Bayesian Evaluation of Informative Hypotheses

Xin Gu
Utrecht University

BAYESIAN EVALUATION OF INFORMATIVE HYPOTHESES

BAYESIAANSE EVALUATIE VAN INFORMATIEVE HYPOTHESES
(met een samenvatting in het Nederlands)

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op vrijdag 3 juni 2016 des ochtends te 10.30 uur

door

Xin Gu

geboren op 7 mei 1987
te Xi'an, China

Promotor:	Prof.dr. H.J.A. Hoijtink
Copromotor:	Dr. J. Mulder

This thesis was accomplished with financial support from the China Scholarship Council (CSC).

Beoordelingscommissie:

Prof.dr. P.A. Boelen
Prof.dr. M.A.L.M. van Assen
Prof.dr. J.P. Fox
Prof.dr. E.J. Wagenmakers
Prof.dr. S. van Buuren

Gu, Xin
Bayesian evaluation of informative hypotheses
Proefschrift Universiteit Utrecht, Utrecht. - Met lit. opg. - Met samenvatting in het Nederlands.
ISBN: 978-94-6299-338-9
Printed by Ridderprint, the Netherlands

To Meng Ma

Contents

Chapter 1

Introduction

This dissertation discusses an alternative for the traditional null hypothesis significance testing: the Bayesian evaluation of informative hypotheses. Informative hypotheses can be constructed using (in)equality constraints among the parameters of a statistical model. The support in the data for these hypotheses can be quantified using the Bayes factor. The last decade has rendered many studies with respect to the evaluation of informative hypotheses by means of the Bayes factor. This approach was pioneered by Klugkist, Laudy, and Hoijtink (2005) in the context of ANOVA models, and was extended to repeated measures (Mulder et al., 2009), contingency tables (Klugkist, Laudy, & Hoijtink, 2010), and multivariate normal linear models (Mulder, Hoijtink, & Klugkist, 2010). However, each of these studies was limited to the evaluation of informative hypotheses in one specific model. This dissertation proposes an approximate Bayesian procedure for the evaluation of informative hypotheses in general statistical models. These models can be structural equation models (Kline, 2011) including, e.g., path models, confirmatory factor analysis models, and latent class models; and, generalized linear (mixed) models (McCullogh & Searle, 2001) including, e.g., multivariate normal linear models, logistic regression models, and multilevel models. The proposed Bayesian methods are implemented into two software packages: BIG for Bayesian evaluation of inequality constrained hypotheses in general statistical models, and BaIn for Bayesian evaluation of informative hypotheses.

Bayesian evaluation of informative hypotheses consists of three steps: construct candidate informative hypotheses based on the expectations of researchers (elaborated in Section 1.1); specify the prior distribution and derive the posterior distribution of the parameters used in the informative hypotheses (elaborated in Section 1.2); and, compute Bayes factors to determine the support in the data for the informative hypotheses of interest (elaborated in Section 1.3). In Section 1.4, a short summary of each of the five upcoming chapters will be given.

1.1 Informative hypotheses

This section begins with a motivating example to explain how an informative hypothesis expresses the expectation of a researcher. Suppose an experiment is conducted to investigate pain sensitivity as a function of morphine tolerance and dependence. The experiment involves four groups of rats which receive infusions four times and are placed on a hot plate after each infusion.

- Group 1 receives saline on all four trials.
- Group 2 receives morphine on all four trials.
- Group 3 receives morphine on the first three trials and saline on the fourth trial.
- Group 4 receives saline on the first three trials and morphine on the fourth trial.

After the fourth infusion, the pain sensitivity of the rats is measured by the paw-lick latency in seconds, i.e., the longer the paw-lick latency, the less the pain sensitivity. Researchers may have the following expectations about this experiment:

- Group 2 is expected to feel the pain as strong as Group 1, because morphine tolerance develops such that the effect of morphine disappears.
- Group 3 is expected to suffer from even stronger pain than Group 1, because morphine dependence occurs.
- Group 4 is expected to experience less pain than Group 1 because of the first use of morphine in the last trial.

These expectations can be represented by an informative hypothesis: $H_i : \theta_3 < \theta_1 = \theta_2 < \theta_4$. where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_4)$ contains the means of the paw-lick latency in the four groups.

The informative hypothesis H_i can be compared to an unconstrained hypothesis

$$H_u : \boldsymbol{\theta} \text{ is unconstrained,} \tag{1.1}$$

or to its complement

$$H_{i_c} : \text{not } H_i. \tag{1.2}$$

The complement of an informative hypothesis H_i specifies the parameter space that is not in agreement with H_i. For example, the complement of $H_i : \theta_1 > 0, \theta_2 > 0$ contains at least one parameter that is not larger than 0. We can also compare H_i to a competing informative hypothesis $H_{i'}$.

1.2 Prior and posterior distributions

The computation of the Bayes factor needs the specification of prior distributions for the parameters used to specify the informative hypothesis. Since the aim of this dissertation is to evaluate informative hypotheses in a general situation, the parameters used in the hypotheses can be either the unbounded location parameters, e.g., the group means and regression coefficients, or the bounded parameters, e.g. the variances, probabilities, and correlations. When informative hypotheses are specified using only inequality constraints, a noninformative prior distribution can be used. Chapter 2 specifies a noninformative normal prior distribution under the unconstrained hypothesis H_u with mean vector of zero and diagonal covariance matrix in which each variance is approaching infinity. Using this prior specification, hypotheses with the same structure, e.g., $H_1 : \theta_1 > 0, \theta_2 > 0$ vs $H_2 : \theta_1 < 0, \theta_2 < 0$ are equally likely a priori. Chapter 3 specifies a noninformative normal prior distribution with a different covariance matrix. The prior covariance matrix in the new method is a product of an infinite number and the estimated covariance matrix of the parameters used in the informative hypothesis. Based on this prior specification, the Bayes factor is invariant to linear transformation of the data. This property is important when comparing, for example, three group means θ_1, θ_2, and θ_3, because the evaluation of hypothesis $H_3 : \theta_1 > \theta_2 > \theta_3$ should be the same as the evaluation of the equivalent hypothesis $H_{3'} : \beta_1 > 0, \beta_2 > 0$ where $\beta_1 = \theta_1 - \theta_2$ and $\beta_2 = \theta_2 - \theta_3$.

Although this noninformative prior performs well for the evaluation of inequality constrained hypotheses, it cannot be used when testing hypotheses formulated using equality constraints because vague priors will result in the Lindley-Bartlett paradox (Lindley, 1957), i.e., the Bayes factor will always favor the equality constrained hypothesis compared to the unconstrained hypothesis regardless of the data. To avoid this paradox, Bayesian statisticians have proposed default priors that render easily computable Bayes factors. Examples are JZS priors (Jeffreys, 1961; Zellner & Siow, 1980; Rouder, Speckman, Sun, Morey, & Iverson, 2009), intrinsic priors (Berger & Pericchi, 1996), expected posterior priors (Pérez & Berger, 2002), and fractional priors (O'Hagan, 1995). Chapter 4 introduces three typical default priors for the one sample t test, and investigates the frequentist properties of these priors in standard situations. It proposes a new method for default prior specification based on the frequentist properties. This methods is generalized in Chapter 5, for the evaluation of informative hypotheses in a general class of statistical models.

The posterior distribution of parameters combines information from the prior and the data. Because in the present context the prior distribution based on either noninformative or default settings contains little information, the posterior distribution depends essentially on the data. Using large sample theory (Gelman, Carlin, Stern, & Rubin, 2004, p. 101-107) the posterior distribution will be approximated by a (multivariate) normal distribution. Although a normal approximation of the posterior distribution has a perfect performance only asymptotically, the simulation studies

in this dissertation will demonstrate that its performance is adequate even for small sample sizes when the goal is to evaluate informative hypotheses.

1.3 Bayes factors

The Bayes factor is defined as the ratio of the marginal likelihoods under two hypotheses of interest (Kass & Raftery, 1995; Hoijtink, 2012, p. 59). The marginal likelihood provides a Bayesian measure of the support in the data for each hypothesis. Therefore, the Bayes factor has a direct interpretation as the relative support in the data for two hypotheses. For example, $BF_{12} = 5$ implies that the support for hypothesis H_1 is five times larger than for H_2 after observing the data. Based on the rule proposed by Kass and Raftery (1995), the degree of support for H_1 versus H_2 is categorized as unconvincing if $BF_{12} \in [1, 3]$, positive if $BF_{12} \in [3, 20]$, strong if $BF_{12} \in [20, 150]$, and very strong if $BF_{12} \in [150, \infty]$.

The Bayes factor for an informative hypothesis against an unconstrained hypothesis can be represented as the ratio of the relative fit and complexity of the informative hypothesis. This will be elaborated in detail in Chapter 2, Chapter 3, and Chapter 5 where two software packages with user manuals are offered for the computation of the Bayes factor. These packages are developed in Fortran 90 for the Bayesian evaluation of inequality constrained hypotheses for general statistical models (BIG) and Bayesian evaluation of informative hypotheses (BaIn). Interested readers can download them at website `http://informative-hypotheses.sites.uu.nl/software/`.

1.4 Outlines of the dissertation

This dissertation extends the existing studies into the Bayesian evaluation of informative hypotheses in three directions. First, it substantially increases the class of statistical models for which informative hypotheses can be evaluated. Second, it explores efficient algorithms to compute the Bayes factor for informative hypotheses and implements them into software packages. Third, it presents new methods for prior specification resulting in Bayes factors with attractive properties.

Chapter 2 proposes an approximate Bayes procedure that can be used for the selection of the best of a set of inequality constrained hypotheses based on Bayes factors in general statistical models. A software package BIG is provided such that applied researchers can use the approach proposed for the analysis of their own data. To illustrate the approximate Bayes procedure and the use of BIG, inequality constrained hypotheses are evaluated in a path model and a logistic regression model. Two simulation studies on the performance of our approximate Bayes procedure show that it results in accurate Bayes factors.

Chapter 3 develops an efficient algorithm for the computation of the Bayes factor when evaluating inequality constrained hypotheses in SEM models. This algorithm

results in substantial improvement of the software package BIG offered in Chapter 2 as it makes BIG much faster and therefore easier to use for applied researchers. Furthermore, this chapter presents two prior specification methods which render Bayes factors with different features.

Chapter 4 discusses the prior specification for Bayesian null hypothesis testing. It investigates the classical type I and type II error probabilities of Bayes factors based on default priors for a Bayesian t test. It is shown that in most typical situations these Bayes factors are asymmetric in information, i.e., they result in unequal error probabilities. Although this asymmetry in information is a natural property of a Bayes factor, severe cases of asymmetry may be undesirable in a default setting because the default priors are not a translation of subjective prior believes. Frequentist calibration is used to obtain Bayes factors with about equal error probabilities.

Chapter 5 focuses on Bayesian evaluation of informative hypotheses that contain both equality and inequality constraints in general statistical models. It generalizes the prior specification methods discussed in Chapter 4 for testing equality constrained hypotheses, and incorporates the approximate Bayesian procedure presented in Chapter 2 and the efficient algorithm developed in Chapter 3 for evaluating inequality constrained hypotheses. All of this is implemented in the software package BaIn that is introduced in this chapter.

Chapter 6 is an application of the approximate Bayesian method proposed in Chapter 2. The informative hypotheses are formulated based on the expectations of biomedical scientists. These hypotheses are evaluated by means of the Bayes factors obtained using the software package BIG.

Chapter 2

Bayesian evaluation of inequality constrained hypotheses[1]

2.1 Introduction

Bayesian evaluation of inequality constrained hypotheses has become an attractive alternative for the evaluation of null hypotheses, because the criticism with respect to the evaluation of the traditional null hypothesis is steadily increasing. This criticism consists of four aspects. First, it is hard to image a population that is accurately described by the null hypothesis "nothing is going on" (Cohen, 1994; Royall, 1997, p.79-81). In addition, Royall claims that a sample size of zero is sufficient to reject the null hypothesis, because there is no population that can be in agreement with the null hypothesis. In Sober's (2002) words this means that the null hypothesis is not plausible and thus is a problematic hypothesis. For example, one can hardly find a population in which three means θ_1, θ_2 and θ_3 are exactly equal, that is, $H_0 : \theta_1 = \theta_2 = \theta_3$. Therefore, H_0 is not a plausible hypothesis, and consequently, data are not needed to be able to reject it. Second, in psychological science most researchers have clear theories or expectations with respect to their population of interest, which can not be expressed by the null hypothesis. Thus, rejection of the null hypothesis does not provide those researchers with an evaluation of their own expectations (van de Schoot, Hoijtink, & Romeijn, 2011). For instance, three means θ_1, θ_2 and θ_3 could be ordered from small to large, but this expectation can not be represented by a null hypothesis. Evaluating a plausible hypothesis like $H_i : \theta_1 < \theta_2 < \theta_3$ produces

[1]This chapter has been published as Gu, X., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, 19(4), 511-527. Author contributions: XG, JM and HH designed the research. MD provided the data. XG performed the data analyses and simulation study, developed the software package, and wrote the paper. JM and HH gave feedback on software development. JM, HH and MD provided extensive feedback on constructing and writing the paper.

more direct and explicit results than traditional null hypothesis evaluation. Inequality constrained hypotheses are a formal representation of the theory or expectation that a researcher has using constraints among the parameters of the statistical model. It fulfills the requirement of constructing plausible (Sober, 2002), specific and thus falsifiable (Popper, 1959) hypotheses. The third criticism is that the null hypothesis significance testing by means of p-values can not quantify the evidence in the data in favor of the hypothesis under investigation (Wagenmakers, 2007). After rejecting ($p<0.05$) or not rejecting ($p>0.05$) the hypothesis, we still do not know the degree to which the hypothesis is true or false. The Bayes factor (Kass & Raftery, 1995), however, can measure the evidence from the data for or against the hypothesis. Using the Bayes factor, which can have values like .5, 1, and 10, the evidence from the data for H_i or its complement "not H_i" can be quantified. This enables researchers to make statements like "after observing the data the support for H_i is half as large as the support for not H_i", "after observing the data the support for H_i is equal to the support for not H_i", and "after observing the data the support for H_i is ten times larger than the support for not H_i". Fourth, the null hypothesis significance testing can only evaluate the null hypothesis against alternative hypotheses, while Bayes factors are able to compare two or more possible non-nested hypotheses, such as, $H_1 : \theta_1 > 0; \theta_2 > 0$, $H_2 : \theta_1 > \theta_2$, and $H_3 : \theta_1 + \theta_2 > 0$.

Bayesian evaluation of inequality constrained hypotheses has almost exclusively been studied in the context of the multivariate normal linear model. For example, Klugkist et al. (2005) developed it for analysis of variance or analysis of covariance models with inequality constraints on the means. The same approach has been applied in repeated measures analysis (Mulder et al., 2009) to evaluate the development of means over time. Applications in the context of the multivariate normal linear model are described in Mulder et al. (2010) and implemented in the software package BIEMS (Mulder, Hoijtink, & de Leeuw, 2012). In addition, there have been a few excursions to other models such as multilevel models (Kato & Hoijtink, 2006; Mulder & Fox, 2013) and models for contingency tables (Klugkist et al., 2010). All these studies used the Bayes factor as the criterion to select the best of competing inequality constrained hypotheses. However, as will be elaborated later in this paper, the formulation of the Bayes factor depends on the statistical model at hand. The procedures that have currently been developed are not generally applicable. For each new statistical model, Bayesian evaluation of informative hypotheses has to be reanalyzed, redeveloped and reprogrammed.

To avoid these repetitions, this paper demonstrates an approximate Bayes procedure for the evaluation of inequality constrained hypotheses that can be applied to a large class of statistical models. The basic principle is that large sample theory allows posterior distributions to be approximated by normal distributions. Explicit discussion of this theory is given by Gelman et al. (2004, p.101-107). As will be shown, this implies that the Bayes factor can be computed based on noninformative prior and approximate normal posterior distributions for many statistical models (a

more precise description will be given below). This renders a generally applicable procedure and will therefore substantially extend the class of statistical models for which inequality constrained hypotheses can be evaluated.

Throughout the paper, inequality constrained hypotheses will be specified as $H_i : \boldsymbol{R}\boldsymbol{\theta} > \boldsymbol{r}$, where $\boldsymbol{R}$ denotes the restriction matrix representing the inequality constrained hypothesis, and $\boldsymbol{\theta}$ and $\boldsymbol{r}$ contain the parameters and constants in the hypothesis, respectively. The number of rows in $\boldsymbol{R}$ is equal to the number of inequality constraints needed to construct H_i denoted by K, and the number of columns is equal to the number of parameters denoted by J. The vector $\boldsymbol{\theta}$ contains J parameters and the vector $\boldsymbol{r}$ contains K constants. For instance, $H_i : \theta_1 < \theta_2 < \theta_3$ is an example with $J = 3$ and $K = 2$, which leads to $\boldsymbol{R} = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}$, $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^{\mathrm{T}}$, and $\boldsymbol{r} = (0, 0)^{\mathrm{T}}$. Note that $\boldsymbol{R}$ should not be used to construct hypotheses using equality constraints like $H_i : \theta_1 = \theta_2$ which would be obtained if $\boldsymbol{R} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$ and $\boldsymbol{r} = (0, 0)^{\mathrm{T}}$, or using about equality or range constraints like $H_i : 0 < \theta_1 < 2$ which would be obtained if $\boldsymbol{R} = (1, -1)^{\mathrm{T}}$ and $\boldsymbol{r} = (0, -2)^{\mathrm{T}}$. The application of our approach is strictly limited to the evaluation of inequality constrained hypotheses.

The outline of this paper is as follows. First, two examples with respect to path modeling and logistic regression modeling are presented in which the expectations of researchers are represented by inequality constrained hypotheses. Subsequently, it will be elaborated that in order to obtain the normal approximation of the posterior distribution, the estimates and covariance matrix of parameters of the statistical model used are needed. For the two examples that will be given, these will be obtained using Mplus (Muthén & Muthén, 2010; `http://www.statmodel.com`) and OpenBUGS (Ntzoufras, 2009; Lunn, Thomas, Best, & Spiegelhalter, 2000; `http://www.openbugs.info`), respectively. Thereafter it is elaborated how estimates and covariance matrix are used to obtain normal approximations of the density of the data, and the resulting posterior under the statistical model at hand. Then it will be elaborated how the Bayes factor can be computed and used as a criterion to evaluate the inequality constrained hypotheses under investigation, after which the two examples will be analyzed using the approach presented in this paper. Thereafter, the performance of the approximate Bayes procedure will be assessed using two simulation studies. The paper is concluded with a short discussion. Appendices will be given that explain the Mplus and OpenBUGS codes used for the analyses of the examples and to describe how the program BIG (Bayesian evaluation of inequality constrained hypotheses for general statistical models) can be used to compute Bayes factors to evaluate inequality constrained hypotheses in general statistical models.

Table 2.1: Descriptives for the variables in the path model

Variable	Mean	S.D.
Antisocial behaviour	1.44	.56
Positive quality	3.03	.92
Negative quality	2.00	.70
Adolescent disclosure	2.76	.64
Deviant peers	1.71	.68

2.2 Two examples of inequality constrained hypotheses

In this section, two examples will be introduced that will be used to illustrate that our approach can be used for the evaluation of inequality constrained hypotheses in a rather general class of statistical models. In each example, the expectations of researchers with respect to the relationships among the variables will be translated into inequality constrained hypotheses. As will be shown in a later section, the Bayes factor can be used to determine the support in the data for these hypotheses.

2.2.1 Example 1: Path modelling

Structural equation modeling (SEM) (Kline, 2011) is popular in the behavioural and social sciences. It incorporates regression, path, and factor analysis models. Here a path model for the prediction of adolescent antisocial behaviour will be used to illustrate the evaluation of inequality constrained hypotheses. The child's family is regarded as the major factor in the evolvement of antisocial behaviour (Deković, Wissink, & Meijer, 2004). To assess the parent-adolescent relationship, three aspects - positive quality, negative quality and adolescent disclosure are frequently investigated. Besides, as children approach adolescence, it is inevitable that the number of their deviant peers becomes a crucial factor of individual antisocial behaviour. The descriptives for each variable are given in Table 2.1. The data consists of N=603 adolescents. Most of the variables are measured on 5-point scale except disclosure which is measured on a 4-point scale. Low scores mean less degree, frequency or quantity of corresponding variables, e.g. for antisocial behaviour the score 1 means the adolescent does not have any antisocial activities in the last 12 months. High scores mean more degree, frequency or quantity, e.g. for disclosure the score 4 means the adolescents tell their parents everything about their activities. The relations among these variables can be represented by the path model presented in Figure 2.1 (van de Schoot, Hoijtink, & Deković, 2010).

Figure 2.1 also shows how the variables are related to each other. Evaluation of inequality hypotheses is only sensible if the parameters involved are comparable, that is, standardized. This can be achieved by standardizing both the independent and

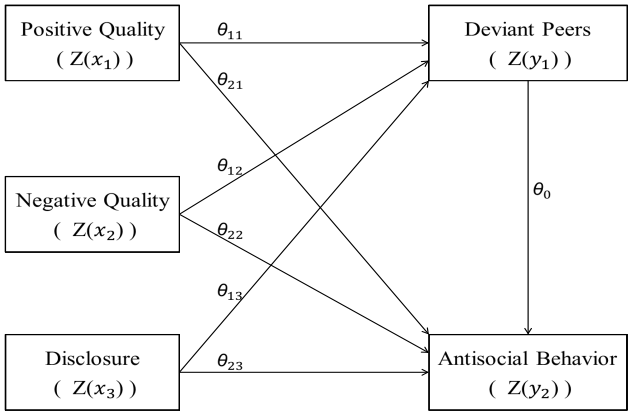

Figure 2.1: Path model for the variables in Example 1

the dependent variables:

$$Z(\boldsymbol{y}_i) = \boldsymbol{\Pi} Z(\boldsymbol{y}_i) + \boldsymbol{\Gamma} Z(\boldsymbol{x}_i) + \boldsymbol{\delta}_i, \tag{2.1}$$

where $Z(.)$ denotes standardization of the argument, $Z(\boldsymbol{y}_i) = (Z(y_{1i}), Z(y_{2i}))^{\mathrm{T}}$, $Z(\boldsymbol{x}_i) = (Z(x_{1i}), Z(x_{2i}), Z(x_{3i}))^{\mathrm{T}}$, $\boldsymbol{\Pi} = \begin{pmatrix} 0 & 0 \\ \theta_0 & 0 \end{pmatrix}$ and $\boldsymbol{\Gamma} = \begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{pmatrix}$ are matrices of regression coefficients, and $\boldsymbol{\delta}_i \sim N(\mathbf{0}, \boldsymbol{\Psi})$ is a vector of residuals with $\boldsymbol{\Psi}$ being the residual covariance matrix. This manner of dealing with scale differences in the structural parameters is straightforward and easy to implement. However, it can be criticized because the data are used twice: once to standardize the dependent and independent variables; and once to evaluate the inequality constrained hypotheses. As is elaborated in Appendix 2.A, there is a more elaborate manner to deal with standardization. However, as will be shown using a simulation study, the results obtained using both approaches are virtually indistinguishable.

Many researchers have expectations that can be represented in the form of inequality constraints among the parameters of a SEM model. Deković et al. (2004) expected that adolescent disclosure is the strongest predictor of antisocial behavior among the parent-adolescent relationship variables, and indicate that the association with deviant peers is the overall strongest determinant of problem behavior in adolescence. We therefore consider the following inequality constrained hypotheses:

$$H_1: \begin{array}{c} \theta_{23} > \{\theta_{21}, \theta_{22}\} \\ \theta_0 > \{\theta_{21}, \theta_{22}, \theta_{23}\} \end{array}, \tag{2.2}$$

Table 2.2: Descriptives for the variables in the LR model

	N		GRE		GPA	
	Total	$admit = 1$	Mean	S.D.	Mean	S.D.
$rank = 1$	61	33	611.80	120.24	3.45	.39
$rank = 2$	151	54	596.03	107.01	3.36	.37
$rank = 3$	121	28	574.88	121.15	3.43	.38
$rank = 4$	67	12	570.15	116.22	3.32	.36
Total	400	127	587.70	115.52	3.39	.38

and

$$H_{1_c} : \text{not } H_1. \tag{2.3}$$

Note that H_1 represents what the researchers consider to be plausible, that is, their expectation. H_{1_c} is the complement of H_1, that is, it expresses what the researchers do not expect. H_1 and H_{1_c} define an arena in which null hypothesis testing is unable to play a significant role. As will be shown in the remaining of the paper, theses hypotheses can be evaluated using the Bayes factor.

2.2.2 Example 2: Logistic regression modelling

The logistic regression (LR) model is the counterpart of the conventional multiple regression model if the dependent variable is binary instead of continuous. Consider an example from UCLA Academic Technology Services (data available on the Web at `http://www.ats.ucla.edu/stat/data/binary.sav`). When a graduate school in the USA assesses the qualification of applicants, general determinants are graduate record examination (GRE) scores and undergraduate grade point average (GPA). As can be seen from the descriptives in Table 2.2, the outcome called "*admit*" is a binary variable, that is, $admit = 1$ means approval and $admit = 0$ means refusal. The independent variable GRE scores are obtained by a standardized test intended to measure the abilities of graduates and takes on values between 200 and 800. The GPA is measured on a 4 point scale and reflects the academic result of graduates throughout their studies. The higher the GRE and GPA scores, the better the applicants.

However, the GPA rendered by undergraduate institutions with high prestige is more convincing than those with low prestige. We accordingly analyze the data among the different institution prestige levels denoted by "*rank*" measured on 4 discrete values. Institutions with a rank of 1 have the highest prestige, while those with a rank of 4 have the lowest. As described in Table 2.2, consequently, the data can be separated into four groups based on the rank.

This example can be modelled using a LR model. To ensure comparability of the

regression coefficients, the independent variables are standardized:

$$\text{logit}(p_{ji}) = \theta_{j0} + \theta_{j1} Z(\text{GRE}_{ji}) + \theta_{j2} Z(\text{GPA}_{ji}). \tag{2.4}$$

There are four rank groups labeled $j = 1, 2, 3, 4$. In each group, $\text{logit}(p_{ji}) = \ln(\frac{p_{ji}}{1-p_{ji}})$, p_{ji} denotes the probability of $admit = 1$, θ_{j0} is the intercept, and θ_{j1} and θ_{j2} are the coefficients of two predictors.

As discussed before, the GPA of candidates from low prestige institutions is less trusted by the evaluator, therefore, the effect of GPA might decrease with ascending rank level. Besides, when the evaluator determines the admission of candidates from high prestige institutions, the convincing GPA is likely to play the principal role. However, for the candidates from low prestige institutions, their GRE scores might be more important. As a result, the hypotheses for this example can be constructed as follows:

$$H_1 : \theta_{12} > \theta_{22} > \theta_{32} > \theta_{42}, \tag{2.5}$$

$$H_2 : \begin{array}{l} \theta_{11} < \theta_{12} \\ \theta_{21} < \theta_{22} \\ \theta_{31} > \theta_{32} \\ \theta_{41} > \theta_{42} \end{array}, \tag{2.6}$$

and

$$H_3 : \begin{array}{l} \theta_{11} < \theta_{12} \\ \theta_{21} < \theta_{22} \\ \theta_{31} > \theta_{32} \\ \theta_{41} > \theta_{42} \\ \theta_{12} > \theta_{22} > \theta_{32} > \theta_{42} \end{array}, \tag{2.7}$$

Note that H_3 contains all the constraints in H_1 and H_2 so that it can be expressed as $H_3 : H_1 \& H_2$. Later in this paper these hypotheses will be mutually evaluated and compared to an unconstrained alternative hypothesis. This will highlight another advantage of hypothesis evaluation using Bayes factors over null hypothesis significant testing. The Bayes factor renders a direct comparison of two or more hypotheses, while null hypothesis significant testing is basically limited to two hypotheses H_0 and H_a.

2.3 Estimates and covariance matrix of the structural parameters

Once the model has been constructed, the parameters in the model can be estimated by various approaches. Let $\boldsymbol{\theta}$ denote a vector containing the structural parameters of the specified model, that is, the parameters that will be used in the formulation of informative hypotheses, and let $\boldsymbol{\zeta}$ denote the nuisance parameters, that is, the

Table 2.3: Estimates and covariance matrix of the structural parameters in the path model from `Mplus`

		$\boldsymbol{\Sigma}_\theta$			
	$\hat{\boldsymbol{\theta}}$	θ_{21}	θ_{22}	θ_{23}	θ_0
θ_{21}	.046	1.04E-3			
θ_{22}	.089	2.25E-4	9.37E-4		
θ_{23}	.126	-4.70E-4	7.53E-5	1.04E-3	
θ_0	.674	-7.21E-5	-2.82E-4	1.78E-4	8.82E-4

parameters that will not be used in the hypotheses. As will be elaborated in the next section, when evaluating constrained hypotheses using our approach, the estimates and covariance matrix of the structural parameters have to be calculated, which are represented by $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_\theta$, respectively. In the following two subsections we will estimate the structural parameters using Mplus (Muthén & Muthén, 2010) and OpenBUGS (Ntzoufras, 2009; Lunn et al., 2000) for both the path model and the LR model. Mplus can be downloaded from `http://www.statmodel.com`, and OpenBUGS can be downloaded from `http://www.openbugs.info`. Both are rather encompassing packages that can be used to obtain estimates and covariance matrix of the structural parameters for very general classes of statistical models.

2.3.1 Example 1 (Continued)

Consider again the example using the path model. Here the structural parameters are $\boldsymbol{\theta} = (\theta_{21}, \theta_{22}, \theta_{23}, \theta_0)$ and the nuisance parameters are $\boldsymbol{\zeta} = (\theta_{11}, \theta_{12}, \theta_{13}, \boldsymbol{\Psi})$. As was elaborated above, the first step in the estimation of the structural parameters is to standardize the variables to ensure that the parameters are comparable. Subsequently, both `Mplus` and `OpenBUGS` are used to obtain the estimates and covariance matrix of these structural parameters. The reason for the use of two softwares is to investigate whether there is a consensus in parameter estimates across programs.

