta} f(\mathbf{x} \,|\, \theta)\pi_1(\theta) d\,\theta. \tag{3.4}$$

Here $\pi_1(\theta)$ represents the prior density for the parameter of interest θ under the alternative hypothesis, i.e., the alternative prior density. (A more detailed description of the Bayesian hypothesis testing framework may be found in, for example, [24] or [19].) In NHSTs, the quantity $m_0(\mathbf{x})$ simply represents the sampling density of the data, say $f(\mathbf{x} \,|\, \theta_0)$, evaluated at the parameter value that defines the null hypothesis, θ_0.

The two-sample t test provides a useful context to discuss the specification of the alternative prior density, $\pi_1(\theta)$. In this setting, the parameter of interest is usually defined to be either the difference in population means, $\mu_2 - \mu_1$, or the standardized difference in population means $\delta = (\mu_2 - \mu_1)/\sigma$. The former is called the effect size, while the latter is called the standardized effect size. Sample estimates of δ are called Cohen's d [58]. (Explicit modeling assumptions on $\mathbf{x}$, μ_1, μ_2, σ^2 and δ are provided in the next section.) For purposes of the present discussion, we assume that the null hypothesis requires that $\delta = 0$, and that under the alternative hypothesis $\delta \neq 0$. In the absence of prior information regarding the value of δ, a common default choice for the alternative prior density on δ is a Cauchy distribution. The Cauchy distribution is a unimodal density that takes its maximum value at 0 and has heavy tails that assign significant mass to large values of the standardized effect size (i.e., $\delta > 1$). When the observational variance σ^2 is unknown, a default prior density on σ is the Jeffreys (or non-informative) prior, given by $p(\sigma^2) \propto 1/\sigma^2$. Although improper, this prior has attractive theoretical properties, provided that it is used as the prior model for the variance under both the null and alternative hypotheses.

If the Jeffreys prior is assumed for the observational variance σ^2 and a Cauchy prior is assumed for δ, then the resulting prior on $\mu_2 - \mu_1$ is called the JZS prior, in deference to Jeffreys, Zellner and Siow [24, 59, 60]. It is the default prior recommended in [1] for one- and two-sample t tests and by [2] in their definition of a sequential Bayes factor (SBF). The JZS prior is an example of a

local alternative prior, or a prior density that is positive at parameter values that are consistent with the null hypothesis.

Intrinsic priors [e.g., 61] are another class of commonly used default priors in Bayesian testing of a normal mean. The operating characteristics of Bayes factors obtained using intrinsic priors and other default local priors are similar to those obtained using the JZS prior. For brevity, we therefore do not consider them separately here.

The focus of this chapter is the description of a new approach to specifying alternative hypotheses in Bayesian tests of a normal mean or difference between means. The approach is based on the use of non-local alternative prior densities (NAPs; [25]). A NAP is a density that exactly equals 0 at parameter values that are consistent with the null hypothesis. For the two-sample t test, this means that the prior density on δ is identically 0 when $\delta = 0$. As we demonstrate below, tests specified with NAPs offer several advantages over tests defined with alternative hypotheses based on local priors. These include the following:

Stronger evidence for true null hypotheses. Because local alternative priors, like the JZS prior, assign high prior probability to parameter values that are consistent with the null hypothesis, data that support the null hypothesis also provide support to the alternative hypothesis. This makes it difficult to obtain evidence that favors a true null hypothesis. We note that accumulating evidence for true null hypotheses is often cited as a primary rationale for the use of Bayes factors in hypothesis testing (e.g., [1]). Ironically, local alternative priors are particularly ill-suited for this task (see, for example, Fig. 3.2). These properties of Bayes factors are discussed below for tests in which sample sizes are fixed at the beginning of a study and for sequential tests.

Comparable or stronger evidence for true alternative hypotheses. Because NAPs assign negligible probability to parameter values that are consistent with the null hypothesis, they are able to assign more prior probability to parameter values that support the alternative hypothesis. When data support the alternative hypothesis, the marginal density for the data under the alternative thus tends to be higher than it is with a local alternative prior, which

increases the Bayes factor in favor of the alternative hypothesis. This is especially true when the NAP assigns high prior probability to standardized effect sizes that are common in the psychological and social sciences.

Smaller Average Sample Number (ASN) in sequential tests. Because NAPs tend to provide more evidence in favor of true null hypotheses and comparable evidence in favor of true alternative hypotheses, sequential tests based on them are likely to reach termination thresholds more quickly. This means that sequential tests based on NAPs often require fewer subjects to make a decision.

Logical consistency. In a properly specified hypothesis test, null and alternative hypotheses are mutually exclusive. That is, if the alternative hypothesis is true, then the null hypothesis cannot be. Despite this truism, local alternative priors assign prior mass to neighborhoods of parameter values that are consistent with the null hypothesis. Indeed, in many cases their densities take their maximum value at the parameter value that defines the null hypothesis. In this regard, the use of NAPs more accurately reflect the prior belief, under the alternative hypothesis, that the tested parameter does not equal a value specified under the null hypothesis.

With regard to the last item, proponents of local alternative prior densities (e.g., [1, 24, 61]) might argue that local alternative priors like the JZS prior reflect a belief that the true parameter value is "close" to the null value. That is, the fact that a hypothesis test is being conducted at all suggests that the tested effect size must not be too large. Thus, it is reasonable for the prior density for the alternative model to take its maximum at the null value. This was the argument originally posited by [24] in proposing a Cauchy prior for the unknown mean of a normal population.

We have two objections to this perspective. First, as we demonstrate below it is generally not feasible to detect small standardized effect sizes without very large sample sizes. As a consequence, investigators who wish to detect small effects are compelled to design studies with large sample sizes. If such studies are planned, investigators can also specify non-local alternative prior

densities that are appropriately scaled to detect the targeted effects. The resulting NAPs are sharply spiked at small effect sizes, which increases the Bayes factor in favor of the alternative hypothesis when it is supported by data. The use of appropriately scaled NAPs can thus lead to savings in sample size and experimental cost.

Second, point null hypotheses are often used to approximate a belief that a standardized effect size is small. When this is the case, the use of local alternative prior densities is even more problematic because they then concentrate prior probability on a range of parameter values that are consistent with the null hypothesis.

The simplest NAP densities are simple alternative hypotheses. For example, in a one-sided test of whether a standardized effect size is zero, a simple alternative hypothesis might be $H_1 : \delta = 0.3\sigma$. We demonstrate below that simple alternative hypotheses make it easy to collect evidence in favor of both true null and true alternative hypotheses, but that they can lack power in detecting true alternative hypotheses defined by other parameter values (e.g., $\delta = 0.15\sigma$). For this reason, we describe a class of continuous NAPs called normal moment densities that are strictly positive at all non-null parameter values.

The rest of the chapter is organized as follows. In the next section, we describe the class of normal moment densities that can be used to define alternative hypotheses for standardized effect sizes in one- and two-sample t and z tests. Unlike simple alternative hypotheses, tests constructed with these alternative prior densities provide support for a range of true alternative hypotheses. They also permit rapid accumulation of evidence in favor of true null hypotheses. Conveniently, these densities lead to explicit expressions for Bayes factors in one- and two-sample z and t tests. In the next two sections we compare the empirical properties of tests defined with NAPs to tests defined with default, objective local alternative priors in fixed and sequential design settings. The third section examines tests in which the sample size is fixed prior to analyses of data. We refer to such tests as fixed design tests. The fourth section examines the performance of NAP-based tests in sequential designs. We conclude with a discussion of results.

3.2 Non-local alternative prior densities

NAPs are probability density functions that take the value 0 at parameter values that are consistent with the null hypothesis [25]. A simple example of a NAP for the test of a normal mean can be described as follows.

Suppose $\mathbf{x} = (x_1, \ldots, x_n)$ are independent and identically distributed (iid) Gaussian random variables with mean μ and known variance σ^2, i.e., $x_i \overset{iid}{\sim} N(\mu, \sigma^2)$. Suppose we wish to test the null hypothesis $H_0 : \mu = 0$ against a two-sided alternative $H_1 : \mu \neq 0$. Let $\phi(a \,|\, m, c^2)$ denote the normal density function evaulated at a with mean m and variance c^2. A NAP density that can be used to define an alternative hypothesis for this test is the *normal moment* prior density, which can be expressed as

$$p_{NM}(\mu \,|\, \mu_0, \tau^2\sigma^2) = \frac{(\mu - \mu_0)^2}{\tau^2\sigma^2}\, \phi(\mu \,|\, \mu_0, \tau^2\sigma^2), \qquad -\infty < \mu < \infty. \tag{3.5}$$

A plot of this density for $\tau^2 = 0.3^2/2 = 0.045$ and $\sigma^2 = 1$ is provided in Fig. 3.1. For comparison, the dashed curve in this plot depicts the Cauchy density with scale parameter $r = \sqrt{2}/2 \approx 0.707$, a local default alternative prior density that is often used to define the alternative hypothesis for this test (see, for example, [2, 62, 63]). We denote the distribution associated with a generic normal moment density (3.5) by $NM(0, \tau^2\sigma^2)$. The density depicted in Fig. 3.1 has modes at $\pm\sqrt{2\sigma^2\tau^2}$. For $\tau^2 = 0.3^2/2$, the modes of the density occur at $\pm 0.3\sigma$, or at standardized effect sizes of ± 0.3. The area of the shaded region in Fig. 3.1, which represents the prior probability assigned to standard effect sizes between $(-0.8, -0.2)$ and $(0.2, 0.8)$, is 0.825. That is, this normal moment prior assigns approximately 83% of its prior probability between "small" (± 0.2) and "large" (± 0.8) effect sizes [58]. These values approximately match the median and interquartile ranges of non-null standardized effect sizes reported in meta-analyses in psychology [64, 65, 66, 67, 68, 69, 70].

The NAP density depicted in Fig. 3.1 assigns only 17.2% of its prior mass to standardized effect sizes less than 0.2 in magnitude, and is identically 0 when the standardized effect size is 0. In contrast, the Cauchy density depicted in this figure assigns about 36.3% of its prior probability

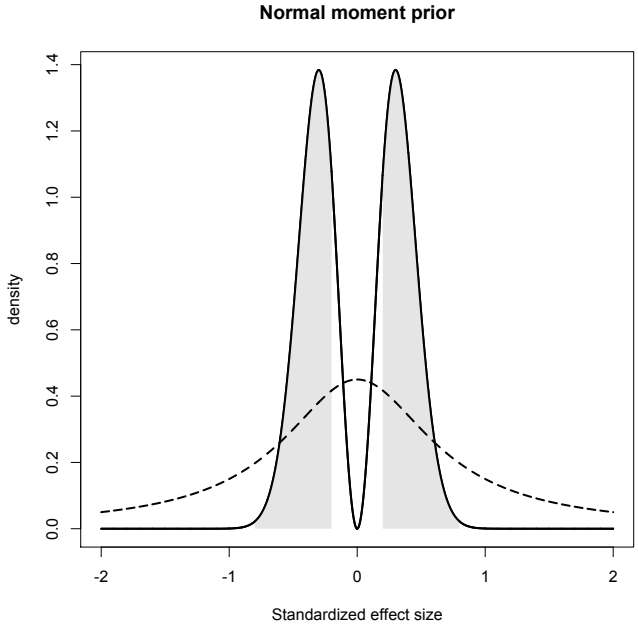


Figure 3.1: Normal moment prior. This is an example of a NAP that can be used to define the alternative hypothesis in test for a normal mean. The shaded area in the figure depicts the prior probability assigned to standardized effect sizes having magnitude between 0.2 and 0.8.

to effect sizes that fall in the range "small" to "large," and assigns 17.5% of its probability to effect sizes that are less than 0.2 in magnitude. The prior probability assigned to standardized effect sizes greater than 1 in magnitude is 0.392. The mode of this density is 0, which corresponds to the null hypothesis.

Throughout the remainder of this paper we define the prior depicted in Fig. 3.1 as the default NAP prior for defining the alternative hypothesis in testing whether the mean of a normal sample, or difference between means of normal samples, is 0. When the observational variance is unknown, we assume the Jeffreys' prior for σ^2 under both null and alternative hypotheses.

From a computational perspective, an advantage of the normal moment alternative prior density in normal models is that it results in closed form expressions for the Bayes factors in both one- and two-sided tests. In contrast, Bayes factors based on the JZS and other default priors (i.e., intrinsic priors) do not have closed-formed expressions and so must be computed using numerical integration routines. For one-sided tests, we use the positive half of the density to define the

alternative hypothesis. That is, for testing $H_1 : \mu > 0$ we define

$$p^{+}_{NM}(\mu \,|\, 0, \tau^2\sigma^2) = 2p_{NM}(\mu \,|\, 0, \tau^2\sigma^2) = \frac{2\mu^2}{\tau^2\sigma^2}\,\phi(\mu \,|\, 0, \tau^2\sigma^2), \qquad \mu > 0. \tag{3.6}$$

A similar definition is used to test $H_1 : \mu < 0$. We denote the distribution corresponding to this density as $NM^{+}(0, \tau^2\sigma^2)$ (or $NM^{-}(0, \tau^2\sigma^2)$ for $\mu < 0$).

We now define the specific assumptions used to perform one and two sample tests for normal means against a two-sided alternative hypothesis, both when the variance is known and unknown. We also provide explicit expressions for the resulting Bayes factors. Expressions for one-sided tests are provided in the supplemental material.

For tests conducted in the psychological sciences with small to moderate sample sizes, and for which no specific prior information regarding the magnitude of standardized effect size is available, we recommend a default value of $\tau^2 = 0.045$.

1. One-sample, known variance test. Suppose $\mathbf{x} = (x_1, \ldots, x_n)$ denote iid $N(\mu, \sigma^2)$ random variables and that σ^2 is known. The Bayes factor of the test $H_1 : \mu \sim NM(0, \tau^2\sigma^2)$ versus $H_0 : \mu = 0$ is given by

$$\text{BF}_{10}(\mathbf{x}) = (n\tau^2 + 1)^{-3/2}(1 + 2w)e^{w}, \tag{3.7}$$

where

$$r = \frac{n\tau^2}{n\tau^2 + 1}, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad Z = \sqrt{n}\bar{x}/\sigma, \quad \text{and} \quad w = rZ^2/2. \tag{3.8}$$

Here, Z is the test statistic used in the frequentist z test.

2. One-sample, unknown variance test. Suppose the conditions in **[1]** hold, except that σ^2 is unknown. Suppose further that the Jeffreys' prior density for σ^2, proportional to $1/\sigma^2$, is assumed under both hypotheses. Then the Bayes factor in favor of the alternative hypothesis

can be expressed as

$$\mathrm{BF}_{10}(\mathbf{x}) = (n\tau^2+1)^{-3/2}\left(\frac{G}{H}\right)^{n/2}\left(1+\frac{qT^2}{H}\right), \tag{3.9}$$

where r and $\bar{x}$ are defined as in (3.8), and

$$q = \frac{rn}{n-1}, \quad S = \sum_{i=1}^{n}(x_i-\bar{x})^2, \quad s^2 = \frac{S}{(n-1)}, \quad T = \frac{\sqrt{n}\bar{x}}{s}, \tag{3.10}$$

$$G = 1+\frac{T^2}{n-1}, \quad \text{and} \quad H = 1+\frac{(1-r)T^2}{(n-1)}. \tag{3.11}$$

Here, T is the test statistic used in the frequentist t test.

3. Two-sample, known variance test. Suppose $\mathbf{x}_1 = (x_{1,1},\ldots,x_{1,n_1})$ denote iid $N(\mu_1,\sigma^2)$ random variables, $\mathbf{x}_2 = (x_{2,1},\ldots,x_{2,n_2})$ iid $N(\mu_2,\sigma^2)$ random variables, $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent of each other, and that σ^2 is known. We assume that the prior density for μ_1 is uniformly distributed on an interval $(-a,a)$ for a large value of a under both hypotheses. Then the Bayes factor for the test $H_1 : \mu_2-\mu_1 \sim NM(0,\tau^2\sigma^2)$ versus $H_0 : \mu_2-\mu_1 = 0$ can be expressed as

$$\mathrm{BF}_{10}(\mathbf{x}_1,\mathbf{x}_2 \,|\, \sigma^2) = (m\tau^2+1)^{-3/2}\,(1+2w)\,e^w, \tag{3.12}$$

where for $i = 1, 2$,

$$\bar{x}_i = \sum_{j=1}^{n_i} x_{j,i}/n_i, \qquad n = n_1+n_2, \qquad m = \frac{n_1 n_2}{n_1+n_2}, \tag{3.13}$$

$$r = \frac{m\tau^2}{m\tau^2+1}, \qquad Z = \sqrt{m}(\bar{x}_2-\bar{x}_1)/\sigma \qquad \text{and} \qquad w = \frac{rZ^2}{2}. \tag{3.14}$$

The value Z is the test statistic in the classical z test. We note that the labeling of samples is arbitrary and the marginal prior density on μ_2 is also approximately uniform on $(-a,a)$.

The Bayes factor is obtained by taking the limit $a \to \infty$.

4. Two-sample, unknown variance test. Suppose the conditions in **[3]** hold, except now that σ^2 is unknown. Suppose further that the Jeffreys' prior density for σ^2 is assumed under both hypotheses. Then the Bayes factor in favor of the alternative hypothesis can be expressed as

$$\mathrm{BF}_{10}(\mathbf{x}_1, \mathbf{x}_2) = (m\tau^2 + 1)^{-3/2} \left(\frac{G}{H}\right)^{(n-1)/2} \left(1 + \frac{qT^2}{H}\right), \tag{3.15}$$

where $\bar{x}_1$, $\bar{x}_2$, r, n, m are defined in (3.13)–(3.14), and

$$q = \frac{r(n-1)}{(n-2)}, \qquad S_i = \sum_{j=1}^{n_i} (x_{i,j} - \bar{x}_i)^2, \qquad S = S_1 + S_2, \tag{3.16}$$

$$T = \frac{\sqrt{m}(\bar{x}_1 - \bar{x}_2)}{\sqrt{S/(n-2)}}, \qquad G = 1 + \frac{T^2}{(n-2)}, \qquad H = 1 + \frac{(1-r)T^2}{(n-2)}. \tag{3.17}$$

Here T is the test statistic in the classical t test. As in **[3]** the labeling of samples is arbitrary, and the Bayes factor is obtained by taking the limit $a \to \infty$.

3.3 Fixed design tests

Classical tests of a normal mean parameter are most commonly based on z or t tests. These tests are designed to control Type I (α) and Type II (β) error probabilities at pre-specified levels. A key disadvantage of these tests is that they do not quantify evidence in favor of true null hypotheses. Instead, they may simply "fail to reject" the null hypothesis. Psychology and other social science researchers often have a need to quantify evidence in favor of true null hypotheses [for example, 1]. Bayes factors provide such a measure.

3.3.1 "Weight of evidence" as a measure of evidence

To summarize the performance of various Bayesian tests, we adopt the measurement scale for evidence based on the natural logarithm of the Bayes factors, $\ln(\mathrm{BF}_{10})$. This quantity, called the "weight of evidence", has the advantage of being on the same scale as the classical likelihood ratio

statistic [19, 24].[1] Because $-\ln(x) = \ln(1/x)$, the weight of evidence in favor of the alternative hypothesis is equal to the negative of the weight of evidence in favor of the null hypothesis (and vice versa). Descriptors for the weight of evidence were proposed by [19] and [24]. Under the former, weight of evidence between 0 and 1 in magnitude is considered "not worth more than a bare mention"; weight of evidence between 1 and 3 is considered "positive"; weight of evidence between 3 and 5 is "strong", and above 5 is labeled as "very strong". At the border between positive and strong (3), the corresponding Bayes factor is about 20, and at the border between strong and very strong, the Bayes factor is about 150. Strong and very strong weights of evidence in favor of the null hypothesis are -3 and -5, or Bayes factors of approximately $1/20$ and $1/150$.

Bayes factors must be multiplied by the prior odds that the null hypothesis is true to determine the posterior odds. If the prior odds are 1 (that is, $\mathbf{P}(H_0) = \mathbf{P}(H_1) = 0.5$), then weight of evidence equal to 3 implies a Bayes factor and posterior odds of about 20, and posterior probability of the alternative hypothesis equal to 0.95. Similarly, weight of evidence of -5 implies a Bayes factor and posterior odds of about 1/150, and posterior probability of the null hypothesis equal to $1 - 0.0066 = 0.9934$. This probability is very close to 1.0, but it is predicated on the assumption that the prior odds are 1.0.

Recent evidence from replication of experiments in psychology and social sciences suggest that the prior probability of a null hypothesis examined in these fields is likely between 0.80–0.95 [8, 9, 23]. If $\mathbf{P}(H_0) = 0.9$, then weight of evidence equal to 3 implies that the posterior probability of the alternative hypothesis is only 0.69, while weight of evidence equal to 5 implies that the posterior probability of the alternative hypothesis is 0.94.

3.3.2 Performance comparison

With the background from the above section in place, we now consider the average weight of evidence that is obtained from a two-sided, one-sample t test that a normal mean is equal to 0 when the true mean is 0. We assume the conditions of test **[2]** above hold. Operating characteristics

[1] [19] propose $2\ln(\mathrm{BF}_{10}(\mathbf{x}))$ as a default measure, but by omitting the factor of 2 their descriptors are more compatible with the measure proposed by [24].

for two-sided z tests and two-sided, two-sample t tests are very similar to those obtained for the two-sided one-sample t test. Corresponding results for these tests are provided in the supplemental materials. The R package BayesFactor [71] was used to compute the Average Sample Number (ASN) for the JZS alternatives.

3.3.2.1 True null hypothesis

Fig. 3.2 displays the average weight of evidence obtained under several alternative hypotheses when the null hypothesis of no effect is true. These curves were based on simulating one-million standard normal random deviates at each sample size. The alternative hypotheses considered in this plot include the following:

1. The default NAP (normal moment) prior with $\tau^2 = 0.3^2/2 = 0.045$ (modes at ± 0.3),
2. The default JZS prior based on a Cauchy with scale $r = \sqrt{2}/2$,
3. A normal moment prior with $\tau^2 = 0.5^2/2 = 0.125$ (modes at ± 0.5),
4. The JZS prior based on a Cauchy with scale $r = 1$,
5. A composite alternative hypothesis that assigns $1/2$ mass to standardized effect sizes of ± 0.3. The mass of a simple hypothesis is split between ± 0.3 to reflect the two-sided specification of the test. Approximate Bayes factors for this test were computed as the ratio of non-central and central t distributions evaluated at simulated t statistics.

Fig. 2 illustrates a critical deficiency of the JZS priors (and related local priors): The use of such priors to define the alternative hypothesis makes it difficult to obtain "very strong" weight of evidence in favor of a true null hypothesis. *For two-sided t tests, the default JZS prior requires about 80,000 subjects, on average, to obtain very strong weight of evidence in favor of a true null hypothesis, and the JZS prior with $r = 1$ requires about 40,000 subjects.* In contrast, the NAP prior with modes at ± 0.3 and ± 0.5 require about 300 and 1200 subjects, on average, for the same purpose.

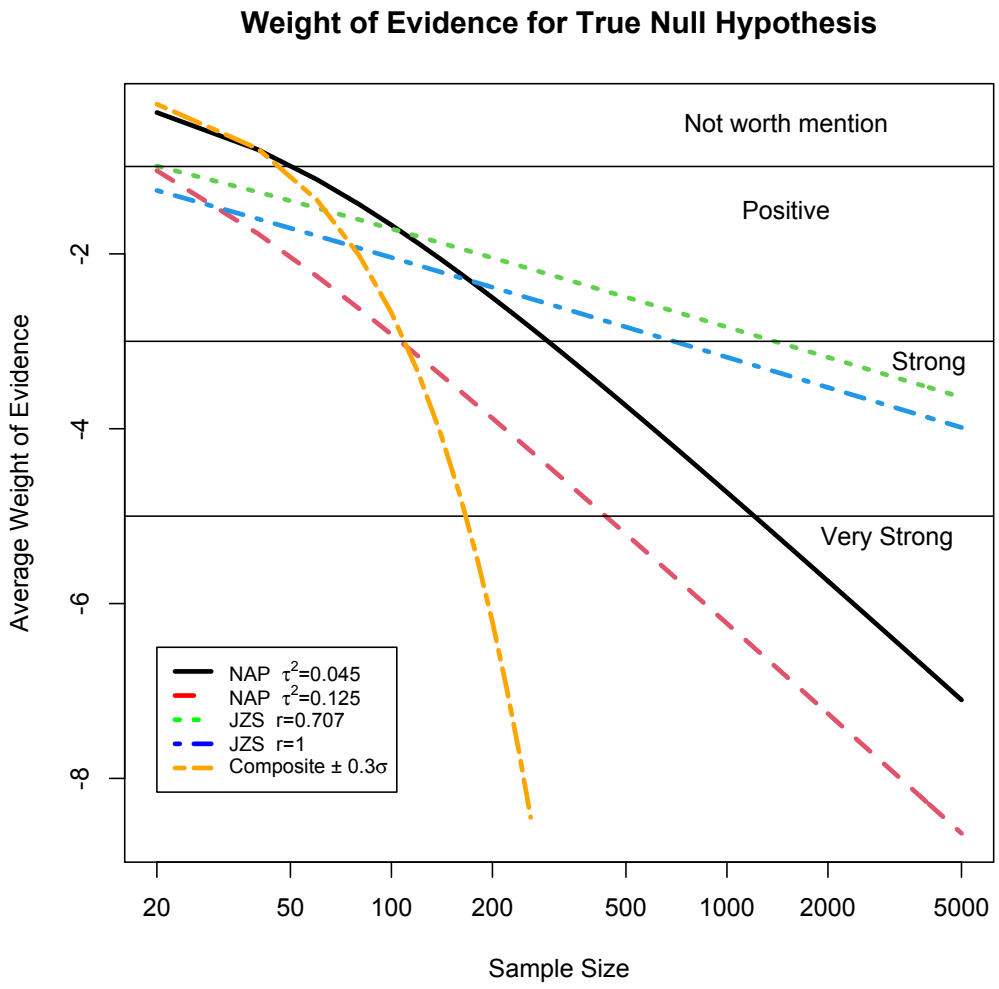


Figure 3.2: Average weight of evidence against alternative hypotheses when the null hypothesis is true. Curves depicted in the plot correspond to normal moment priors with modes at ± 0.3 and ± 0.5; the JZS prior with scale $\sqrt{2}/2$ and 1; and a composite alternative hypothesis that places one-half mass at $\pm 0.3\sigma$. *The horizontal axis is displayed on the logarithmic scale because of the large differences in samples sizes required by the different methods to obtain, on average, strong or very strong weight of evidence against each alternative hypothesis* The JZS priors do not, on average, yield very strong weight of evidence until sample sizes exceed 40,000.

Obtaining even strong weight of evidence in favor of a true null hypothesis is difficult when standard JZS priors are used to define the alternative hypothesis. On average, 1,400 subjects are required to obtain strong weight of evidence when the default JZS prior is used, and on average 750 subjects are needed when the JZS prior with scale $r = 1$ is used to define the alternative

hypothesis. In contrast, alternative hypotheses defined with NAP priors require about 300 subjects at the default scaling, and 110 subjects if the prior mode is set to 0.5.

Like the continuous NAP priors, the composite hypothesis that places one-half point mass at $\pm 0.3\sigma$ is also able to quickly obtain evidence in favor of a true null hypothesis. Indeed, because this nonlocal composite hypothesis places no prior mass in the interval $(-0.3, 0.3)$, it accumulates evidence in favor of a true null faster than the normal moment priors do.

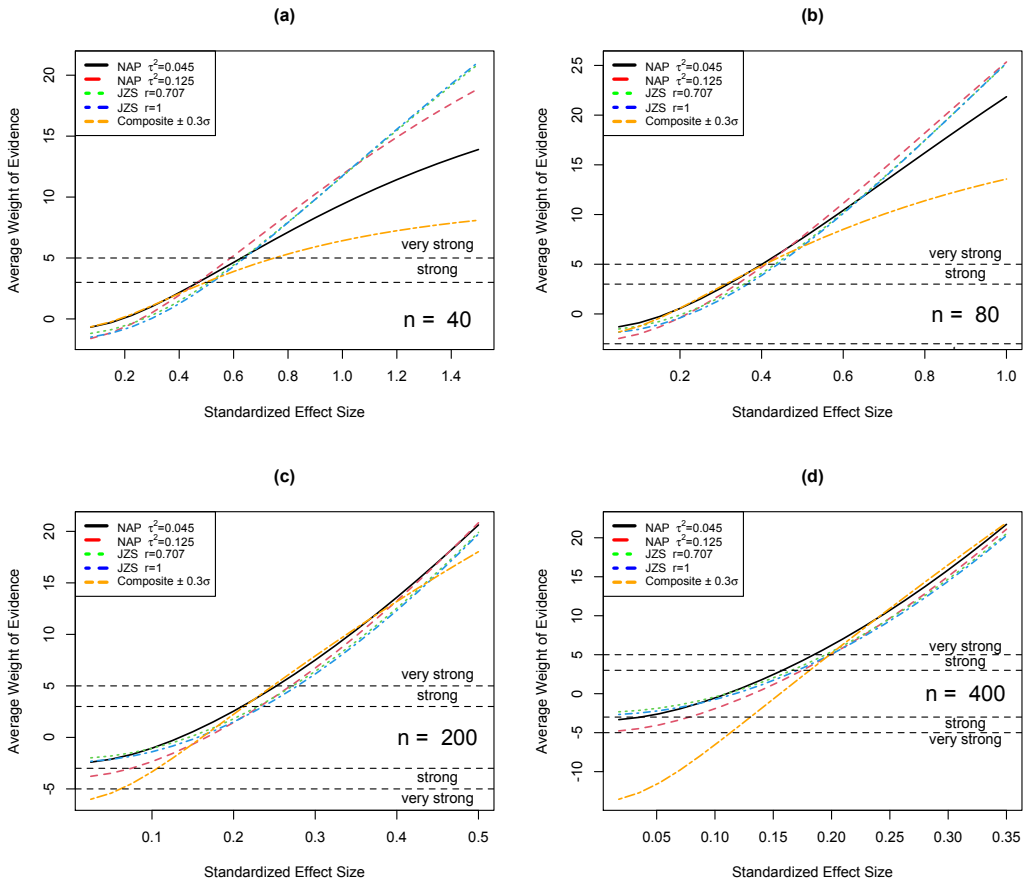


Figure 3.3: Weight of evidence for true alternative hypotheses. Curves depicted in the plots denote the average weight of evidence versus true effect size when the alternative hypothesis was defined by various NAP and JZS densities.

3.3.2.2 True alternative hypotheses

What is the cost that NAP priors when used to detect true alternative hypotheses? As it turns out, not too much for normal moment alternative prior specifications, but more with the two-point composite alternative hypothesis. Fig. 3.3 shows the average weights of evidence obtained under these prior specifications for a range of sample sizes in fixed-design tests as a function of the true standardized effect sizes. It shows that the NAP priors (based on normal moment priors) achieve strong or very strong weight of evidence in favor of the alternative hypothesis for smaller standardized effect sizes than the JZS priors do. Alternative hypotheses defined with the JZS priors provided, on average, higher weights of evidence for larger standardized effect sizes, but this additional evidence tends to occur when the evidence for the alternative hypothesis provided by the NAP priors was also very strong. For sample sizes greater than about 40 and standardized effect sizes between about 0.10 and 0.65, the default NAP prior produces, on average, higher weight of evidence against the null hypotheses than do default JZS priors.

The properties of tests defined using the two-point composite hypothesis are more ambiguous. For standardized effect sizes within $(-0.15\sigma, 0.15\sigma)$ (or $1/2$ the magnitude of the simple alternatives that comprise the composite alternative), tests based on the composite hypothesis provide, on average, support for a false null hypothesis (that is, a negative weight of evidence). For sample sizes of 200 and 400, the average weight of evidence in favor of the false null hypothesis is even very strong for smaller effect sizes. This phenomenon is not unexpected, however, because the null hypothesis in these cases is "closer" to the data-generating parameter than the composite alternative is. The composite alternative hypothesis also provides, on average, substantially less weight of evidence for large standardized effect sizes. This happens because the composite alternative hypothesis assigns no prior probability to standardized effect sizes greater than 0.3σ in magnitude.

Local priors, like the JZS prior, provide more support for very small standardized effect sizes. However, strong evidence in favor of very small standardized effect sizes can only be obtained with very large sample sizes. When the sample size is 500 and the standardized effect size is less than 0.10, all four of the Bayes factors based on alternative hypothesis defined by the JZS and

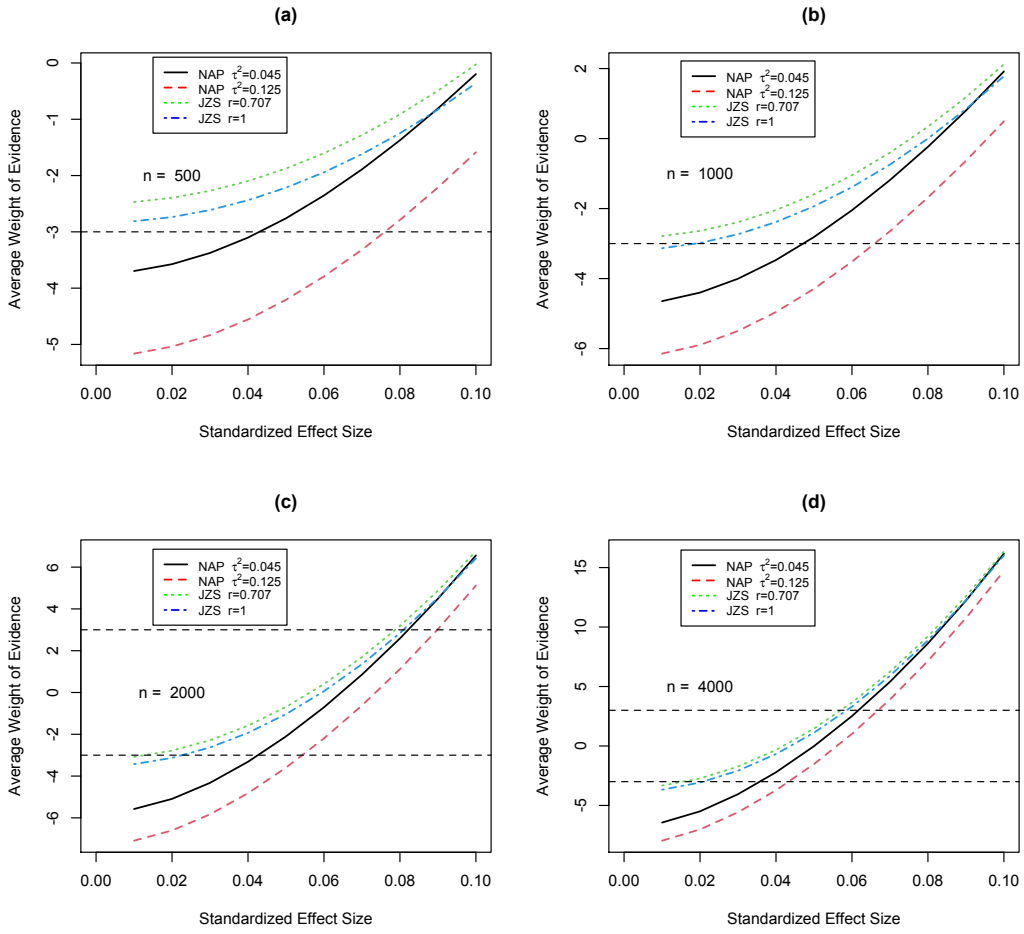


Figure 3.4: Weight of evidence for true alternative hypotheses with very small effect sizes. Curves depicted in the plots denote the average weight of evidence versus true effect size when the alternative hypothesis was defined by various NAP and JZS prior densities. Dashed lines at ± 3 provide boundaries for strong support of the alternative hypothesis (> 3) or null hypothesis (< -3).

NAP priors in Fig. 3.4(a) yield average weights of evidence that are negative, thus favoring the null hypothesis of no effect. Indeed, for standardized effect sizes less than about 0.045, use of the default NAP prior provides, on average, "strong" support for the null hypothesis, and when the standardized effect size is less than about 0.023 the NAP prior with mode at 0.5 provides "very strong" support for the null hypothesis. This misleading performance of the NAP priors for true standardized effect sizes less than 0.05 persists, and even degrades, for sample sizes up to 4,000. When the sample size is 1,000 (Fig. 3.4(b)), the default NAP prior and the JZS priors begin to

show positive support (i.e., $\log(\mathrm{BF}_{10}(\mathbf{x})) > 1$) for standardized effect sizes greater than about 0.09. None of the alternative models depicted in Fig. 3.4(b) provide, on average, strong support for the alternative hypothesis for any standardized effect size less than 0.1. If the sample size is increased to 2,000 (Fig. 3.4(c)), then the JZS priors and default NAP prior provide, on average, strong evidence for standardized effect sizes greater than about 0.08, and positive evidence for effect sizes greater than about 0.065 (JZS) or 0.07 (default NAP). Increasing the sample size to 4,000 (Fig. 3.4(d)) yields a similar pattern, except that "very strong" weight of evidence is obtained, on average, for standardized effect sizes greater than about 0.07 if the default NAP or JZS priors are used to define the alternative hypothesis.

The conclusions from Figs. 3.2–3.4 might be simply stated as follows. Alternative hypotheses defined with NAP priors can provide strong or very strong weight of evidence in favor of true null hypotheses for small or moderate sample sizes (i.e., < 400). In many practical settings (i.e., $n < 2000$), JZS or other local priors cannot. For small to medium standardized effects (i.e., in $(0.2 - 0.5)$), alternatives defined with the default NAP prior provide, on average, slightly higher weight evidence for small to moderate sample sizes than do JZS priors with standard scale specifications. For medium or larger standardized effect sizes (> 0.6), alternative hypotheses defined with JZS priors provide higher average weights of evidence, with all specifications providing strong or very strong weights of evidence for sample sizes greater than 40. Alternative hypotheses defined with JZS priors provided higher average weight of evidence for very small effect sizes (i.e., < 0.10), but require large or very large sample sizes (> 2000) to provide strong support. Nearly identical conclusions apply to two-sample t tests and z tests.

3.3.3 An Application to incidental disfluency studies

To illustrate the use of NAP-based Bayes factors on real data, we applied them to replications of an incidental disfluency study, one of the 28 studies included in the "Many Labs 2" project [72]. Data for this example are available from the Open Science Framework (OSF) (`https://osf.io/8cd4r/`). For purposes of illustration, we restrict attention to Bayes factors based

on default NAP and JZS priors.

In the original disfluency study, [73] investigated whether a slow, analytical, and deliberate processing style can be activated by metacognitive experiences of difficulty or disfluency during the process of reasoning. To test this hypothesis, participants in the study were asked questions after reading two-statement syllogisms presented in either a hard-to-read or an easy-to-read font. Forty-one undergraduates from Princeton University completed a questionnaire that contained one of six syllogistic reasoning problems. The syllogisms were selected based on their accuracy rates established in prior research: two were easy, two were moderately difficult, and two were very difficult. Alter's original study compared responses based on the two moderately difficult syllogisms. Participants were randomly assigned to a questionnaire printed in either a hard-to-read (disfluent) or an easy-to-read (fluent) font. Each questionnaire contained 6 questions and the number of questions correctly answered by each participant was recorded as the response.

The study was subsequently replicated 13 times by researchers in multiple countries. To minimize differences between replications, "English in-lab" questionnaires were used in the following analyses (i.e., on the OSF website, "English" from the "Language" column and "In a lab" from the "setting" column). In total, these studies collected 2,580 responses, 1,268 from the fluent condition and 1,312 from the disfluent condition.

In previous analyses of these data, authors of the original and replication studies used two-sample t tests to test the null hypothesis that the mean number of correct responses from the two conditions were the same. Following their lead, we assume that the sample means of correct responses under fluent and disfluent conditions are independently and normally distributed with means μ_f and μ_d, respectively, and unknown common variances σ^2/n_f and σ^2/n_d, where n_f and n_d are the numbers of subjects responding under the fluent and disfluent conditions. The tested hypotheses can then be expressed in frequentist terms as

$$H_0 : \mu_f - \mu_d = 0 \qquad \text{vs.} \qquad H_1 : \mu_f - \mu_d \neq 0. \tag{3.18}$$

Prior	Weight of evidence
Default NAP	-4.33
Default JZS	-2.81

Table 3.1: Weight of evidence accumulated by the default NAP and JZS priors in favor of H_1 in (3.18) in a fixed-design test.

The P-value for the two-sample t-test is 0.43, which does not support the rejection of the null hypothesis of no effect. Neither does it provide an interpretable summary of evidence in favor of the null.

Taking a Bayesian perspective, we computed Bayes factors from these data by using default NAP and JZS priors on the standardized difference $(\mu_f - \mu_d)/\sigma$ to define the alternative hypothesis H_1.

Table 3.1 displays the weight of evidence accumulated by the each test using all 2,580 responses. The table shows that both priors favor the null hypothesis. The Bayes factor based on the default JZS prior fails to provide "strong" evidence against the null, whereas the NAP-based Bayes factor does. These values correspond to odds (i.e. Bayes factors) of about 17:1 in favor of H_0 using the JZS prior, and 76:1 using the NAP-based Bayes factor. In other words, over four times more support is obtained in favor of the null hypothesis when the NAP is used to define the alternative hypothesis.

3.4 Sequential tests

Unlike fixed sample size tests, sequential testing procedures are designed to terminate as soon as compelling evidence has been collected in favor of either the null or alternative hypothesis. After each subject or group of subjects is observed, they employ a rule that determines whether to (i) continue to collect data, (ii) stop data collection and reject the null hypothesis, or (iii) stop data collection and reject the alternative hypothesis. An important advantage of sequential designs is that they offer a potential mechanism for reducing the number of subjects that are needed to perform statistical tests.

Sequential tests have been developed extensively since their introduction by Wald in the 1940s [e.g., 12]. For a comprehensive review of developments in the statistical theory underlying sequential probability ratio tests (SPRT), see [74]. More recently, sequential designs have been proposed for application in psychology and other social sciences by [2], [3], [54], and [63].

[2] proposed a sequential Bayes factors (SBF) in which data are collected until the Bayes factor crosses predefined thresholds. They discuss a variety of prior densities on the standardized effect size that might be used to define the alternative model and Bayes factor, but recommend as a default the JZS prior with scale parameter $r = \sqrt{2}/2$. [3] discuss a modification of Wald's SPRT that applies to two-sample t tests [75]. The thresholds for making a decision in this test are chosen to maintain Type I and Type II error control. Hajnal's test is based on computing the ratio of non-central t and F sampling densities under the alternative hypotheses to central t and F densities that apply under the null. [3] do not provide objective criteria or default values for the standardized effect sizes that define the non-centrality parameters in these tests. [63] discuss the connections between [2] and [3], pointing out that the thresholds for the Bayes factors in the former can be adjusted to control Type I and II error probabilities. Readers interested in more detailed descriptions of these and related sequential testing procedures are encouraged to consult [2] and [3].

[54] propose a modification of the SPRT that they call the modified sequential probability ratio test (MSPRT). There are three innovations of this test. First, unlike the SPRT and SBF, the maximum sample size for the MSPRT is set in advance. Second, if a decision has not been reached after the maximum sample size is determined, a decision threshold is used at the end of the test to determine whether to the accept or reject each hypothesis. This threshold is estimated numerically so that the Type II error probability is minimized under the constraint that the targeted Type I error is maintained. Finally, the simple alternative hypothesis for the test is determined using the uniformly most powerful Bayesian test (UMPBT; [16]) at the maximum sample size, say N. In the test of whether a normal mean equals 0, the UMPBT alternative is of order $N^{-1/2}$. Successful application of the MSPRT thus implicitly depends on the selection of a maximum sample size that

is commensurate with the anticipated magnitude of the standardized effect size. For example, if $N = 10,000$, the point alternative hypothesis for the standardized effect size in a MSPRT for a one-sided z-test of size 0.05 is 0.01645σ ($= 1.645\sigma/\sqrt{10000}$). If the magnitude of the anticipated standardized effect size is substantially larger than this, then the UMPBT default value should not be used. [54] do not provide guidance on the selection of alternative values or maximum sample size for the test. Because the alternative hypotheses defined in this procedure depend on the maximum sample size specified for the test, it is difficult to compare its operating characteristics to the other sequential procedures, and so it is not considered in the comparisons below.

Results presented for fixed design tests suggest that the use of the JZS prior (as well as other local alternative priors) to define alternative hypotheses makes it difficult to accumulate evidence in favor of true null hypotheses and "very small" effect sizes. In sequential tests, this means that sequential procedures may not reach termination criteria before available sample sizes are expended when the null hypothesis is true. To resolve this difficulty, we propose using the default NAP prior to define the alternative hypothesis in these tests.

We now explore this proposal in the two contexts suggested by [2] and [3]. First, we examine the Bayesian approach and the SBF proposed in [2]. In this test, data is accumulated until the weight of evidence exceeds specified thresholds. After this, we examine the performance of both methods viewed from the frequentist perspective of [3] in which SPRT-type thresholds are determined so as to maintain specified Type I and Type II error probabilities.

3.4.1 Sequential design with symmetric evidence thresholds

3.4.1.1 Performance comparison

In this section, we again consider a one-sample two-sided t test of a normal mean μ, with $H_0 : \mu = 0$. In each simulated replication of the test, samples are collected until the weight of evidence in favor of the alternative hypothesis exceeds 3 or 5 or the weight of evidence against the alternative hypothesis is less than -3 or -5. The performances of the tests are summarized over

50,000 simulations of the tests at each standardized effect size.[2]

Several alternative hypotheses were considered. Following [2], we computed Bayes factors using the default JZS prior with scale parameter $r = \sqrt{2}/2$. We also examined the default NAP prior with $\tau^2 = 0.045$. Bayes factors for these procedures were computed using formulae provided in section two. We refer to the sequential testing procedures based on these prior assumptions as SBF-JZS and SBF-NAP. To facilitate comparisons with [3], we also tested the SPRT proposed in [75] with a simple alternative hypothesis defined to be a point mass at concentrated on standardized effect sizes of ± 0.3. This value matches the mode of the default NAP prior and matches the composite hypothesis examined in the previous section. As noted in [3], the Bayes factors from these tests are most efficient when the true standardized effect size is close to the assumed alternative hypothesis. For ease of exposition, we refer to the sequential test based on Hajnal's approximate Bayes factor with alternative hypothesis equal to a standardized effect size of magnitude d as the "Hajnal(d)" test.

3.4.1.2 True null hypothesis

Fig. 3.5 presents the boxplot of sample sizes and the ASN required by the SBF-NAP, SBF-JZS and Hajnal(0.3) tests to exceed thresholds of ± 3 and ± 5 when the null hypothesis is true. As the plots show, the SBF-JZS test typically requires significantly more samples to reach a decision. In the case of thresholds of ± 3, the ASN's for the SBF-JZS test and Hajnal(0.3) test were 968 and 99, respectively, while the ASN for the SBF-NAP test was 239. For a threshold of ± 5, the corresponding ASN's were 54,833, 164, and 1,026, respectively. These trends mimic those observed for fixed design studies.

[2]To manage simulation time for SBF-JZS, the sample size at each sequential step is increased following [2]. For a sequential comparison at the next step, we add 1 new sample until the total sample size (n) reaches 100, 5 new samples until n reaches 1000, 10 new samples until n reaches 2500, 20 new samples until n reaches 5000, and 50 new samples afterwards.

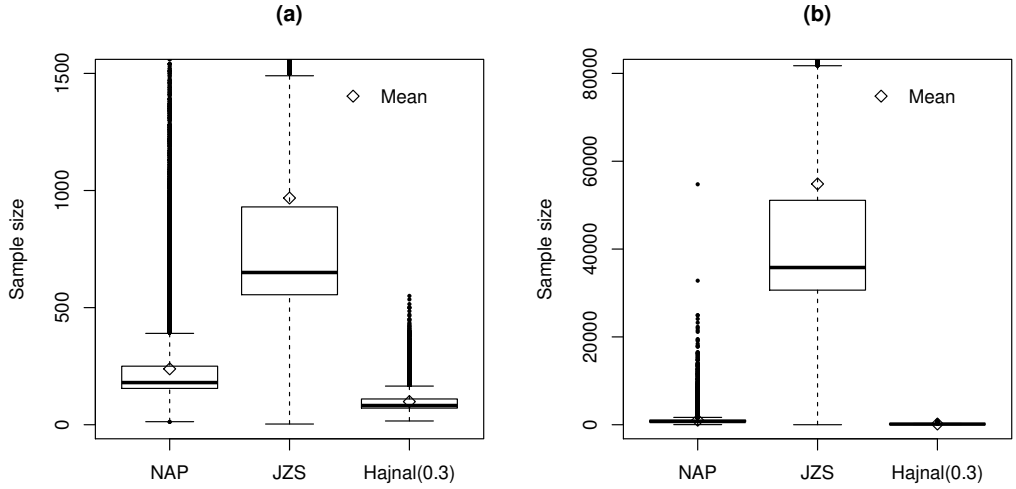


Figure 3.5: ASN for sequential procedures under a true null hypothesis. The plots are truncated at 1500 and 80,000 to enhance comparisons at moderate sample sizes. Panel (a) provides a boxplot estimate of the distribution of sample sizes required for the SBF-NAP, SBF-JZS and Hajnal(0.3) procedures to cross an exceedance threshold of ± 3. About 0.3% percent of SBF-NAP tests and 11% of SBF-JZS tests required more than 1500 samples to reach a decision. All Hajnal(0.3) tests terminated by 550 samples. Panel (b) provides the corresponding boxplots when the exceedance threshold is ± 5. About 12% of SBF-JZS tests required more than 80,000 samples to reach a decision. The black diamonds show the ASN's for each procedure. All SBF-NAP tests reached a decision by 54750 samples, and all Hajnal(0.3) tests terminated by observation 980.

3.4.1.3 True alternative hypothesis

The increased efficiency of the SBF-NAP and Hajnal(0.3) tests under the null hypothesis is offset by decreased power to detect smaller standardized effect sizes. This phenomenon is illustrated in Fig. 3.6. The panels on the left side of this figure represent the ASN and power achieved by the three sequential tests when an evidence threshold of ± 3 was imposed, while the panels on the right correspond to evidence thresholds of ± 5.

The general take-away from this figure is that the SBF-JZS provides substantially better power than the SBF-NAP test for standardized effect sizes less than 0.25 (left) or 0.10 (right). However, the cost of the additional power can be very high in terms of the ASN required to reach a decision. For instance, SBF-JZS requires ASN's that are greater than 50,000 to reach a decision for stan-

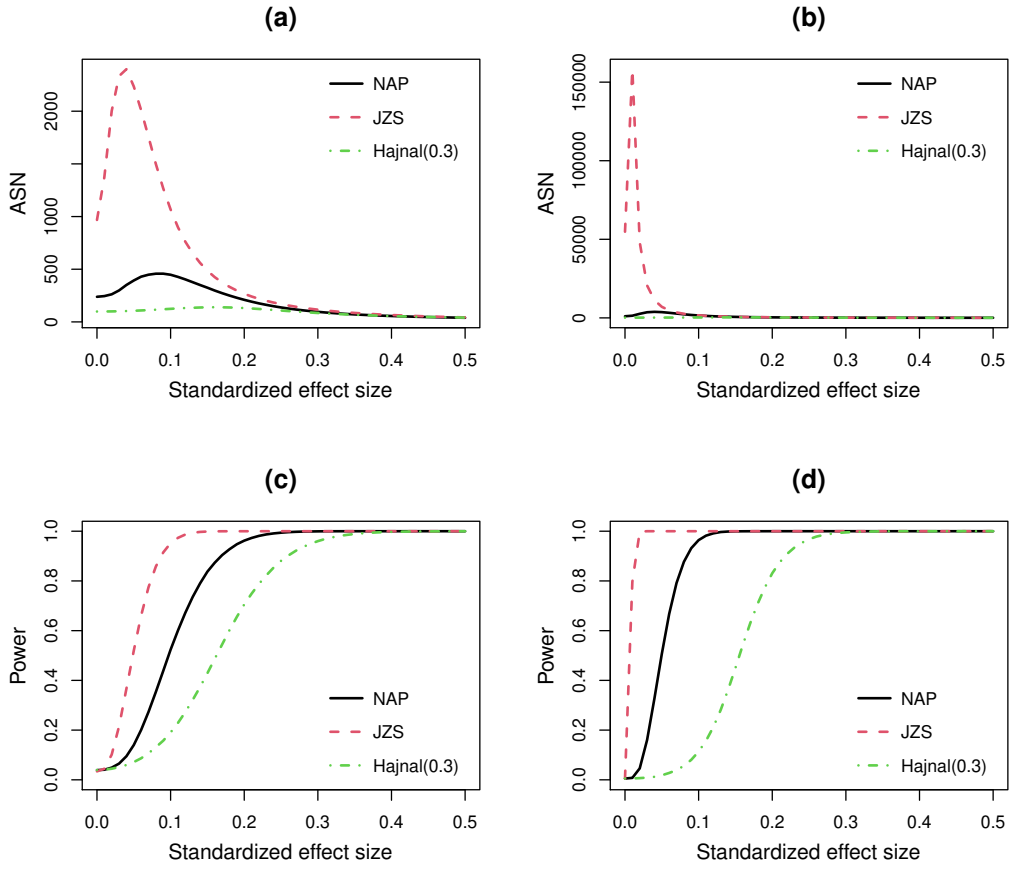


Figure 3.6: Operating characteristics under true alternative hypotheses. Panels (a) and (b) depict the ASN's for three sequential tests when the exceedance thresholds are ± 3 and ± 5, respectively, versus the data-generating value of the standardized effect size. Panels (c) and (d) provide the corresponding probabilities that each test rejects the null hypothesis as a function of the standardized effect size.

dardized effect sizes less than about 0.02 and weight of evidence thresholds of ± 5, even though the power at these smaller effect sizes can be well below 0.5.

3.4.1.4 Sequential analysis of the incidental disfluency study

In this section we compare the performances of the SBF-JZS and the SBF-NAP priors using the disfluency data described earlier. For brevity, we again only compare the default choices of NAP and JZS priors.

To perform a sequential analysis of the data collected in the 13 replicated studies, we assume

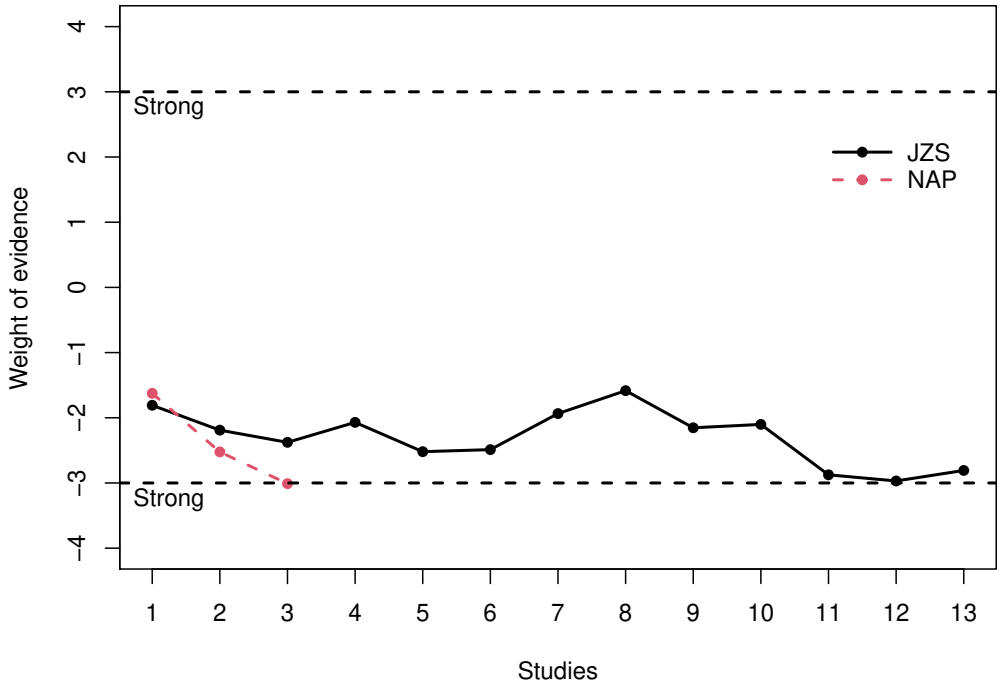


Figure 3.7: A comparison of the SBF-JZS and the SBF-NAP with symmetric "strong" thresholds in case of the replicated incidental disfluency data. For each prior the natural logarithm of the Bayes factor in favor of the alternative hypothesis that incidental disfluency activates a deliberate, analytic processing style is calculated. The curves corresponding to each prior depicts the sequentially calculated values after observing each of the 13 studies until they exceed ± 3. The horizontal axis displays the studies in the assumed order they were observed.

for illustration purposes that data from these studies was collected sequentially according to study number, and that all data from each study was collected simultaneously.

Given this ordering, we calculated the weight of evidence against the null hypothesis specified in (3.18) after data from each study "arrived." The weight of evidence was then computed using all available data. If the weight of evidence for one hypothesis was strong (i.e., > 3 or < -3), that test was terminated. The time courses for the accumulation of weight of evidence for the SBF-NAP and SBF-JZS procedures are displayed in Figure 3.7.

From the figure we see that weight of evidence from the SBF-JZS approaches, but never

crosses, the strong weight of evidence threshold. In contrast, the SBF-NAP test provides strong weight of evidence in favor of the null hypothesis after Study 3, using only 588 of the 2,580 combined study participants. Thus, application of the SBF-NAP procedure uses nearly 2,000 fewer subjects to conclude that there is a negligible disfluency effect, while at the same time providing stronger evidence in favor of this conclusion.

3.4.2 Sequential design with the SPRT thresholds

3.4.2.1 Performance comparison

The sequential probability ratio test, as proposed by [12], is based on comparing the likelihood ratio between a simple null and a simple alternative hypothesis and terminating an experiment when the likelihood ratio strongly favors one of the two. More specifically, let $x_1, x_2, \ldots$ represent independent, identically distributed realizations from a distribution with density function $f(x;\theta)$ under both hypotheses. Suppose the null hypothesis H_0 stipulates that $\theta = \theta_0$ and the alternative hypothesis H_1 that $\theta = \theta_1$. Then the likelihood ratio statistic in favor of the alternative hypothesis based on the first n observations may be expressed as

$$L(\theta_0, \theta_1; \mathbf{x}_n) = \prod_{i=1}^{n} \frac{f(x_i;\theta_1)}{f(x_i;\theta_0)}. \tag{3.19}$$

Wald's SPRT continues data collection until $L(\theta_0, \theta_1; \mathbf{x}_n) > A$ and the null hypothesis is rejected, or $L(\theta_0, \theta_1; \mathbf{x}_n) < B$ and the alternative hypothesis is rejected. The decision thresholds are defined as

$$A = \frac{1-\beta}{\alpha} \qquad \text{and} \qquad B = \frac{\beta}{1-\alpha}. \tag{3.20}$$

Typical design parameters in the social sciences and medicine often assume that Type I and Type II errors fall in the range $(0.005, 0.05)$ and $(0.05, 0.2)$, respectively. It follows that the SPRT thresholds for $(\alpha, \beta) = (0.05, 0.2)$ are $A = 16$ and $B = 0.21$, and for $(\alpha, \beta) = (0.005, 0.05)$ are $A = 190$ and $B = 0.05$.

[63] point out that the SPRT can be modified for use with composite hypotheses by replacing the likelihood ratio with the Bayes factor between hypotheses. [75] and [3] extended the SPRT to t tests by replacing the likelihood ratio for normally distributed data with unknown means and common variance by the ratio of a non-central t density to a central t density, evaluated at the t statistic for the experiment (e.g., $t = \sqrt{n}\bar{x}/s$). [63] provided numerical comparisons of the Hajnal test to the SPRT based on the Bayes factor defined with the JZS prior (and several other prior choices). We now extend this comparison to include the SPRT obtained by using the default normal moment prior (default NAP) density to define the alternative hypothesis. Before doing so, however, it is useful to compare the SPRT thresholds to the symmetric thresholds examined in the previous section.