As illustrated in Figure 2.1, $Z(y_1)$ regresses on $Z(x_1), Z(x_2), Z(x_3)$, and $Z(y_2)$ regresses on $Z(y_1), Z(x_1), Z(x_2), Z(x_3)$. In Appendix 2.B, the commands in `Mplus`

Table 2.4: Estimates and covariance matrix of the structural parameters in the path model from `OpenBUGS`

		$\boldsymbol{\Sigma}_\theta$			
	$\hat{\boldsymbol{\theta}}$	θ_{21}	θ_{22}	θ_{23}	θ_0
θ_{21}	.046	1.54E-3			
θ_{22}	.090	3.37E-4	1.39E-3		
θ_{23}	.127	-6.87E-4	1.15E-4	1.53E-3	
θ_0	.674	-1.01E-4	-4.07E-4	2.58E-4	1.30E-3

Table 2.5: Estimates and covariance matrix of the structural parameters in the LR model from `Mplus`

		$\boldsymbol{\Sigma}_\theta$							
	$\hat{\boldsymbol{\theta}}$	θ_{11}	θ_{12}	θ_{21}	θ_{22}	θ_{31}	θ_{32}	θ_{41}	θ_{42}
θ_{11}	.30	9.91E-2							
θ_{12}	.42	-3.59E-2	9.18E-2						
θ_{21}	.17	0	0	3.36E-2					
θ_{22}	.30	0	0	-8.91E-3	3.39E-2				
θ_{31}	.40	0	0	0	0	7.35E-2			
θ_{32}	.16	0	0	0	0	-3.48E-2	6.98E-2		
θ_{41}	.29	0	0	0	0	0	0	.11	
θ_{42}	.39	0	0	0	0	0	0	-2.58E-2	.13

needed to estimate the parameters of this model are displayed and annotated, whereas the commands for `OpenBUGS` are omitted in this model. Executing `Mplus` and `OpenBUGS` with the data and commands renders the estimates and covariance matrix of the structural parameters. The results obtained in `Mplus` and `OpenBUGS` are displayed in Table 2.3 and 2.4, respectively, and show that the parameter estimations from different software packages are almost identical.

2.3.2 Example 2 (Continued)

Consider again the example using the logistic regression model. To ensure that the regression coefficients are comparable, the predictors in the LR model have to be standardized. In the hypotheses H_1, H_2 and H_3, the coefficients of each predictor are treated as the structural parameters, that is $\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}, \theta_{31}, \theta_{32}, \theta_{41}, \theta_{42})$, whereas the intercepts are the nuisance parameters, that is, $\boldsymbol{\zeta} = (\theta_{10}, \theta_{20}, \theta_{30}, \theta_{40})$. To estimate these coefficients and their covariance matrix for the LR model, both `Mplus` and `OpenBUGS` are used.

In Appendix 2.C, the `OpenBUGS` code needed to obtain $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_\theta$ is displayed and annotated, whereas the `Mplus` code is left out. Implementation of `Mplus` and `OpenBUGS` renders the estimates and covariance matrix of the structural parameters under consideration. The results are displayed in Table 2.5 and 2.6 using each software, and the similar conclusion of consistent estimation in `Mplus` and `OpenBUGS` can be drawn.

2.4 Density of the data, prior and posterior distribution

The density of the data plays an important role both in classical and Bayesian inference. It is a formal representation of the information contained in the data with respect to the unknown model parameters. For general statistical models, the density

Table 2.6: Estimates and covariance matrix of the structural parameters in the LR model from `OpenBUGS`

		$\boldsymbol{\Sigma}_\theta$							
	$\hat{\boldsymbol{\theta}}$	θ_{11}	θ_{12}	θ_{21}	θ_{22}	θ_{31}	θ_{32}	θ_{41}	θ_{42}
θ_{11}	.32	9.94E-2							
θ_{12}	.45	-4.07E-2	.10						
θ_{21}	.17	0	0	3.40E-2					
θ_{22}	.31	0	0	-8.93E-3	3.48E-2				
θ_{31}	.42	0	0	0	0	7.74E-2			
θ_{32}	.17	0	0	0	0	-3.70E-2	7.40E-2		
θ_{41}	.30	0	0	0	0	0	0	.12	
θ_{42}	.41	0	0	0	0	0	0	-2.93E-2	.14

is $f(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{\zeta})$, where $\boldsymbol{X}$ denotes the data, and $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ denote the structural and nuisance parameters, respectively. For example, consider the path model in the previous section:

$$\boldsymbol{y}_i = \boldsymbol{\Pi}\boldsymbol{y}_i + \boldsymbol{\Gamma}\boldsymbol{x}_i + \boldsymbol{\delta}_i, \quad \boldsymbol{\delta}_i \sim N(\boldsymbol{0}, \boldsymbol{\Psi}). \tag{2.8}$$

The data are $\boldsymbol{X} = \{\boldsymbol{y}_i, \boldsymbol{x}_i\}$, the structural parameters are $\boldsymbol{\theta} = (\theta_0, \theta_{21}, \theta_{22}, \theta_{23})$, and the nuisance parameters are $\boldsymbol{\zeta} = (\theta_{11}, \theta_{12}, \theta_{13}, \boldsymbol{\Psi})$.

When evaluating inequality constrained hypotheses of the form $\boldsymbol{R}\boldsymbol{\theta} > \boldsymbol{r}$, we can specify non-informative normal prior distributions because the complexity of H_i (an important component of the Bayes factor, see the next section) is independent of the specification of the prior mean and variance. A thorough discussion of this issue is given in Chapter 10 of Hoijtink (2012, p.195). Therefore, the prior distribution of the structural parameters can be chosen as:

$$h(\boldsymbol{\theta}) = N(\boldsymbol{0}, \boldsymbol{\Sigma}_\infty), \tag{2.9}$$

where $\boldsymbol{0} = (0, \ldots, 0)^{\mathrm{T}}$, and $\boldsymbol{\Sigma}_\infty$ means that the variance of each parameter is approaching infinity, whereas each covariance is equal to zero.

There are several advantages of using prior distribution (2.9). First, it conjugates to the normally approximated posterior distribution shown at the end of this section. Second, the impact of prior distribution (2.9) on the posterior distribution is negligible, therefore, for any sample size the posterior distribution only depends on the data. Third, for an equivalent set of hypotheses, for example $H_1 : \theta_1 > \theta_2 > \theta_3$ and other five hypotheses in which the parameters have different orders, all hypotheses are equally likely a priori, that is, $1/6$ of the prior distribution is in agreement with each hypothesis. The concept of equivalent hypotheses will be elaborated later on in this section. Other non-informative priors can also be specified without being conjugate, such as uniform priors with very large and symmetric bounds and t-distributions with very large variance. However, estimates and covariance matrix of structural parameters will not be affected by choosing other non-informative priors, because the

posterior distribution is virtually independent of non-informative priors, and the proportion of prior distribution in agreement with each hypothesis that is a member of equivalent set is still unchangeable. For example, when specifying a non-informative uniform prior distribution, that is, with a lower bound of -10000 and an upper bound of 10000, for the structural parameters of LR model in OpenBUGS, the estimates of structural parameters are $\boldsymbol{\theta} = (0.32, 0.45, 0.17, 0.31, 0.42, 0.17, 0.29, 0.42)$, which are almost identical to the results in Table 2.6.

Note that the normal prior distribution (2.9) can straightforwardly be applied to statistical models in which the structural parameters are unbounded. Examples are structural equation models (Kline, 2011), the generalized linear model (McCullagh & Nelder, 1989) which includes, among others, linear and non-linear regression models, analysis of variance models, and models for the analysis of repeated measures, generalized mixed models (McCullogh & Searle, 2001), which includes among others multilevel models (Hox, 2010; De Leeuw & Meijer, 2008) and log-linear models (Agresti, 2007, Chapter 7; Azen & Walker, 2010, Chapter 7).

Equation (2.9) can not straightforwardly be applied to statistical models in which the structural parameters are bounded. Examples of bounded parameters are variances (which have a lower bound of zero), probabilities (which are bounded between 0 and 1), and correlations. Suitable prior distributions for these parameters have to account for their bounded nature. If, for example, the data are contained in a 2 x 2 contingency table (an example will be given in the second simulation study presented later in this paper), the prior distribution of the four cell-probabilities should be Dirichlet (Klugkist et al., 2010) rather than normal. However, as will be elaborated in the next section, the prior distribution is used to compute the complexity of the inequality constrained hypotheses under consideration. This implies that our approach can also be applied to models in which the structural parameters are bounded, as long as the complexity of the inequality constrained hypotheses of interest computed using (2.9) is the same as the complexity computed using a non-informative prior distribution that accounts for the bounded nature of the structural parameters (like the Dirichlet distribution in the contingency table example). As is elaborated in Hoijtink (2012, Chapter 10.3), this holds for all hypotheses that belong to an equivalent set.

A hypothesis belongs to an equivalent set if: each element of $\boldsymbol{R}$ is either a 1, -1, or 0; if the sum of the elements in each row of $\boldsymbol{R}$ equals zero; if the first row of $\boldsymbol{R}$ can be divided into J/M subsets of the same size such that the other rows are permutations of these subsets (note that M denotes the number of 1's in the first row of $\boldsymbol{R}$); and if $\boldsymbol{r} = \boldsymbol{0}$. The following are examples of hypotheses belonging to equivalent sets:

- $\theta_1 > \theta_2 > \theta_3 > \theta_4$ for which $\boldsymbol{R} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}$. As can be seen, each element of $\boldsymbol{R}$ is either a 1, -1, or 0. The sum of the elements in each row equals zero. The first row can be divided in $J/M = 4/1 = 4$ subsets each

containing one number and the second and third row are permutations of these four subsets.

- $\theta_1 - \theta_2 > \theta_3 - \theta_4 > \theta_5 - \theta_6$ for which $\boldsymbol{R} = \begin{pmatrix} 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 & 1 \end{pmatrix}$. As can be seen, each element of $\boldsymbol{R}$ is either a 1, -1, or 0. The sum of the elements in each row equals zero. The first row can be divided in $J/M = 6/2 = 3$ subsets each containing two numbers, that is, $\{1, -1\}, \{-1, 1\}$ and $\{0, 0\}$, and the second row is a permutation of these three subsets.
- $\theta_1 > \theta_2, \theta_3 > \theta_4$ for which $\boldsymbol{R} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}$.
- $\theta_1 > \{\theta_2, \theta_3, \theta_4\}$ for which $\boldsymbol{R} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{pmatrix}$.

The following are examples of hypotheses that do not belong to an equivalent set:

- $\theta_1 - \theta_2 > \theta_2 - \theta_3$ for which $\boldsymbol{R} = (1, -2, 1)$. Because $\boldsymbol{R}$ contains the value 2 this hypothesis does not belong to an equivalent set.
- $\theta_1 - \theta_2 > .5$ for which $\boldsymbol{R} = (1, -1)$. Because $\boldsymbol{r}$ contains the value .5 this hypothesis does not belong to an equivalent set.
- $\theta_1 - \theta_2 > \theta_3 - \theta_4$ with $\theta_1 + \theta_2 > \theta_3 + \theta_4$ for which $\boldsymbol{R} = \begin{pmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \end{pmatrix}$. Note that the first row of $\boldsymbol{R}$ can be divided into $J/M = 4/2 = 2$ subsets of two numbers, that is, $\{1, -1\}$ and $\{-1, 1\}$. However because the second row cannot be obtained by permuting these two subsets, this hypothesis does not belong to an equivalent set.

The posterior distribution integrates the information contained in both the density of the data and the prior with respect to the structural parameters. In order to evaluate informative hypotheses for statistical models in general, large sample theory (Gelman et al., 2004, p.101-107) is used. A fundamental principle of large sample theory is the asymptotic normality of the posterior distribution. This implies that the posterior distribution can be approximated by a normal distribution:

$$g(\boldsymbol{\theta}|\boldsymbol{X}) \approx N(\hat{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_\theta). \tag{2.10}$$

where $\hat{\boldsymbol{\theta}}$ denotes the estimates of structural parameters, and $\boldsymbol{\Sigma}_\theta$ denotes their covariance matrix.

Using normal approximations for the posterior distribution of the structural parameters of statistical models, the evaluation of informative hypotheses has become feasible for many statistical models without the need to reformulate, reevaluate and recompute the Bayes factor for each new statistical model.

Table 2.7: Degree of evidence

BF_{ia}	evidence in favor of H_i
1 to 3	anecdotal
3 to 20	positive
20 to 150	strong
> 150	very strong

2.5 Bayes factor

In this paper, the Bayes factor is used to select the best of a set of competing inequality constrained hypotheses. The Bayes factor of an inequality constrained hypothesis H_i against an unconstrained hypothesis H_a can be represented as the ratio of the posterior and prior probability that the inequality constraints hold (Mulder et al., 2010; Hoijtink, 2012, p.51):

$$BF_{ia} = f_i/c_i, \tag{2.11}$$

where c_i called complexity is the proportion of the prior distribution (2.9) in agreement with H_i, and f_i called fit is the proportion of the posterior distribution (2.10) in agreement with H_i. In addition to BF_{ia}, we can also compute the Bayes factor of a hypothesis versus its complement. This Bayes factor can be written as:

$$BF_{ii_c} = \frac{f_i}{c_i} / \frac{1 - f_i}{1 - c_i}. \tag{2.12}$$

The Bayes factor can be interpreted as the amount of evidence from the data in favor of hypothesis H_i against hypothesis H_a. For example, $BF_{ia} = 10$ indicates that after observing the data, the support for H_i is ten times stronger than the support for H_a. To interpret the strength of evidence according to BF_{ia}, Kass and Raftery (1995) proposed rules as given in Table 2.7. Note that, the number 1 is an important reference value for the interpretation of the Bayes factor. If $BF_{ia} > 1$, H_i obtains more support from the data than H_a, and if $BF_{ia} < 1$, H_a obtains more support from the data than H_i. This interpretation of the Bayes factor also applies to BF_{ii_c} and $BF_{ii'}$, the latter will be elaborated below.

In this paper, both c_i and f_i are estimated by sampling form the prior and posterior distribution, respectively. Note that the posterior is approximated by a multivariate normal distribution, for which Gibbs sampler can be particularly used. The Gibbs sampler obtains draws from the posterior distribution and allows those sample draws to be summarized to obtain a full description of the posterior distribution of model parameters. According to the previous paragraph, c_i is the proportion of the prior sample in agreement with H_i, and f_i is the proportion of posterior sample in agreement with H_i. In order to obtain accurate estimates of c_i and f_i, the decomposition of the Bayes factor presented in Chapter 10 of Hoijtink (2012) is used. The technical

details of the Gibbs sampler and decomposition of the Bayes factor are presented in Appendix 2.D.

A software package BIG can be used for the computation of the Bayes factor using estimates and covariance matrix of the structural parameters and the hypothesis H_i of interest as input. This package can be downloaded from `http://informative-hypotheses.sites.uu.nl/software/`, and a user manual is provided in Appendix 2.E. Execution of BIG renders BF_{ia} and BF_{ii_c} accompanied by their Monte Carlo error (MC error), that is, an estimate of the standard deviation of each Bayes factor due to sampling (Hoijtink, 2012, p.211) and a 95% credible interval.

If the evaluation of two or more constrained hypotheses is of interest, the program BIG has to be run once for every hypothesis, The Bayes factor that compares two hypotheses H_i and $H_{i'}$ can be obtained as:

$$BF_{ii'} = BF_{ia}/BF_{i'a} = \frac{f_i}{c_i}/\frac{f_{i'}}{c_{i'}}. \tag{2.13}$$

As can be seen, using BF_{ia} and $BF_{i'a}$ obtained using `BIG`, $BF_{ii'}$ can straightforwardly be computed. To present the information in the Bayes factors computed for a set of three or more hypotheses in an accessible manner, the Bayes factors can be transformed into posterior model probabilities (PMPs) (Hoijtink, 2012, p.52). Assuming that a priori of each hypothesis is equally likely, these PMPs can be computed as:

$$PMP_i = \frac{BF_{ia}}{\sum_i BF_{ia}}. \tag{2.14}$$

The PMPs are a representation of the support in the data for each hypothesis on a scale between 0 and 1. These PMPs convey the same information as the corresponding Bayes factors and will be helpful if two or more Bayes factors have to be evaluated.

2.6 Results for the two examples

2.6.1 Example 1 (Continued)

The inequality constrained hypothesis H_1 is compared with its complement H_{1_c} in the path model predicting adolescent antisocial behaviour. BF_{11_c} can be computed using BIG described in Appendix 2.E. Running the program with the parameter estimates and covariance matrix in Table 2.3 renders $BF_{11_c} = 39.89$. This implies that the support in the data for H_1 is 39.89 times larger than the support for H_{1_c}. Based on the rules of Kass and Raftery (1995) displayed in Table 2.7, $BF_{11_c} = 39.89$ is strong evidence in favor of H_1. The 95% credible interval of BF_{11_c} is $(32.56, 46.84)$. This reflects that the MC error is rather small and does not affect our conclusions.

In a psychological perspective this implies that adolescent disclosure is the strongest determinant of antisocial behavior among the three parent-adolescent relationships,

Table 2.8: Bayes Factors and Posterior Model Probabilities

H_i	BF_{ia}	95% credible interval	PMP_i
H_1	1.72	(1.49, 1.96)	.24
H_2	1.89	(1.62, 2.16)	.26
H_3	3.65	(3.02, 4.33)	.50

that is, positive quality, negative quality and adolescent disclosure, and that deviant peers is overall the most prominent predictor of problem behavior in adolescence.

It should be noted that H_1 can also be compared to other competing models, for instance, $H_2 : \theta_{21} > \theta_{22} > \theta_{23} > \theta_0$. The selection of competing hypotheses depends on the expectations of researchers. In the path model, hypothesis H_1 against its complement H_{1c} is evaluated using BF_{11_c}, whereas in the LR model presented in the next section, three competing hypotheses will be evaluated via Bayes factors and posterior model probabilities.

2.6.2 Example 2 (Continued)

With respect to the LR model predicting the probability of being admitted to graduate school, three inequality constrained hypotheses are considered. To determine the best hypothesis among H_1, H_2 and H_3, the Bayes factor of each inequality constrained hypothesis against the unconstrained hypothesis will be calculated using (2.11). This can be achieved by running the BIG three times using the parameter estimates and covariance matrix displayed in Table 2.6. After obtaining these Bayes factors, posterior model probabilities will be computed. The results are listed in Table 2.8.

As can be seen in Table 2.8, $BF_{3a} = 3.65$ is larger than others, which reflects that H_3 is most supported by the data. This can also be seen from the corresponding PMPs: H_3 has the largest PMP. According to the rules shown in Table 2.7, $BF_{3a} = 3.65$ is positive evidence in favor of H_3, whereas both $BF_{1a} = 1.72$ and $BF_{1a} = 1.89$ are only anecdotal evidence, that is, they are not convincing evidence in favor of the corresponding hypotheses. Note that, using (2.13) we obtain that $BF_{31} = 2.12$ and $BF_{32} = 1.93$, that is, H_3 is not yet convincingly better than H_1 and H_2. Note finally that the 95% credible intervals (see Table 2.8) reflecting the size of the MC error are relatively small. This implies that our conclusions are not affected by the MC error.

In conclusion, although not yet convincingly, H_3 is preferred. This suggests that the impact of GPA declines with ascending rank level and that the GRE score of the applicant from low prestige institutions is more important than the GRE score for applicants from high prestige institutions.

2.7 Performance of normal approximations

Normal approximations are often used in statistics. Examples are Wald's test if a parameter is zero (Gourieroux, Holly, & Monfort, 1982) and the derivation of Akaike's information criterion (Akaike, 1973). The performance of normal approximations depends on the true model and the sample size. Therefore, normal approximations may be not generally applicable and the accuracy of normal approximations needs to be assessed. In this section, we conduct two simulation studies comparing the Bayes factors based on a normal approximation of the posterior distribution to the Bayes factors based on the true posterior distribution. In the first study the Bayes factor computed using BIG is compared to the Bayes factor computed using BIEMS (Mulder et al., 2012) in the context of a multiple regression model. In the second study the Bayes factor computed using BIG is compare to the Bayes factor computed using ContingencyTable (Klugkist et al., 2010) in the context of inequality constrained hypotheses for contingency tables. BIEMS and ContingencyTable can be downloaded from `http://informative-hypotheses.sites.uu.nl/software/`.

2.7.1 Multiple regression

In the first study the following regression model is considered

$$y_i = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \epsilon_i, \tag{2.15}$$

where, θ_0 is the intercept, θ_1 and θ_2 are the regression coefficients relating each predictor to y_i, and $\epsilon_i \sim N(0, \sigma^2)$ is the residual. Three inequality constrained hypotheses are evaluated using the Bayes factor: $H_1 : \theta_1 > 0, \theta_2 > 0$, $H_2 : \theta_1 > \theta_2$, and $H_3 : H_1 \& H_2$. Data sets with sizes 20, 40, 80 and 160 are generated using the following specifications:

1. $\theta_0 = 0, \theta_1 = \theta_2 = 0, \sigma^2 = 1, \mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1, \rho = 0$
2. $\theta_0 = 0, \theta_1 = \theta_2 = .4472, \sigma^2 = .6, \mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1, \rho = 0$
3. $\theta_0 = 0, \theta_1 = \theta_2 = .378, \sigma^2 = .6, \mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1, \rho = .4$
4. $\theta_0 = 0, \theta_1 = 2\theta_2 = .5656, \sigma^2 = .6, \mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1, \rho = 0$
5. $\theta_0 = 0, \theta_1 = 2\theta_2 = .4924, \sigma^2 = .6, \mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1, \rho = .4$
6. $\theta_0 = 0, 2\theta_1 = \theta_2 = .5656, \sigma^2 = .6, \mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1, \rho = 0$
7. $\theta_0 = 0, 2\theta_1 = \theta_2 = .4924, \sigma^2 = .6, \mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1, \rho = .4$

Note that μ_1 and μ_2 denote the means of x_1 and x_2, respectively, and σ_1^2, σ_2^2 and ρ denote the variances and correlation of x_1 and x_2, separately. As can be seen, in

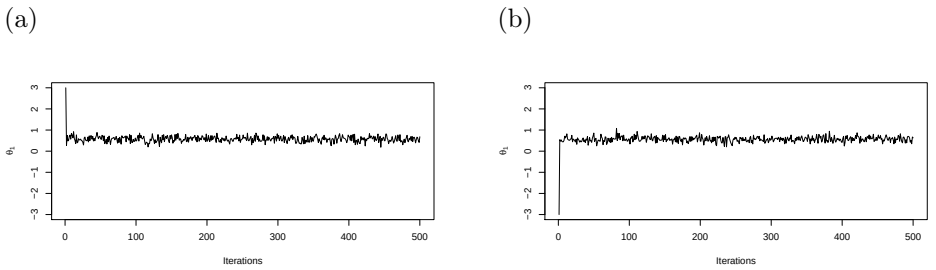

Figure 2.2: The first 500 iterations in Markov chain for θ_1 in BIG with the goal to evaluate H_2

the first population the proportion of variance explained is 0 and in the other six populations the proportion of variance explained is .4. It should be noted that data generation was executed without using Monte-Carlo replications. In fact, data was generated using the program BIEMS in such a way that the estimates of θ_0, θ_1, θ_2, σ^2, μ_1, μ_2, σ_1^2, σ_2^2 and ρ based on the generated data are exactly equal to their population values. Take the first population for example, x_{1i}, x_{2i} and ϵ_i are both generated from the normal distribution $N(0, 1)$. Thereafter, we standardize the samples of x_{1i}, x_{2i} and ϵ_i so that their means and variances are precisely 0 and 1, respectively. Finally, y_i is obtained by substituting the standardized samples of x_{1i}, x_{2i} and ϵ_i and true values of θ_0, θ_1 and θ_2 into (2.15).

After generating the data, estimates and covariance matrix of θ_1 and θ_2 can be obtained using OpenBUGS. Subsequently, running the program BIG renders the Bayes factor that is the ratio of fit and complexity. The fit and complexity are computed by sampling from posterior and prior distributions, respectively. As shown in Appendix 2.D, a particular MCMC algorithm called Gibbs sampler is adopted. Before the sample obtained from Gibbs sampler is used, there are two important steps, that is, discarding the burn-in phase and checking the convergence. We only consider the convergence of the sample from unconstrained posterior distribution used to evaluate $H_2 : \theta_1 > \theta_2$, because parameters in non-informative prior (2.9) are independent such that convergence is not an issue. For a multivariate posterior normal distribution, often within a small number of iterations the effect of the initial values vanishes and sample converges to the desired distribution. This can perhaps be best verified graphically. Figure 2.2 depicts the values sampled against the iteration number for θ_1 in population 4 with a data size of 40. Note that different starting values of -3 and 3 are chosen in the sub-figure (a) and (b), respectively. As can be seen in Figure 2.2 the Markov chain converges very rapidly, therefore, a burn-in period of 100 iterations should be more than sufficient to remove the effect of the initial values. Thereafter, the convergence of the remaining iterations is demonstrated in Table 2.9, in which the Gelman and Rubin's R statistic (Gelman et al., 2004) is rather close to

Table 2.9: Convergence of Gibbs sampler for θ_2 in BIG with the goal to evaluate H_2

Data	Sample	Population 3			Population 6		
size	size	mean	standard error	R	mean	standard error	R
	1000	0.384	6.89E-3	1.02	0.570	6.29E-3	1.01
20	5000	0.373	3.10E-3	1.01	0.561	2.88E-3	1
	40000	0.376	1.09E-3	1	0.564	1.01E-3	1
	1000	0.380	2.16E-3	1.02	0.567	1.96E-3	1.01
160	5000	0.377	9.75E-4	1	0.565	9.00E-4	1
	40000	0.378	3.42E-4	1	0.565	3.15E-4	1

1 even for a sample size of 1000 (which is the least number of iterations used in our Gibbs sampler according to the rule in Table 2.16 of Appendix 2.D). This implies that the remaining iterations in the Gibbs sampler converge very well and can be further used for computing the fit. Note that Table 2.9 only displays the results for θ_2 in population 3 and population 6 because convergence information for other populations are the same.

Bayes factors for H_1, H_2 and H_3 can be computed using both BIG and BIEMS. Note that in BIEMS we specified non-informative prior distribution for all the parameters whereas in BIG we use non-informative prior distributions for the structural parameters as specified in equation (2.9). The Bayes factors obtained with BIG and BIEMS are displayed in Tables 2.10 and 2.11 for two of the populations investigated. We present only the results for these two populations because the results obtained for the other populations are identical. As can been seen in Table 2.10 and 2.11, Bayes factors obtained using BIG and BIEMS are quite similar in every population even though the sample size may be as small as 20. This supports the use of normal approximations for the computation of the Bayes factor in regression models.

Table 2.10: Comparison of Bayes factors computed using BIG and BIEMS (population: $\theta_0 = 0, \theta_1 = \theta_2 = .4472, \sigma^2 = .6, \mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1, \rho = 0$)

data	BF_{1a}		BF_{2a}		BF_{3a}	
size	BIG	BIEMS	BIG	BIEMS	BIG	BIEMS
20	3.880	3.912	1.007	.996	3.951	3.881
40	3.971	4.019	1.006	1.003	3.930	4.031
80	4.001	3.977	1.003	.999	3.951	4.049
160	4.000	4.017	1.003	.997	3.963	3.945

Table 2.11: Comparison of Bayes factors computed using BIG and BIEMS (population: $\theta_0 = 0, 2\theta_1 = \theta_2 = .4924, \sigma^2 = .6, \mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1, \rho = .4$)

data	BF_{1a}		BF_{2a}		BF_{3a}	
size	BIG	BIEMS	BIG	BIEMS	BIG	BIEMS
20	3.425	3.525	.504	.483	1.944	1.791
40	3.792	3.790	.304	.295	1.234	1.188
80	3.932	3.952	.132	.131	.541	.494
160	3.973	3.957	.030	.031	.119	.119

2.7.2 Contingency tables

In the second simulation study inequality constrained hypotheses are evaluated for data collected in a contingency table. This contingency table is displayed in Table 2.12. A Factor A with levels 1 and 2 is crossed by a Factor B with levels 1 and 2. Denote the entries of the table by x_{11}, x_{12}, x_{21} and x_{22}, which denote the number of persons in each entry. Denote the corresponding probabilities by p_{11}, p_{12}, p_{21} and p_{22}. Note that the posterior distribution of these probabilities is Dirichlet (Klugkist et al., 2010). Using a normal approximation of this posterior will be a strong test of the viability of our approach. Also note that in the contingency table it holds that $p_{11} + p_{12} + p_{21} + p_{22} = 1$. To deal with this dependency, we computed a normal approximation to the posterior distribution of p_{11}, p_{12}, and p_{21}. Whenever p_{22} is used in an inequality constrained hypotheses it is implicitly replaces by $1 - p_{11} - p_{12} - p_{21}$.

The following inequality constrained hypotheses (which belongs to an equivalent set) will be evaluated: $H_1 : p_{11} > p_{12}, p_{21} < p_{22}$. Data have been generated from the following populations with sample sizes N equal to 20, 40, 80 and 160 such that $x_{11}/N = p_{11}$, $x_{12}/N = p_{12}$, $x_{21}/N = p_{21}$ and $x_{22}/N = p_{22}$.