For $(\alpha, \beta) = (0.05, 0.2)$, the Bayes factor thresholds are $A = 16$ and $B = 0.21$, with $\ln(A) = 2.77$ and $\ln(B) = -1.56$. The latter value represents the threshold at which the alternative hypothesis is rejected. It is substantially smaller in magnitude than the thresholds of -3 and -5 examined previously. With prior odds equal to 1, weight of evidence equal to -1.56 implies that the posterior probability of the alternative hypothesis is 0.17, which might be considered too high for rejection. The use of this less stringent threshold for "accepting" the null hypothesis reduces the ASN required by the SBF-JZS test. Values of $(\alpha, \beta) = (0.005, 0.05)$ yield weight-of-evidence thresholds that are more similar to those studied in the previous section. With prior odds equal to 1, the alternative hypothesis is not rejected unless it has posterior probability less than 0.05, and the null hypothesis is not rejected unless it has posterior probability less than 0.0052.

3.4.2.2 True null hypothesis

Fig. 3.8 depicts the ASN for three sequential tests when Type I and Type II error probabilities were constrained to $(0.05, 0.20)$ (left panel) and $(0.005, 0.05)$ (right panel). As in the previous section, all three sequential tests were designed to test the null hypothesis that the mean of a sample of normal random variables with unknown variance was equal to 0. The boxplots in this figure were based on 50,000 replications of each test. The decision thresholds for each test were

set according to (3.20), and data for each test were simulated under the null hypothesis that the standardized effect size was 0.

The three tests included in the plot include the SPRT based on the Bayes factor obtained by defining the alternative hypothesis with the default NAP on the standardized effect size (i.e., a normal moment prior with $\tau^2 = 0.045$), the SPRT based on the Bayes factor obtained by defining the alternative hypothesis with the default JZS prior on the standardized effect size ($r = \sqrt{2}/2$), and the [3] version of Hajnal's two-sided t-test with a composite hypothesis that assigned one-half probability to $\pm 0.3\sigma$.

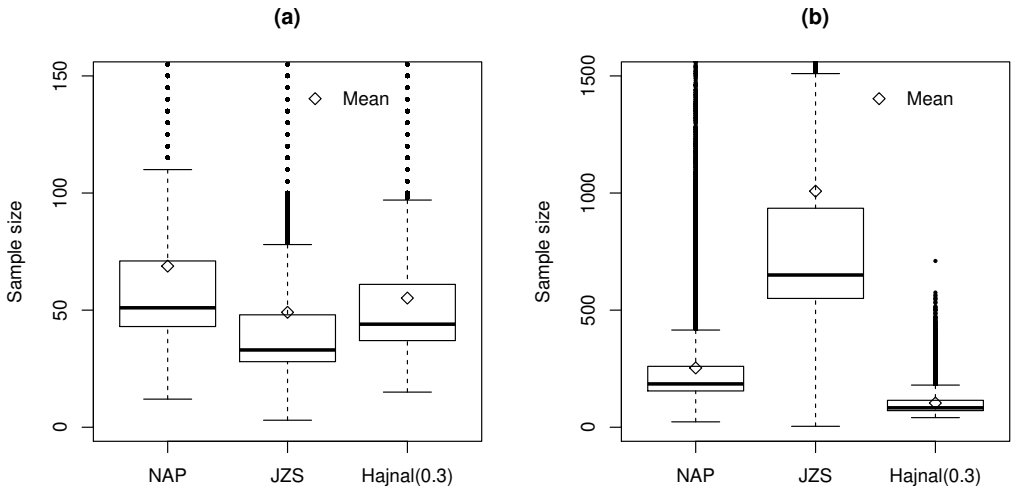


Figure 3.8: ASN for SPRT procedures when the null hypothesis is true. Panel (a) provides a boxplot estimate of the distribution of sample sizes required for the SBF-NAP, SBF-JZS and Hajnal(0.3) procedures to cross Wald's decision thresholds at $\alpha = 0.05$ and $\beta = 0.2$. The plot is truncated at 150 samples (5.49% of SBF-NAP tests, 3.35% of SBF-JZS tests, and 1.75% of Hajnal(0.3) tests required more than 150 samples). Panel (b) provides the corresponding estimate when Wald's decision thresholds were based on $\alpha = 0.005$ and $\beta = 0.05$. The plot is truncated at 1500 samples (0.54% of SBF-NAP and 11.1% of SBF-JZS tests required more than 1500 samples; none of Hajnal(0.3) tests did). The black diamonds show the ASN for each procedure.

The left panel of Fig. 3.8 shows that the test based on the JZS alternative required the smallest mean and median ASN when the targeted Type I and Type II errors were 0.05 and 0.2, respectively.

The realized Type I errors for the tests were 0.035, 0.043, and 0.044 for the alternative hypotheses defined by the JZS, composite, and NAP priors.

The right panel depicts similar findings when the targeted Type I and Type II errors were 0.005 and 0.05, respectively. With thresholds again determined from (3.20), the ASN required by the JZS test jumps significantly at the more stringent significance threshold, requiring an average of over 1,000 observations before reaching a decision. The NAP and composite tests required an average of 253 and 103 observations, respectively.

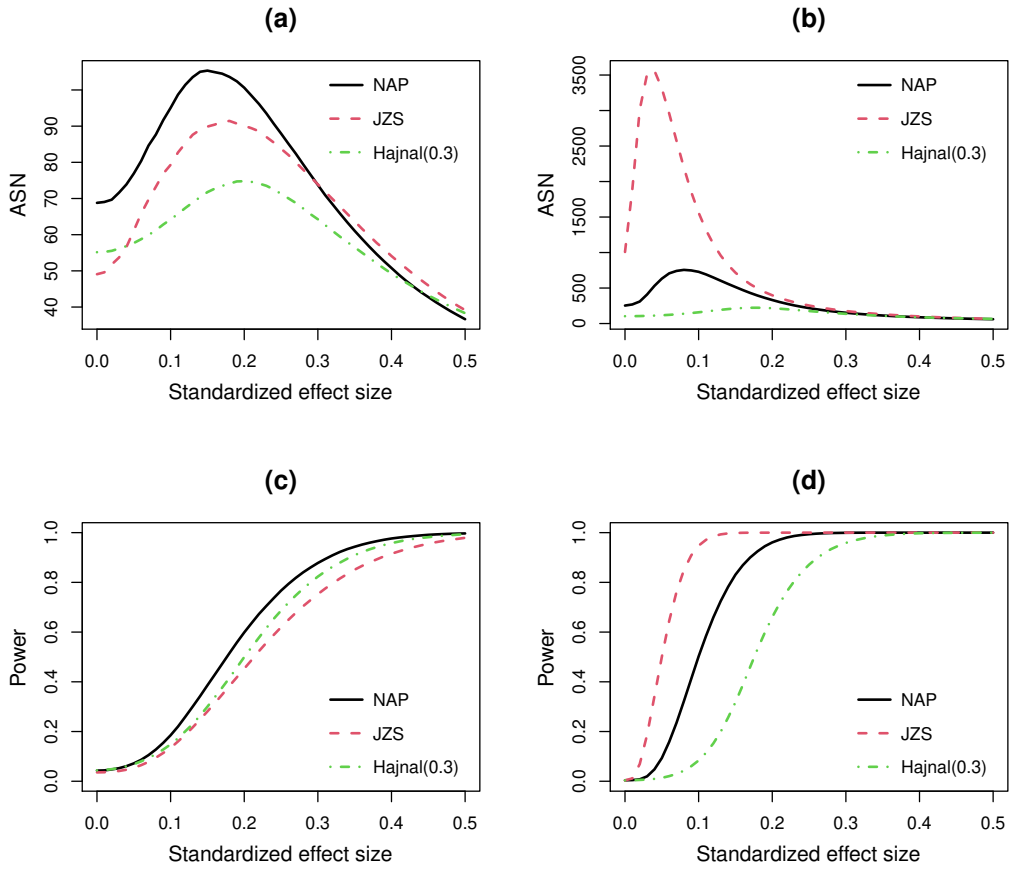


Figure 3.9: Operating characteristics under true alternative hypotheses. Panels (a) and (b) depict the ASN for three SPRT procedures based on Wald's decision thresholds for $(\alpha, \beta) = (0.05, 0.2)$ and $(0.005, 0.05)$, respectively, versus the data-generating value of the standardized effect size. Panels (c) and (d) provide the probability that each procedure rejected the null hypothesis as a function of the standardized effect size.

3.4.2.3 True alternative hypothesis

Fig. 3.9 provides the ASN and power of each of the three sequential tests as a function of true standardized effect size. As in Fig. 3.8, the panels on the left (a,c) reflect the operating characteristics of the test with targeted Type I and Type II error probabilities equal to 0.05 and 0.2, while panels (b,d) targeted to error probabilities of 0.005 and 0.05.

From panels (a) and (c), we see that the NAP prior requires, on average, a higher number of samples to reach a decision for standardized effect sizes less than about 0.3 (JZS) or 0.42 (composite), although it provides better power over the range of standardized effect sizes depicted. True standardized effect sizes of 0.27, 0.29, and 0.33 are needed for the NAP, composite, and JZS to reach their targets of 80%. For the composite alternative hypothesis, this value is close to the point mass alternatives used to define the test.

Panels (b) and (d) reveal a somewhat different trend for the more stringent tests. With error probability targets of $(0.005, 0.05)$, the ASN for the test defined with the JZS alternative can be as large as 3,500. However, these larger sample sizes provide higher power, with 95% power achieved for standardized effect sizes greater than 0.1, whereas the tests defined with the composite and NAP priors only provide 95% power for standardized effect sizes greater than 0.29 and 0.19, respectively. As in the less stringent test, the composite hypothesis achieves its targeted power at the point mass alternatives used in its definition.

3.4.2.4 Sequential analysis of the incidental disfluency study (continued)

We previously examined the efficacy of the SBF-NAP and SBF-JZS tests in accumulating strong evidence in favor of the null hypothesis against a disfluency effect using symmetric exceedance thresholds. We now consider a similar analysis using the weak ($(\alpha, \beta) = (0.05, 0.20)$) and stringent $(\alpha, \beta) = (0.005, 0.005)$ Wald thresholds. Note that the weight-of-evidence curves displayed in Fig. 3.7 for the SBF-JZS and SBF-NAP tests do not change according to the termination thresholds that are used.

The weight of evidence thresholds that correspond to the less stringent criterion of $(\alpha, \beta) =$

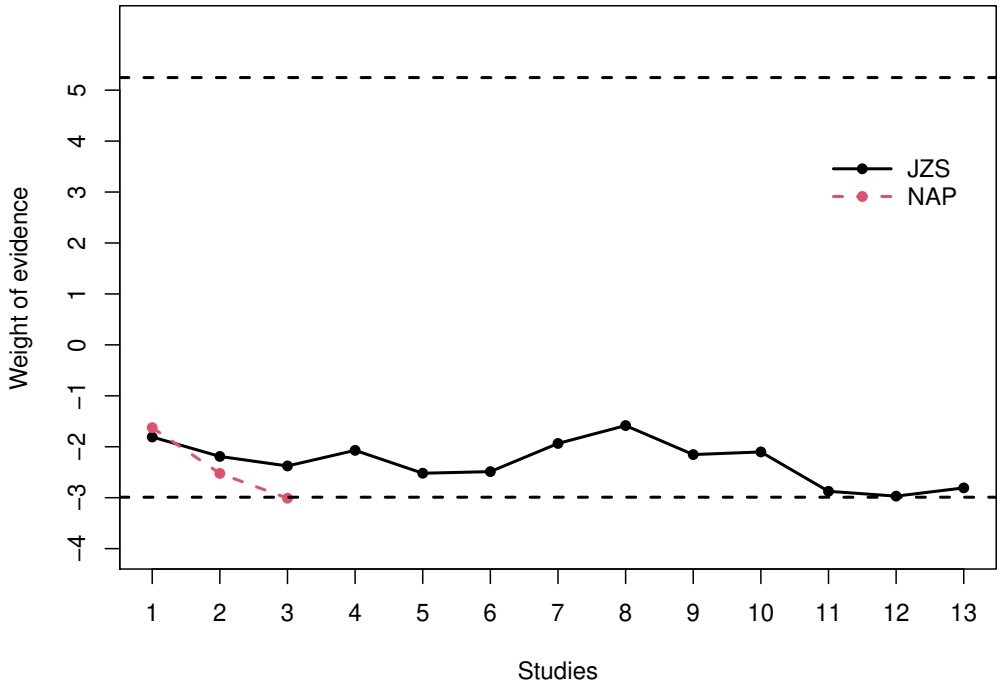


Figure 3.10: A comparison of the SBF-JZS and the SBF-NAP with the SPRT thresholds in case of the replicated incidental disfluency data. For each prior the natural logarithm of the Bayes factor in favor of the alternative hypothesis that incidental disfluency activates a deliberate, analytic processing style is calculated. The curves corresponding to each prior depicts the sequentially calculated values after observing each of the 13 studies until they exceed the SPRT thresholds corresponding to $(\alpha, \beta) = (0.005, 0.05)$. The horizontal axis displays the studies in the assumed order they were observed.

$(0.05, 0.20)$ are $A = 2.77$ and $B = -1.56$. As Fig. 3.7 suggests, both the SBF-JZS and SBF-NAP prior fall below the lower threshold after the first study (weights of evidence equal to -1.81 and -1.63, respectively).

The weight of evidence thresholds that correspond to the more stringent criterion of $(\alpha, \beta) = (0.005, 0.05)$ are $A = 5.24$ and $B = -2.99$. Because the lower threshold is close to -3.0, the conclusions from the last section apply here also: The SBF-NAP test terminates after the third study in favor of the null hypothesis and uses only 588 subjects, while the SBF-JZS does not terminate even after responses from all 2580 subjects are accumulated.

Prior	Strong	Very strong
Default NAP	294	1,208
Default JZS	1,445	79,424

Table 3.2: Average sample numbers required for fixed-design tests under true null hypotheses. This table displays the minimum sample sizes required for Bayes factors to achieve, on average, strong ($\log(BF_{01}) \geq 3$) or very strong weight of evidence ($\log(BF_{01}) \geq 5$) in favor of true null hypotheses.

3.5 Discussion

This chapter has explored the use of non-local alternative prior densities, or NAP's, to define alternative models in Bayesian z and t tests. From a subjective perspective, evidence suggests that NAPs approximate the marginal distribution of non-null effect sizes observed in the psychology and social science literature [64, 65, 66, 67, 68, 69, 70]. Viewed more objectively, the operating characteristics of Bayesian tests based on NAP's provide an opportunity for researchers to more rapidly accumulate evidence in favor of true null hypotheses and alternative hypotheses in which standardized effect sizes are moderate in magnitude.

Table 2 illustrates this effect when the null hypothesis is true. Sample sizes required to obtain strong weight of evidence, on average, are nearly 5 times larger using the JZS specification than the NAP specification. To obtain very strong weight of evidence, the sample size required by the JZS specification needs to be 65 times larger. Table 3 provides a similar comparison when the alternative hypothesis is true. Evidence for small and medium standardized effect sizes accumulates faster, although the gains for these alternative hypotheses is less pronounced. Bayes factors based on default JSZ priors outperform those based on default NAP priors for large effect sizes, an advantage that increases with increasing standardized effect size. Of course, the sample sizes required to detect large effects tend to be fairly small no matter which alternative is specified.

Tables 4 and 5 demonstrate that similar trends persist for sequential tests based on Bayes factors. In the case of tests with symmetric thresholds, however, the smaller ASN's achieved by the

Prior	Small effect		Medium effect		Large effect	
	Strong	Very strong	Strong	Very strong	Strong	Very strong
Default NAP	225	335	37	56	21	31
Default JZS	267	379	42	62	19	28

Table 3.3: Average sample numbers required for fixed-design tests under true alternative hypotheses. This table displays the average sample sizes required for Bayes factors to achieve strong ($\log(BF_{10}) \geq 3$) or very strong weight of evidence ($\log(BF_{10}) \geq 5$) for small (0.2), medium (0.5) and large (0.8) standardized effect sizes.

NAP-based Bayes factors should be balanced against the fact that these tests have high probability of generating evidence in favor of the null hypothesis when the magnitude of a standardized effect sizes is less than 0.1.

Perhaps related to this trade-off, [76] argue that Bayes factors can either favor a point null hypothesis (Issue 9) or an alternative hypothesis (Issue 10). With regard to the latter, they cite [25] and express concern that evidence is accumulated asymmetrically in favor of the alternative model. [77] correctly point out that 'the claim that something is absent is more difficult to support than the claim that something is present, at least when one is uncertain about the size of the phenomenon that is present. Consider, for instance, the null hypothesis "There is no animal in this room," tested against the alternative hypothesis: "There is an animal in this room, but it could be as small as an ant or as big as a cow". Now if the "effect" is of medium size (say a cat), it can be quickly

Priors	Symmetric thresholds	
	Strong	Very strong
Default NAP	238	1,026
Default JZS	968	54,832

Table 3.4: Average sample numbers for sequential tests under true null hypotheses. Columns refer to the average sample sizes required for Bayes factors to exceed, on average, strong ($|\log(BF_{01})| \geq 3$) or very strong weight of evidence ($|\log(BF_{01})| \geq 5$) thresholds when termination thresholds are symmetric.

Priors	Symmetric thresholds	
	Strong	**Very strong**
Default NAP	458	3,853
Default JZS	2,399	158,235

Table 3.5: Maximum average sample numbers for sequential tests under true alternative hypotheses. This table does not reflect the power of the tests, which for standardized effect sizes less than 0.2 is greater for the default JZS prior with symmetric thresholds. Columns list the maxiumum of the ASN required for Bayes factors to exceed, on average, strong ($|\log(BF_{10})| \geq 3$) or very strong weight of evidence ($|\log(BF_{10})| \geq 5$) thresholds. The power and standardized effect sizes at which these values obtain can be discerned from Fig. 6.

discovered and H1 then receives decisive support. But if a cursory inspection does not reveal any animal, then support for H0 will only be weak (after all, it is easy to miss an ant). Now there is a way to collect strong evidence for H0, but it requires more effort – a systematic search with a magnifying glass, for instance.'

Theoretical support for this statement can be found in the pioneering work of [78] and [79], who showed that likelihood ratios and Bayes factors in favor of true null hypotheses and true alternative hypotheses increase exponentially fast with sample size when the parameter spaces associated with the two hypotheses are separated. Sub-exponential convergence occurs when the parameter defining one hypothesis falls on the boundary between the spaces. This is the case with NHSTs, where the null parameter value is not separated from parameter values that define alternative hypotheses. One objective of NAP-based tests is to approximately "separate" the hypotheses. This goal is complicated by the desire to avoid discontinuities in the prior densities that define the alternative hypotheses. For example, assigning positive prior density to say, 0.3, and 0 density to all smaller values may not make sense.

The comments of [77] illustrate this principle well. If only animals larger than cats are considered–so that the hypotheses are well separated–then one can test "no animal present" versus "animal present" very quickly. If ants and even smaller animals count, then the null hypothesis is not well separated from the alternative and testing takes longer. For the NAP-based tests proposed in this

chapter, Bayes factors in favor of true null hypotheses increase at a rate of $n^{3/2}$. For local alternative hypotheses, this rate is only $\sqrt{n}$ [25]. In contrast, the rate for any true alternative hypothesis, which is always distinct from the null value, increases exponentially fast with n.

The default NAP-based tests proposed in this chapter should not be categorized as objective Bayesian tests because they explicitly target the detection of standardized effect sizes of most interest in psychology and other social sciences. Nevertheless, it is interesting to examine their properties using criteria that are sometimes used to judge the performance of objective Bayesian tests. As summarized in, for example, [80] and [81], such criteria include basic (Bayesian) consistency, model selection consistency, intrinsic consistency, information consistency, predictive matching, and scale-location invariance.

NAP-based tests satisfy basic and intrinsic consistency since they are Bayesian tests that do not depend on arbitrary normalizing constants, training sample sizes, or other arbitrary effects that do not disappear with increasing sample sizes. Model selection consistency requires that the posterior probability of the true model converges to 1 as the sample size increases. The NAP-based z and t tests proposed here satisfy this criterion. The Hajnal tests do not.

The NAP-based z tests proposed in this chapter satisfy information consistency. That is, they are able to obtain unbounded evidence against the null hypothesis for arbitrarily extreme observations based on any given sample size. When the observational variance is known, an arbitrarily large sample mean (or difference in sample means) can provide arbitrarily high evidence against the null hypothesis, regardless of the sample size.

NAP-based t tests are not information consistent. We do not regard this as a shortcoming of the tests, however. In our view, it should not necessarily be possible to obtain unbounded information in favor of an alternative hypothesis using a finite sample of measurements if the properties of the measuring device or error structure (e.g., the variance) are not known. This is particularly true when prior knowledge suggests that the value of the tested parameter under the null and alternative hypotheses are not too dissimilar.

As a final comment on this issue, we note that lack of information consistency for NAP-based t tests cannot be attributed to the improper prior assumed for the observational variance. Even if a proper inverse gamma distribution is assumed on the observational variance, NAP-based t tests do not attain information consistency (see Theorem S2.8 of Supplemental Materials). That is, formal Bayesian tests based on fully specified statistical models with proper priors on all unknown parameters may not satisfy the information consistency criterion.

Because the NAP-based tests are functions of z and t statistics, they inherit the invariance properties of those test statistics.

Predictive matching requires that Bayes factors between models based on "minimal" sample sizes should approximately equal 1. Exact predictive matching requires that they exactly equal 1. Minimal sample sizes can be loosely interpreted as the smallest sample size that makes maximum likelihood estimation possible for all parameters in all models. In the case of a one-sample t test, for example, the minimal sample size necessary to estimate the mean and variance is 2 if improper priors are specified on both parameters.

Predictive matching and information consistency are antithetical for minimal sample sizes. Predictive matching requires that the Bayes factor remain close to 1 whenever a minimal sample has been obtained, while the information consistency requires that the Bayes factor can become unbounded for extreme data. Given the discussion above, it is therefore not surprising that NAP-based z tests are not predictive matching and that NAP-based t tests are. In the former case, the minimal sample size is 1 and the Bayes factor grows exponentially with the magnitude of a single observation. For one sample t tests and minimal sample size of 2, the NAP-based Bayes factors range between $(1 + 2\tau^2)^{-3/2}$ and $(1 + 4\tau^2)/\sqrt{1 + 2\tau^2}$. For the default value $\tau^2 = 0.045$, the corresponding range is approximately $(0.88, 1.13)$.

Our interpretation of these results is that predictive matching and information consistency desiderata are not useful as general criteria for defining Bayesian tests. On one hand, a single large normal observation with known variance can provide very strong evidence against a null hypothesis that a normal mean equals 0. If the minimal sample size is 1, then accepting such evidence

violates the predictive matching criterion. On the other, the posterior probability against a null hypothesis of no effect should not necessarily become arbitrarily small based on a finite sample when there is uncertainty regarding the precision of the values that were measured.

This chapter has concentrated on default NAP-based tests in which targeted standardized effect sizes fall in the range (0.2,0.8). However, in some testing contexts specific prior information regarding the magnitude of a standardized effect size may be known. For instance, a researcher may wish to detect a very small standardized effect size (e.g., < 0.2). In such cases, we recommend defining $\tau^2 = \delta_p^2/2$, where δ_p denotes the prior estimate of the standardized effect size or difference in standardized effect sizes.

To illustrate, suppose the magnitude of a standardized effect size is expected to be approximately $\delta_p = 0.05$ in a one-sample test of a normal mean with unknown variance. Then a good choice for τ^2 is $.05^2/2 = .00125$. A plot of this NAP density is provided in Fig. 3.11. If a NAP-based test is conducted with a normal moment prior with this value of τ^2, then the average weight of evidence from a fixed-design test with $n = 4000$ observations is slightly greater than 4.0. In contrast, the average weight of evidence for the default NAP-based test with $\tau^2 = 0.045$ is approximately -0.017, and for the default JZS-based test is 1.46 (see Fig. 4d). Thus, by including subjective prior information into a test, an investigator can substantially increase the evidence collected in favor of a very small standardized effect size.

Although this chapter has focused on two-sided tests, one-sided tests can also be conducted using formulae for Bayes factors provided in the supplemental information. The NAP prior used for one-sided tests are twice as large as the densities used for two-sided tests for either positive (or negative) standardized effect sizes, and 0 for negative (or positive) standardized effects. This implies that the weight of evidence in favor of a true alternative in a one-sized test can be as much as $0.69 = \ln(2)$ higher than in a two-sided test, and that the average weight of evidence in favor of true null hypotheses can also be higher, particularly when the sign of the sample mean of data disagrees with sign of the standardized effect size assumed under the alternative. Figures summarizing simulation studies for one-sided tests are provided in the supplemental materials.

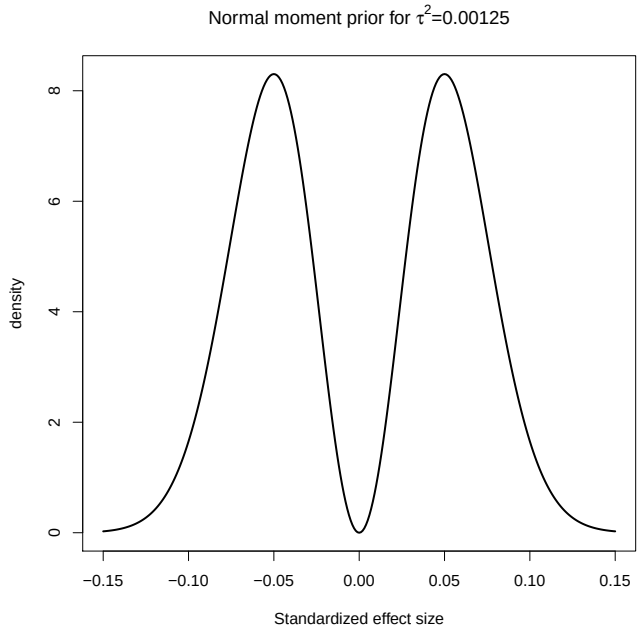


Figure 3.11: Normal moment prior for detecting a very small standardized effect. This normal moment prior density has peaks at ± 0.05 and places most of its prior mass on standardized effect sizes with magnitudes in the interval $(0.02, 0.10)$.

R functions [82] for implementing the NAP and Hajal tests described in this chapter are available at CRAN and GitHub.

4. EFFICIENT ALTERNATIVES FOR BAYESIAN HYPOTHESIS TESTS FOR PROPORTIONS

4.1 Bayesian Approaches for Testing Two Proportions

There exists two primary Bayesian approaches for testing proportion(s): (a) assuming prior distribution(s) directly on the proportion(s), and (b) take a logistic regression perspective and specify prior(s) on the logit transformed proportion(s) [83, 84, 85]. To illustrate, we focus on the two-sample proportion test. For notation purpose, suppose we respectively observe y_1 and y_2 successes out of n_1 and n_2 samples independently drawn from two populations where n_1 and n_2 are prefixed. We assume

$$y_1 \sim \text{Binomial}\left(n_1, p_1\right), \quad \text{and} \quad y_2 \sim \text{Binomial}\left(n_2, p_2\right), \tag{4.1}$$

where p_1 and p_2 are population proportions of interest that are unknown and we want to test

$$H_0 : p_1 = p_2 \quad \text{vs.} \quad H_1 : p_1 \neq p_2. \tag{4.2}$$

In the Bayesian paradigm, we assume prior distributions on the respective model parameters under H_0 and H_1 that reflect our prior beliefs on the parameters. To quantify evidence accumulated from the observed data, we then compute the Bayes factor in favor of the alternative hypothesis to choose between the two competing hypotheses.

4.1.0.0.1 Independent Beta approach. There are two common choices when specifying priors on the proportion scale. The simplest way is assuming independent Beta priors on the proportions under both hypotheses [84, 85, 86]. Henceforth, we refer to this as the "Independent Beta (IB) approach". For default purposes [84] and [85] suggest the Uniform prior, a special case of the Beta distribution. In this approach, the marginal density under the alternative, and hence the Bayes factor, can be obtained in closed form.

4.1.0.0.2 Difference approach. Another way of specifying priors on the proportion scale is to parameterize p_1 and p_2 as

$$p_1 = \zeta + \frac{\eta}{2}, \quad \text{and} \quad p_2 = \zeta - \frac{\eta}{2}. \tag{4.3}$$

Then the hypothesis testing problem in (4.2) is the same as testing

$$H_0 : \eta = 0 \quad \text{vs.} \quad H_1 : \eta \neq 0. \tag{4.4}$$

In order to specify priors, we note that η and ζ respectively take values in $(-1, 1)$ and $(0, 1)$ and they are dependent. Particularly, given η, ζ lies between $(|\eta|/2, 1 - |\eta|/2)$ in order for p_1 and p_2 to be valid proportions. One can then hierarchically specify priors on η and $\zeta \,|\, \eta$. A common choice is to assume $\eta \sim N\left(0, \sigma_\eta^2\right)$ truncated between $(-1, 1)$, and $\zeta \,|\, \eta \sim N(0, \sigma_\zeta^2)$ and truncated between $(|\eta|/2, 1 - |\eta|/2)$ [85]. Henceforth, this is referred to as the "Diff-Local approach". Due to the involved parameterization the marginal density under the alternative cannot be obtained in closed form and it needs to be calculated numerically.

4.1.0.0.3 Logit approach. Taking a Logistic regression perspective, this approach transforms the proportions on the Logit scale as

$$\text{Logit}(p_1) = \beta + \frac{\psi}{2}, \quad \text{and} \quad \text{Logit}(p_2) = \beta - \frac{\psi}{2}, \tag{4.5}$$

where $\text{Logit}(x) = \ln(x/(1-x))$. This implies

$$\beta = \frac{1}{2}\Big[\text{Logit}(p_1) + \text{Logit}(p_2)\Big], \quad \text{and} \quad \psi = \text{Logit}(p_1) - \text{Logit}(p_2). \tag{4.6}$$

Here β is interpreted as the mean of the log odds ratios, and ψ is the difference of the log odds. Following this setup, we test

$$H_0 : \psi = 0 \quad \text{vs.} \quad H_1 : \psi \neq 0 \tag{4.7}$$

for testing the equality of proportions as in (4.2). Unlike the IB and the Diff approach, here we assume priors on the transformed parameters β and ψ. This induces priors on the population proportions, the actual parameter of interest, due to the formulations

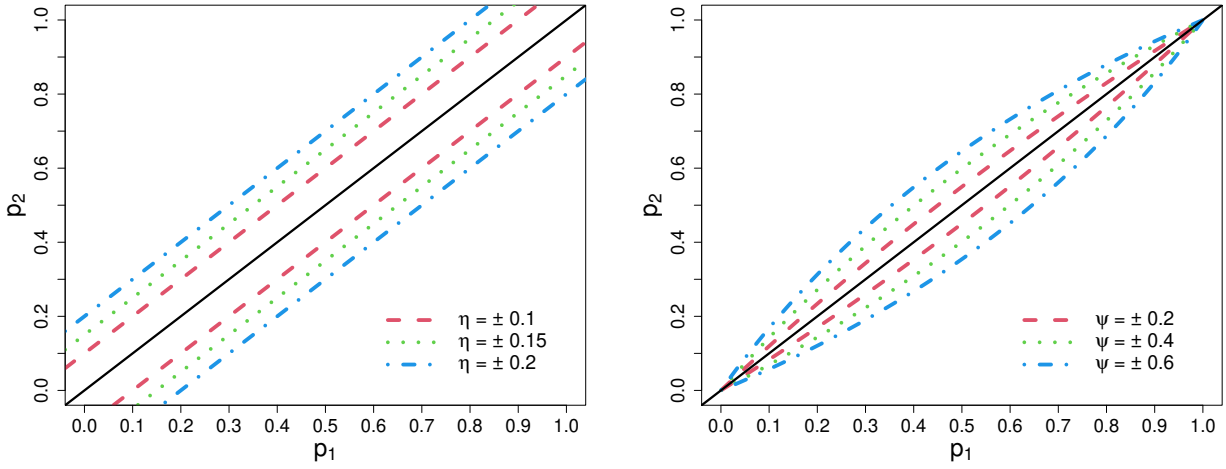
$$p_1 = \left(1 + e^{\beta+\psi/2}\right)^{-1}, \quad \text{and} \quad p_2 = \left(1 + e^{\beta-\psi/2}\right)^{-1}. \tag{4.8}$$

While specifying priors, note that β can also be interpreted as the Logit of the common population proportion when the null hypothesis is true. In the literature, a more recommended prior on proportion is the Beta distribution, particularly its non-informative variants like the Uniform or the Jeffreys prior [24]. [83] and [87] recommends the choice of $N(0, \sigma_\beta^2)$ prior on β under both hypotheses, and $N(0, \sigma_\psi^2)$ prior on ψ independent of β under H_1. To make the prior on β coherent with non-informative recommendations, one can either approximate it by assuming normal priors with appropriate σ_β (≈ 1.5 for Uniform and ≈ 3 for Jeffreys) or assuming the prior density

$$f(\beta) = \frac{1}{\text{Beta}(a, b)} \frac{e^{a\beta}}{(1 + e^{\beta})^{a+b}}, \quad \text{for } -\infty < \beta < \infty, \tag{4.9}$$

on β. It has been observed that the test is robust with respect to the choice of prior on β, but crucially depends on the prior on ψ. (4.9) induces an exact Beta distribution with shape parameters a and b on $\text{Logit}^{-1}(\beta)$, the common population proportion under true null hypothesis. Due to the parameterization (4.8), this specifies a dependent prior on (p_1, p_2) unlike in the IB approach where the priors are independent. They suggest a default value of 1 for σ_β and σ_ψ. Figures 2, 3 and D1 (in the appendix) in [85] provides visuals of priors on (p_1, p_2) from all these approaches. Due to the non-linear Logit transformation, the marginal density under the alternative cannot be obtained in closed form and it needs to be calculated numerically. Henceforth, this is referred to as the "Logit-Local approach".

4.1.0.0.4 Fundamental difference between the two approaches. Specifying prior(s) on the proportion scale and taking a logistic regression perspective are two different approaches that are often taken for one- and two-sample proportion test. Although the null hypotheses in two approaches



(a) $p_1 - p_2 = \eta$ (b) $\text{Logit}(p_1) - \text{Logit}(p_2) = \psi$

Figure 4.1: Contours of proportion pairs (p_1, p_2) satisfying a pre-specified difference η (on the left) and a pre-specified log-odds ψ (on the right). The solid black like denotes the proportion pairs consistent with the null hypotheses under respective approaches.

correspond to each other, the proportions that are consistent with the alternative in the two approaches are fundamentally different. For example, let us consider two-sample proportion tests. Figure 4.1 shows proportion pairs that satisfies $p_1 - p_2 = \eta$ and $\text{Logit}(p_1) - \text{Logit}(p_2) = \psi$ for different values of η and ψ. Comparing the figures we note that if p_1 and p_2 are moderate (near 0.5), the two approaches treat them similarly. When they are both small or large, the two approaches treat them very differently. For example, let $p_1 = 0.1$ and $p_2 = 0.05$. Their absolute difference is 0.05 which is small in the absolute sense. The absolute difference of log-odds is 0.74 which is substantially large. Thus a Logit approach is particularly suitable in detecting true proportions that are both very small/large and unequal.

4.2 Non-local Alternative Prior Densities for Proportion Tests

NAPs are probability density functions that take the value 0 at parameter values that are consistent with the null hypothesis [25]. [88] have recently described the use of "non-local" alternative hypotheses in Bayesian hypothesis testing of population mean(s) based on one- and two-samples. Under the alternative, they propose normal moment priors, a special case of the "non-local" den-

sity. The resulting class of Bayesian hypothesis tests permits more rapid accumulation of evidence in favor of both true null hypotheses and alternative hypotheses that are compatible with standardized effect sizes of most interest in psychology over classical testing procedures or their "local" alternatives. Below, we describe NAPs for one- and two-sample proportion tests.

4.2.1 One-sample Proportion Tests

Suppose we observe y successes out of n samples drawn from a population where n is prefixed. Mathematically, $y \sim$ Binomial (n, p) where p the true population proportion of interest and is unknown. For a prespecified p_0, one-sample proportion tests compare $H_0 : p = p_0$ against a two-sided alternative $H_1 : p \neq p_0$.

4.2.1.0.1 NAP on proportion. A NAP that can be used to define an alternative hypothesis on the proportion p for this test is the *Beta Moment Prior* density of order m, which can be expressed as

$$f_{BM}\left(p \,|\, p_0, K, \pi, m\right) = c_{BM}(p - p_0)^{2m} p^{K\pi}(1-p)^{K(1-\pi)}, \quad \text{for } p \in (0,1). \tag{4.10}$$

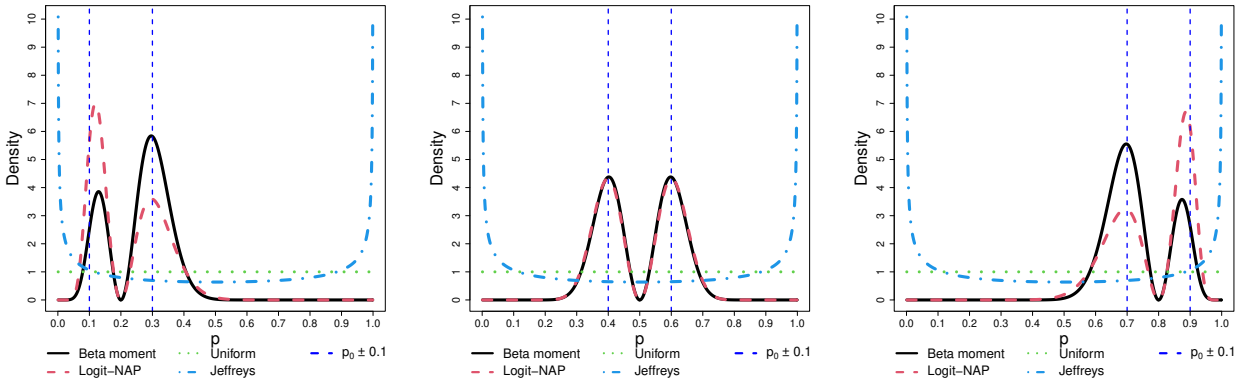
Here m and K are positive integers, and the Beta kernel of this moment prior is parameterized to have its mode at π taking value in $(0, 1)$. The proportionality constant $c_{BM} = 1/\mathrm{P}(0,0)$ where

$$\mathrm{P}(a,b) = \sum_{j=0}^{2m} (-1)^j \binom{2m}{j} {p_0}^{2m-j}\, \mathrm{B}\left(a + j + K\pi + 1, b + K(1-\pi) + 1\right), \tag{4.11}$$

for $a \geq 0$ and $b \geq 0$ with $\mathrm{B}(a,b)$ being the Beta function. Note that, the NAP density is exactly 0 at the null hypothesized value p_0. This makes the prior logically consistent under the alternative. The moment order m in the prior controls how much around the null hypothesized value we want to penalize under the alternative. We recommend $m = 1$ as the default which penalizes the least amount. Sensitivity of the prior to the choice of m is important, but for maintaining clarity hereon we fixed it at 1. Although π can be chosen based on prior belief, setting it to p_0, the hypothesized value under H_0, maximizes the NAP density for all p and K. So we recommend this for default implementation. Then K remains as the only tuning parameter of the prior and it controls the

modes of the NAP. As K increases, the modes get closer to p_0, and vice versa. This reflects the amount of difference from the null hypothesized value that we want to detect when H_1 is true. In many sensitive real-life applications a difference of ± 0.1 on the proportion scale can have substantial impact. So for default purposes, we suggest tuning K such that both the modes of the Beta moment prior are within $p_0 \pm 0.1$. Nonetheless, practitioners are recommended to appropriately tune K based on what they want.

Figure 4.2 provides example plots of the NAP for different null hypothesized values. For each null hypothesized value p_0, K is appropriately chosen to place both modes of the prior within $p_0 \pm 0.1$. For comparison, in the same figures we also plot the marginal density on proportion in the Logit-NAP, the Uniform and the Jeffreys priors. The Uniform and the Jeffreys are local default alternative priors that are often used to define the alternative hypothesis for this test. The



(a) $H_0 : p_0 = 0.2$ (b) $H_0 : p_0 = 0.5$ (c) $H_0 : p_0 = 0.8$

Figure 4.2: Beta moment prior densities and marginal densities on proportion in the Logit-NAP for one-sample proportion tests. Figures (4.2a)–(4.2c) are examples of NAPs that can be used to define the alternative hypothesis when the hypothesized values p_0 under the null are 0.2, 0.5 and 0.8, respectively. The blue dashed vertical lines denote $p_0 \pm 0.1$. The hyperparameters in each prior are chosen so that both the modes are within $p_0 \pm 0.1$. The hyperparameter values are respectively $K = 45, 50, 40$ and $\tau = .55/\sqrt{2}, .4/\sqrt{2}, .6/\sqrt{2}$. The Uniform and Jeffreys priors are also plotted for comparison.

NAPs assign about 0.01% prior probability between $p_0 \pm 0.01$ and 80–90% probability between

absolute differences 0.05 and 0.2. In comparison, the local alternative priors assign about $10\times$ to $20\times$ more prior probability between $p_0 \pm 0.01$. For tests conducted in the psychological sciences and other real-life applications with small to moderate sample sizes, and for which no specific prior information regarding the magnitude of standardized effect size is available, we recommend a default value of K such that the NAP modes are within $p_0 \pm 0.1$.

From computational standpoint, an advantage of the Beta moment prior density is that it results in closed form expressions for the Bayes factors in both one- and two-sided tests. We now define the specific assumptions used to perform the tests and provide explicit expressions for the resulting Bayes factors.

Bayes factor using the Beta moment prior. Suppose $y \sim$ Binomial (n, p). The Bayes factor of the test $H_1 : p \sim BM(p_0, K, \pi, m)$ versus $H_0 : p = p_0$ is given by

$$\text{BF}_{10}(y) = \frac{\text{P}(y, n-y)}{p_0^y (1-p_0)^{n-y}\, \text{P}(0,0)}, \tag{4.12}$$

where $\text{P}(a, b)$ is as in (4.11).

4.2.1.0.2 NAP in the Logit approach. Taking a Logistic regression perspective, we can perform one-sample proportion tests on the Logit scale. A NAP that can be used to define an alternative hypothesis is using the normal moment prior [88]. That is, first we define $\beta = \text{Logit}(p)$, $\beta_0 = \text{Logit}(p_0)$, and then assume

$$\beta \sim NM\left(\beta_0, \tau^2, m\right). \tag{4.13}$$

This denotes the normal moment prior density of order $m\,(\geq 1)$ given by

$$f_{NM}\left(x \,|\, \mu, \tau^2, m\right) = c_{NM} \left(\frac{(x-\mu)^2}{\tau^2}\right)^m \phi\left(x \,|\, \mu, \tau^2\right), \qquad \text{for } -\infty < x < \infty, \tag{4.14}$$

where $\phi\left(x \,|\, \mu, \tau^2\right)$ denotes the normal density function evaluated at x with mean μ and variance τ^2, and the proportionality constant $c_{NM} = \prod_{i=0}^{m-1}(1+2i)$. Henceforth, we refer to this specification as the Logit-NAP.

The density (4.14) has two parameters. The moment order m has the same interpretation as in (4.10) and we set it to 1 for default implementation. τ controls the modes of the density. Choosing this parameter is crucial as it reflects the $\log$-odds value, and in turn the proportion value, that we want to detect if the alternative is true. For a positive u, $\tau^2 = u^2/2m$ places the modes of $NM(\mu, \tau^2, m)$ at $\mu \pm u$. Figure 1 in [88] depicts the density when $\mu = 0$ and $u = 0.3$ for $m = 1$. Given u, the population proportions that the modes correspond to are

$$p_+ = \Big[1 + \exp\Big(- \beta_0 - u\Big)\Big]^{-1}, \quad \text{and} \quad p_- = \Big[1 + \exp\Big(- \beta_0 + u\Big)\Big]^{-1}. \tag{4.15}$$

The proportion p being of primary interest, we propose choosing u such that both the absolute differences $|p_+ - p_0|$ and $|p_- - p_0|$ does not exceed a desired value. Henceforth, we fix the desired difference to 0.1. This can be tuned according to the need of users and the problem at hand by choosing a different value than 0.1 or by directly choosing a desired u. Figure 4.2 provides example plots of the NAP for different null hypothesized values. For each null hypothesized value p_0, K is appropriately chosen to place both modes of the prior within $p_0 \pm 0.1$. For comparison, in the same figures we also plot the marginal density on proportion in the Logit-NAP, the Uniform and the Jeffreys priors. The Uniform and the Jeffreys are local default alternative priors that are often used to define the alternative hypothesis for this test. Figure 4.2 provides example plots of the Logit-NAP for different null hypothesized values. For each null hypothesized value p_0, τ is appropriately chosen to place both modes of the marginal density on proportion within $p_0 \pm 0.1$. For comparison, in the same figures we also plot the Beta moment, Uniform and Jeffreys priors.

4.2.1.0.3 Test-statistic based approach. Bayesian hypothesis tests are driven by the Bayes factors. Upon specification of prior beliefs in Null Hypothesis Significance Tests (NHSTs), it quantifies from data the odds of marginal evidence in favor of the alternative. This is also the factor by which prior odds of the hypotheses are updated to obtain the posterior odds. In NHSTs Bayes factors crucially depend on the prior specified under the alternative. More so, as the number of parameters increase, it gets more complex and computationally expensive to evaluate the marginal

densities. In a seminal work [89] proposed to defining Bayes factors by directly modeling distributions of test statistics. Being a pivotal quantity, its sampling distribution does not involve any unknown parameters when the null hypothesis is true. Under the alternative, the distribution becomes a "non-central" version of the null distribution, and introduces a non-centrality parameter that must be taken into account when modeling. In standard testing problems involving χ^2, F, t and Z test-statistics, the non-centrality parameter is much lower-dimensional than the actual parameter space.

We propose a similar approach for one-sample proportion tests. The Z-statistic for one-sample proportion tests is given by

$$Z_1 = \frac{\hat{p} - p_0}{\hat{\sigma}}, \quad \text{where } \hat{\sigma}^2 = \frac{\hat{p}(1-\hat{p})}{n}. \tag{4.16}$$

Here $\hat{p} = y/n$ is the sample proportion. A frequentist approach identifies that for a large sample size n, under the true null hypothesis Z_1 approximately follows the standard normal distribution. Under a true alternative hypothesis Z_1 approximately follows $N(\sqrt{n}\,(p-p_0)/\sqrt{p(1-p)}, 1)$ where p is the true and unknown proportion of interest. $\delta_1 = (p - p_0)/\sqrt{p(1-p)}$ is interpreted as the standardized effect size and is the only unknown parameter in the model. We propose a test-statistic based approach for one-sample proportion tests and model the Z-statistic under the alternative hypothesis using the normal moment prior density as in (4.14) [88]. That is, for testing $H_0 : p = p_0$ against $H_1 : p \neq p_0$, we first compute the Z-statistic from the available data and model it under each hypothesis as

$$Z_1 \sim N(0, 1), \quad \text{under } H_0, \tag{4.17}$$

$$Z_1 \sim NM(0, \tau^2, m), \quad \text{under } H_1. \tag{4.18}$$

Henceforth, we refer to this approach as the Z-NAP. As before, we set m to 1 for default implementation and τ controls the modes of the density. Choosing this parameter is crucial as it reflects the proportion value that we want to detect under true alternative hypotheses. Precisely, $\tau^2 = u^2/2m$

places the mode of $NM(0, \tau^2, m)$ at $\pm u$. For choosing τ, we use the large sample approximation of the sampling density of Z_1 and propose $u = \sqrt{n}\,\delta_1^*$. δ_1^* reflects the standardized effect size we want to detect when the alternative is true. A value of δ_1 corresponds to two population proportions and they are the roots of the quadratic equation

$$ap^2 - bp + c = 0, \tag{4.19}$$

where

$$a = 1 + \delta_1^2, \quad b = 2p_0 + \delta_1^2, \quad c = p_0^2. \tag{4.20}$$

Denote by $p_+(\delta_1)$ and $p_-(\delta_1)$ the proportions δ_1 corresponds to. Since the proportion p is of primary interest, we propose choosing δ_1^* such that the absolute differences $|p_+(\delta_1^*) - p_0|$ and $|p_-(\delta_1^*) - p_0|$ does not exceed a desired value. Henceforth, we fix this desired difference on the proportion scale to 0.1 and tune δ_1^* accordingly. Nonetheless, this can be tuned according to the need of users and the problems at hand.

From a computation standpoint, in the test-statistic based approach we only need to calculate the likelihood ratio between the two models under respective hypotheses. Below we provide specific assumption and explicit expression of the resulting likelihood ratio.

Likelihood Ratio based on Z-statistic. Suppose $y \sim$ Binomial (n, p) and define the test-statistic Z_1 as in (4.16). The likelihood ratio of $Z_1 \sim NM(0, \tau^2, m)$ under H_1 versus $Z_1 \sim N(0, 1)$ under H_0 is given by

$$\mathrm{LR}_{10}(Z_1) = |Z_1|^{2m}\ \tau^{-(2m+1)}\ \exp\left(rZ_1^2/2\right), \tag{4.21}$$

where $r = 1 - 1/\tau^2$.

4.2.2 Two-sample Proportion Tests

Consider the notations as in (4.1). Two-sample proportion tests compare proportions of interest from two independent populations and test for their equality. While (4.2) and (4.4) tests for it on

the scale of proportion, (4.7) tests for the same null hypothesis on the logit scale. Particularly, we propose the Diff-NAP and the Logit-NAP. In Diff-NAP we specify priors on the proportion scale while the Logit-NAP specifies prior on the logit scale. Below we describe the use of NAP densities in these approaches.

4.2.2.0.1 NAP in the Difference approach. The Diff approach uses the parameterization (4.3) and tests $H_0 : \eta = 0$ against $H_1 : \eta \neq 0$ to test for the equality of the two proportions. For testing $H_0 : \eta = \eta_0$ versus $H_1 : \eta \neq \eta_0$, a NAP that can be used to define an alternative hypothesis is as follows:

$$\frac{\eta + 1}{2} \sim BM\left(\frac{\eta_0 + 1}{2}, K, \pi, m\right), \tag{4.22}$$

$$\zeta \,|\, \eta \sim \text{Unif}\left(\frac{|\eta|}{2}, 1 - \frac{|\eta|}{2}\right), \tag{4.23}$$

where Unif(a, b) denotes the Uniform distribution in the interval (a, b). Henceforth, we refer to this specification as the Diff-NAP. (4.22)–(4.23) specifies a joint prior density on the parameters (ζ, η) which induces a joint prior on (p_1, p_2). The Beta moment prior density in (4.22) evaluates to 0 at η_0, the null hypothesized value. This feature translates to the joint prior on (p_1, p_2) and the prior density equals to 0 at the diagonal $p_1 = p_2$, the proportion pairs consistent with the null hypothesis. This induces a non-locality around the diagonal on the proportion scale.

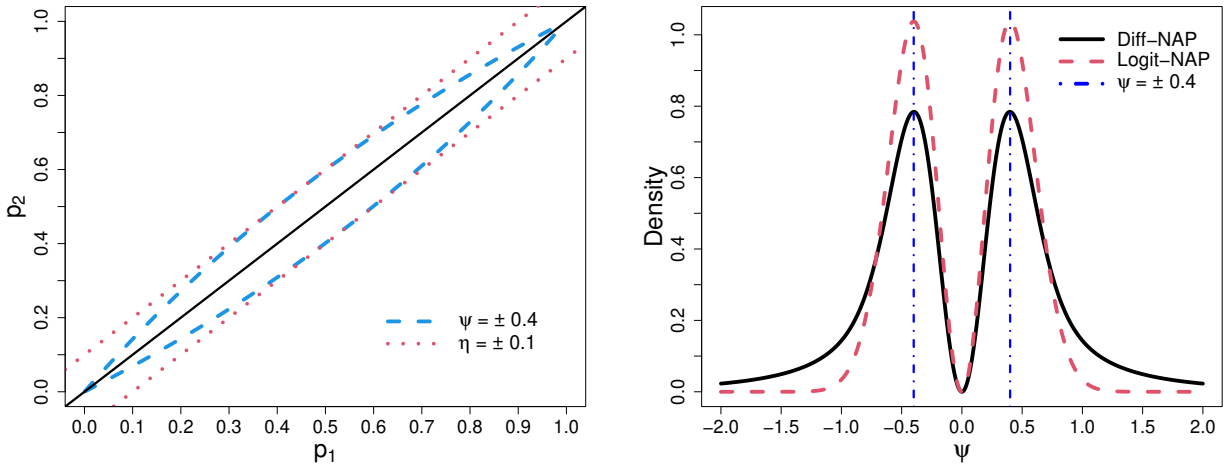
As in one-sample tests, we need to choose K, π and m. Note that the interpretations of the parameters are the same as in Section 4.2.1. We recommend $\pi = (\eta_0 + 1)/2$ and $m = 1$ for default implementations. Then (4.22) only has the tuning parameter K. It controls the mode of the prior on η, the proportion difference of interest. As K increases, the modes are placed closer to η_0, and vice versa. Like in Section 4.1 the marginal density under the alternative, and hence the Bayes factor, is not available in closed form and it needs to be calculated numerically. The choice of K for default implementation is discussed below.

4.2.2.0.2 NAP in the Logit approach. The Logit approach transforms the proportions on the Logit scale and tests $H_0 : \psi = 0$ against $H_1 : \psi \neq 0$ for testing the equality of proportions. We

propose the use of normal moment prior on ψ as the NAP to define the alternative hypothesis for this test [88]; that is, under the alternative we assume

$$\psi \sim NM\left(0, \tau^2, m\right). \tag{4.24}$$

The same priors on β as in Section 4.1 is assumed under both hypotheses. Henceforth, we refer to this specification as the Logit-NAP. m is set to 1 for default purposes. τ controls the mode of the prior on ψ. Larger the value of τ, the closer the modes are to 0, and vice versa. The modes reflect the difference in log-odds that we want to detect when the null hypothesis is not true. Like in Section 4.1 the marginal density under the alternative, and hence the Bayes factor, is not available in closed form and it needs to be calculated numerically. The choice of τ for default implementation is discussed below.



(a) Approximately matching ψ^* with η^*

(b) Marginal prior on log-odds

Figure 4.3: On the left, Figure 4.3a shows the proportion pairs (p_1, p_2) satisfying a pre-specified difference of $\eta = \pm 0.1$ and a pre-specified log-odds of $\psi = \pm 0.4$. The solid black like denotes the proportion pairs consistent with the null hypotheses, which are the same in both Diff-NAP and Logit-NAP for two-sample proportion tests.

4.2.2.0.3 Choosing hyperparameters in Diff-NAP and Logit-NAP. There are multiple ways one can set the hyperparameters K is Diff-NAP and τ is Logit-NAP for default implementations. We note the fundamental differences between specifying priors on the proportion scale and on the Logit scale. The proportions are of primary interest and suppose we aim to detect an absolute difference of η^* on the proportion scale. To make the two approaches comparable, we tune the target log-odds ψ^* under alternative such that the maximum absolute difference between all proportion pairs satisfying the log-odds ψ^* is approximately η^*. For default implementations we suggest $\eta^* = 0.1$ which implies $\psi^* \approx 0.4$. This is presented in Figure 4.3a. Then we choose the hyperparameters K and τ such that the modes of the marginal priors on the log-odds in the two approaches are at ± 0.4. This corresponds to $\tau = 0.4/\sqrt{2}$ in the Logit-NAP and $K \approx 280$ in the Diff-NAP. Figure 4.3b shows the marginal priors on the log-odds and Figure 4.4 shows the joint prior on the proportions (p_1, p_2) corresponding to the default choices in the two approaches. Nonetheless, η^* and δ^* can be varied according to the need of users and the problems at hand.

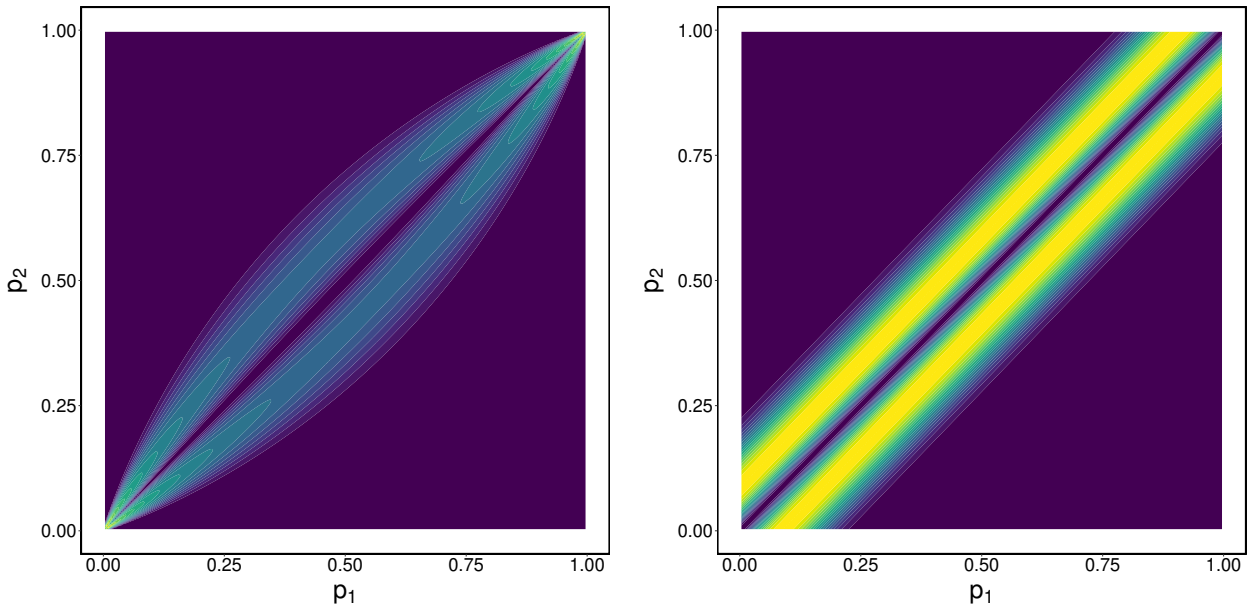
4.2.2.0.4 Test-statistic based approach. In two-sample proportion tests, there are two parameters in the model, p_1 and p_2. For testing in the Bayesian way, we accordingly need to specify priors on them under each hypothesis. This can get somewhat complicated as we have seen from the Logit and the Diff approach both using local priors and NAP. Like in one-sample tests here we propose an approach based on the Z-statistic that is often and widely used to conduct large sample two-sample proportion tests. The test-statistic is given by

$$Z_2 = \frac{\hat{p}_1 - \hat{p}_2}{\hat{\sigma}_P}, \quad \text{or} \quad Z_2 = \frac{\hat{p}_1 - \hat{p}_2}{\hat{\sigma}_{UP}}. \tag{4.25}$$

$\hat{\sigma}_{UP}$ and $\hat{\sigma}_P$ are respectively unpooled and pooled estimate of the standard error of $\hat{p}_1 - \hat{p}_2$. They are defined as

$$\hat{\sigma}^2_{UP} = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}, \quad \text{and} \tag{4.26}$$

$$\hat{\sigma}^2_P = \hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right), \tag{4.27}$$



(a) Logit-NAP

(b) Diff-NAP

Figure 4.4: The default joint NAP prior assigned to (p_1, p_2) for two-sided two-sample proportion tests. Figure (a) on the left panel corresponds to the prior in the Logit-NAP, and Figure (b) on the right panel corresponds to the prior in the Diff-NAP. The brighter the color, the higher is the prior density there. The hyperparameters are $\tau = 0.4/\sqrt{2}$ in the Logit-NAP and $K = 280$ in the Diff-NAP. These default values are chosen so that the modes of the marginal prior on the log-odds are at ± 0.4. Following Figure 4.3a this implies a maximum difference of ± 0.1 on the proportion scale.

where $\hat{p} = (n_1\hat{p}_1 + n_2\hat{p}_2)/(n_1 + n_2)$ is the pooled estimate of the population proportions. For large sample two-sample proportion tests a frequentist approach assumes that the group sizes n_1 and n_2 are increasing such that their ratio $n_1/n_2 \to c$. c is a positive fraction and indicates the balance of information coming from two groups. Under this assumption, Z_2 approximately follows the standard normal distribution under the true null hypothesis. Alternatively, when the null is not true, Z_2 approximately follows $N(\sqrt{n}_1\, \delta_2, \sigma^2)$. δ_2 is interpreted as the standardized effect size. δ_2 and σ^2 are slightly different depending on whether we use the unpooled or pooled estimate of standard error when calculating the test-statistic. $\delta_2 = (p_1 - p_2)/\sqrt{p_1(1-p_1) + c\, p_2(1-p_2)}$ if the unpooled estimate is used. If the pooled estimate is used, it equals to $(p_1 - p_2)/\sqrt{p(1-p)(1+c)}$ where $p = (c\, p_1 + p_2)/(1 + c)$. With this in the background, we propose a test-statistic based

approach for two-sample proportion tests and model the Z statistic under the alternative hypothesis using the normal moment prior density as in (4.14) [88]. To this, we first compute the Z-statistic from the available data and model it under each hypothesis as

$$Z_2 \sim N(0,1), \quad \text{under } H_0, \tag{4.28}$$

$$Z_2 \sim NM(0,\tau^2,m), \quad \text{under } H_1. \tag{4.29}$$

Henceforth, we refer to this specification as the Z-NAP. As before, we set m to 1 for default implementation and τ controls the modes of the density. Recall that $\tau^2 = u^2/2m$ places the mode of $NM(0,\tau^2,m)$ at $\pm u$. For choosing τ, we use the large sample approximation of the sampling density of Z and propose $u = \sqrt{n}_1\,\delta_2^*$. δ_2^* reflects the standardized effect size that we want to detect when the null is not true. Given δ_2, the population proportions p_1 and p_2 that it corresponds are given by the equation

$$\frac{(p_1-p_2)^2}{p_1(1-p_1)+c\,p_2(1-p_2)} = \delta_2, \quad \text{if the unpooled estimate is used, or} \tag{4.30}$$

$$\frac{(p_1-p_2)^2}{p(1-p)(1+c)} = \delta_2, \quad \text{if the pooled estimate is used.} \tag{4.31}$$

Let $(p_1(\delta_2), p_2(\delta_2))$ denotes all the proportion pairs that δ_2 corresponds to. Since the proportions p_1 and p_2 are of primary interest, we propose choosing δ_2^* such that the maximum of the absolute difference $|p_1(\delta_2^*) - p_2(\delta_2^*)|$ among all the choices of proportion pairs does not exceed a desired value. Henceforth, we fix this maximum allowed difference on the proportion scale to 0.1 and tune δ_2^* accordingly. Nonetheless, this can be varied according to the need of users and the problems at hand.

From a computation standpoint, in the test-statistic based approach we only need to calculate the likelihood ratio between the two models under respective hypotheses. This provides a significant computational advantage compared to the Diff-NAP and Logit-NAP where the marginals under the alternative are not available in closed forms and need to be calculated numerically. Below

we provide specific assumption and explicit expression of the resulting likelihood ratio.

Likelihood Ratio based on Z-statistic. Suppose y_1 and y_2 are independent samples from Binomial (n_1, p_1) and Binomial (n_2, p_2), respectively. Define the test-statistic Z_2 as in (4.25). The likelihood ratio of $Z_2 \sim NM(0, \tau^2, m)$ under H_1 versus $Z_2 \sim N(0, 1)$ under H_0 is the same as in (4.21) for one-sample tests and is given by

$$\mathrm{LR}_{10}(Z_2) = |Z_2|^{2m} \ \tau^{-(2m+1)} \ \exp\left(rZ_2^2/2\right), \tag{4.32}$$

where $r = 1 - 1/\tau^2$.

4.3 Weight of Evidence comparison in Fixed design tests

Classical tests of a population proportion parameter are either based on exact tests or large sample approximate tests based on Z-statistics. These tests are designed to control Type I (α) and Type II (β) error probabilities at prespecified levels. A key disadvantage of these tests is that they do not quantify evidence in favor of true null hypotheses. Instead, they may simply "fail to reject" the null hypothesis. Psychology and other social science researchers often have a need to quantify evidence in favor of true null hypotheses [for example, 1]. Bayes factors provide such a measure.

To summarize the performance of various Bayesian tests, we adopt the measurement scale for evidence based on the natural logarithm of the Bayes factors, $\ln(\mathrm{BF}_{10})$. This quantity, called the "weight of evidence", has the advantage of being on the same scale as the classical likelihood ratio statistic [19, 24].[1] Because $-\ln(x) = \ln(1/x)$, the weight of evidence in favor of the alternative hypothesis is equal to the negative of evidence in favor of the null hypothesis (and vice versa). Descriptors for the weight of evidence were proposed by [19] and [24]. Under the former, weight of evidence between 0 and 1 in magnitude is considered "not worth more than a bare mention"; weight of evidence between 1 and 3 is considered "positive"; weight of evidence between 3 and 5 is "strong", and above 5 is labeled as "very strong". At the border between positive and strong

[1] [19] propose $2\ln(\mathrm{BF}_{10}(\mathbf{x}))$ as a default measure, but by omitting the factor of 2 their descriptors are more compatible with the measure proposed by [24].

(3), the corresponding Bayes factor is about 20, and at the border between strong and very strong, the Bayes factor is about 150. Strong and very strong weights of evidence in favor of the null hypothesis are -3 and -5, or Bayes factors of approximately $1/20$ and $1/150$.

Bayes factors must be multiplied by the prior odds that the null hypothesis is true to determine the posterior odds. If the prior odds are 1 (that is, $\mathbf{P}(H_0) = \mathbf{P}(H_1) = 0.5$), then weight of evidence equal to 3 implies a Bayes factor and posterior odds of about 20, and posterior probability of the alternative hypothesis equal to 0.95. Similarly, weight of evidence of -5 implies a Bayes factor and posterior odds of about 1/150, and posterior probability of the null hypothesis equal to $1 - 0.0066 = 0.9934$. This probability is very close to 1.0, but it is predicated on the assumption that the prior odds are 1.0.

Recent evidence from replication of experiments in psychology and social sciences suggest that the prior probability of a null hypothesis examined in these fields is likely between 0.80–0.95 [8, 9, 23]. If $\mathbf{P}(H_0) = 0.9$, then weight of evidence equal to 3 implies that the posterior probability of the alternative hypothesis is only 0.69, while weight of evidence equal to 5 implies that the posterior probability of the alternative hypothesis is 0.94. With this background in place, for two-sided, one- and two-sample proportion tests we now compare the average weight of evidences from different approaches described in Section 4.2.1 and 4.2.2 respectively hold for one- and two-sample proportion tests.

4.3.1 One-sample Proportion Tests

We assume the conditions mentioned in Section 4.2.1 hold. Figure 4.5 displays the average weight of evidence obtained under different approaches when the null hypothesis $H_0 : p = 0.2$ is true. These curves were based on simulating one-million random samples at each sample size. The alternative hypotheses considered in this plot include the following:

1. Beta moment prior with $K = 45$,

2. Logit-NAP with $\tau^2 = (0.55^2)/2$ (modes at ± 0.55),

3. Under alternative Z_1 is modeled as $NM(0, \tau^2, 1)$ with $\tau^2 = (0.12)^2\, n/2$ (modes at $\pm 0.12\,\sqrt{n}$),

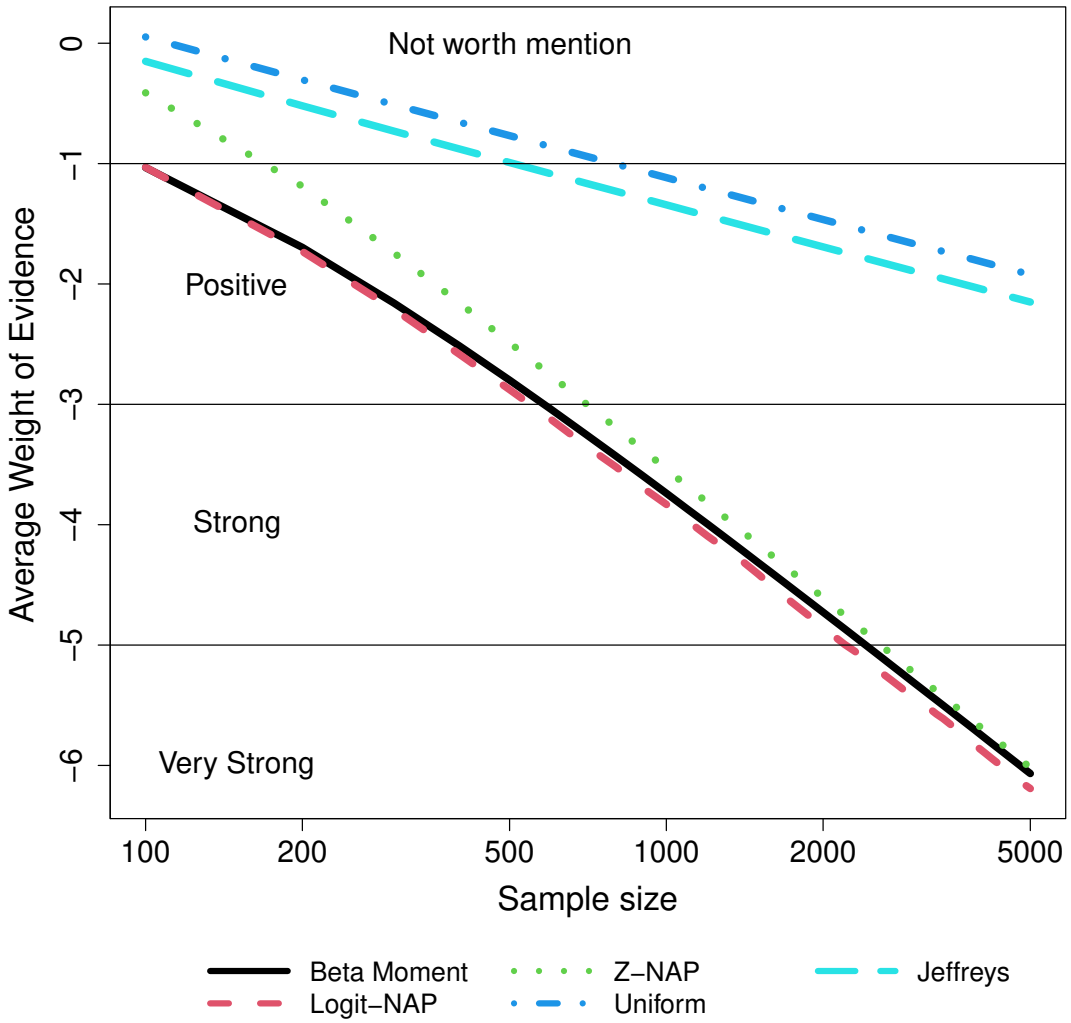


Figure 4.5: Average weight of evidence in two-sided one-sample proportion tests of $H_0 : p = 0.2$ against alternative hypotheses when the null hypothesis is true. *The horizontal axis is displayed on the logarithmic scale because of the large differences in samples sizes required by the different methods to obtain, on average, strong or very strong weight of evidence against each alternative hypothesis.*

4. The Uniform prior on $(0, 1)$

5. The Jeffreys prior, that is, Beta distribution with both shape parameters 0.5.

4.3.1.0.1 True Null Hypothesis. Figure 4.5 illustrates a critical deficiency of the local priors: The use of such priors to define the alternative hypothesis makes it difficult to obtain "very strong"

weight of evidence in favor of a true null hypothesis. The Uniform prior requires about 2.4 million subjects, on average, to obtain very strong weight of evidence in favor of a true null hypothesis, and the Jeffreys prior requires about 1.5 million subjects. In contrast, the NAP approaches, namely the Beta moment prior, Logit-NAP and Z-NAP, require about 2500 subjects, on average, for the same purpose.

Obtaining even strong weight of evidence in favor of a true null hypothesis is difficult when standard Uniform and Jeffreys priors are used to define the alternative hypothesis. On average, 50,000 subjects are required to obtain strong weight of evidence when the Uniform prior is used, and on average 30,000 subjects are needed when the Jeffreys prior is used to define the alternative hypothesis. In contrast, the Beta moment prior and the Logit-NAP require about 600 subjects, and 700 subjects if the prior mode is set to 0.5.

4.3.1.0.2 True Alternative Hypothesis. Here we discuss the cost that NAP approaches pay to detect true null hypotheses. Figure 4.6 shows the average weights of evidence obtained under these prior specifications for a range of sample sizes in fixed-design tests as a function of the true proportions. For a sample size of 200, the Uniform and Jefferys prior achieve strong or very strong weight of evidence in favor of the alternative hypothesis than the NAP approaches do when the $p \lessapprox 0.14$ ($\lessapprox$ stands for approximately less than or equal to. $\gtrapprox$ is similarly interpreted.). When $p \gtrapprox 0.27$, all the approaches except Z-NAP produce strong to very strong evidence. As sample size increases, the Z-NAP produces strong and very strong weight of evidence than the other approaches when $p \lessapprox 0.14$. For sample sizes 400 or larger, all the methods achieve similar weight of evidence when $p \gtrapprox 0.27$.

Local priors, namely the Uniform and Jeffreys priors, provide more support for proportions very close to 0.2. However, strong evidence in favor of these proportions can only be obtained with very large sample sizes. When the sample size is 500 and the proportion is within ± 0.02, all the Bayes factors in Figure 4.7 yield average weights of evidence that are negative, thus favoring the null hypothesis of no effect. Indeed, for proportions within the range of ± 0.02 around the null, use of the NAP approaches provide, on average, "positive" support for the null hypothesis.

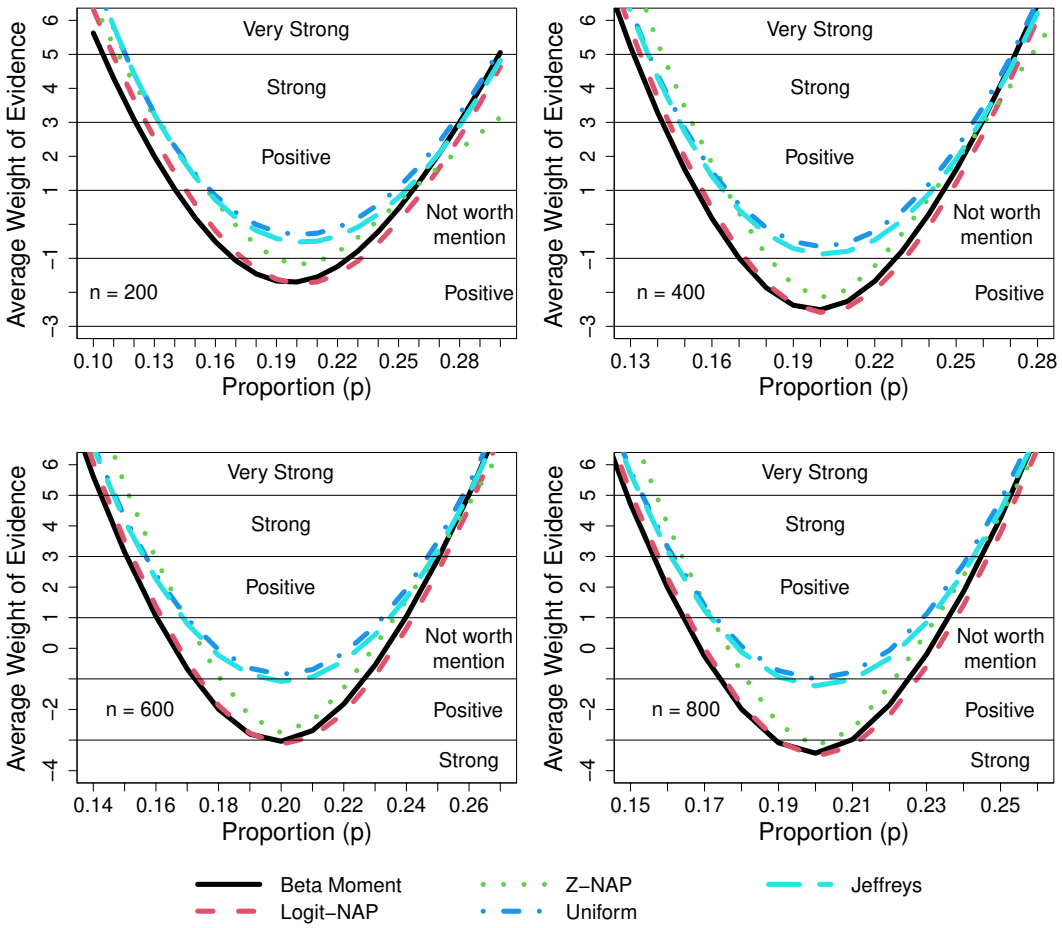


Figure 4.6: Average weight of evidence in two-sided one-sample proportion tests of $H_0 : p = 0.2$ for true alternative hypotheses. Curves depicted in the plots denote the average weight of evidence versus true population proportion when different local and NAP approaches are used.

This misleading performance of the NAP priors for true proportions within the range of ± 0.01 around the null persists, and even degrades, for sample sizes up to 4,000. When the sample size is 2,000, the NAP approaches and the JZS priors begin to show positive support (i.e., $\ln(\text{BF}_{10}) > 1$) for proportions outside the range of ± 0.02. None of the NAP models depicted here provide, on average, strong support for the alternative hypothesis for any standardized effect size less than 0.1. On the other hand, the local priors do attain for proportions outside the range of ± 0.025. If the sample size is increased to 4,000, then the Uniform and Jeffreys priors provide, on average, strong evidence for proportions outside the range of ± 0.02, and positive evidence for proportions outside

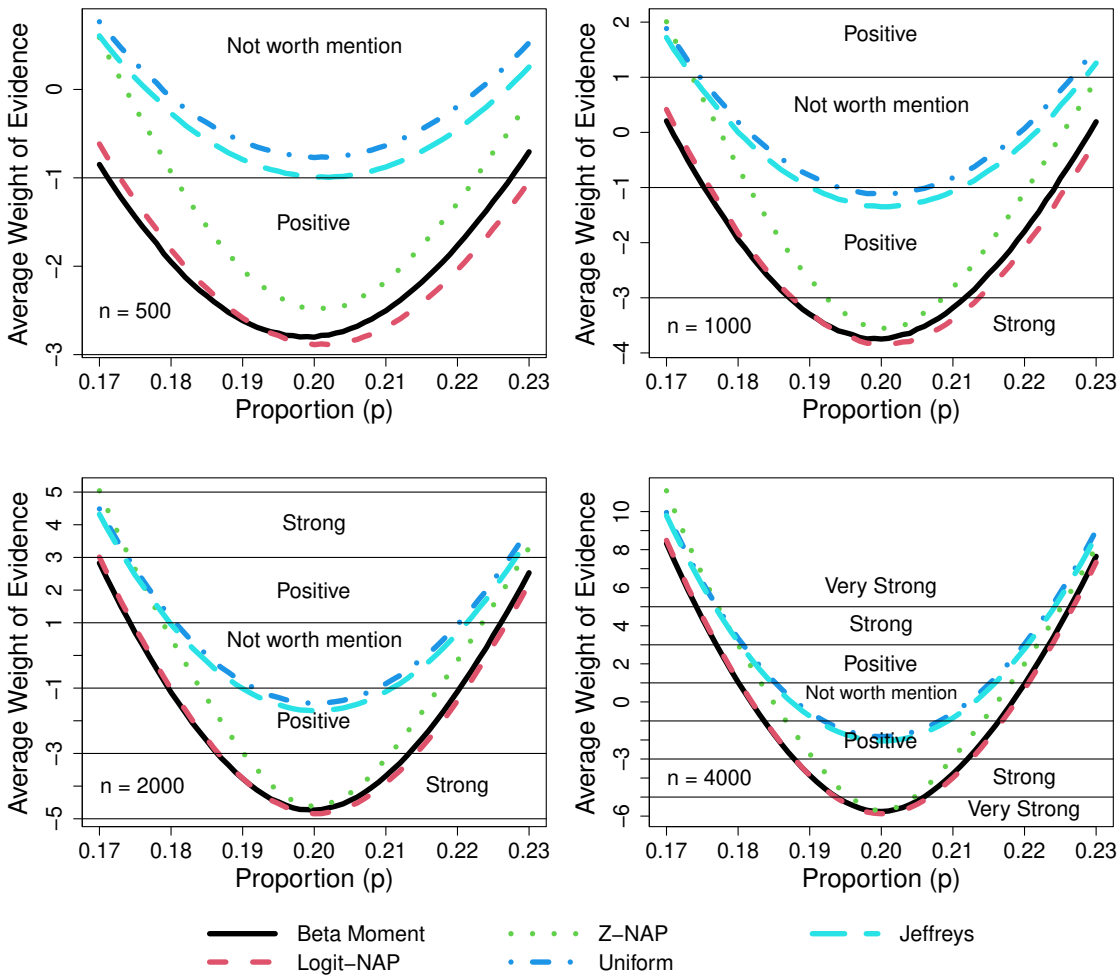


Figure 4.7: Weight of evidence for true alternative hypotheses with proportions ± 0.03 around the null $H_0 : p = 0.2$. Curves depicted in the plots denote the average weight of evidence versus true proportions for different approaches.

the range of ± 0.015. For NAP approaches, the proportions need to be outside the range of ± 0.02 to attain positive evidence and outside ± 0.025 to attain strong evidence.

The conclusions from Figures 4.5–4.7 might be simply stated as follows. NAP approaches can provide strong or very strong weight of evidence in favor of true null hypotheses for small or moderate sample sizes (i.e., $n \approx 700$). In many practical settings (i.e., $n < 2000$), the local priors cannot. For proportions outside a range of about ± 0.05 around the null, all the approaches on average achieve strong evidence for sample sizes greater than 600. Alternative hypotheses defined

with local priors provide higher average weight of evidence for proportions very close to the null (i.e., within a range of about ± 0.02), but require large sample sizes ($\gtrapprox 2000$ to provide positive support and $\gtrapprox 4000$ to provide strong support).

4.3.2 Two-sample Proportion Tests

We assume the conditions mentioned in Section 4.2.2 hold. Figure 4.8 displays the average weight of evidence obtained under different approaches when the null hypothesis $H_0 : p_1 = p_2$ is true. We investigate the impact of common population proportion and vary it from small to moderate as 0.1, 0.2, 0.3, and 0.5. These curves were based on simulating $10,000$ random samples at each sample size. The different alternative hypotheses and approaches considered in this plot include the following:

1. IB approach with Beta$(1, 1)$ prior,
2. Logit-Local with Unif$(0, 1)$ prior on β and $N(0, 1)$ prior on ψ,
3. Diff-Local with $N(0.5, 0.5^2)$ prior on ζ and $N(0, 0.2^2)$ prior on η,
4. Diff-NAP with $K = 280$,
5. Logit-NAP with Unif$(0, 1)$ prior on β and $NM(0, \tau^2, 1)$ prior on ψ with $\tau^2 = (0.4)^2/2$ (modes of the NAP at ± 0.4),
6. Under the alternative, the un-pooled Z_2 is modeled as $NM(0, \tau^2, 1)$ with $\tau^2 = (0.15)^2\, n/2$ (modes of the NAP at $\pm 0.15\, \sqrt{n}$).

4.3.2.0.1 True Null Hypothesis. As in one-sample tests, Figure 4.8 illustrates a similar drawback of the local priors: The use of such priors to define the alternative hypothesis makes it difficult to obtain "very strong" weight of evidence in favor of a true null hypothesis. When the common proportion is 0.1, the IB and Diff-Local requires about 69,000 subjects, on average, to obtain very strong weight of evidence in favor of a true null hypothesis. The Logit-Local requires about 1.4 million subjects for the same. In contrast, the NAP approaches Diff-NAP, Logit-NAP and Z-NAP

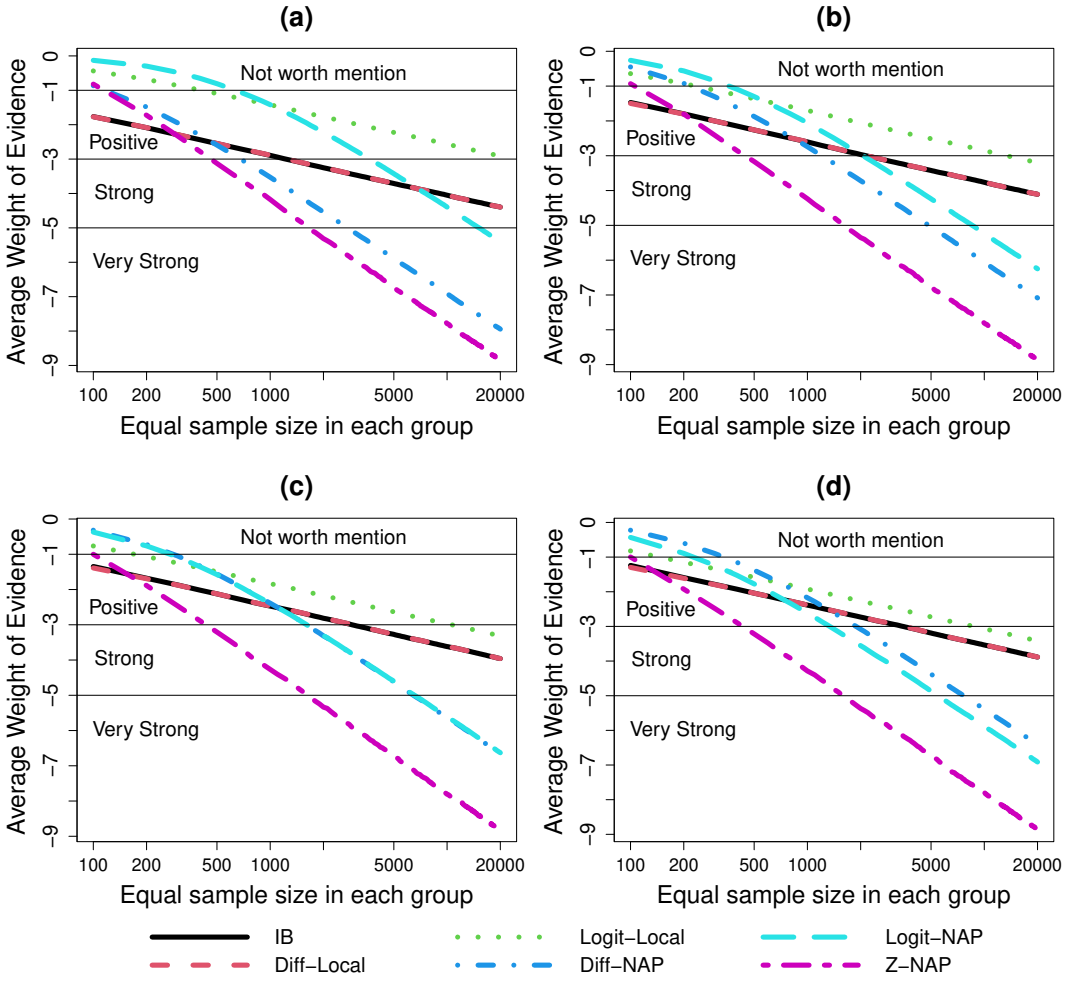


Figure 4.8: Average weight of evidence in two-sided two-sample proportion tests of $H_0 : p_1 = p_2$ against alternative hypotheses when the null hypothesis is true. Figures (a)–(d) respectively assumes common proportions 0.1, 0.2, 0.3, and 0.5. *The horizontal axis is displayed on the logarithmic scale because of the large differences in samples sizes required by the different methods to obtain, on average, strong or very strong weight of evidence against each alternative hypothesis.*

respectively require about 2800, 16000, and 1700 subjects, on average, for the same purpose. This shows a very clear performance gap between the test-statistic based approach and the other approaches. In fact, as the common proportion increases, the gap only gets wider. The Logit-Local accumulates evidence a little faster while the other approaches requires more samples to attain very strong evidence in favor of the null. Nonetheless, the NAP approaches still provides a significantly smaller number of samples as compared to the local priors for this. When the common proportion

is 0.3, the IB and Diff-Local each requires about 160,000 subjects to attain very strong evidence. The Logit-Local requires about 60,000 subjects for the same. Compared to this, the Diff-NAP and Logit-NAP each requires 7000 samples while the Z-NAP requiring only 1700 subjects.

In terms of achieving very strong evidence in favor of the null, there seems to be a consensus within the NAP approaches and the local priors, in the sense, the NAP approaches require smaller number of samples to achieve very strong evidence, on average, than the local priors. But this does not hold if we only want to obtain strong evidence. In this case, the crucial difference between specifying priors on the proportion scale and the Logit scale becomes prominent. This is more prominent when the common population is very small. For example, when the common population proportion is 0.1, Figure 4.8(a) shows that the IB approach requires lesser number of samples compared to the Logit-NAP to achieve strong evidence. If we compare within each type of specifying priors, Logit-NAP requires lesser samples than the Logit-Local and Diff-NAP needs smaller samples than the IB. For example, when the common population proportion is 0.1, the IB and the Diff-Local requires about 1300 samples to attain strong evidence as compared to 700 samples by the Diff-NAP. On the other hand, the Logit-Local requires about 25,000 samples as compared to 3700 samples by the Logit-NAP. Compared to all these approaches, Z-NAP does better than all and requires only about 500 sample for the same. As the common proportion increases (Please see Figure 4.8(b)–(d)), the difference between the two types of prior specifications reduces. For example, when the common population proportion is 0.3, the IB and the Diff-Local requires about 3000 samples to attain strong evidence as compared to 1700 samples by the Diff-NAP. On the other hand, the Logit-Local requires about 10,600 samples as compared to 1700 samples by the Logit-NAP. Compared to all these approaches, Z-NAP does better than all and requires only about 500 sample for the same.

4.3.2.0.2 True Alternative Hypothesis. Here we discuss the cost that NAP approaches pay to detect true null hypotheses. Figures 4.9–4.11 respectively correspond to $p_1 = 0.1, 0.2,$ and 0.5. Each figure shows the average weight of evidence obtained by the approaches for a range of sample sizes in fixed-design tests as a function of p_2 varied within $(p_1, p_1 + 0.1)$. For a prefixed $p_1 = 0.1$

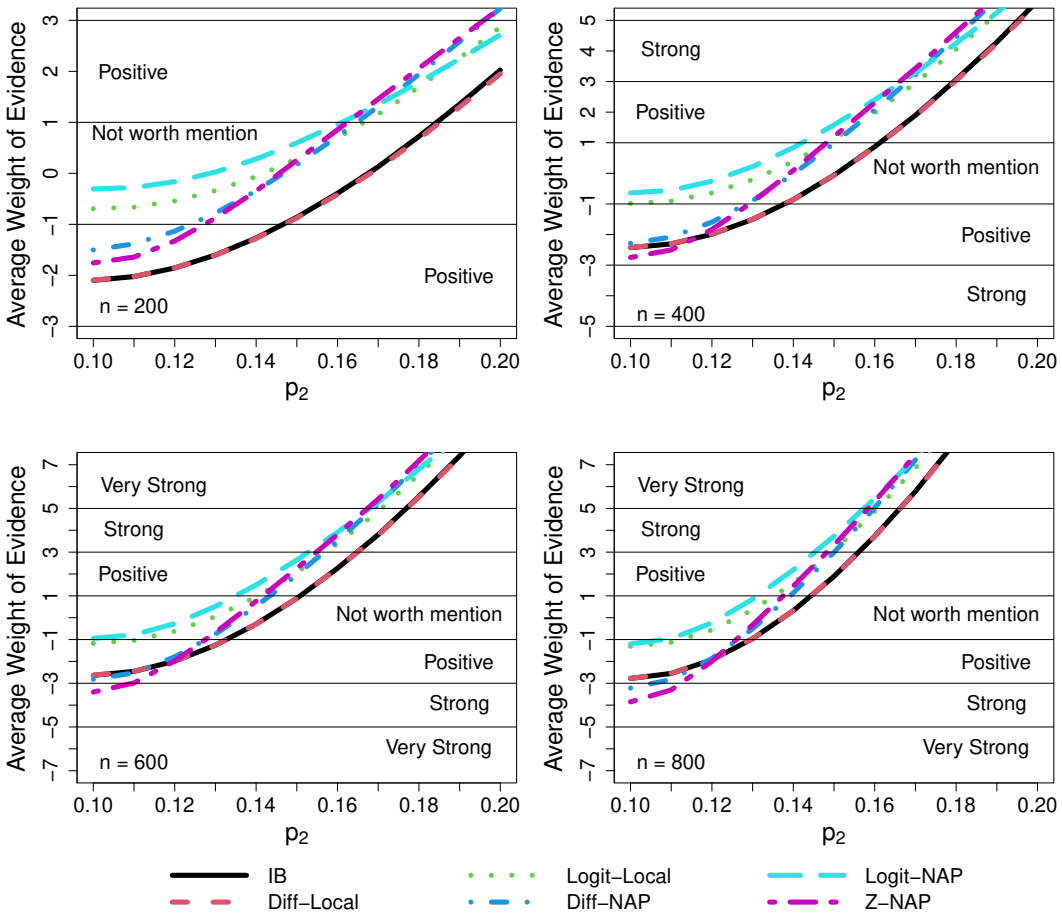


Figure 4.9: Average weight of evidence in two-sided two-sample proportion tests of $H_0 : p_1 = p_2$ for true alternative hypotheses. For a prefixed $p_1 = 0.1$ curves depicted in the plots denote the average weight of evidence versus true population proportion p_2 varied within $(p_1, p_1 + 0.1)$ when different approaches are used.

and 200 samples, Figures 4.9 shows that the IB and Diff-Local consistently produce lesser weight of evidence than others. On the other hand, the Logit-Local and Logit-NAP achieves positive evidence for p_2 larger than about 0.16 but the evidence they accumulate for p_2 lesser than 0.16 is not worth a mention. Compared to this, even with only 200 samples the Diff-NAP and Z-NAP is able to achieve a positive evidence both for proportions within $(0.1, 0.12)$ and larger than about 0.16. As sample size increases to 800, all the approaches except the IB and the Diff-Local achieves very strong weight of evidence for p_2 larger than about 0.16 and strong weight of evidence larger

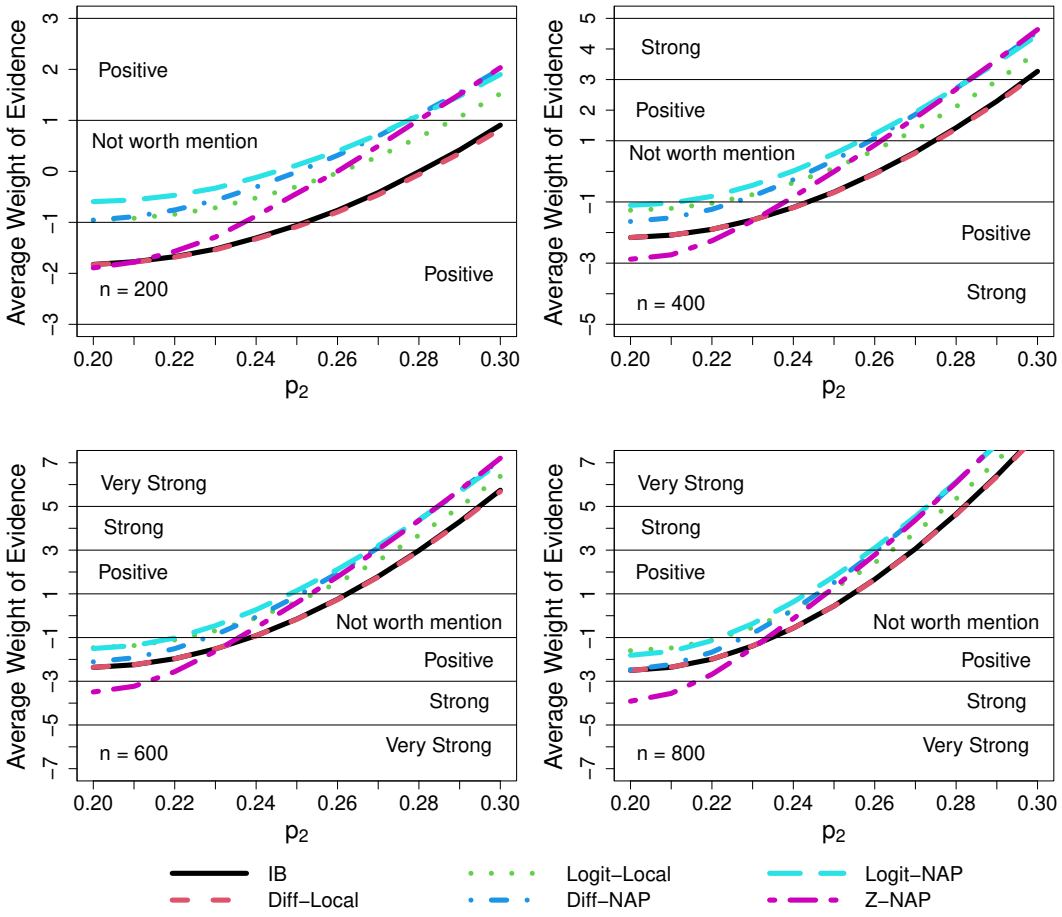


Figure 4.10: Average weight of evidence in two-sided two-sample proportion tests of $H_0 : p_1 = p_2$ for true alternative hypotheses. For a prefixed $p_1 = 0.2$ curves depicted in the plots denote the average weight of evidence versus true population proportion p_2 varied within $(p_1, p_1 + 0.1)$ when different approaches are used.

than about 0.15. On the other end near 0.1, the null, the Diff-NAP and the Z-NAP accumulates evidence faster with increase in sample size and achieves strong evidence in favor of the null. The proportion values being very small, this again highlights the difference in evidence accumulation between specifying priors on the logit scale and the proportion scale, and modeling of the test-statistic.

As the prefixed p_1 increases to 0.5, an agreement within the NAP and the local priors shows up. For a sample size of 200, only IB and Diff-Local achieves positive evidence near the null

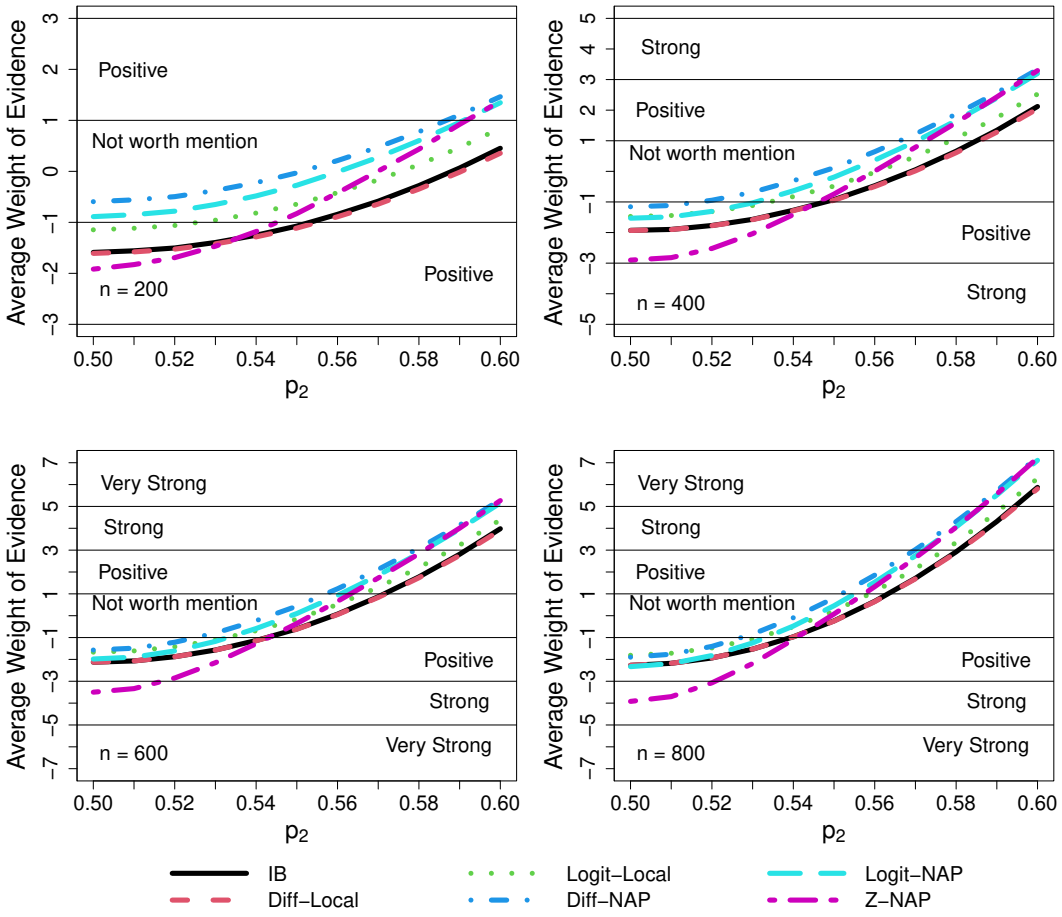


Figure 4.11: Average weight of evidence in two-sided two-sample proportion tests of $H_0 : p_1 = p_2$ for true alternative hypotheses. For a prefixed $p_1 = 0.5$ curves depicted in the plots denote the average weight of evidence versus true population proportion p_2 varied within $(p_1, p_1 + 0.1)$ when different approaches are used.

while Logit-Local barely achieves so, and Logit-NAP and Diff-NAP achieve evidence that are not worth mention. For p_2 near 0.6, all the NAP approaches achieves positive evidence while the local priors fall short. As the sample size increases to 800 all the methods except the Z-NAP achieves positive evidence. Compared to others, the Z-NAP performs more consistently. For sample size 200 it achieves the highest evidence in favor of the null and, as other NAP approaches, achieves positive evidence for p_2 larger than 0.59. When sample size increases to 800, the Z-NAP is the only approach that achieves strong evidence in favor of the null and also achieves higher evidence

over the local priors and is as good as the other NAPs.

The conclusions from Figures 4.8–4.11 can be summarized as follows. Whether under true null or true alternative hypothesis, performance of Logit based approaches and priors specified on proportion scale is sensitive to the actual value of proportions. This performance difference less prominent if sufficient sample size is provided and a decision is reached after attaining very strong evidence. In such cases, the NAP approaches can provide strong or very strong weight of evidence in favor of true null and alternative hypotheses. In many practical settings, the local priors cannot. The difference between the approaches becomes more vivid when both proportions are small or large and decision is reached after strong evidence is attained. Remarkably, the Z-NAP is almost agnostic to this issue. Since it is based on the test-statistic, we only need to specify a prior on the single non-centrality parameter under the alternative. This makes the method simple to specifying priors. If a prior is correctly specified, this also makes it less sensitive to the chosen prior.

4.4 An Application to the *New England Journal of Medicine* studies

To illustrate the use of NAP-based Bayes factors on real data, we applied them to 39 tests results from two-sided two-sample proportion tests from articles published in the New England Journal of Medicine in 2015 [90]. Out of a total of 207 articles published in the journal, there are results from 39 statistical tests that compared two population proportions and resulted in null results. To demonstrate the benefits of using NAP based approaches, we consider the same 39 tests and compare weight of evidences attained by the methods described in Section 4.3.2. Data for this example are available from GitHub.

In a recent reanalysis of this data, for each study [85] performed two-sided two-sample proportion tests to test the null hypothesis that the two population proportions are equal. Following their lead, in each study we assume that the number of "successes" y_1 from Population 1 and y_2 from Population 2 are independently distributed as Binomial(n_1, p_1) and Binomial(n_2, p_2), respectively. Here p_1 and p_2 are population proportions of interest and are unknown. n_1 and n_2 are observed sample size from Population 1 and 2, respectively, and are prefixed. The tested hypotheses can

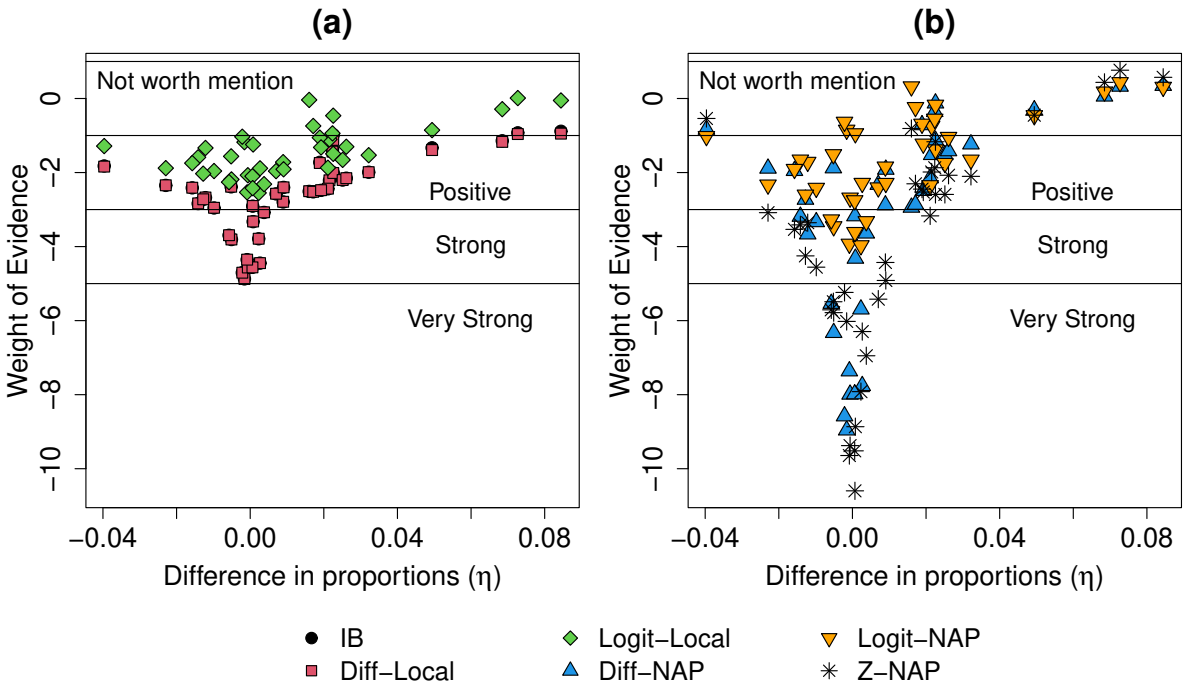


Figure 4.12: Weight of evidence achieved by all approaches in favor of H_1 in (4.33) in fixed-design tests. The horizontal axis represents difference in proportions estimated from the sample. The left panel shows weight of evidence using the local priors and the right panel shows the same obtained using the NAP based approaches.

then be expressed in frequentist terms as

$$H_0 : p_1 - p_2 = 0 \qquad \text{vs.} \qquad H_1 : p_1 - p_2 \neq 0. \tag{4.33}$$

For 39 studies we computed the P-values for the Fisher's exact test and the Pearson's Chi-squared test statistic. The lowest P-value is 0.11, which does not support the rejection of the null hypothesis of equality of proportions at 0.005 or 0.05 level of significance. Neither does it provide an interpretable summary of evidence in favor of the null. Following [85], we take a Bayesian perspective and compute Bayes factors from these data by using default Diff-NAP, Logit-NAP and Z-NAP. Figures 4.12–4.13 display the weight of evidence accumulated by each approach. We separately analyze impact of specifying priors on proportion or on the Logit scale. The figures have the

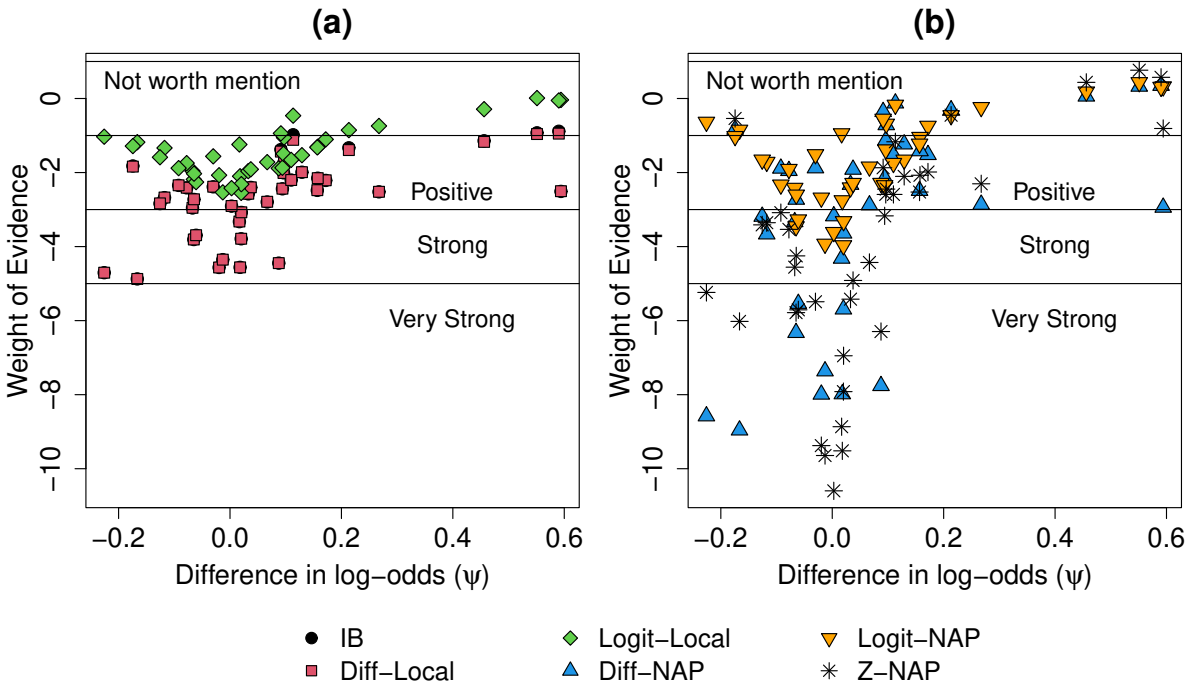


Figure 4.13: Weight of evidence achieved by all approaches in favor of H_1 in (4.33) in fixed-design tests. The horizontal axis represents difference in log-odds estimated from the sample. The left panel shows weight of evidence using the local priors and the right panel shows the same obtained using the NAP based approaches.

weight of evidence on the vertical axis with difference in proportions estimated from the samples on the horizontal axis in Figure 4.12 and with difference in log-odds estimated from the samples on the horizontal axis in Figure 4.13. The figures show that for 31 studies the Logit-Local attained a positive evidence in favor of the null and for other 8 studies the evidence accumulated is not worth a mention. Compared to the Logit-Local, the IB and Diff-Local consistently achieved a higher evidence in favor of the null for all studies. They attained strong evidence for 11 studies, positive evidence for 25 studies, and evidence that are not worth a mention for 3 studies. Figure 4.12(b) shows that when the estimated proportion difference in close to 0, the NAP based approaches attained higher evidence in favor of the null as compared to their respective local approaches. It shows that the Diff-NAP and Z-NAP achieved very strong evidence for 9 and 14 studies, strong evidence for 6 and 9 studies, positive evidence for 16 and 10 studies, evidence not worth a mention

for 8 and 6 studies. Compared to this, the Logit-NAP could not attain very strong evidence for any study. It achieved strong evidence for 6 studies, positive evidence for 20 studies, evidence not worth a mention for 13 studies. Figures 4.12–4.13 also highlights the difference between specifying priors on the proportion and on the Logit scale. Compared to all these approaches, the Z-NAP has a more consistent pattern of evidence accumulation irrespective of whether we plot it as a function of estimated difference in proportions or $\log$-odds. In both cases, at the null $H_0 : \eta = 0$ or $H_0 : \psi = 0$ the Z-NAP attains the highest amount of evidence in favor of the null compared to the other approaches. As we move away from the null on both sides, the evidence in favor of the null gets weaker. We find that for the studies with observed proportion difference outside about $(-0.03, 0.03)$, the achieved evidence is not worth a mention. This suggests a possible presence of tiny proportion differences and the sample sizes observed are not large enough to detect them. Nonetheless, the Z-NAP provides twice or more support in favor of the null hypothesis as compared to the IB and Diff-Local for 21 studies and as compared to the Logit-Local for 33 studies.

4.5 Discussion

This chapter has explored the use of non-local alternative prior densities, or NAP's, to define alternative models in Bayesian proportion tests based on one- and two-samples. From a subjective perspective, evidence suggests that NAPs approximate the marginal distribution of non-null effect sizes observed in the psychology and social science literature [64, 65, 66, 67, 68, 69, 70]. Viewed more objectively, the operating characteristics of Bayesian tests based on NAP's provide an opportunity for researchers to more rapidly accumulate evidence in favor of true null hypotheses and alternative hypotheses in which standardized effect sizes are moderate in magnitude.

For fixed design experiments, tests defined based on Z statistic using a normal moment prior model allow strong or very strong weight of evidence to be collected in favor of true null hypotheses after only an average of 500 subjects (strong weight of evidence) or 1,700 subjects (very strong weight of evidence). In contrast, tests designed using default local priors require on average at least $3\times$ or $35\times$ more subjects to obtain strong or very strong weight of evidence in favor of a

true null hypothesis. NAP alternative specifications, particularly Z-NAP also provide similar or stronger support, on average, for true proportion difference of 0.05 or larger. When true difference is smaller than that, the use of NAP often provide misleading evidence *in favor* of false null hypotheses. In these cases, tests based on local priors can provide positive evidence in favor of the alternative hypothesis, but achieving strong weight of evidence often requires more observations. Researchers interested in detecting smaller standardized effect sizes should plan on very large samples and should consider tailoring both the sample size and the prior or the NAP model in Z-NAP (for example, placing the modes of the NAP) used define the alternative hypothesis in their studies accordingly.

5. HURDLE NETWORK MODEL FOR ZERO-INFLATED DIRECTED NETWORK USING LATENT DYNAMIC SHRINKAGE PROCESS

5.1 Introduction

Statistical modeling of networks has been of active interest for many years. Recent technological advancements in diverse areas of studies have made it easier than ever to collect data on many individuals over time. As a result, the static and dynamic modeling of networks has grabbed some renewed attention. To name a few, some real life examples include functional connectivity network among brain regions, interactions between people in a social network, email communication networks, citation network among research articles or authors, network of co-purchased products, and bilateral trade flows among countries. The modeling approaches for network data can be broadly classified into two categories: models that do not use latent variables, and those that do. The network models that fall into the first category are, for example, Exponential random graph models (ERGMs), the quadratic assignment procedure (QAP), and stochastic actor oriented models (SAOMs). The latent variables in the second type of models introduce various forms of dependence between the individuals and the edges among them. Two particular models belonging to this class, namely the stochastic block models (SBMs) and the latent space models (LSMs), have seen significant developments in recent time. The SBMs assume that the individuals in the network belong to blocks or groups and the individuals can interact both within and between the groups. In this case, both the groups and the number of groups are unknown and a key objective is to estimate both of them together with edge probabilities and their memberships to each group. On the other hand, the LSMs assumes that the individuals in a network lie in a K-dimensional Euclidean space and the presence of an edge between any two individuals depends on their positions. This geometric approach of modeling provides a visual representation and interpretation of individuals in network or relational data.

There exists a host of literature on the LSMs that have been applied to the data arising from

different real-life scenarios [28, 91, 92, 93, 94, 95, 96]. To our interest, [93] proposes a class of models to analyze social network data following a conditional independence approach. Given the latent positions of individuals in an Euclidean space, the model assumes that the probability of a presence of a tie or an observed value between two individuals is inversely proportional to the distance between their positions. This model has two key attractions. First, marginalizing over the latent positions induces low-rank network structure which provides a convenient way of inducing generalized dependence among the individuals. Second, the latent positions are treated as model parameters and they are estimated from the data. These estimated positions can provide crucial insight on the underlying network structure influenced by the individuals.

Although these methods have been important in setting the premise of network data modeling, they can be improved upon in several aspects. In this research, we motivate ourselves from the bilateral trade flows observed among 29 countries from 1994 to 2013 specific to the apparel industry. The presence or absence of trades and the trade volumes in the presence of trades are observed between each pair of 29 countries. So we assume that there is an underlying network structure in the data driven by the countries. We refer to the first network as the binary network and the second as the continuous network. Beside the presence of a network structure, the data have some features which are of particular interest. These are (1) dynamic evolution of the network structure influencing both binary and continuous networks, (2) an abundance of unobserved trades among many country pairs, (3) available covariates specific to countries and pairs of countries. Several methods relying on the Gaussian random walk on the latent positions have been proposed to account for the dynamic evolution [26, 27]. But this often restricts the dynamic dependence to a Markov structure. Individual strategies exist in the literature that can separately model a binary or continuous network. In the context of bilateral trade flows, [28] proposed independently modeling binary and continuous networks sequential at each time point. This approach essentially assumes that there is one stochastic process that governs the incidence of trade and another that governs the volume of trade. But this approach can be inefficient when the proportion of presence and absence of trades become unbalanced. Also, it is counter-intuitive to assume that two independent under-

lying processes are responsible for the two networks where the same set of countries are involved in both of them.

In this research, we propose the *Hurdle Network Model* for zero-inflated network data with two key modifications. First, on the latent variables we assume a dynamic shrinkage process prior as proposed by [29]. This lets us jointly model their dynamic evolution using continuous scale mixtures of Gaussian distributions in a global-local framework. In latent space, this performs desirable shrinkage as global-local priors, while providing local adaptivity when necessary. This allows for an adaptive way of modeling trend in a time series data. Second, we assume there is a single stochastic process which governs both the binary and the continuous networks. More precisely, we assume the probability that a trade is present in a binary network is a strictly increasing function of the mean process in a continuous network. This lets us jointly model the two networks. Performance summaries from simulation study and the application on the bilateral trade data show significant improvement of the joint modeling strategy over the independent modeling and other competitive strategies.

The rest of the chapter is organized as follows. In Section 5.2, we propose the model and specify priors on model parameters. The strategy leverages on the latent dynamic shrinkage process prior and is aimed at learning dynamic structure present on the latent space. We discuss motivations behind the choices and specify pre-specified choices of hyperparameters. Section 5.3 and Section 5.4 respectively evaluate the performance of Hurdle-Net through simulation studies and an application to the bilateral trade flows data from the apparel industry. Finally, we conclude in Section 5.5 and discuss the contribution of this research.

5.2 Methodology

In this section we propose the Hurdle Network Model (Hurdle-Net) for zero-inflated network data that are observed over time. Although the proposed model is motivated from bilateral trade flows data, the model is applicable to network data arising from sources where binary and continuous measurements are observed over time between each pair in a same set of individuals. To this, in Section 5.2.1 we introduce notations to propose the model for general purposes. In Section 5.2.2,

we present the Hurdle Net using node-specific latent variables. The model allows us to jointly model binary and continuous network data, and the latent variable induces a low-rank dependence among the nodes at any given time. For inference on the model parameters we take a Bayesian approach. To this, in Section 5.2.3 we discuss the choice of a dynamic horseshoe process as the prior on the latent variables. We conclude the section in Section 5.2.4 by discussing the choices of prior for other parameters and the posterior sampling.

5.2.1 Notations

Suppose there is a fixed set of n nodes for which we observe the network data over T time points. Without loss of generality let us fix a time $t = 1, \ldots, T$. Let $\boldsymbol{Y}_t = ((Y_{ijt}))_{n\times n}$ denotes the continuous directed network data observed among the nodes at time t. Thus $\boldsymbol{Y}_t$ is a $n \times n$ matrix where $Y_{ijt} \in \mathbb{R}$, referred as the continuous edge value, denotes the observed value of a continuous variable of interest from node i to j at time t. Because $\boldsymbol{Y}_t$ is a directed network, Y_{ijt}s are asymmetric in i and j. Further, we also have a set of covariates that might be predictive of the network response $\boldsymbol{Y}_t$. Without loss of generality, let us fix $i = 1, \ldots, n$ and $j \neq i$. At time t, let $\boldsymbol{x}_{it} \in \mathbb{R}^{p_1}$ denotes a p_1-dimensional *node-specific* covariates for the node i, and $\boldsymbol{x}_{i\bullet j,t} \in \mathbb{R}^{p_2}$ denotes a p_2-dimensional *pair-specific* covariate for the pair of nodes i and j. The pair-specific covariates are symmetric in i and j in our data; that is, $\boldsymbol{x}_{i\bullet j,t} = \boldsymbol{x}_{j\bullet i,t}$. Combining the intercept and the two types of covariate information, we denote $\boldsymbol{x}_{ijt} = (1, \boldsymbol{x}_{it}, \boldsymbol{x}_{jt}, \boldsymbol{x}_{i\bullet j,t})^{\mathrm{T}}$ denotes the $(2p_1 + p_2 + 1)$ dimensional covariate information corresponding to the data Y_{ijt} at time t. Let $\|\boldsymbol{x}\|$ denote the ℓ_2 norm of a vector $\boldsymbol{x}$.