1. $p_{11} = p_{22} = 0.3, p_{12} = p_{21} = 0.2$,
2. $p_{11} = p_{22} = 0.2, p_{12} = p_{21} = 0.3$,
3. $p_{11} = p_{22} = 0.35, p_{12} = p_{21} = 0.15$,
4. $p_{11} = p_{22} = 0.15, p_{12} = p_{21} = 0.35$.

Table 2.12: A hypothetical contingency table

		Factor B	
		Level 1	Level 2
Factor A	Level 1	x_{11}	x_{12}
	Level 2	x_{21}	x_{22}

Table 2.13: Comparison of Bayes factors computed using BIG and ContingencyTable (Conti denotes ContingencyTable)

	$p_{11} = p_{22} = 0.3$ $p_{12} = p_{21} = 0.2$		$p_{11} = p_{22} = 0.2$ $p_{12} = p_{21} = 0.3$		$p_{11} = p_{22} = 0.35$ $p_{12} = p_{21} = 0.15$		$p_{11} = p_{22} = 0.15$ $p_{12} = p_{21} = 0.35$	
sample size	BIG	Conti	BIG	Conti	BIG	Conti	BIG	Conti
20	2.173	2.117	.264	.304	3.278	3.126	2.61E-2	4.95E-2
40	2.657	2.613	.129	.147	3.752	3.664	2.52E-3	5.83E-3
80	3.211	3.186	3.79E-2	4.53E-2	3.971	3.953	3.97E-5	1.30E-4
160	3.710	3.708	4.71E-3	5.45E-3	4.001	4.001	0	2.04E-8

The results are displayed in Table 2.13.

As can be seen in Table 2.13, even for smaller sample sizes the Bayes factors obtained from BIG and ContingencyTable are very similar. This is further support for our assertion that normal approximations of posterior distributions render valid inferences if the goal is to evaluate inequality constrained hypotheses. It even works in a context where the true posterior distribution rather non-normal as is the case for a Dirichlet distribution.

2.8 Discussion

This study developed an approach for the evaluation of inequality constrained hypotheses in a large class of statistical models. As was discussed in the introduction, null hypothesis significance testing has a number of drawbacks: the null hypothesis is not a plausible and realistic hypothesis; p values can not be used as a measure of support for the null hypothesis; and only two hypotheses can be compared at the same time. As was shown in this paper by means of two examples, Bayesian evaluation of inequality constrained hypotheses addressed these problems in an appropriate manner. The method proposed in this research substantially extends the class of models to which Bayesian evaluation of inequality constrained hypotheses can be applied. We approximate the posterior distribution of structural parameters in any model by a normal distribution. This leads to an easy and straightforward tool for the computation of the Bayes factor for inequality constrained hypotheses. Our approximate Bayes procedure is implemented in the software package BIG that computed Bayes factors using estimates and covariance matrix of the structural parameters as input. To illustrate our approach, we evaluated several inequality constrained hypotheses in two examples with respect to path modelling and logistic regression modelling. The resulting Bayes factors quantified the evidence from the data in favor of the hypothesis compared to its complement or to other hypotheses.

The performance of normal approximations to posterior distributions was evaluated by means of a comparison of the resulting Bayes factors with the Bayes factors obtained from the true distributions. First, we computed Bayes factors of inequality

constrained hypotheses in a regression model using BIG for normal approximations and BIEMS for true distributions. The comparison of these two approaches demonstrated that BIG with normal approximations performs as well as BIEMS in regression models. Secondly, we evaluated inequality constrained hypotheses using the data in a contingency table and compared the Bayes factors obtained from BIG and ContingencyTable. The results showed that there were no notable difference between the Bayes factors from BIG and ContingencyTable. This indicates that normal approximations can also be used in models with non-normal posterior distributions. Eventually, it can be concluded that our approach based on normal approximations nicely performs not only in normal models but also in non-normal models.

2.A A comparison of two standardization approaches

In many statistical models structural parameters have to be standardized when they are compared in hypotheses. To obtain standardized structural parameters, the dependent and independent variables are often standardized using their observed means and variances. This approach is widely and easily applied by researchers. However, a possible flaw is that the data is used twice, that is, both to standardize the variables and to evaluate the inequality constrained hypotheses. An alternative approach is to directly obtain the estimates and covariance matrix of the standardized structural parameters.

The simulation study with respect to the regression model presented in the paper will be used to compare both approaches. For this simple regression model estimates and covariance matrix of standardized regression coefficients are obtained using OpenBUGS without standardizing the data. Consider again this regression model:

$$y_i = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \epsilon_i, \tag{2.16}$$

where θ_0 is the intercept, θ_1 and θ_2 are the regression coefficients, and $\epsilon_i \sim N(0, \sigma^2)$, that is, the residual has a normal distribution with mean 0 and variance σ^2. OpenBUGS can be used to sample θ_0, θ_1, θ_2 and σ as well as the means of x_1 and x_2 denoted by μ_1 and μ_2, respectively, and their standard deviations and covariance denoted by σ_1, σ_2 and $\rho\sigma_1\sigma_2$, separately. Note that ρ is the correlation between x_1 and x_2. Thereafter, a sample of parameter vectors indexed by $t = 1, \ldots, T$ is obtained from OpenBUGS using a Gibbs sampler. For each of these T parameter vectors the standardized parameter estimates can be computed using

$$\begin{aligned} Z(\theta_{1t}) &= \theta_{1t} \cdot \sigma_{1t}/\sigma_{yt}, \\ Z(\theta_{2t}) &= \theta_{2t} \cdot \sigma_{2t}/\sigma_{yt} \end{aligned} \tag{2.17}$$

where σ_{yt} is the sample of the standard deviation of y and can be computed based on the variance equation of the regression model:

$$\sigma_{yt}^2 = \theta_{1t}^2\sigma_{1t}^2 + \theta_{2t}^2\sigma_{2t}^2 + 2\theta_{1t}\theta_{2t}\rho\sigma_{1t}\sigma_{2t} + \sigma_t^2. \tag{2.18}$$

Table 2.14: Comparison of parameter variances and Bayes factors computed using two standardization approaches
(population: $\theta_0 = 0, \theta_1 = \theta_2 = .4472, \sigma^2 = .6, \mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1, \rho = 0$)

sample	$VAR(\theta_1)$		$VAR(\theta_2)$		BF_{1a}		BF_{2a}		BF_{3a}	
size	stdp	stdd	stdp	stdd	stdp	stdd	stdp	stdd	stdp	stdd
20	.02865	.03972	.02877	.03999	3.949	3.880	1.006	1.007	3.921	3.951
40	.01358	.01704	.01380	.01726	3.973	3.971	.999	1.006	3.986	3.930
80	.00667	.00804	.00677	.00808	4.001	4.001	.993	1.003	3.945	3.951
160	.00326	.00388	.00333	.00392	4.001	4.000	.994	1.003	3.952	3.963

Thus, a sample of $Z(\theta_1)$ and $Z(\theta_2)$ has been generated in OpenBUGS, which can be used to compute the estimates and covariance matrix of $Z(\theta_1)$ and $Z(\theta_2)$ to feed to BIG.

To study the performance of standardization for coefficients, we compared the approach in which the data is standardized and the approach in which the parameters are standardized on the basis of the variance of the regression coefficients and the Bayes factor. Consider the same data sets and hypotheses used in the first simulation study presented in the paper. Since the results in all 7 tables are rather similar, we will only present the Table 2.14 and 2.15 corresponding to the two tables presented for the first simulation study. Note that the labels "stdp" and "stdd" in the tables means the values below are obtained using the standardization of parameters and the data, respectively.

As can be seen in these tables, the variance of the standardized coefficients obtained using standardized parameters is smaller than those obtained using standardized data, whereas the Bayes factors are rather similar. This provides support for our assertion that in the context of the evaluation of inequality constrained hypotheses it is sufficient to standardize the data to obtain estimates and covariance matrix of standardized parameters without having to go through the effort to directly standardize the parameters as exemplified in this section.

Table 2.15: Comparison of coefficient variances and Bayes factors computed using two standardization approaches
(population: $\theta_0 = 0, 2\theta_1 = \theta_2 = .4924, \sigma^2 = .6, \mu_1 = \mu_2 = 0, \sigma_1^2 = \sigma_2^2 = 1, \rho = .4$)

sample	$VAR(\theta_1)$		$VAR(\theta_2)$		BF_{1a}		BF_{2a}		BF_{3a}	
size	stdp	stdd	stdp	stdd	stdp	stdd	stdp	stdd	stdp	stdd
20	.03624	.04712	.04712	.04707	3.495	3.425	.497	.504	1.943	1.944
40	.01767	.02033	.01541	.02035	3.805	3.792	.304	.304	1.219	1.234
80	.00878	.00960	.00754	.00956	3.955	3.932	.135	.132	.545	.541
160	.00434	.00463	.00371	.00464	3.973	3.973	.036	.030	.120	.119

2.B Mplus Command File

```
DATA: FILE IS PATH.dat; !  load data
VARIABLE: NAMES ARE Zx1 Zx2 Zx3 Zy1 Zy2;
MODEL: !  define the relation between the variables
Zy1 ON Zx1 Zx2 Zx3;
Zy2 ON Zx1 Zx2 Zx3 Zy1;
OUTPUT: TECH3; !  show covariance matrix in the output
SAVEDATA: TECH3 IS PATH.txt; !  save the covariance matrix in PATH.txt
```

2.C OpenBUGS and R Command Files

After specifying the following model and loading the standardized data in OpenBUGS, the estimates of structural parameters can be obtained by clicking the "stats" button in "Sample Monitor Tool".

```
model {
# The density of data of the logistic regression model are specified.
# Note that there are four groups denoted by rank.
# The number of observed data in four groups are n1=61, n2=151, n3=121
# and n4=67 as shown in Table 2.2.

# rank=1
for (i in 1:n1){
logit(p1[i]) <- theta0[1]+theta[1,1]*Zx1[i,1]+theta[1,2]*Zx1[i,2]
Zx1[i,3] ~ dbern(p1[i])}

# rank=2
for (i in 1:n2){
logit(p2[i]) <- theta0[2]+theta[2,1]*Zx2[i,1]+theta[2,2]*Zx2[i,2]
Zx2[i,3] ~ dbern(p2[i])}

# rank=3
for (i in 1:n3){
logit(p3[i]) <- theta0[3]+theta[3,1]*Zx3[i,1]+theta[3,2]*Zx3[i,2]
Zx3[i,3] ~ dbern(p3[i])}

# rank=4
for (i in 1:n4){
```

```
logit(p4[i]) <- theta0[4]+theta[4,1]*Zx4[i,1]+theta[4,2]*Zx4[i,2]
Zx4[i,3] ~ dbern(p4[i])}

# Below a noninformative normal prior is specified for all parameters

for(j in 1:4){
theta0[j] ~ dnorm(0.0,1.0E-5)
theta[j,1] ~ dnorm(0.0,1.0E-5)
theta[j,2] ~ dnorm(0.0,1.0E-5)}
}
```

When running OpenBUGS, researchers have to specify a burn-in period and sample size in the "Update Tool". After updating the sample, sufficient burn-in period and the convergence of the MCMC sample can be checked in the diagnostics part of "Sample Monitor Tool", analogous to how we check burn-in period in Figure 2.2. In the current model, we use a sample of 30000 with a burn-in period of 3000, both of which are desirable values according to the diagnostic. Subsequently, using the "coda" button in the same menu in OpenBUGS, the sample of parameters is given in "CODAchain 1" and labeled in "CODA index". In combination with the R2WinBUGS package in R, the covariance matrix of parameters can be obtained using the following R code.

```
## Note that R package "R2WinBUGS" must be loaded.  To extract
## the information with CODA format in OpenBUGS, "CODAchain 1" and
## "CODA index" produced by OpenBUGS CODA button should be saved
## as "CODAchain1.txt" and "CODAindex.txt".  Set the work directory
## where CODAchain1.txt and CODAindex.txt are saved, e.g.,"c:/openbugs"
   setwd("c:/openbugs")
## Read the coda information.
   output<-read.openbugs(stem = "c://openbugs//")
   x<-data.matrix(output,rownames.force=NA)
   y<-cov(x)
## Write the output in "covariance.txt".
   write(t(y),file='covariance.txt',ncolumns=ncol(y))
```

2.D Constrained Gibbs Sampler

When inequality constrained hypotheses are formulated using a large number of constraints, the true values of their complexity and fit may be extremely small. To accurately estimate such a small probability, a very large sample is needed. For ex-

ample, for hypothesis $H_1 : \theta_1 > \ldots > \theta_{10}$, the complexity is $c_1 = 1/10!$ according to the principle of equivalent set as elaborated before. A sample of more than 20 million draws is required to obtained a fair estimate of c_1 (Hoijtink, 2012, p.207), which may take so much time that efficient evaluation of H_1 is not feasible. This problem can be solved by means of a decomposition of the Bayes factor:

$$BF_{ia} = BF_{i_1,a} \times BF_{i_2,i_1} \times \cdots \times BF_{i_K,i_{K-1}} \tag{2.19}$$

where $i_k, k = 1, \ldots, K$ denotes a hypothesis using the constraints in the first k rows of $\boldsymbol{R}$, and $BF_{i_k,i_{k-1}}$ is defined by:

$$BF_{i_k,i_{k-1}} = \frac{f_{i_k,i_{k-1}}}{c_{i_k,i_{k-1}}}, \tag{2.20}$$

where $c_{i_k,i_{k-1}}$ denotes the proportion of the prior distribution of $H_{i_{k-1}}$ in agreement with H_{i_k}, and $f_{i_k,i_{k-1}}$ denotes the proportion of the posterior distribution of $H_{i_{k-1}}$ in agreement with H_{i_k}. Note that each of the $c_{i_k,i_{k-1}}$ and $f_{i_k,i_{k-1}}$ is much larger than c_i and f_i, and can be accurately computed using a relatively small sample from the corresponding prior and posterior distributions. As stated earlier, the complexity for $H_1 : \theta_1 > \ldots > \theta_{10}$ is $c_1 = 1/10!$ and can be decomposed by $c_1 = \frac{1}{2}\frac{1}{3}\cdots\frac{1}{10}$. Each component in the product is substantially larger than the value of $1/10!$ and needs a sample of, say, 9,600 from the prior distribution, which is much smaller than 20 millon needed without decomposition. The sufficient sample size for accurate estimations of decomposed complexities and fits is displayed in Table 2.16 (Hoijtink, 2012, p.154).

After decomposing the Bayes factor, the Gibbs sampler is used to compute $c_{i_k,i_{k-1}}$ and $f_{i_k,i_{k-1}}$. The basic principle of the Gibbs sampler is to generate a sample for each parameter from prior or posterior distribution conditionally on the current values of all the others. Suppose the number of structural parameters is J, and denote the size of the Gibbs sample by T. Sampling from the prior and posterior distributions of H_{i_k} can be achieved in the following steps:

1. Provide initial values for $\boldsymbol{\theta}^0$ that are in agreement with the constraints H_{i_k}.
2. Initialize the sample size $T = 1000$.
3. Repeat the next step $T + 100$ times for both sampling from the prior and posterior distributions, where 100 denotes the first 100 iterations, that is, the burn-in phase of the Gibbs sampler that are discarded.
4. Do for $j = 1, \ldots, J$: Sample θ_j from its distribution conditional on the current values of the other θs. In combination with the $k - 1$ constraints that are currently active, the current values of the other θs can be used to determine a lower bound L and upper bound U for θ_j. Using inverse probability sampling (Klugkist et al., 2005), it is straightforward to obtain a sample from a truncated normal distribution within L and U using three sub-steps:

Table 2.16: Gibbs sample size T determination

bound of $c_{i_k,i_{k-1}}$ or $f_{i_k,i_{k-1}}$	.5-1	.166-.5	.042-.166	.008-.042	.0014-.008	.0002-.0014	0-0.0002
T	1,000	3,000	9,600	120,000	360,000	2,520,000	10,000,000

- Specify the conditional distribution of θ_j:

$$(\theta_j|\theta_i, all\ i \neq j) \sim N(\hat{\theta}_j + \sum_{i \neq j} c_{ji}(\theta_i - \hat{\theta}_i), [(\boldsymbol{\Sigma}_\theta^{-1})_{jj}]^{-1}) \tag{2.21}$$

  where $\hat{\boldsymbol{\theta}}$ is the estimates of $\boldsymbol{\theta}$, and c_{ji} is the element of $I - [diag(\boldsymbol{\Sigma}_\theta^{-1})]^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}$ and $\boldsymbol{\Sigma}_\theta$ is the covariance matrix of $\boldsymbol{\theta}$. A more detailed discussion is given by Gelman et al. (2004, p.579)

- Sample a random number ν via a uniform distribution on the interval [0,1].
- Compute $\theta_j = \Phi_{\theta_j}^{-1}[\Phi_{\theta_j}(L) + \nu(\Phi_{\theta_j}(U) - \Phi_{\theta_j}(L))]$, where Φ_{θ_j} is the cumulative distribution function of (2.21) and $\Phi_{\theta_j}^{-1}$ is the inverse cumulative distribution function.

5. Discard the first 100 iterations and compute $c_{i_k,i_{k-1}}$ and $f_{i_k,i_{k-1}}$, that is, the proportion of the prior and posterior with the first $k-1$ constraints in agreement with the first k constraints, respectively. After estimation of $c_{i_k,i_{k-1}}$ and $f_{i_k,i_{k-1}}$ the rules displayed in Table 2.16 are used to determine whether the number of iterations T should be reset to ensure accurate computation of the Bayes factor. If T is reset, restart the computation in Step 3. If T is not reset, the computation of the Bayes factor is finished.

2.E User manual for BIG.exe

The evaluation of inequality constrained hypotheses can be executed using BIG.exe as long as structural parameter estimates and covariance matrix are obtained. It computes Bayes factor based on the decomposition presented in Chapter 10 of Hoijtink (2012), which was elaborated in Appendix 2.D. This decomposition ensures accurate estimates of the Bayes factor. The software package BIG is available on the website `http://informative-hypotheses.sites.uu.nl/software/`. To run BIG.exe only an input file named Input.txt is needed. This appendix is illustrated using the input and output files for Example 1. For each new analysis the user has to modify the input file. For both examples given in this paper the input and output files are provided with the BIG software. The input file for Example 1 is presented below:

Number of structural parameters and constraints
4 3

```
Estimates of parameters
0.046   0.089   0.126   0.674
Covariance matrix of parameters
  1.04E-3    2.25E-4  -4.70E-4  -7.21E-5
  2.25E-4    9.37E-4   7.53E-5  -2.82E-4
 -4.70E-4    7.53E-5   1.04E-3   1.78E-4
 -7.21E-5   -2.82E-4   1.78E-4   8.82E-4
Restriction matrix (R-r)
-1   0   1   0    0
 0  -1   1   0    0
-1   0   0   1    0
 0  -1   0   1    0
 0   0  -1   1    0
```

The first line is the label for the next line on which the numbers of structural parameters and constraints used in the hypothesis have to be recorded. Below the label on the third line the estimates of parameters displayed in Table 2.3 will be given, and below the label on the fifth line the covariance matrix of parameters displayed in Table 2.3 will be given as well. The label restriction matrix reflects that the constraints in hypothesis (2.2) will be recorded below using $\boldsymbol{R\theta} > \boldsymbol{r}$. The first J columns belong to $\boldsymbol{R}$, where J is the number of structural parameters, and the last column belongs to $\boldsymbol{r}$. The meaning of the K lines (one for each restriction) following the label Restriction matrix can be elaborated using a few examples:

- 1 0 -1 0 0 denote that $\theta_1 - \theta_3 > 0$, that is, $\theta_1 > \theta_3$
- -1 0 1 0 0 denote that $-\theta_1 + \theta_3 > 0$, that is, $\theta_1 < \theta_3$
- 0 1 0 0 .5 denote that $\theta_2 > .5$
- 0 -1 0 0 -1 denote that $-\theta_2 > -1$, that is, $\theta_2 < 1$
- a b c d e denotes that $a\theta_1 + b\theta_2 + c\theta_3 + d\theta_4 > e$

It can be seen that five constraints in hypothesis (2.2) leads to five rows in restriction matrix. Note that each column corresponds to one parameter, and that their order should be in line with the order of the estimates and covariance matrix of the parameters in the input file.

Executing the BIG.exe with the example input file renders the following output file:

Result:

Fits	Numbers of iterations
0.924	1000

0.851	1000
1.000	1000
1.000	1000
1.000	1000

Complexities	Numbers of iterations
0.510	3000
0.655	1000
0.616	1000
0.816	1000
0.503	1000

Total fits	Total complexities
0.786	0.084

BFia	MC error (standard deviation)	2.5percentile	97.5percentile
9.310	0.497	8.417	10.343

BFic	MC error (standard deviation)	2.5percentile	97.5percentile
39.891	3.624	32.559	46.842

As can be found in the beginning of the output file, the fits and complexities computed in each step of the decomposition of the Bayes factor, which was elaborated in Appendix 2.D, can be found. The corresponding numbers of iterations used for the computation of each fit and complexity are displayed in the same line below the label Numbers of iterations. The number of iterations is determined by the rules displayed in Table 2.16 (Hoijtink, 2012, p.154). Multiplying all the fits renders the fit of the Bayes factor BF_{ia}, labeled by Total fit, and the complexity of BF_{ia} can be obtained in the same way, labeled by Total complexity. This corresponds to f_i and c_i in (2.11) and (2.12). Subsequently, the Bayes factor of H_i versus H_a is displayed with the label BFia, followed by its MC error and 2.5 and 97.5 percentile in the same line. In the last line, the Bayes factor of H_i versus H_{i_c} is shown with the label BFic, followed by its MC error and 2.5 and 97.5 percentile as well. Note that the MC error is the standard deviation of the corresponding Bayes factor, and the 2.5 and 97.5 percentile give a 95% credible interval for the Bayes factor due to sampling error.

Chapter 3

An efficient program for the evaluation of inequality constrained hypotheses using Bayes factors in structural equation models[1]

3.1 Introduction

Applied researchers have become increasingly interested in the evaluation of inequality constrained hypotheses, because the traditional null hypothesis is often not a realistic representation of the population of interest (Cohen, 1994; Royall, 1997, p. 79-81). In structural equation models, researchers may have explicit theories or expectations, for example, if an independent variable has positive effect on a dependent variable or what might be the most representative indicator for a latent variable. These expectations can be represented by inequality constrained hypotheses among the model parameters. Inequality constrained hypotheses can be evaluated using either the frequentist approach by means of p values (see, e.g., Silvapulle & Sen, 2004; van de Schoot et al., 2010) or the Bayesian approach by means of Bayes factors (see, e.g., van de Schoot, Hoijtink, Hallquist, & Boelen, 2012; Klugkist et al., 2005; Hoijtink, 2012). In this paper, the Bayes factor (Kass & Raftery, 1995) is used as a criterion for assessing the hypotheses because p values can only reject a null hypothesis. Bayes factors on

[1]This chapter will be submitted as Gu, X., Hoijtink, H., Mulder, J., & Rosseel, Y. An efficient program for the evaluation of inequality constrained hypotheses using Bayes factor in structural equation models.
Author contributions: XG, HH, and JM designed the research. XG developed the software package, performed the data analyses and simulation study, and wrote the paper. HH, JM and YR gave feedback on software development. HH, JM and YR provided extensive feedback on constructing and writing the paper.

the other hand are able to measure the relative evidence in the data between multiple non-nested hypotheses containing inequality constraints (Wagenmakers, 2007). For this reason, Bayes factors can be viewed as a more generally applicable tool for statistical hypothesis testing than classical p values.

During the past decade, Bayesian evaluation of inequality constrained hypotheses has been studied for various statistical models. Besides statistical theory development, these studies rendered software packages that can be used by applied researchers, see Hoijtink (2012, p. 179) for an overview. As a pioneer, Klugkist et al. (2005) presented a Bayesian approach to evaluate analysis of (co)variance models (ANOVA or ANCOVA) with inequality constraints on the means. The study for ANOVA models was further developed by Kuiper and Hoijtink (2010) for the comparison of means using both Bayesian and non-Bayesian methods. This research resulted in a software package `ConfirmatoryANOVA` (Kuiper, Klugkist, & Hoijtink, 2010). Thereafter, Mulder et al. (2010) extended the previous study to multivariate linear models (MANOVA, repeated measures, multivariate regression), which is implemented in the software package `BIEMS` (Mulder et al., 2012). More recently, Gu, Mulder, Deković, and Hoijtink (2014) explored a Bayesian procedure which is incorporated in the software package `BIG` that can be used for the evaluation of inequality constrained hypotheses in a very general class of statistical models. However, analyses with `BIG` are computationally intensive and may take a lot of time.

This paper provides a new algorithm for the computation of Bayes factors, which substantially reduces the computational time. The resulting software is still referred to as `BIG`, because it has a similar function as its previous version. Furthermore, this paper proposes two prior specification methods which result in two Bayes factors with different features (Gu et al., 2014; Mulder, 2014a). `BIG` renders both Bayes factors such that researchers can choose either of them to evaluate inequality constrained hypotheses. It should be noted that `BIG` needs the estimates and covariance matrix of the parameters under consideration, which can be obtained using the `lavaan` package (Rosseel, 2012) in `R` for the analysis of structural equation models (SEM). Although other softwares such as `Mplus` and `OpenBUGS` can deal with SEM models as well, throughout this paper `lavaan` is used as the basis for analyses with `BIG`.

In what follows, Section 3.2 shortly introduces SEM models and defines inequality constrained hypotheses. For the evaluation of inequality constrained hypotheses, the Bayes factor as a criterion is briefly introduced in Section 3.3. Subsequently, Section 3.4 specifies prior and posterior distributions which are the determinants of the Bayes factor. To illustrate how to evaluate inequality constrained hypotheses using our program, Section 3.5 analyzes two classic SEM models: confirmatory factor analysis and multiple regression models with latent variables. Furthermore, the properties of the prior distributions specified in Section 3.4 are discussed in Section 3.6. The procedure for the computation of Bayes factors is presented in Section 3.7 in which seven sub-sections describe the principles and algorithms used. Thereafter, Section 3.8 conducts a simulation study to investigate the performance of our program. Fi-

nally, a user manual is provided in Appendix 3.B such that researchers can use the implementation in `BIG` successfully for the analysis of their own data.

3.2 Inequality constrained structural equation models

3.2.1 Structural equation models

The structural equation model (SEM) mainly consists of two components, i.e., the measurement model which expresses the relations between latent variables and their indicators, and the structural model which expresses the relations between endogenous and exogenous (latent) variables. The measurement model can be written by

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{\Lambda}_y \boldsymbol{\eta} + \boldsymbol{\epsilon}_y \\ \boldsymbol{x} &= \boldsymbol{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\epsilon}_x \end{aligned} \tag{3.1}$$

where $\boldsymbol{y}$ and $\boldsymbol{x}$ denote the vectors of endogenous and exogenous observed variables, respectively, $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ denote the vectors of endogenous and exogenous latent variables, respectively, $\boldsymbol{\Lambda}_y$ and $\boldsymbol{\Lambda}_x$ are the corresponding matrices of factor loadings, and the measurement errors $\boldsymbol{\epsilon}_y$ and $\boldsymbol{\epsilon}_x$ have zero means and covariance matrices $\boldsymbol{\Psi}_{\epsilon_y}$ and $\boldsymbol{\Psi}_{\epsilon_x}$, respectively.

The structural model represents the relations among latent variables:

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\delta}, \tag{3.2}$$

where $\mathbf{B}$ and $\boldsymbol{\Gamma}$ are matrices of regression coefficients, and $\boldsymbol{\delta}$ with mean of $\mathbf{0}$ and covariance matrix of $\boldsymbol{\Psi}_\delta$ is the error term. In addition,

$$\boldsymbol{\Phi}_\eta = (\mathbf{I} - \mathbf{B})^{-1}(\boldsymbol{\Gamma}\boldsymbol{\Phi}_\xi\boldsymbol{\Gamma}^T + \boldsymbol{\Psi}_\delta)(\mathbf{I}^T - \mathbf{B}^T)^{-1}, \tag{3.3}$$

where $\boldsymbol{\Phi}_\eta$ and $\boldsymbol{\Phi}_\xi$ are the covariance matrices of the latent variables $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$, respectively. Note that both $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ may contain observed variables if one wants to model the relationship between observed variables. This can be done by creating single-indicator latent variables (with a fixed factor loading of 1, and zero measurement error) corresponding to each observed variable.

The general framework of SEM is described by equations (3.1) and (3.2) which can be specified using `lavaan` syntax (Rosseel, 2012) in `R`. As can be seen from (3.1), (3.2) and (3.3), the non-fixed elements in $\{\boldsymbol{\Lambda}_y, \boldsymbol{\Lambda}_x, \mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\Psi}_{\epsilon_y}, \boldsymbol{\Psi}_{\epsilon_x}, \boldsymbol{\Psi}_\delta, \boldsymbol{\Phi}_\xi\}$ of a specific SEM model can be collected in a parameter vector $\boldsymbol{\lambda}$. The density of the data is given by $f(\boldsymbol{X}|\boldsymbol{\lambda})$, where $\boldsymbol{X}$ denotes the data (Bollen, 1989). Furthermore, the non-fixed parameters can be divided into $\boldsymbol{\lambda} = \{\boldsymbol{\theta}, \boldsymbol{\zeta}\}$, where $\boldsymbol{\theta}$ denotes the target parameters that will appear in the inequality constrained hypotheses elaborated in the next section, and $\boldsymbol{\zeta}$ denotes the nuisance parameters that will not.