5.2.2 Hurdle Network Model for zero-inflated directed networks

Given a set of n nodes, suppose we observe the continuous network data $\boldsymbol{Y}_t$ among them at times $t = 1, \ldots, T$ where a significant proportion of continuous edge values exactly equal to 0. Such is the case in the bilateral trade flows data from the apparel industry where 30% of the Y_{ijt} values are 0. To take this into account in the model, we introduce a binary network $\boldsymbol{\Delta}_t$ similar to $\boldsymbol{Y}_t$. Here $\boldsymbol{\Delta}_t$ is a $n \times n$ adjacency matrix where δ_{ijt} is a binary variable with 1 and 0 respectively

indicating the presence and absence of a directed edge from node i to j at time t. From here on, we consider T pair of networks $(\boldsymbol{\Delta}_t, \boldsymbol{Y}_t)$ as the observed data.

Hurdle Net is based on two key assumptions. First, following [93] at each time point we take a conditional independence approach for modeling dependence among the nodes in the networks. To this, define $\boldsymbol{Z}_t = (\boldsymbol{z}_{1t}, \ldots, \boldsymbol{z}_{nt})^{\mathrm{T}}$ where $\boldsymbol{z}_{it} \in \mathbb{R}^K$ denotes the latent position of node i at time t in the K-dimensional Euclidean space. Given $\boldsymbol{Z}_t$, Hurdle-Net assumes the following:

Conditional independence in the binary network. The presence or absence of a tie from node i to j is independent of all other ties in the network. For $i \neq i^{'}$ or $j \neq j^{'}$ this means, δ_{ijt} is conditionally independent of $\delta_{i^{'}j^{'}t}$ given $\boldsymbol{z}_{it}$ and $\boldsymbol{z}_{jt}$.

Conditional probability of a tie in the binary network. The conditional probability that a tie is present from node i to j is proportional to their similarity in the latent space. The greater the similarity between $\boldsymbol{z}_{it}$ and $\boldsymbol{z}_{jt}$ the larger is the probability that a tie is present, and vice versa.

Conditional independence in the continuous network. The continuous value observed from node i to j is independent of all other values in the network. For $i \neq i^{'}$ or $j \neq j^{'}$ this means, Y_{ijt} is conditionally independent of $Y_{i^{'}j^{'}t}$ given $\boldsymbol{z}_{it}$ and $\boldsymbol{z}_{jt}$.

Conditional mean in the continuous network. The expected continuous edge value from node i to j is proportional to their similarity in the latent space. The greater the similarity between $\boldsymbol{z}_{it}$ and $\boldsymbol{z}_{jt}$ the larger is the expected value, and vice versa.

Second, we assume that both δ_{ijt} and Y_{ijt} are governed by the same underlying mechanism and propose a joint model for the two networks. Combining these assumptions, we consider the following parametric model: for all $1 \leq i \neq j \leq n$ and $t = 1, \ldots, T$,

$$Y_{ijt} \,|\, \delta_{ijt} = 1, \boldsymbol{x}_{ijt}, L_{ijt}, \boldsymbol{\beta}, \sigma \overset{ind}{\sim} N\left(\boldsymbol{x}_{ijt}^{\mathrm{T}}\boldsymbol{\beta} + L_{ijt}, \sigma^2\right), \tag{5.1}$$

$$\delta_{ijt} \,|\, \boldsymbol{x}_{ijt}, L_{ijt}, \boldsymbol{\beta} \overset{ind}{\sim} \text{Bernoulli}\left(\Phi\left(f\left(\boldsymbol{x}_{ijt}^{\mathrm{T}}\boldsymbol{\beta} + L_{ijt}\right)\right)\right), \tag{5.2}$$

where Φ is the standard normal cdf. We refer to (5.1) as the *continuous model* and it models the continuous edge value given the presence of a tie, while (5.2) is referred to as the *probit model* and it models the presence or absence of a tie. Thus (5.1) and (5.2) jointly model the binary and continuous networks, and we collectively refer to them as the *Hurdle Network Model (Hurdle-Net)*. As for the other components of the model, $\boldsymbol{x}_{ijt}$ denotes the vector of observed covariates corresponding to each tie, $\boldsymbol{\beta}$ is the static regression coefficient (including the intercept), σ^2 is the error variance, L_{ijt} is the latent term, and $f : \mathbb{R} \mapsto \mathbb{R}$ is a strictly increasing function. The choices for L_{ijt} and f and their importance are discussed below.

Several models have been proposed in the literature for modeling dynamic networks (interested readers please refer to [97]). Briefly, these models can be broadly classified into two types: latent space models (LSM) and stochastic block models (SBM). LSM assumes that the probability of a tie, or the expected continuous edge value is proportional to their relative positions in the latent space. On the other hand, SBM assumes that the nodes can be partitioned into some latent classes or blocks. Here we take the first approach. In particular, we consider the projection model as in [93] and use the latent term $a_i \, \boldsymbol{v}_{it}^{\mathrm{T}} \boldsymbol{v}_{jt}$. Here $a_i > 0$ is the activity level of node i (*parent node*) and $\boldsymbol{v}_{it}$ is a unit-length K-dimensional Euclidean space associated with node i at time t. This parameterization represents the nodes at each time point as points on the K-dimensional unit sphere. In a binary network, at each time point it encourages the presence of a tie from node i to j if the angle between $\boldsymbol{v}_{it}$ and $\boldsymbol{v}_{jt}$ are small (that is, $\boldsymbol{v}_{it}^{\mathrm{T}} \boldsymbol{v}_{jt} > 0$), discourages if the angle is obtuse (that is, $\boldsymbol{v}_{it}^{\mathrm{T}} \boldsymbol{v}_{jt} < 0$), and stays neutral if the angle is right angle (that is, $\boldsymbol{v}_{it}^{\mathrm{T}} \boldsymbol{v}_{jt} = 0$). Similar interpretation can also be drawn in the context of a continuous network. Following the suggestion of [93] in a static model, one can choose $\boldsymbol{z}_{it} = a_i \, \boldsymbol{v}_{it}$, which means $a_i = \|\boldsymbol{z}_{it}\|$ and $\boldsymbol{v}_{it} = \boldsymbol{z}_{it} / \|\boldsymbol{z}_{it}\|$. Although the presence of a_i is crucial and allows for asymmetry in the model, one can similarly consider a_j, the activity level of node j (*child node*), in the model. But that can possibly lead to a different inference. So to ensure invariant inference from either modeling $(\boldsymbol{\Delta}_t, \boldsymbol{Y}_t)$ or $\left(\boldsymbol{\Delta}_t^{\mathrm{T}}, \boldsymbol{Y}_t^{\mathrm{T}}\right)$, we define

$$L_{ijt} = \alpha \, \frac{\boldsymbol{z}_{it}^{\mathrm{T}} \boldsymbol{z}_{jt}}{\|\boldsymbol{z}_{jt}\|} + (1 - \alpha) \, \frac{\boldsymbol{z}_{it}^{\mathrm{T}} \boldsymbol{z}_{jt}}{\|\boldsymbol{z}_{it}\|}, \qquad \text{for } 0 \leq \alpha \leq 1. \tag{5.3}$$

α close to 1 implies the effect of parent nodes is dominant in the network, whereas a value close to 0 implies the same for the child nodes.

For joint modeling of the two networks, we assume that a single underlying process governs both networks. More precisely, Hurdle-Net assumes that the probability of a tie in the probit model is a strictly increasing function (f) of the expected edge value in the continuous model. So the success probability in (5.2) is a composition of Φ and f, denoted by $\Phi \circ f$. We note that, for any strictly increasing f, $\Phi \circ f$ can be uniquely represented by an 'S'-shaped function $f^* : \mathbb{R} \mapsto (0, 1)$. The latter class of functions can be derived as a special case of the generalized logistic function. So, with a slight abuse of notation, we replace $\Phi \circ f$ in (5.2) by f (different from the f in $\Phi \circ f$) which has the parametric form

$$f(x) = \left(1 + e^{(a-bx)}\right)^{-1/\gamma}, \quad \text{for } a \in \mathbb{R} \text{ and } b, \gamma > 0. \tag{5.4}$$

Here f can be considered as a generalized link function which provides a data-driven generalized way of modeling the edge probability in the binary network. This significantly reduces the complexity of the proposed model, and removes the necessity of specifying a fixed link function (for example, the logit or the probit) at the expense of only three more parameters.

In Hurdle-Net $\boldsymbol{\beta}$, σ, α, a, b, and γ are static model parameters, and $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T$ are dynamic latent positions. All of these parameters are of key interest and they need to be estimated. We take a Bayesian route and hierarchically specify priors on the parameters for posterior inference. The choice of priors and the motivation behind them are discussed below in the following subsections.

5.2.3 Latent dynamic shrinkage process

Given the latent positions, Hurdle-Net is conditionally independent across the ties over time. To estimate the positions, for identifiability we assume that $\boldsymbol{Z}_1$ is lower triangular with positive diagonal elements. Then we *apriori* specify the dynamic shrinkage process (DSP) to model their dynamic evolution [29]. The motivation behind DSP stems from a Bayesian adaptation of trend filtering models [98, 99, 100]. First, a suitable difference of the time varying model parameter is

defined. The order of difference relates to sharpness in the change of slope of the latent position between subsequent time points. Then a global-local continuous shrinkage prior is assumed on the differences. For specifying the process of order 1, let us define

$$\boldsymbol{\Omega}_1 = \boldsymbol{Z}_1, \text{ and } \boldsymbol{\Omega}_t = \boldsymbol{Z}_t - \boldsymbol{Z}_{t-1}, \text{ for } t > 1, \tag{5.5}$$

Also, for any t denote $\boldsymbol{\Omega}_t = (\boldsymbol{\omega}_{1t}, \ldots, \boldsymbol{\omega}_{nt})^{\mathrm{T}}$. Then for $i = 1, \ldots, n$, $t = 1, \ldots, T$, and $k = 1, \ldots, K$, we hierarchically specify the prior as follows:

$$\begin{aligned}
&\omega_{itk} \,|\, \tau_0, \{\tau_j\}, \{\lambda_{js}\} \overset{ind}{\sim} N\left(0, \tau_0^2 \tau_i^2 \lambda_{it}^2\right), \\
&h_{it} = \log\left(\tau_0^2 \tau_i^2 \lambda_{it}^2\right), \\
&h_{i1} = \mu_0 + \mu_i + \eta_{i1}, \text{ and } h_{it} = \mu_0 + \mu_i + \phi_i\left(h_{i,t-1} - \mu_0 - \mu_i\right) + \eta_{it}, \\
&\mu_0, \mu_i, \eta_{it} \overset{iid}{\sim} Z\left(c, s, 0, 1\right),
\end{aligned} \tag{5.6}$$

where $Z\left(c, s, \mu_z, \sigma_z\right)$ denotes the Z-distribution [29]. In general, one can specify a process of order $D \in \mathbb{N}$ by defining $\boldsymbol{\Omega}_t = \boldsymbol{\Delta}^D \boldsymbol{Z}_t$ in (5.5), where $\boldsymbol{\Delta}^D$ is the difference operator of order D. Then one can specify the same hierarchical model (5.6) on the differences. As discussed in [29], the Z-distribution is a general class of distributions and contains many important shrinkage distributions in the literature (Please refer to Table 1 in [29]). To our interest, $Z\left(1/2, 1/2, 0, 1\right)$ corresponds to the Horseshoe prior on the differences ω_{itk}'s [101]. The probability distribution is known to have many attractive shrinkage properties. The positions across different latent dimensions are assumed independent and identically distributed. The shrinkage behavior of the differences $\boldsymbol{\omega}_{it}$ determines the dynamics of the latent positions. When the differences are shrunk toward 0, the dynamics of the positions are locally linear. On the other hand, large values of $\boldsymbol{\omega}_{it}$ results in large changes of slopes in the dynamics of positions. Thus the hyperparameters τ_0, τ_i's, and λ_{it}'s are crucial in determining the shrinkage behavior of the differences. For each i, the dynamic structure in the prior comes from the dependence between local scale parameters λ_{it} where λ_{it} is informed by its past values $\{\lambda_{is}\}_{s<t}$. λ_{it} near 0 implies aggressive shrinkage which indicates no change in the

dynamics $\boldsymbol{z}_{it}$, and a large value indicates significant absolute change in $\boldsymbol{z}_{it}$ from the previous time point. This temporally adaptive shrinkage behavior controls the smoothness and adaptivity in the dynamics of the latent positions. Henceforth, for any difference order D, we collectively refer to (5.5)–(5.6) with $c = s = 1/2$ as the Dynamic Horseshoe Shrinkage (DHS) process of order D with adaptive shrinkage, and is denoted by Adaptive-DHS(D). In this specification, setting $\phi_i = 0$ for all i does not allow for adaptive shrinkage. This specification is referred to as the DHS process of order D with non-adaptive shrinkage, and is denoted by NonAdaptive-DHS(D).

5.2.4 Other priors and sampling from the posterior

In this section we specify priors on other model parameters and discuss sampling from the posterior distribution. We assume a non-informative normal prior $N\left(\mathbf{0}, \sigma_\beta^2 \boldsymbol{I}\right)$ with large σ_β on $\boldsymbol{\beta}$, the Jeffrey's prior on σ^2 which is proportional to $1/\sigma^2$, the Uniform prior in $[0, 1]$ on α, a non-informative normal prior $N\left(0, \sigma_a^2\right)$ with large σ_a on a, and the Uniform prior in $[\varepsilon, L]$ on b and γ where $\varepsilon > 0$ and L is large. For the results presented here, we set $\sigma_\beta = \sigma_a = L = 10^5$ and $\varepsilon = 10^{-5}$.

For conducting Bayesian inference, we note that all the parameters in the model are continuous. We take advantage of the probabilistic programming language **`Stan`** and use the No-U-turn Hamiltonian Monte Carlo sampling to draw samples from the joint posterior distribution [102]. We further use serial tempering as a warm-start to better initialize the Markov chain, which leads to quicker MCMC convergence. Suppose we have prespecified a sequence of increasing positive fractions $\{\lambda_1, \ldots, \lambda_G\}$ with $\lambda_G < 1$. The sampling procedure has two stages: burn-in and final sampling. Based on the prefixed sequence we divide the burn-in stage into G sequential burn-in. At step g, we implement **`Stan`** with the likelihood equals to the actual likelihood raised to the power λ_g. At step 1 of burn-in, the starting point of the sampling algorithm is randomly chosen. At subsequent steps, it is set to be the last sample drawn at the previous step. At the final sampling stage we initialize at the last sample drawn at step G of burn-in and implement **`Stan`** with the actual likelihood to get desired number of posterior samples.

5.3 Simulation study

In this section, we compare performances of different methods through simulation studies. We set up the simulation experiment to mimic some features of the real data while making it misspecified from the proposed model in other aspects. We set the number of time points T to 10. Following the real-data described in Section 5.4 we set the number of node-specific covariates p_1 to 3 and the number of pair-specific covariates p_2 to 2. To mimic some features of the real data, we set the true regression coefficient $\boldsymbol{\beta}$, the parameter in the latent term α, the noise standard deviation σ, the true generalized link function f are set to their estimates from the Hurdle-Net+Adaptive-DHS(1) that provides the best prediction performance (Please see Section 5.4.3). Following this the latent dimension K is prefixed to 5. To generate the true node-specific latent attributes $\boldsymbol{Z}_t$'s,

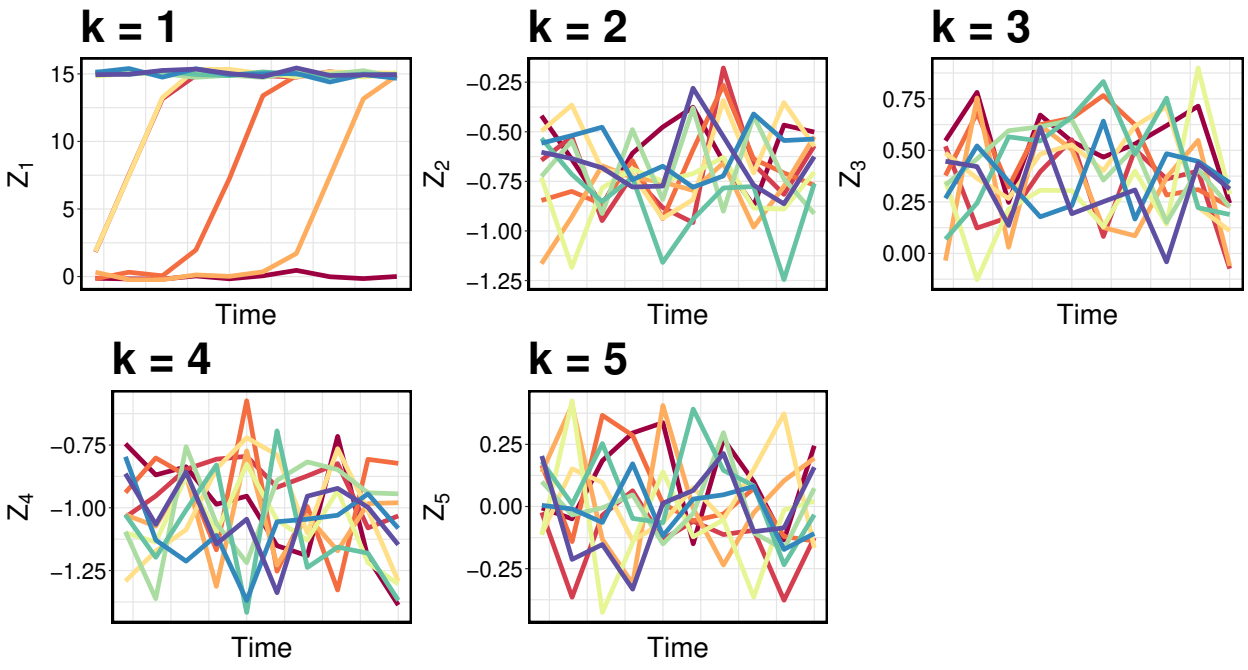


Figure 5.1: Simulated latent positions at 10 time points for $n = 10$. Data from the first 9 time points are used to fit the models. The predictive performance is tested at the last time point.

we mimic their estimate presented in Section 5.4.3 (Please see 5.10). We assume that there exists two groups of equal size of the n expected Z_{it1} in the first latent dimension. Over time 80% of the

nodes in the cluster closer to the origin moves to the second group. The times when the transitions start is chosen at random. The group-means in other latent dimensions are assumed to have only one group but the group-mean values are different from each other. Finally Z_{itk}'s are simulated by adding a random noise to it. The random noises are generated from normal distribution with mean 0 and standard deviation 0.2. Figure 5.1 shows the simulated latent variables for $n = 10$ for the 5 latent dimensions.

In addition to the regression coefficient $\boldsymbol{\beta}$, the estimation and prediction of edge probabilities and expected nonzero means for observed edge occurrences are also of interest. Parameter estimates are obtained using the posterior mean. We compute the Normalized Mean Squared Error (MSE) to quantify the performance. For an estimate $\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$ and its true value $\boldsymbol{\theta}_0 \in \mathbb{R}^d$, the normalized squared error is defined as $\mathbb{E}\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\|^2 / d$. For example, to compute this for the latent term, we vectorize $\{L_{ijt}\}$ as

$$\boldsymbol{L} = (L_{121}, \ldots, L_{n,n-1,1}, L_{122}, \ldots, L_{n,n-1,2}, \ldots, L_{12T}, \ldots, L_{n,n-1,T})^{\mathrm{T}}.$$

To quantify the estimation and prediction performance of the edge probabilities and expected nonzero means, they are vectorized and then its normalized MSE is calculated. To compute the normalized MSE and MSPE for the edge probabilities, we use the binary vector of the observed occurrences as the true value. We compare the performance of the Network Hurdle Model with several of its simplifications. These are as follows:

1. Hurdle-Net model with Adaptive-DHS(1) prior on the latent positions. This is denoted by Hurdle-Net+Adaptive-DHS(1).

2. Hurdle-Net model with NonAdaptive-DHS(1) prior on the latent positions. This is denoted by Hurdle-Net+NonAdaptive-DHS(1). This evaluates the importance of adaptive shrinkage in the DHS for modeling dynamic evolution of the latent positions.

3. Hurdle-Net model with Adaptive-DHS(0) prior on the latent positions. This is denoted by Hurdle-Net+Adaptive-DHS(0). This evaluates whether modeling the first order difference

is beneficial to capture the trend in the latent dynamics. In the presence of a strong trend, we expect Hurdle-Net+Adaptive-DHS(1) to provide significant improvement over this specification.

4. Hurdle-Net model with NonAdaptive-DHS(0) prior on the latent positions. This is denoted by Hurdle-Net+NonAdaptive-DHS(0). This evaluates the importance of modeling first order difference and of allowing for adaptive shrinkage.

5. Independent modeling of binary and continuous network using the probit and Gaussian model as in (5.2) and (5.1), respectively. In each model, we assume the Adaptive-DHS(1) prior on the latent positions. This is denoted by Indep+Adaptive-DHS(1). This evaluates the importance of a joint modeling approach for the two networks.

6. Hurdle-Net model with static latent positions. We assume a Horseshoe prior on the latent positions and this is denoted by Hurdle-Net+HS. This evaluates the importance of dynamic latent positions.

7. Hurdle-Net model without latent positions. This is denoted by Hurdle-Net+NoLatent. This evaluates the importance of latent positions.

The data is simulated from 20 replicated studies. We add an intercept and set it to 7 when simulating from the continuous model to approximately match the proportion of observed nonzero edge occurrences in the simulated data with the real data in Section 5.4. For the 50 replicated studies the observed nonzero edge occurrence percentage varies from $66 - 70\%$. For each replicated study, we fit each of the 7 models on the data until the 9^{th} time point and then predict at the 10^{th} time point. To compare fitting and prediction performances, we vary n as 10, 20 and 30.

Figures 5.2–5.3 together illustrate the efficacy of the proposed methods, namely Hurdle-Net+Adaptive-DHS(1) and Hurdle-Net+NonAdaptive-DHS(1), over other methods. Figure 5.2 shows that on average, Hurdle-Net+Adaptive-DHS(1), Hurdle-Net+NonAdaptive-DHS(1) and Indep+Adaptive-DHS(1) achieve similar MSEs in regression coefficient estimation. The MSEs obtained by them

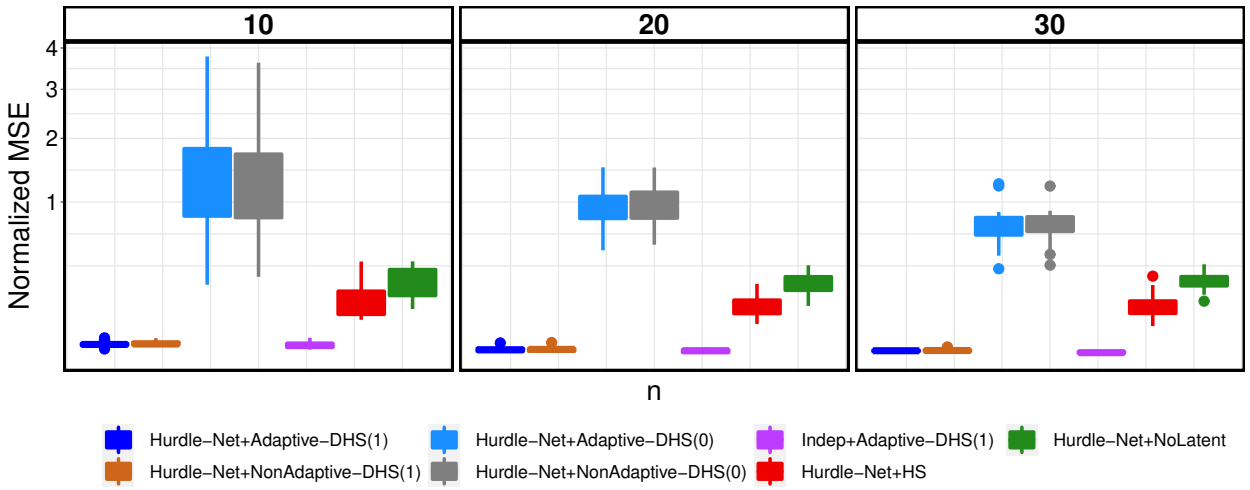


Figure 5.2: Boxplots of normalized Mean Squared Errors (MSEs) of regression coefficient estimated based on the posterior mean from different methods in replicated studies.

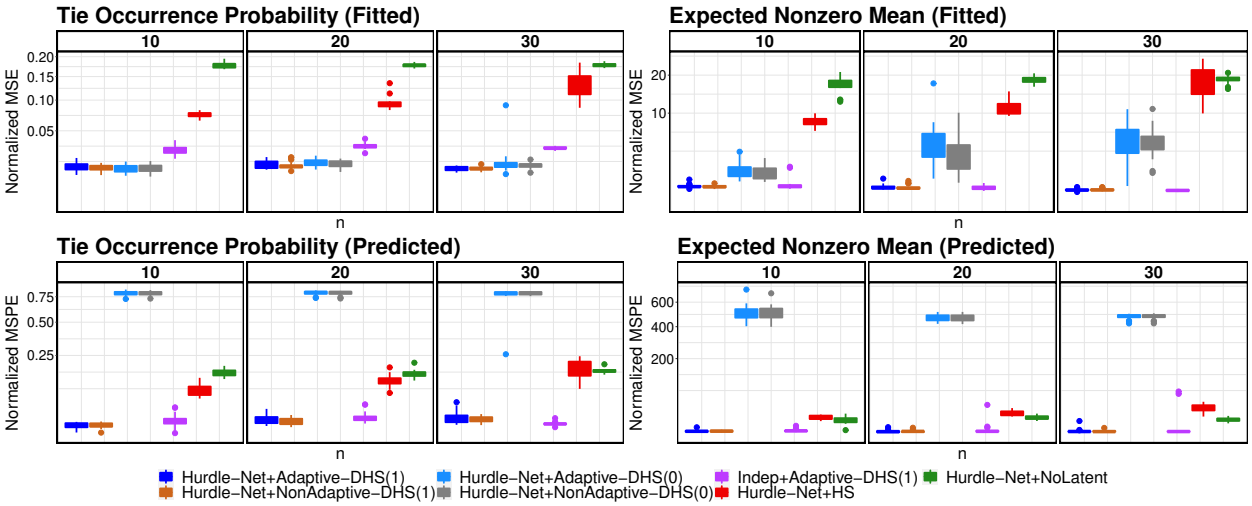


Figure 5.3: Mean squared error (MSE) and mean squared prediction error (MSPE) of parameters and terms in the model. Heights of the bars are the MSEs or MSPEs with error bars denoting ± 1 standard errors around it. Bars depicted in the plot correspond to the 7 methods under comparison.

are significantly better than the other methods. In this figure, we find that the normalized MSEs of Hurdle-Net+Adaptive-DHS(0) and Hurdle-Net+NonAdaptive-DHS(0) is particularly worse. This shows the significance of modeling the first order difference along the time scale of node-specific latent attributes. In Figure 5.3, similar differences in performance are also observed in estimation

and prediction of expected nonzero means and tie occurrence probabilities. The top and bottom row in this figure respectively analyzes the performance on the training and test set. In estimating tie occurrence probabilities, Indep+Adaptive-DHS(1), Hurdle-Net+HS, and Hurdle-Net+NoLatent have worse MSEs than the other methods uniformly over the number of nodes. Their performances does not improve even as n increases. As for the other four methods, the MSEs of Hurdle-Net+Adaptive-DHS(1) and Hurdle-Net+NonAdaptive-DHS(1) have higher MSEs as compared to the Hurdle-Net+Adaptive-DHS(0) and Hurdle-Net+NonAdaptive-DHS(0) when the number of nodes is as low as 10. In this case the method does not benefit from shrinking the latent positions to the positions at the previous time point. As the number of nodes increases to 30, the benefit of shrinkage is revealed and Hurdle-Net+Adaptive-DHS(1) and Hurdle-Net+NonAdaptive-DHS(1) have significantly better MSEs over all the methods. In estimating the expected nonzero mean for observed tie occurrences, when the number of nodes is 10 the Hurdle-Net+Adaptive-DHS(1) and Hurdle-Net+NonAdaptive-DHS(1) have the lowest MSEs and the latter is as good as Indep+Adaptive-DHS(1). As the n increases to 30, Hurdle-Net+Adaptive-DHS(1), Hurdle-Net+NonAdaptive-DHS(1), Hurdle-Net+Adaptive-DHS(0), Hurdle-Net+NonAdaptive-DHS(0) and Indep+Adaptive-DHS(1) perform similar to each other and has significantly lower MSEs than Hurdle-Net+HS, and Hurdle-Net+NoLatent.

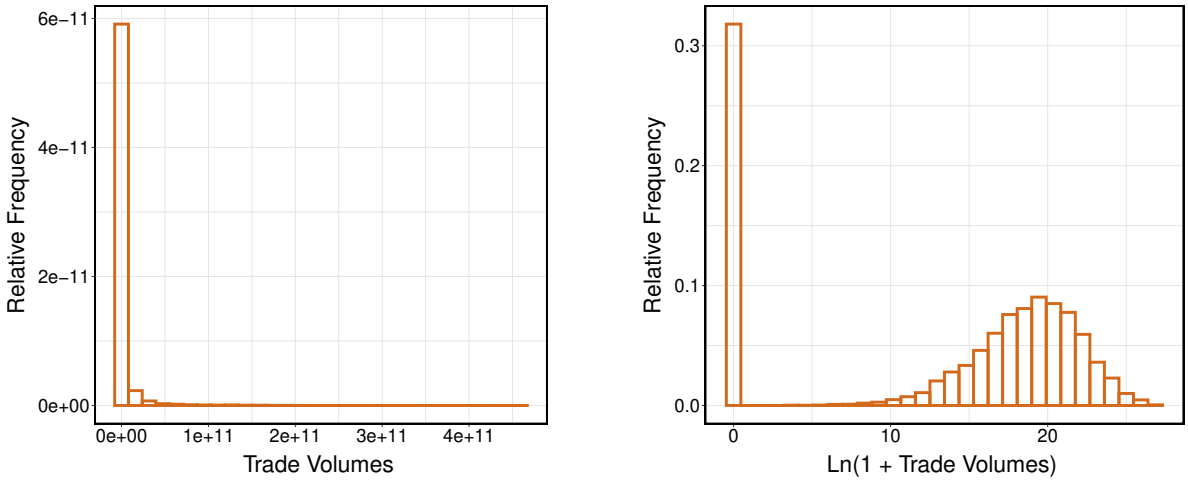
The second row of Figure 5.3 presents the predictive performance of each model at the last time point. In predicting both tie occurrence probabilities and expected nonzero mean for occurred ties, Hurdle-Net+Adaptive-DHS(0) and Hurdle-Net+NonAdaptive-DHS(0) have at least 6 times the MSPEs of the other methods. The MSPEs are about 26 times higher than others when predicting the expected nonzero mean for occurred ties. Compared to this, Hurdle-Net+Adaptive-DHS(1) and Hurdle-Net+NonAdaptive-DHS(1) are consistently having smaller MSPEs with smaller standard errors.

Combining Figures 5.2–5.3 we feel it is safe to conclude that the performances of Hurdle-Net+Adaptive-DHS(1) and Hurdle-Net+NonAdaptive-DHS(1) is very consistent and robust over all simulation settings. In some scenarios they perform as good as the independent approach while

in other scenarios the shrinkage of the node-specific latent attributes to its previous time points can provide substantial improvement in the performance with lower standard error. These highlight the importance of dynamic node-specific latent attributes and their adaptive dynamic shrinkage. As for the comparison with an independent modeling approach, an improved performance can be observed in estimation and prediction of edge probabilities, particularly when an unbalanced number of presence and absence of edges are observed in the binary network. In such cases, a joint modeling strategy borrows information from the continuous model and provide lower MSEs and MSPEs with lower standard errors. As for the performance difference between Hurdle-Net+Adaptive-DHS(1) and Hurdle-Net+NonAdaptive-DHS(1), the non-adaptive variant evaluates the importance of allowing adaptive shrinkage in the $\log$-variance model. Whether this can prove to be superior depends on the complexity of the model and informativeness of the underlying latent dynamics. Although the two methods are performing quite similar to each other in the simulation setting considered here, Hurdle-Net+Adaptive-DHS(1) shows a significant improvement in the real-data analysis on the bilateral trade data from the apparel industry. This discussion is further continued in the next section.

5.4 Application to international trade of apparel industry

To illustrate the use of Hurdle-Net+Adaptive-DHS(1) on a real-data, we applied them to the international trade data from the apparel industry. The data contains import volumes between each pair of 29 countries from 1994 to 2013. Combining all the years, trades occurred between 70% of the country pairs. For the rest of the pairs, the countries did not trade. In this application, we model the dynamics of the trade network among the 29 countries, determine the importance of available covariates, and predict trade occurrence probabilities and trade volumes in presence of a trade. For comparison we apply the methods presented in Section 5.3 and compare their predictive performances. For the purpose of modeling, we transform the observed trade volumes on the $\log$ scale to satisfy the normality assumption better in the continuous model (5.1). The histogram of the actual trade volumes is presented in Figure 5.4(a). Figure 5.4(b) shows the histogram of $\ln(1 + \text{trade volumes})$. To compare predictive performance, we use the data from 1994 to 2012 as



(a) Original scale

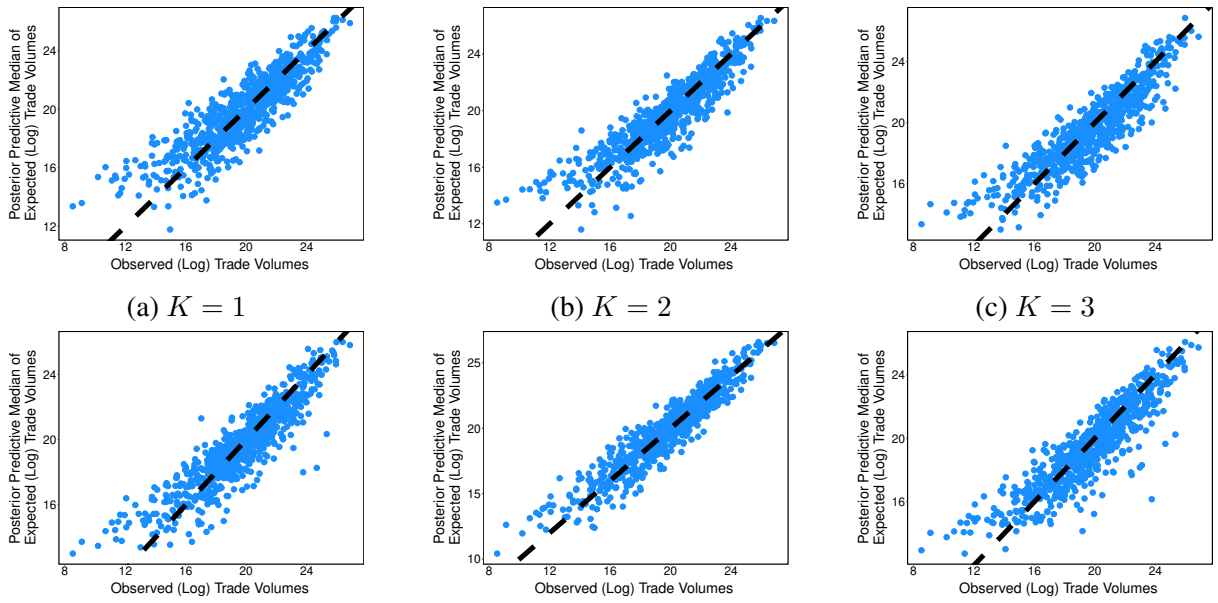
(b) After $\ln$ transformation

Figure 5.4: Histogram of the observed trade volumes before and after the $\ln$ transformation. The histogram on the left is for the actual observed trade volumes. On the right, it shows the histogram of $\ln(1 + \text{trade volumes})$. On the left of this histogram we see a bar of approximate height 0.3. This corresponds to the country pairs with unobserved trade occurrences.

the training data and then use the data in 2013 as the test data. First, we fit Hurdle-Net+Adaptive-DHS(1) on the training data with latent dimension K varied from 1 through 6. Then we choose the value of K that provides the best prediction in 2013. Next, we fit all other models on the training data using this best value of K and compare predictive performances of the 7 models. In this modeling $n = 29$ and $T = 19$. For a fixed pair of countries (i, j) and time t, we assume δ_{ijt} denotes the presence or absence of a trade from country i to j, and the continuous variable Y_{ijt} denotes the $\log$ of trade volume from country i to j. For each country we have their GDPs, populations and areas as the node-specific covariates, and distances between capitals and labour provisions as the pair-specific covariates. Thus $p_1 = 3$, $p_2 = 2$, and $\boldsymbol{x}_{ijt}$ denotes 8 available covariates corresponding to each pair of countries.

5.4.1 Performance for varied latent dimensions

To determine the impact of varying latent dimensions, we fit Hurdle-Net+Adaptive-DHS(1) to the training data for varied latent dimensions and then compare their predictive performance on the test set. Figure 5.5–5.6 compares the MSPEs corresponding to latent dimensions 1 through 6



(a) $K = 1$ (b) $K = 2$ (c) $K = 3$

(d) $K = 4$ (e) $K = 5$ (f) $K = 6$

Figure 5.5: Comparison of posterior predictive median of expected log(trade volume) and the observed log(trade volume) in 2013 for the country pairs between which trades occurred. The figures show the performance of Hurdle-Net+Adaptive-DHS(1) for prefixed latent dimension K varied from 1 through 6. For observed trade occurrences, x-axis denotes log(trade volume) observed between the country pairs, and y-axis denotes the posterior predictive median of expected log(trade volume) for them. The dashed black line denotes the $y = x$ line for reference.

for continuous and binary prediction, respectively. We separately compare each prediction performance by calculating the normalized MSPEs as defined in Section 5.3.

Figure 5.5 shows the continuous prediction performance at 2013 for Hurdle-Net+Adaptive-DHS(1) for varied K. The vertical axis in the figure denotes the predicted $\log$ trade volumes for the observed trade occurrences in the test set. The horizontal axis denotes the observed $\log$ trade volumes for the observed trade occurrences in the test set. For reference we overlay the plot with the $y = x$ line. The more tightly the points are around the line, the better is the prediction. Similarly Figure 5.6 shows the binary prediction performance. On the horizontal axis 0 and 1 denotes the observed absence and presence of trades. Against each of them, on the vertical axis we show the boxplot of the posterior predictive median of trade occurrence probabilities. For the country pairs with no trades we expect the boxplot to be near 0 and for the other we expect it

Table 5.1: Mean squared prediction errors in 2013 for Hurdle-Net+Adaptive-DHS(1) with prefixed latent dimensions 1 through 6.

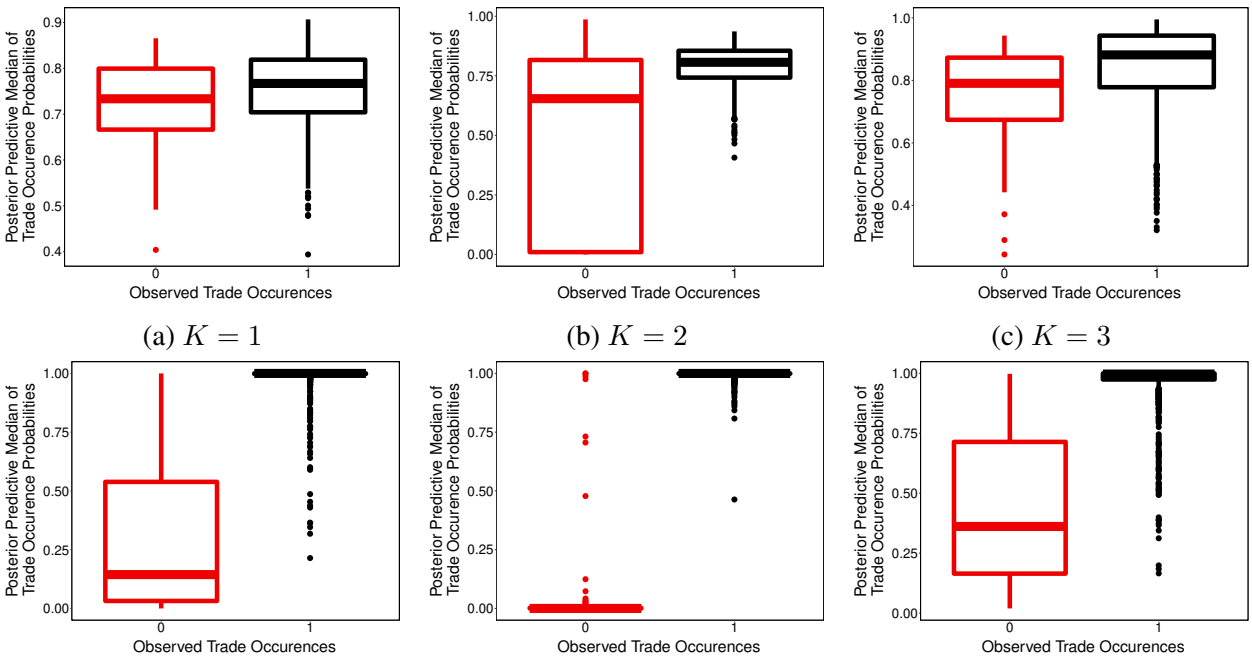
Latent dimension (K)	Continuous prediction	Binary prediction
1	2.02	0.10
2	1.67	0.07
3	1.71	0.09
4	1.66	0.02
5	0.97	0.007
6	1.83	0.04

Note. For checking accuracy in continuous prediction, we compute the normalized MSPEs between predicted and observed log(trade volume) values for observed trade occurrences. Also for checking accuracy in binary prediction, we compute normalized MSPEs between predicted trade occurrence probabilities and observed trade occurrences.

to be near 1. The closer the boxplots are to 0 and 1 respectively, the better the model is able to predict the presence or absence of trades. Finally, Table 5.2 shows the normalized MSPEs calculated for both binary and continuous prediction and it clearly shows the superiority in the performance of Hurdle-Net+Adaptive-DHS(1). The figures show that both the continuous and the binary prediction improve as we increase the latent dimension. At $K = 5$ the MSPE reaches the lowest in both prediction and then increases again at $K = 6$. Table 5.1 presents the MSPEs in both predictions. It shows that at $K = 5$ the MSPEs in continuous and binary predictions have reduced 42% and 65%, respectively, compared to at $K = 4$. This suggests that a low-rank latent structure of rank 5 is required in addition to the observed covariates to explain the variance in the observed data. This implies there are covariates other than the 8 that are considered here and are responsible for explaining the dynamic of the network.

5.4.2 Model comparison

Next, we provide a comparative analysis and focus on the same 7 models described in Section 5.3. As in the section above, we train the models on the training set from 1994 to 2012



(a) $K = 1$ (b) $K = 2$ (c) $K = 3$

(d) $K = 4$ (e) $K = 5$ (f) $K = 6$

Figure 5.6: Comparison of posterior predictive median of trade occurrence probability and the observed trade occurrence in 2013 for all pairs of 29 countries. The figures show the performance of Hurdle-Net+Adaptive-DHS(1) for prefixed latent dimension K varied from 1 through 6. in each figure, 0 and 1 on x-axis refers to observed and unobserved trades among country pairs. y-axis denotes the posterior predictive median of trade occurrence probabilities for those pairs.

and check their binary and continuous prediction accuracy in 2013. We fit each model (except Hurdle-Net+NoLatent) with prefixed latent dimension $K = 5$ and then compare their binary and continuous prediction for the year 2013. This is presented in Figure 5.7–5.8.

Figure 5.7 shows the continuous prediction performance at 2013 for different models for prefixed latent dimension $K = 5$. The vertical axis in the figure denotes the predicted $\log$ trade volumes for the observed trade occurrences in the test set. The horizontal axis denotes the observed $\log$ trade volumes for the observed trade occurrences in the test set. For reference we overlay the $y = x$ line on the plot. The more tightly the points are around the line, the better is the prediction. The figure shows that the points are better wrapped around the $y = x$ line for Hurdle-Net+Adaptive-DHS(1) than the methods. Hurdle-Net+Adaptive-DHS(0) and Hurdle-Net+NonAdaptive-DHS(0) particularly show a constant under estimation in the prediction. Similarly Figure 5.8 shows the bi-

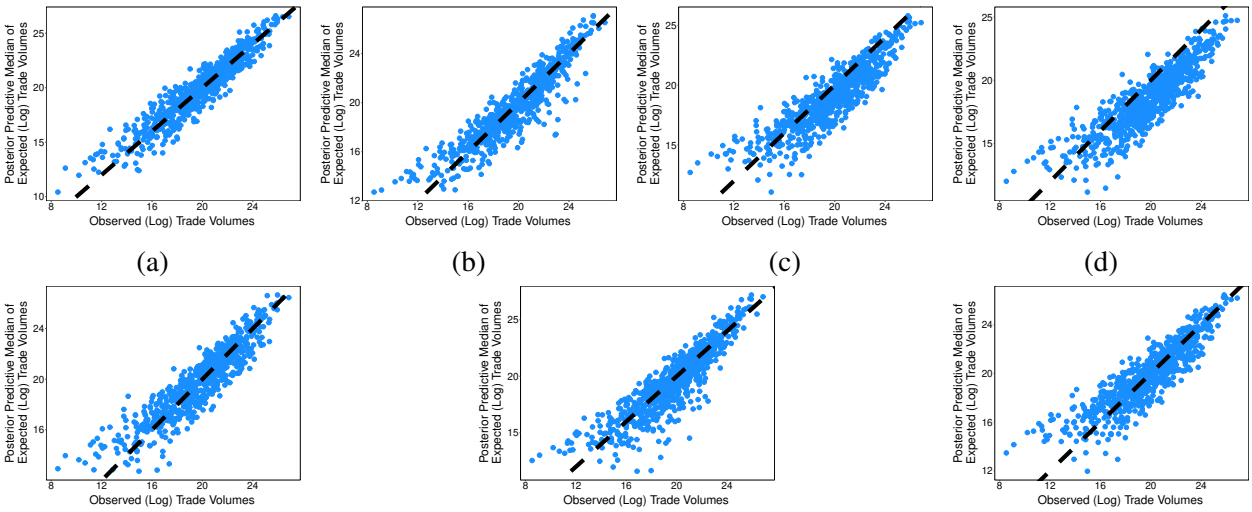


Figure 5.7: Comparison of posterior predictive median of expected log(trade volume) and the observed log(trade volume) in 2013 for the country pairs between which trades occurred. The figures show the performance of Hurdle-Net+Adaptive-DHS(1) for prefixed latent dimension K varied from 1 through 6. For observed trade occurrences, x-axis denotes log(trade volume) observed between the country pairs, and y-axis denotes the posterior predictive median of expected log(trade volume) for them. The dashed black line denotes the $y = x$ line for reference.

nary prediction performance. On the horizontal axis 0 and 1 denotes the observed absence and presence of trades. Against each of them, on the vertical axis we show the boxplot of the posterior predictive median of trade occurrence probabilities. For the country pairs with no trades we expect the boxplot to be near 0 and for the other we expect it to be near 1. The closer the boxplots are to 0 and 1 respectively, the better the model is able to predict the presence or absence of trades. Hurdle-Net+Adaptive-DHS(1) shows the best performance among the models considered here. Indep+Adaptive-DHS(1) very accurately predicts the probabilities in the presence of trades, but does poorly for those that are absent. The other methods predict similar trade occurrence probabilities for trades that are both present and absent, and thus perform poorly. This clearly shows the benefit of shrinking the latent positions of countries to their positions at the previous time point in explaining the dynamics of the trade network. Besides, [28] provides a similar analysis of international trade flows data in a different context. They propose independent modeling of binary and continuous networks where the each model is fitted sequentially at the time points. This strat-

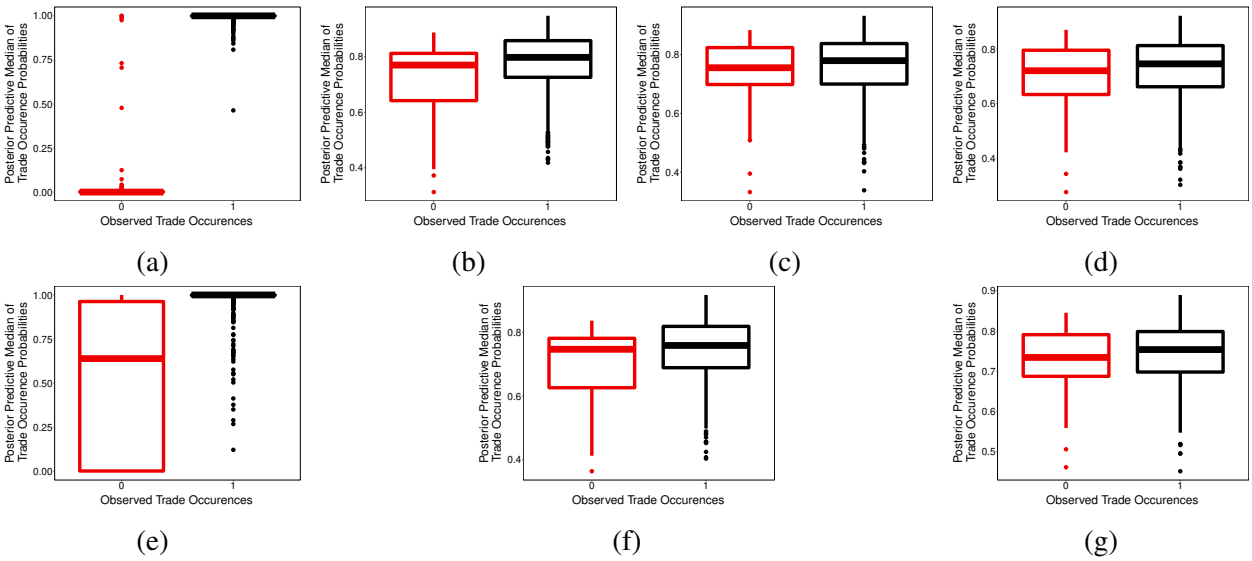


Figure 5.8: Comparison of posterior predictive median of trade occurrence probability and the observed trade occurrence in 2013 for all pairs of 29 countries. The figures show the performance of Hurdle-Net+Adaptive-DHS(1) for prefixed latent dimension K varied from 1 through 6. in each figure, 0 and 1 on x-axis refers to observed and unobserved trades among country pairs. y-axis denotes the posterior predictive median of trade occurrence probabilities for those pairs.

Table 5.2: Mean squared prediction errors in 2013 for different models with prefixed latent dimension $K = 5$.

Models	**Continuous prediction**	**Binary prediction**
Hurdle-Net+Adaptive-DHS(1)	0.97	0.007
Hurdle-Net+NonAdaptive-DHS(1)	1.36	0.10
Hurdle-Net+Adaptive-DHS(0)	2.41	0.10
Hurdle-Net+NonAdaptive-DHS(0)	3.62	0.12
Indep+Adaptive-DHS(1)	1.45	0.04
Hurdle-Net+HS	1.89	0.11
Hurdle-Net+NoLatent	1.87	0.11

egy is similar to Indep+Adaptive-DHS(1) that is considered here and it clearly overestimates the trade occurrences (Please see Figure 5.8). Hurdle-Net+HS on the other hand overestimates trade occurrence probabilities. This necessitates the dynamic modeling of a latent structure. Finally, the

predicted probabilities from Hurdle-Net+Adaptive-DHS(0), Hurdle-Net+NonAdaptive-DHS(0), and Hurdle-Net+NoLatent are predicting similar chance of an occurrence for observed and unobserved trades. This indicates the importance of modeling successive differences of latent positions. In fact the boxplots highlights that the performance without accounting for this is doing no better than Hurdle-Net+NoLatent where no latent dependence is considered in modeling. Finally, Table 5.2 shows the normalized MSPEs calculated for both binary and continuous prediction and it clearly shows the superiority in the performance of Hurdle-Net+Adaptive-DHS(1). For example, the MSPEs for the independent modeling has $1.5\times$ the error in continuous prediction and $5.7\times$ the error in binary prediction as compared to Hurdle-Net+Adaptive-DHS(1). This is consistent with our visual conclusion that we discussed above shows the superior performance of Hurdle-Net+Adaptive-DHS(1) and Hurdle-Net+NonAdaptive-DHS(1) over others. This goes to show that the data support the assumption of a strong latent dynamics and a common underlying mechanism in determining the behavior of the two networks.

5.4.3 Interpreting the Parameter estimates

Apart from providing good predictive performance, another equally desired goal in this research is to interpreting model parameters for statistical inference. In fact, two of these parameters are of primary interest: (a) the regression coefficient $\boldsymbol{\beta}$ which provides importance of country and country pair-specific covariates, and (b) the latent positions $(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T)$ which provides dynamics of 29 countries in explaining the inherent structure not explained by the covariates in the model. Following the performance comparison in the previous section, we now discuss the statistical inference of these parameters obtained using Hurdle-Net+Adaptive-DHS(1).

Figure 5.9 presents the boxplot of the posterior samples of the regression coefficients corresponding to the node- and pair-specific covariates. It shows that the GDP and area of the exporter, GDP of the importer, and the Labour Provision has a statistically significant positive effect in determining the international trade network. On the other hand, population of the exporter and distance between the capitals of two trading countries have a statistically significant negative effect in determining the trade pattern. Although, the population and the area of the importer show a positive

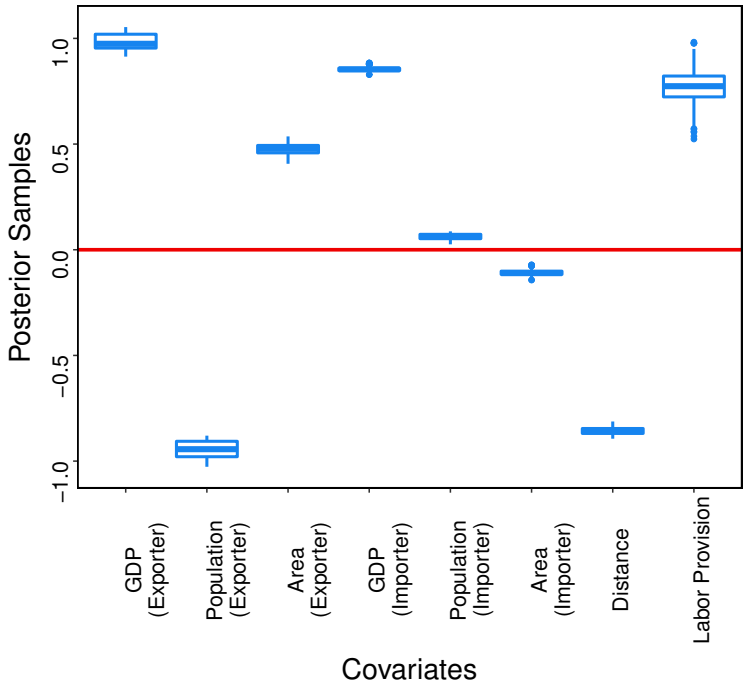


Figure 5.9: Boxplot of the posterior samples for the regression coefficient. The vertical axis represents the magnitude of each component of the regression coefficient and the horizontal axis shows the 8 covariates. For each component the figure shows the boxplot of the posterior samples. The horizontal solid red line denotes the line $y = 0$ and it denotes the absence of effect.

and negative effect, their magnitudes seem to be very small.

Figure 5.10–5.11 presents the inference of country specific latent variables $\{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_T\}$. At time point t, $\boldsymbol{Z}_t$ denotes the latent positions of 29 countries in the 5-dimensional latent Euclidean space as a 29×5 matrix. The rows correspond to countries and columns correspond to dimensions in the latent space. For better interpretation and representation, we order the columns of these matrices in decreasing order of their variances calculated combining all time points. This reorders the latent dimensions based on their importance in explaining the variability in the observed data. We also order the rows in an increasing order of the distance of latent positions at 1994, the first time point. Figure 5.10 provides heatmaps of these latent position matrices at the years 1994, 1998, 2002, 2006, 2010 and 2013 to show the dynamics of the latent positions (The heatmaps at all the time points are provides in the supplemental). The heatmaps from 1994 to 2010 are positions estimated from the training data. The heatmap at 2013 shows the predicted positions

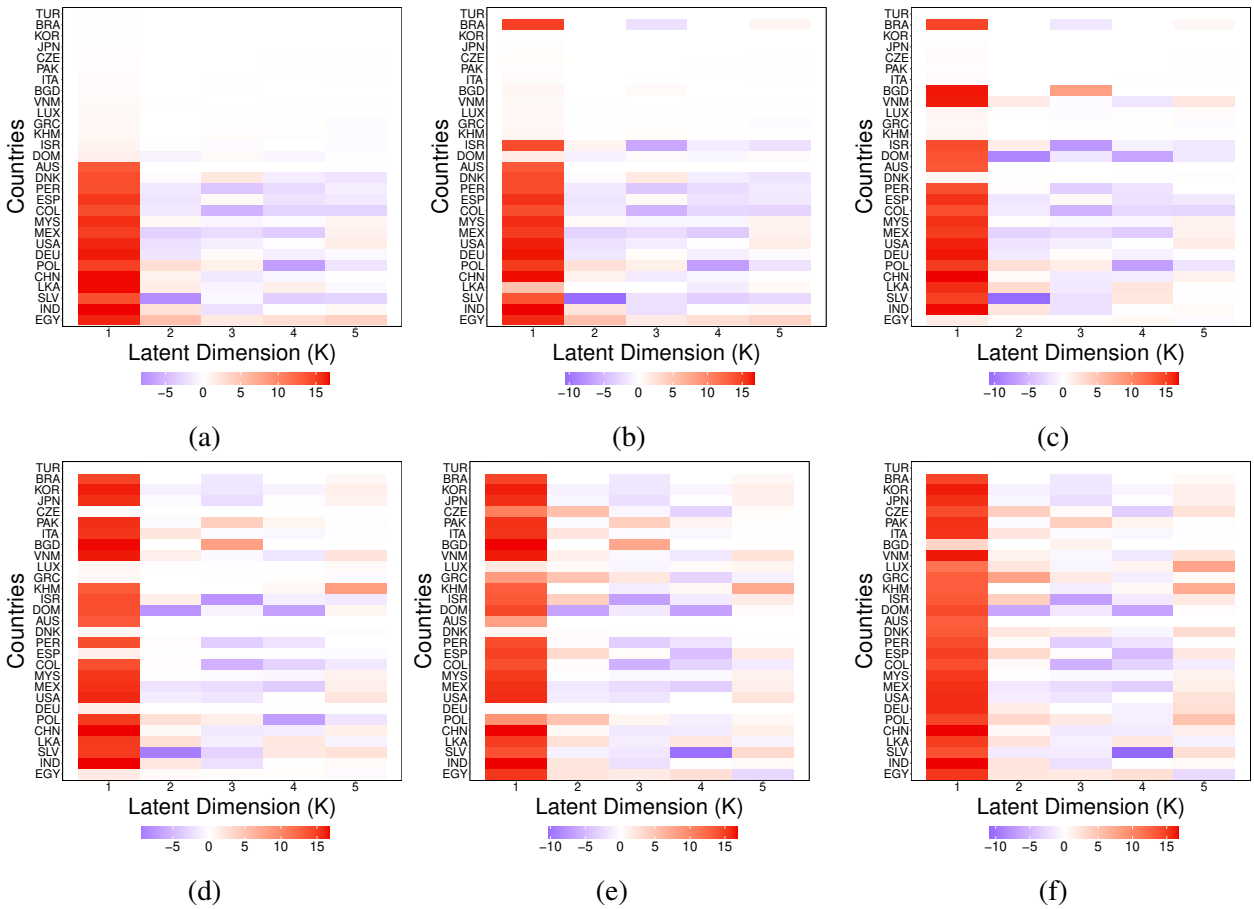
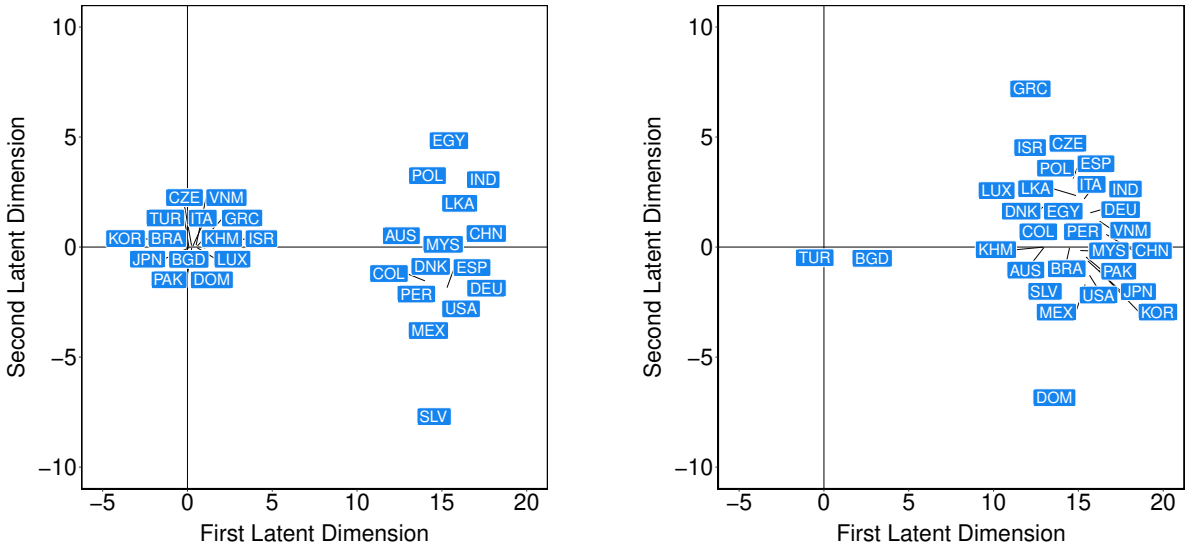


Figure 5.10: Heatmaps of latent positions $\boldsymbol{Z}_t$. Figures (a)–(e) correspond to the estimated positions at years 1994 1998 2002 2006, and 2010. Figure (f) shows the predicted latent positions of the countries in 2013. The columns are ordered based on the decreasing variance calculated combining all times points. The rows are ordered based on the increasing distance of countries from the origin in 1994, the first time point.

based on the fitted model. For estimation we use the posterior mean and the posterior median is used for prediction. Figure 5.10(a) very clearly shows two clusters of countries in the latent space and the clusters have about equal sizes. Before we interpret the dynamics, we note that the parameter α in (5.3) indicates the overall role of parent node in the binary and continuous networks. In this data, the parent and child nodes are interpreted as exporters and importers. The fitted Hurdle-Net+Adaptive-DHS(1) estimates α approximately 1. This means the exporter countries play the more significant role in deciding the presence or absence trades or the trade volumes. For interpreting the latent position we note that, with the estimate $\alpha \approx 1$, $L_{ijt} \approx \boldsymbol{z}_{it}^{\mathrm{T}} \boldsymbol{z}_{jt} / \|\boldsymbol{z}_{jt}\|$. For



(a) (b)

Figure 5.11: A scatterplot of estimated and predicted latent positions of 29 countries in 1994 and 2013, respectively. This plot is based on the first two latent dimensions from Figure 5.10 and shows the clusters of countries in the 2-dimensional latent Euclidean space that accounts for the first and second largest variance in latent contribution. Figure (a) on the left shows the estimated latent positions in 1994 and Figure (b) on the right shows the predicted latent positions in 2013.

a fixed pair of countries (i, j), the magnitude of the latent contribution is determined by $\boldsymbol{z}_{jt}$, the distance of the importer from the origin. The closer the importer is to the origin, the more is the latent contribution magnitude, and vice versa. This means the closer the countries are to the origin, the more the covariates fail to explain their log import volumes. In 1994, such is the case for the 14 countries corresponding to the first 14 rows (Please see Figrue 5.10(a)). This suggests in 1994, the covariates fails to explain the log import volumes of those countries when 15 countries in the other group trade with them. As we move forward in time, the included covariates starts to explain more. Finally, in 2013 only Turkey and Bangladesh are the only two countries left closest to the origin. This suggests the covariates still fail to explain the log import volumes of Turkey and Bangladesh when 27 countries in the other group trade with them.

5.5 Discussion

This chapter has explored the use of the dynamic shrinkage process prior on node-specific dynamic latent attributes in a dynamic network model. In many real-life applications we observe data on a fixed set of individuals over a time period. Because the dynamics involve the same set of individuals, a key interest in these applications is to take their network structure into account in the modeling. Often the continuous network that we observe has excessive zeros as observations. This can occur for a variety of reasons. For example, in microbiome data this occurs due to limitations of instruments used for a continuous measurement. Here zeros represent the measuring thresholds in those instruments. In this research we focused on bilateral trade data observed from the apparel industry. The observed zeros in this network means the absence of trades between country pairs. This highlights the importance the observed zeros and suggests taking this into account while modeling. To this, we re-frame this as data observed from a binary and a continuous network, and propose a Hurdle-Net model for modeling them jointly. Often there are many available covariates that aim to account for the variability in the observed data. But part of the goal is to figure out whether the available covariates are statistically significant. In cases when they are not, a latent contribution can be assumed in the modeling to account for the unexplained variance in the data. For modeling binary and continuous network data in presence of node- and pair-specific covariates, we assume node-specific latent attributes to account for any unexplained latent contribution. The latent terms in the model both bring in the network structure in the model and also account for the unexplained variance in the data. Using a latent space approach, we first independently model the two networks conditional on the latent attributes. Then we model the latent attributes using a dynamic shrinkage process apriori and independently across the nodes. All the model parameters and hyperparameters in the priors are continuous. This led us to take advantage of Stan and perform Bayesian inference using No-U-Turn Hamiltonian Monte Carlo sampling. Results presented in Section 5.3 shows that the method is quite robust and can prove to be significantly beneficial over the methods under comparison at times. Another advantage of the model is the interpretability of the parameters, particularly the regression coefficient and the

node-specific latent attributes. In application to the bilateral trade data, the results presented in Section 5.4 shows impressive performance over the other methods. It shows statistical significance of covariates that are realistic and covariate like the Labor Provision that is found relevant in the Economics literature [103]. The node-specific latent attributes shows clear cluster structure that accounts for the variance unexplained by the available covariates. The latent dynamic contribution helps us in making good prediction in the absence of all important covariates. Nonetheless, it begs to discover the covariates to account for the latent structure and the unexplained trades between countries.

6. SUMMARY OF THESIS

In this thesis we attempt to lay a foundation of efficiently choosing priors for two different problems: (1) designing Bayesian hypothesis tests for detecting presence or absence of hypothesized effects with lowered costs, and (2) efficient modeling of dynamic zero-inflated network data. Our contributions cover the computational and methodological aspects of this problem and touches upon theoretical aspects in some cases.

The costs of conducting experiments to test hypothesized effects is often related directly to the number of tested items or participants. In Chapter 2 we describe a modified sequential probability ratio test that can be used to reduce the average sample size required to perform statistical hypothesis tests at specified levels of significance and power. Examples are provided for z tests, t tests, and tests of binomial success probabilities. A description of a software package to implement the test designs is provided. We compare the sample sizes required in fixed design tests conducted at 5% significance levels to the average sample sizes required in sequential tests conducted at 0.5% significance levels, and we find that the two sample sizes are approximately equal.

Bayesian hypothesis testing procedures have gained increased acceptance in recent years. A key advantage that Bayesian tests have over classical testing procedures is their potential to quantify information in support of true null hypotheses. Ironically, default implementations of Bayesian tests prevent the accumulation of strong evidence in favor of true null hypotheses because associated default alternative hypotheses assign a high probability to data that are most consistent with a null effect. In Chapter 3–4 we propose the use of "non-local" alternative hypotheses to resolve this paradox. The resulting class of Bayesian hypothesis tests permits more rapid accumulation of evidence in favor of both true null hypotheses and alternative hypotheses that are compatible with standardized effect sizes of most interest in psychology. The prior used to define the alternative hypothesis in Chapter 2 is a special instance of this class of priors.

In the context of modeling zero-inflated directed networks Chapter 5 has proposed a Hurdle Network Model and explored the use of the dynamic shrinkage process prior on node-specific

dynamic latent attributes. In the model the latent terms both bring in the network structure and also account for the unexplained variance in the data. Using a latent space approach, we independently model the two networks conditional on the latent attributes. Another advantage of the model is the interpretability of the parameters, particularly the regression coefficient and the node-specific latent attributes. The latent dynamic contribution helps us in making good prediction in the absence of all important covariates.

REFERENCES

[1] J. Rouder, P. Speckman, D. Sun, and R. Morey, "Bayesian t tests for accepting and rejecting the null hypothesis," *Psychonomic Bulletin and Review*, vol. 16, pp. 225–237, 2009.

[2] F. Schönbrodt, E.-J. Wagenmakers, M. Zehetleitner, and M. Perugini, "Sequential hypothesis testing with bayes factors: Efficiently testing mean differences," *Psychological Methods*, vol. 22(2), pp. 322–339, 2017.

[3] M. Schnuerch and E. Erdfelder, "Controlling decision errors with minimal costs: The sequential probability ratio t-test," *Psychological Methods*, vol. 25, p. 206Ð226, 2020.

[4] V. Amrhein, S. Greenland, and B. McShane, "Scientists rise up against statistical significance," 2019.

[5] B. B. McShane, D. Gal, A. Gelman, C. Robert, and J. L. Tackett, "Abandon statistical significance," *The American Statistician*, vol. 73, no. sup1, pp. 235–245, 2019.

[6] H. Pike, "Statistical significance should be abandoned, say scientists," *BMJ*, vol. 364, 2019.

[7] V. Savalei and E. Dunn, "Is the call to abandon p-values the red herring of the replicability crisis?," *Frontiers in Psychology*, vol. 6, p. 245, 2015.

[8] Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, 2015.

[9] V. E. Johnson, R. D. Payne, T. Wang, A. Asher, and S. Mandal, "On the reproducibility of psychological science," *Journal of the American Statistical Association*, vol. 112, no. 517, pp. 1–10, 2017. PMID: 29861517.

[10] D. J. Benjamin, J. O. Berger, M. Johannesson, et al., and V. E. Johnson, "Redefine statistical significance," *Nature Human Behavior*, vol. 2, no. 1, pp. 6–10, 2018.

[11] V. E. Johnson, "Revised standards for statistical evidence," *Proceedings of the National Academy of Sciences*, vol. 110, no. 48, pp. 19313–19317, 2013.

[12] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, 1945.

[13] B. Chattopadhyay and K. Kelley, "Estimating the standardized mean difference with minimum risk: Maximizing accuracy and minimizing cost with sequential estimation," *Psychological Methods*, vol. 22, no. 1, pp. 94–113, 2016.

[14] K. Kelley, F. B. Darku, and B. Chattopadhyay, "Accuracy in parameter estimation for a general class of effect sizes: A sequential approach," *Psychological Methods*, vol. 23, no. 2, pp. 226–243, 2018.

[15] K. Kelley, F. B. Darku, and B. Chattopadhyay, "Sequential accuracy in parameter estimation for population correlation coefficients," *Psychological Methods*, vol. 24, no. 4, pp. 492–515, 2019.

[16] V. Johnson, "Uniformly most powerful Bayesian tests," *The Annals of Statistics*, vol. 41, no. 4, pp. 1716 – 1741, 2013.

[17] A. Nikooienejad and V. E. Johnson, "On the existence of uniformly most powerful bayesian tests with application to non-central chi-squared tests," *Bayesian Anal.*, 2020. Advance publication.

[18] D. Siegmund, *Sequential Analysis: Tests and confidence intervals*. Springer-Verlag New York, 1985.

[19] R. E. Kass and A. E. Raftery, "Bayes factors," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795, 1995.

[20] Z. Dienes, "Bayesian versus orthodox statistics: Which side are you on?," *Perspectives on Psychological Science*, vol. 6, pp. 274–290, 2011.

[21] A. Etz and J. Vandekerckhove, "Introduction to Bayesian inference for psychology," *Psychonomic Bulletin and Review*, vol. 25(1), pp. 5–34, 2018.

[22] J. Cover, M. Curd, and C. Pincock, *Philosophy of Science: The Central Issues, 2nd edition*. New York: W.W. Norton and Company, 2012.

[23] A. Dreber, T. Pfeiffer, J. Almenberg, S. Isaksson, B. Wilson, Y. Chen, B. A. Nosek, and M. Johannesson, "Using prediction markets to estimate the reproducibility of scientific research," *Proceedings of the National Academy of Sciences*, vol. 112, no. 50, pp. 15343–

15347, 2015.

[24] H. Jeffreys, *Theory of Probability*. New York: Oxford University Press, 1961.

[25] V. E. Johnson and D. Rossell, "On the use of non-local prior densities in bayesian hypothesis tests," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 72, no. 2, pp. 143–170, 2010.

[26] P. Sarkar and A. Moore, "Dynamic social network analysis using latent space models," in *Advances in Neural Information Processing Systems* (Y. Weiss, B. Schölkopf, and J. Platt, eds.), vol. 18, MIT Press, 2006.

[27] P. Sarkar, S. M. Siddiqi, and G. J. Gordon, "A latent space approach to dynamic embedding of co-occurrence data," in *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics* (M. Meila and X. Shen, eds.), vol. 2 of *Proceedings of Machine Learning Research*, (San Juan, Puerto Rico), pp. 420–427, PMLR, 21–24 Mar 2007.

[28] M. D. Ward, J. S. Ahlquist, and A. Rozenas, "Gravity's rainbow: A dynamic latent space model for the world trade network," *Network Science*, vol. 1, no. 1, pp. 95–118, 2013.

[29] D. R. Kowal, D. S. Matteson, and D. Ruppert, "Dynamic shrinkage processes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 81, no. 4, pp. 781–804, 2019.

[30] D. M. Oppenheimer and B. Monin, "The retrospective gambler's fallacy: Unlikely events, constructing the past, and multiple universes," *Judgment and Decision Making*, vol. 4, no. 5, p. 326, 2009.

[31] R. Klein, K. Ratliff, M. Vianello, R. Adams Jr, S. Bahník, M. Bernstein, *et al.*, "Investigating variation in replicability: a "many labs" replication project. open science framework," 2014.

[32] G. G. Kingsbury and D. J. Weiss, "A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure," in *New horizons in testing: Latent trait test theory and computerized adaptive testing* (D. J. Weiss, ed.), pp. 257–283, New York, NY: Academic Press, 1983.

[33] T. J. H. M. Eggen, "Item selection in adaptive testing with the sequential probability ratio test," *Applied Psychological Measurement*, vol. 23, no. 3, pp. 249–261, 2015.

[34] S. W. Nydick, "The sequential probability ratio test and binary item response models," *Journal of Educational and Behavioral Statistics*, vol. 39, no. 3, pp. 203–230, 2014.

[35] Y. I. Chang, "Application of sequential probability ratio test to computerized criterion-referenced testing," *Sequential Analysis*, vol. 23, no. 1, pp. 45–61, 2004.

[36] T. L. Lai, "Sequential analysis: Some classical problems and new challenges," *Statistica Sinica*, vol. 11, no. 2, pp. 303–351, 2001.

[37] T. L. Lai, "Likelihood ratio identities and their applications to sequential analysis," *Sequential Analysis*, vol. 23, no. 4, pp. 467–497, 2004.

[38] T. L. Lai, *Sequential Analysis*, pp. 1–6. American Cancer Society, 2008.

[39] J. Bartroff, M. Finkelman, and T. L. Lai, "Modern sequential analysis and its applications to computerized adaptive testing," *Psychometrika*, vol. 73, no. 3, pp. 473–486, 2008.

[40] S. Bar and J. Tabrikian, "A sequential framework for composite hypothesis testing," *IEEE Transactions on Signal Processing*, vol. 66, no. 20, pp. 5484–5499, 2018.

[41] A. Wald and J. Wolfowitz, "Optimum character of the sequential probability ratio test," *The Annals of Mathematical Statistics*, vol. 19, no. 3, pp. 326–339, 1948.

[42] T. W. Anderson, "A modification of the sequential probability ratio test to reduce the sample size," *The Annals of Mathematical Statistics*, vol. 31, no. 1, pp. 165–197, 1960.

[43] T. L. Lai, *Handbook of Sequential Analysis*, ch. Asymptotic optimality of generalized sequential likelihood ratio tests in some classical sequential testing procedures, pp. 121–144. Dekker: New York, 1991.

[44] M. Kulldorff, R. L. Davis, M. Kolczak†, E. Lewis, T. Lieu, and R. Platt, "A maximized sequential probability ratio test for drug and vaccine safety surveillance," *Sequential Analysis*, vol. 30, no. 1, pp. 58–78, 2011.

[45] M.-C. Shih, T. L. Lai, J. F. Heyse, and J. Chen, "Sequential generalized likelihood ratio tests for vaccine safety evaluation," *Statistics in Medicine*, vol. 29, no. 26, pp. 2698–2708, 2010.

[46] D. L. Demets and K. G. Lan, "Interim analysis: The alpha spending function approach," *Statistics in medicine*, vol. 13, no. 13-14, pp. 1341–1352, 1994.

[47] J. L. Haybittle, "Repeated assessment of results in clinical trials of cancer treatment," *British Journal of Radiology*, vol. 44, no. 526, pp. 793–797, 1971.

[48] C. Jennison and B. Turnbull, *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC Interdisciplinary Statistics, Boca Raton, FL: CRC Press, 1999.

[49] P. C. O'Brien and T. R. Fleming, "A multiple testing procedure for clinical trials," *Biometrics*, vol. 35, pp. 549–556, 1979.

[50] R. Peto, "Discussion of 'On the allocation of treatments in sequential medical trials' by J. A. Bather and 'The search for optimality in clinical trials' by P. Armitage," *International Statistical Review/Revue Internationale de Statistique*, vol. 53, no. 1, pp. 31–34, 1985.

[51] S. J. Pocock, "Group sequential methods in the design and analysis of clinical trials," *Biometrika*, vol. 64, no. 2, pp. 191–199, 1977.

[52] D. Siegmund, "Boundary crossing probabilities and statistical applications," *The Annals of Statistics*, vol. 14, no. 2, pp. 361–404, 1986.

[53] K. Anderson, "gsdesign: Group sequential design," *R package version*, pp. 2–9, 2014.

[54] S. Pramanik, V. Johnson, and A. Bhattacharya, "A modified sequential probability ratio test," *Journal of Mathematical Psychology*, vol. 101, 2021.

[55] G. Fechner, *Elements of Psychophysics*. New York: Holt, Rinehart & Winston, 1966.

[56] S. Stevens, "On the psychophysical law," *Psychological Review*, vol. 64, pp. 153–181, 1957.

[57] T. Augustin, "Stevens' power law and the problem of meaningfulness," *Acta Psychologica*, vol. 128, 2008.

[58] J. Cohen, *Statistical power analysis for the behavioral sciences, 2nd edition*. Hillsdale, N.J.: Erlbaum, 1988.

[59] A. Zellner and A. Siow, "Posterior odds ratio for selected regression hypotheses," in *Bayesian Statistics 1* (J. Bernardo, M. DeGroot, D. Lindley, and A. Smith, eds.), pp. 585–603, Valencia: University Press, 1980.

[60] A. Zellner and A. Siow, *Basic Issues in Econometrics*. Chicago: University of Chicago, 1986.

[61] J. Berger and L. Pericchi, "On the justification of default and intrinsic Bayes factors," in *Modelling and Prediction Honoring Seymour Geisser* (J. Lee, W. Johnson, and A. Zellner, eds.), pp. 173–204, New York: Springer, 1996.

[62] R. Morey and J. Rouder, "Bayesfactor: Computation of Bayes factors for common designs," in *https://CRAN.R-project.org/package=BayesFactor*, 2015.

[63] A. Stefan, F. Schönbrodt, N. Evans, and E.-J. Wagenmakers, "Efficiency in sequential testing: Comparing the sequential probability ratio test and sequential bayes factor." Unpublished paper, 2021.

[64] M. Bakker, A. van Dijk, and J. M. Wicherts, "The rules of the game called psychological science," *Perspectives on Psychological Science*, vol. 7(6), pp. 543–554, 2012.

[65] C. Anderson, J. Lindsay, and B. Bushman, "Research in the psychological laboratory," *Current Directions in Psychological Science*, vol. 8, pp. 3–9, 1999.

[66] J. Hall, "How big are nonverbal sex differences?," in *Sex differences and similarities in communication* (D. Canary and K. Dindia, eds.), pp. 155–177, Mahwah, N.J.: Erlbaum, 1998.

[67] M. Lipsey and D. Wilson, "The efficacy of psychological, educational and behavioral treatment: Confirmation from meta-analysis," *American Psychologist*, vol. 48, pp. 1181–1209, 1993.

[68] J. Meyer, S. Finn, L. Eyde, G. Kay, K. Moreland, R. Dies, others, and G. Reed, "Psychological testing and psychological assessment: A review of evidence and issues," *American Psychologist*, vol. 56, pp. 128–156, 2001.

[69] F. Richard, C. Bond, and J. Stokes-Zoota, "One hundred of years of social psychology quantitatively described," *Review of General Psychology*, vol. 7, pp. 331–363, 2003.

[70] R. Tett, J. Meyers, and N. Roese, "Applications of meta-analysis: 1987-1992," *International Review of Industrial and Organizational Psychology*, vol. 9, pp. 71–112, 1994.

[71] R. Morey, J. Rouder, T. Jamil, S. Urbanek, K. Forner, and A. Ly, *Package "BayesFactor"*. R Foundation for Statistical Computing, 2018.

[72] R. A. Klein, M. Vianello, F. Hasselman, B. G. Adams, J. Reginald B. Adams, S. Alper, M. Aveyard, J. R. Axt, M. T. Babalola, Štěpán Bahník, R. Batra, M. Berkics, M. J. Bernstein, D. R. Berry, O. Bialobrzeska, E. D. Binan, K. Bocian, M. J. Brandt, R. Busching, A. C. Rédei, H. Cai, F. Cambier, K. Cantarero, C. L. Carmichael, F. Ceric, J. Chandler, J.-H. Chang, A. Chatard, E. E. Chen, W. Cheong, D. C. Cicero, S. Coen, J. A. Coleman, B. Collisson, M. A. Conway, K. S. Corker, P. G. Curran, F. Cushman, Z. K. Dagona, I. Dalgar, A. D. Rosa, W. E. Davis, M. de Bruijn, L. D. Schutter, T. Devos, M. de Vries, C. Doğulu, N. Dozo, K. N. Dukes, Y. Dunham, K. Durrheim, C. R. Ebersole, J. E. Edlund, A. Eller, A. S. English, C. Finck, N. Frankowska, M. Ángel Freyre, M. Friedman, E. M. Galliani, J. C. Gandi, T. Ghoshal, S. R. Giessner, T. Gill, T. Gnambs, Ángel Gómez, R. González, J. Graham, J. E. Grahe, I. Grahek, E. G. T. Green, K. Hai, M. Haigh, E. L. Haines, M. P. Hall, M. E. Heffernan, J. A. Hicks, P. Houdek, J. R. Huntsinger, H. P. Huynh, H. IJzerman, Y. Inbar, Åse H. Innes-Ker, W. Jiménez-Leal, M.-S. John, J. A. Joy-Gaba, R. G. Kamiloğlu, H. B. Kappes, S. Karabati, H. Karick, V. N. Keller, A. Kende, N. Kervyn, G. Knežević, C. Kovacs, L. E. Krueger, G. Kurapov, J. Kurtz, D. Lakens, L. B. Lazarević, C. A. Levitan, J. Neil A. Lewis, S. Lins, N. P. Lipsey, J. E. Losee, E. Maassen, A. T. Maitner, W. Malingumu, R. K. Mallett, S. A. Marotta, J. Međedović, F. Mena-Pacheco, T. L. Milfont, W. L. Morris, S. C. Murphy, A. Myachykov, N. Neave, K. Neijenhuijs, A. J. Nelson, F. Neto, A. L. Nichols, A. Ocampo, S. L. O'Donnell, H. Oikawa, M. Oikawa, E. Ong, G. Orosz, M. Osowiecka, G. Packard, R. Pérez-Sánchez, B. Petrović, R. Pilati, B. Pinter, L. Podesta, G. Pogge, M. M. H. Pollmann, A. M. Rutchick, P. Saavedra, A. K. Saeri, E. Salomon, K. Schmidt, F. D. Schönbrodt, M. B. Sekerdej, D. Sirlopú, J. L. M. Skorinko, M. A. Smith, V. Smith-Castro, K. C. H. J. Smolders, A. Sobkow, W. Sowden, P. Spachtholz, M. Srivastava, T. G. Steiner, J. Stouten, C. N. H. Street, O. K. Sundfelt, S. Szeto, E. Szumowska, A. C. W. Tang, N. Tanzer, M. J. Tear, J. Theriault, M. Thomae, D. Torres, J. Traczyk, J. M. Tybur, A. Ujhelyi, R. C. M. van

Aert, M. A. L. M. van Assen, M. van der Hulst, P. A. M. van Lange, A. E. van 't Veer, A. Vásquez-Echeverría, L. A. Vaughn, A. Vázquez, L. D. Vega, C. Verniers, M. Verschoor, I. P. J. Voermans, M. A. Vranka, C. Welch, A. L. Wichman, L. A. Williams, M. Wood, J. A. Woodzicka, M. K. Wronska, L. Young, J. M. Zelenski, Z. Zhijia, and B. A. Nosek, "Many labs 2: Investigating variation in replicability across samples and settings," *Advances in Methods and Practices in Psychological Science*, vol. 1, no. 4, pp. 443–490, 2018.

[73] A. L. Alter, D. M. Oppenheimer, N. Epley, and R. N. Eyre, "Overcoming intuition: metacognitive difficulty activates analytic reasoning.," *Journal of Experimental Psychology: General*, vol. 136, no. 4, p. 569, 2007.

[74] D. Siegmund, *Sequential analysis: Tests and confidence intervals*. New York, NY: Springer Science & Business Media, 2013.

[75] J. Hajnal, "A two-sample sequential t-test," *Biometrika*, vol. 48, pp. 65–75, 1961.

[76] J. Tendeiro and H. Kiers, "A review of issues about null hypothesis Bayesian testing," *Psychological Methods*, vol. 24(6), pp. 774–795, 2019.

[77] D. van Ravenzwaaij and E.-J. Wagenmakers, "Advantages masquerading as "issues" in Bayesian hypothesis testing: A commentary on tendeiro and kiers (2019)." to appear in *Psychological Methods*, 2019.

[78] R. Bahadur, "Rates of convergence of estimates and test statistics," *Annals of Mathematical Statistics*, vol. 38(2), pp. 303–324, 1967.

[79] R. Bahadur and P. Bickel, "Asymptotic optimality of bayes' test statistics," tech. rep., The University of Chicago, 1967.

[80] M. Bayarri, J. Berger, A. Forte, and G. Garcia-Donato, "Criteria for Bayesian model choice with application to variable selection," *Annals of Statistics*, vol. 40(3), pp. 1550–1577, 2012.

[81] G. Consonni, D. Fouskakis, B. Liseo, and I. Ntzoufras, "Prior distributions for objective Bayesian analysis," *Bayesian Analysis*, vol. 13(2), pp. 627–679, 2018.

[82] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[83] R. E. Kass and S. K. Vaidyanathan, "Approximate bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 54, no. 1, pp. 129–144, 1992.

[84] E. Gunel and J. Dickey, "Bayes factors for independence in contingency tables," *Biometrika*, vol. 61, no. 3, pp. 545–557, 1974.

[85] F. Dablander, K. Huth, Q. F. Gronau, A. Etz, and E.-J. Wagenmakers, "A puzzle of proportions: Two popular bayesian tests can yield dramatically different conclusions," *Statistics in Medicine*, vol. n/a, no. n/a, 2021.

[86] T. Jamil, A. Ly, R. D. Morey, J. Love, M. Marsman, and E.-J. Wagenmakers, "Default "gunel and dickey"bayes factors for contingency tables," *Behavior Research Methods*, vol. 49, pp. 638–652, 2017.

[87] Q. F. Gronau, A. Raj K. N., and E.-J. Wagenmakers, "Informed bayesian inference for the a/b test," *Journal of Statistical Software*, vol. 100, no. 17, p. 1–39, 2021.

[88] S. Pramanik and V. Johnson, "Efficient alternatives for bayesian hypothesis tests in psychology," *Psychological Methods*, 2022.

[89] V. Johnson, "Bayes factors based on test statistics," *Journal of the Royal Statistical Society: Series B*, vol. 67, pp. 689–701, 2005.

[90] R. Hoekstra, R. Monden, D. van Ravenzwaaij, and E.-J. Wagenmakers, "Bayesian reanalysis of null results reported in medicine: Strong yet variable evidence for the absence of treatment effects," *PLOS ONE*, vol. 13, pp. 1–9, 04 2018.

[91] R. J. Fletcher, M. A. Acevedo, B. E. Reichert, K. E. Pias, and W. M. Kitchens, "Social network models predict movement and connectivity in ecological landscapes," *Proceedings of the National Academy of Sciences*, vol. 108, no. 48, pp. 19282–19287, 2011.

[92] T. R. Henry, D. Banks, D. Owens-Oas, and C. Chai, "Modeling community structure and topics in dynamic text networks," *Journal of Classification*, vol. 36, no. 2, pp. 322–349, 2019.

[93] P. D. Hoff, A. E. Raftery, and M. S. Handcock, "Latent space approaches to social network

analysis," *Journal of the American Statistical Association*, vol. 97, no. 460, pp. 1090–1098, 2002.

[94] P. D. Hoff and M. D. Ward, "Modeling dependencies in international relations networks," *Political Analysis*, vol. 12, no. 2, pp. 160–175, 2004.

[95] P. M. Krafft, J. Moore, B. A. Desmarais, and H. M. Wallach, "Topic-partitioned multinetwork embeddings.," in *NIPS*, pp. 2816–2824, 2012.

[96] M. D. Ward, R. M. Siverson, and X. Cao, "Disputes, democracies, and dependencies: A reexamination of the kantian peace," *American Journal of Political Science*, vol. 51, no. 3, pp. 583–601, 2007.

[97] B. Kim, K. H. Lee, L. Xue, and X. Niu, "A review of dynamic network models with latent variables," *Statistics Surveys*, vol. 12, no. none, pp. 105 – 135, 2018.

[98] J. R. Faulkner and V. N. Minin, "Locally adaptive smoothing with markov random fields and shrinkage priors," *Bayesian analysis*, vol. 13, no. 1, p. 225, 2018.

[99] S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky, "\ell_1 trend filtering," *SIAM review*, vol. 51, no. 2, pp. 339–360, 2009.

[100] R. J. Tibshirani, "Adaptive piecewise polynomial estimation via trend filtering," *The Annals of Statistics*, vol. 42, no. 1, pp. 285 – 323, 2014.

[101] C. M. CARVALHO, N. G. POLSON, and J. G. SCOTT, "The horseshoe estimator for sparse signals," *Biometrika*, vol. 97, no. 2, pp. 465–480, 2010.

[102] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," *Journal of Statistical Software, Articles*, vol. 76, no. 1, pp. 1–32, 2017.

[103] D. LeClercq, D. Samaan, and R. Robertson, "Labor provisions in trade agreements: Recasting the protectionist debate," *Available at SSRN 3668916*, 2020.

[104] R. Hankin, "The gauss hypergeometric function," in *https://CRAN.R-project.org/package=BayesFactor*, 2016.

[105] N. E. Korotkov and A. N. Korotkov, *Integrals Related to the Error Function*. CRC Press,

2020.

[106] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*. Academic press, 2014.

APPENDIX A

SUPPLEMENTARY MATERIAL: A MODIFIED SEQUENTIAL PROBABILITY RATIO TEST

A.1 Introduction

We consider null hypothesis significance tests (NHSTs) where the maximum number of samples (N) is specified and in which we wish to control Type I and Type II error probabilities at specified levels α and β, respectively.

Let X be a random variable having density $f(x;\theta)$ under both the null and alternative hypotheses, and let θ, $\theta \in \Theta$, denote the parameter of interest. Let $f(\mathbf{x}_n;\theta)$ denote the joint sampling density of the observation $\mathbf{x}_n = \{x_1, \ldots, x_n\}$ for some sample size n, and let $\pi_i(\theta)$ denote the prior density assigned to θ under H_i (for $i = 0, 1$). Then the marginal density $m_i(\mathbf{x}_n)$ of the data under H_i (for $i = 0, 1$) is defined as

$$m_i(\mathbf{x}_n) = \int_\Theta f(\mathbf{x}_n;\theta)\pi_i(\theta)d\theta. \tag{A.1}$$

For a given point alternative hypothesis $H_1 : \theta = \theta_1$, we define the likelihood ratio (LR) as

$$L(\theta_1, \theta_0; n) = \frac{f(\mathbf{x}_n;\theta_1)}{f(\mathbf{x}_n;\theta_0)}. \tag{A.2}$$

When there is no ambiguity regarding the values of (θ_0, θ_1), we simply write $L_n \equiv L(\theta_1, \theta_0; n)$. The Bayes factor (BF) in favor of H_1 is defined as $BF_{10}(\mathbf{x}_n) = m_1(\mathbf{x}_n)/m_0(\mathbf{x}_n)$.

Following [16], the uniformly most powerful Bayesian test (UMPBT) for evidence threshold $\delta > 0$ in favor of the alternative H_1 against a fixed null H_0, denoted by UMPBT(δ), is a Bayesian hypothesis test in which the Bayes factor for the test satisfies the following inequality for any $\theta_t \in \Theta$ and for all alternative hypotheses $H_2 : \theta \sim \pi_2(\theta)$:

$$\mathbf{P}_{\theta_t}[BF_{10}(\mathbf{x}) > \delta] \geq \mathbf{P}_{\theta_t}[BF_{20}(\mathbf{x}) > \delta]. \tag{A.3}$$

That is, the UMPBT maximizes the probability that the Bayes factor against a fixed null hypothesis exceeds a specified threshold. Following equation (A.3) for one-parameter exponential family models, the UMPBT alternative is defined as the alternative θ_1 which maximizes $\mathbf{P}_{\theta_t}[BF_{10}(\mathbf{x}) > \delta]$ among all prior densities on θ, $\theta \in \Theta$. A list of the UMPBT alternatives for common statistical tests can be found in the supporting information file of [11].

In tests of a simple null against a composite alternative, there is often a correspondence between the rejection regions of Bayesian testing rules using a UMPBT alternative and classical uniformly most powerful (UMP) tests (when such tests exist). Given a δ, the UMPBT(δ) alternative is optimal in the sense that it maximizes the probability that the Bayes factor in favor of the alternative exceeds a specified threshold δ. In such cases, δ can be determined by matching the rejection region of the test to that of the classical Neyman-Pearson UMP test of size α. This naturally induces a one-to-one correspondence between the size of the test (α) and the Bayesian evidence threshold (δ).

In the rest of the discussion, we refer to the UMP test as the fixed-design test.

A.2 The Modified Sequential Probability Ratio Test (MSPRT)

Given N, α, and β, suppose we are interested in testing a simple null against a one-sided alternative, i.e.,

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta > \theta_0 \quad \text{or} \quad \theta < \theta_0, \tag{A.4}$$

where θ is a scalar parameter defining $f(x;\theta)$. We further assume that $f(x;\theta)$ belongs to a one-parameter exponential family. Then, following the preceding discussion, we can obtain the UMPBT alternative by matching the UMPBT's rejection region to that of the fixed-design test using N samples. Doing so leads to the definition of the UMPBT alternative hypothesis and the evidence threshold. Once the alternative is determined, we can compute the likelihood ratio (or Bayes factor) in favor of the alternative as we observe data sequentially. For each n, let L_n denote the likelihood ratio as defined in equation (1) in the main article. As in the case of SPRTs, we define the acceptance and rejection threshold for L_n by $B = \dfrac{\beta}{1-\alpha}$ and $A = \dfrac{1-\beta}{\alpha}$, respectively. Using this notation, the conduct of the MSPRT can be defined by Algorithm 1.

Algorithm 1 : MSPRT

For $n = 1, \ldots, N$

1. **Stop** and **reject** H_0 **if** $L_n \geq A$
2. **Stop** and **accept** H_0 **if** $L_n \leq B$
3. **Collect the next** data point **if** $B < L_n < A$

If **no decision** has been made after collecting $\boldsymbol{N}$ observations, **terminate** the procedure and **reject** H_0 **if** $L_N \geq \gamma$; otherwise, **accept** H_0.

The threshold γ, which we refer to as the termination threshold, is chosen to be the smallest number that preserves the targeted size of the test α. In general, numerical procedures are required to determine the value of γ. We can implement this procedure using the `R` package `MSPRT`. A more detailed illustration for common tests is provided in Section A.2.

A.3 Examples

A.3.1 One-sample z test for a population mean

Suppose $X_1, \ldots, X_N$ are *i.i.d.* $N(\mu, \sigma^2)$ random variables, σ^2 is known, and we wish to test

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu > \mu_0. \tag{A.5}$$

Following [16], the UMPBT(δ) alternative is defined as

$$\mu_{1N} = \underset{\mu > \mu_0}{\arg\min} \left[\frac{\sigma^2 \log \delta}{N(\mu - \mu_0)} + \frac{(\mu + \mu_0)}{2} \right] = \mu_0 + \sigma \sqrt{\frac{2 \log \delta}{N}}. \tag{A.6}$$

By matching the rejection region from the UMPBT with that of the fixed-design test, we obtain the evidence threshold as

$$\delta = \exp\left(\frac{z_\alpha^2}{2}\right), \tag{A.7}$$

where z_α is the $100(1-\alpha)$th quantile of the standard normal distribution. Substituting this in

(A.6), we get the UMPBT alternative

$$\mu_{1N} = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{N}}. \tag{A.8}$$

The alternative corresponds to the rejection boundary for the fixed-design test of size α based on N observations.

Using the alternative, we compute L_n as

$$L_n = \frac{f(\mathbf{x}_n; \mu_{1N})}{f(\mathbf{x}_n; \mu_0)} = \exp\left[\frac{(\mu_{1N} - \mu_0)}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{n(\mu_{1N}^2 - \mu_0^2)}{2\sigma^2}\right]. \tag{A.9}$$

After γ is obtained, the MSPRT is then conducted according to Algorithm 1 in Section A.2.

A.3.2 One-sample t test for a population mean

Now suppose the conditions of Section A.3.1apply, but σ^2 is not known.

A UMPBT does not exist in this case. For this reason, we instead use the approximate data-dependent UMPBT(δ) alternative defined in [11] as

$$\mu_{1N} = \mu_0 + s_N \sqrt{\frac{\nu\delta^*}{N}} \tag{A.10}$$

where $s_N^2 = \dfrac{1}{N-1}\displaystyle\sum_{i=1}^{N}(x_i - \bar{x}_N)^2$, $\nu = N - 1$, and $\delta^* = \delta^{2/N} - 1$.

Based on the maximum sample size N, the condition for matching the rejection regions of the UMPBT and the fixed-design t test can be derived as

$$\sqrt{\nu\delta^*} = t_{\alpha;N-1}^2 \qquad \Leftrightarrow \qquad \delta = \left[\frac{t_{\alpha;N-1}^2}{\nu} + 1\right]^{\frac{N}{2}}, \tag{A.11}$$

where $t_{\alpha;N-1}$ is the $100(1-\alpha)$th quantile of a t distribution with degrees of freedom (df) $N - 1$.

From observed data, we obtain the UMPBT alternative at step n as

$$\mu_{1n} = \mu_0 + t_{\alpha;N-1}\frac{s_n}{\sqrt{N}}, \tag{A.12}$$

for $n = 2, ..., N$.

Using this alternative, we define the integrated likelihood function (or Bayes factor) L_n according to

$$L_n = \left[\frac{1 + \left(\frac{n}{n-1}\right)t_{0,n}^2}{1 + \left(\frac{n}{n-1}\right)t_{1,n}^2}\right]^{\frac{n}{2}}, \tag{A.13}$$

where $t_{0,n} = \dfrac{\bar{x}_n - \mu_0}{s_n}$ and $t_{1,n} = \dfrac{\bar{x}_n - \mu_{1n}}{s_n}$.

We obtained this integrated likelihood by imposing the noninformative prior $\pi(\sigma^2) \propto 1/\sigma^2$ on the unknown variance parameter.

Once γ is determined numerically, the MSPRT is conducted according to Algorithm 1 in Section A.2.

A.3.3 One-sample test for a binomial proportion

Suppose $X_1, \ldots, X_N$ represent *i.i.d.* Bernoulli observations with success probability p, and for some p_0 we wish to test

$$H_0 : p = p_0 \quad \text{vs.} \quad H_1 : p > p_0. \tag{A.14}$$

To design the MSPRT, we must determine the alternative hypothesis that will be used to compute L_n. We can accomplish this most easily by first examining the form of the fixed design test's rejection region. Based on the maximum sample size N, that test rejects H_0 if

$$\sum_{i=1}^{N} X_i > c_0, \tag{A.15}$$

where c_0 is defined as

$$c_0 = \inf\left\{c \middle| P_{H_0}\left(\sum_{i=1}^{N} x_i > c\right) \leq \alpha\right\}. \tag{A.16}$$

Following [16], the UMPBT(δ) alternative value of p is defined as

$$p_{1N}(\delta) = \underset{p>p_0}{\arg\min} \;\; h_N(p, \delta), \tag{A.17}$$

where

$$h_N(p, \delta) = \frac{\log \delta - N\left[\log(1-p) - \log(1-p_0)\right]}{\log\left(\frac{p}{1-p}\right) - \log\left(\frac{p_0}{1-p_0}\right)}. \tag{A.18}$$

For a given (p, δ), the rejection region for the UMPBT(δ) test is

$$\sum_{i=1}^{N} X_i > h_N(p, \delta). \tag{A.19}$$

Thus, the rejection region from the fixed-design test can be matched to that of the UMPBT by solving

$$h_N\Big(p_{1N}(\delta), \delta\Big) = c_0 \tag{A.20}$$

for δ. This solution provides the evidence threshold for the test.

In practice, the discrete nature of binomial data causes the Type I error of the test to be less than the targeted α. In order to achieve the exact α in a classical test, one must use a randomized test. The randomized test can be described as follows: with probability ψ, reject H_0 if $\sum_{i=1}^{N} x_i > (c_0 - 1)$, and with probability $(1 - \psi)$, reject H_0 if $\sum_{i=1}^{N} x_i > c_0$. The value of ψ is determined according to

$$\psi = \left[\alpha - P_{H_0}\left(\sum_{i=1}^{N} x_i > c_0\right)\right] \Big/ P_{H_0}\left(\sum_{i=1}^{N} x_i = c_0\right). \tag{A.21}$$

This suggests that we obtain the UMPBT alternative according to the following modification. Noting that the fixed-design randomized test involves two rejection regions, namely $\Big(c_0 - 1, N\Big]$

and $\Big(c_0, N\Big]$, and recalling (A.6), we solve

$$h_N\Big(p_{1N}(\delta_L), \delta_L\Big) = c_0 - 1 \quad \text{and} \quad h_N\Big(p_{1N}(\delta_U), \delta_U\Big) = c_0. \tag{A.22}$$

In contrast to z and t tests, using these values we define the UMPBT alternative as a mixture distribution of two points $p_{1N}(\delta_L) \equiv p_{1N,L}$ and $p_{1N}(\delta_U) \equiv p_{1N,U}$ with mixing probabilities ψ and $(1-\psi)$, respectively. Then we obtain L_n as a weighted likelihood function defined by

$$L_n = \psi \frac{f(\mathbf{x}_n; p_{1N,L})}{f(\mathbf{x}_n; p_0)} + (1-\psi)\frac{f(\mathbf{x}_n; p_{1N,U})}{f(\mathbf{x}_n; p_0)}, \tag{A.23}$$

where

$$\frac{f(\mathbf{x}_n; p)}{f(\mathbf{x}_n; p_0)} = \left[\frac{1-p}{1-p_0}\right]^n \left[\frac{p(1-p_0)}{p_0(1-p)}\right]^{\sum_{i=1}^n x_i}. \tag{A.24}$$

After γ has been numerically obtained, the MSPRT can be implemented using Algorithm 1 in Section A.2.

A.4 Examples with `MSPRT`: A user's guide

We have created an `R` package named `MSPRT` to implement the MSPRT conveniently. We illustrate the use of the test in the following examples. We assume throughout that `MSPRT` has been loaded into the `R` command environment.

A.4.1 Designing and implementing a MSPRT

A key function in the package is `design.MSPRT()`. Given N, α, β, and other parameters, this function finds the MSPRT. Recall from Algorithm 1 that finding the MSPRT requires finding the termination threshold γ. The function `design.MSPRT()` does this. It also provides an option (through the argument `theta1`) to find the performance of the resulting MSPRT at a user-defined point alternative.

A.4.1.1 One-sample z test for a population mean

Our first illustration of the MSPRT is for a simple z test. For concreteness, suppose we wish to test $H_0 : \mu = 3$ against the alternative hypothesis $H_1 : \mu > 3$ for a fixed $\sigma = 1.5$ with a maximum of $N = 30$ patients in a $\alpha = 0.5\%$ test with Type II error of approximately $\beta = 0.2$. There are two steps in the testing process: design and implementation.

In the design step, we calculate the termination threshold and the operating characteristics of the MSPRT. To do this, we use the functions `design.MSPRT()` and `OCandASN.MSPRT()`, respectively. The function `design.MSPRT()` is used to determine the termination threshold and evaluate the performance of the MSPRT when the null hypothesis is true. The required commands follow:

```
> design.out = design.MSPRT(test.type = "oneZ", theta0 = 3, sigma = 1.5,
                            N.max = 30)
> design.out$TypeI.attained    ## Type I error probability
[1] 0.005
> design.out$EN[1]    ## avg. sample size under the null
[1] 14.24063
> design.out$theta.UMPBT    ## UMPBT alternative
[1] 3.70542
> design.out$termination.threshold    ## termination threshold
[1] 27.911
```

In this code snippet, the values `TypeI.attained`, `EN[1]`, and `termination.threshold` represent the Type I error probability, the average sample size required for reaching a decision when the null hypothesis is true, and the termination threshold of the MSPRT, respectively.

Normally, we must find the operating characteristics of the test at several alternative values. For the UMPBT alternative (equal to 3.7054 in this case), these values can be obtained by giving the following command.

```
> OC.out = OCandASN.MSPRT(theta = 3.7054,
                          design.MSPRT.object = design.out)
> OC.out$acceptH0.prob   ##Type II error at the UMPBT alternative
[1] 0.509086
> OC.out$EN   ##avg. sample size at the UMPBT alternative
[1] 25.29154
```

The values returned from this function call include (but are not restricted to) `acceptH0.prob` and `EN`. They are interpreted as the Type II error probability and the average sample size required by the designed MSPRT for reaching a decision when the UMPBT alternative is true, respectively.

Finally, it may be necessary to obtain the operating characteristics at arbitrary values of the alternative hypothesis. Again for concreteness, suppose we wish to determine the operating characteristics for $\mu = 4$ (for example). Then the following command may be given.

```
> OC.out = OCandASN.MSPRT(theta = 4, design.MSPRT.object = design.out)
> OC.out$acceptH0.prob   ##Type II error at the the desired alternative
[1] 0.151229
> OC.out$EN   ##avg. sample size at the desired alternative
[1] 22.67337
```

The output from this command may be interpreted as before.

Next, in the implementation phase we can apply the test to a sequence of observed values. To illustrate this procedure, we simulate the observed values as follows:

```
> set.seed(1)
> x = rnorm(n = 30, mean = 5, sd = 1.5)
```

Given these values, the MSPRT stopping criteria can be tested with the command `implement.MSPRT()`. Note that the object `design.out` is obtained using the `design.MSPRT()` command as above.

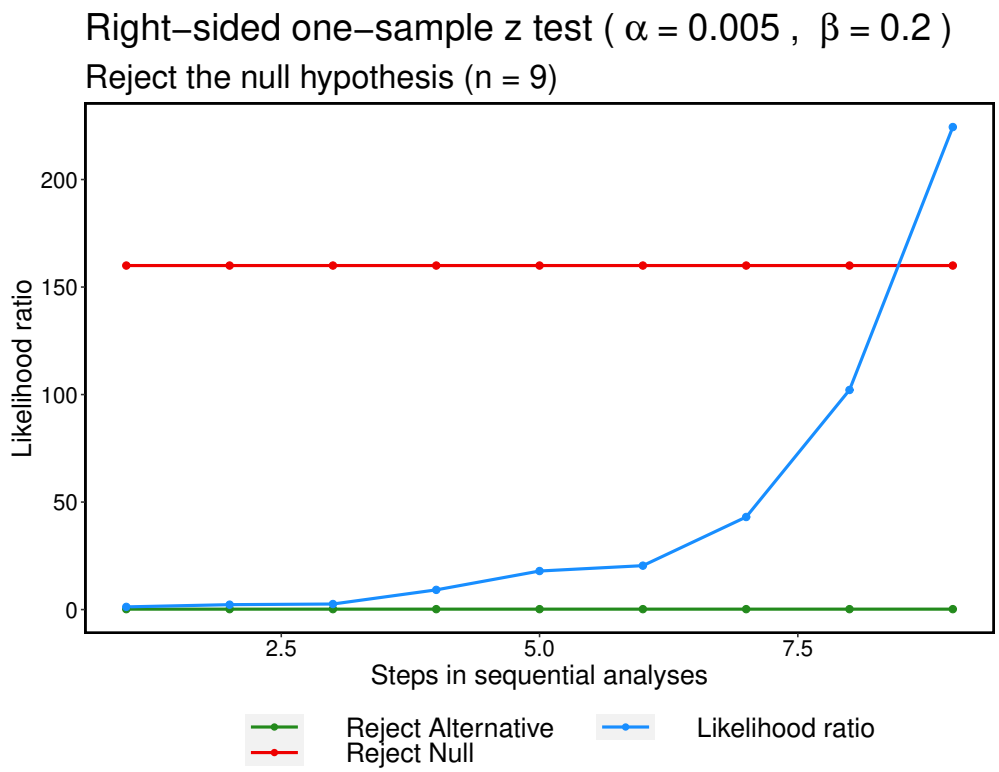


Figure A.1: One-sample z test that a population mean equals 3. Hypothesis test of $H_0 : \mu = 3$ vs. $H_1 : \mu > 3$ with σ known to be 1.5. Sequential comparison plot of the MSPRT obtained in Section A.4.1.1.

```
> implement.out = implement.MSPRT(obs = x,
                                  design.MSPRT.object = design.out)
> implement.out$decision    ##decision
[1]  "reject.null"
> implement.out$n    ##number of observations required to reach
                     ## the decision
[1]  9
```

This output shows that the null hypothesis is rejected after the 9^{th} observation.

If `plot.it = 2` (the default), the call to `implement.MSPRT()` also returns a sequential comparison plot similar to that depicted in Figure A.1. This particular plot shows that L_n crosses the “reject null" threshold on the 9^{th} observation, at which time the null hypothesis is rejected.

A.4.1.2 One-sample t test for a population mean

Our next illustration of the MSPRT is for a *t* test. For concreteness, suppose we again wish to test $H_0 : \mu = 3$ against an alternative hypothesis $H_1 : \mu > 3$ for an unknown σ with a maximum of $N = 30$ patients in a $\alpha = 0.5\%$ test with Type II error of approximately $\beta = 0.2$. Again there are two steps in the testing process: design and implementation.

In the design step, we calculate the termination threshold and the operating characteristics of the MSPRT. To do this, we again use the functions `design.MSPRT()` and `OCandASN.MSPRT()`, respectively. The function `design.MSPRT()` is used to determine the termination threshold and evaluate the performance of the MSPRT when the null hypothesis is true. The required commands follow:

```
> design.out = design.MSPRT(test.type = "oneT", theta0 = 3, N.max = 30)
> design.out$TypeI.attained    ## Type I error probability
[1] 0.005
> design.out$EN[1]    ## avg. sample size under the null
[1] 14.60748
> design.out$termination.threshold    ## termination threshold
[1] 34.02
```

The values `TypeI.attained`, `EN[1]`, and `termination.threshold` can be interpreted as before.

Once we have obtained the MSPRT design, it may be necessary to obtain the operating characteristics of the test at arbitrary values of the alternative hypothesis. Again for concreteness, suppose we wish to determine the operating characteristics for $\mu = 4$. We can do that by using the following command.

```
> OC.out = OCandASN.MSPRT(theta = 4, design.MSPRT.object = design.out)
> OC.out$acceptH0.prob   ##Type II error at the the desired alternative
[1] 0.006113
```

```
> OC.out$EN   ##avg. sample size at the desired alternative
[1] 22.39615
```

The values can be interpreted as in the previous section.

Next, in the implementation phase we can apply the test to a sequence of observed values. To illustrate this procedure, we use the same `x` as in Section A.4.1.1:

```
> set.seed(1)
> x = rnorm(n = 30, mean = 5, sd = 1.5)
```

Given these values, the MSPRT stopping criteria can be tested with the command `implement.MSPRT()`. Note that the object `design.out` is obtained using the `design.MSPRT()` command as above.

```
> implement.out = implement.MSPRT(obs = x, design.MSPRT.object = design.out)
> implement.out$decision   ##decision
[1]  "reject.null"
> implement.out$n   ##number of observations required to reach decision
[1]  22
```

Output from these commands shows that the null hypothesis is rejected after the 22^{nd} observation.

If `plot.it = 2` (the default), the call to `implement.MSPRT()` also returns a sequential comparison plot similar to that depicted in Figure A.2. This particular plot show that L_n crosses the "reject null" threshold on the 22^{nd} observation, at which time the null hypothesis is rejected.

A.4.1.3 One-sample test of a binomial proportion

We next consider the MSPRT for a proportion test. For concreteness, suppose we wish to test $H_0 : p = 0.2$ against the alternative hypothesis $H_1 : p > 0.2$ with a maximum of $N = 30$ patients in a $\alpha = 0.5\%$ test with Type II error of approximately $\beta = 0.2$. Again we go through the two steps in the testing process: design and implementation.

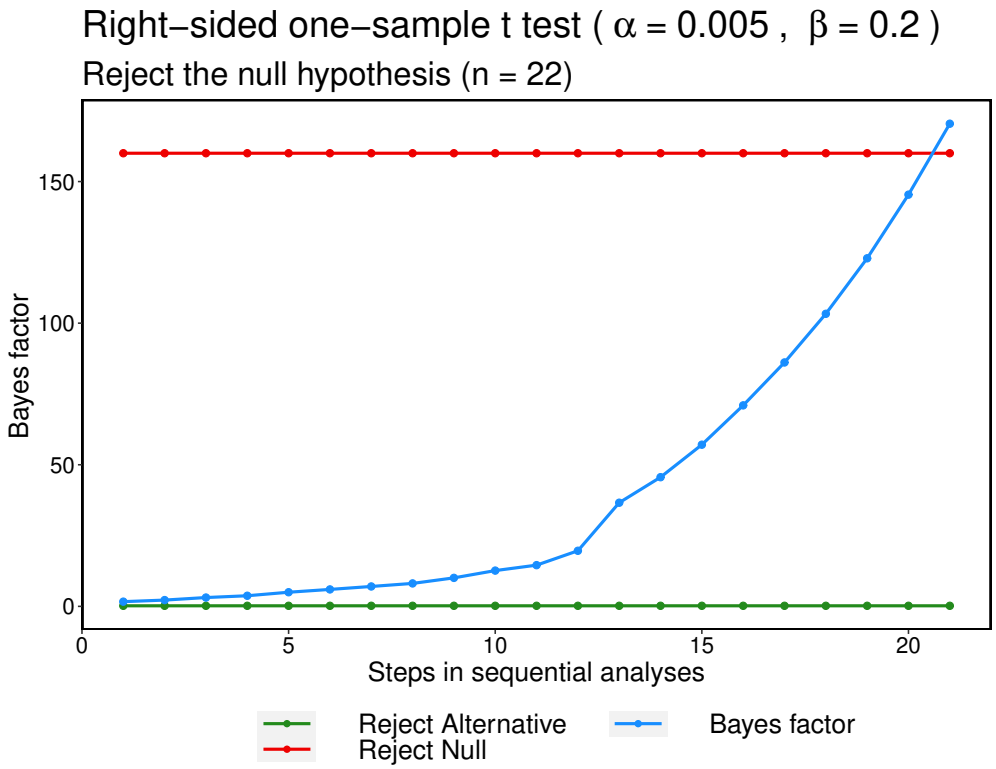


Figure A.2: One-sample *t* test that a population mean equals 3. Hypothesis test of $H_0 : \mu = 3$ vs. $H_1 : \mu > 3$ when σ is assumed unknown. Sequential comparison plot of the MSPRT obtained in Section A.4.1.2.

In the design step, we calculate the termination threshold and the operating characteristics of the MSPRT. To do this, we again use functions `design.MSPRT()` and `OCandASN.MSPRT()`, respectively. The commands follow:

```
> design.out = design.MSPRT(test.type = "oneProp", theta0 = 0.2, N.max = 30)
> design.out$Type1.attained   ##Type I error probability
[1] 0.002946
> design.out$EN[1]   ##avg. sample size under the null
[1] 12.9514
> design.out$UMPBT$theta   ##two points of the UMPBT alternative
[1] 0.3666727 0.4000178
> design.out$UMPBT$mix.prob   ##mixing probability for the UMPBT alternative
[1] 0.2959777 0.7040223
> design.out$termination.threshold   ##termination threshold
[1] 13.21
```

The values `TypeI.attained`, `EN[1]`, and `termination.threshold` can be interpreted as before. The values of `UMPBT$theta` and `UMPBT$mix.prob` together specify the UMPBT alternative used by the MSPRT. In this case the alternative is 0.3667 and 0.4 with approximate probabilities 0.296 and 0.704, respectively.

Once we have the MSPRT design, we can use `OCandASN.MSPRT()` to compute the operating characteristics of that MSPRT. For concreteness, suppose we wish to determine the operating characteristics for $p = 0.3$. The following commands do this.

```
> OC.out = OCandASN.MSPRT(theta = 0.3, design.MSPRT.object = design.out)
> OC.out$acceptH0.prob   ##Type II error at the the desired alternative
[1] 0.920718
> OC.out$EN   ##avg. sample size at the desired alternative
[1] 20.1515
```

The values returned from this function call have the same interpretation as before.

Next, in the implementation phase we can apply the test to a sequence of observed values. To illustrate this procedure, we simulate the observed binary values as follows:

```
> set.seed(1)
> x = rbinom(n = 30, size = 1, prob = 0.2)
```

Given these values, the MSPRT stopping criteria can be tested with the command `implement.MSPRT()`. Note that the object `design.out` is obtained using the `design.MSPRT()` command as above.

```
> implement.out = implement.MSPRT(obs = x, design.MSPRT.object = design.out)
> implement.out$decision   ##decision
[1] "reject.alt"
> implement.out$n   ##number of observations required to reach decision
[1] 15
```

This output shows that the alternative hypothesis is rejected after using the 15^{th} observation. In particular, the sequential test plot in Figure A.3 shows the sequential trajectory of L_n until the alternative hypothesis is rejected.

A.4.1.4 Two-sample z test for a difference in two population means

Let, μ_1 and μ_2 be the population means of two groups of patients, respectively. Suppose we want to test $H_0 : \mu_1 - \mu_2 = 0$ against the alternative hypothesis $H_1 : \mu_1 - \mu_2 > 0$ for a known common population variance of $\sigma = 1.5$. Assume that we can observe a maximum of 30 patients from each group (that is, $N_1 = N_2 = 30$). We set $\alpha = 0.5\%$ and the Type II error level $\beta = 0.2$.