3.2.2 Inequality constrained hypotheses

Inequality constrained hypotheses express the expectations of researchers among the (standardized) target parameters in SEM. For example, hypothesis $H_1 : \theta_1 > \theta_2$ where θ_1 and θ_2 are the coefficients of the predictors ξ_1 and ξ_2, respectively, implies that the predictor ξ_1 is stronger than ξ_2. The general form of an inequality constrained hypothesis H_i is given by

$$H_i : \boldsymbol{R}_i\boldsymbol{\theta} > \boldsymbol{r}_i, \tag{3.4}$$

where $\boldsymbol{R}_i$ is the restriction matrix containing inequality constraints, and $\boldsymbol{\theta}$ and $\boldsymbol{r}_i$ denote the target parameter vector and constant vector in H_i, respectively. We assume that the number of inequality constraints is K and the number of target parameters is J. Therefore, $\boldsymbol{R}_i$ is a $K \times J$ matrix, and the lengths of $\boldsymbol{\theta}$ and $\boldsymbol{r}_i$ are J and K, respectively. For instance, $H_2 : \theta_1 > \theta_2 > \theta_3$ is an example with $J = 3$ and $K = 2$, which leads to $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^{\mathrm{T}}$ and an augmented matrix:

$$[\boldsymbol{R}_2|\boldsymbol{r}_2] = \left[\begin{array}{ccc:c} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \end{array}\right].$$

The augmented matrix $[\boldsymbol{R}_i|\boldsymbol{r}_i]$ should be implemented as an input file in `BIG`. It should be emphasised that the application of our program is strictly limited to the evaluation of inequality constrained hypotheses. This implies that, equality constrained hypotheses like $H_3 : \theta_1 = \theta_2$, about equality constrained hypotheses like $H_4 : \theta_1 \approx \theta_2$, and range constrained hypotheses like $H_5 : 0 < \theta < 1$, can not be processed by our program. In addition, the inequality constrained hypothesis H_i can not contain contradicting constraints. For example, both $H_6 : \theta > 1$, $\theta < -2$ and $H_7 : \theta_1 > \theta_2$, $\theta_2 > \theta_1$ are invalid in our program. Our program will automatically check whether the hypothesis specified by researchers in the input file is valid or not. This will be elaborated in Section 3.7.2 and Appendix 3.B. If the hypothesis specified in the input file is invalid, the program will write a warning message in the output file.

The hypothesis H_i is often compared to an unconstrained hypothesis

$$H_u : \boldsymbol{\theta} \in \mathbb{R}^J, \tag{3.5}$$

where $\mathbb{R}^J$ denotes the J-dimensional real vector space, or to its complement

$$H_{i_c} : \text{not } H_i. \tag{3.6}$$

Furthermore, we can evaluate H_i against a competing hypothesis

$$H_{i'} : \boldsymbol{R}_{i'}\boldsymbol{\theta} > \boldsymbol{r}_{i'}. \tag{3.7}$$

The evaluation of these hypotheses can be conducted using Bayes factors, which will be elaborated in the next section.

When specifying inequality constrained hypotheses in SEM models, the target parameters may need to be standardized. For example, if hypothesis $H_1 : \theta_1 > \theta_2$ compares two regression coefficients to determine which predictor is stronger, then the coefficients θ_1 and θ_2 should be standardized to be comparable. The standardization of target parameters can be achieved by standardizing the observed and latent variables in SEM models. However, this manner might be criticized because the data is used twice, once for standardization and once for evaluation of the hypothesis (Gu et al., 2014). The `lavaan` package (Rosseel, 2012) provides an alternative approach that can directly obtain estimates and covariance matrix of standardized target parameters. This paper uses the alternative standardization approach in lavaan. To keep the notation simple, in this paper $\boldsymbol{\theta}$ will be used to denote both unstandardized and standardized target parameters.

3.3 Bayes factor

The Bayes factor of H_i against H_u is defined as the ratio of two marginal likelihoods (Jeffreys, 1961; Kass & Raftery, 1995; Hoijtink, 2012):

$$BF_{iu} = \frac{m_i(\boldsymbol{X})}{m_u(\boldsymbol{X})} = \frac{\int\int f(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{\zeta})\pi_i(\boldsymbol{\theta},\boldsymbol{\zeta})d\boldsymbol{\theta}d\boldsymbol{\zeta}}{\int\int f(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{\zeta})\pi_u(\boldsymbol{\theta},\boldsymbol{\zeta})d\boldsymbol{\theta}d\boldsymbol{\zeta}}, \tag{3.8}$$

where $\pi_i(\boldsymbol{\theta},\boldsymbol{\zeta})$ and $\pi_u(\boldsymbol{\theta},\boldsymbol{\zeta})$ denote the prior distribution under H_i and H_u (will be specified in the next section), respectively, and $f(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{\zeta})$ denotes the density of $\boldsymbol{X}$ given $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ (see Bollen, 1989). Furthermore, from equation (3.8) it follows that the Bayes factor of H_i against H_{i_c} can be obtained as $BF_{ii_c} = BF_{iu}/BF_{i_cu}$, and the Bayes factor of H_i against $H_{i'}$ is $BF_{ii'} = BF_{iu}/BF_{i'u}$.

The Bayes factor BF_{iu} quantifies the relative evidence in the data in favor of hypothesis H_i against H_u. For example $BF_{iu} = 2$ indicates that the support in the data for H_i is twice as large as the support for H_u. A general guideline for the interpretation of the Bayes factor is that $BF_{iu} \in (1, 3]$ indicates evidence for H_i that is not worth mentioning, and $BF_{iu} \in (3, 20]$, $BF_{iu} \in (20, 150]$ and $BF_{iu} > 150$ indicate positive, strong and very strong evidence for H_i, respectively (Kass & Raftery, 1995). Note that if $BF_{iu} < 1$ which implies evidence against H_i, the strength of this evidence is quantified using the rule above for the reciprocal of BF_{iu}. Furthermore, Bayes factors BF_{ii_c} and $BF_{ii'}$ can also be interpreted using the same rule. Although this rule renders a proposal to interpret the Bayes factor, it is not suggested using it strictly because this interpretation is a rough descriptive statement with respect to the standards of evidence, which could very well be modified based on the research context. For this reason users can judge by themselves when the evidence in the data is positive, strong or decisive in favor or against a hypothesis based on the observed Bayes factor.

Formula (3.8) can be simplified to (Klugkist & Hoijtink, 2007):

$$BF_{iu} = \frac{f_i}{c_i}, \tag{3.9}$$

where

$$c_i = \iint_{\theta \in \Theta_i} \pi_u(\boldsymbol{\theta}, \boldsymbol{\zeta}) d\boldsymbol{\theta} d\boldsymbol{\zeta} = \int_{\theta \in \Theta_i} \pi_u(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{3.10}$$

called relative complexity (Mulder, 2014b), is the proportion of the prior distribution (specified in the next section) in agreement with H_i, and

$$f_i = \iint_{\theta \in \Theta_i} \pi_u(\boldsymbol{\theta}, \boldsymbol{\zeta} | \boldsymbol{X}) d\boldsymbol{\theta} d\boldsymbol{\zeta} = \int_{\theta \in \Theta_i} \pi_u(\boldsymbol{\theta} | \boldsymbol{X}) d\boldsymbol{\theta}, \tag{3.11}$$

called relative fit, is the proportion of the posterior distribution (specified in the next section) in agreement with H_i. Here $\Theta_i = \{\boldsymbol{\theta} | \boldsymbol{R}_i \boldsymbol{\theta} > \boldsymbol{r}_i\}$ denotes the parameter space constrained by H_i, and $\boldsymbol{\zeta}$ is not constrained. The complexity implies how specific a hypothesis is, and the fit implies how much the data supports a hypothesis relative to H_u. The more specific the hypothesis, the less the complexity, while the more the support from the data, the larger the fit. The derivation of equation (3.9) can be found in Mulder (2014b). Equation (3.9) shows that the Bayes factor of an inequality constrained hypothesis H_i against an unconstrained hypothesis H_u can be represented as the ratio of the fit and complexity of H_i. This representation facilitates our development of the software for the evaluation of inequality constrained hypotheses.

Based on BF_{iu}, the Bayes factor BF_{ii_c} for H_i against H_{i_c}, and $BF_{ii'}$ for two competing hypotheses H_i and $H_{i'}$ can also be derived. Noting that the proportion of prior and posterior distributions in agreement with H_{i_c} are $1 - c_i$ and $1 - f_i$, respectively, it follows that

$$BF_{ii_c} = \frac{f_i}{c_i} / \frac{1 - f_i}{1 - c_i}. \tag{3.12}$$

Analogously, $BF_{ii'}$ can be obtained by

$$BF_{ii'} = BF_{iu} / BF_{i'u} = \frac{f_i}{c_i} / \frac{f_{i'}}{c_{i'}}. \tag{3.13}$$

Furthermore, an accessible manner for comparing a set of hypotheses is to transform Bayes factors into posterior model probabilities (PMPs). The PMPs are a representation of the support in the data for each hypothesis on a scale between 0 and 1. Assuming equal prior probabilities for the hypotheses, we obtain PMPs for all the competing hypotheses excluding H_u using (Hoijtink, 2012, p. 52)

$$PMP_i = \frac{BF_{iu}}{\sum_i BF_{iu}} \quad \text{for} \quad i = 1, \ldots, I_N, \tag{3.14}$$

where I_N denotes the number of competing hypotheses. The execution of our program renders both Bayes factors (3.9) and PMPs (3.14). As was shown in (3.9), the Bayes factor for H_i against H_u depends on the complexity and fit for which the prior and posterior distributions of $\boldsymbol{\theta}$ under H_u need to be specified, respectively. The specification of prior and posterior distributions will be introduced in the next section.

3.4 Prior and posterior distributions

3.4.1 Noninformative normal prior distributions

The specification of prior distributions is an important step in Bayesian hypothesis testing. As can be seen from equation (3.10), only a prior of $\boldsymbol{\theta}$ for the unconstrained hypothesis needs to be specified when evaluating inequality constrained hypotheses. In this paper, for the target parameters under H_u we specify noninformative normal priors of the form:

$$\pi_u(\boldsymbol{\theta}) = N(\mathbf{0}, \omega\boldsymbol{\Sigma}_s), \tag{3.15}$$

where $\boldsymbol{\Sigma}_s$ is the prior covariance *s*tructure of $\boldsymbol{\theta}$, and the positive number $\omega \to \infty$. Note that in `BIG` we specify $\omega = 10000$ by default. This manner of prior construction has several properties. First, the normal prior distribution is a conjugate prior when assuming that the posterior distribution is approximately normal, which will be elaborated in Section 3.4.2. Second, the use of noninformative priors results in a posterior distribution that is completely determined by the data for any sample size. Although very vague priors are not recommended when testing hypotheses with equality constraints due to Lindley-Bartlett's paradox (Lindley, 1957), such priors can be used for testing inequality constrained hypotheses (Klugkist et al., 2005). Third, the complexity is invariant to the actual specification of the prior mean such that a mean vector of $\mathbf{0}$ can be used. This property is proven in Section 3.6. In our program, we use two prior distributions with means of $\mathbf{0}$ and different covariance structures.

Gu et al. (2014) proposed the noninformative normal prior distribution with identity covariance structure

$$\pi_u^1(\boldsymbol{\theta}) = N(\mathbf{0}, \omega\boldsymbol{I}), \tag{3.16}$$

where $\mathbf{0} = (0, \ldots, 0)^{\mathrm{T}}$ and $\boldsymbol{I}$ is an identity matrix. Using this prior distribution every combination of values of target parameters is a priori equally probable.

Mulder (2014b) proposed to set the prior covariance structure equal to the posterior covariance structure. In our noninformative normal prior this implies

$$\pi_u^2(\boldsymbol{\theta}) = N(\mathbf{0}, \omega\hat{\boldsymbol{\Sigma}}_\theta), \tag{3.17}$$

where $\hat{\boldsymbol{\Sigma}}_\theta$ is the estimated covariance matrix of the (standardized) target parameters that can be obtained in the `lavaan` package (Rosseel, 2012).

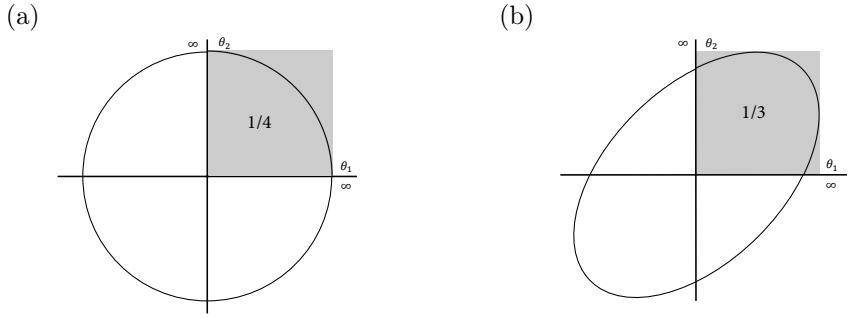

Figure 3.1: Two prior distributions of $H_1 : \theta_1 > 0, \theta_2 > 0$. Note that the grey area describes the constrained parameter space under H_1. The circle and ellipse denote the 95% iso-density control, and the numbers in the figures denote the complexities of H_1 under different priors.

The complexity (3.10) based on $\pi_u^1(\boldsymbol{\theta})$ may differ from $\pi_u^2(\boldsymbol{\theta})$. Consider the hypothesis $H_1 : \theta_1 > 0, \theta_2 > 0$. Figure 3.1 illustrates the complexities of H_1 under prior distributions

$$\pi_u^1(\boldsymbol{\theta}) \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \omega \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) \quad \text{and} \quad \pi_u^2(\boldsymbol{\theta}) \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \omega \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right).$$

As can be seen, the complexity obtained using $\pi_u^1(\boldsymbol{\theta})$ in Figure 3.1 (a) is $1/4$ which is smaller than the complexity obtained using $\pi_u^2(\boldsymbol{\theta})$ in Figure 3.1 (b), i.e., $1/3$. This illustrates how the measure of complexity depends on the covariance structure $\boldsymbol{\Sigma}_s$ in the prior distribution. The prior distributions (3.16) and (3.17) are used to compute the complexity in our program, because both of them have attractive properties which will be discussed in Section 3.6. In what follows, the posterior distribution is specified to obtain the fit (3.11).

3.4.2 Normal approximations to posterior distributions

In order to compute Bayes factors for inequality constrained hypotheses in SEM models, the asymptotic normality of the posterior distribution is used based on Laplace's method (DiCiccio, Kass, Raftery, & Wasserman, 1997; Gelman et al., 2004, p. 101-107). As elaborated in the beginning of this section, the posterior distribution only depends on the density of the data $f(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{\zeta})$ when using prior (3.16) or (3.17) for $\omega \to \infty$. The posterior distribution can be approximated by:

$$\pi_u(\boldsymbol{\theta}|\boldsymbol{X}) \approx N(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}_\theta), \tag{3.18}$$

where $\hat{\boldsymbol{\theta}}$ denotes the estimates of the target parameters, and $\hat{\boldsymbol{\Sigma}}_\theta$ is their covariance matrix. Both of them can be obtained in `lavaan` using estimation methods, such as

least square estimation and maximum likelihood estimation (Rosseel, 2012). Furthermore, to obtain standardized $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}_\theta$ `lavaan` provides approaches to standardize the observed variables and to directly standardize the target parameters. The performance of these two approaches of standardization was discussed in Gu et al. (2014), which showed that the variances of standardized parameters obtained using two approaches are different, whereas the resulting Bayes factors are similar. Now that the prior and posterior distributions have been specified, the Bayes factor can be obtained using (3.9). We will in the next section illustrate how to evaluate inequality constrained hypotheses using Bayes factors, and in Section 3.7 elaborate on how to compute Bayes factors technically in our program.

3.5 Examples

In this section, our procedure of evaluating inequality constrained hypotheses will be illustrated using two classic SEM examples. One example concerns confirmatory factor analysis (CFA), and the other example concerns multiple regression model.

3.5.1 Confirmatory factor analysis

In the first example, we reanalyze a dataset built into `lavaan` called HolzingerSwineford1939 (Rosseel, 2012). This dataset is taken from the Holzinger and Swineford 1939 (H&S) study, which is a commonly used example in factor analysis. The raw dataset consists of scores of 301 seventh and eighth grade students from the Pasteur School (n=145) and Grant-White School (n=156) who participated in 26 psychological aptitude tests. In our example, only a subset with 9 variables of the complete data is extracted to measure 3 correlated latent variables, each with three indicators, i.e.,

- a visual factor (ξ_1) is measured by visual perception (x_1), cubes (x_2) and lozenges (x_3).
- a textual factor (ξ_2) is measured by paragraph comprehension (x_4), sentence completion (x_5) and word meaning (x_6)
- a speed factor (ξ_3) is measured by addition (x_7), counting of dots (x_8) and discrimination of straight and curved capitals (x_9).

The descriptives for the observed variables are given in Table 3.1, whereas the relations between latent variables and their indicators are formulated in the next paragraph and expressed using path notation (without showing measurement errors) in Figure 3.2.

The confirmatory factor analysis model for the H&S data can be represented as:

$$\boldsymbol{x} = \boldsymbol{\Lambda}_x \boldsymbol{\xi} + \boldsymbol{\epsilon}_x, \tag{3.19}$$

Table 3.1: Descriptives for the variables in the confirmatory factor analysis

Variable		Mean	S.D.
visual perception	x_1	4.94	1.17
cubes	x_2	6.09	1.18
lozenges	x_3	2.25	1.13
paragraph	x_4	3.06	1.16
sentence	x_5	4.34	1.29
word mean	x_6	2.19	1.10
addition	x_7	4.19	1.09
dots	x_8	5.53	1.01
straight curved	x_9	5.37	1.01

where $\boldsymbol{x} = (x_1, \ldots, x_9)^T$ denotes observed variables, $\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_3)^T$ denotes latent variables,

$$\boldsymbol{\Lambda}_x^T = \begin{pmatrix} \theta_1 & \theta_2 & \theta_3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \theta_4 & \theta_5 & \theta_6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \theta_7 & \theta_8 & \theta_9 \end{pmatrix} \tag{3.20}$$

is a matrix of factor loadings, and $\boldsymbol{\epsilon}_x$ is a 3×1 vector of measurement errors with $\boldsymbol{\epsilon}_x \sim N(0, \boldsymbol{\Psi}_{\epsilon_x})$ and $\boldsymbol{\Psi}_{\epsilon_x}$ being its covariance matrix. The covariance matrix of observed variables is given by:

$$\boldsymbol{\Sigma}_x = \boldsymbol{\Lambda}_x \boldsymbol{\Phi}_\xi \boldsymbol{\Lambda}_x^T + \boldsymbol{\Psi}_{\epsilon_x}, \tag{3.21}$$

where the factor covariance matrix $\boldsymbol{\Phi}_\xi$ is a symmetric matrix:

$$\boldsymbol{\Phi}_\xi = \begin{pmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{12} & \phi_{22} & \phi_{23} \\ \phi_{13} & \phi_{23} & \phi_{33} \end{pmatrix}. \tag{3.22}$$

Because the confirmatory factor analysis model is a measurement model without a structural model, we can simply specify this model using `lavaan` syntax in `R` (see Appendix 3.A). To ensure that the target parameters are comparable, we standardize them all. As is elaborated in Appendix 3.A, `lavaan` provides both the standardized estimates and covariance matrix of target parameters. Recall that this is all the information that `BIG` needs to compute Bayes factors. Furthermore, in factor analysis models, indicators are required to both identify the model and set a metric for latent variables. This can be typically achieved either by standardizing the variances of latent variables or by constraining one factor loading per latent variable to 1. In this example, the former way is chose.

Factor loadings indicate the degree of correspondence between the factor and the indicator, with higher loadings making the indictor more representative of the factor.

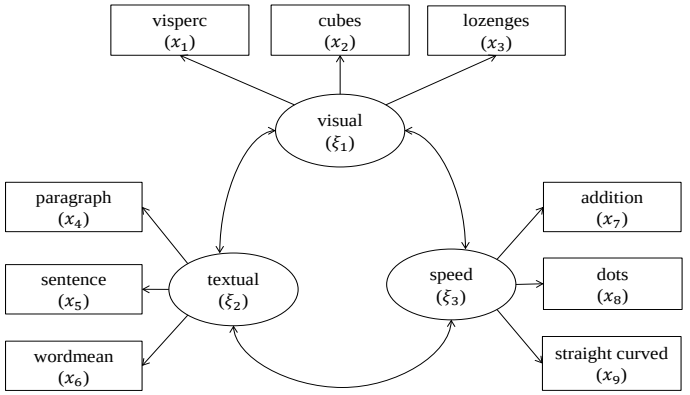

Figure 3.2: Confirmatory factor analysis

Researchers might be interested in the issue which indicator plays the most important role in defining a factor. For instance, the first indicator of every factor may be expected to be strongest, which can be represented by the following hypothesis

$$H_1 : \begin{array}{c} \theta_1 > \{\theta_2, \theta_3\} \\ \theta_4 > \{\theta_5, \theta_6\} \\ \theta_7 > \{\theta_8, \theta_9\} \end{array} . \tag{3.23}$$

We can also test a hypothesis with respect to the structure of the correlations between the latent variables. For example, we can evaluate whether the correlation between visual and textual is larger than the correlation either between visual and speed or between textual and speed:

$$H_2 : \phi_{12} > \{\phi_{13}, \phi_{23}\}. \tag{3.24}$$

Using `BIG` (the user manual of `BIG` can be found in Appendix 3.B) to compute Bayes factors for H_1 against H_u or H_{1_c} renders $BF_{1u} = 0.088$ or $BF_{11_c} = 0.085$ under prior (3.16), and $BF_{1u} = 0.087$ or $BF_{11_c} = 0.084$ under prior (3.17). For H_2 against H_u or H_{2_c}, `BIG` renders $BF_{2u} = 1.34$ or $BF_{22_c} = 1.61$ under prior (3.16), and $BF_{2u} = 1.35$ or $BF_{22_c} = 1.63$ under prior (3.17). These results imply that hypothesis H_1 is not supported by the data, and the evidence from the data for H_2 is not convincing because BF_{2u} or BF_{22_c} is quite close to 1.

Table 3.2: Descriptives for the variables in the multiple regression model

Variable	Mean	S.D.
y_1	1.06	0.16
y_2	1.05	0.15
x_{11}	1.43	0.30
x_{12}	1.33	0.24
x_{21}	2.84	0.43
x_{22}	2.91	0.38
x_{31}	2.54	0.34
x_{32}	2.47	0.32
x_4	2.12	0.31

3.5.2 Multiple regression with latent variables

In a study reported by Warren, White, and Fuller (1974) (data available on the Web at `http://tinyurl.com/warren-1974`), a sample of 98 managers of farmer cooperatives was selected with the objective of studying managerial behavior. They postulated that a latent variable manager performance (η) was predicted by three correlated latent variables, i.e., knowledge (ξ_1), orientation (ξ_2) and satisfaction (ξ_3), and an observed variable training (x_4). The latent variables η, ξ_1, ξ_2, and ξ_3 were measured based on qualitative and quantitative answers to identical questionnaires collected from a random sample of managers in farmer cooperatives. These variables are assumed to be measured with error, and the errors of measurement were computed using the split halves procedure (Warren et al., 1974) for all variables subject to measurement error:

- η is measured by y_1 and y_2,
- ξ_1 is measured by x_{11} and x_{12},
- ξ_2 is measured by x_{21} and x_{22},
- ξ_3 is measured by x_{31} and x_{32}.

The observed variables are described in Table 3.2 and the graphical specification of this structural equation model is found in Figure 3.3.

As can be seen from Figure 3.3, the relations of the variables can be represented by a multiple regression model with η, ξ_1, ξ_2, and ξ_3 that are latent. The measurement model is given by

$$\begin{aligned} \boldsymbol{y} &= \Lambda_y \eta + \boldsymbol{\epsilon}_y \\ \boldsymbol{x} &= \Lambda_x \boldsymbol{\xi} + \boldsymbol{\epsilon}_x, \end{aligned} \tag{3.25}$$

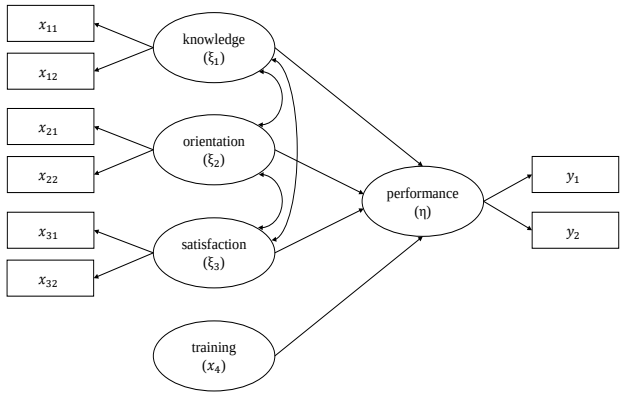

Figure 3.3: Multiple regression with latent variables

where $\boldsymbol{x} = (x_{01}, x_{02}, x_{11}, x_{12}, x_{21}, x_{22}, x_{31}, x_{32})^T$ denotes observed variables, and η and $\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_3)^T$ are latent variables. For the structural model, we have

$$\eta = \theta_0 + \theta_1\xi_1 + \theta_2\xi_2 + \theta_3\xi_3 + \theta_4 x_4 + \delta, \tag{3.26}$$

where θ_0 is the intercept, θ_1, θ_2, θ_3, and θ_4 are regression coefficients, and $\delta \sim N(0, \sigma^2)$ is the residual. This regression model is analyzed in `lavaan` (see Appendix 3.A). We standardize the coefficients to make them comparable. Using the standardized estimates and covariance matrix of these coefficients from `lavaan`, `BIG` can compute Bayes factors.

The hypothesis we evaluated is based on the results obtained by Warren et al. (1974) It states that knowledge is the strongest predictor followed by orientation, training and satisfaction. The resulting hypothesis is

$$H_3 : \theta_1 > \theta_2 > \theta_4 > \theta_3. \tag{3.27}$$

This hypothesis can be compared to, for example, knowledge is stronger than orientation followed by satisfaction and training:

$$H_4 : \theta_1 > \theta_2 > \theta_3 > \theta_4, \tag{3.28}$$

and training is stronger than satisfaction followed by orientation and knowledge:

$$H_5 : \theta_4 > \theta_3 > \theta_2 > \theta_1. \tag{3.29}$$

Table 3.3: Bayes factors and PMPs of H_3, H_4 and H_5

	BF_{iu}		PMPs	
	$\pi_u^1(\boldsymbol{\theta})$	$\pi_u^2(\boldsymbol{\theta})$	$\pi_u^1(\boldsymbol{\theta})$	$\pi_u^2(\boldsymbol{\theta})$
H_3	5.337	9.461	0.804	0.784
H_4	1.294	2.597	0.195	0.215
H_5	0.006	0.011	0.001	0.001

The results of the evaluation of these three hypotheses using `BIG` are displayed in Table 3.3 (the user manual of `BIG` can be found in Appendix 3.B). As can be seen, there is evidence in favor of H_3, no convincing evidence for H_4, and evidence against H_5. Furthermore, it can be seen from the PMPs introduced in (3.14) that H_3 receives the largest support from the data.

3.6 Properties of two prior distributions

This section discusses the properties of prior distributions $\pi_u^1(\boldsymbol{\theta})$ and $\pi_u^2(\boldsymbol{\theta})$ proposed in Section 3.4.1. First of all, the following theorem proves that when using noninformative normal prior the complexity is independent of the prior mean. This property enables us to simply specify a mean vector of $\mathbf{0}$ for $\pi_u^1(\boldsymbol{\theta})$ and $\pi_u^2(\boldsymbol{\theta})$.

Theorem 1: If $\pi_u(\boldsymbol{\theta}) = N(\boldsymbol{\theta}_0, \omega\boldsymbol{\Sigma}_s)$, the limit of $P(\boldsymbol{R}_i\boldsymbol{\theta} > \boldsymbol{r}_i|\pi_u(\boldsymbol{\theta}))$ is independent of $\boldsymbol{\theta}_0$ for $\omega \to \infty$. Note that $P(\boldsymbol{R}_i\boldsymbol{\theta} > \boldsymbol{r}_i|\pi_u(\boldsymbol{\theta}))$ is the complexity of $H_i : \boldsymbol{R}_i\boldsymbol{\theta} > \boldsymbol{r}_i$ under prior $\pi_u(\boldsymbol{\theta})$.

Proof: Let $\boldsymbol{\beta} = \boldsymbol{R}_i\boldsymbol{\theta} - \boldsymbol{r}_i$, then

$$P(\boldsymbol{R}_i\boldsymbol{\theta} > \boldsymbol{r}_i|\pi_u(\boldsymbol{\theta})) = P(\boldsymbol{\beta} > \mathbf{0}|\pi_u(\boldsymbol{\beta})) \tag{3.30}$$

with

$$\pi_u(\boldsymbol{\beta}) = N(\boldsymbol{R}_i\boldsymbol{\theta_0} - \boldsymbol{r}_i, \omega\boldsymbol{R}_i\boldsymbol{\Sigma}_s\boldsymbol{R}_i^T). \tag{3.31}$$

Therefore,

$$\begin{aligned} & \lim_{\omega\to\infty} P(\boldsymbol{R}_i\boldsymbol{\theta} > \boldsymbol{r}_i|\pi_u(\boldsymbol{\theta})) \\ = & \lim_{\omega\to\infty} P(\boldsymbol{\beta} > \mathbf{0}|\boldsymbol{\beta} \sim N(\boldsymbol{R}_i\boldsymbol{\theta_0} - \boldsymbol{r}_i, \omega\boldsymbol{R}_i\boldsymbol{\Sigma}_s\boldsymbol{R}_i^T) \\ = & \lim_{\omega\to\infty} P(\boldsymbol{\beta} > \mathbf{0}|\boldsymbol{\beta} \sim N(\frac{\boldsymbol{R}_i\boldsymbol{\theta_0} - \boldsymbol{r}_i}{\sqrt{w}}, \boldsymbol{R}_i\boldsymbol{\Sigma}s\boldsymbol{R}_i^T) \\ = & P(\boldsymbol{\beta} > \mathbf{0}|\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{R}_i\boldsymbol{\Sigma}_s\boldsymbol{R}_i^T) \end{aligned} \tag{3.32}$$

is independent of $\boldsymbol{\theta}_0$. □

Through the proof of Theorem 1, it can be seen that the complexity depends on the

restriction matrix $\boldsymbol{R}_i$ and covariance structure $\boldsymbol{\Sigma}_s$. This implies the prior distributions $\pi_u^1(\boldsymbol{\theta})$ and $\pi_u^2(\boldsymbol{\theta})$ with different covariance structures have different properties with respect to the complexity.