In the design step, we calculate the termination threshold and the operating characteristics of the MSPRT. As before, the function `design.MSPRT()` is used to determine the termination threshold and evaluate the performance of the MSPRT when the null hypothesis is true. The required commands are as follows:

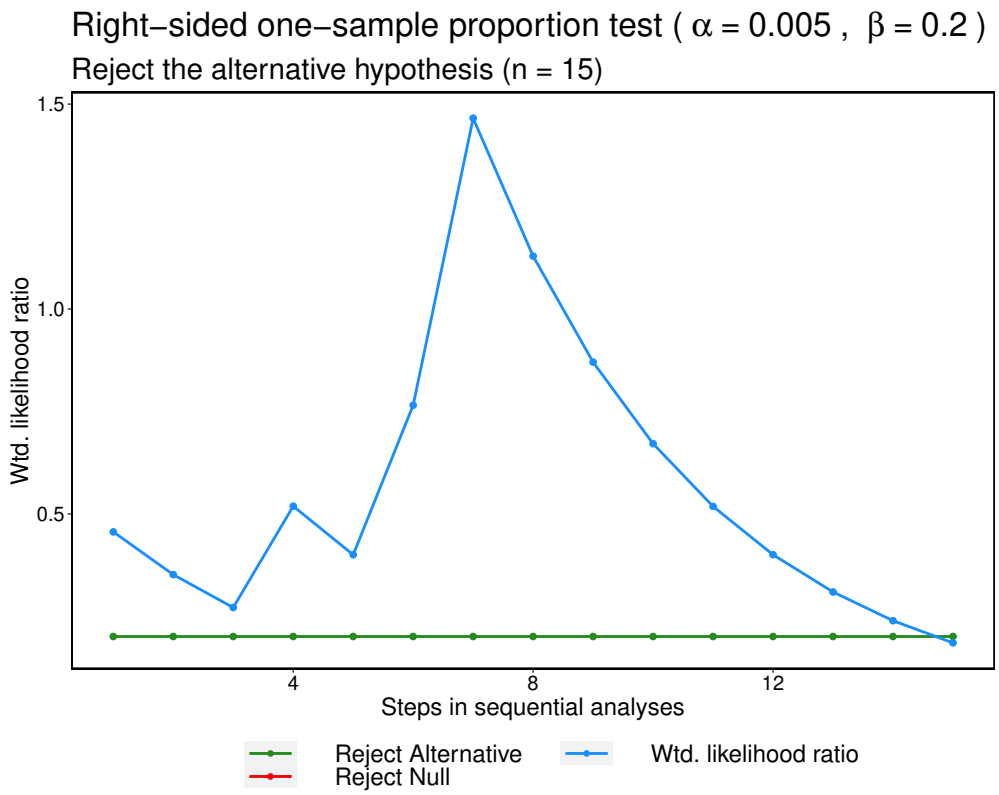


Figure A.3: One-sample test that a binomial proportion equals 0.2. Hypothesis test of $H_0 : p = 0.2$ vs. $H_1 : p > 0.2$. Sequential comparison plot of the MSPRT as in Section A.4.1.3.

```
> design.out = design.MSPRT(test.type = "twoZ", sigma1 = 1.5,
                            sigma2 = 1.5, N1.max = 30, N2.max = 30)
> design.out$Type1.attained   ##Type 1 error probability
[1] 0.005
> design.out$EN$H0
$Group1   ##avg. sample size from Group 1 under the null
[1] 14.22938

$Group2   ##avg. sample size from Group 2 under the null
[1] 14.22938
> design.out$theta.UMPBT   ##UMPBT alternative
[1] 0.9976144
> design.out$termination.threshold   ##termination threshold
[1] 27.885
```

In this code snippet, the values `TypeI.attained`, `EN$H0`, and `termination.threshold` respectively represent the Type I error probability, the average sample size required from Group 1 and 2 under the null hypothesis, and the termination threshold of the designed MSPRT.

Normally, we must also find the operating characteristics of the test at several alternative values. For the UMPBT alternative (equal to 0.9976 in this case), these values can be obtained by giving the following command.

```
> OC.out = OCandASN.MSPRT(theta = 0.9976144,
                          design.MSPRT.object = design.out)
> OC.OC.out$acceptH0.prob   ##Type II error at the UMPBT alternative
[1] 0.509531
> OC.out$EN1   ##avg. sample size from Group 1 at the UMPBT alternative
[1] 25.31669
> OC.out$EN2   ##avg. sample size from Group 2 at the UMPBT alternative
[1] 25.31669
```

The values returned from this function call include `theta`, `acceptH0.prob`, `EN1`, and `EN1`. They are interpreted as the effect size where the performace is evaluated, the Type II error probability, the average sample size required from Group 1 at the UMPBT alternative, and the average sample size required from Group 2 at the UMPBT alternative, respectively.

To obtain the operating characteristics at arbitrary values of the alternative hypothesis, suppose we wish to determine the operating characteristics for $\mu_1 - \mu_2 = 2$. Then the following command may be given.

```
> OC.out = OCandASN.MSPRT(theta = 2, design.MSPRT.object = design.out)
> OC.out$acceptH0.prob   ##Type II error at the desired alternative
[1] 0.007961
> OC.out$EN1   ##avg. sample size from Group 1 at the desired alternative
[1] 16.17953
> OC.out$EN2   ##avg. sample size from Group 2 at the desired alternative
[1] 16.17953
```

The output from this command may be interpreted as before.

Next, in the implementation phase we can apply the test to two sequences of observed values from both groups. To illustrate this procedure, suppose that we simulate the observed values from Group 1 and 2 as follows:

```
> set.seed(1)
> x1 = rnorm(n = 30, mean = 0.998, sd = 1.5)
> x2 = rnorm(n = 30, mean = 0, sd = 1.5)
```

Given these values, the MSPRT stopping criteria can be tested with the command `implement.MSPRT()`. Note that the object `design.out` is obtained using the `design.MSPRT()` command as above.

```
> implement.out = implement.MSPRT(obs1 = x1, obs2 = x2,
                                  design.MSPRT.object = design.out)
```

```
> implement.out$decision   ##decision
[1] "reject.alt"
> implement.out$n1   ##number of observations required from Group 1
[1] 30
> implement.out$n2   ##number of observations required from Group 2
[1] 30
```

This output shows that the alternative hypothesis is rejected after using the maximum number of available samples from each group.

If `plot.it = 2` (the default), the call to `implement.MSPRT()` also returns a sequential comparison plot similar to that depicted in Figure A.4. This particular plot shows that L_n reaches $N = 30$ without reaching a decision. But the likelihood ratio at the maximum sample size is approximately $L_{30} = 16.74$ (stored in `implement.out$LR`), which is below the termination threshold 27.885. So the test rejects the alternative after observing 30 samples from each group.

A.4.1.5 Two-sample t test for a difference in two population means

Assume the exact setup as in Section A.4.1.4, and suppose we want to test $H_0 : \mu_1 - \mu_2 = 0$ against $H_1 : \mu_1 - \mu_2 > 0$, but the common population variance is unknown.

In the design step, we calculate the termination threshold and the operating characteristics of the MSPRT. The required commands follow:

```
> design.out = design.MSPRT(test.type = "twoT", N1.max = 30, N2.max = 30)
> design.out$Type1.attained   ##Type 1 error probability
[1] 0.005
> design.out$EN$H0
$Group1   ##avg. sample size from Group 1 under the null
[1] 13.93484
$Group2   ##avg. sample size from Group 2 under the null
[1] 13.93484
```

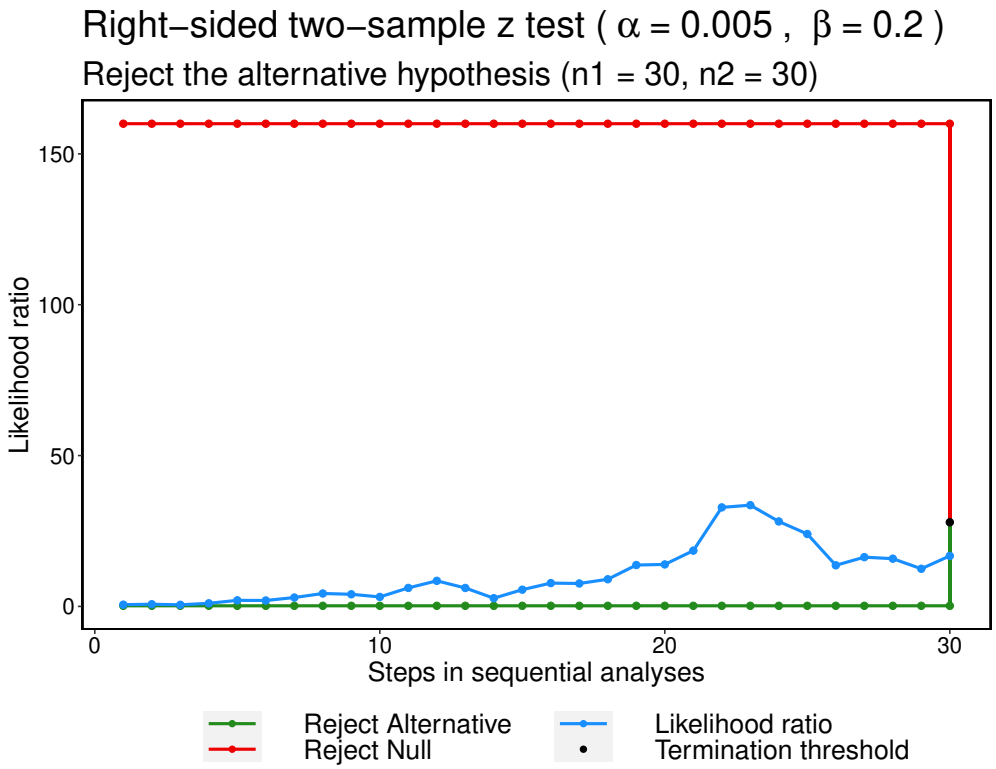


Figure A.4: Two-sample z test that the difference in population means is 0. Hypothesis test of $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_1 : \mu_1 - \mu_2 > 0$ with known common population standard deviation 1.5. Sequential comparison plot of the MSPRT obtained in Section A.4.1.4.

```
> design.out$termination.threshold   ##termination threshold
[1] 33.243
```

In this code snippet, the values `Type1.attained`, `EN$H0`, and `termination.threshold` represent the Type I error probability, the average sample size required from each group under the null hypothesis, and the termination threshold of the designed MSPRT, respectively.

To obtain the operating characteristics at arbitrary values of the alternative hypothesis, say, $\mu_1 - \mu_2 = 2$, the following command may be given.

```
> OC.out = OCandASN.MSPRT(theta = 2, design.MSPRT.object = design.out)
> OC.out$acceptH0.prob   ##Type II error at the UMPBT alternative
[1] 4.9e-05
> OC.out$EN1   ##avg. sample size from Group 1 at the desired alternative
[1] 15.61961
> OC.out$EN2   ##avg. sample size from Group 2 at the desired alternative
[1] 15.61961
```

The output from this command may be interpreted as before.

Next, in the implementation phase we can apply the test to two sequences of observed values from both groups. To illustrate this procedure, we use the same `x1` and `x2` as in Section A.4.1.4:

```
> set.seed(1)
> x1 = rnorm(n = 30, mean = 0.998, sd = 1.5)
> x2 = rnorm(n = 30, mean = 0, sd = 1.5)
```

Given these values, the MSPRT stopping criteria can be tested with the command `implement.MSPRT()`. Note that the value of `termination.threshold` is obtained using the `design.MSPRT()` command above.

```
> implement.out = implement.MSPRT(obs1 = x1, obs2 = x2,
```

```
                                    design.MSPRT.object = design.out)
> implement.out$decision   ##decision
[1] "reject.null"
> implement.out$n1   ##number of observations required from Group 1
[1] 30
> implement.out$n2   ##number of observations required from Group 2
[1] 30
```

This output shows that the null hypothesis is rejected after observing the maximum available number of 30 patients from each group.

If `plot.it = 2` (the default), the call to `implement.MSPRT()` also returns a sequential comparison plot similar to that depicted in Figure A.4.1.5. This particular plot shows that L_n reaches $N = 30$ without reaching a decision. But the likelihood ratio at the maximum sample size is approximately $L_{30} = 40.615$ (stored in `implement.out$LR`), which is above the termination threshold 33.243. So the test rejects the null after observing 30 samples from each group.

A.4.2 Results from simulation studies

In this section we describe in more detail the simulation results from the main article. We examine one-sample tests for a binomial proportion, *z* tests and *t* tests of size $\alpha = 0.05$ and 0.005. For simplicity, we examine one-sided tests with alternative hypotheses of the form $H_1 : \theta > \theta_0$. We also assume that the targeted power of the test is 80% (i.e., $\beta = 0.2$). Two-sided tests, tests of alternative hypotheses of the form $H_1 : \theta < \theta_0$, and tests with different Type I or Type II errors are handled similarly. We compare the MSPRTs to standard fixed-design tests having the same size α, sample size N, and Type II error $\beta = 0.2$. Given N and α for fixed-design tests, we define θ_a, the fixed-design alternative, as the alternative parameter value (effect size) that provides the specified β.

Figures A.6 through A.8 display the average proportion of the fixed-design sample size N needed in a MSPRT to achieve nearly equivalent Type I and Type II errors. In all plots, Type I errors are maintained. The subplots on the right depict that average power achieved at the corresponding

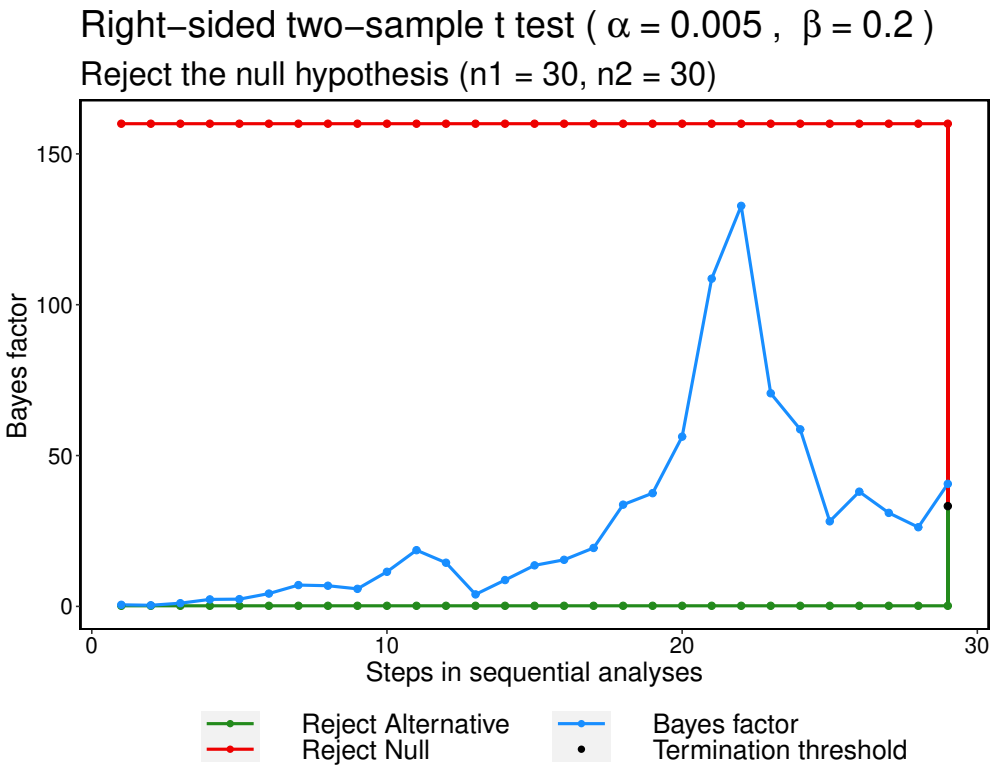


Figure A.5: Two-sample t test that the difference in population means is 0. Hypothesis test of $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_1 : \mu_1 - \mu_2 > 0$ with unknown common population standard deviation. Sequential comparison plot of the MSPRT obtained in Section A.4.1.5.

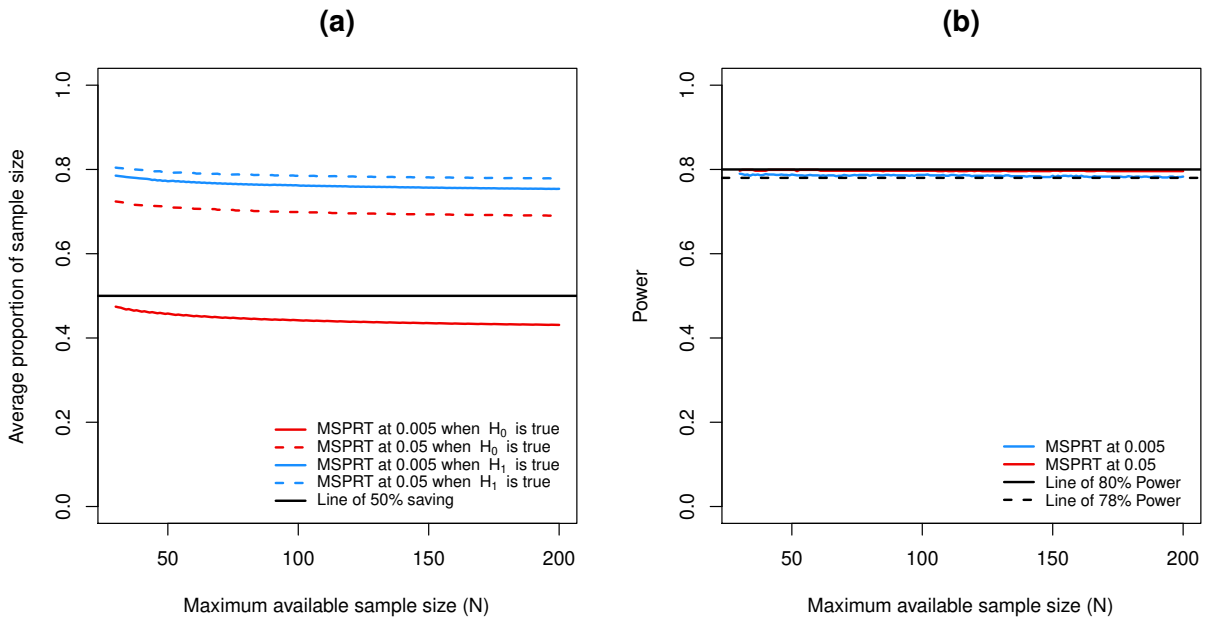


Figure A.6: One-sample z test that a population mean equals 0. Hypothesis test of $H_0 : \theta = 0$ vs. $H_1 : \theta > 0$. The curves in the left plot represent the average proportion of the maximum sample size (N) used before the MSPRT terminates in favor of the null or alternative hypothesis. The plot on the right displays the average power of the test against its targeted value of 0.8. In both plots, the operating characteristics under the alternative are evaluated at the corresponding fixed-design alternatives.

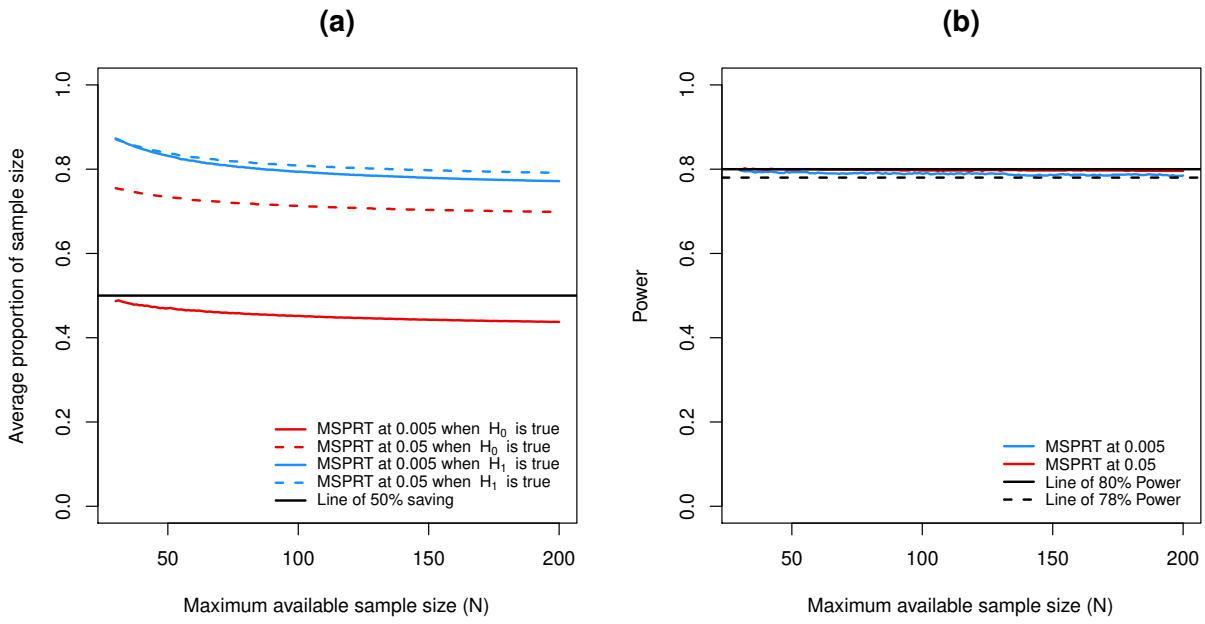


Figure A.7: One-sample t test that a population mean is 0. Hypothesis test of $H_0 : \theta = 0$ vs. $H_1 : \theta > 0$. In contrast to Figure A.6, the population standard deviation is assumed to be unknown. The curves in the left plot represent the average proportion of the maximum sample size (N) used before the MSPRT terminates in favor of the null or alternative hypothesis. The plot on the right displays the average power of the test against its targeted value of 0.8. In both plots, the operating characteristics under the alternative are evaluated at the corresponding fixed-design point alternatives.

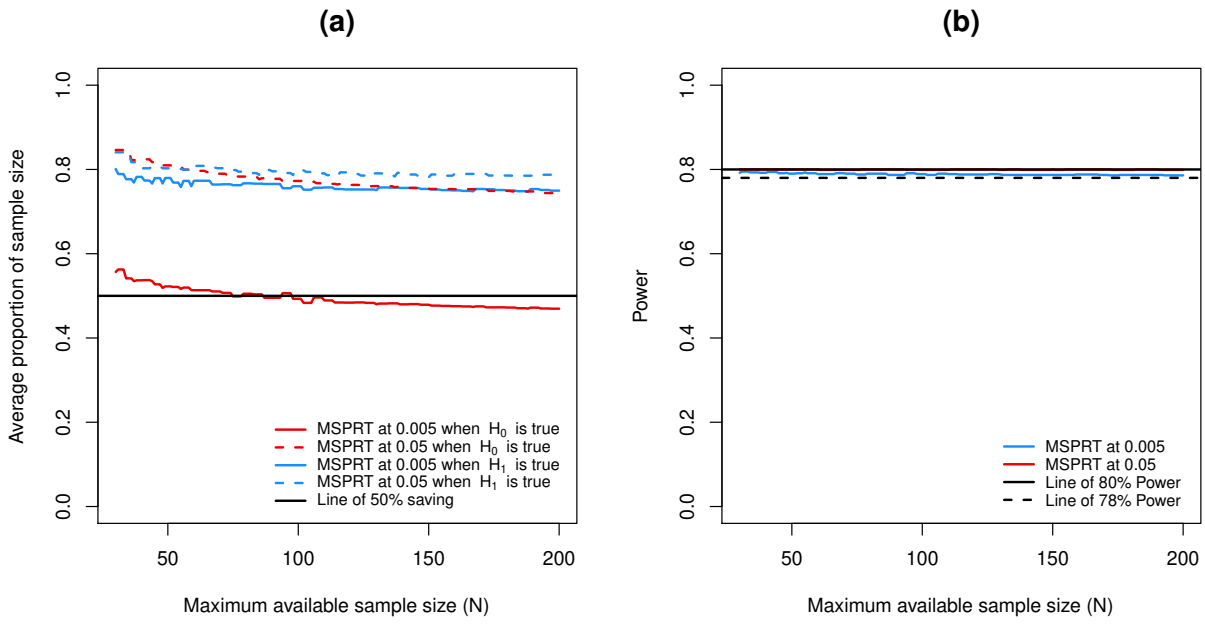


Figure A.8: One-sample test that a binomial proportion equals 0.2. Hypothesis test of $H_0 : \theta = 0.2$ vs. $H_1 : \theta > 0.2$. The curves in the left plot represent the average proportion of the maximum sample size (N) used before the MSPRT terminates in favor of the null or alternative hypothesis. The plot on the right displays the average power of the test against its targeted value of 0.8. In both plots, the operating characteristics under the alternative are evaluated at the corresponding fixed-design point alternatives.

fixed-design point alternatives.

The plot provided in Figure A.6 for a one-sided z test is nearly indistinguishable from the corresponding plots obtained for one-sample t tests and tests of a binomial proportion. For the one-sample z test and the proportion test, we get curves similar to those in Figure A.6. In the case of the proportion test, the discreteness of binomial data causes some non-monotonicity in the proportion of the maximum sample size that is required to reach a decision. This feature of the plot corresponds to the non-monotonicity of power curves for fixed-design tests when sample sizes are increased. For a given a choice of N, the `R` package `MSPRT` finds an "ideal" maximum sample size that accounts for this non-monotonicity. We refer to these values as the "effective sample sizes." In the proportion test, we illustrate the figure using only those values as the maximum sample sizes. This point is further discussed in Section A.4.4.

We next provide the results from simulation studies to examine the potential benefit that the MSPRT can provide in offsetting the increase of sample size that would be incurred if the bar for declaring a result "statistically significant" were moved from $p < 0.05$ to $p < 0.005$. Specifically, we compare the sample sizes needed to achieve statistical significance at the 5% level in standard fixed-design tests to the average sample size needed to achieve statistical significance at the 0.5% level using the MSPRT.

From results cited in the article, this comparison is straightforward if the null hypothesis is true. If not, care must be taken to make sure that the same alternative hypotheses are compared at both levels of significance under the fixed and MSPRT designs. To make this comparison, we determine the θ^* that achieves the targeted Type II error in a fixed-design test of size 0.05. For that θ^*, we next determine the N^* needed to achieve the same Type II error in a fixed-design test of size $\alpha = 0.005$. We then define that N^* to be the maximum sample size for the MSPRT.

A.4.3 Computing the UMPBT alternative

The UMPBT alternative is a key component of the MSPRT design. In this section we illustrate how this alternative can be obtained using the `R` package.

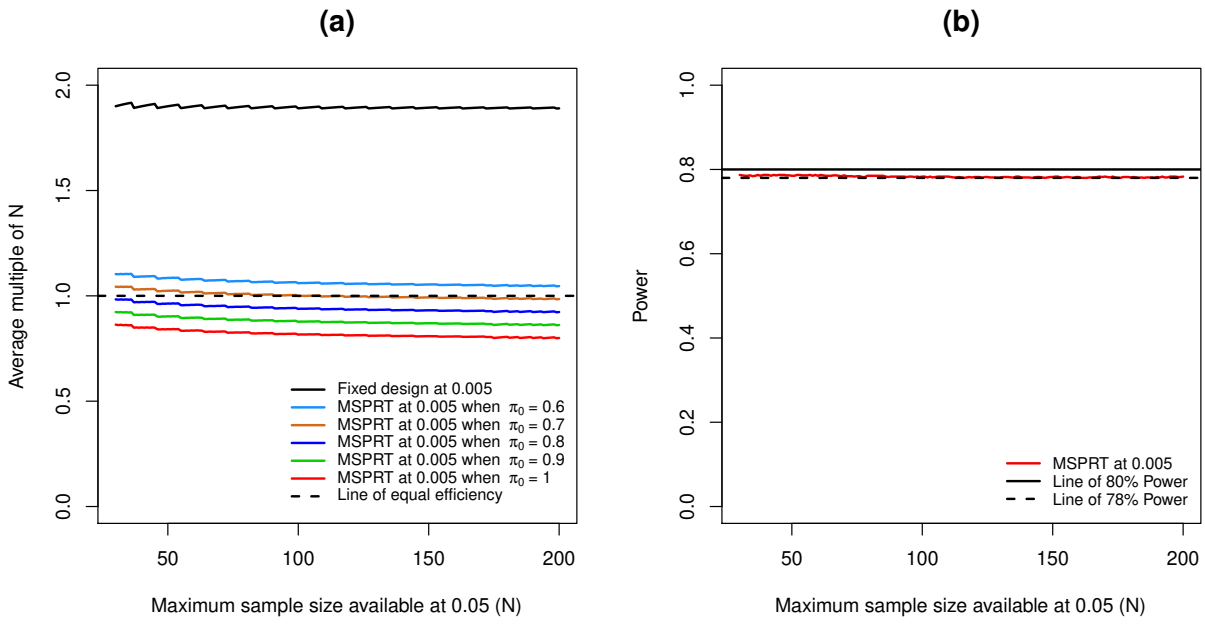


Figure A.9: One-sample z test that a population mean equals 0. Curves in the left plot represent the average multiple of the sample size in a fixed-design test of size 0.05 required in a MSPRT of size 0.005 of approximately the same power. Average sample sizes are dependent on the proportion of tested null hypotheses that are true. The MSPRT maintains a Type I error of 0.005, and its power at θ^* approximately equals 0.8 for the indicated proportion of N^* (the sample size of the corresponding fixed-design test). The power of the MSPRT is depicted in the plot on the right.

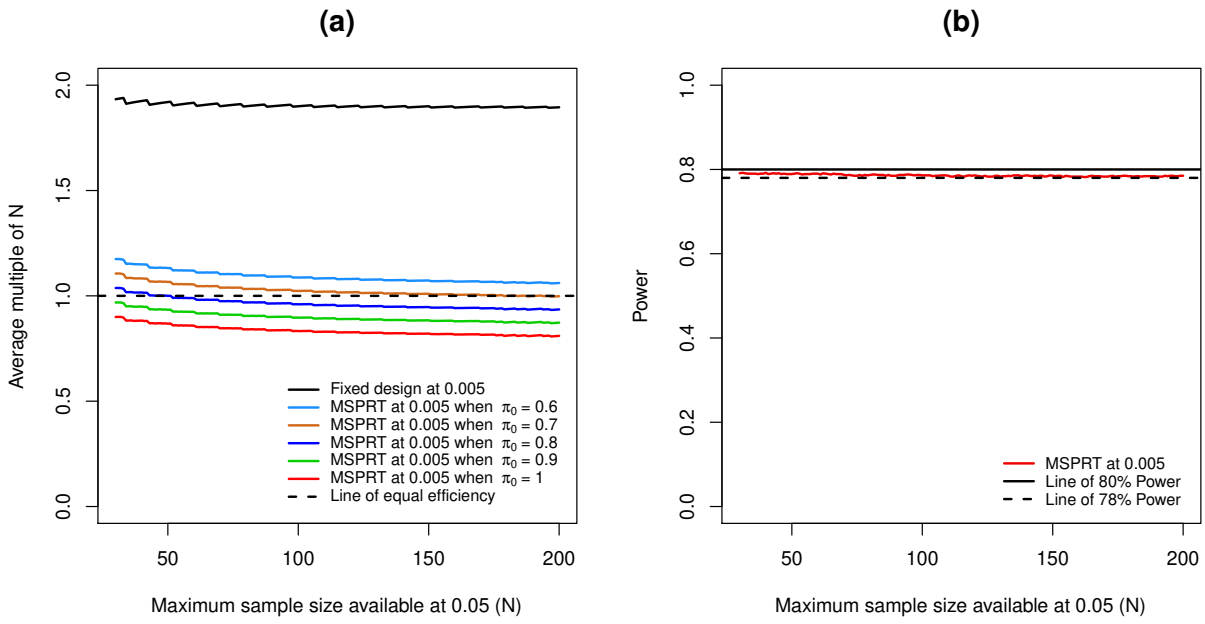


Figure A.10: One-sample t test that a population mean is 0. Curves in the left plot represent the average multiple of the sample size in a fixed-design test of size 0.05 required in a MSPRT of size 0.005 of approximately the same power. Average sample sizes are dependent on the proportion of tested null hypotheses that are true. The MSPRT maintains a Type I error of 0.005, and its power at θ^* approximately equals 0.8 for the indicated proportion of N^* (the sample size of the corresponding fixed-design test). The power of the MSPRT is depicted in the plot on the right.

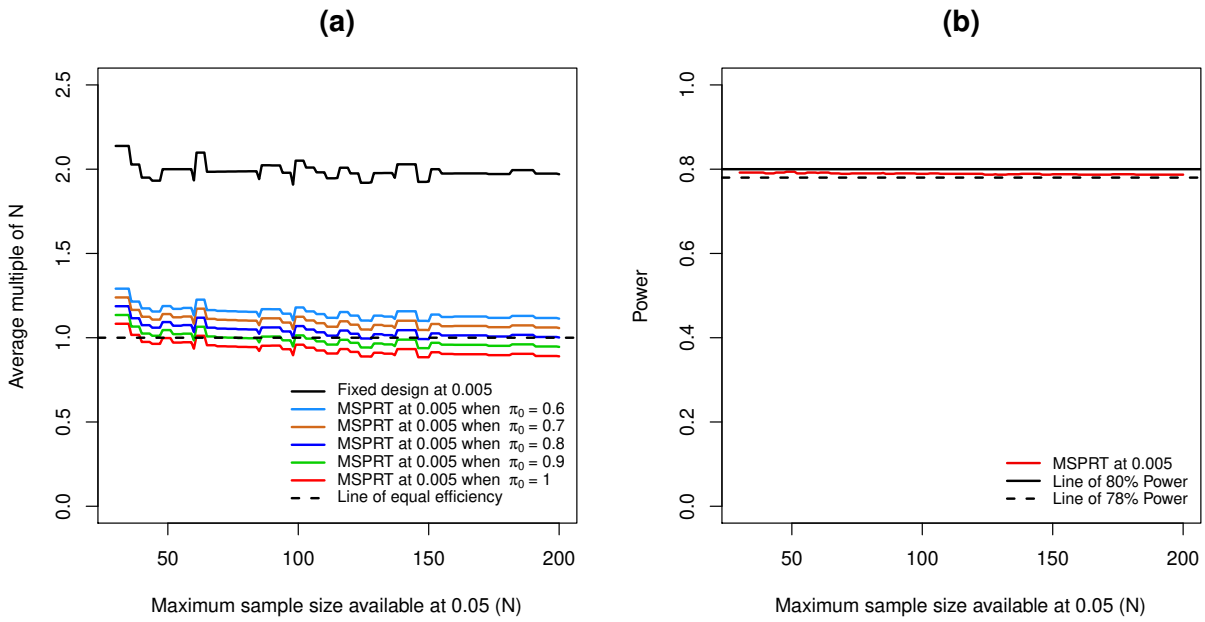


Figure A.11: One-sample test that a binomial proportion equals 0.2. Curves in the left plot represent the average multiple of the sample size in a fixed-design test of size 0.05 required in a MSPRT of size 0.005 of approximately the same power. Average sample sizes are dependent on the proportion of tested null hypotheses that are true. This proportion (π_0) is coded by color, as indicated. The MSPRT maintains a Type I error of 0.005, and its power at θ^* approximately equals 0.8 for the indicated proportion of N^* (the sample size of the corresponding fixed-design test). The power of the MSPRT is depicted in the plot on the right.

A.4.3.1 The z test for a population mean

Consider the test in Section A.4.1.1. To find the UMPBT alternative for a z test, we can use the function `UMPBT.alt()`. This command is executed as follows:

```
> UMPBT.alt(test.type = "oneZ", theta0 = 3, N = 30, Type1 = 0.005, sigma = 1.5)
[1]  3.7054
```

A.4.3.2 The t *test for a population mean*

Similar to the z test, the function `UMPBT.alt()` also calculates the alternative for a t test. From (A.12), it follows that the alternative is data-dependent. Thus, we need to compute the UMPBT alternative after acquiring each data point. In order to do that, we need to specify either the sequentially observed data or the standard deviation (i.e., $s = \sqrt{\sum(x_i - \bar{x})^2/(n-1)}$) of the data.

Consider again the test in Section A.4.1.2 with data `x`:

```
> set.seed(1)
> x = rnorm(n = 30, mean = 5, sd = 1.5)
```

Suppose we want to find the UMPBT alternative after observing the fifth data value. We then need to specify either the data `x[1:5]` or the standard deviation (sd) of these data, which is roughly 1.44, in `UMPBT.alt()`. The required commands are as follows:

```
> UMPBT.alt(test.type = "oneT", theta0 = 3, N = 30, Type1 = 0.005, obs = x[1:5])
[1] 3.725457
>
> sd(x[1:5])   ##sd of the data x[1:5]
[1] 1.441559
>
> UMPBT.alt(test.type = "oneT", theta0 = 3, N = 30, Type1 = 0.005,
```

```
            sd.obs = 1.441559)
[1] 3.725457
```

A.4.3.3 Test for a binomial proportion

In Table 1 of the main article we mentioned that the UMPBT alternative used by the MSPRT is a mixture distribution of two points. The function `UMPBT.alt()` numerically computes this mixture. For illustration, consider the testing problem in Section A.4.1.3. We calculate the alternative for this case with the following command:

```
> UMPBT.alt(test.type = "oneProp", theta0 = 0.2, N = 30, Type1 = 0.005)
$theta
[1] 0.3666727 0.4000178

$mix.prob
[1] 0.2959777 0.7040223
```

From the output, we see that the UMPBT alternative is a mixture distribution of the two points 0.3667 and 0.4 with probabilities 0.296 and 0.704, respectively. This output corresponds to the solutions of (A.22) and the value of ψ defined in (A.21). The alternative illustrated above is a slight modification of what is originally defined as the UMPBT point alternative in [16]. Note that the original alternative is always the second component (`theta[2]` in the previous output) of the UMPBT alternative used by the MSPRT. This output corresponds to the solution of (A.20).

A.4.3.4 Two-sample z *test for a difference in two population means*

We again consider the testing problem in Section A.4.1.4. To find the UMPBT alternative for a two-sample z test, we can similarly use the function `UMPBT.alt()`. This command is executed as follows:

```
> UMPBT.alt(test.type = "twoZ", N1 = 30, N2 = 30, Type1 = 0.005,
            sigma1 = 1.5, sigma2 = 1.5)
[1] 0.9976144
```

A.4.3.5 Two-sample t *test for a difference in two population means*

Similar to the two-sample z test, the function `UMPBT.alt()` calculates the alternative for a two-sample t test. From [11], it follows that the alternative is data-dependent. Thus, we need to compute the UMPBT alternative after acquiring each data point from both groups. In order to calculate this, we need to specify either the sequentially observed data from two groups or the estimated pooled standard deviation.

We again consider the testing problem in Section A.4.1.5 with data `x1` and `x2`:

```
> set.seed(1)
> x1 = rnorm(n = 30, mean = 0.998, sd = 1.5)
> x2 = rnorm(n = 30, mean = 0, sd = 1.5)
```

Suppose we want to find the UMPBT alternative after observing the fifth observation from each group. We then need to specify either the data (`x1[1:5]` and `x2[1:5]`) itself or the estimated pooled standard deviation of these data, which is roughly 1.005, in `umpbt.twoT()`. The commands are as follows:

```
> UMPBT.alt(test.type = "twoT", N1 = 30, N2 = 30, Type1 = 0.005,
            obs1 = x1[1:5], obs2 = x2[1:5])
[1] 1.004799
>
> sqrt(((5-1)*var(x1[1:5]) + (5-1)*var(x2[1:5]))/(5+5-2))  ## estimated pooled sd
[1] 1.461191
>
> UMPBT.alt(test.type = "twoT", N1 = 30, N2 = 30, Type1 = 0.005, pooled.sd = 1.4611
[1] 1.004799
```

A.4.4 Obtaining the "effective sample size" in a proportion test

Because of the discreteness issue in a proportion test, power does not increase monotonically with N when Type I error is exactly maintained. We recommend choosing N to make the expected sample size as small as possible. To accomplish this, a function named `effective.N()` is defined in the MSPRT package.

To illustrate this function, suppose we want to test $H_0 : p = 0.2$ against $H_1 : p > 0.2$ at $\alpha = 0.005$ with at most 30 samples. Given this choice of $N = 30$, we use `effectiveN.oneProp()` to determine the maximum sample size that should be used in designing the MSPRT. The command to do this is as follows:

```
> effectiveN.oneProp(N = 30, theta0 = 0.2)
[1] 28
```

From the output, we see that the recommended design is based on $N = 28$ rather than $N = 30$. If `plot.it = T` (the default), the call to `effective.N()` also returns a plot similar to that depicted in Figure A.12. This plot shows the way an efficient N is chosen, based on decreasing point UMPBT alternatives. The green circled points correspond to the possible choices of N. The largest is chosen as the "effective" N.

A.4.5 Finding N^*

In the main article we compared tests conducted at two levels of significance, 0.05 and 0.005. The comparison was based on the number of samples needed to achieve the higher significance level while still maintaining a prespecified power for the fixed point alternative. In those comparisons we set the point alternative to be the fixed-design alternative for the 5% test.

To determine the fixed-design alternative in a z test for testing $H_0 : \mu = 0$ with known $\sigma = 1$, $\alpha = 0.05$ and $\beta = 0.2$, the following command can be used:

```
> fixed_design.alt(test.type = "oneZ", theta0 = 0, sigma = 1, N = 30,
                   Type1 = 0.05, Type2 = 0.2)
```

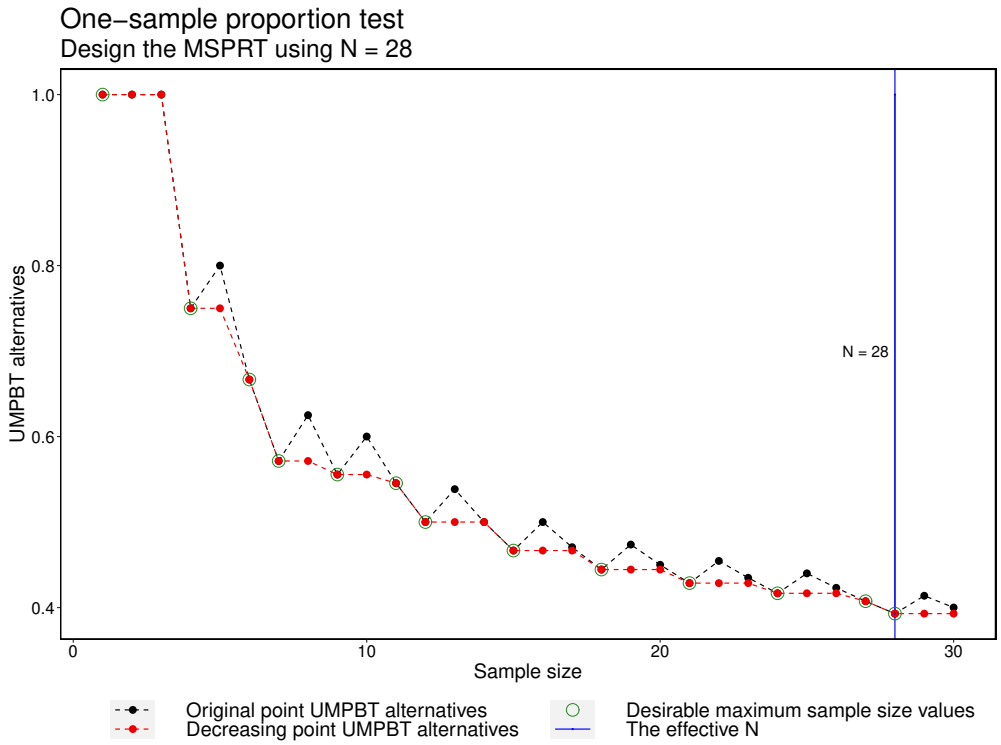


Figure A.12: The "effective" N for testing $H_0 : p = 0.2$ at $\alpha = 0.005$.

```
[1] 0.4539661
```

Now consider finding N^*. Suppose we know N for the 5% test, and we want at least 80% power (the default) at the fixed-design alternative with $\alpha = 0.05$ (that is, at 0.454 for the z test described as above). Given these constraints, the function `find.samplesize()` defined in the MSPRT package finds the required N^*. In this case, the increased sample size in the z test for $N = 30$ for the MSPRT of size 0.5% can be found using the following command:

```
> Nstar(test.type = "oneZ", N = 30)
[1] 57
```

The output reveals that we need 57 samples, about twice the value of N, to achieve the higher significance level of 0.005 while maintaining approximately the same 80% power at the alternative 0.454. If `plot.it = T` (the default), the call to `find.samplesize()` also returns a plot

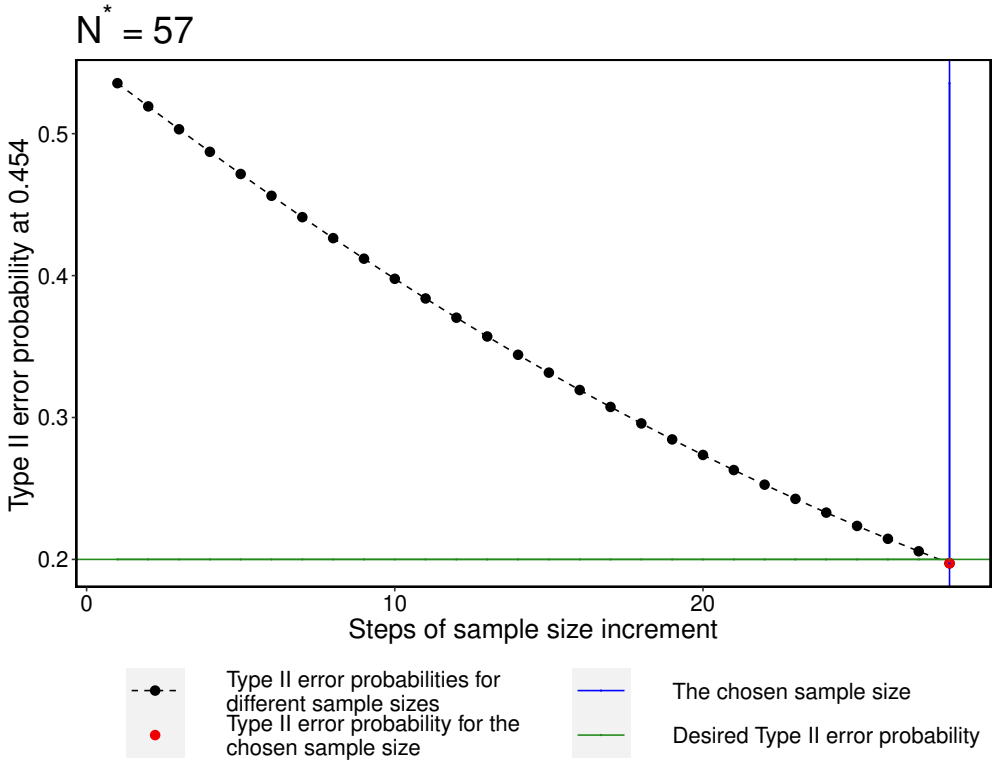


Figure A.13: Finding N^*.

similar to that depicted in Figure A.13. This plot shows that we at least need 57 samples (red point) to meet our requirements.

APPENDIX B

SUPPLEMENTARY MATERIAL: EFFICIENT ALTERNATIVES FOR BAYESIAN HYPOTHESIS TESTS IN PSYCHOLOGY

The supplemental materials address several topics not covered in the main article. First, we provide analytic expressions for Bayes factors in one-sided tests for normal means and differences in normal means. Following this, we provide proofs of these theorems and those stated in the main article. Finally, we provide summaries of operating characteristics for z and two-sample t tests that show that these tests perform similarly to the one-sample t test discussed in the main article.

B.1 Bayes factors for one-sided tests

S1. One-sample, one-sided, known variance test. Assume the conditions of **[1]** in the main article hold, except that now $H_1 : \mu \sim NM^+(0, \tau^2\sigma^2)$. Then the Bayes factor in favor of H_1 can be expressed as

$$\mathrm{BF}_{10}(\mathbf{x}) = 2(n\tau^2 + 1)^{-3/2} \left[(1 + 2w)e^w \left(1 - \mathcal{N}(\sqrt{w/2})\right) + \sqrt{\frac{2w}{\pi}} \right], \tag{B.1}$$

where

$$r = \frac{n\tau^2}{1 + n\tau^2}, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad Z = \sqrt{n}\bar{x}/\sigma, \quad \text{and} \quad w = rZ^2/2, \tag{B.2}$$

and $\mathcal{N}(x)$ is the standard normal distribution function.

S2. One-sample, one-sided, unknown variance test. Suppose the conditions in **[2]** of the main article hold, except now that σ^2 is unknown. Suppose further that the Jeffreys' prior density is assumed under both hypotheses. In this case, the closed-form expression for the Bayes factor is more complicated because it depends on the Gauss hypergeometric function, ${}_2F_1(a, b, c, x)$ and the beta function, $\mathrm{B}(a, b)$. These functions are available in many statistical and math-

ematical software packages, including **R** (see package "hypergeo" for ${}_2F_1$ [104]). Using these functions, the Bayes factor in favor of the alternative hypothesis can be expressed as

$$\mathrm{BF}_{10}(\mathbf{x}) = \begin{cases} c_1 \left[f_1 d_1^2 \left(1 - \mathcal{T}_{2\nu-1}\left(-d_1\sqrt{2\nu-1}\right)\right) + f_2\, d_1 |d_1|^{2(1-\nu)} + \right. & \\ \quad \left. f_3\, |d_1|^{3-2\nu} \right] & \text{if } \bar{x} < 0, \\ c_1 \mathrm{B}\left(3/2, \nu - 3/2\right) & \text{if } \bar{x} = 0, \\ c_1 \left[f_1 d_1^2 \left(1 - \mathcal{T}_{2\nu-1}\left(-d_1\sqrt{2\nu-1}\right)\right) + f_2\, d_1 |d_1|^{2(1-\nu)} + \right. & \\ \quad \left. f_3\, |d_1|^{3-2\nu} + 2 f_4 |d_1|^3 \right] & \text{if } \bar{x} > 0, \end{cases} \tag{B.3}$$

where

$$c_1 = \frac{4\Gamma\left(\nu\right)}{\sqrt{\pi}(n\tau^2+1)^{3/2}\Gamma\left(n/2\right)}, \tag{B.4}$$

$$q = \frac{rn}{n-1}, \quad S = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad s^2 = \frac{S}{(n-1)}, \quad T = \frac{\sqrt{n}\bar{x}}{s}, \tag{B.5}$$

$$G = 1 + \frac{T^2}{n-1}, \quad \text{and} \quad H = 1 + \frac{(1-r)T^2}{(n-1)}. \tag{B.6}$$

Variables $\bar{x}$, S, r, T, G, and H are defined in (B.2, B.5, B.6), $\nu = (n+3)/2$, and $d_1 = \sqrt{r}T/\sqrt{(n-1)H}$. The variables $f_1 - f_4$ are defined as

$$f_1 = \mathrm{B}\left(\nu - 1/2, 1/2\right), \quad f_2 = \frac{{}_2F_1\left(\nu, \nu-1; \nu; -1/d_1^2\right)}{(\nu-1)}, \tag{B.7}$$

$$f_3 = \frac{{}_2F_1\left(\nu, \nu-3/2; \nu-1/2; -1/d_1^2\right)}{(2\nu-3)}, \quad f_4 = \frac{{}_2F_1\left(\nu, 3/2; 5/2; -d_1^2\right)}{3}. \tag{B.8}$$

The function $\mathcal{T}_{2\nu-1}(\cdot)$ denotes the cumulative distribution function of a Student t random variable on $(2\nu-1)$ degrees of freedom.

S3. Two-sample, one-sided, known variance test. Assume the conditions in **[3]** of the main article hold, except that now $H_1 : \mu_2 - \mu_1 \sim NM^+(0, \tau^2\sigma^2)$. Then the Bayes factor in favor

of H_1 can be expressed as

$$\mathrm{BF}_{10}(\mathbf{x}_1, \mathbf{x}_2) = 2\left(m\tau^2 + 1\right)^{-3/2}\left[e^w\,(1+2w)\left(1 - \mathcal{N}\left(\sqrt{w/2}\right)\right) + \sqrt{\frac{2w}{\pi}}\right], \qquad \text{(B.9)}$$

where

$$\bar{x}_i = \sum_{j=1}^{n_i} x_{j,i}/n_i, \qquad n = n_1 + n_2, \qquad m = \frac{n_1 n_2}{n_1 + n_2}, \qquad \text{(B.10)}$$

$$r = \frac{m\tau^2}{m\tau^2 + 1}, \qquad Z = \sqrt{m}(\bar{x}_2 - \bar{x}_1)/\sigma \qquad \text{and} \qquad w = \frac{rZ^2}{2}. \qquad \text{(B.11)}$$

and $\mathcal{N}$ is again the standard normal distribution function.

S4. Two-sample, one-sided, unknown variance test. Suppose the conditions in **[4]** of the main article hold, except now that σ^2 is unknown. Suppose further that the Jeffreys' prior density for σ^2 is assumed under both hypotheses. Then the Bayes factor in favor of the alternative hypothesis can be expressed as

$$\mathrm{BF}_{10}(\mathbf{x}_1, \mathbf{x}_2) = \begin{cases} c_1\left[f_1 d_1^2\left(1 - \mathcal{T}_{2\nu-1}\left(-d_1\sqrt{2\nu-1}\right)\right) + f_2\, d_1|d_1|^{2(1-\nu)} + \right. & \\ \quad \left. f_3\,|d_1|^{3-2\nu}\right] & \text{if } \bar{x}_2 < \bar{x}_1, \\ c_1 \mathrm{B}\left(3/2, \nu - 3/2\right) & \text{if } \bar{x} = 0, \\ c_1\left[f_1 d_1^2\left(1 - \mathcal{T}_{2\nu-1}\left(-d_1\sqrt{2\nu-1}\right)\right) + f_2\, d_1|d_1|^{2(1-\nu)} + \right. & \\ \quad \left. f_3\,|d_1|^{3-2\nu} + 2f_4|d_1|^3\right] & \text{if } \bar{x}_2 > \bar{x}_1, \end{cases} \qquad \text{(B.12)}$$

where

$$c_1 = \frac{2\Gamma\left(\nu\right)}{\sqrt{\pi}(m\tau^2+1)^{3/2}\Gamma\left((n-1)/2\right)}, \qquad \text{(B.13)}$$

$$T = \frac{\sqrt{m}(\bar{x}_1 - \bar{x}_2)}{\sqrt{S/(n-2)}}, \qquad G = 1 + \frac{T^2}{(n-2)}, \qquad H = 1 + \frac{(1-r)T^2}{(n-2)}. \qquad \text{(B.14)}$$

and $\nu = n/2 + 1$. The variables $f_1 - f_4$ are as defined in (B.7,B.8), but with $d_1 = \sqrt{r}T/\sqrt{(n-2)H}$.

B.2 Proofs of theorems for one-sample tests

B.2.1 Variance known

B.2.1.1 Two-sided tests

Suppose $\mathbf{x} = (x_1, \ldots, x_n)$ are i.i.d. observations from a $N(\mu, \sigma^2)$ distribution with σ^2 known. The null hypothesis specifies that $H_0 : \mu = 0$. Under H_1, we assume that μ is drawn from a normal moment prior density specified by

$$p_{NM}(\mu \,|\, \tau^2, \sigma^2) = \frac{1}{\sqrt{2\pi}\tau^3\sigma^3}\mu^2 \exp\left(-\frac{\mu^2}{2\tau^2\sigma^2}\right) \quad \text{for } \mu \in \mathbb{R}. \tag{B.15}$$

Theorem B.2.1. *Under the null hypothesis $H_0 : \mu = 0$ and σ^2 known, the marginal density of $\mathbf{x}$ is given by*

$$m_0(\mathbf{x} \,|\, \sigma^2) = c \exp\left(-\frac{n\bar{x}^2}{2\sigma^2}\right), \tag{B.16}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad S = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad \textit{and} \quad c = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{S}{2\sigma^2}\right). \tag{B.17}$$

Proof: The marginal density under the data under the simple null hypothesis is simply the sampling density of the data. Thus,

$$\begin{aligned} m_0(\mathbf{x} \,|\, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) && \text{(B.18)} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{S}{2\sigma^2} - \frac{n(\bar{x} - \mu)^2}{2\sigma^2}\right). && \text{(B.19)} \end{aligned}$$

Noting $\mu = 0$ under H_0, the result follows.

■

Theorem B.2.2. *Under the alternative hypothesis H_1 that μ is drawn* a priori *from the normal*

moment prior (B.15) and σ^2 known, the marginal density of $\mathbf{x}$ is given by

$$m_1(\mathbf{x} \,|\, \sigma^2) = \frac{c\, a^{3/2}}{\tau^3 \sigma^2} \left(\sigma^2 + a n^2 \bar{x}^2\right) \exp\left[-\frac{an\bar{x}^2}{2\tau^2\sigma^2}\right], \tag{B.20}$$

where $a = 1/(n + \tau^{-2})$ and c is defined in (B.17).

Proof: Substituting the expression for the sampling density of the data obtained in the proof of Theorem B.2.1, multiplying by the prior on μ, and integrating to obtain the marginal density leads to

$$\begin{aligned} m_1(\mathbf{x} \,|\, \sigma^2) &= \int_{-\infty}^{\infty} \frac{c}{\sqrt{2\pi}\tau^3\sigma^3} \mu^2 \exp\left(-\frac{\mu^2}{2\tau^2\sigma^2}\right) \exp\left(-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right) d\mu & \text{(B.21)} \\ &= \int_{-\infty}^{\infty} \frac{c}{\sqrt{2\pi}\tau^3\sigma^3} \mu^2 \exp\left[-\frac{1}{2\sigma^2}\left(\frac{\mu^2}{\tau^2} + n(\bar{x}-\mu)^2\right)\right] d\mu. & \text{(B.22)} \end{aligned}$$

Because

$$\frac{\mu^2}{\tau^2} + n(\bar{x}-\mu)^2 = \frac{1}{a}\left(\mu - an\bar{x}\right)^2 + \frac{an\bar{x}^2}{\tau^2}, \tag{B.23}$$

it follows that

$$\begin{aligned} m_1(\mathbf{x} \,|\, \sigma^2) &= \int_{-\infty}^{\infty} \frac{c}{\sqrt{2\pi}\tau^3\sigma^3} \mu^2 \exp\left\{-\frac{1}{2\sigma^2}\left[\frac{1}{a}\left(\mu - an\bar{x}\right)^2 + n\bar{x}^2 - an^2\bar{x}^2\right]\right\} d\mu & \text{(B.24)} \\ &= \frac{\sqrt{a}c}{\tau^3\sigma^2} \exp\left(-\frac{an\bar{x}^2}{2\tau^2\sigma^2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi a}\sigma} \mu^2 \exp\left[-\frac{(\mu - an\bar{x})^2}{2a\sigma^2}\right] d\mu. & \text{(B.25)} \end{aligned}$$

The integral represents the second moment of a normal distribution with mean $an\bar{x}$ and variance $a\sigma^2$. Thus

$$m_1(\mathbf{x} \,|\, \sigma^2) = \frac{\sqrt{a}c}{\tau^3\sigma^2}\left[a\sigma^2 + (an\bar{x})^2\right] \exp\left(-\frac{an\bar{x}^2}{2\tau^2\sigma^2}\right). \tag{B.26}$$

■

Theorem B.2.3. *Under the assumptions of Thm B.2.1 and B.2.2, the Bayes factor in favor of the*

alternative hypothesis H_1 against the null hypothesis H_0 is given by

$$\mathrm{BF}_{10}(\mathbf{x} \,|\, \sigma^2) = (n\tau^2 + 1)^{-3/2} \left(1 + r\, T^2\right) \exp\left(\frac{r\, T^2}{2}\right), \tag{B.27}$$

where $r = n\tau^2/\left(n\tau^2 + 1\right)$ and $T = \sqrt{n}\bar{x}/\sigma$.

Proof: Following the definition of the Bayes factor and substituting the expression for the marginal density of $\mathbf{x}$ from Thm B.2.1 and B.2.2 leads to

$$\begin{aligned}
& \mathrm{BF}_{10}(\mathbf{x} \,|\, \sigma^2) && \text{(B.28)} \\
= \; & \frac{m_1(\mathbf{x} \,|\, \sigma^2)}{m_0(\mathbf{x} \,|\, \sigma^2)} && \text{(B.29)} \\
= \; & \frac{1}{\sigma^2(n\tau^2 + 1)^{3/2}} \left[\sigma^2 + \frac{n\bar{x}^2}{1 + (n\tau^2)^{-1}}\right] \exp\left(\frac{n^2\tau^2\bar{x}^2}{2\sigma^2(n\tau^2 + 1)}\right) && \text{(B.30)} \\
= \; & (n\tau^2 + 1)^{-3/2} \left(1 + r\, T^2\right) \exp\left(\frac{r\, T^2}{2}\right). && \text{(B.31)}
\end{aligned}$$

■

B.2.1.2 One-sided tests

Assume the conditions of the two-sided test again hold, except that we now wish to test $H_0 : \mu = 0$ versus $H_1 : \mu > 0$. To this end, under H_1 we assume that μ is drawn from a normal moment prior truncated on $(0, \infty)$. The density is specified by

$$p_{NM}(\mu \,|\, \tau^2, \sigma^2) = \frac{\sqrt{2}}{\sqrt{\pi}\tau^3\sigma^3}\mu^2 \exp\left(-\frac{\mu^2}{2\tau^2\sigma^2}\right) \quad \text{for } \mu > 0. \tag{B.32}$$

Under this setup we note that the marginal density of $\mathbf{x}$ under the null hypothesis $H_0 : \mu = 0$ is the same as in Theorem B.2.1.

Theorem B.2.4. *Under the alternative hypothesis H_1 that μ is drawn* a priori *from the normal*

moment prior (B.32) and σ^2 known, the marginal density $m_1(\mathbf{x} \mid \sigma^2)$ of $\mathbf{x}$ is given by

$$\frac{c}{(n\tau^2+1)^{3/2}} \exp\left(-\frac{d^2}{n\tau^2}\right) \left[\left(2d^2+1\right)\{1-\textit{erf}(-d)\} + \frac{2d}{\sqrt{\pi}} \exp\left(-d^2\right)\right], \tag{B.33}$$

where a is as in Theorem B.2.2, c is as in Theorem B.2.1, and $d = \sqrt{a}\, n\bar{x}/\sqrt{2}\,\sigma$.

Proof: Substituting the expression for the sampling density of the data obtained in the proof of Theorem B.2.1, multiplying by the prior (B.32) on μ, and integrating to obtain the marginal density leads to

$$m_1(\mathbf{x} \mid \sigma^2) = \int_0^\infty \frac{c\sqrt{2}}{\sqrt{\pi}\tau^3\sigma^3}\mu^2 \exp\left(-\frac{\mu^2}{2\tau^2\sigma^2}\right) \exp\left(-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right) d\mu. \tag{B.34}$$

Using the identity (B.23) and using 2.1.3.1 from [105] leads to

$$\begin{aligned}
m_1(\mathbf{x} \mid \sigma^2) &= \frac{c\sqrt{2}}{\sqrt{\pi}\tau^3\sigma^3} \exp\left(-\frac{an\bar{x}^2}{2\tau^2\sigma^2}\right) \int_0^\infty \mu^2 \exp\left[-\frac{(\mu - an\bar{x})^2}{2a\sigma^2}\right] d\mu && \text{(B.35)}\\
&= \frac{c\sqrt{2}}{\sqrt{\pi}\tau^3\sigma^3} \exp\left(-\frac{an\bar{x}^2}{2\tau^2\sigma^2}\right) \times && \text{(B.36)}\\
&\quad \left[\frac{\sqrt{\pi}a^{3/2}\sigma^3}{\sqrt{2}}\left(\frac{an^2\bar{x}^2}{\sigma^2}+1\right)\left\{1-\text{erf}\left(-\frac{\sqrt{a}n\bar{x}}{\sqrt{2}\sigma}\right)\right\} + \right. && \text{(B.37)}\\
&\quad \left. a^2 n\bar{x}\sigma^2 \exp\left(-\frac{an^2\bar{x}^2}{2\sigma^2}\right)\right] && \text{(B.38)}\\
&= \frac{ca^{3/2}}{\tau^3} \exp\left(-\frac{an\bar{x}^2}{2\tau^2\sigma^2}\right) \times && \text{(B.39)}\\
&\quad \left[\left(\frac{an^2\bar{x}^2}{\sigma^2}+1\right)\left\{1-\text{erf}\left(-\frac{\sqrt{a}n\bar{x}}{\sqrt{2}\sigma}\right)\right\} + \frac{\sqrt{2a}\, n\bar{x}}{\sqrt{\pi}\sigma} \exp\left(-\frac{an^2\bar{x}^2}{2\sigma^2}\right)\right] && \text{(B.40)}\\
&= \frac{c}{(n\tau^2+1)^{3/2}} \exp\left(-\frac{d^2}{n\tau^2}\right) \times && \text{(B.41)}\\
&\quad \left[\left(2d^2+1\right)\{1-\text{erf}\,(-d)\} + \frac{2d}{\sqrt{\pi}} \exp\left(-d^2\right)\right]. && \text{(B.42)}
\end{aligned}$$

■

Theorem B.2.5. *Under the assumptions stated above, the Bayes factor $\text{BF}_{10}(\mathbf{x} \mid \sigma^2)$ in favor of the*

alternative hypothesis H_1 against the null hypothesis H_0 is given by

$$\left(n\tau^2+1\right)^{-3/2}\exp\left(\frac{rT^2}{2}\right)\left[\left(rT^2+1\right)\left(1-\textit{erf}\left(-\frac{\sqrt{r}T}{\sqrt{2}}\right)\right)+\frac{\sqrt{2r}T}{\sqrt{\pi}}\exp\left(-\frac{rT^2}{2}\right)\right], \tag{B.43}$$

where $r = n\tau^2/\left(n\tau^2+1\right)$ and $T = \sqrt{n}\bar{x}/\sigma$.

Proof: Following the definition of the Bayes factor and substituting the expression for the marginal density of $\mathbf{x}$ from Thm B.2.1 and B.2.4 leads to

$$\begin{aligned}
& \mathrm{BF}_{10}(\mathbf{x}\,|\,\sigma^2) && \text{(B.44)}\\
= \;& \frac{m_1(\mathbf{x}\,|\,\sigma^2)}{m_0(\mathbf{x}\,|\,\sigma^2)} && \text{(B.45)}\\
= \;& \left(n\tau^2+1\right)^{-3/2}\exp\left(\frac{n^2\tau^2\bar{x}^2}{2\sigma^2(n\tau^2+1)}\right)\times && \text{(B.46)}\\
& \left[\left(\frac{n^2\tau^2\bar{x}^2}{\sigma^2(n\tau^2+1)}+1\right)\left(1-\mathrm{erf}\left(-\frac{n\tau\bar{x}}{\sigma\sqrt{2\left(n\tau^2+1\right)}}\right)\right)+\right. && \text{(B.47)}\\
& \left.\frac{\sqrt{2}n\tau\bar{x}}{\sigma\sqrt{\pi\left(n\tau^2+1\right)}}\exp\left(-\frac{n^2\tau^2\bar{x}^2}{2\sigma^2(n\tau^2+1)}\right)\right] && \text{(B.48)}\\
= \;& \left(n\tau^2+1\right)^{-3/2}\exp\left(\frac{rT^2}{2}\right)\times && \text{(B.49)}\\
& \left[\left(rT^2+1\right)\left(1-\mathrm{erf}\left(-\frac{\sqrt{r}T}{\sqrt{2}}\right)\right)+\frac{\sqrt{2r}T}{\sqrt{\pi}}\exp\left(-\frac{rT^2}{2}\right)\right]. && \text{(B.50)}
\end{aligned}$$

■

B.2.2 Variance unknown

B.2.2.1 Two-sided tests

As in for the two-sided t test, let $\mathbf{x} = (x_1, \ldots, x_n)$ denote i.i.d. observations from a $N(\mu, \sigma^2)$ distribution, but assume now that σ^2 unknown. We again wish to test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$. Under H_1, the prior on μ, given σ^2, is again specified as a normal moment prior (B.15). To complete the model specification, under both H_0 and H_1 we also assume an inverse gamma prior

on σ^2, parameterized here as

$$\pi(\sigma^2 \,|\, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-\alpha-1} \exp\left(-\frac{\beta}{\sigma^2} \right) \quad \text{for } \sigma^2 > 0, \tag{B.51}$$

with shape parameter $\alpha\,(>0)$ and scale parameter $\beta\,(>0)$.

Theorem B.2.6. *Under these assumptions and assuming H_0 to be true, the marginal density of the data $m(\mathbf{x})$ is given by*

$$m_0(\mathbf{x}) = \frac{(2\pi)^{-n/2}\,\beta^\alpha\,\Gamma(n/2+\alpha)}{\Gamma(\alpha)} \left[\frac{S + n\bar{x}^2}{2} + \beta \right]^{-n/2-\alpha}, \tag{B.52}$$

where S is as defined in (B.17).

Proof: Since H_0 is a point null hypothesis, the prior on μ is a degenerate distribution with all the mass at μ_0. So the marginal density $m(\mathbf{x})$ can be expressed as

$$m_0(\mathbf{x}) = \int \pi(\sigma^2 \,|\, \alpha, \beta) \prod_{i=1}^{n} \phi(x_i \,|\, 0, \sigma^2)\, d\sigma^2 \tag{B.53}$$

$$= \frac{(2\pi)^{-n/2}\,\beta^\alpha}{\Gamma(\alpha)} \int (\sigma^2)^{-n/2-\alpha-1} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 - \frac{\beta}{\sigma^2} \right] d\sigma^2. \tag{B.54}$$

Noting that the above integral with respect to σ^2 is proportional to an Inverse-gamma density yields (B.52).

■

Theorem B.2.7. *Under the assumptions above and assuming H_1 is true, the marginal density of the data $m(\mathbf{x})$ can be expressed as*

$$m_1(\mathbf{x}) = c^* \left[\frac{S + d\bar{x}^2}{2} + \beta \right]^{-n/2-\alpha-1} \left[\frac{S + d\bar{x}^2}{2} + \beta + nd\tau^2\bar{x}^2 \left(\frac{n}{2} + \alpha \right) \right], \tag{B.55}$$

where $\bar{x}$, S are as in (B.17), and

$$c^* = \frac{(2\pi)^{-n/2}\,\beta^{\alpha}\,\Gamma(n/2+\alpha)}{(n\tau^2+1)^{3/2}\,\Gamma(\alpha)}, \quad \textit{and} \quad d = \frac{n}{n\tau^2+1}. \tag{B.56}$$

Proof: The marginal density $m_1(\mathbf{x})$, given τ^2, can be obtained by integrating (B.20) over the prior on σ^2, leading to

$$m_1(\mathbf{x}) = \int \pi(\sigma^2 \,|\, \alpha, \beta) \times \frac{c}{\sigma^2(n\tau^2+1)^{3/2}} \exp\left[-\frac{d\bar{x}^2}{2\sigma^2}\right]\left(\sigma^2 + nd\tau^2\bar{x}^2\right) d\sigma^2 \tag{B.57}$$

$$= \frac{(2\pi)^{-n/2}\beta^{\alpha}}{(n\tau^2+1)^{3/2}\,\Gamma(\alpha)} \int (\sigma^2)^{-n/2-\alpha-2} \exp\left[-\frac{(S+d\bar{x}^2+2\beta)}{2\sigma^2}\right]\left(\sigma^2 + nd\tau^2\bar{x}^2\right) d\sigma^2. \tag{B.58}$$

Noting that the integrals with respect to σ^2 are proportional to an Inverse-gamma density results in

$$m_1(\mathbf{x}) = \frac{(2\pi)^{-n/2}\beta^{\alpha}}{(n\tau^2+1)^{3/2}\,\Gamma(\alpha)} \left[\Gamma\left(\frac{n}{2}+\alpha\right)\left\{\frac{S}{2}+\frac{d\bar{x}^2}{2}+\beta\right\}^{-n/2-\alpha} + \right. \tag{B.59}$$

$$\left. nd\tau^2\bar{x}^2\,\Gamma\left(\frac{n}{2}+\alpha+1\right)\left\{\frac{S}{2}+\frac{d\bar{x}^2}{2}+\beta\right\}^{-n/2-\alpha-1}\right] \tag{B.60}$$

$$= \quad c^*\left[\frac{S+d\bar{x}^2}{2}+\beta\right]^{-n/2-\alpha-1}\left[\frac{S+d\bar{x}^2}{2}+\beta+nd\tau^2\bar{x}^2\left(\frac{n}{2}+\alpha\right)\right]. \tag{B.61}$$

■

Theorem B.2.8. *Under the assumptions of Thm B.2.6 and B.2.7, the Bayes factor in favor of the alternative hypothesis H_1 against the null hypothesis H_0 is given by*

$$\mathrm{BF}_{10}(\mathbf{x}) = (n\tau^2+1)^{-3/2}\left(\frac{G}{H}\right)^{n/2+\alpha}\left(1+\frac{q\,T^2}{H}\right), \tag{B.62}$$

where

$$r = \frac{n\tau^2}{n\tau^2+1}, \qquad q = \frac{2r\,(n/2+\alpha)}{n-1}, \qquad T = \frac{\sqrt{n}\bar{x}}{\sqrt{S/(n-1)}}, \tag{B.63}$$

$$G = 1 + \frac{T^2}{n-1} + \frac{2\beta}{S}, \qquad H = 1 + \frac{(1-r)\,T^2}{n-1} + \frac{2\beta}{S}. \tag{B.64}$$

and S is as in (B.17).

Proof: Following the definition of the Bayes factor and substituting the expression for the marginal density of $\mathbf{x}$ from Thm B.2.1 and B.2.2 leads to

$$\begin{aligned}
& \mathrm{BF}_{10}(\mathbf{x}) && \text{(B.65)}\\
= \;& \frac{m_1(\mathbf{x})}{m_0(\mathbf{x})} && \text{(B.66)}\\
= \;& (n\tau^2+1)^{-3/2}\left[\frac{(S+n\bar{x}^2)/2+\beta}{(S+d\bar{x}^2)/2+\beta}\right]^{n/2+\alpha}\left[1+\frac{n^2\tau^2\bar{x}^2\,(n/2+\alpha)}{(n\tau^2+1)\left((S+d\bar{x}^2)/2+\beta\right)}\right] && \text{(B.67)}\\
= \;& (n\tau^2+1)^{-3/2}\left[\frac{1+T^2/(n-1)+2\beta/S}{1+T^2/\{(n\tau^2+1)(n-1)\}+2\beta/S}\right]^{n/2+\alpha} \times && \text{(B.68)}\\
& \left[1+\frac{2\tau^2 n(n/2+\alpha)}{(n\tau^2+1)}\frac{T^2/(n-1)}{1+T^2/\{(n\tau^2+1)(n-1)\}+2\beta/S}\right] && \text{(B.69)}\\
= \;& (n\tau^2+1)^{-3/2}\left(\frac{G}{H}\right)^{n/2+\alpha}\left(1+\frac{q\,T^2}{H}\right). && \text{(B.70)}
\end{aligned}$$

■

B.2.2.2 One-sided tests

Assume the conditions of the two-sided test hold, except that we now wish to test $H_0 : \mu = 0$ versus $H_1 : \mu > 0$. The prior on μ given σ^2 under H_1 is specified as (B.32), a normal moment prior truncated on $(0, \infty)$. To complete the model specification, under both H_0 and H_1 we assume an inverse gamma prior on σ^2 defined by (B.51). Under these assumptions and assuming H_0 to be true, the marginal density of the data $m_0(\mathbf{x})$ is the same as in Theorem B.2.6.

Theorem B.2.9. *Under the assumptions above and assuming H_1 is true, the marginal density of*

the data $m_1(\mathbf{x})$ *can be expressed as*

$$\begin{cases} c^* \left(f_1 d^2 \left(1 - F_{2\nu-1}\left(-d\sqrt{2\nu-1}\right)\right) + f_2\, d\, |d|^{2(1-\nu)} + f_3\ |d|^{3-2\nu} \right) & \text{if } \bar{x} < 0, \\ c^* \mathrm{B}\left(3/2, \nu - 3/2\right) & \text{if } \bar{x} = 0, \\ c^* \left(f_1 d^2 \left(1 - F_{2\nu-1}\left(-d\sqrt{2\nu-1}\right)\right) + f_2\, d\, |d|^{2(1-\nu)} + \right. & \\ \quad \left. f_3\ |d|^{3-2\nu} + 2 f_4\, |d|^3 \right) & \text{if } \bar{x} > 0, \end{cases} \tag{B.71}$$

where $\bar{x}$*,* S *are as in (B.17),* $d = \sqrt{a}\, n\bar{x}/\sqrt{2A_1}$*,* $\nu = (n+3)/2 + \alpha$*,* $a = 1/\left(n + \tau^{-2}\right)$*,*

$$c^* = \frac{(2\pi)^{-n/2}\, 4\beta^\alpha\, \Gamma(\nu)}{\sqrt{\pi}(n\tau^2+1)^{3/2}\, \Gamma(\alpha) A_1^{n/2+\alpha}}, \quad A_1 = \beta + \frac{S}{2} + \frac{an\bar{x}^2}{2\tau^2}, \tag{B.72}$$

$$f_1 = \mathrm{B}\left(\nu - 1/2, 1/2\right), \quad f_2 = \frac{{}_2F_1\left(\nu, \nu-1; \nu; -1/d^2\right)}{(\nu-1)}, \tag{B.73}$$

$$f_3 = \frac{{}_2F_1\left(\nu, \nu-3/2; \nu-1/2; -1/d^2\right)}{(2\nu-3)}, \quad f_4 = \frac{{}_2F_1\left(\nu, 3/2; 5/2; -d^2\right)}{3}, \tag{B.74}$$

$\mathrm{B}(\cdot,\cdot)$ *is the Beta function,* $F_{2\nu-1}$ *is the cdf of the Student's* t *distribution (center 0 and scale 1) with degrees of freedom* $2\nu - 1$*, and* ${}_2F_1$ *is the Gauss hypergeometric function.*

Proof: Substituting the expression for the sampling density of the data obtained in the proof of Theorem B.2.1, multiplying by the priors on $\mu \,|\, \sigma^2$ and σ^2, and integrating to obtain the marginal density given τ^2 leads to

$$m_1(\mathbf{x}) = \int_0^\infty \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\sigma^2\right)^{-\alpha-1} \exp\left(-\frac{\beta}{\sigma^2}\right) \frac{\sqrt{2}}{\sqrt{\pi}\tau^3\sigma^3} \mu^2 \exp\left(-\frac{\mu^2}{2\tau^2\sigma^2}\right) \times \tag{B.75}$$

$$\left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{S}{2\sigma^2} - \frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right) d\sigma^2\, d\mu \tag{B.76}$$

$$= \frac{\beta^\alpha \Gamma\left(\nu\right)}{2^{(n-1)/2}\, \pi^{(n+1)/2}\, \tau^3\, \Gamma(\alpha)} \int_0^\infty \mu^2 \left[\beta + \frac{S}{2} + \frac{1}{2}\left(\frac{\mu^2}{\tau^2} + n(\bar{x}-\mu)^2\right)\right]^{-\nu} d\mu. \tag{B.77}$$

where $\nu = (n+3)/2 + \alpha$. Define $A_1 = \beta + S/2 + an\bar{x}^2/2\tau^2$ and

$$A_2 = \frac{\beta^\alpha \Gamma(\nu)}{2^{(n-1)/2}\,\pi^{(n+1)/2}\,\tau^3\,\Gamma(\alpha)\,A_1^\nu}. \tag{B.78}$$

Using the identity (B.23) and some algebraic simplifications lead to

$$\begin{aligned}
m_1(\mathbf{x}) =& A_2 \int_0^\infty \mu^2 \left(1 + \frac{(\mu - an\bar{x})^2}{2aA_1}\right)^{-\nu} d\mu && \text{(B.79)}\\
=& A_2 \int_{-an\bar{x}}^\infty (u + an\bar{x})^2 \left(1 + \frac{u^2}{2aA_1}\right)^{-\nu} du && \text{(B.80)}\\
=& A_2 \left(a^2n^2\bar{x}^2\, I_0(-an\bar{x}) + 2an\bar{x}\, I_1(-an\bar{x}) + I_2(-an\bar{x})\right) && \text{(B.81)}\\
=& A_2 \left(m_{10} + m_{11} + m_{12}\right), && \text{(B.82)}
\end{aligned}$$

where

$$I_k(g) = \int_g^\infty u^k \left(1 + \frac{u^2}{2aA_1}\right)^{-\nu} du \quad \text{for } g \in \mathbb{R}, k \geq 0, \tag{B.83}$$

$$m_{10} = a^2n^2\bar{x}^2\, I_0(-an\bar{x}), \quad m_{11} = 2an\bar{x}\, I_1(-an\bar{x}), \quad m_{12} = I_2(-an\bar{x}). \tag{B.84}$$

For $I_0(-an\bar{x})$, first doing a change of variable with $w/\sqrt{2\nu-1} = u/\sqrt{2aA_1}$ and then some algebraic simplifications lead to

$$\begin{aligned}
I_0(-an\bar{x}) &= \int_{-an\bar{x}}^\infty \left(1 + \frac{u^2}{2aA_1}\right)^{-\nu} du && \text{(B.85)}\\
&= \left(\frac{2aA_1}{2\nu-1}\right)^{1/2} \int_{-n\bar{x}\sqrt{\frac{(2\nu-1)a}{2A_1}}}^\infty \left(1 + \frac{w^2}{2\nu-1}\right)^{-((2\nu-1)+1)/2} dw && \text{(B.86)}\\
&= \sqrt{2aA_1}\, \mathrm{B}\left((2\nu-1)/2, 1/2\right) \left[1 - F_{2\nu-1}\left(-n\bar{x}\sqrt{\frac{(2\nu-1)a}{2A_1}}\right)\right], && \text{(B.87)}
\end{aligned}$$

where $\mathrm{B}(\cdot,\cdot)$ is the Beta function and $F_{2\nu-1}$ is the cdf of the Student's t distribution (center 0 and

scale 1) with degrees of freedom $2\nu - 1$. Following this we get

$$m_{10} = a^2 n^2 \bar{x}^2 \, I_0 \left(-an\bar{x}\right) = \left(2aA_1\right)^{3/2} f_1 d^2 \left(1 - F_{2\nu-1}\left(-d\sqrt{2\nu-1}\right)\right), \tag{B.88}$$

where $d = \sqrt{a}\, n\bar{x}/\sqrt{2A_1}$ and $f_1 = \mathrm{B}\left(\nu - 1/2, 1/2\right)$. For $I_1(-an\bar{x})$ and $I_2(-an\bar{x})$, we note that for integers k,

$$I_k(g) = \begin{cases} I_k\left(|g|\right) & \text{if } g \geq 0, \text{ or } g < 0 \text{ and } k \text{ is odd,} \\ I_k\left(|g|\right) + 2J_k\left(|g|\right) & \text{if } g < 0 \text{ and } k \text{ is even,} \end{cases} \tag{B.89}$$

where for $g > 0$,

$$I_k(g) = \int_g^\infty u^k \left(1 + \frac{u^2}{2aA_1}\right)^{-\nu} du = \frac{1}{2}\int_{g^2}^\infty w^{(k+1)/2-1} \left(1 + \frac{w}{2aA_1}\right)^{-\nu} dw, \tag{B.90}$$

$$J_k(g) = \int_0^g u^k \left(1 + \frac{u^2}{2aA_1}\right)^{-\nu} du = \frac{1}{2}\int_0^{g^2} w^{(k+1)/2-1} \left(1 + \frac{w}{2aA_1}\right)^{-\nu} dw. \tag{B.91}$$

Using equations 3.194.1–3.194.3 from [106] to this leads to

$$I_k(g) = \begin{cases} \frac{g^{k+1-2\nu}\left(2aA_1\right)^\nu}{2\nu-k-1} \, {}_2F_1\left(\nu, (2\nu-k-1)/2; (2\nu-k+1)/2; -2aA_1/g^2\right) & \text{if } g > 0, \\ \left(2aA_1\right)^{(k+1)/2} \mathrm{B}\left((k+1)/2, (2\nu-k-1)/2\right) & \text{if } g = 0, \end{cases} \tag{B.92}$$

and

$$J_k\left(g\right) = \frac{g^{k+1}}{k+1} \, {}_2F_1\left(\nu, (k+1)/2; (k+3)/2; -g^2/2aA_1\right) \quad \text{for } g > 0, \tag{B.93}$$

where ${}_2F_1$ is the Gauss hypergeometric function. To our interest, this results in

$$I_1(an\left|\bar{x}\right|) = \begin{cases} \frac{\left(an|\bar{x}|\right)^{2(1-\nu)}\left(2aA_1\right)^\nu}{2(\nu-1)} \, {}_2F_1\left(\nu, \nu-1; \nu; -2A_1/an^2\bar{x}^2\right) & \text{if } \bar{x} \neq 0, \\ 2aA_1 \, \mathrm{B}\left(1, \nu-1\right) & \text{if } \bar{x} = 0. \end{cases} \tag{B.94}$$

This leads to

$$m_{11} = \begin{cases} 2an\bar{x}\, I_1\left(an\,|\bar{x}|\right) & \text{if } \bar{x} \neq 0, \\ 0 & \text{if } \bar{x} = 0 \end{cases} = \begin{cases} \left(2aA_1\right)^{3/2}\ f_2\, d\,|d|^{2(1-\nu)} & \text{if } \bar{x} \neq 0, \\ 0 & \text{if } \bar{x} = 0. \end{cases} \tag{B.95}$$

where $f_2 = {}_2F_1\left(\nu, \nu-1; \nu; -1/d^2\right)/\left(\nu-1\right)$. Similarly, it also results in

$$I_2(an\,|\bar{x}|) = \begin{cases} \left(2aA_1\right)^{3/2}\ f_3\ |d|^{3-2\nu} & \text{if } \bar{x} \neq 0, \\ \left(2aA_1\right)^{3/2}\ \mathrm{B}\left(3/2, \nu-3/2\right) & \text{if } \bar{x} = 0, \end{cases} \tag{B.96}$$

and

$$J_2(an\,|\bar{x}|) = \left(2aA_1\right)^{3/2}\ f_4\ |d|^3\,, \tag{B.97}$$

where $f_3 = {}_2F_1\left(\nu, \nu-3/2; \nu-1/2; -1/d^2\right)/\left(2\nu-3\right)$ and $f_4 = {}_2F_1\left(\nu, 3/2; 5/2; -d^2\right)/3$. This leads to

$$m_{12} = \begin{cases} \left(2aA_1\right)^{3/2}\ f_3\ |d|^{3-2\nu} & \text{if } \bar{x} < 0, \\ \left(2aA_1\right)^{3/2}\ \mathrm{B}\left(3/2, \nu-3/2\right) & \text{if } \bar{x} = 0, \\ \left(2aA_1\right)^{3/2}\ \left(f_3\ |d|^{3-2\nu} + 2f_4\,|d|^3\right) & \text{if } \bar{x} > 0. \end{cases} \tag{B.98}$$

Finally, (B.71) follows by combining m_{10}, m_{11} and m_{12}.

■

Theorem B.2.10. *Under the assumptions specified above, the Bayes factor* $\mathrm{BF}_{10}(\mathbf{x})$ *in favor of the alternative hypothesis* H_1 *against the null hypothesis* H_0 *is given by*

$$\begin{cases} C_1\left(f_1 d^2\left(1 - F_{2\nu-1}\left(-d\sqrt{2\nu-1}\right)\right) + f_2\, d\,|d|^{2(1-\nu)} + f_3\ |d|^{3-2\nu}\right) & \textit{if } \bar{x} < 0, \\ C_1 \mathrm{B}\left(3/2, \nu-3/2\right) & \textit{if } \bar{x} = 0, \\ C_1\left(f_1 d^2\left(1 - F_{2\nu-1}\left(-d\sqrt{2\nu-1}\right)\right) + f_2\, d\,|d|^{2(1-\nu)} + \right. & \\ \quad \left. f_3\ |d|^{3-2\nu} + 2f_4\,|d|^3\right) & \textit{if } \bar{x} > 0, \end{cases} \tag{B.99}$$

where

$$C_1 = \frac{4\Gamma(\nu)}{\sqrt{\pi}(n\tau^2+1)^{3/2}\Gamma(n/2+\alpha)}, \tag{B.100}$$

$\bar{x}$, S are as in (B.17), $\nu = (n+3)/2+\alpha$, T, r, G and H are as in (B.63)–(B.64), $d = \sqrt{r}T/\sqrt{(n-1)H}$, and f_1 to f_4 are as in (B.73)–(B.74) with d is it is defined here.

Proof: Following the definition of the Bayes factor we know that $\mathrm{BF}_{10}(\mathbf{x}) = m_1(\mathbf{x})/m_0(\mathbf{x})$. While substituting the expression for the marginal density of $\mathbf{x}$ from Thm B.2.6 and B.2.9 we note that

$$\frac{c^*}{m_0(\mathbf{x})} = \frac{4\Gamma(\nu)}{\sqrt{\pi}(n\tau^2+1)^{3/2}\Gamma(n/2+\alpha)}\left(\frac{\beta+S/2+n\bar{x}^2/2}{\beta+S/2+n\bar{x}^2/2(n\tau^2+1)}\right)^{n/2+\alpha} \tag{B.101}$$

$$= \frac{4\Gamma(\nu)}{\sqrt{\pi}(n\tau^2+1)^{3/2}\Gamma(n/2+\alpha)}\left(\frac{G}{H}\right)^{n/2+\alpha}. \tag{B.102}$$

Also, d as in Theorem B.2.9 can be rewritten as

$$d = \frac{\sqrt{rn}\bar{x}}{\sqrt{2\left(\beta+S/2+n\bar{x}^2/2(n\tau^2+1)\right)}} = \frac{\sqrt{r}T}{\sqrt{(n-1)H}}. \tag{B.103}$$

(B.99) directly follows from combining these.

■

B.3 Proofs of theorems of two-sample tests

B.3.1 Variance known

B.3.1.1 Two-sided tests

Suppose $\mathbf{x_1} = (x_{1,1}, \ldots, x_{1,n_1})$ and $\mathbf{x_2} = (x_{2,1}, \ldots, x_{1,n_2})$ are observations from i.i.d. $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ distributions, respectively, and we wish to test $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$. To this end, we assume that under both H_0 and H_1, the prior on μ_1 is $U(-a, a)$ for some large a. Under H_1, we further assume that $\mu_2 = \mu_1 + \delta$, where

$$p(\delta \,|\, \tau^2, \sigma^2) = \frac{1}{\sqrt{2\pi}\tau^3\sigma^3}\delta^2 \exp\left(-\frac{\delta^2}{2\tau^2\sigma^2}\right), \tag{B.104}$$

a normal moment prior on the difference between the means μ_1 and μ_2. We let $\phi(\cdot \mid \mu, \sigma^2)$ denote a normal density function with mean μ and variance σ^2. With a uniform prior on μ_1 and sufficiently large a, we note that the marginal distributions described below are invariant with respect to the labeling of samples.

Theorem B.3.1. *Under the assumptions above and assuming H_1 is true and that σ^2 is known, the marginal density of the data $m_1(\mathbf{x}_1, \mathbf{x_2} \mid \sigma^2)$ is given by*

$$\frac{\sqrt{2\pi}c_1}{\sigma\sqrt{n(m\tau^2+1)^3}}\left(\sigma^2 + \frac{m^2\tau^2(\bar{x}_1-\bar{x}_2)^2}{m\tau^2+1}\right)\exp\left\{-\frac{1}{2\sigma^2}\left[\frac{m(\bar{x}_1-\bar{x}_2)^2}{m\tau^2+1}\right]\right\}, \tag{B.105}$$

where we define the following quantities for $i = 1, 2$:

$$\bar{x_i} = \sum_{j=1}^{n_i} x_{j,i}/n_i, \qquad S_i = \sum_{j=1}^{n_i}(x_{j,i}-\bar{x_i})^2, \qquad n = n_1 + n_2 \tag{B.106}$$

$$m = \frac{n_1 n_2}{n_1+n_2}, \qquad c_1 = \frac{1}{2a}(2\pi\sigma^2)^{-(n_1+n_2)/2}\exp\left[-\frac{1}{2\sigma^2}(S_1+S_2)\right]. \tag{B.107}$$

Proof: When not indicated otherwise, we assume that all sums and products extend from $i = 1$ to 2, and that integrals extend from $-\infty$ to ∞. We also define

$$c_2(\delta) = c_1\, p(\delta \mid \tau^2, \sigma^2).$$

The marginal density $m_1(\mathbf{x}_1, \mathbf{x_2} \mid \sigma^2)$ (ignoring dependence on σ^2 and τ^2) can be expressed as

$$\begin{aligned}
& m_1(\mathbf{x}_1, \mathbf{x_2} \mid \sigma^2) && \text{(B.108)}\\
= & \int\int_{-a}^{a} \frac{p(\delta \mid \tau^2, \sigma^2)}{2a}\prod_{j=1}^{n_1}\phi(x_{1,j} \mid \mu_1, \sigma^2)\prod_{j=1}^{n_2}\phi(x_{2,j} \mid \mu_1+\delta, \sigma^2)d\mu_1\, d\delta && \text{(B.109)}\\
\doteq & \int\int c_2(\delta)\exp\left\{-\frac{1}{2\sigma^2}\left[n_1(\bar{x_1}-\mu_1)^2 + n_2(\bar{x_2}-\mu_1-\delta)^2\right]\right\}d\mu_1\, d\delta. && \text{(B.110)}
\end{aligned}$$

Defining $b = [n_1\bar{x}_1 + n_2(\bar{x}_2 - \delta)]/n$, completing the square in μ_1 and integrating leads to

$$= \int\int c_2(\delta) \exp\left\{-\frac{1}{2\sigma^2}\left[n(\mu_1 - b)^2 - nb^2 + n_1\bar{x}_1^2 + n_2(\bar{x}_2 - \delta)^2\right]\right\} d\mu_1\, d\delta \quad \text{(B.111)}$$

$$= \int \frac{c_1\delta^2}{\tau^3\sigma^2\sqrt{n}} \exp\left\{-\frac{1}{2\sigma^2}\left[\frac{\delta^2}{\tau^2} + n_1\bar{x}_1^2 + n_2(\bar{x}_2 - \delta)^2 - nb^2\right]\right\} d\delta. \quad \text{(B.112)}$$

Completing the square in δ and defining $d = [m(\bar{x}_2 - \bar{x}_1)]$ and $f = (m + 1/\tau^2)$ leads to

$$= \int \frac{c_1\delta^2}{\tau^3\sigma^2\sqrt{n}} \exp\left\{-\frac{1}{2\sigma^2}\left[m(\bar{x}_1 - \bar{x}_2)^2 - \frac{d^2}{f} + f\left(\delta - \frac{d}{f}\right)^2\right]\right\} d\delta \quad \text{(B.113)}$$

Noting that the integral is proportional to the second moment of a normal density with mean d/f and variance σ^2/f results in

$$= \frac{\sqrt{2\pi}c_1}{\tau^3\sigma\sqrt{nf}}\left(\frac{\sigma^2}{f} + \frac{d^2}{f^2}\right) \exp\left\{-\frac{1}{2\sigma^2}\left[\frac{m(\bar{x}_1 - \bar{x}_2)^2}{m\tau^2 + 1}\right]\right\}. \quad \text{(B.114)}$$

$$= \frac{\sqrt{2\pi}c_1}{\sigma\sqrt{n(m\tau^2+1)^3}}\left(\sigma^2 + \frac{m^2\tau^2(\bar{x}_1 - \bar{x}_2)^2}{m\tau^2 + 1}\right) \exp\left\{-\frac{1}{2\sigma^2}\left[\frac{m(\bar{x}_1 - \bar{x}_2)^2}{m\tau^2 + 1}\right]\right\} \quad \text{(B.115)}$$

■

Theorem B.3.2. *Under the assumptions of Theorem B.3.1, but now assuming H_0 to be true, the marginal density of the data is given by*

$$m_0(\mathbf{x}_1, \mathbf{x}_2 \,|\, \sigma^2) = \frac{\sqrt{2\pi}\sigma c_1}{\sqrt{n}} \exp\left\{-\frac{1}{2\sigma^2}\left[m(\bar{x}_1 - \bar{x}_2)^2\right]\right\}. \quad \text{(B.116)}$$

Proof: Using the proof of Theorem 1, divide equation (B.112) by $p(\delta \,|\, \tau^2, \sigma^2)$ and set $\delta = 0$ to obtain the marginal density of the data under H_0 after marginalizing over $\mu \sim U(-a, a)$. Simplifying the result in the exponential term yields (B.116).

■

Theorem B.3.3. *Under the assumptions of Thm B.3.1 and B.3.2, the Bayes factor in favor of the alternative hypothesis H_1 against the null hypothesis H_0 is given by*

$$\mathrm{BF}_{10}(\mathbf{x}_1, \mathbf{x}_2 \,|\, \sigma^2) = (m\tau^2 + 1)^{-3/2} \left(1 + rT^2\right) \exp\left(\frac{rT^2}{2}\right), \tag{B.117}$$

where $r = 1/\left(1 + (m\tau^2)^{-1}\right)$ and $T = \sqrt{m}(\bar{x}_2 - \bar{x}_1)/\sigma$.

Proof: Following the definition of the Bayes factor and substituting the expression for the marginal density of $(\mathbf{x}_1, \mathbf{x}_2)$ from Thm B.3.1 and B.3.2 leads to

$$\begin{aligned}
& \mathrm{BF}_{10}(\mathbf{x}_1, \mathbf{x}_2 \,|\, \sigma^2) && \text{(B.118)} \\
= \; & \frac{m_1(\mathbf{x}_1, \mathbf{x}_2 \,|\, \sigma^2)}{m_0(\mathbf{x}_1, \mathbf{x}_2 \,|\, \sigma^2)} && \text{(B.119)} \\
= \; & \frac{1}{\sigma^2(m\tau^2 + 1)^{3/2}} \left[\sigma^2 + \frac{m^2\tau^2(\bar{x}_1 - \bar{x}_2)^2}{m\tau^2 + 1}\right] \exp\left[\frac{m^2\tau^2(\bar{x}_1 - \bar{x}_2)^2}{2\sigma^2(m\tau^2 + 1)}\right] && \text{(B.120)} \\
= \; & (m\tau^2 + 1)^{-3/2} \left(1 + rT^2\right) \exp\left(\frac{rT^2}{2}\right). && \text{(B.121)}
\end{aligned}$$

■

B.3.1.2 One-sided tests

Assume the conditions for the two-sided, two-sample z test hold, except that we now wish to test $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_2 > \mu_1$. To this end, under both H_0 and H_1 we similarly assume the $U(-a, a)$ prior on μ_1 is for some large a. Under H_1 we still assume that $\mu_2 = \mu_1 + \delta$ except the prior on δ is assumed to be a normal moment prior truncated on $(0, \infty)$. The density is specified by

$$p_+(\delta \,|\, \tau^2, \sigma^2) = \frac{\sqrt{2}}{\sqrt{\pi}\tau^3\sigma^3}\delta^2 \exp\left(-\frac{\delta^2}{2\tau^2\sigma^2}\right) \quad \text{for } \delta > 0. \tag{B.122}$$

Under this setup we note that the marginal density of $\mathbf{x}$ under the null hypothesis $H_0 : \mu = 0$ is the same as in Theorem B.3.2.

Theorem B.3.4. *Under the assumptions above and assuming H_1 is true and that σ^2 is known, the*

marginal density of the data $m_1(\mathbf{x}_1, \mathbf{x_2} \,|\, \sigma^2)$ *is given by*

$$\frac{\sqrt{2\pi}\sigma c_1}{\sqrt{n}\,(m\tau^2+1)^{3/2}} \exp\left(-\frac{d_1^2}{m\tau^2}\right)\left[\left(2d_1^2+1\right)\left(1-\textit{erf}(-d_1)\right)+\frac{2d_1}{\sqrt{\pi}}\exp\left(-d_1^2\right)\right], \qquad \text{(B.123)}$$

where m, $\bar{x}_1$, $\bar{x}_2$, S_1, S_2, c_1 *are as in (B.106)–(B.107),* $d = m(\bar{x}_2 - \bar{x}_1)$, $f = (m + 1/\tau^2)$, *and* $d_1 = d/\sigma\sqrt{2f}$.

Proof: The marginal density $m_1(\mathbf{x}_1, \mathbf{x_2} \,|\, \sigma^2)$ (ignoring dependence on σ^2 and τ^2) can be expressed as

$$m_1(\mathbf{x}_1, \mathbf{x_2} \,|\, \sigma^2) = \int_0^\infty \int_{-a}^{a} \frac{p_+(\delta \,|\, \tau^2, \sigma^2)}{2a} \prod_{j=1}^{n_1} \phi(x_{1,j} \,|\, \mu_1, \sigma^2) \prod_{j=1}^{n_2} \phi(x_{2,j} \,|\, \mu_1 + \delta, \sigma^2) d\mu_1 \, d\delta \qquad \text{(B.124)}$$

Marginalizing over μ_1 and following (B.112) leads to

$$m_1(\mathbf{x}_1, \mathbf{x_2} \,|\, \sigma^2) \doteq \int_0^\infty \frac{2c_1\delta^2}{\tau^3\sigma^2\sqrt{n}} \exp\left\{-\frac{1}{2\sigma^2}\left[\frac{\delta^2}{\tau^2} + n_1\bar{x}_1^2 + n_2(\bar{x}_2-\delta)^2 - nb^2\right]\right\} d\delta, \quad \text{(B.125)}$$

where c_1 is as in (B.107). Completing the square in δ and defining $d = [m(\bar{x}_2 - \bar{x}_1)]$ and $f = (m + 1/\tau^2)$ lead to

$$\begin{aligned}
& m_1(\mathbf{x}_1, \mathbf{x_2} \,|\, \sigma^2) && \text{(B.126)}\\
= \; & \int_0^\infty \frac{2c_1\delta^2}{\tau^3\sigma^2\sqrt{n}} \exp\left\{-\frac{1}{2\sigma^2}\left[m(\bar{x}_1-\bar{x}_2)^2 - \frac{d^2}{f} + f\left(\delta - \frac{d}{f}\right)^2\right]\right\} d\delta && \text{(B.127)}\\
= \; & \frac{2c_1}{\tau^3\sigma^2\sqrt{n}} \exp\left\{-\frac{1}{2\sigma^2}\left[\frac{m(\bar{x}_1-\bar{x}_2)^2}{m\tau^2+1}\right]\right\} \int_0^\infty \delta^2 \exp\left[-\frac{f\,(\delta - d/f)^2}{2\sigma^2}\right] d\delta. && \text{(B.128)}
\end{aligned}$$

Using 2.1.3.1 from [105] and some algebraic simplifications result in

$$
\begin{aligned}
& m_1(\mathbf{x}_1, \mathbf{x_2} \mid \sigma^2) && \text{(B.129)} \\
= \; & \frac{2c_1}{\tau^3\sigma^2\sqrt{n}} \exp\left\{-\frac{1}{2\sigma^2}\left[\frac{m(\bar{x}_1-\bar{x}_2)^2}{m\tau^2+1}\right]\right\} \times && \text{(B.130)} \\
& \left[\frac{\sqrt{\pi}\sigma^3}{\sqrt{2}f^{3/2}}\left(\frac{d^2}{\sigma^2 f}+1\right)\left\{1-\text{erf}\left(-\frac{d}{\sigma\sqrt{2f}}\right)\right\}+\frac{d\sigma^2}{f^2}\exp\left(-\frac{d^2}{2\sigma^2 f}\right)\right] && \text{(B.131)} \\
= \; & \frac{\sqrt{2\pi}\sigma c_1}{\sqrt{n}\left(m\tau^2+1\right)^{3/2}}\exp\left(-\frac{d_1^2}{m\tau^2}\right)\left[\left(2d_1^2+1\right)\left(1-\text{erf}\left(-d_1\right)\right)+\frac{2d_1}{\sqrt{\pi}}\exp\left(-d_1^2\right)\right] && \text{(B.132)}
\end{aligned}
$$

■

Theorem B.3.5. *Under the assumptions stated above, the Bayes factor* $\text{BF}_{10}(\mathbf{x}_1, \mathbf{x}_2 \mid \sigma^2)$ *in favor of the alternative hypothesis* H_1 *against the null hypothesis* H_0 *is given by*

$$
\left(m\tau^2+1\right)^{-3/2}\exp\left(\frac{rT^2}{2}\right)\left[\left(rT^2+1\right)\left(1-\textit{erf}\left(-\frac{\sqrt{r}T}{\sqrt{2}}\right)\right)+\frac{\sqrt{2r}T}{\sqrt{\pi}}\exp\left(-\frac{rT^2}{2}\right)\right], \tag{B.133}
$$

where $r = 1/\left(1+(m\tau^2)^{-1}\right)$ *and* $T = \sqrt{m}(\bar{x}_2-\bar{x}_1)/\sigma$.

Proof: Following the definition of the Bayes factor and substituting the expression for the marginal

density of $(\mathbf{x}_1, \mathbf{x}_2)$ from Thm B.3.4 and B.3.2 leads to

$$
\begin{aligned}
& \mathrm{BF}_{10}(\mathbf{x}_1, \mathbf{x}_2 \,|\, \sigma^2) && \text{(B.134)} \\
= \;& \frac{m_1(\mathbf{x}_1, \mathbf{x}_2 \,|\, \sigma^2)}{m_0(\mathbf{x}_1, \mathbf{x}_2 \,|\, \sigma^2)} && \text{(B.135)} \\
= \;& \left(m\tau^2 + 1\right)^{-3/2} \exp\left(\frac{m^2\tau^2(\bar{x}_2 - \bar{x}_1)^2}{2\sigma^2(m\tau^2+1)}\right) \times && \text{(B.136)} \\
& \left[\left(\frac{m^2\tau^2(\bar{x}_2 - \bar{x}_1)^2}{\sigma^2(m\tau^2+1)} + 1\right)\left(1 - \mathrm{erf}\left(-\frac{m\tau(\bar{x}_2 - \bar{x}_1)}{\sigma\sqrt{2(m\tau^2+1)}}\right)\right) + \right. && \text{(B.137)} \\
& \left. \frac{\sqrt{2}m\tau(\bar{x}_2 - \bar{x}_1)}{\sigma\sqrt{\pi(m\tau^2+1)}} \exp\left(-\frac{m^2\tau^2(\bar{x}_2 - \bar{x}_1)^2}{2\sigma^2(m\tau^2+1)}\right)\right] && \text{(B.138)} \\
= \;& \left(m\tau^2 + 1\right)^{-3/2} \exp\left(\frac{rT^2}{2}\right) \times && \text{(B.139)} \\
& \left[\left(rT^2 + 1\right)\left(1 - \mathrm{erf}\left(-\frac{\sqrt{r}T}{\sqrt{2}}\right)\right) + \frac{\sqrt{2r}T}{\sqrt{\pi}} \exp\left(-\frac{rT^2}{2}\right)\right]. && \text{(B.140)}
\end{aligned}
$$

■

B.3.2 Variance unknown

B.3.2.1 Two-sided tests

We now consider the case where the variance σ^2 is not known. In this case, we assume that σ^2 is drawn *a priori* from an inverse gamma density parameterized as in (B.51).

Theorem B.3.6. *Under the assumptions stated above and assuming H_1 is true, the marginal density of the data is given by*

$$
m_1(\mathbf{x}_1, \mathbf{x}_2) = \frac{c_3 c_4^{-(n+1)/2-\alpha}}{(m\tau^2+1)^{3/2}}\left[c_4 + \frac{m^2\tau^2(\bar{x}_1 - \bar{x}_2)^2}{m\tau^2+1}\left(\frac{n-1}{2} + \alpha\right)\right], \qquad \text{(B.141)}
$$

where $\bar{x}_1$, $\bar{x}_2$, S_1, S_2, n, m are defined in (B.106-B.107), and

$$
c_3 = \frac{(2\pi)^{-(n-1)/2}\,\beta^\alpha}{2a\sqrt{n}\,\Gamma(\alpha)}\,\Gamma\left(\frac{n-1}{2} + \alpha\right), \quad \textit{and} \quad c_4 = \frac{m(\bar{x}_1 - \bar{x}_2)^2}{2(m\tau^2+1)} + \frac{S_1 + S_2}{2} + \beta. \qquad \text{(B.142)}
$$

Proof: The marginal density $m_1(\mathbf{x}_1, \mathbf{x}_2)$ (ignoring dependence on τ^2) can be expressed as

$$m_1(\mathbf{x}_1, \mathbf{x}_2) = \int\int\int \pi(\sigma^2 \,|\, \alpha, \beta)\, \pi(\mu_1 \,|\, a)\, \pi(\delta \,|\, \tau^2, \sigma^2) \times \tag{B.143}$$

$$\prod_{j=1}^{n_1} \phi(x_{1,j} \,|\, \mu_1, \sigma^2) \prod_{j=1}^{n_2} \phi(x_{2,j} \,|\, \mu_1 + \delta, \sigma^2)\, d\mu_1\, d\delta\, d\sigma^2. \tag{B.144}$$

To this we note that, given σ^2 the integral with respect to μ_1 and δ is identical to (B.109). From Theorem B.3.1, and noting that the integrals with respect to σ^2 are proportional to an inverse gamma density yields

$$\begin{aligned} m_1(\mathbf{x}_1, \mathbf{x}_2) = \; & \frac{(2\pi)^{-(n_1+n_2-1)/2}\beta^\alpha}{2a\sqrt{n(m\tau^2+1)^3}\,\Gamma(\alpha)} \int (\sigma^2)^{-(n_1+n_2+1)/2-\alpha-1} \left(\sigma^2 + \frac{m^2\tau^2(\bar{x}_1 - \bar{x}_2)^2}{m\tau^2+1}\right) \\ & \times \exp\left[-\frac{1}{\sigma^2}\left\{\frac{m(\bar{x}_1-\bar{x}_2)^2}{2(m\tau^2+1)} + \frac{S_1+S_2}{2} + \beta\right\}\right] d\sigma^2 \qquad \text{(B.145)} \\ = \; & \frac{(2\pi)^{-(n_1+n_2-1)/2}\beta^\alpha}{2a\sqrt{n(m\tau^2+1)^3}\,\Gamma(\alpha)} \left[\Gamma\left(\frac{n_1+n_2+1}{2} + \alpha - 1\right) c_4^{-(n_1+n_2+1)/2-\alpha+1}\right. \\ & + \left. \frac{m^2\tau^2(\bar{x}_1-\bar{x}_2)^2}{m\tau^2+1}\,\Gamma\left(\frac{n_1+n_2+1}{2} + \alpha\right) c_4^{-(n_1+n_2+1)/2-\alpha}\right] \qquad \text{(B.146)} \\ = \; & \frac{c_3 c_4^{-(n+1)/2-\alpha}}{\sqrt{n}} \left[c_4 + \frac{m^2\tau^2(\bar{x}_1-\bar{x}_2)^2}{m\tau^2+1}\left(\frac{n-1}{2} + \alpha\right)\right]. \qquad \text{(B.147)} \end{aligned}$$

■

Theorem B.3.7. *Under the assumptions of Theorem B.3.6, but now assuming H_0 to be true, the marginal density of the data is given by*

$$m_0(\mathbf{x}_1, \mathbf{x}_2) = c_3 \left[\frac{m(\bar{x}_1-\bar{x}_2)^2}{2} + \frac{S_1+S_2}{2} + \beta\right]^{-(n-1)/2-\alpha}, \tag{B.148}$$

where c_3 is as in (B.142).

Proof: The marginal density $m_0(\mathbf{x}_1, \mathbf{x}_2)$ can be expressed as

$$m_0(\mathbf{x}_1, \mathbf{x}_2) = \int \int \pi(\sigma^2 \,|\, \alpha, \beta)\, \pi(\mu \,|\, a) \prod_{i=1}^{2} \prod_{j=1}^{n_i} \phi(x_{i,j} \,|\, \mu, \sigma^2)\, d\mu\, d\sigma^2. \tag{B.149}$$

To this we note that, given σ^2 the integral with respect to μ is the same as the marginal $m_0(\mathbf{x}_1, \mathbf{x}_2 \,|\, \sigma^2)$ in Theorem (B.3.2). Using (B.116) and noting that the integrals with respect to σ^2 are proportional to an Inverse-gamma density results in

$$\begin{aligned}
m_0(\mathbf{x}_1, \mathbf{x}_2) = \quad & \frac{(2\pi)^{-(n_1+n_2-1)/2} \beta^{\alpha}}{2a\sqrt{n}\,\Gamma(\alpha)} \int (\sigma^2)^{-(n_1+n_2-1)/2-\alpha-1} \times & \text{(B.150)} \\
& \exp\left[-\frac{1}{\sigma^2} \left\{ \frac{m(\bar{x}_1 - \bar{x}_2)^2}{2} + \frac{S_1 + S_2}{2} + \beta \right\} \right] d\sigma^2 & \text{(B.151)} \\
= \quad & \frac{(2\pi)^{-(n_1+n_2-1)/2} \beta^{\alpha}}{2a\sqrt{n}\,\Gamma(\alpha)} \Gamma\left(\frac{n_1 + n_2 - 1}{2} + \alpha \right) \times & \text{(B.152)} \\
& \left[\frac{m(\bar{x}_1 - \bar{x}_2)^2}{2} + \frac{S_1 + S_2}{2} + \beta \right]^{-(n_1+n_2-1)/2-\alpha} & \text{(B.153)} \\
= \quad & c_3 \left[\frac{m(\bar{x}_1 - \bar{x}_2)^2}{2} + \frac{S_1 + S_2}{2} + \beta \right]^{-(n-1)/2-\alpha}. & \text{(B.154)}
\end{aligned}$$

■

Theorem B.3.8. *Under the assumptions of Thm B.3.6 and B.3.7, the Bayes factor in favor of the alternative hypothesis H_1 against the null hypothesis H_0 is given by*

$$\mathrm{BF}_{10}(\mathbf{x}_1, \mathbf{x}_2) = (m\tau^2 + 1)^{-3/2} \left(\frac{G_2}{H_2} \right)^{(n-1)/2+\alpha} \left(1 + \frac{qT_2^2}{H_2} \right), \tag{B.155}$$

where $\bar{x}_1$, $\bar{x}_2$, S_1, S_2, n, m are defined in (B.106)–(B.107), and

$$r = \frac{m\tau^2}{m\tau^2 + 1}, \qquad q = \frac{2r\,((n-1)/2 + \alpha)}{n-2}, \qquad S = S_1 + S_2, \tag{B.156}$$

$$T = \frac{\sqrt{m}(\bar{x}_1 - \bar{x}_2)}{\sqrt{S/(n-2)}}, \qquad G = 1 + \frac{T^2}{n-2} + \frac{2\beta}{S}, \qquad H = 1 + \frac{(1-r)T^2}{n-2} + \frac{2\beta}{S}. \tag{B.157}$$

Proof: Following the definition of the Bayes factor and substituting the expression for the marginal density of $(\mathbf{x}_1, \mathbf{x}_2)$ from Thm B.3.6 and B.3.7 leads to

$$
\begin{aligned}
& \mathrm{BF}_{10}(\mathbf{x}_1, \mathbf{x}_2) && \text{(B.158)} \\
= \; & \frac{m_1(\mathbf{x}_1, \mathbf{x}_2)}{m_0(\mathbf{x}_1, \mathbf{x}_2)} && \text{(B.159)} \\
= \; & \frac{1}{(m\tau^2+1)^{3/2}} \left[\frac{m(\bar{x}_1 - \bar{x}_2)^2/2 + S/2 + \beta}{m(\bar{x}_1 - \bar{x}_2)^2/2(m\tau^2+1) + S/2 + \beta} \right]^{(n-1)/2+\alpha} \times && \text{(B.160)} \\
& \left[1 + \frac{m^2\tau^2(\bar{x}_1 - \bar{x}_2)^2 \left((n-1)/2 + \alpha\right)/(m\tau^2+1)}{m(\bar{x}_1 - \bar{x}_2)^2/2(m\tau^2+1) + S/2 + \beta} \right] && \text{(B.161)} \\
= \; & (m\tau^2+1)^{-3/2} \left[\frac{1 + T^2/(n-2) + 2\beta/S}{1 + T^2/\left\{(n-2)(m\tau^2+1)\right\} + 2\beta/S} \right]^{(n-1)/2+\alpha} \times && \text{(B.162)} \\
& \left[1 + \frac{2m\tau^2\left((n-1)/2 + \alpha\right)}{(m\tau^2+1)} \frac{T^2/(n-2)}{1 + T^2/\left\{(n-2)(m\tau^2+1)\right\} + 2\beta/S} \right] && \text{(B.163)} \\
= \; & (m\tau^2+1)^{-3/2} \left(\frac{G}{H} \right)^{(n-1)/2+\alpha} \left(1 + \frac{qT^2}{H} \right). && \text{(B.164)}
\end{aligned}
$$

■

B.3.2.2 One-sided tests

Assume the conditions of the two-sample, two-sided t test hold, except that we now wish to test $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_2 > \mu_1$. To this end, under both H_0 and H_1 we similarly assume the $U(-a, a)$ prior on μ_1 is for some large a. Under H_1 we still assume that $\mu_2 = \mu_1 + \delta$, but the prior on δ given σ^2 is assumed to be a normal moment prior truncated on $(0, \infty)$ whose density is defined by (B.122). To complete the model specification, under both H_0 and H_1 we again assume an inverse gamma prior on σ^2 defined by (B.51). Under these assumptions and assuming H_0 to be true, the marginal density of the data $m_0(\mathbf{x}_1, \mathbf{x}_2)$ is the same as in Theorem B.3.7.