The prior distribution $\pi_u^1(\boldsymbol{\theta})$ in (3.16) specifies equal probabilities for every possible value in the parameter space such that it is neutral with respect to competing inequality constrained hypotheses that belong to an equivalent set (Hoijtink, 2012). A formal definition of the equivalent set is proposed by Hoijtink (2012, p. 202). An equivalent set consists of equivalent hypotheses $H_{i1}, \ldots, H_{iQ}$ for which $H_{i1} \bigcup \ldots \bigcup H_{iQ}$ encompasses 100% of the parameter space and $H_{iq} \bigcap H_{iq'} = \varnothing$ for any $q \neq q'$, where $q, q' = 1, \ldots, Q$ denotes the index of the equivalent hypotheses of H_i. These equivalent hypotheses have the same complexity, because they are equally likely a priori under $\pi_u^1(\boldsymbol{\theta})$ (Hoijtink, 2012, p. 48). This further suggests that the complexity of a hypothesis from an equivalent set is $1/Q$. In this paper, we modify the definition of the equivalent set so that it has less conditions without sacrificing properties when using $\pi_u^1(\boldsymbol{\theta})$.

A hypothesis $H_i : \boldsymbol{R}_i\boldsymbol{\theta} > \boldsymbol{r}_i$ is a member of an equivalent set if it has two characteristics:

1. $\boldsymbol{D} = \boldsymbol{R}_i\boldsymbol{R}_i^T$ has a rank of K.

2. $\frac{\boldsymbol{D}_{kk'}}{\boldsymbol{D}_{kk}}$ equals either 0, $\frac{1}{2}$ or $-\frac{1}{2}$ for $k, k' = 1, \ldots, K$ and $k' \neq k$.

Examples of hypotheses that belong to an equivalent set are

- $H_1 : \theta_1 > 0, \theta_2 > 0$ for which $\boldsymbol{R}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $\boldsymbol{D} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. With different inequality signs $Q = 4$ equivalent hypotheses, for example $H_1^{'} : \theta_1 < 0, \theta_2 > 0$, belong to the same equivalent set. The complexity for each hypothesis is $c_1 = 1/4$.

- $H_2 : \theta_1 > \theta_2 > \theta_3$ for which $\boldsymbol{R}_2 = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$ and $\boldsymbol{D} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$. This hypothesis is one of a set of six equivalent hypotheses, for example, $H_2^{'} : \theta_2 > \theta_1 > \theta_3$, and therefore the complexity for each hypothesis is $c_2 = 1/6$.

- $H_3 : \theta_1 > \{\theta_2, \theta_3\}$ for which $\boldsymbol{R}_3 = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$ and $\boldsymbol{D} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$. The other two hypotheses that belong to the same equivalent set are $H_3^{'} : \theta_2 > \{\theta_1, \theta_3\}$ and $H_3^{''} : \theta_3 > \{\theta_1, \theta_2\}$. Each equivalent hypothesis has a complexity of $c_3 = 1/3$.

- $H_4 : 2\theta_1 > \theta_2 + \theta_3 > 2\theta_4$ for which $\boldsymbol{R}_4 = \begin{bmatrix} 2 & -1 & -1 & 0 \\ 0 & 1 & 1 & -1 \end{bmatrix}$ and $\boldsymbol{D} = \begin{bmatrix} 4 & -2 \\ -2 & 4 \end{bmatrix}$. It has five equivalent hypotheses, e.g., $H_4^{'} : \theta_2 + \theta_3 > 2\theta_4 > 2\theta_1$, and its complexity is $c_4 = 1/6$.

Note that based on the definition of equivalent set in Hoijtink (2012) H_1 and H_4 are not members of equivalent sets. Examples of hypotheses that do not belong to an equivalent set are:

- $H_5 : \theta_1 > 0, \theta_2 > 0, \theta_1 > \theta_2$ for which $\boldsymbol{R}_5 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & -1 \end{bmatrix}$ and $\boldsymbol{D} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 1 & -1 & 2 \end{bmatrix}$ with a rank of 2.
- $H_6 : \theta_1 > 2\theta_2 > \theta_3$ for which $\boldsymbol{R}_6 = \begin{bmatrix} 1 & -2 & 0 \\ 0 & 2 & -1 \end{bmatrix}$ and $\boldsymbol{D} = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix}$, because $\frac{\boldsymbol{D}_{12}}{\boldsymbol{D}_{11}} = -\frac{4}{5}$ is not equal to 0, $\frac{1}{2}$ or $-\frac{1}{2}$.

Although prior distribution $\pi_u^1(\boldsymbol{\theta})$ results in the same complexity of the hypotheses from an equivalent set, the complexity based on this prior is not invariant for linear transformations of the data (Mulder, 2014a). Linear transformations of the data are often considered in repeated measures models. As an example we compute the complexity of the inequality constrained hypothesis $H_2 : \theta_1 > \theta_2 > \theta_3$ with restriction matrix

$$\boldsymbol{R}_2 = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$$

for the repeated measures data $\boldsymbol{y} = (\boldsymbol{y}_{1i}, \boldsymbol{y}_{2i}, \boldsymbol{y}_{3i}) \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}_y)$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^T$ is a mean vector and $\boldsymbol{\Sigma}_y$ is a covariance matrix of the data. In this case, the complexity of H_2 using $\pi_u^1(\boldsymbol{\theta})$ is $c_2 = 1/6$ because of the equivalent set. Now the data is transformed according to $\boldsymbol{R}_2\boldsymbol{y} = \boldsymbol{z} \sim N(\boldsymbol{\gamma}, \boldsymbol{\Sigma}_z)$ with $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)^T = (\theta_1 - \theta_2, \theta_2 - \theta_3)^T$. In terms of $\boldsymbol{\gamma}$, we obtain $H_2 : \gamma_1 > 0, \gamma_2 > 0$. Hence, using prior distribution (3.16) for $\boldsymbol{\gamma}$ the complexity of H_2 is $c_2 = 1/4$. This indicates that the complexity of a hypothesis with constraints on the parameters is generally different from the complexity of a hypothesis with constraints on the parameter differences, although two hypotheses represent the same theory.

The complexity under prior distribution $\pi_u^2(\boldsymbol{\theta})$ in (3.17) is invariant to linear one-to-one transformations of the data. Inspired by Mulder (2014a), Theorem 2 proves this invariance of complexity for the hypothesis with respect to the mean parameters in repeated measures. The proof only focuses on the repeated measure model, because the property of invariance is important when comparing the means.

Theorem 2: The complexity of $H_i : \boldsymbol{R}_i\boldsymbol{\theta} > \boldsymbol{r}_i$ when using $\pi_u^2(\boldsymbol{\theta})$ is invariant for

linear one-to-one transformation of the repeated measures data $\boldsymbol{y} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma}_y)$.

Proof: For the repeated measures, the covariance matrix of $\boldsymbol{\theta}$ is approximated by $\hat{\boldsymbol{\Sigma}}_\theta = \boldsymbol{S}_Y/n$, where $\boldsymbol{S}_Y = (\boldsymbol{Y} - \mathbf{1}\bar{\boldsymbol{y}}^T)^T(\boldsymbol{Y} - \mathbf{1}\bar{\boldsymbol{y}}^T)$ with $\bar{\boldsymbol{y}}$ being the sample means of $\boldsymbol{Y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)$. Following (3.17) the prior distribution for $\boldsymbol{\theta}$ is $\pi_u^2(\boldsymbol{\theta}) = N(0, \frac{\omega}{n}\boldsymbol{S}_Y)$.

Consider a linear one-to-one transformation $\boldsymbol{L}\boldsymbol{y} = \boldsymbol{z} \sim N(\boldsymbol{\gamma}, \boldsymbol{\Sigma}_z)$, where $\boldsymbol{L}$ is a $J \times J$ full rank matrix, and $\boldsymbol{\gamma} = \boldsymbol{L}\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}_z = \boldsymbol{L}\boldsymbol{\Sigma}_y\boldsymbol{L}^T$. After linear transformation, similarly, the covariance matrix of $\boldsymbol{\gamma}$ is approximated by $\hat{\boldsymbol{\Sigma}}_\gamma = \boldsymbol{S}_Z/n$, where $\boldsymbol{S}_Z = (\boldsymbol{Z} - \mathbf{1}\bar{\boldsymbol{z}}^T)^T(\boldsymbol{Z} - \mathbf{1}\bar{\boldsymbol{z}}^T)$ with $\bar{\boldsymbol{z}}$ being the sample means of $\boldsymbol{Z} = (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n)$. Note that $\boldsymbol{S}_Z = \boldsymbol{L}(\boldsymbol{Y} - \mathbf{1}\bar{\boldsymbol{y}}^T)^T(\boldsymbol{Y} - \mathbf{1}\bar{\boldsymbol{y}}^T)\boldsymbol{L}^T = \boldsymbol{L}\boldsymbol{S}_Y\boldsymbol{L}^T$ which implies $\hat{\boldsymbol{\Sigma}}_\gamma = \boldsymbol{L}\hat{\boldsymbol{\Sigma}}_\theta\boldsymbol{L}^T$, then the prior distribution for $\boldsymbol{\gamma}$ becomes $\pi_u^2(\boldsymbol{\gamma}) = N(0, \frac{\omega}{n}\boldsymbol{L}\boldsymbol{S}_Y\boldsymbol{L}^T)$

Let $\boldsymbol{\beta}_1 = \boldsymbol{R}_i\boldsymbol{\theta} - \boldsymbol{r}_i$ and $\boldsymbol{\beta}_2 = \boldsymbol{R}_i\boldsymbol{L}^{-1}\boldsymbol{\gamma} - \boldsymbol{r}_i$ with

$$\pi_u^2(\boldsymbol{\beta}_1) = N(0, \frac{\omega}{n}\boldsymbol{R}_i\boldsymbol{S}_Y\boldsymbol{R}_i^T), \tag{3.33}$$

and

$$\pi_u^2(\boldsymbol{\beta}_2) = N(0, \frac{\omega}{n}\boldsymbol{R}_i\boldsymbol{L}^{-1}\boldsymbol{S}_Z(\boldsymbol{R}_i\boldsymbol{L}^{-1})^T) = N(0, \frac{\omega}{n}\boldsymbol{R}_i\boldsymbol{S}_Y\boldsymbol{R}_i^T) \tag{3.34}$$

then we have

$$\begin{aligned} P(\boldsymbol{R}_i\boldsymbol{\theta} > \boldsymbol{r}_i|\pi_u^2(\boldsymbol{\theta})) &= P(\boldsymbol{\beta}_1 > 0|\pi_u^2(\boldsymbol{\beta}_1)) = P(\boldsymbol{\beta}_2 > 0|\pi_u^2(\boldsymbol{\beta}_2)) \\ &= P(\boldsymbol{R}_i\boldsymbol{L}^{-1}\boldsymbol{\gamma} > \boldsymbol{r}_i|\pi_u^2(\boldsymbol{\gamma})) \end{aligned} \tag{3.35}$$

which manifests that the complexity is invariant. □

Based on the discussion above, it can be concluded that prior distribution $\pi_u^1(\boldsymbol{\theta})$ is neutral for every value of parameters such that the complexities of hypotheses from an equivalent set are equal. However, the complexity with respect to such prior is not invariant when transforming the data. Conversely, prior distribution $\pi_u^2(\boldsymbol{\theta})$ has a benefit for the invariance of complexity, but it may favor one or more specific inequality constrained subspaces a priori under the unconstrained hypothesis.

As was shown in Section 3.4.1, the complexities from two prior distributions may be different. This is analogous to the classical information criterions AIC (Akaike, 1973) and BIC (Burnham & Anderson, 2002). The AIC has the complexity term which consists of two times of the number of parameters, whereas the BIC penalizes $log(n)$ times the number of parameters. There is not a consistent answer whether AIC or BIC should be preferred in model selection. Similarly, there is not a consistent answer to the question that which prior distribution is better. This may depend on the statistical model at hand. Therefore, our program renders Bayes factors obtained using both prior distributions.

3.7 Computation of Bayes factors

As was elaborated in Section 3.3, the Bayes factor is a ratio of fit and complexity. Because both prior distributions $\pi_u^1(\boldsymbol{\theta}) = N(0, \omega \boldsymbol{I})$ and $\pi_u^2(\boldsymbol{\theta}) = N(0, \omega \hat{\boldsymbol{\Sigma}}_\theta)$ and the posterior distribution $\pi_u(\boldsymbol{\theta}|\boldsymbol{X}) \approx N(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}_\theta)$ are normal distributions, for notational convenience each of them can be denoted by

$$p(\boldsymbol{\theta}) = N(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta). \tag{3.36}$$

Thus, the complexity and fit can be represented by the following probability

$$P(H_i) = P(\boldsymbol{R}_i\boldsymbol{\theta} > \boldsymbol{r}_i) = \int_{\boldsymbol{R}_i\boldsymbol{\theta} > \boldsymbol{r}_i} p(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{3.37}$$

This probability can be estimated by sampling from the prior or posterior distribution using the Gibbs sampler (Gelman et al., 2004).

Before presenting the core algorithm of the Gibbs sampler, we shall present two pre-steps of the sampling procedure which can efficiently reduce the computing time. First, the Bayes factor is decomposed in Section 3.7.1 such that less iterations of the Gibbs sampler are needed to accurately estimate the complexity and fit. Second, the target parameters are transformed in Section 3.7.2 such that in each iteration of the Gibbs sampler less time is needed. Thereafter, Section 3.7.3 introduces the constrained Gibbs sampling procedure based on the transformed parameters. After obtaining the samples of transformed parameters, decomposed complexities and fits can be estimated using two methods proposed in Section 3.7.4. Furthermore, the sample size of the Gibbs sampler for accurate estimation of the complexity and fit is discussed in Section 3.7.5. Section 3.7.6 summarizes the constrained Gibbs sampling procedure by which we estimate the complexity and fit, and thus the Bayes factor.

3.7.1 Decomposition of the Bayes factor

When hypothesis H_i is formulated using a relatively large number of inequality constraints, accurately estimating the complexity and fit can be computationally intensive. Consider, for example, the complexity of $H_1 : \theta_1 >, \ldots, > \theta_{10}$ under prior $\pi_u^1(\boldsymbol{\theta})$. As stated in Section 3.6, H_1 belongs to an equivalent set and its complexity is $c_1 = 1/J! = 1/10!$, that is, a really small value with the need of more than 20 million Gibbs sampler draws (Hoijtink, 2012, p. 207) to ensure the deviation of the estimate is almost never over 10%. Directly estimating this complexity may not be feasible or extremely time-consuming. Consequently, when computing the Bayes factor for hypotheses with relatively large K, a decomposition of the Bayes factor is needed (Klugkist et al., 2010):

$$BF_{iu} = BF_{i_1,u} \times BF_{i_2,i_1} \times \cdots \times BF_{i_K,i_{K-1}}, \tag{3.38}$$

where $i_k, k = 1, \ldots, K$ denotes a hypothesis using the constraints in the first k rows of $\boldsymbol{R}_i$. More specifically, $BF_{i_k,i_{k-1}}$ is defined by:

$$BF_{i_k,i_{k-1}} = \frac{f_{i_k,i_{k-1}}}{c_{i_k,i_{k-1}}}. \tag{3.39}$$

Let H_{i_k} denote the hypothesis using constraints in the first k rows of $\boldsymbol{R}_i$, then $c_{i_k,i_{k-1}}$ and $f_{i_k,i_{k-1}}$ denote the probabilities of prior and posterior distributions in agreement with H_{i_k} conditional on $H_{i_{k-1}}$, respectively. Then, the complexity and fit can be expressed by

$$c_i = \prod_{k=1}^{K} c_{i_k,i_{k-1}} \text{ and } f_i = \prod_{k=1}^{K} f_{i_k,i_{k-1}}. \tag{3.40}$$

Let

$$P(H_{i_k}|H_{i_{k-1}}) = P(\boldsymbol{R}_{i_k}\boldsymbol{\theta} > \boldsymbol{r}_{i_k}|\boldsymbol{R}_{i_1}\boldsymbol{\theta} > \boldsymbol{r}_{i_1}, \ldots, \boldsymbol{R}_{i_{k-1}}\boldsymbol{\theta} > \boldsymbol{r}_{i_{k-1}}) \tag{3.41}$$

denote either $c_{i_k,i_{k-1}}$ or $f_{i_k,i_{k-1}}$, then the probability (3.37) for c_i and f_i becomes

$$P(H_i) = P(H_{i_1}) \times P(H_{i_2}|H_{i_1}) \times \cdots \times P(H_{i_K}|H_{i_{K-1}}). \tag{3.42}$$

Because each of the probabilities in (3.42) is larger than $P(H_i)$ especially when K is large, accurately estimating $c_{i_k,i_{k-1}}$ or $f_{i_k,i_{k-1}}$ requires much less draws from the Gibbs sampler compared to directly estimating c_i or f_i. For example, the decomposed complexities for $H_1 : \theta_1 >, \ldots, > \theta_{10}$ are $c_{1_1,1_0} = \frac{1}{2}$, $c_{1_2,1_1} = \frac{1}{3}$, $\cdots$, $c_{1_{10},1_9} = \frac{1}{10}$, which can be accurately estimated using, e.g., less than 9,600 draws for $c_{1_{10},1_9}$ based on the rule proposed by Hoijtink (2012, p. 154). Although every probability in (3.42) needs to be estimated, the total sample size for decomposed c_i or f_i is still less than that without decomposition because the sample size for accurate estimation increases dramatically as K increases. This will be illustrated in Section 3.7.5. Before introducing the method for the computation of the probability (3.41), we transform the target parameters such that the inequality constrained hypothesis has a simple form, which will be elaborated in the next section.

3.7.2 Transformation of target parameters

This section simplifies the form of the hypothesis H_i using parameter transformation $\boldsymbol{\beta} = \boldsymbol{R}_i\boldsymbol{\theta} - \boldsymbol{r}_i$ such that $H_i : \boldsymbol{R}_i\boldsymbol{\theta} > \boldsymbol{r}_i$ becomes $H_i : \boldsymbol{\beta} > 0$ and the decomposed complexity or fit shown in (3.41) becomes

$$P(H_{i_k}|H_{i_{k-1}}) = P(\beta_k|\beta_1 > 0, \ldots, \beta_{k-1} > 0). \tag{3.43}$$

This transformation was also used in Mulder (in press). It has three benefits in terms of the efficiency of estimating the decomposed complexity and fit. First, the subset

of vector $\boldsymbol{\beta}$ that needs to be sampled has a length that is less than or equal to J (the length of $\boldsymbol{\theta}$). Take hypothesis $H_1 : \theta_1 > \theta_2 > \theta_3$ for example. The transformation $(\beta_1, \beta_2)^T = (\theta_1 - \theta_2, \theta_2 - \theta_3)^T$ leads to $H_1 : \beta_1 > 0, \beta_2 > 0$. Therefore, we only need to sample $\boldsymbol{\beta}$ with a length of 2. Although for another example $H_2 : \theta_1 > 0, \theta_2 > 0, \theta_1 > \theta_2$ the length of $\boldsymbol{\beta}$, where $(\beta_1, \beta_2, \beta_3)^T = (\theta_1, \theta_2, \theta_1 - \theta_2)^T$, is larger than the length of $\boldsymbol{\theta}$, only a subset $(\beta_1, \beta_2)^T$ needs to be sampled because $\beta_3 = \beta_1 - \beta_2$. This issue will be further explained in the following paragraph. Second, it is more straightforward to define the conditional probability in (3.43) than in (3.41), because each β has a lower bound of 0 if it is constrained, whereas if θ is constrained, a lower and upper bound has to be determined which will take much effort especially when K is relatively large. It will be shown in Section 3.7.3 how the constrained $\boldsymbol{\beta}$ can be sampled. Third, the conditional probability $P(\beta_k | \beta_1 > 0, \ldots, \beta_{k-1} > 0)$ can analytically be determined, which will be further discussed in Section 3.7.4.

Since $\boldsymbol{\theta}$ has a multivariate normal distribution (3.36), after the linear transformation, $\boldsymbol{\beta}$ also has a multivariate normal distribution $p(\boldsymbol{\beta}) = N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$, where $\boldsymbol{\mu}_\beta = \boldsymbol{R}_i \boldsymbol{\mu}_\theta - \boldsymbol{r}_i$ and $\boldsymbol{\Sigma}_\beta = \boldsymbol{R}_i \boldsymbol{\Sigma}_\theta \boldsymbol{R}_i^T$. It should be noted that if $\boldsymbol{R}_i$ is of full row rank, then the elements of $\boldsymbol{\beta}$ is linearly independent, otherwise the elements of $\boldsymbol{\beta}$ are not independent. Take, for example, hypothesis

$$H_3 : \begin{array}{l} \theta_1 > \theta_3 \\ \theta_1 > \theta_4 \\ \theta_2 > \theta_3 \\ \theta_2 > \theta_4 \end{array} \quad \text{with} \quad [\boldsymbol{R}_3 | \boldsymbol{r}_3] = \left(\begin{array}{cccc|c} 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \end{array} \right). \tag{3.44}$$

The matrix $\boldsymbol{R}_3$ has a rank of 3 and the transformation

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \boldsymbol{R}_3 \boldsymbol{\theta} - \boldsymbol{r}_3 = \begin{pmatrix} \theta_1 - \theta_3 \\ \theta_1 - \theta_4 \\ \theta_2 - \theta_3 \\ \theta_2 - \theta_4 \end{pmatrix} \tag{3.45}$$

implies that $\beta_4 = -\beta_1 + \beta_2 + \beta_3$. Without loss of generality, we suppose the rank of $\boldsymbol{R}_i$ is M and let

$$\boldsymbol{\beta} = (\bar{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}) = (\bar{\beta}_1, \ldots, \bar{\beta}_M, \tilde{\beta}_{M+1}, \ldots, \tilde{\beta}_K), \tag{3.46}$$

where $\bar{\boldsymbol{\beta}}$ contains M independent elements of $\boldsymbol{\beta}$, and $\tilde{\boldsymbol{\beta}}$ is a linear combination of the elements of $\bar{\boldsymbol{\beta}}$. This implies that we only need to sample $\bar{\boldsymbol{\beta}}$ from its distribution. The distribution of $\bar{\boldsymbol{\beta}}$ is $p(\bar{\boldsymbol{\beta}}) = N(\boldsymbol{\mu}_{\bar{\beta}}, \boldsymbol{\Sigma}_{\bar{\beta}})$ with $\boldsymbol{\mu}_{\bar{\beta}} = \bar{\boldsymbol{R}}_i \boldsymbol{\mu}_\theta - \bar{\boldsymbol{r}}_i$ and $\boldsymbol{\Sigma}_{\bar{\beta}} = \bar{\boldsymbol{R}}_i \boldsymbol{\Sigma}_\theta \bar{\boldsymbol{R}}_i^T$, where $\bar{\boldsymbol{R}}_i$ is a full row rank matrix that consists of M rows of $\boldsymbol{R}_i$ and $\bar{\boldsymbol{r}}_i$ is the corresponding constant vector. Although $\bar{\boldsymbol{R}}_i$ may not be unique, any set of linearly independent M rows of $\boldsymbol{R}_i$ can be chosen because the order of constraints does not affect the evaluation of the hypothesis.

The specification of $\bar{\boldsymbol{R}}_i$, $\bar{\boldsymbol{r}}_i$, and the linear combination of $\bar{\boldsymbol{\beta}}$ that renders $\tilde{\boldsymbol{\beta}}$ can be achieved using elementary row operations (Gaussian elimination) for the matrix $\boldsymbol{R}_i$. The detailed procedure is given as follows:

1. Set an identity matrix $\boldsymbol{C}$ with a rank of $\max\{K, J\}$. Initialize $\boldsymbol{A} = \boldsymbol{R}_i$, $M = K$ and $\boldsymbol{d} = (1, 2, \ldots, K)$ to record the swap of constraints in $\boldsymbol{R}_i$.
2. Repeat step (i), (ii) and (iii) for $k = 1, \cdots, K$.
   (i) If $\boldsymbol{A}_{k,k} = 0$ and $\boldsymbol{A}_{k',k} \neq 0$ where $k' > k$, then swap the kth row with the k'th row in $\boldsymbol{A}$ and $\boldsymbol{C}$, and swap d_k and $d_{k'}$ in $\boldsymbol{d}$.
   (ii) If $\boldsymbol{A}_{k,k} \neq 0$ after step (i), then let $\boldsymbol{A}_{k,j} = \boldsymbol{A}_{k,j}/\boldsymbol{A}_{k,k}$ and $\boldsymbol{C}_{k,j} = \boldsymbol{C}_{k,j}/\boldsymbol{C}_{k,k}$ for $j = 1, \cdots, J$.
   (iii) Let $\boldsymbol{A}_{k',j} = \boldsymbol{A}_{k',j} - \boldsymbol{A}_{k,j}\boldsymbol{A}_{k',k}$ and $\boldsymbol{C}_{k',j} = \boldsymbol{C}_{k',j} - \boldsymbol{C}_{k,j}\boldsymbol{C}_{k',k}$ for all $k' \neq k$ and $j = 1, \cdots, J$.
3. For $k = 1, \cdots, K$, if $\sum_{j=1}^{J} |\boldsymbol{A}_{k,j}| = 0$ then $M = M - 1$.
4. For $k = 1, \cdots, K$, if $\sum_{j=1}^{J} |\boldsymbol{A}_{k,j}| = 0$ and $\sum_{j=1}^{J} |\boldsymbol{A}_{k',j}| \neq 0$ where $k' > k$, then swap the kth row with the k'th row in $\boldsymbol{A}$ and $\boldsymbol{C}$, and swap d_k and $d_{k'}$ in $\boldsymbol{d}$.
5. Let $\boldsymbol{R}_i = (\boldsymbol{R}_{i,d_1}, \ldots, \boldsymbol{R}_{i,d_K})^T$ and $\boldsymbol{r}_i = (r_{d_1}, \ldots, r_{d_K})$, where $\boldsymbol{R}_{i,d_k}$ denotes the d_kth row of $\boldsymbol{R}_i$. Then let $\boldsymbol{\beta} = \boldsymbol{R}_i\boldsymbol{\theta} > \boldsymbol{r}_i$ in which $\bar{\boldsymbol{\beta}}$ corresponds to the first M elements in $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$ corresponds to the remaining part.

After conducting this procedure, we obtain the rank of $\boldsymbol{R}_i$, i.e., M, and $[\bar{\boldsymbol{R}}_i|\bar{\boldsymbol{r}}_i]$ which contains the first M rows of $[\boldsymbol{R}_i|\boldsymbol{r}_i]$. Furthermore, the dependence in $\boldsymbol{\beta}$ can be expressed by

$$\begin{aligned} C_{M+1,d_1} \cdot \beta_1 + \quad \cdots \quad + C_{M+1,d_K} \cdot \beta_K &= r_{d_{M+1}}, \\ \vdots & \\ C_{K,d_1} \cdot \beta_1 + \quad \cdots \quad + C_{K,d_K} \cdot \beta_K &= r_{d_K}. \end{aligned} \tag{3.47}$$

For example, for the hypothesis H_3 shown in (3.44), executing the procedure above renders

$$\begin{aligned} [\boldsymbol{A}|\boldsymbol{C}] &= \left(\begin{array}{cccc:cccc} 1 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 1 \end{array}\right) \\ &\rightarrow \left(\begin{array}{cccc:cccc} 1 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & -1 & -1 & 1 & 1 & 0 \\ 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 \end{array}\right) \end{aligned} \tag{3.48}$$

and $\boldsymbol{d} = (1, 3, 2, 4)$ which means the second and third rows have been swapped. Since there are three non-zero rows in $\boldsymbol{A}$ after Gaussian elimination, the rank of $\boldsymbol{R}_i$ is $M = 3$ and the first three rows of $\boldsymbol{R}_i$ are independent because they correspond to the non-zero rows. Furthermore, according to (3.47) the last row of $\boldsymbol{C}$ after Gaussian elimination indicates $\beta_1 - \beta_2 - \beta_3 + \beta_4 = 0$, i.e., $\beta_4 = -\beta_1 + \beta_2 + \beta_3$.

As elaborated earlier, our program can not handle hypotheses that contain equality constraints, about equality constraints and range constraints. The procedure of transforming parameters can be used to check whether the hypothesis under evaluation can be handled by `BIG`. The invalid hypothesis corresponds to the situation that all the elements in the kth row of $\boldsymbol{C}$ where $k > M$ have the same sign after the procedure above is executed. For example, the execution of our procedure for the range hypothesis $H_4 : 0 < \theta_1 < \theta_2 < 1$ results in

$$[\boldsymbol{A}|\boldsymbol{C}] = \left(\begin{array}{cc:ccc} 1 & 0 & 1 & 0 & 0 \\ -1 & 1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 1 \end{array}\right) \rightarrow \left(\begin{array}{cc:ccc} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{array}\right). \tag{3.49}$$

As can be seen, the rank of $\boldsymbol{R}_4$ for H_4 is 2, and all the elements in the last row of $\boldsymbol{C}$ are positive. This corresponds to a hypothesis that contains one or more equality, about equality or range constraints.

After the transformation of parameters, the probability $P(\beta_k|\beta_1 > 0, \ldots, \beta_{k-1} > 0)$ from equation (3.43) can be estimated using the constrained Gibbs sampler. This will be discussed in the next section.