Theorem B.3.9. *Under the assumptions stated above and assuming H_1 is true, the marginal den-*

sity of the data $m_1(\mathbf{x}_1, \mathbf{x}_2)$ *is given by*

$$\begin{cases} c^* \left(f_1 d_1^2 \left(1 - F_{2\nu-1}\left(-d_1\sqrt{2\nu-1}\right)\right) + f_2\, d_1\, |d_1|^{2(1-\nu)} + f_3\, |d_1|^{3-2\nu} \right) & \text{if } \bar{x}_2 < \bar{x}_1, \\ c^* \mathrm{B}\left(3/2, \nu - 3/2\right) & \text{if } \bar{x}_2 = \bar{x}_1, \\ c^* \left(f_1 d_1^2 \left(1 - F_{2\nu-1}\left(-d_1\sqrt{2\nu-1}\right)\right) + f_2\, d_1\, |d_1|^{2(1-\nu)} + \right. & \\ \quad \left. f_3\, |d_1|^{3-2\nu} + 2 f_4\, |d_1|^3 \right) & \text{if } \bar{x}_2 > \bar{x}_1, \end{cases} \tag{B.165}$$

where $\bar{x}_1$, $\bar{x}_2$, S_1, S_2, n, m *are defined in (B.106-B.107),* $S = S_1 + S_2$, $d = m\,(\bar{x}_2 - \bar{x}_1)$, $f = m + \tau^{-2}$, $d_1 = d/\sqrt{2fA_1}$, $\nu = n/2 + \alpha + 1$,

$$c^* = \frac{2^{3/2}(2\pi)^{-n/2}\,\beta^\alpha\,\Gamma(\nu)}{a\Gamma(\alpha)\sqrt{n}(m\tau^2+1)^{3/2}\,A_1^{(n-1)/2+\alpha}}, \quad A_1 = \beta + \frac{S}{2} + \frac{d^2}{2m\,(m\tau^2+1)}, \tag{B.166}$$

$$f_1 = \mathrm{B}\left(\nu - 1/2, 1/2\right), \quad f_2 = \frac{{}_2F_1\left(\nu, \nu-1; \nu; -1/d_1^2\right)}{(\nu-1)}, \tag{B.167}$$

$$f_3 = \frac{{}_2F_1\left(\nu, \nu-3/2; \nu-1/2; -1/d_1^2\right)}{2\nu-3}, \quad f_4 = \frac{{}_2F_1\left(\nu, 3/2; 5/2; -d_1^2\right)}{3}, \tag{B.168}$$

$\mathrm{B}(\cdot,\cdot)$ *is the Beta function,* $F_{2\nu-1}$ *is the cdf of the Student's* t *distribution (center 0 and scale 1) with degrees of freedom* $2\nu - 1$*, and* ${}_2F_1$ *is the Gauss hypergeometric function.*

Proof: The marginal density $m_1(\mathbf{x}_1, \mathbf{x}_2)$ (ignoring dependence on τ^2) can be expressed as

$$m_1(\mathbf{x}_1, \mathbf{x}_2) = \int_0^\infty \int_0^\infty \int_{-a}^a p_+(\delta \mid \tau^2, \sigma^2)\, \pi(\sigma^2 \mid \alpha, \beta)\, \pi(\mu_1 \mid a) \times \tag{B.169}$$

$$\prod_{j=1}^{n_1} \phi(x_{1,j} \mid \mu_1, \sigma^2) \prod_{j=1}^{n_2} \phi(x_{2,j} \mid \mu_1 + \delta, \sigma^2)\, d\mu_1\, d\sigma^2\, d\delta. \tag{B.170}$$

Following (B.125) integrating with respect to μ_1, and then integrating with respect to σ^2 leads to

$$m_1(\mathbf{x}_1, \mathbf{x_2}) \doteq \int_0^\infty \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-\alpha-1} \exp\left(-\frac{\beta}{\sigma^2}\right) \frac{(2\pi\sigma^2)^{-n/2}}{2a} \exp\left(-\frac{S}{2\sigma^2}\right) \times \tag{B.171}$$

$$\frac{2\delta^2}{\tau^3\sigma^2\sqrt{n}} \exp\left\{-\frac{1}{2\sigma^2}\left[\frac{\delta^2}{\tau^2} + n_1\bar{x}_1^2 + n_2(\bar{x}_2 - \delta)^2 - nb^2\right]\right\} d\sigma^2\, d\delta \tag{B.172}$$

$$= \frac{(2\pi)^{-n/2}\beta^\alpha\Gamma(n/2+\alpha+1)}{a\Gamma(\alpha)\tau^3\sqrt{n}} \times \tag{B.173}$$

$$\int_0^\infty \delta^2 \left[\beta + \frac{1}{2}\left\{S + \frac{d^2}{m(m\tau^2+1)} + f\left(\delta - \frac{d}{f}\right)^2\right\}\right]^{-(n/2+\alpha+1)} d\delta. \tag{B.174}$$

Define $\nu = n/2 + \alpha + 1$, $A_1 = \beta + S/2 + d^2/2m(m\tau^2+1)$ and

$$A_2 = \frac{(2\pi)^{-n/2}\beta^\alpha\Gamma(\nu)}{a\Gamma(\alpha)\tau^3\sqrt{n}\, A_1^\nu}. \tag{B.175}$$

Then $m_1(\mathbf{x}_1, \mathbf{x_2})$ simplifies to

$$m_1(\mathbf{x}_1, \mathbf{x_2}) = A_2 \int_0^\infty \delta^2 \left(1 + \frac{f(\delta - d/f)^2}{2A_1}\right)^{-\nu} d\mu \tag{B.176}$$

$$= A_2 \int_{-d/f}^\infty \left(u + \frac{d}{f}\right)^2 \left(1 + \frac{fu^2}{2A_1}\right)^{-\nu} du \tag{B.177}$$

$$= A_2 \left(\frac{d^2}{f^2} I_0\left(-\frac{d}{f}\right) + \frac{2d}{f} I_1\left(-\frac{d}{f}\right) + I_2\left(-\frac{d}{f}\right)\right) \tag{B.178}$$

$$= A_2 \left(m_{10} + m_{11} + m_{12}\right), \tag{B.179}$$

where

$$I_k(g) = \int_g^\infty u^k \left(1 + \frac{fu^2}{2A_1}\right)^{-\nu} du \quad \text{for } g \in \mathbb{R}, k \geq 0, \tag{B.180}$$

$$m_{10} = \frac{d^2}{f^2} I_0\left(-\frac{d}{f}\right), \quad m_{11} = \frac{2d}{f} I_1\left(-\frac{d}{f}\right), \quad m_{12} = I_2\left(-\frac{d}{f}\right). \tag{B.181}$$

For $I_0(-d/f)$, first doing a change of variable with $w/\sqrt{2\nu-1} = \sqrt{f}u/\sqrt{2A_1}$ and then some algebraic simplifications lead to

$$\begin{aligned} I_0\left(-d/f\right) &= \int_{-d/f}^{\infty}\left(1+\frac{fu^2}{2A_1}\right)^{-\nu} du && \text{(B.182)} \\ &= \left(\frac{2A_1}{f(2\nu-1)}\right)^{1/2}\int_{-d\sqrt{\frac{(2\nu-1)}{2fA_1}}}^{\infty}\left(1+\frac{w^2}{2\nu-1}\right)^{-((2\nu-1)+1)/2} dw && \text{(B.183)} \\ &= \left(\frac{2A_1}{f}\right)^{1/2}\mathrm{B}\left((2\nu-1)/2,1/2\right)\left[1-F_{2\nu-1}\left(-d\sqrt{\frac{(2\nu-1)}{2fA_1}}\right)\right], && \text{(B.184)} \end{aligned}$$

where $\mathrm{B}(\cdot,\cdot)$ is the Beta function and $F_{2\nu-1}$ is the cdf of the Student's t distribution (center 0 and scale 1) with degrees of freedom $2\nu-1$. Following this we get

$$m_{10} = \frac{d^2}{f^2}\,I_0\left(-\frac{d}{f}\right) = \left(\frac{2A_1}{f}\right)^{3/2} f_1\, d_1^2\,\left(1-F_{2\nu-1}\left(-d_1\sqrt{2\nu-1}\right)\right), \qquad \text{(B.185)}$$

where $d_1 = d/\sqrt{2fA_1}$ and $f_1 = \mathrm{B}\left(\nu-1/2,1/2\right)$. To calculate $I_1(-d/f)$ and $I_2(-d/f)$ we again use (B.89)–(B.93). Using these we get

$$I_1(|d|/f) = \begin{cases} \frac{|d|^{2(1-\nu)}(2A_1)^{\nu}}{f^{2-\nu}\,2(\nu-1)}\,{}_2F_1\left(\nu,\nu-1;\nu;-2fA_1/d^2\right) & \text{if } \bar{x}_2 \neq \bar{x}_1, \\ \frac{2A_1}{f}\,\mathrm{B}\left(1,\nu-1\right) & \text{if } \bar{x}_2 = \bar{x}_1. \end{cases} \qquad \text{(B.186)}$$

This leads to

$$m_{11} = \begin{cases} (2d/f)\;I_1\left(|d|/f\right) & \text{if } \bar{x}_2 \neq \bar{x}_1, \\ 0 & \text{if } \bar{x}_2 = \bar{x}_1 \end{cases} = \begin{cases} \left(2A_1/f\right)^{3/2}\;f_2\,d_1\,|d_1|^{2(1-\nu)} & \text{if } \bar{x}_2 \neq \bar{x}_1, \\ 0 & \text{if } \bar{x}_2 = \bar{x}_1. \end{cases} \qquad \text{(B.187)}$$

where $f_2 = {}_2F_1\left(\nu,\nu-1;\nu;-1/d_1^2\right)/\left(\nu-1\right)$. Similarly, it also results in

$$I_2(|d|/f) = \begin{cases} \left(2A_1/f\right)^{3/2}\;f_3\;|d_1|^{3-2\nu} & \text{if } \bar{x}_2 \neq \bar{x}_1, \\ \left(2A_1/f\right)^{3/2}\;\mathrm{B}\left(3/2,\nu-3/2\right) & \text{if } \bar{x}_2 = \bar{x}_1, \end{cases} \qquad \text{(B.188)}$$

and

$$J_2(|d|/f) = (2A_1/f)^{3/2}\ f_4\ |d_1|^3\,, \tag{B.189}$$

where $f_3 = {}_2F_1\left(\nu, \nu - 3/2; \nu - 1/2; -1/d_1^2\right)/\left(2\nu - 3\right)$ and $f_4 = {}_2F_1\left(\nu, 3/2; 5/2; -d_1^2\right)/3$. This leads to

$$m_{12} = \begin{cases} (2A_1/f)^{3/2}\ f_3\ |d_1|^{3-2\nu} & \text{if } \bar{x}_2 < \bar{x}_1, \\ (2A_1/f)^{3/2}\ \mathrm{B}\left(3/2, \nu - 3/2\right) & \text{if } \bar{x}_2 = \bar{x}_1, \\ (2A_1/f)^{3/2}\ \left(f_3\ |d_1|^{3-2\nu} + 2f_4\,|d_1|^3\right) & \text{if } \bar{x}_2 > \bar{x}_1. \end{cases} \tag{B.190}$$

Finally, (B.71) follows by combining m_{10}, m_{11} and m_{12}.

■

Theorem B.3.10. *Under the assumptions stated above, the Bayes factor* $\mathrm{BF}_{10}(\mathbf{x}_1, \mathbf{x}_2)$ *in favor of the alternative hypothesis* H_1 *against the null hypothesis* H_0 *is given by*

$$\begin{cases} C_1\left(f_1 d_1^2\left(1 - F_{2\nu-1}\left(-d_1\sqrt{2\nu-1}\right)\right) + f_2\, d_1\,|d_1|^{2(1-\nu)} + f_3\ |d_1|^{3-2\nu}\right) & \textit{if } \bar{x}_2 < \bar{x}_1, \\ C_1 \mathrm{B}\left(3/2, \nu - 3/2\right) & \textit{if } \bar{x}_2 = \bar{x}_1, \\ C_1\left(f_1 d_1^2\left(1 - F_{2\nu-1}\left(-d_1\sqrt{2\nu-1}\right)\right) + f_2\, d_1\,|d_1|^{2(1-\nu)} + \right. & \\ \quad \left. f_3\ |d_1|^{3-2\nu} + 2f_4\,|d_1|^3\right) & \textit{if } \bar{x}_2 > \bar{x}_1, \end{cases} \tag{B.191}$$

where

$$C_1 = \frac{2\Gamma\left(\nu\right)}{\sqrt{\pi}(m\tau^2+1)^{3/2}\Gamma\left((n-1)/2+\alpha\right)}, \tag{B.192}$$

where $\bar{x}_1$, $\bar{x}_2$, S, n, m, ν *are as in Theorem B.3.9,* T, r, G *and* H *are as in (B.156)–(B.157),* $d_1 = \sqrt{r}T/\sqrt{(n-2)H}$, *and* f_1 *to* f_4 *are as in (B.167)–(B.168) with* d_1 *is as it is defined here.*

Proof: Following the definition of the Bayes factor we know that $\mathrm{BF}_{10}(\mathbf{x}_1, \mathbf{x}_2) = m_1(\mathbf{x}_1, \mathbf{x}_2)/m_0(\mathbf{x}_1, \mathbf{x}_2)$. While substituting the expression for the marginal density of $(\mathbf{x}_1, \mathbf{x}_2)$ from Theorem B.3.7 and

B.3.9 we note that

$$\frac{c^*}{m_0(\mathbf{x}_1,\mathbf{x}_2)} = \frac{2\Gamma(\nu)}{\sqrt{\pi}(m\tau^2+1)^{3/2}\Gamma((n-1)/2+\alpha)} \times \tag{B.193}$$

$$\left(\frac{\beta+S/2+m(\bar{x}_2-\bar{x}_1)^2/2}{\beta+S/2+m(\bar{x}_2-\bar{x}_1)^2/2(m\tau^2+1)}\right)^{(n-1)/2+\alpha} \tag{B.194}$$

$$= \frac{2\Gamma(\nu)}{\sqrt{\pi}(m\tau^2+1)^{3/2}\Gamma((n-1)/2+\alpha)}\left(\frac{G}{H}\right)^{(n-1)/2+\alpha}. \tag{B.195}$$

Also, d_1 as in Theorem B.3.9 can be rewritten as

$$d_1 = \frac{m\tau(\bar{x}_2-\bar{x}_1)}{\sqrt{2(m\tau^2+1)\left(\beta+S/2+m(\bar{x}_2-\bar{x}_1)^2/2(m\tau^2+1)\right)}} = \frac{\sqrt{r}T}{\sqrt{(n-2)H}}. \tag{B.196}$$

(B.12) directly follows from combining these.

■

B.4 Operating characteristics of z and t tests

The operating characteristics for one-sample z, and two-sample z and t tests are similar to those cited in the main article for one-sample t tests. For purposes of comparison, plots similar to those found in the main article are presented below.

B.4.1 Fixed design tests

Fig. B.1 displays the operating characteristics of the one-sample z and two-sample t tests under a true null hypothesis. For the two-sample t test, equal sample sizes were assumed drawn from both populations, and the sample size appearing on the horizontal axis refers to the sample size for each sample. This figure is comparable to Fig. 2 in the main article for the default choices of the NAP and JZS priors.

For the same tests, Fig. B.2–B.4 displays the weight of evidence for different effect sizes under the alternative hypothesis as sample size varies. These figures are comparable to Fig. 3 in the main article for the composite alternative placing one-half mass at $\pm 0.3\sigma$ and different choices of the

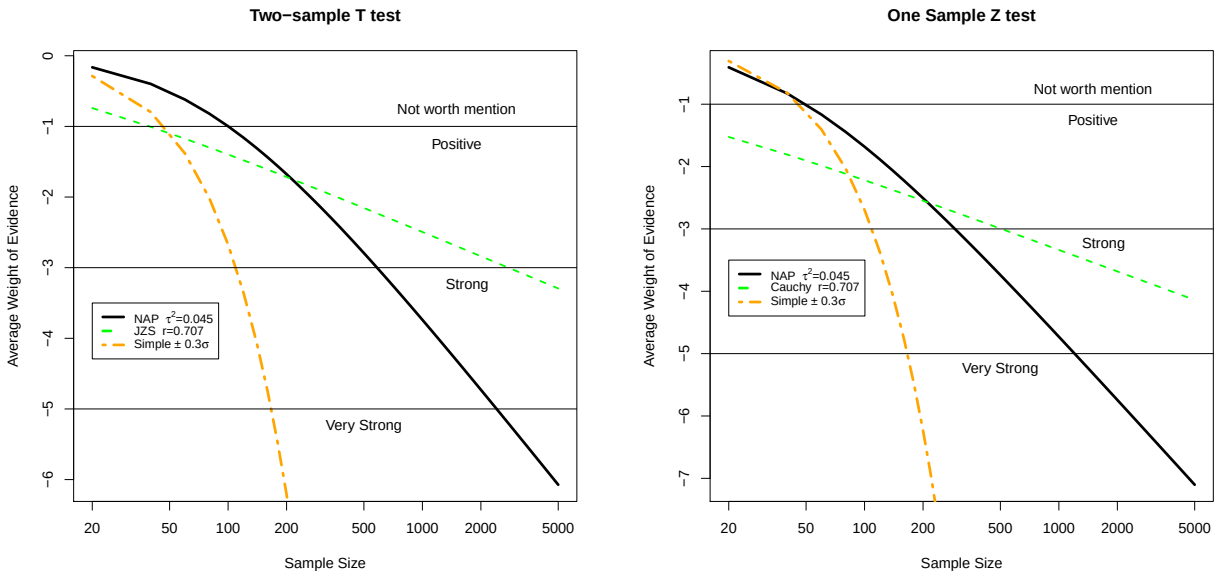


Figure B.1: Weight of evidence for true null hypotheses in two-sample t test and one-sample z test. The black curves represent the average weight of evidence for the default NAP priors, while the dashed green curve the default JZS prior. The dashed orange curve depicts the average weight of evidence obtained when the alternative hypothesis assigned one-half mass to $\pm 0.3\sigma$.

NAP and JZS priors.

B.4.2 Sequential tests

Fig. B.5–B.16 display the operating characteristics of the Hajnal(0.3), default SBF-NAP and default SBF-JZS tests. The results presented below correspond to one-sample z tests and two-sample z and t tests under a true null and alternative hypothesis. For the two-sample tests, equal sample sizes were assumed drawn from both populations. For these tests, the ASN refers to the sample size from each group required on average by the sequential tests. As in the main article, two types of exceedance thresholds were considered: (a) symmetric exceedance thresholds of ± 3 and ± 5, and (b) SPRT thresholds with (α, β) equal to $(0.05, 0.2)$ and $(0.005, 0.05)$. The figures are comparable to Fig. 5–8 in the main article.

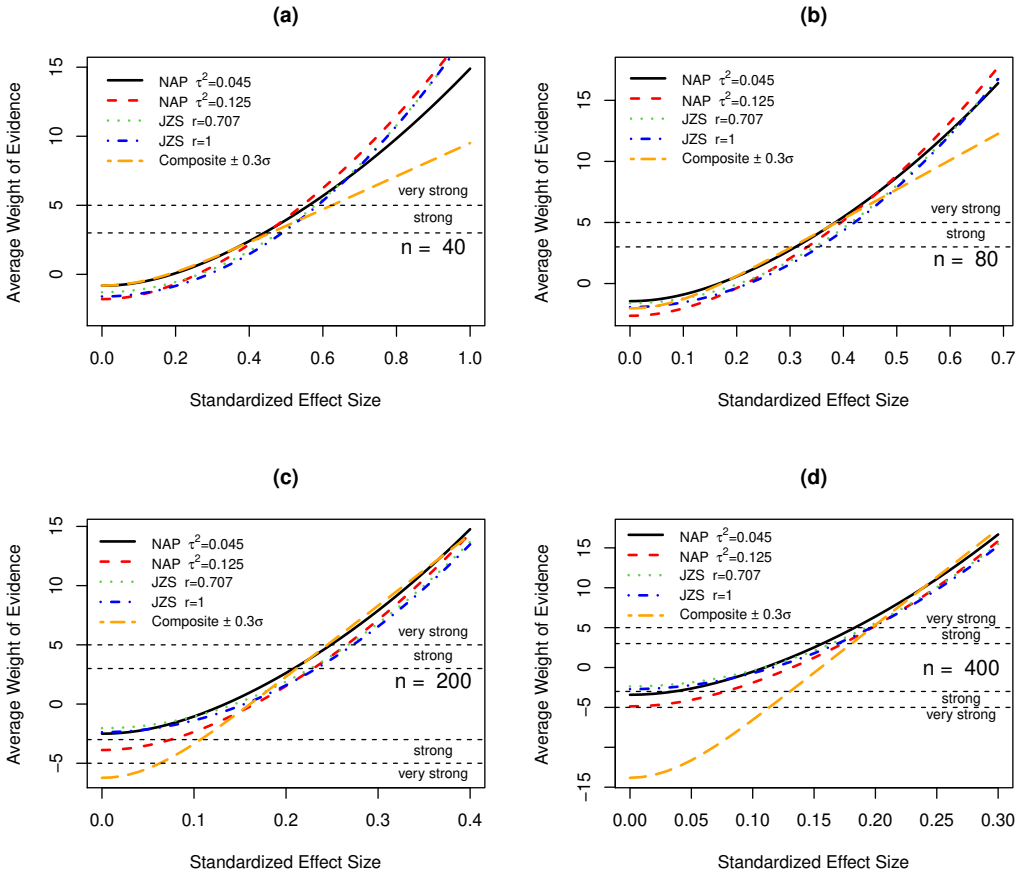


Figure B.2: Weight of evidence for true alternative hypotheses in one-sample z test. Curves depicted in the plots denote the average weight of evidence versus true effect size when the alternative hypothesis was defined by various NAP and JZS densities.

B.4.2.1 Numerical evaluation of symmetric evidence thresholds

Fig. B.5–B.9 display the operating characteristics of Hajnal(0.3), and the SBF-NAP and SBF-JZS with their default choices. The results presented below correspond to the symmetric exceedance thresholds of ± 3 and ± 5. The figures are comparable to Fig. 5–6 in the main article.

B.4.2.2 Numerical evaluation of SPRT thresholds

Fig. B.11–B.16 display the operating characteristics of Hajnal(0.3), default SBF-NAP and default SBF-JZS tests. The results presented below correspond to the SPRT thresholds with (α, β) equal to $(0.05, 0.2)$ and $(0.005, 0.05)$. The figures are comparable to Fig. 7–8 in the main article.

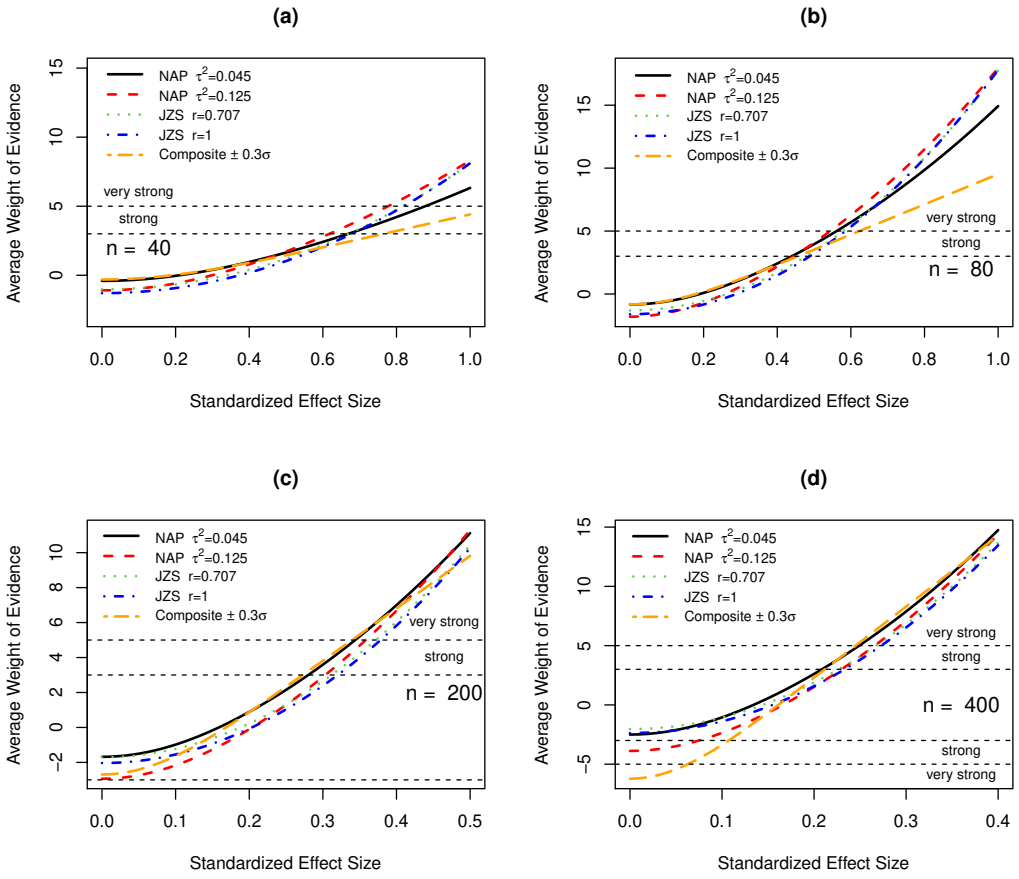


Figure B.3: Weight of evidence for true alternative hypotheses in two-sample z test. Curves depicted in the plots denote the average weight of evidence versus true effect size when the alternative hypothesis was defined by various NAP and JZS densities.

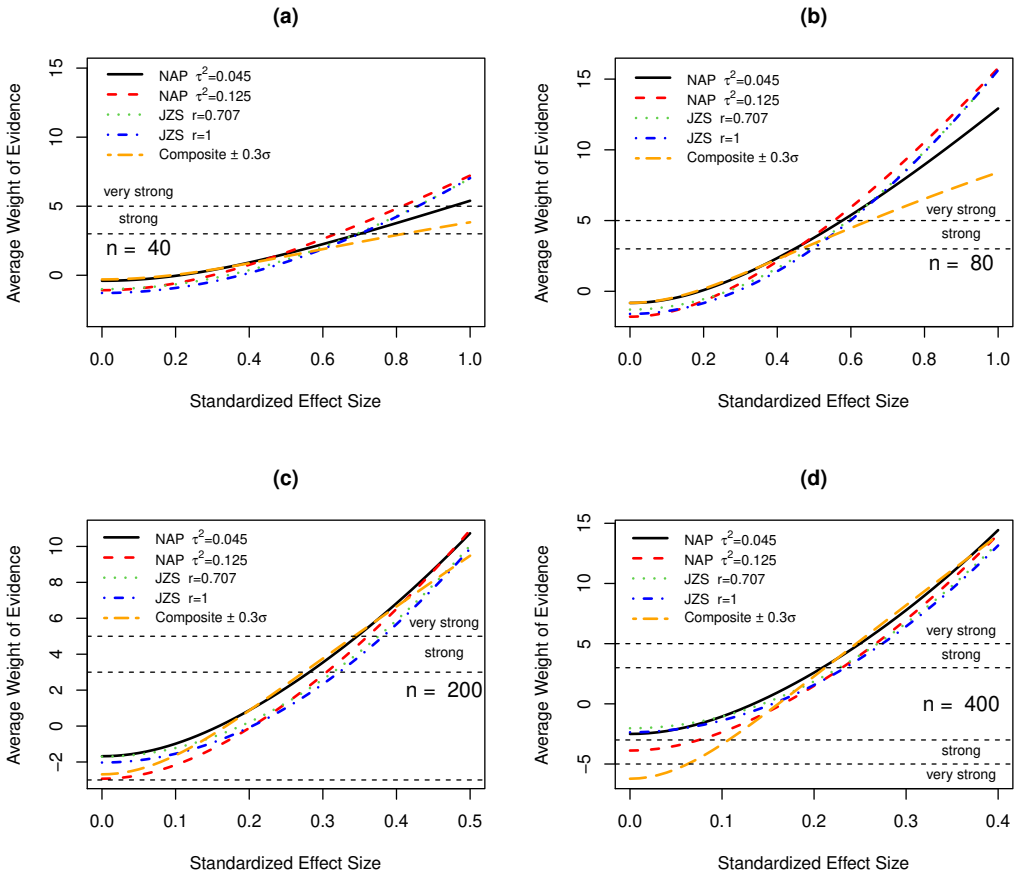


Figure B.4: Weight of evidence for true alternative hypotheses in two-sample t test. Curves depicted in the plots denote the average weight of evidence versus true effect size when the alternative hypothesis was defined by various NAP and JZS densities.

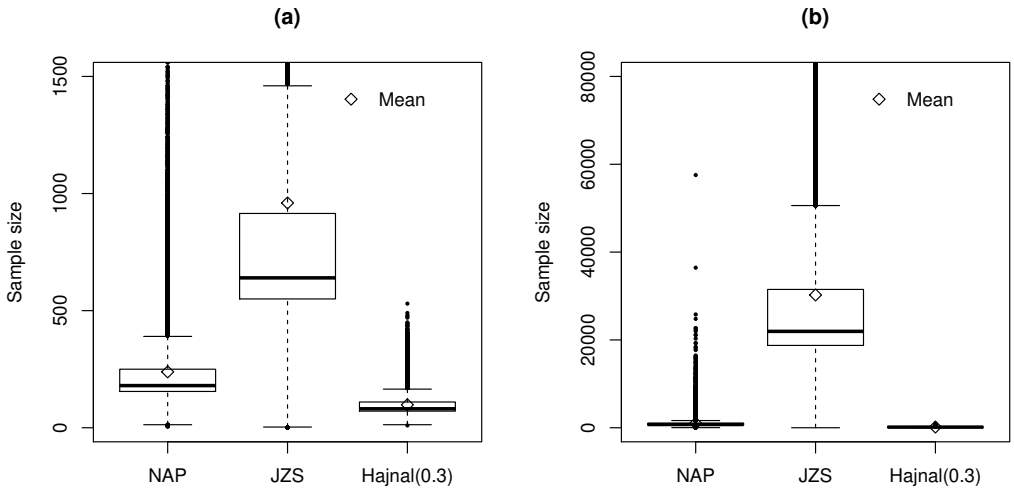


Figure B.5: ASN for sequential procedures under a true null hypothesis in one-sample z test. The plots are truncated at 1500 and 80,000 to enhance comparisons at moderate sample sizes. Panel (a) provides a boxplot estimate of the distribution of sample sizes required for the SBF-NAP, SBF-JZS and Hajnal(0.3) procedures to cross an exceedance threshold of ± 3. About 0.3% percent of SBF-NAP tests and 11% of SBF-JZS tests required more than 1500 samples to reach a decision. All Hajnal(0.3) tests terminated by 530 samples. Panel (b) provides the corresponding boxplots when the exceedance threshold is ± 5. About 4% of SBF-JZS tests required more than 80,000 samples to reach a decision. The black diamonds show the ASN's for each procedure. All SBF-NAP tests reached a decision by 57550 samples, and all Hajnal(0.3) tests terminated by observation 985.

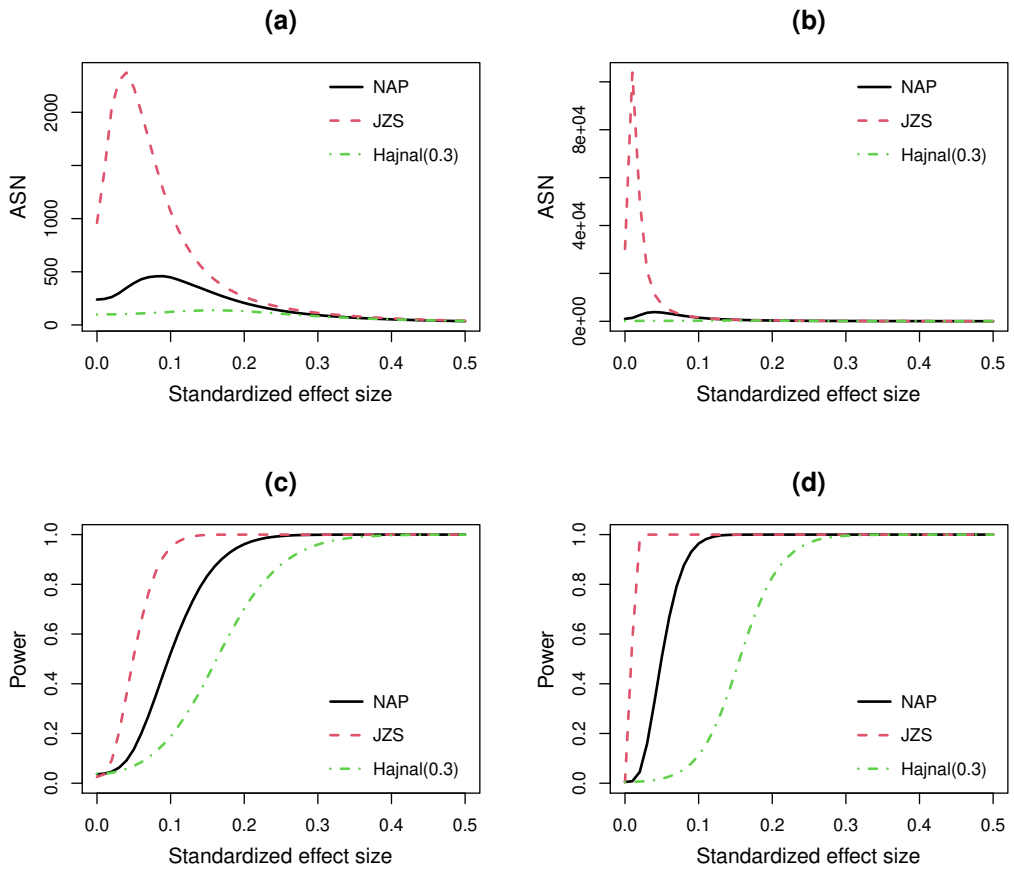


Figure B.6: Operating characteristics under true alternative hypotheses in one-sample z test. Panels (a) and (b) depict the ASN's for three sequential tests when the exceedance thresholds are ± 3 and ± 5, respectively, versus the data-generating value of the standardized effect size. Panels (c) and (d) provide the corresponding probabilities that each test rejects the null hypothesis as a function of the standardized effect size.

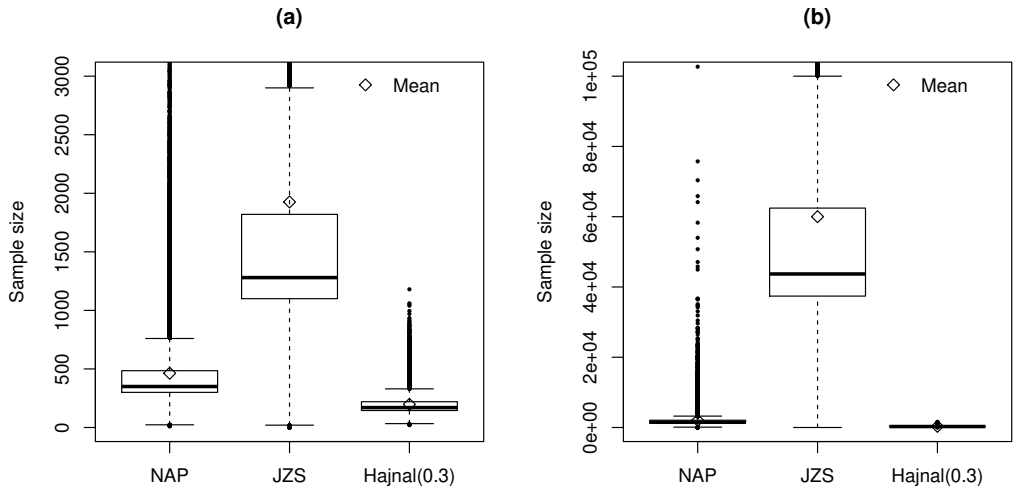


Figure B.7: ASN for sequential procedures under a true null hypothesis in two-sample z test. The plots are truncated at 3000 and 100,000 to enhance comparisons at moderate sample sizes. Panel (a) provides a boxplot estimate of the distribution of sample sizes required from each group for the SBF-NAP, SBF-JZS and Hajnal(0.3) procedures to cross an exceedance threshold of ± 3. About 0.3% of SBF-NAP tests and 11% of SBF-JZS tests required more than 3000 samples from each group to reach a decision. All Hajnal(0.3) tests terminated by 1180 samples. Panel (b) provides the corresponding boxplots when the exceedance threshold is ± 5. About 0.002% of SBF-NAP tests and 10% of SBF-JZS tests required more than 100,000 samples from each group to reach a decision. The black diamonds show the ASN's for each procedure. All Hajnal(0.3) tests terminated by 1600 observations from each group.

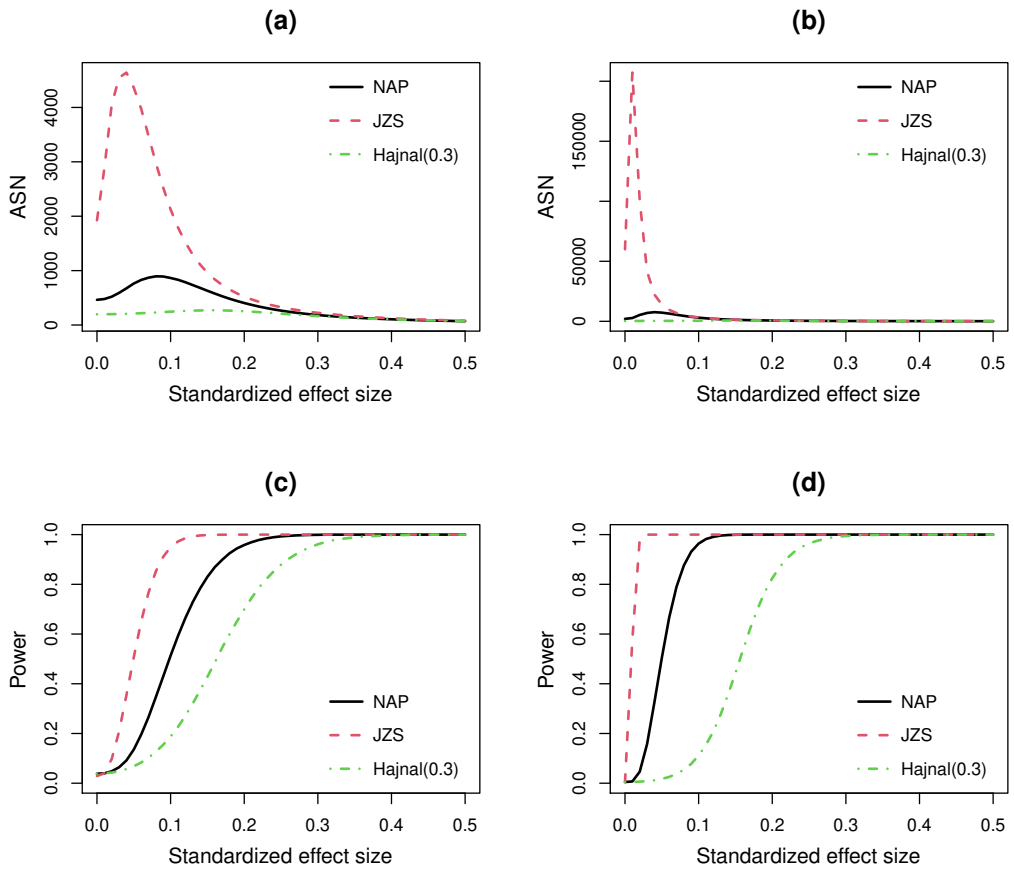


Figure B.8: Operating characteristics under true alternative hypotheses in two-sample z test. Panels (a) and (b) depict the ASN's for three sequential tests when the exceedance thresholds are ± 3 and ± 5, respectively, versus the data-generating value of the standardized effect size. Panels (c) and (d) provide the corresponding probabilities that each test rejects the null hypothesis as a function of the standardized effect size.

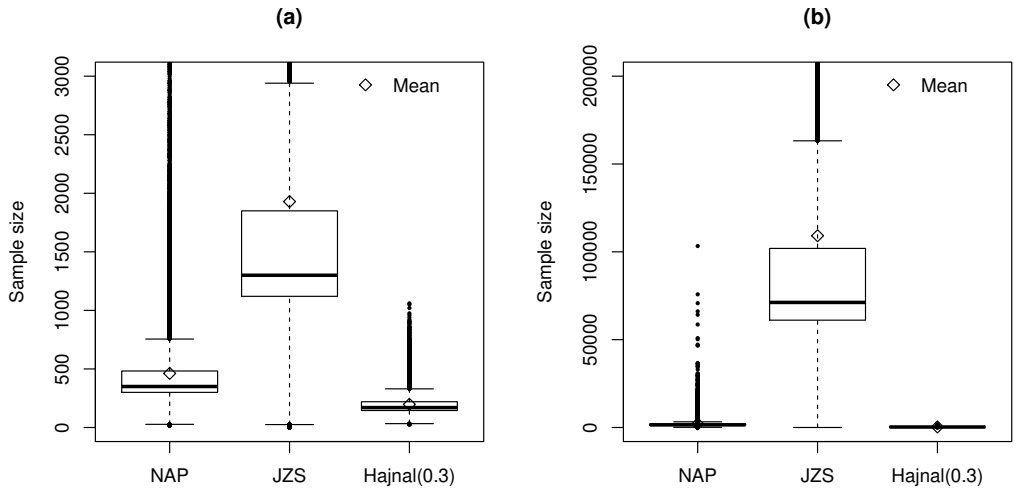


Figure B.9: ASN for sequential procedures under a true null hypothesis in two-sample t test. The plots are truncated at 3000 and 200,000 to enhance comparisons at moderate sample sizes. Panel (a) provides a boxplot estimate of the distribution of sample sizes required from each group for the SBF-NAP, SBF-JZS and Hajnal(0.3) procedures to cross an exceedance threshold of ± 3. About 0.3% of SBF-NAP tests and 11% of SBF-JZS tests required more than 3000 samples from each group to reach a decision. All Hajnal(0.3) tests terminated by 1060 samples. Panel (b) provides the corresponding boxplots when the exceedance threshold is ± 5. About 8% of SBF-JZS tests required more than 200,000 samples from each group to reach a decision. The black diamonds show the ASN's for each procedure. All SBF-NAP tests reached a decision by 103300 samples, and all Hajnal(0.3) tests terminated by 1610 samples from each group.

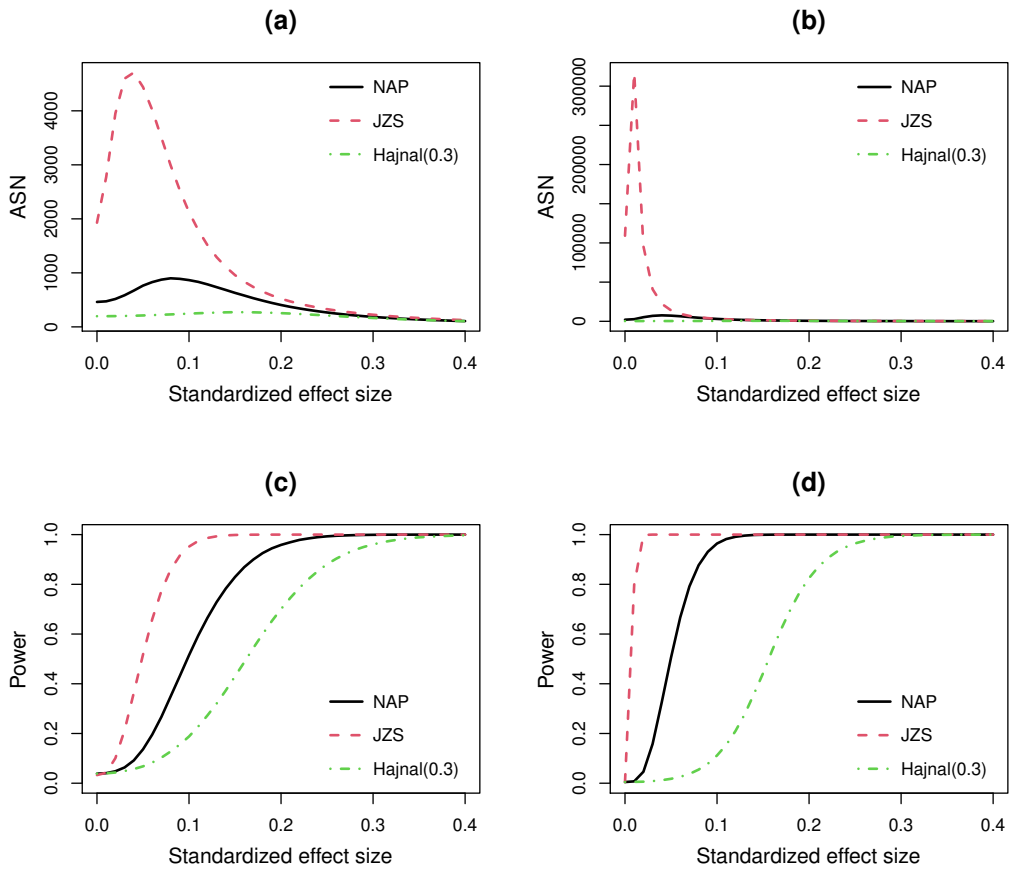


Figure B.10: Operating characteristics under true alternative hypotheses in two-sample t. Panels (a) and (b) depict the ASN's for three sequential tests when the exceedance thresholds are ± 3 and ± 5, respectively, versus the data-generating value of the standardized effect size. Panels (c) and (d) provide the corresponding probabilities that each test rejects the null hypothesis as a function of the standardized effect size.

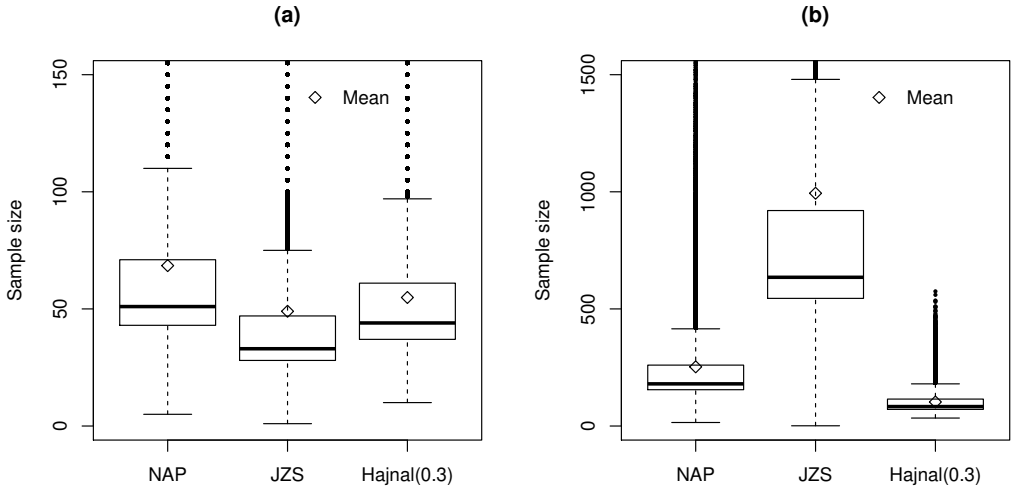


Figure B.11: ASN for SPRT procedures when the null hypothesis is true in one-sample z test. Panel (a) provides a boxplot estimate of the distribution of sample sizes required for the SBF-NAP, SBF-JZS and Hajnal(0.3) procedures to cross Wald's decision thresholds at $\alpha = 0.05$ and $\beta = 0.2$. The plot is truncated at 150 samples (5.3% of SBF-NAP tests, 3.33% of SBF-JZS tests, and 1.71% of Hajnal(0.3) tests required more than 150 samples). Panel (b) provides the corresponding estimate when Wald's decision thresholds were based on $\alpha = 0.005$ and $\beta = 0.05$. The plot is truncated at 1500 samples (0.52% of SBF-NAP and 10.76% of SBF-JZS tests required more than 1500 samples; none of Hajnal(0.3) tests did). The black diamonds show the ASN for each procedure.

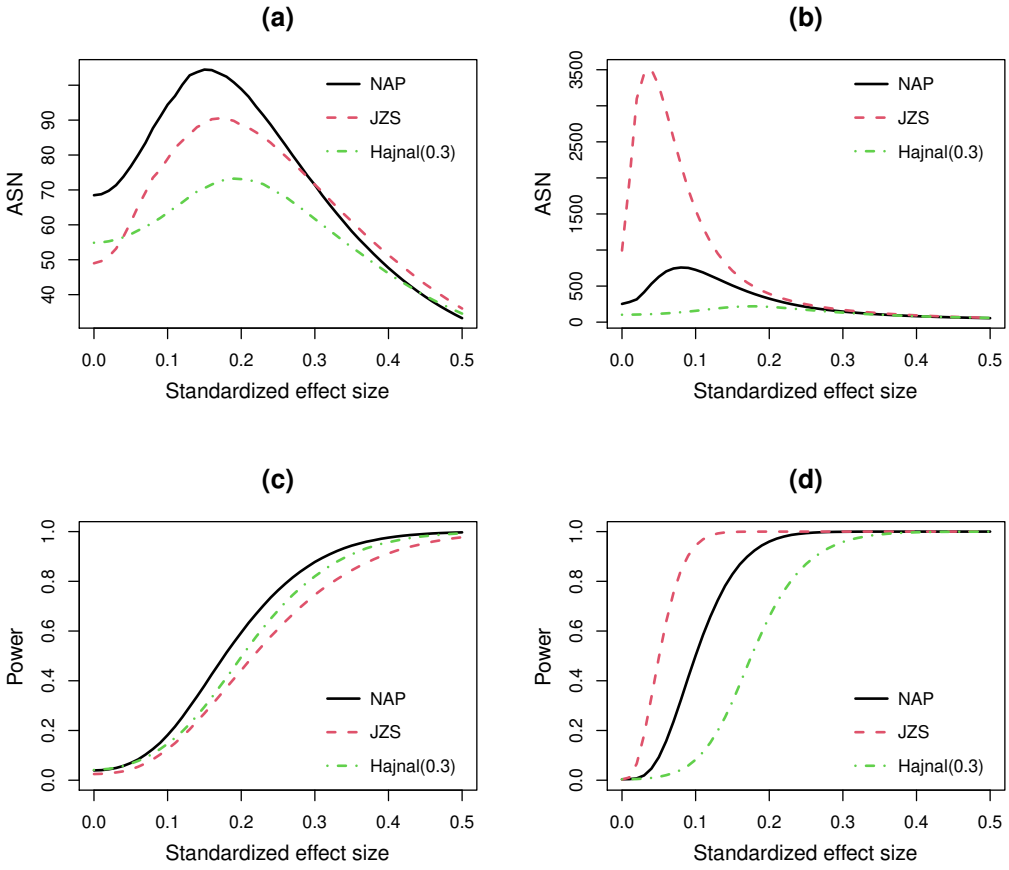


Figure B.12: Operating characteristics under true alternative hypotheses in one-sample z test. Panels (a) and (b) depict the ASN for three SPRT procedures based on Wald's decision thresholds for $(\alpha, \beta) = (0.05, 0.2)$ and $(0.005, 0.05)$, respectively, versus the data-generating value of the standardized effect size. Panels (c) and (d) provide the probability that each procedure rejected the null hypothesis as a function of the standardized effect size.

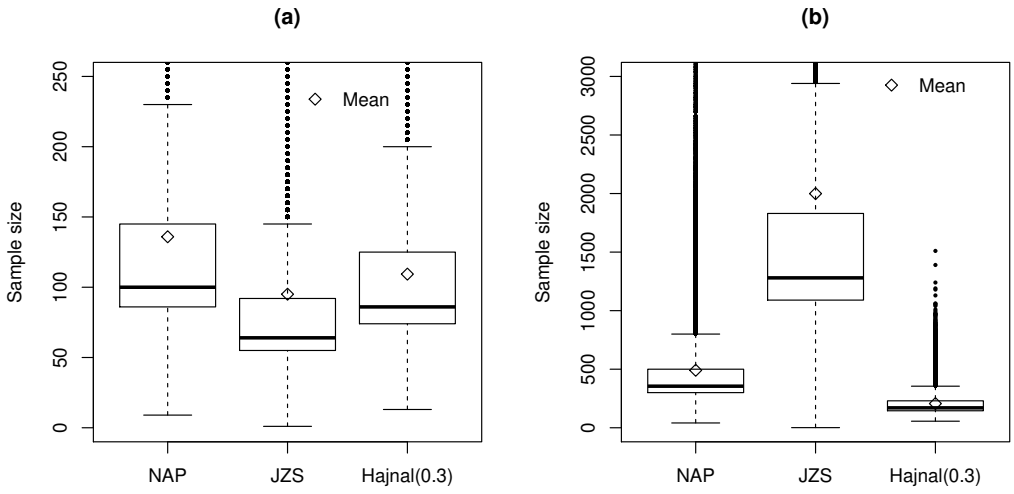


Figure B.13: ASN for SPRT procedures when the null hypothesis is true in two-sample z test. Panel (a) provides a boxplot estimate of the distribution of sample sizes from each group required for the SBF-NAP, SBF-JZS and Hajnal(0.3) procedures to cross Wald's decision thresholds at $\alpha = 0.05$ and $\beta = 0.2$. The plot is truncated at 250 samples (7.68% of SBF-NAP tests, 4.37% of SBF-JZS tests, and 3.3% of Hajnal(0.3) tests required more than 250 samples). Panel (b) provides the corresponding estimate when Wald's decision thresholds were based on $\alpha = 0.005$ and $\beta = 0.05$. The plot is truncated at 3000 samples (0.47% of SBF-NAP and 10.89% of SBF-JZS tests required more than 1500 samples; none of Hajnal(0.3) tests did). The black diamonds show the ASN for each procedure.

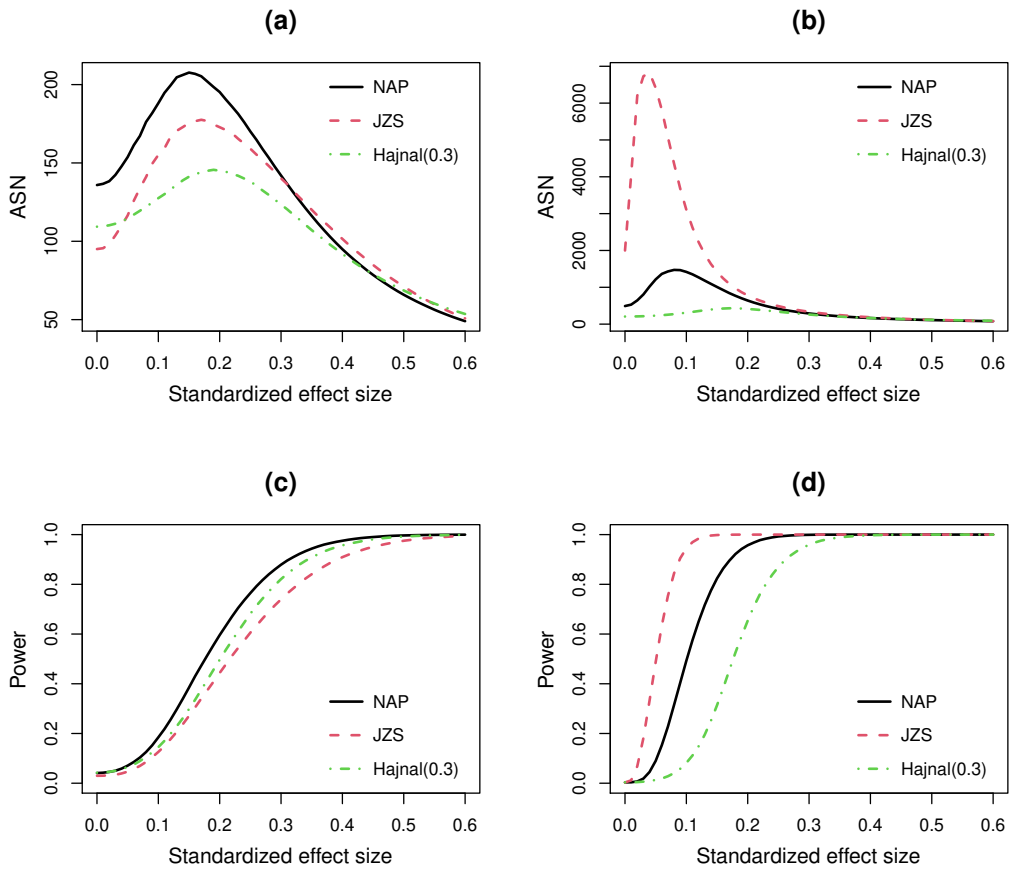


Figure B.14: Operating characteristics under true alternative hypotheses in two-sample z test. Panels (a) and (b) depict the ASN for three SPRT procedures based on Wald's decision thresholds for $(\alpha, \beta) = (0.05, 0.2)$ and $(0.005, 0.05)$, respectively, versus the data-generating value of the standardized effect size. Panels (c) and (d) provide the probability that each procedure rejected the null hypothesis as a function of the standardized effect size.

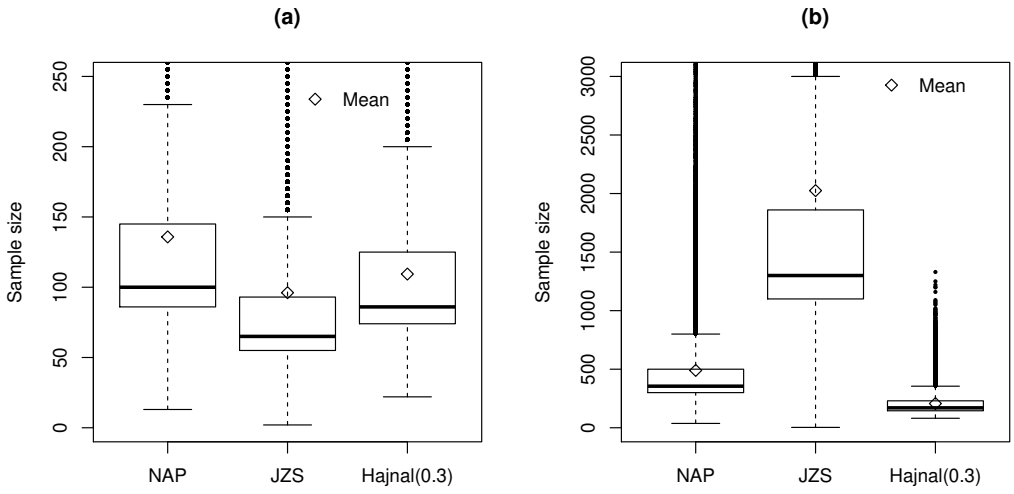


Figure B.15: ASN for SPRT procedures when the null hypothesis is true in two-sample t test. Panel (a) provides a boxplot estimate of the distribution of sample sizes from each group required for the SBF-NAP, SBF-JZS and Hajnal(0.3) procedures to cross Wald's decision thresholds at $\alpha = 0.05$ and $\beta = 0.2$. The plot is truncated at 250 samples (7.82% of SBF-NAP tests, 4.4% of SBF-JZS tests, and 3.26% of Hajnal(0.3) tests required more than 250 samples). Panel (b) provides the corresponding estimate when Wald's decision thresholds were based on $\alpha = 0.005$ and $\beta = 0.05$. The plot is truncated at 3000 samples (0.47% of SBF-NAP and 11.18% of SBF-JZS tests required more than 1500 samples; none of Hajnal(0.3) tests did). The black diamonds show the ASN for each procedure.

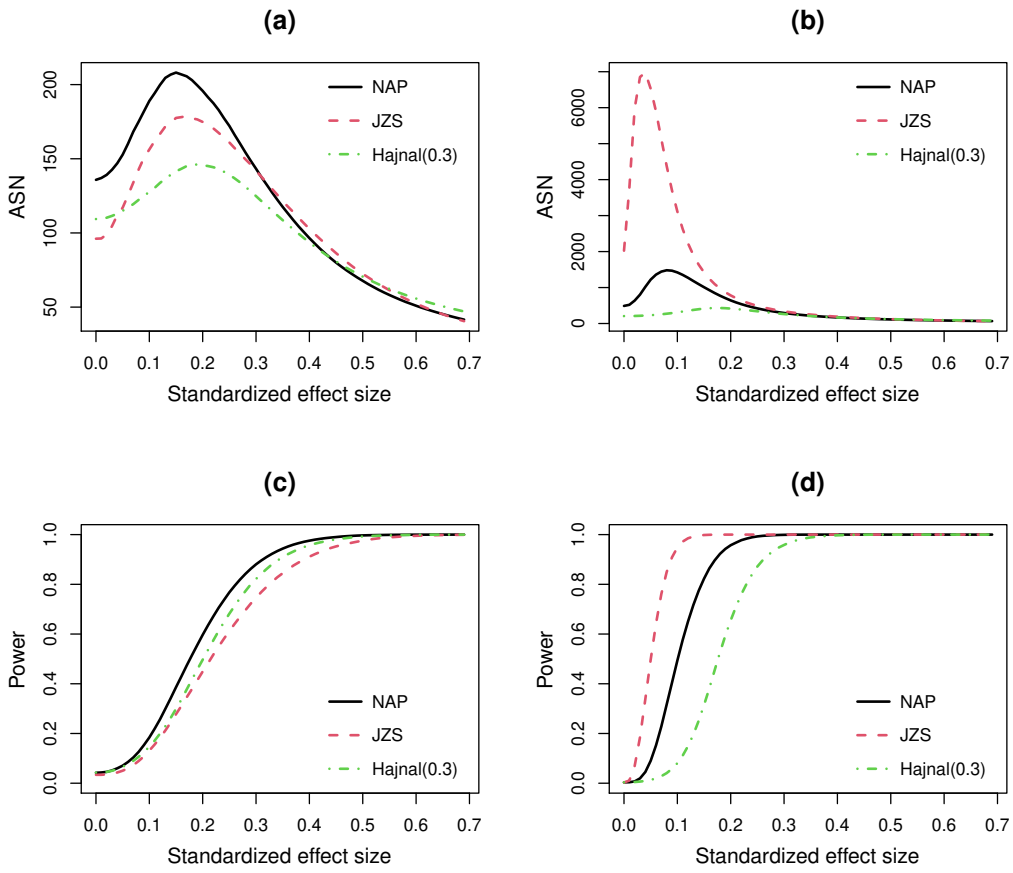


Figure B.16: Operating characteristics under true alternative hypotheses in two-sample t test. Panels (a) and (b) depict the ASN for three SPRT procedures based on Wald's decision thresholds for $(\alpha, \beta) = (0.05, 0.2)$ and $(0.005, 0.05)$, respectively, versus the data-generating value of the standardized effect size. Panels (c) and (d) provide the probability that each procedure rejected the null hypothesis as a function of the standardized effect size.