3.7.3 Constrained Gibbs sampler

The constrained Gibbs sampler is applied to estimate each decomposed complexity and fit. The basic principle of the Gibbs sampler is to sequentially generate a sample for each β conditionally on the current values of all the others. As was elaborated before, only $\bar{\boldsymbol{\beta}}$ needs to be sampled, and the sample of $\tilde{\boldsymbol{\beta}}$ can be computed using the sample of $\bar{\boldsymbol{\beta}}$. Since $\bar{\boldsymbol{\beta}}$ is normally distributed, the conditional distribution of any parameter of $\bar{\boldsymbol{\beta}}$ given the remaining parameters is also normal. In each iteration, $\bar{\beta}_k^t$, where t denotes the iteration index of the Gibbs sampler and $k = 1, \ldots, M$, can be sampled from the following conditional distribution

$$p(\bar{\beta}_k^t|\bar{\beta}_{l\neq k}^t) = N(\mu_{\bar{\beta}_k} + \sum_{l=1}^{k-1} b_{kl}(\bar{\beta}_l^t - \mu_{\bar{\beta}_l}) + \sum_{l=k+1}^{M} b_{kl}(\bar{\beta}_l^{t-1} - \mu_{\bar{\beta}_l}), [(\boldsymbol{\Sigma}_{\bar{\beta}}^{-1})_{kk}]^{-1}) \tag{3.50}$$

where $\mu_{\bar{\beta}_k}$ is the mean of $\bar{\beta}_k$ in this full conditional distribution, b_{kl} is the element at the kth row and lth column in the matrix $\boldsymbol{B}_{M\times M} = \boldsymbol{I} - [diag(\boldsymbol{\Sigma}_{\bar{\beta}}^{-1})]^{-1}\boldsymbol{\Sigma}_{\bar{\beta}}^{-1}$ with $\boldsymbol{\Sigma}_{\bar{\beta}}$ being the covariance matrix of $\bar{\boldsymbol{\beta}}$ and $\boldsymbol{I}$ being a $M \times M$ identity matrix, and $(\boldsymbol{\Sigma}_{\bar{\beta}}^{-1})_{kk}$

is the element at the kth row and kth column in $\boldsymbol{\Sigma}_{\bar{\beta}}^{-1}$. The derivation of equation (3.50) can be found in Gelman et al. (2004, p. 579).

The estimation of probability $P(\beta_k|\beta_1 > 0, \ldots, \beta_{k-1} > 0)$ requires a sample of $\bar{\boldsymbol{\beta}} = (\bar{\beta}_1, \ldots, \bar{\beta}_M)$ from the prior or posterior distribution that is in agreement with the first $k-1$ constraints $\beta_1 > 0, \ldots, \beta_{k-1} > 0$. Using the current value of $\boldsymbol{\beta}$ and the linear restriction if R_i is not of full row rank, a lower bound L and a upper bound U of $\bar{\boldsymbol{\beta}}$ can be specified. More specifically, if $k \leq M+1$ then $(\bar{\beta}_1, \cdots, \bar{\beta}_k)$ are sampled with a lower bound of $L = 0$ and no upper bound, and other βs are not constrained. If $k > M+1$, all $\bar{\boldsymbol{\beta}}$ have a lower bound of $L = 0$, and $(\tilde{\beta}_{M+1} > 0, \ldots, \tilde{\beta}_{k-1} > 0)$ will be used to define a further lower bound and a upper bound of $\bar{\boldsymbol{\beta}}$ based on their dependence. Using inverse probability sampling (Gelfand, Smith, & Lee, 1992), it is straightforward to obtain a sample from equation (3.50) constrained in (L, U) according to the following two steps.

(i) Randomly generate a number ν via a uniform distribution on the interval [0,1].

(ii) Compute $\bar{\beta}_k = \Phi_{\bar{\beta}_k}^{-1}[\Phi_{\bar{\beta}_k}(L) + \nu(\Phi_{\bar{\beta}_k}(U) - \Phi_{\bar{\beta}_k}(L))]$, where $\Phi_{\bar{\beta}_k}$ is the cumulative distribution function of (3.50) and $\Phi_{\bar{\beta}_k}^{-1}$ is the inverse cumulative distribution function.

Running the Gibbs sampler for $t = 1, \ldots, T$ iterations renders a sample of each component of $\bar{\boldsymbol{\beta}} = (\bar{\beta}_1, \ldots, \bar{\beta}_M)$. As elaborated in Section 3.7.2, $\tilde{\boldsymbol{\beta}}$ is linearly dependent on $\bar{\boldsymbol{\beta}}$. Thus, we can also obtain a sample of $\tilde{\boldsymbol{\beta}}$ using the sample of $\bar{\boldsymbol{\beta}}$ and equation (3.47).

The choice of burn-in period and the check of convergence are important steps in the Gibbs sampler. In our method, however, we specify the prior distribution and approximate the posterior distribution with a multivariate normal distribution. Therefore, convergence is not an issue because the sample from multivariate normal distribution converges very fast even if the initial value is far away from the prior or posterior mode. This is explicitly illustrated in Gu et al. (2014), which applies the constrained Gibbs sampler to multivariate normal distributions as well. In addition, Gu et al. (2014) also shows that within a burn-in period of 100 iterations the effect of the initial values is eliminated and the sample converges to the desired distribution. Thus, we discard the first 100 iterations as a burn-in phase of the Gibbs sampler. In the next section, two methods for estimating the decomposed complexity and fit are presented based on the samples of $\boldsymbol{\beta}$ obtained in this section.

3.7.4 Two methods for estimating complexity and fit

In this section, we propose two approaches to estimate the probability (3.43) after obtaining the samples of $\boldsymbol{\beta}$ of size T from either prior or posterior distribution. A

straightforward manner is counting the number of samples in agreement with $\beta_k > 0$:

$$P(\beta_k > 0|\beta_1 > 0, \ldots, \beta_{k-1} > 0) = T^{-1} \sum_{t=1}^{T} I(\beta_k^t > 0|\beta_1^t > 0, \ldots, \beta_{k-1}^t > 0), \quad (3.51)$$

where $I(\cdot)$ denotes the indicator function which is 1 if the argument is true and 0 otherwise.

Particularly for estimating this probability with respect to the first M decomposed constraints $\bar{\boldsymbol{\beta}} > 0$, we adopt a more efficient approach inspired by Gelfand et al. (1992) and used in Morey, Rouder, Pratte, and Speckman (2011), and Mulder (in press). The principle of this method is that the density of the univariate β_k can be approximated by the average of its full conditional density constructed using the current sample of all the other βs. This implies the probability $P(\beta_k > 0)$ given the density of β_k can be approximated by the average of $P(\beta_k > 0)$ given the conditional density based on different samples. Consequently, using the constrained samples for $\beta_1, \ldots, \beta_{k-1}$ in the conditional density, we obtain

$$\begin{aligned} & P(\beta_k > 0|\beta_1 > 0, \ldots, \beta_{k-1} > 0) \\ = \; & T^{-1} \sum_{t=1}^{T} P(\beta_k > 0|\beta_1^t > 0, \ldots, \beta_{k-1}^t > 0, \beta_{k+1}^t, \ldots, \beta_K^t). \end{aligned} \quad (3.52)$$

This probability can easily be computed because the conditional distribution (3.50) of β_k is a univariate normal distribution that is easily integrated for $\beta_k > 0$.

It should be emphasised that this method is not applicable for estimating decomposed complexities or fits for which $k > M$, because $\tilde{\beta}_k$ for $k > M$ is a linear combination of $\bar{\beta}_1, \ldots, \bar{\beta}_M$, which means $\tilde{\beta}_k$ is a point given $\bar{\beta}_1, \ldots, \bar{\beta}_M$. Therefore in this case equation (3.51) will be used. Despite of this limitation, the new method (3.52) is still attractive because it increases the accuracy of the estimation for a give sample size of the Gibbs sampler. This will be elaborated in the next paragraph. This implies that fewer iterations of the Gibbs sampler are needed to obtain an acceptable accuracy. Consequently, for estimating the decomposed complexities and fits in our program, the new method (3.52) is used when $k \leq M$, whereas the approach shown in (3.51) is used when $k > M$.

To investigate the performance of the two methods, we shall consider a series of hypotheses $H_1 : \theta_1 > \ldots > \theta_J$ for $J = 3, \ldots, 5$ and estimate the complexities under $\pi_u^1(\theta) = N(\mathbf{0}, \omega \boldsymbol{I})$, where $\mathbf{0}$ is a zero vector with a length of J, $\boldsymbol{I}$ is a $J \times J$ identity matrix, and $\omega \to \infty$. The true value of c_1 with respect to prior $\pi_u^1(\theta)$ in these hypotheses is known as $c_1^{True} = 1/J!$. We estimate the complexities of H_1 1000 times using each method when the sample size of the Gibbs sampler is $T = 50, 500$, and 5000. This results in $c_{11}^{(s)}$ and $c_{12}^{(s)}$ based on methods (3.51) and (3.52), respectively, where $s = 1, \ldots, 1000$. Thereafter, we compute the mean squared relative error (MSRE),

Table 3.4: MSRE of estimate using two methods

True	c_i=0.166 (J=3)		c_i=4.17E-2 (J=4)		c_i=8.33E-3 (J=5)	
	MSRE_1	MSRE_2	MSRE_1	MSRE_2	MSRE_1	MSRE_2
$T = 50$	7.76E-2	8.37E-3	0.140	3.36E-2	0.272	9.64E-2
$T = 500$	9.25E-3	7.62E-4	1.61E-2	3.34E-3	2.44E-2	8.96E-3
$T = 5000$	5.28E-4	7.78E-5	1.49E-3	3.38E-4	2.46E-3	9.15E-4

$MSRE_1 = \frac{1}{1000}\sum_{s=1}^{1000}(\frac{c_1^{True}-c_{11}^{(s)}}{c_1^{True}})^2$ and $MSRE_2 = \frac{1}{1000}\sum_{s=1}^{1000}(\frac{c_1^{True}-c_{12}^{(s)}}{c_1^{True}})^2$, to measure the accuracy of the estimation using methods (3.51) and (3.52), respectively.

Table 3.4 displays the MSREs of the estimate for c_1. As can be seen in Table 3.4, the MSREs from method (3.52) MSRE_2 are much smaller than that from method (3.51) MSRE_1. This implies that method (3.52) needs a smaller sample size of the Gibbs sampler to attain the same accuracy. Furthermore, it can be seen that the MSREs decreases as sample size increases, and small complexity $c_i = 0.166$ needs more sample size than large complexity c_i =8.33E-3 to obtain the same magnitude of MSREs. This implies we can determine sample size T for both methods (3.51) and (3.52) based on the acceptable estimation accuracy and the size of the probability under estimation. This will be discussed in the next section.

3.7.5 Sample size determination for the Gibbs sampler

This section discusses the sample size T of the Gibbs sampler needed to accurately estimate $P(\beta_k > 0|\beta_1 > 0, \ldots, \beta_{k-1} > 0)$, which has a true value P^{True}. As stated earlier, this probability is estimated using method (3.51) if $k > M$, and method (3.52) if $k \leq M$. For method (3.51), Hoijtink (2012, p. 154) proposes a rule to determine the sample size T_1 needed to accurately estimate the complexity or fit, which is shown in the top panel of Table 3.5. The criterion is that the 95% central credibility interval for the estimate has lower and upper bounds that are less than 10% different from the true value. The first row in Table 3.5 displays the true probabilities P^{True} that needs to be estimated. In addition, L-95% and U-95% demonstrate the lower and upper bounds of the 95% central credibility interval when using the corresponding T_1 above.

For method (3.52), we present a new rule to determine the sample size T_2 based on a more strict accuracy criterion, that is, the differences between both L-95% and U-95%, and P^{True} are less than 5%. We let $N(\mu_{\beta_k}, \sigma^2_{\beta_k})$ denote the distribution of β_k in $P(\beta_k|\beta_1 > 0, \ldots, \beta_{k-1} > 0)$, where μ_{β_k} is the mean and $\sigma^2_{\beta_k}$ is the variance. Then equation (3.52) becomes

$$\begin{aligned} P(\beta_k|\beta_1 > 0, \ldots, \beta_{k-1} > 0) &= P(\beta_k > 0|\beta_k \sim N(\mu_{\beta_k}, \sigma^2_{\beta_k})) \\ &= P(\beta_k > 0|\beta_k \sim N(\hat{\lambda}_k, 1)) \end{aligned} \tag{3.53}$$

Table 3.5: Gibbs sample size determination

P^{True}	0.166	4.17E-2	8.33E-3	1.39E-3	1.98E-4	2.48E-5
T_1	3,000	9,600	120,000	360,000	2,520,000	20,160,000
L-95%	0.154	3.8E-2	7.8E-3	1.27E-3	1.82E-4	2.3E-5
U-95%	0.180	4.6E-2	8.9E-3	1.52E-3	2.17E-4	2.7E-5
T_2	4,000	8,000	12,000	18,000	25,000	32,000
L-95%	0.159	3.97E-2	7.93E-3	1.32E-3	1.89E-4	2.37E-5
U-95%	0.175	4.36E-2	8.74E-3	1.46E-3	2.08E-4	2.60E-5

where $\hat{\lambda}_k = \mu_{\beta_k}/\sigma_{\beta_k}$ is the standardized population mean of β_k. The principle of the sample size determination for method (3.52) is based on two facts. First, in the Gibbs sampler, we obtain T_2 samples of β_k from $N(\mu_{\beta_k}, \sigma^2_{\beta_k})$ or standardized β_k from $N(\hat{\lambda}_k, 1)$. This implies that the distribution of the standardized sample mean of β_k, denoted by λ_k, is $N(\hat{\lambda}_k, \frac{1}{T_2})$. Second, the probability $P(\beta_k|\beta_1 > 0, \ldots, \beta_{k-1} > 0)$ is a one-to-one correspondence function of $\hat{\lambda}_k$. For example, if $\hat{\lambda}_k = 0$, we obtain a probability of 1/2, and conversely if the true value of the probability is 1/6, we would expect a $\hat{\lambda}_k$ of -0.97. These enable us to determine the sample size T_2 needed to accurately estimate $P(\beta_k > 0|\beta_1 > 0, \ldots, \beta_{k-1} > 0)$ given a true value P^{True} using the following steps.

1. Compute $\hat{\lambda}_k$ such that $P(\beta_k > 0|\beta_k \sim N(\hat{\lambda}_k, 1)) = P^{True}$, and initialize $T_2 = 1000$.

2. Sample λ_k 10000 times from $N(\hat{\lambda}_k, \frac{1}{T_2})$, and then obtain 10000 estimates of $P(\beta_k > 0|\beta_k \sim N(\hat{\lambda}_k, 1))$.

3. Using 10000 estimates of $P(\beta_k > 0|\beta_k \sim N(\hat{\lambda}_k, 1))$, compute their 95% central credibility interval (L, U).

4. If either $\frac{|L-P^{True}|}{P^{True}} > 5\%$ or $\frac{|U-P^{True}|}{P^{True}} > 5\%$, then $T_2 = T_2 + 1000$ and go to Step 2.

The bottom panel of Table 3.5 displays the sample size T_2 and the resulting L-95% and U-95% from the procedure above given corresponding P^{True}.

In `BIG`, Table 3.5 is adopted to determine the sample size T_1 and T_2 of the Gibbs sampler for estimating each decomposed complexity and fit based on methods (3.51) and (3.52). Because T_1 or T_2 is large enough to accurately estimate the corresponding P^{True} in the first row of Table 3.5, it will also be sufficient to estimate a probability that is larger than this P^{True}. We estimate $P(\beta_k|\beta_1 > 0, \ldots, \beta_{k-1} > 0)$ with a starting sample size $T_1 = 3000$ if $k > M$ or $T_2 = 4000$ if $k \leq M$, and gradually reset T_1 or T_2 based on Table 3.5 until the estimate of the complexity or fit is larger than

the corresponding P^{True}. Note that if the estimate is still less than 2.48E-5 when using the corresponding T_1 or T_2, we specify $T_1 = 100,000,000$ or $T_2 = 100,000$.

3.7.6 Summary of the computation of the Bayes factor

This section summarizes the computation of the Bayes factor for H_i against H_u, which is a ratio of the fit and complexity. The following steps describe how our program computes the complexity and fit, and therefore the Bayes factor.

1. Transform $\boldsymbol{\theta}$ into $\boldsymbol{\beta}$ using the procedure shown in Section 3.7.2. Then, we obtain $(\bar{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}})$ and M the rank of $\boldsymbol{R}_i$.

2. Repeat Step $3, \ldots, 10$ for $k = 1, \ldots, K$ to estimate each $P(\beta_k > 0 | \beta_1 > 0, \ldots, \beta_{k-1} > 0)$ for the decomposed complexity $c_{i_k, i_{k-1}}$ and fit $f_{i_k, i_{k-1}}$.

3. Initialize the sample size of the Gibbs sampler as $T_2 = 4000$ if $k \leq M$ and $T_1 = 3000$ if $k > M$, and initialize $\boldsymbol{\beta} = 0$.

4. Repeat Step 5 and 6 for $t = 1, \ldots, T_2 + 100$ iterations if $k \leq M$ or for $t = 1, \ldots, T_1 + 100$ iterations if $k > M$, where 100 denotes the first 100 iterations, that is, a burn-in phase of the Gibbs sampler.

5. If $k \leq M + 1$, then define a boundary $(L, U) = (0, \infty)$ for $\bar{\beta}_1, \ldots, \bar{\beta}_{k-1}$ and no boundary for $\bar{\beta}_k, \ldots, \bar{\beta}_K$. Thereafter, sequentially generate a sample of $\bar{\boldsymbol{\beta}}^t$ from the truncated distribution of (3.50) as previously described in Step (i) and (ii) in Section 3.7.3.

6. If $k > M + 1$, then define a boundary (L, U) for $\bar{\beta}_1, \ldots, \bar{\beta}_M$ using the linear relation between the $\bar{\boldsymbol{\beta}} > 0$ and $\tilde{\boldsymbol{\beta}} > 0$. Thereafter, sequentially generate a sample of $\bar{\boldsymbol{\beta}}^t$ from the truncated distribution of (3.50) as previously described in Step (i) and (ii) in Section 3.7.3. Then a sample of $\tilde{\boldsymbol{\beta}}^t$ is obtained by means of its linear dependence on $\bar{\boldsymbol{\beta}}^t$.

7. Discard all the iterations for which $t \leq 100$ to account the burn-in period as discussed in Section 3.7.3.

8. If $k \leq M$, compute the probability $P(\beta_k > 0 | \beta_1 > 0, \ldots, \beta_{k-1} > 0)$ $= T_2^{-1} \sum_{t=101}^{T_2+100} P(\beta_k > 0 | \beta_k \sim N(\mu_{\beta_k}^t, (\sigma_{\beta_k}^2)^t))$ using method (3.52) in Section 3.7.4.

9. If $k > M$, compute the probability $P(\beta_k > 0 | \beta_1 > 0, \ldots, \beta_{k-1} > 0)$ $= T_1^{-1} \sum_{t=101}^{T_1+100} I(\beta_k^t > 0 | \beta_1^t > 0, \ldots, \beta_{k-1}^t > 0)$ using method (3.51) in Section 3.7.4.

10. If $P(\beta_k|\beta_1 > 0, \ldots, \beta_{k-1} > 0)$ obtained in Step 8 or 9 is less than the reference value that corresponds to the current T_2 or T_1 in Table 3.5, respectively, then reset T_2 or T_1 using the value of the next column in the table and restart the procedure from Step 4. If not, the estimation of $P(\beta_k|\beta_1 > 0, \ldots, \beta_{k-1} > 0)$ is completed, which renders the decomposed complexity $c_{i_k,i_{k-1}}$ or fit $f_{i_k,i_{k-1}}$. This was elaborated in Section 3.7.5.

11. The complexity and fit can be computed by $c_i = \prod_{k=1}^{K} c_{i_k,i_{k-1}}$ and $f_i = \prod_{k=1}^{K} f_{i_k,i_{k-1}}$ shown in Section 3.7.1. Then, the Bayes factor for H_i against H_u is $BF_{iu} = f_i/c_i$.

3.8 Simulation study

In this section, the performance of our program is investigated via the comparison with the software `BIEMS` (Mulder et al., 2012) for calculating Bayes factors in multivariate normal models. We compute Bayes factors obtained from both `BIEMS` and `BIG` for two inequality constrained hypotheses in a regression model. Consider a regression model in which a dependent variable is predicted by four independent variable:

$$y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \theta_3 x_{i3} + \theta_4 x_{i4} + \epsilon_i, \tag{3.54}$$

where y_i and x_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, 4$ are the dependent variable and independent variables, respectively, θ_0 is the intercept, θ_j for $j = 1, \ldots, 4$ denotes the regression coefficients, and $\epsilon_i \sim N(0, \sigma^2)$ denotes the residual for case i with σ^2 denoting the residual variance. For this regression model we evaluate two hypotheses, $H_1 : \theta_1 > \theta_2 > \theta_3 > \theta_4$ and $H_2 : \theta_1 > \theta_3, \theta_1 > \theta_4, \theta_2 > \theta_3, \theta_2 > \theta_4$. Note that the restriction matrix $\boldsymbol{R}_1$ for H_1 is of full row rank, whereas $\boldsymbol{R}_2$ for H_2 is not. To illustrate the performance of the program, data sets of sizes 20 and 80 are generated using `BIEMS` from six populations in which the intercept is $\theta_0 = 0$, the residual variance is $\sigma^2 = 0.6$, and the means and standard deviations of independent variables are 0 and 1, respectively. Furthermore, the target parameters θ_j for $j = 1, \cdots, 4$ and the correlations among independent variables $\rho_{j'j}$ for $j' < j$ are specified such that the proportion of variance explained equals 0.4. Based on these assumptions, we consider the following six populations.

1. $\theta_1 = \theta_2 = \theta_3 = \theta_4 = 0.316$, $\rho_{12} = \rho_{13} = \rho_{14} = \rho_{23} = \rho_{24} = \rho_{34} = 0$;
2. $\theta_1 = \theta_2 = \theta_3 = \theta_4 = 0.2$, $\rho_{14} = \rho_{23} = 0.5$, $\rho_{12} = \rho_{13} = \rho_{24} = \rho_{34} = -0.5$;
3. $\theta_1 = 2\theta_2 = 2\theta_3 = 3\theta_4 = 0.498$, $\rho_{12} = \rho_{13} = \rho_{14} = \rho_{23} = \rho_{24} = \rho_{34} = 0$;
4. $\theta_1 = 2\theta_2 = 2\theta_3 = 3\theta_4 = 0.624$, $\rho_{12} = \rho_{34} = 0.5$, $\rho_{13} = \rho_{14} = \rho_{23} = \rho_{24} = -0.5$;
5. $3\theta_1 = 2\theta_2 = 2\theta_3 = \theta_4 = 0.498$, $\rho_{12} = \rho_{13} = \rho_{14} = \rho_{23} = \rho_{24} = \rho_{34} = 0$;

Table 3.6: Bayes factors computed using BIEMS and BIG. Note that $\texttt{BIEMS}_1$ and $\texttt{BIEMS}_2$ indicate priors $\pi_u^1(\boldsymbol{\theta})$ and $\pi_u^2(\boldsymbol{\theta})$ are specified in BIEMS, respectively, and $\texttt{BIG}_1$ and $\texttt{BIG}_2$ indicate priors $\pi_u^1(\boldsymbol{\theta})$ and $\pi_u^2(\boldsymbol{\theta})$ are specified in BIG, respectively.

	H_1				H_2			
Population	$\texttt{BIEMS}_1$	$\texttt{BIG}_1$	$\texttt{BIEMS}_2$	$\texttt{BIG}_2$	$\texttt{BIEMS}_1$	$\texttt{BIG}_1$	$\texttt{BIEMS}_2$	$\texttt{BIG}_2$
$1 : n = 20$	0.995	0.994	1.03	0.977	1.00	1.02	0.984	0.993
$1 : n = 80$	1.02	0.994	1.01	0.977	0.987	1.02	0.994	0.993
$2 : n = 20$	1.72	1.71	1.01	0.975	1.35	1.35	0.986	0.997
$2 : n = 80$	1.68	1.71	1.02	0.975	1.33	1.35	0.986	0.997
$3 : n = 20$	3.53	3.99	3.57	3.93	2.03	2.15	2.02	2.10
$3 : n = 80$	6.88	6.95	6.85	6.82	2.61	2.60	2.56	2.54
$4 : n = 20$	2.38	2.61	6.89	7.78	1.49	1.68	4.70	5.08
$4 : n = 80$	5.72	5.74	17.39	17.11	2.39	2.39	7.35	7.24
$5 : n = 20$	0.117	0.068	0.117	0.066	0.214	0.146	0.213	0.144
$5 : n = 80$	0.0014	0.0013	0.0014	0.0012	0.0054	0.0056	0.0065	0.0057
$6 : n = 20$	0.032	0.011	0.039	0.011	0.316	0.215	0.240	0.160
$6 : n = 80$	0	0.0004	0	0.0003	0.0061	0.0059	0.0049	0.0049

6. $3\theta_1 = 2\theta_2 = 2\theta_3 = \theta_4 = 0.624$, $\rho_{13} = \rho_{24} = 0.5$, $\rho_{12} = \rho_{14} = \rho_{23} = \rho_{34} = -0.5$.

Note that the data generated from BIEMS under each population result in the estimates of parameters that are exactly equal to their population values. After generating the data from each population, we standardize the observed variables y_i and x_{ij} for $j = 1, \ldots, 4$. This makes the resulting Bayes factors from BIEMS and BIG comparable. To compute the Bayes factor using BIG, the standardized data from each population has to be saved as a data file, e.g., regression.dat in the working directory of R program.

Table 3.6 displays the Bayes factors obtained using BIEMS and BIG. Note that in BIEMS the prior distributions are specified using the same manner as in BIG, i.e., equations (3.16) and (3.17). As can be seen, the Bayes factors obtained from BIEMS and BIG are rather similar for each population and prior specification. This implies our program based on a normal approximation of the posterior distribution performs well for the evaluation of inequality constrained hypotheses, even if the sample size is as small as 20 in the regression model with four predictors.

3.9 Discussion

Inequality constrained hypotheses provide a representation of a researcher's theory with respect to the relations between the parameters of interest in SEM models. We developed a program BIG that can evaluate theses hypotheses. The input of BIG consists of the estimates and covariance matrix of target parameters obtained from the R package lavaan, and one or more restriction matrices representing a researcher's expectations. The output from BIG consists of Bayes factors, which measure the

evidence from the data for a hypothesis, and posterior model probabilities, which can be used to compare two or more hypotheses.

BIG as discussed in this paper has two major improvements compared to its previous version (Gu et al., 2014). First of all, BIG specifies two noninformative normal prior distributions for the target parameters in inequality constrained hypotheses. The first prior distribution specifies that every combination of values is equally likely for the target parameters. This implies that the complexity of a hypothesis only depends on the hypothesis and not on the data. The second prior distribution has a data based covariance matrix which results in the complexity that is invariant to linear one-one transformation of the data. BIG renders the Bayes factors under both prior specifications. Secondly, BIG is much more efficient than its previous version, which makes it faster and therefore easier to use for applied researchers.

Acknowledgment The third author was supported by a Veni Grant provided by the Netherlands Organization for Scientific Research (NWO).

3.A Estimates and covariance matrix obtained using lavaan

BIG uses the estimates and covariance matrix of target parameters to compute Bayes factors. These can be obtained from the R package lavaan (Rosseel, 2012). This appendix illustrates how to obtain the estimates and covariance matrix of target parameters using the two examples discussed in Section 3.5.

First of all, researchers need to install the version 0.5-18 or higher version of lavaan by starting R and typing `install.packages("lavaan")`. Note that R should be upgraded to R.3.1.0 or a higher version. The user manual of the latest version of lavaan can be found at
http://cran.r-project.org/web/packages/lavaan/lavaan.pdf.
The following R syntax renders the estimates and covariance matrix for the CFA model presented in Section 3.5.1.

```
# Load lavaan package.
  library(lavaan)
# Specify the CFA model.
  CFA.model <- 'visual  =~ x1 + x2 + x3
                textual =~ x4 + x5 + x6
                speed   =~ x7 + x8 + x9'
  fit<-cfa(CFA.model,data=HolzingerSwineford1939)

# Obtain standardized estimates of parameters
  standardizedSolution(fit)
# Obtain standardized covariance matrix of parameters.
  ZVCOV <- lavInspect(fit, "vcov.std.all")
```

```
  ZVCOV[1:9,1:9]     # For target parameters in (3.23)
  ZVCOV[19:21,19:21] # For target parameters in (3.24)
```

The output of `standardizedSolution(fit)` for the CFA model is

```
       lhs op     rhs est.std    se      z pvalue
1   visual =~      x1   0.772 0.055 14.041      0
2   visual =~      x2   0.424 0.060  7.105      0
3   visual =~      x3   0.581 0.055 10.539      0
4  textual =~      x4   0.852 0.023 37.776      0
5  textual =~      x5   0.855 0.022 38.273      0
6  textual =~      x6   0.838 0.023 35.881      0
7    speed =~      x7   0.570 0.053 10.714      0
8    speed =~      x8   0.723 0.051 14.309      0
9    speed =~      x9   0.665 0.051 13.015      0
⋮
22  visual ~~ textual   0.459 0.064  7.189      0
23  visual ~~   speed   0.471 0.073  6.461      0
24 textual ~~   speed   0.283 0.069  4.117      0
```

Note that the label `visual =~ x1` denotes the factor loading θ_1 relating x_1 to ξ_1 and the label `visual ~~ textual` denotes the covariance ϕ_{12} between ξ_1 and ξ_2. We only show the results for nine factor loadings used in (3.23) and three covariances used in (3.24). The standardized estimates of the target parameters are given in the column under `est.std`. For example, the estimate of θ_4 is 0.852 in the row of `textual =~ x4`, and the estimate of ϕ_{23} is 0.283 in the row of `textual ~~ speed`.

The output of `ZVCOV` contains the standardized covariance matrix of the target parameters. We only show the covariance matrix `ZVCOV[19:21,19:21]` of ϕ_{12}, ϕ_{13}, and ϕ_{23}:

```
                 visual~~textual  visual~~speed  textual~~speed
 visual~~textual    0.0040678110   0.0007276616     0.001156340
 visual~~  speed    0.0007276616   0.0053037342     0.001480068
textual~~  speed    0.0011563398   0.0014800678     0.004723718
```

The following `R` syntax renders the estimates and covariance matrix for the regression model in Section 3.5.2.

```
# Load lavaan package.
  library(lavaan)
# Set R working director where the data is saved.
  setwd("C:/Example2")
# Read data "example2.dat".
  performance<-read.table("example2.dat",header=TRUE)
```

```
# Specify the regression model.
  perform.model<-'
              # measurement model
                 kno =~ x11+x12
                 ori =~ x21+x22
                 sat =~ x31+x32
                 per =~ y1+y2
              # regressions
                 per ~ kno + ori + sat + tra'
  fit<-sem(perform.model,data=performance)

# Obtain standardized estimates and covariance matrix
  standardizedSolution(fit)
  ZVCOV <- lavInspect(fit, "vcov.std.all")
  ZVCOV[9:12,9:12] # For target parameters in (3.27), (3.28), (3.29)
```

The output of `standardizedSolution(fit)` for the regression model is

```
   lhs op rhs est.std    se     z pvalue
⋮
9  per  ~ kno   0.478 0.161 2.960  0.003
10 per  ~ ori   0.336 0.165 2.030  0.042
11 per  ~ sat   0.151 0.105 1.440  0.150
12 per  ~ tra   0.286 0.084 3.403  0.001
⋮
```

Note that the label `per ~ kno` denotes the coefficient θ_1 which relates η to ξ_1 in the regression model (3.26). We only show the results for the four regression coefficients used in (3.27), (3.28), and (3.29). The standardized estimates of the target parameters are given in the column under `est.std`. For example, the estimate of θ_1 is 0.478 in the row of `per ~ kno`, and the estimate of θ_4 is 0.286 in the row of `per ~ tra`.

The output of `ZVCOV[9:12,9:12]` renders the standardized covariance matrix of $\theta_1, \ldots, \theta_4$:

```
             per~kno       per~ori       per~sat       per~tra
per~kno  0.026034895 -0.0223249106 -0.0050273595 -0.0011610045
per~ori -0.022324911  0.0273346337  0.0043904540 -0.0007619234
per~sat -0.005027359  0.0043904540  0.0110250662 -0.0002713825
per~tra -0.001161004 -0.0007619234 -0.0002713825  0.0070519650
```

The standardized estimates and covariance matrix of target parameters obtained in `lavaan` can be used as input for `BIG`. This will be shown in the user manual in Appendix 3.B.

3.B User Manual of BIG

`BIG` is a `Fortran 90` program developed in Microsoft Visual Studio 2005 with the IMSL 5.0 `Fortran` numerical library. This software package is free and available at http://informative-hypotheses.sites.uu.nl/software/. The downloadable folder contains an executable file `BIG`.exe, and text files Input.txt and Output.txt for the two examples used in this paper. This section provides the user manual of `BIG` such that researchers can use it for the evaluation of inequality constrained hypotheses by means of Bayes factors. The input for `BIG` contains the estimates and covariance matrix of target parameters obtained in `lavaan` and the restriction matrix $[\boldsymbol{R}_i|\boldsymbol{r}_i]$ for each hypothesis under consideration. With this input, running `BIG` renders the Bayes factor and PMP for each hypothesis. We will use the example from Section 3.5.2 to illustrate the use of `BIG`.

3.B.1 Input file

The Input.txt and `BIG`.exe files have to be located in the same folder. The input file, e.g., for the regression model from Section 3.5.2 can be found below:

```
1   #Numbers of target parameters and hypotheses under consideration
2   4 3
3   #Estimates of parameters
4   0.784 0.550 0.248 0.471
5   #Covariance matrix of parameters
6    0.026034895 -0.0223249106 -0.0050273595 -0.0011610045
7   -0.022324911  0.0273346337  0.0043904540 -0.0007619234
8   -0.005027359  0.0043904540  0.0110250662 -0.0002713825
9   -0.001161004 -0.0007619234 -0.0002713825  0.0070519650
10  #Number of constraints in hypothesis 1
11  3
12  #Restriction matrix (R|r) for hypothesis 1
13  1 -1  0  0   0
14  0  1  0 -1   0
15  0  0 -1  1   0
16  #Number of constraints in hypothesis 2
17  3
18  #Restriction matrix (R|r) for hypothesis 2
19  1 -1  0  0   0
```

```
20  0  1 -1  0   0
21  0  0  1 -1   0
22  #Number of constraints in hypothesis 3
23  3
24  #Restriction matrix (R|r) for hypothesis 3
25  -1  1  0  0   0
26   0 -1  1  0   0
27   0  0 -1  1   0
```

Note that the structure of the input file cannot be changed. Both the lines containing annotation starting with # and the lines with numbers have to be presented. As can be seen on the second line, there are 4 target parameters in the regression model and 3 competing hypotheses with respect to those parameters. On the fourth line, the estimates of parameters obtained from `lavaan` are given, and on line six through nine the covariance matrix is given. The eleventh line shows that hypothesis 1 can be specified using 3 constraints. Next, there are three lines under the label `#Restriction matrix (R|r) for hypothesis 1`, each of which expresses a constraint in hypothesis 1. This will be elaborated in detail in the next paragraph. Because the second line shows that 3 hypotheses have to be evaluated, we need to specify two extra hypotheses (hypothesis 2 and hypothesis 3) for which the numbers of constraints and the restriction matrices can be placed in a similar fashion as for hypothesis 1.

As was shown in Section 3.2.2, an inequality constrained hypothesis H_i can be formulated by $\boldsymbol{R}_i\theta > \boldsymbol{r}_i$. Each constraint $\boldsymbol{R}_{ik}\boldsymbol{\theta} > r_{ik}$ for $k = 1, \ldots, K$ in the hypothesis can be written as $R_{ik1}\theta_1 + \ldots + R_{ikJ}\theta_J > r_{ik}$, where K and J are numbers of constraints and parameters in H_i, respectively. Note that every parameter should be moved to the left hand side of the inequality sign ">", and the constant should be moved to the right hand. In the restriction matrix (R|r), the constraint $\boldsymbol{R}_{ik}\boldsymbol{\theta} > r_{ik}$ can be expressed by the line

R_{ik1} R_{ik2} ... R_{ikJ} r_{ik}.

For example,

- $\theta_1 + \theta_2 + \theta_3 > 0$ corresponds to
  1 1 1 0
- $\theta_1 - 2\theta_2 + 3\theta_3 > 0.5$ corresponds to
  1 -2 3 0.5
- $\theta_1 - 2 > \theta_2 - \theta_3$ corresponds to
  1 -1 1 2
- $\theta_1 > \theta_2 > \theta_3$ corresponds to
  1 -1 0 0
  0 1 -1 0

- $\theta_1 - \theta_2 > \theta_3 - \theta_4 > \theta_5 - \theta_6$ corresponds to
  1 -1 -1 1 0 0 0
  0 0 1 -1 -1 1 0

Thus, below the label `#Restriction matrix (R|r) for hypothesis 1`, the three lines

```
1 -1  0  0   0
0  1  0 -1   0
0  0 -1  1   0
```

represent the hypothesis $\theta_1 > \theta_2 > \theta_4 > \theta_1$ in the regression model.

It should be noted that the equality, about equality, and range constrained hypotheses can not be evaluated using `BIG`. Therefore, the restriction matrix (R|r)

```
 1 -1   0
-1  1   0
```

is not allowed, because it implies an equality constrained hypothesis $\theta_1 = \theta_2$. The restriction matrix (R|r)

$$\begin{matrix} 1 & -1 & -d \\ -1 & 1 & d \end{matrix}$$

is not allowed, because it implies an about equality constrained hypothesis $|\theta_1 - \theta_2| < d$, where d represents the tolerable deviation. The restriction matrix (R|r)

```
 1  0   0
-1  1   0
 0 -1   1
```

is not allowed, because it implies a range constrained hypothesis $0 < \theta_1 < \theta_2 < 1$. The restriction matrix (R|r)

```
 1  1
-1  0
```

is not allowed, because it implies $\theta_1 > 1$ and $\theta_1 < 0$ which contradict each other. If the restriction matrix (R|r) contains any equality, about equality, range, or contradicting constraints, executing `BIG` will produce an error message:

```
WARNING: Hypothesis i contains equality, about equality, range, or
contradicting constraints!
```

Besides the input of inappropriate hypotheses, there are four possible ways of making errors in the Input.txt file. First, one may by accident delete the annotate line starting with `#`. This results in the error message:

```
WARNING: Miss an annotate line in Input.txt!
```

Second, the length of the estimates and the rank of the covariance matrix of parameters are not in line with the number of target parameters specified in the second line. This results in the error message:

```
WARNING: An error below "#Estimates of parameters" in Input.txt!
```

or

```
WARNING: An error below "#Covariance matrix of parameters" in
```

`Input.txt!`
Third, the number of lines below `#Restriction matrix (R|r)` is not in line with the number below `#Number of constraints in hypothesis i`. This results in the error message:
`WARNING: An error below #Restriction matrix (R|r) for hypothesis i in Input.txt!`
Fourth, when two or more hypotheses are under consideration, one may forget to specify the number of constraints and the restriction matrix (R|r) for every hypothesis. This results in the error message:
`Hypothesis i needs to be specified in Input.txt!`
If an unknown problem occurs when running `BIG`.exe, please send your Input.txt to `x.gu@uu.nl`.

3.B.2 Output file

Executing `BIG`.exe renders a text file Output.txt in the same folder. If there already exists an Output.txt, it will be overwritten by the new one. Output.txt not only displays Bayes factors and PMPs for inequality constrained hypotheses, but also the decomposed fits and complexities with the corresponding numbers of iterations in Gibbs sampler. The output file corresponding to Input.txt shown in the previous section is:

```
Result for hypothesis  1
     Fits                numbers of iterations
    0.6881                     4000
    0.4330                     4000
    0.7631                     4000
Complexities (prior 1)   numbers of iterations
    0.5015                     4000
    0.3379                     4000
    0.2514                     4000
Complexities (prior 2)   numbers of iterations
    0.4970                     4000
    0.1908                     4000
    0.2534                     4000
Total fit      Complexity (prior1)     Complexity (prior 2)
  0.2274             0.0426                  0.0240
BFiu (prior 1)    BFiu (prior 2)
  5.3367             9.4612

BFic (prior 1)    BFic (prior 2)
  6.6129             11.9512
```

```
Result for hypothesis  2
⋮

Result for hypothesis  3
⋮

Result of PMP for each hypothesis
PMP (prior1)      PMP (prior2)    for hypothesis  1
  0.8041             0.7840
PMP (prior1)      PMP (prior2)    for hypothesis  2
  0.1950             0.2152
PMP (prior1)      PMP (prior2)    for hypothesis  3
  0.0009             0.0009
```

The output file contains the Bayes factors and PMPs for each hypothesis under consideration. The interpretations of Bayes factors and PMPs are elaborated in Section 3.3 in this paper. As shown in Section 3.3, the Bayes factor can be computed by multiplying the decomposed fits divided by decomposed complexities. For the result of hypothesis 1, first of all three decomposed fits are displayed below the label `Fits`, and the corresponding numbers of iterations used for the computation of the fits are shown on the right side. Then the decomposed complexities under prior distribution (3.16) are presented below the label `Complexities (prior 1)`, which is followed by the complexities under prior distribution (3.17). The numbers of iterations used to obtain these complexities are placed in the corresponding line. Thereafter, the fit, and complexities under prior (3.16) and (3.17) for hypothesis 1 can be obtained by multiplying the decomposed fits and complexities, which are shown under the labels `Total fit`, `Complexity (prior1)`, and `Complexity (prior 2)`, respectively. Based on the fit and two complexities, `BIG` computes the Bayes factors under both prior distributions and displays them below `BFiu (prior 1)` and `BFiu (prior 2)` for H_i against H_u, and below `BFic (prior 1)` and `BFic (prior 2)` for H_i against H_{i_c}. We omit the results for hypothesis 2 and hypothesis 3, because they have the same form as hypothesis 1. Finally, the PMPs is printed, which can be obtained based on the results of the Bayes factors above. For each hypothesis, its PMPs under two prior distributions are written in the line below `PMP (prior1)` and `PMP (prior2)`.

Chapter 4

Error Probabilities in Default Bayesian Hypothesis Testing[1]

4.1 Introduction

We shall focus on the well-known t test of an effect in a normally distributed population with unknown variance, i.e., $x_i \sim N(\theta, \sigma^2)$, for $i = 1, \ldots, n$, where θ denotes the population effect and σ^2 denotes the population variance. We will test the null hypothesis, $H_0 : \theta = 0$, which assumes that the population effect equals zero against the alternative hypothesis, $H_1 : \theta \neq 0$, which assumes that the population effect is unequal to zero. In a Bayesian framework, we have to specify prior distributions of the free parameters under both hypotheses. These priors reflect which values are assumed to be most likely for the free parameters before observing the data. Therefore, a prior must be specified for the variance under H_0, denoted by $\pi_0(\sigma^2)$, and a joint prior must be specified for the effect and the variance under H_1, denoted by $\pi_1(\theta, \sigma^2)$. A Bayesian hypothesis test can then be formulated as

$$H_0 : \theta = 0, \pi_0(\sigma^2) \text{ versus } H_1 : \pi_1(\theta, \sigma^2). \tag{4.1}$$

Note that under H_0, the restriction $\theta = 0$ can also be viewed as a prior distribution with point mass at zero.

A natural way to perform a Bayesian hypothesis test is using the Bayes factor. The Bayes factor is defined as the ratio of the marginal likelihoods under H_0 and H_1,

[1]This chapter has been published as Gu, X., Hoijtink, H., & Mulder, J. (2015). Error probabilities in default Bayesian hypothesis testing. *Journal of Mathematical Psychology*. doi:10.1016/j.jmp.2015.09.001.
Author contributions: XG, HH, and JM designed the research. XG performed the data analyses and simulation study, and wrote the paper. JM and HH provided extensive feedback on constructing and writing the paper.

i.e.,

$$B_{01} = \frac{m_0(\mathbf{x})}{m_1(\mathbf{x})}. \tag{4.2}$$

The marginal likelihood, $m_t(\mathbf{x})$ for $t = 0, 1$, is the probability of observing the data $\mathbf{x}$ under H_t given the prior π_t. Thus, the Bayes factor B_{01} quantifies how much more likely the data were generated under the null hypothesis H_0 with prior π_0 in comparison to the alternative hypothesis H_1 with prior π_1. Therefore, the Bayes factor is typically interpreted as a relative measure of evidence in the data between two hypotheses. If B_{01} is greater than, equal to, or smaller than 1, this implies that there is more, equal, or less evidence for H_0 relative to H_1, respectively. For example, if $B_{01} = 10$ this implies that the data were 10 times more likely to come from H_0 than from H_1, which clearly implies evidence in favor of H_0 against H_1.

Although type I and type II error probabilities, i.e., the probability of incorrectly selecting H_1 while H_0 is true and the probability of incorrectly selecting H_0 while H_1 is true, respectively, are fundamental elements in classical hypothesis testing, classical error probabilities are often not of focal interest to Bayesians. One of the reasons is that we do not have to make a dichotomous decision when interpreting Bayes factors. For example, when observing $B_{01} = 10$, a researcher can judge for him or herself whether this is 'positive' or 'strong' support for H_0 against H_1. Thus, we do not need cut-off values as in classical hypothesis testing where we decide to reject or not reject H_0 against H_1 depending on whether the p value is smaller or larger than a prespecified significance level α. Suggestions have been made how to qualify Bayes factors (Jeffreys, 1961; Kass & Raftery, 1995), e.g., a Bayes factor B_{01} between 3 and 20 should be interpreted as 'positive' evidence for H_0 against H_1. These suggestions however should not be used as strict rules but more as rough guidelines when interpreting Bayes factors.

Despite the fact that we do not need to make a dichotomous decision in Bayesian hypothesis testing, error probabilities do play a central role in hypothesis testing using the Bayes factor. We shall make this more explicit using the following calibration scheme. First we generate a hypothesis based on equal prior probabilities, i.e., $P(H_0) = P(H_1) = .5$. Second, parameters are generated based on the prior density π_t under the hypothesis H_t that is generated in the first step, for $t = 0$ or 1. Third, data is generated with sample size n according to the normal distribution $N(\theta, \sigma^2)$ where θ and σ^2 are taken from the second step. The Bayes factor B_{01} is then computed for these data. If we then select H_0 if $B_{01} > 1$ and select H_1 if $B_{01} < 1$, we would minimize the sum of the type I and the type II error probabilities on average (e.g., Berger, 1985). Thus, in addition to the intuitive interpretation of the Bayes factor as the relative evidence between two hypotheses, testing hypotheses using the Bayes factor also satisfies an important frequentist argument.

Although this decision rule minimizes the average sum of the error probabilities, the separate error probabilities are not minimized. Therefore, the unknown type I

error probability may be very different from the unknown type II error probability, i.e., $p_0 = P(B_{01} < 1|H_0) \neq P(B_{01} > 1|H_1) = p_1$. If this is the case, the Bayes factor has a tendency to either select H_0 or H_1. We shall refer to this as asymmetry in information.

Garcia-Donato and Chen (2005) proposed a correction to the decision rule to ensure that the error probabilities are equal. They proposed to select H_0 if $B_{01} > c$ and select H_1 if $B_{01} < c$, where the value $c > 0$ is calibrated such that $P(B_{01} < c|H_0) = P(B_{01} > c|H_1)$. Despite the intuitive appeal of this decision rule from a frequentist perspective, there is no Bayesian justification for this method. The reason is that the asymmetry in information in the Bayes factor comes naturally from the chosen priors under H_0 and H_1. It may be that it is easier to generate data under the prior under the null hypothesis, π_0, that is consistent with data that is generated under H_1 than to obtain data that is generated under the prior under the alternative hypothesis, π_1, that is consistent with data generated under H_0. If this would be the case, the Bayes factor does exactly what it is supposed to do: it would select H_1 more often if H_0 would be true than it would select H_0 if H_1 would be true. Consequently, the type I error probability would be larger than the type II error probability. If the priors under H_0 and H_1 are carefully chosen based on the prior beliefs of the researcher, asymmetry in information is a natural property of the Bayes factor. Therefore it seems more reasonable to select either H_0 or H_1 depending on whether B_{01} is larger or smaller than 1, respectively, instead of comparing the observed Bayes factor with the observed c.

In this paper we focus on Bayesian hypothesis testing using so-called default Bayes factors. We shall use the term default Bayes factor when a prior is used that is not directly related to the substantive expectations of the researcher. Default priors typically contain little information and have distributional forms that ensure that the Bayes factor is relatively easy to compute. A potential issue with default Bayes factors lies in its interpretation. The potential issue is that the outcome of a default Bayes factor is a default quantification of the relative evidence between two hypotheses. This default outcome may be very different than the subjective relative evidence in the data between the hypotheses if priors were used that are based on the researcher's substantive beliefs. For example a popular default prior is to set a Cauchy prior for the standardized effect under H_1 centered around 0 with scale 1 (Rouder et al., 2009), and set noninformative improper Jeffreys priors for the variances under both hypotheses. This prior has good theoretical properties. For example, it avoids the information paradox, see Liang, Paulo, Molina, Clyde, and Berger (2008). This Cauchy prior however implies that we expect that there is 50% chance to find an absolute effect that is larger than 1 (i.e., an effect that is larger than 1 or smaller than -1) before observing the data. In psychological research however we hardly ever observe absolute effects larger than 1, and therefore, it is not realistic that the effect follows this Cauchy distribution if H_1 would be true. Consequently, the relative evidence as quantified by the default Bayes factor based on this Cauchy prior may have been very different

from the Bayes factor that would have been obtained when the researcher would have carefully formulated a prior based on external substantive knowledge.

In this paper we investigate the error probabilities of commonly used default priors in typical situations in psychological research where the effect is either zero, small, medium, or large (corresponding to standardized effects of 0, .2, .5, or .8 according to Cohen, 1992) while considering different sample sizes of $n = 20$, 50, and 100. Note that error probabilities for larger samples are not very interesting because as the sample size grows to infinity the error probabilities go to zero. In the case of limited data, which is typical in psychological research, understanding the (classical) error probabilities is useful because of the following three reasons.

First, default Bayes factors are based on default priors which typically do not reflect the prior beliefs of a researcher. For this reason it is useful to know whether a default Bayes factor has a tendency to either select H_0 or H_1 in standard situations encountered in psychological research because there is no reason to either prefer H_0 or H_1 more than the other from a subjective point of view because the priors are not based on subjective prior beliefs.

Second, as was mentioned above Bayes factors minimize the sum of the error probabilities when generating data under the respective models and priors. Default Bayes factors are typically not based on proper priors from which we can sample. For example, the priors of the nuisance parameters can be improper (such as in the Cauchy prior approach) or the priors are based on the observed data (such as in the intrinsic Bayes factor (Berger & Pericchi, 1996) or the fractional Bayes factor (O'Hagan, 1995)). Therefore we do not know under which conditions (priors) the sum of the error probabilities is minimized when using default Bayes factors.

Third, from the error probabilities we will learn which of the two models (i.e., the null or alternative) is best in predicting data coming from the other model. Thus we will find out whether (i) H_1 is better in predicting data that come from H_0 or (ii) whether H_0 is better in predicting data that come from H_1. Because H_0 is nested in H_1 one might expect that scenario (i) is more likely than scenario (ii). On the other hand, Bayes factors automatically correct for model complexity and therefore it is not automatically true that a Bayes factor has a tendency to prefer the larger unrestricted model under H_1. Furthermore, we will look at typical scenario's in psychological research where the standardized effect under H_1 is most likely between 0 and 1 (e.g., Cohen, 1992). For example a medium effect of .5 may be "closer" to H_0 than to H_1 when using a Cauchy prior for the standardized effect under H_1 with a scale of 1. This suggests that the default Bayes factor based on this Cauchy prior has a tendency to actually prefer H_0 when generating medium effects under H_1.

As will be shown certain default Bayes factors may result in very different type I and type II error probabilities. If a researcher finds this undesirable there are two possible solutions. The first solution is to construct a prior that corresponds to one's subjective beliefs. The second solution, which may be useful in the case of limited prior knowledge, is to use a default Bayes factor that is close to being symmetric in

information. At the end of the paper a method is discussed how the default prior can be tuned so that we obtain a Bayes factor that results in almost equal error probabilities in certain scenario's. Note that combining Bayesian and frequentist properties is not new (e.g., Good, 1992; Berger, Boukai, & Wang, 1997). So far, however, we have not seen that the two approaches are combined in this manner. Morey, Wagenmakers, and Rouder (in press) criticized the way of choosing default priors based on frequentist error probabilities in two aspects. First, it may render inconsistent Bayes factors. As will be shown in Section 4.6 of this paper, however, the tuned Bayes factor based on our method are consistent and are not very different from typical default Bayes factors. Second, it renders Bayes factors that behave like a classical test statistic (we will elaborate this in Section 4.5). We agree that the Bayesian test with tuned Bayes factors will render a similar decision as the frequentist test where type I and type II errors are set to be equal. Nevertheless, the proposed method offers a way to obtain equal error probabilities and a reasonable default quantification of the relative support for two hypotheses (this will be elaborated in Section 4.6 and Section 4.7).

Three different default Bayes factors will be considered for testing an effect in a normal population: a Bayes factor based on Zellner's g prior (Zellner, 1986), a Bayes factor based on an inverse gamma mixture of g priors (e.g., Liang et al., 2008), which implies a Cauchy distribution of the standardized effect under H_1, and an adjustment of O'Hagan's fractional Bayes factor (O'Hagan, 1995), which was recently proposed by Mulder (2014b). Each of these Bayes factors contains a tuning parameter which directly influences the prior variance of the effect θ under H_1. The prior variance of θ plays a key role in the Bayes factor which can be seen for example from the Jeffreys-Lindley paradox (Jeffreys, 1961; Lindley, 1957). Symmetry in information will be investigated for default choices of the tuning parameters.

This paper is organized as follows. Section 4.2 presents an empirical example which will be used to illustrated our method. Thereafter, three different Bayes factors are introduced in Section 4.3. In Section 4.4 the type I and type II error probabilities are investigated of these Bayes factors in a default setting. Section 4.5 shows how we can tune these Bayes factors such that they are symmetric in information. Section 4.6 illustrates the consistency of the tuned Bayes factors. Then a simulation study is conducted to investigate the error probabilities based on the tuned Bayes factors in Section 4.7. Section 4.8 revisits the empirical example based on the tuned default Bayes factors. We end this paper with a discussion.

4.2 Empirical example

We reanalyze the t test example used in Howell (2012, p. 196). An experiment is conducted to assess whether therapeutic touch (a widely used nursing practice) practitioners are able to identify which of their hands is below the experimenter's

under blinded condition. The experiment involved 28 testing sessions of 10 trials. For chance performance we expect an average of 5 correct trials out of 10. The difference between the observed score from 0 to 10 and chance score 5 is assumed to be normally distributed and denoted by $x_i \sim N(\theta, \sigma^2)$, where $i = 1, ..., n$ and $n = 28$. To investigate whether the participants made correct decisions by chance, we apply the t test to the following hypotheses: $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. If H_0 is selected, this implies that there is no difference between the observed score and chance score. If H_1 is selected, we would conclude there is a difference. From the report of Howell (2012, p. 196), the sample mean equals 0.607, the sample standard deviation equals 1.663, and the standardized effect equals 0.365. Throughout this paper we will use this as an illustrative running example.

4.3 Bayes factor

The Bayes factor for comparing a null hypothesis H_0 against an alternative hypothesis H_1 is defined as the ratio of their marginal likelihoods (Kass & Raftery, 1995):

$$BF_{01} = \frac{m_0(\mathbf{x})}{m_1(\mathbf{x})} = \frac{\int f(\mathbf{x}|\theta = 0, \sigma^2)\pi_0(\sigma^2)d\sigma^2}{\iint f(\mathbf{x}|\theta, \sigma^2)\pi_1(\theta, \sigma^2)d\theta d\sigma^2}, \tag{4.3}$$

where $\mathbf{x} = (x_1, \cdots, x_n)'$ and the likelihood of the data is given by

$$f(\mathbf{x}|\theta, \sigma^2) = (2\pi)^{-n/2}\sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}[ns^2 + n(\bar{x} - \theta)^2]\right\}, \tag{4.4}$$

with sample mean $\bar{x}$ and sums of squares $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. The integrals in (4.3) can be computed analytically or numerically depending on the distributional form of the priors π_0 and π_1. The outcome then quantifies the relative evidence between the two hypotheses.

Below three different default Bayes factors are described, which have been proposed in the literature where the prior of θ under H_1 is centered at zero and a single tuning parameter is chosen to specify the prior variance. We consider the Bayes factor based on Zellner's (1986) g prior with tuning parameter g, the Bayes factor based on a mixture of g priors (Liang et al., 2008) with tuning parameter r, and the prior adjusted default Bayes factor (Mulder, 2014b) with tuning parameter b.

4.3.1 The Bayes factor based on Zellner's g prior

In the context of regression models Zellner's (1986) g prior is widely used in Bayesian hypothesis testing and model selection. It specifies a normally conditional prior distribution for θ with a mean of 0 and a variance that contains a scalar parameter g. For the one sample t-test (4.1), Zellner's g prior is defined as $\pi_1(\theta, \sigma^2|g) =$

$\pi_1(\theta|\sigma^2, g)\pi_1(\sigma^2)$, with

$$\pi_1(\theta|\sigma^2, g) = N(0, g\sigma^2/n), \quad \pi_1(\sigma^2) \propto \sigma^{-2}. \tag{4.5}$$

Under H_0, the prior equals $\pi_0(\sigma^2) \propto \sigma^{-2}$.

An appealing aspect of the g prior is that the resulting Bayes factor has an analytic expression. By substituting (4.5) into (4.3), the Bayes factor based on Zellner's g prior (ZBF) can be obtained as

$$ZBF_{01}(\mathbf{x}, g) \quad = \quad (1+g)^{-\frac{n-1}{2}}(1+\tfrac{g}{1+(\bar{x}/s)^2})^{\frac{n}{2}}. \tag{4.6}$$

Note that $\bar{x}/s$ reflects the observed standardized effect (Cohen, 1992). One can tune the amount of prior information in a relatively simple way via g, where a large (small) value for g implies a very vague (informative) prior with a large (small) variance.

A popular choice is to set $g = n$ which results in the so-called 'unit information prior' (Kass & Wasserman, 1995), where the amount of prior information corresponds to the information of one observation. Other choices for g that have been recommended in the literature are generally smaller than n. For an overview, see Liang et al. (2008), for example. Throughout this paper, we will let g vary in the interval $(0, n]$ so that the amount of prior information is never less than the amount of information in one observation. Furthermore, we shall focus on two extreme choices: $g = n$ ('unit information') and $g = 1$. The latter choice implies that the amount of information in the prior is equal to amount of information in the data.

4.3.2 The Bayes factor based on a mixture of g priors

Despite the popularity and usefulness of the g prior, it does have an undesirable property, i.e., it is information inconsistent. This implies that when the standardized effect $\bar{x}/s$ goes to infinity the Bayes factor ZBF_{01} does not go to zero, which would be expected in this extreme situation, but instead it converges to a constant. A way to avoid the problem is by putting a probability distribution on g. The resulting prior on θ under H_1 is then a mixture of g priors (Liang et al., 2008). The choice of an inverse-Gamma$(1/2, r^2/2)$ on g is becoming increasingly popular (e.g., Rouder et al., 2009). Thus the prior distribution is given by

$$\pi_1(\theta|\sigma^2, g) = N(0, g\sigma^2/n), \quad \pi_1(g|r) = \text{inv-Gamma}(1/2, r^2/2), \quad \pi_1(\sigma^2) \propto \sigma^{-2}, \tag{4.7}$$

and again we set $\pi_0(\sigma^2) \propto \sigma^{-2}$ under H_0. The Bayes factor can then be obtained by integrating the ZBF in (4.6) over the prior of g, i.e.,

$$MBF_{10}(\mathbf{x}, r) \quad = \quad \int ZBF_{10}(g)\pi_1(g|r)dg. \tag{4.8}$$

Note that MBF_{01} can be obtained as a reciprocal of MBF_{10}. Although an analytical expression for equation (4.8) does not exist, we can obtain the outcome using either

a Laplace approximation (see Liang et al., 2008) or an approximated Savage-Dickey method (see Morey et al., 2011).

In this prior, we can obtain the conditional prior for θ given σ^2 by integrating out g in the joint prior, i.e., $\pi_1(\theta|\sigma^2, r) = \int \pi_1(\theta|\sigma^2, g)\pi_1(g|r)dg$. This implies that the standardized effect θ/σ has a Cauchy prior centered at 0 with scale parameter $r/\sqrt{n}$. Throughout this paper we will let r vary in the interval $(0, \sqrt{n}]$ because $r = \sqrt{n}$ also corresponds to 'unit information', similar as by setting $g = n$ in the ZBF. For the MBF, we consider two default choices: $r = \sqrt{n}$ (Rouder et al., 2009) and $r = \sqrt{n}/2$ (http://bayesfactorpcl.r-forge.r-project.org/). Note that both choices do not reflect very realistic prior distributions for substantive researchers because the prior probability of observing an absolute effect larger than 1 would be equal to 50% and 29.5%, respectively, when setting $r = \sqrt{n}$ and $r = \sqrt{n}/2$. These probabilities are substantially larger than what we would expect in substantive research.

4.3.3 Prior adjusted default Bayes factors

The prior adjusted default Bayes factor was introduced by Mulder (2014b) as a modification of O'Hagan's (1995) fractional Bayes factor (FBF). In the FBF a fraction b is taken from the likelihood of the data, i.e., $f(\mathbf{x}|\theta, \sigma^2)^b$, for default prior specification, resulting in a marginal updated prior for θ under H_1 of $\pi_1(\theta|\mathbf{x}^b) = t(\bar{x}, s^2/(nb-1), nb-1)$, i.e., a Student t density with mean $\bar{x}$, scale parameter $s^2/(nb-1)$, and degrees of freedom $nb-1$. Thus, if b is large, the prior variance will be small, and vice versa. However it has been advocated that the prior under H_1 should be symmetrical around 0 and nonincreasing for $|\theta|$ because 0 is the focal point of our hypothesis test (e.g., Jeffreys, 1961; Berger & Delampady, 1987). For this reason, Mulder (2014b) proposed an adjustment such that the underlying prior has a $t(0, s^2/(nb-1), nb-1)$ distribution, which is centered at 0 and nonincreasing in $|\theta|$. The resulting prior adjusted default Bayes factor (DBF) for the Bayesian t test can then be expressed as

$$DBF_{01}(\mathbf{x}, b) = \frac{\Gamma(n/2)}{\Gamma((n-1)/2)} \Big/ \frac{\Gamma(nb/2)}{\Gamma((nb-1)/2)} (1 + (\bar{x}/s)^2)^{-n/2}. \tag{4.9}$$

Throughout this paper we will let b vary between $[\frac{2}{n}, 1]$. Note that the minimal choice of $\frac{2}{n}$ corresponds to the amount of information of two observations, which is the minimal number of observations needed to obtain a proper updated prior. As can be seen, the value $b = \frac{1}{n}$ is not allowed because it implies a density with zero degrees of freedom in the updated prior, which is improper. For the DBF, we consider two default choices: $b = 2/n$ and $b = 1/\sqrt{n}$, where the latter choice was proposed by O'Hagan (1995), with the goal of reducing the sensitivity of the Bayes factor to the prior distribution.

4.3.4 Application of default choices to the empirical example

Different choices of g, r, and b may result in Bayes factors that do not have a consistent preference towards either H_0 or H_1. Consider the empirical example from Section 4.2 in which $\bar{x} = 0.607, s = 1.663$, and a standardized effect of 0.365 for $n = 28$. Using ZBF with $g = 1$ (Default 1) renders $ZBF_{01} = 0.606$ which is in favor of H_1, and with $g = n$ (Default 2) renders $ZBF_{01} = 1$ which does not prefer any hypothesis. In addition, the choices of $r = \sqrt{n}/2$ (Default 1) and $r = \sqrt{n}$ (Default 2) render $MBF_{01} = 0.83$ and $MBF_{01} = 1.26$, respectively, and the choices of $b = 2/n$ (Default 1) and $b = 1/\sqrt{n}$ (Default 2) render $DBF_{01} = 1.12$ and $DBF_{01} = 0.46$, respectively.

Based on these outcomes we cannot determine which hypothesis is preferred by the data because different default Bayes factors differ in their preference towards either H_0 or H_1 based on the observed data. In these default Bayes factors, the underlying default priors do not directly reflect substantive believes about the model parameters, and therefore, we cannot choose either one of these outcomes to get an idea in which direction the data are most likely pointing. In this situation it would be insightful to know whether these default Bayes factors have a tendency to prefer either H_0 or H_1. For example if the default Bayes factor MBF with $r = \sqrt{n}$ has a tendency to prefer H_0 for example (which implies that the type I error probability is smaller than the type II error probability), the observed (small) preference towards H_0 ($MBF_{01} = 1.26$) might very well be caused by the asymmetry of the default Bayes factor. For this reason it would be interesting to investigate how large the type I and type II error probabilities are for these default Bayes factors. This is discussed in the following section.

4.4 Error probabilities of default Bayes factors

By means of the empirical example above it was illustrated that the choices of g, r, and b play an important role in hypothesis testing. In this section we investigate whether Bayes factors based on certain default choices for g, r, and b have a tendency to either select H_0 or H_1 based on their type I and type II error probabilities. This will be done by calibrating Bayes factors under H_0 and H_1.

4.4.1 Sampling distributions of the Bayes factor under H_0 and H_1

The three Bayes factors can be used as a test statistic where we select H_0 when $BF_{01} > 1$ and select H_1 when $BF_{01} < 1$. Because of the clear relation with classical hypothesis testing, it is interesting to investigate the sampling distribution of each Bayes factor for a given population. This sampling distribution can be obtained in three steps: (i) sample K data sets $\mathbf{x}^{(k)}$ with fixed sample size n, for $k = 1, \ldots, K$ either from H_0 or from H_1; (ii) compute the sample mean $\bar{x}^{(k)}$, the sample standard deviation $s^{(k)}$, and the resulting Bayes factor $BF_{01}^{(k)}$ via (4.6), (4.8), or (4.9) for all

K data sets; and (iii) plot the distribution of the Bayes factor based on the sampled outcomes.

Let us assume $\sigma^2 = 1$ under both H_0 and H_1, and under H_1 the effect is equal to $\theta = .4$. Thus, under H_0 we generate data according to $x_i \sim N(0, 1)$, and under H_1, we generate data according to $x_i \sim N(.4, 1)$. Note that we do not sample data by sampling effects from the default priors under H_0 and H_1. The reason is that these priors do not represent substantive beliefs, and that we cannot sample the variance σ^2 because noninformative improper priors are used.

The sampling distributions are displayed in Figure 4.1 under H_0 (solid line) and under H_1 (dashed line) for different choices for g: Panel (a) displays the sampling distribution of the logarithm of the ZBF for H_0 against H_1 for $g = 1$ and panel (b) displays the sampling distribution for $g = n = 50$. In addition, the sampling distribution in panel (c) is obtained based on $g = 9.4$, which will be discussed in Section 4.5.

Figure 4.1 shows that the distribution of the ZBF highly depends on g. More specifically, when $g = 1$, the ZBF is distributed around relatively small values, i.e., the means of the sampling distribution of ZBF_{01} under H_0 and H_1 are 1.15 and 0.33, respectively. When $g = n$, the ZBF is distributed around relatively large values, i.e, the means of the sampling distribution under H_0 and H_1 are 5.07 and 0.72, respectively. This suggests a preference towards H_0 when $g = n$, and a preference towards H_1 in the case of $g = 1$. When $g = 9.4$ there is not clear preference towards either one of the two hypotheses. This will be made more explicit in Section 4.4.3.

4.4.2 Error probabilities for default choices of g, r, and b

The preference towards either H_0 or H_1 can be observed more precisely by computing the error probabilities. We shall define the error probability, p_0, of H_0 as the probability that $BF_{01} < 1$ given that H_0 is true, i.e., $p_0 = P(BF_{01} < 1|H_0)$, which corresponds to the type I error in the classical sense, and the error probability, p_1, of H_1 as the probability that $BF_{01} > 1$ given that H_1 is true, i.e., $p_1 = P(BF_{01} > 1|H_1)$, the type II error probability. Regarding the three Bayes factors discussed earlier, we let p_t^Z, p_t^M, and p_t^D, for $t = 0$ or 1, denote the corresponding error probabilities in ZBF, MBF, and DBF, respectively.

The error probabilities can be obtained from the sampling distribution of the Bayes factor as the proportion of samples generated under a hypothesis resulting in preferring the other hypothesis. Table 4.1 provides an overview of the error probabilities for the ZBF, MBF, and DBF based on their default choices for g, r, and b, respectively. As can be seen in Table 4.1 (rows p_0, p_1, and $p_0 + p_1$), default choices for g, r, and b result in unequal error probabilities under H_0 and H_1. For example, it can be seen that $g = 1$ renders a small $p_1^Z = 0.052$ but a large $p_0^Z = 0.244$, whereas $g = n$ renders a large $p_1^Z = 0.214$ but a small $p_0^Z = 0.048$. These values can also be found in Figure 4.1 where the dark grey area represents p_0^Z and the light grey area represents

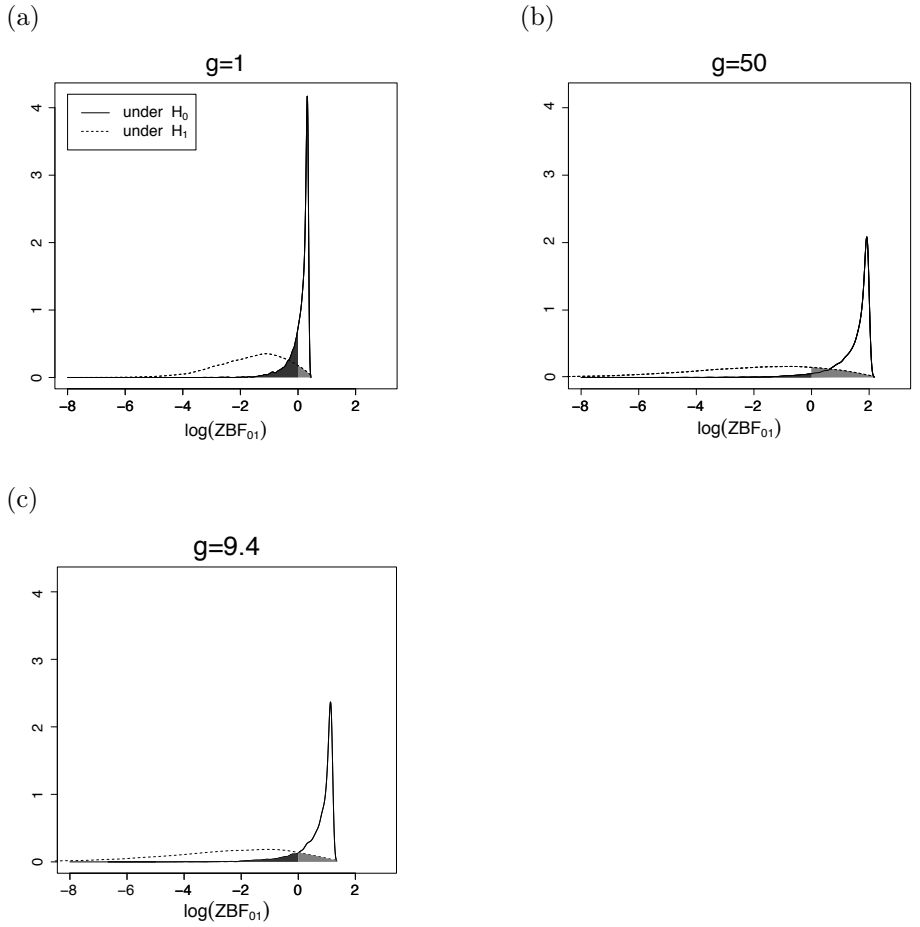

Figure 4.1: Sampling distributions of the logarithm of the ZBF_{01} based on $g = 1$ (panel a), $g = 50$ (panel b) and $g = 9.4$ (panel c) where $\theta = .4$ is assumed under H_1. The dark grey area represents the error probability under H_0 and the light grey area represents the error probability under H_1.

Table 4.1: Error probabilities of the Bayes factors using different tuning parameters ($n = 50$, $\theta = 0$ and $\sigma^2 = 1$ under H_0, and $\theta = 0.4$ and $\sigma^2 = 1$ under H_1) with error probabilities defined as $p_0 = P(BF_{01} < 1|H_0)$ and $p_1 = P(BF_{01} > 1|H_1)$, and $\tilde{p}_0 = P(BF_{01} < \frac{1}{3}|H_0)$ and $\tilde{p}_1 = P(BF_{01} > 3|H_1)$.

	ZBF		MBF		DBF	
	$g = 1$	$g = n$	$r = \sqrt{n}/2$	$r = \sqrt{n}$	$b = \frac{2}{n}$	$b = \frac{1}{\sqrt{n}}$
p_0	0.244	0.048	0.079	0.035	0.043	0.151
p_1	0.052	0.214	0.213	0.315	0.238	0.090
$p_0 + p_1$	0.296	0.262	0.292	0.350	0.281	0.241
$\tilde{p}_0$	0.016	0.014	0.017	0.010	0.011	0.041
$\tilde{p}_1$	0.000	0.068	0.085	0.153	0.087	0.000
$\tilde{p}_0 + \tilde{p}_1$	0.016	0.082	0.102	0.163	0.098	0.041

p_1^Z. These error probabilities clearly show that the ZBF has a tendency to select H_0 if $g = n$, and it has a tendency to select H_1 if $g = 1$. A similar pattern can be seen for both default choices of b in the DBF. Furthermore, both default choices of r in the MBF result in considerably larger type II errors than type I errors. In this specific setting, the MBF with $r = \sqrt{n}$ seems to be most asymmetric in information with a type II error probability that is 9 times larger than the type I error probability. In sum, all default choices result in Bayes factors that are asymmetric in information in this situation which would be typical in psychological research.

One may object to the choice of "selecting" a hypothesis based on a Bayes factor that is larger or smaller than 1 because Bayes factors close to 1, e.g., $B_{01} = 1.5$, do not imply any clear evidence towards one specific hypothesis anyway. For this reason we also looked at the error probabilities when the wrong hypothesis receives three times more evidence from the data than the true hypothesis, i.e., $\tilde{p}_0 = P(BF_{01} < \frac{1}{3}|H_0)$ and $\tilde{p}_1 = P(BF_{01} > 3|H_1)$. The interval $(\frac{1}{3}, 3)$ can then been as a "no decision" region (similar as in Berger, Brown, and Wolpert (1994)). The results can be found in Table 4.1 (rows $\tilde{p}_0$, $\tilde{p}_1$, and $\tilde{p}_0 + \tilde{p}_1$). As can be seen the direction of the asymmetry is the same as when selecting a hypothesis based on a Bayes factor larger or smaller than 1. This suggests that the direction (and the severity) of the asymmetry in information of default Bayes factors does not change a lot when using other cut-off values than 1. For this reason we shall continue to use the cut-off value of 1 in the definition of error probabilities throughout this paper.

4.4.3 Error probabilities for other choices of g, r, and b

In the previous section we showed that default choices for the tuning parameters result in default Bayes factors with unequal error probabilities. In this section we investigate the error probabilities for other choices of the tuning parameters than

the default choices. We will discuss this in detail for all values of b in the interval $(\frac{1}{n}, 1]$ in the DBF for different effects under H_1 and different sample sizes n. The value for b that results in equal error probabilities, i.e., $p_0^D = p_1^D$, will be denoted by b^*. Similarly, the tuning parameters that results in equal error probabilities for the ZBF and MBF will be denoted by g^* and r^*, respectively, such that $p_0^Z = p_1^Z$ and $p_0^M = p_1^M$.

Figure 4.2 displays the plots of the error probabilities p_0^D (solid line) and p_1^D (dashed line) for $b \in (\frac{1}{n}, 1]$, for $\sigma^2 = 1$, a fixed sample size of $n = 20$, 50, or 100, and an effect size under H_1 of $\theta = 0.2$, 0.5, or 0.8, which typically corresponds to a small, medium or large effect size, respectively (Cohen, 1992). Because all three default Bayes factors in (4.6), (4.8), and (4.9) only depend on the standardized effect in the data, $\frac{\bar{x}}{s}$, the resulting error probabilities also only depend on the standardized effect in the population, the tuning parameter, and the sample size. This implies for example that $\theta = 1$ and $\sigma = 2$ under H_1 would result in the same plots as Figure 4.2 (b), (e), and (h) for the respective sample sizes based on $\theta = .5$ and $\sigma = 1$ under H_1. In the remaining part of this paper, we will only consider the case that $\sigma = 1$ when generating data which implies that θ corresponds to a standardized effect. Also note that similar types of figures can also be obtained as in Figure 4.2 for the ZBF and the MBF by letting g and r vary.

The results in Figure 4.2 can be summarized as follows. First, when b increases, p_0^D increases and p_1^D decreases. This suggests an intersection point exists for $b = b^*$ for a given standardized effect size and sample size such that $p_0^D = p_1^D$. We discuss the existence of b^*, g^*, and r^* in Appendix 4.A. Second, for a given standardized effect size a larger sample size results in the reduction of b^* and the corresponding error probabilities. Note that the default choices $b = \frac{2}{n}$ and $\frac{1}{\sqrt{n}}$ also decrease when increasing the sample size. The figure suggests b^* approaches zero and the resulting error probabilities $p_0^D = p_1^D$ also go to zero as the sample size increases. Third, for a given sample size a larger standardized effect size results in a smaller b^* and smaller error probabilities. Fourth, the error probability under H_0 is independent of the actual effect because it is calibrated under H_0. Note that the error probability under H_0 only slightly depends on the sample size for a fixed b.

Hence, the tuning parameters b^*, g^*, and r^* that result in equal error probabilities depend on the sample size n and the standardized effect size θ under H_1. Therefore, we will denote them as functions of θ and n, i.e., $b^*(\theta, n)$, $g^*(\theta, n)$, and $r^*(\theta, n)$. Figure 4.3 displays these functions by varying θ from 0 to 1 for a fixed sample size of $n = 20$, 50, or 100. As can be seen in Figure 4.3 (a), the plot indicates that $b^*(\theta, n)$ is a decreasing function of the standardized effect size and sample size. Furthermore, Figure 4.3 (b) and (c) display $g^*(\theta, n)$ and $r^*(\theta, n)$, respectively. As can be seen, $g^*(\theta, n)$ and $r^*(\theta, n)$ are increasing functions of the standardized effect size θ and the sample size n. Note that the allowed regions for b, g, and r, i.e., $[\frac{2}{n}, 1]$, $(0, n]$ and $(0, \sqrt{n}]$, respectively, are also taken into account in these plots. Therefore, the minimal

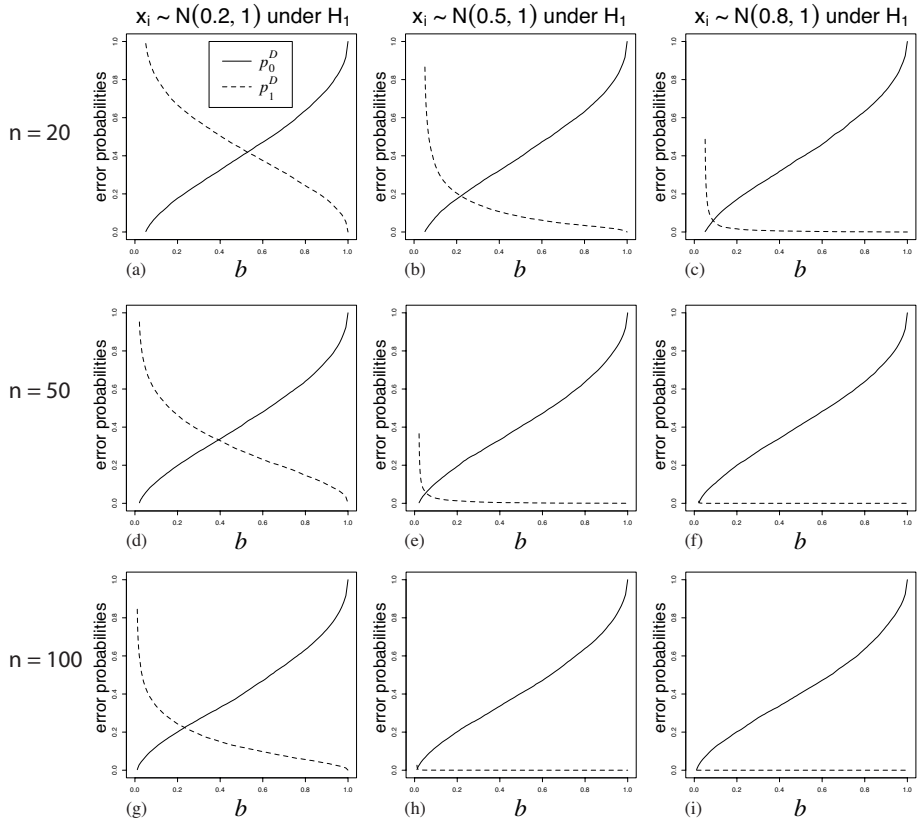

Figure 4.2: Error probabilities p_0^D (solid line) and p_1^D (dashed line) when $b \in (\frac{1}{n}, 1]$, based on an effect size under H_1 of $\theta = .2$, $.5$, or $.8$ and $\sigma^2 = 1$ (displayed in the columns) and a sample size of $n = 20$, 50, or 100 (displayed in the rows). In the intersection b^* the error probabilities are equal.

$b^* = \frac{2}{n}$, and the maximal $g^* = n$ and $r^* = \sqrt{n}$ are obtained if the standardized effect is assumed to be larger than a certain threshold value (e.g., 0.73, 0.52, and 0.39 for $n = 20$, 50, and 100, respectively, in the case of g^*).

As an example, Figure 4.1 (c) shows the sampling distributions of the tuned ZBF with $g = 9.4$ and $n = 50$ under H_0 and under a fixed standardized effect of $\theta = .4$ under H_1. As can be seen, the error probabilities are equal which was not the case for the default choices as can be seen in Figure 4.1 (a) and (b). This can also be seen in Figure 4.3 (b) that under $n = 50$ and $\theta = .4$ the tuned parameter $g^* = 9.4$.

The plots in Figure 4.3 show an interesting characteristic of the tuning parameter that results in equal error probabilities. As can be seen in Figure 4.3 (b), for example, a large (small) standardized effect under H_1 results in a g^* that is also large (small), which, in turn, implies a large (small) prior variance in (4.6). This relationship between the standardized effect θ and g^* corresponds exactly with how g would be chosen based on the expected standardized effect under H_1 in prior specification based on substantive expectations: If we would expect a large (small) standardized effect under H_1, we want the prior variance for θ under H_1 to be large (small), which can be achieved by setting a large (small) value for g. This relationship also holds for r^* and b^*. Thus, it can be concluded that well-specified subjective priors result in Bayes factors with good frequency properties (in the sense that the type I error probability is close to the type II error probability) for the range of effects that are likely under the specified prior.

4.4.4 Final remarks about default choices of the tuning parameters

In this section we observed that default choices of the tuning parameters may result in default Bayes factors that are highly asymmetry in information when $n = 50$ and the true standardized effect is of medium size under H_1. The MBF with $r = \sqrt{n}$ (Default 2) which resulted in the largest asymmetry with a type I error probability that is 9 times smaller than the type II error probability. This asymmetry can be explained by the fact that this MBF uses a Cauchy prior for the standardized effect with a scale parameter of 1. Under this prior absolute standardized effects larger than 1 are equally likely a priori as absolute standardized effects smaller than 1 (Morey et al., in press). Consequently, medium effects can on average be better predicted by H_0 than that zero effects can be predicted by H_1. Furthermore, this MBF resulted in the largest sum of the error probabilities .350 (which is an important property in Bayesian hypothesis testing). For this reason, this MBF (with tuning parameter $r = \sqrt{n}$) is not the preferred choice for default Bayesian hypothesis testing based on the error probabilities. The DBF (with Default 2) resulted in the least asymmetric results and the smallest sum of the error probabilities of .241 in this specific scenario. To get a better idea about the error probabilities of the default Bayes factors in other scenario's a more thorough simulation study will be conducted.

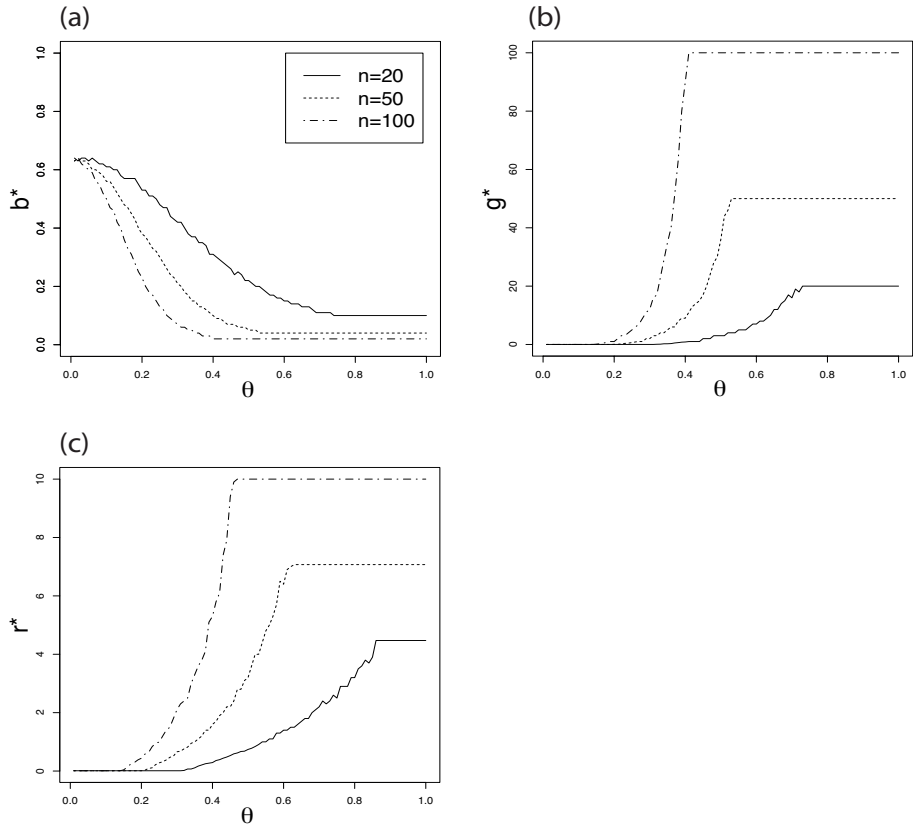

Figure 4.3: Examples of b^*, g^*, and r^* that result in equal error probabilities as a function of the expected standardized effect under H_1 and sample size n.

Before discussing this simulation study we present a method on how to set the tuning parameter so that the default Bayes factor is approximately symmetric in information in a specific scenario. As was noted by Morey et al. (in press) two issues may arise with this approach. First, the resulting default Bayes factor may not be consistent in the sense that the evidence towards a true alternative hypothesis will not go to infinity as the sample size goes to infinity. In the following two sections we will show how default Bayes factors can be obtained that are approximately symmetric in information and also consistent. The second potential issue is that the resulting default Bayes factor that is approximately symmetric in information behaves essentially as a classical test statistic. We come back to this in the following section and in the discussion.

4.5 A new default choice for the tuning parameters

In the previous section we observed that the tuning parameters can be chosen such that the resulting default Bayes factors are symmetric in information for a specific effect under H_1. We can use this result to tune the default Bayes factors such that it is symmetric in information for a specific effect. The problem is however that the effect under H_1 is unknown. Because of this uncertainty we suggest to specify a distribution of effects for which the default Bayes factor is symmetric in information. This distribution of effects, which will be denoted by $\pi^*(\theta)$, should reflect for which effects we want the default Bayes factor to have equal error probabilities. The rule for setting the tuning parameter is

Rule: For a distribution of the standardized effect under H_1, $\pi^*(\theta)$, and a given sample size n, choose $g_{\pi^*} = E_{\pi^*(\theta)}[g^*(\theta, n)]$, $r_{\pi^*} = E_{\pi^*(\theta)}[r^*(\theta, n)]$, and $b_{\pi^*} = E_{\pi^*(\theta)}[b^*(\theta, n)]$.

We consider two choices for $\pi^*(\theta)$. The first option is a uniform distribution in the interval $[-1, 1]$, i.e., $\pi^*(\theta) = U(-1, 1)$, which implies that small, medium, and large effects are equally likely. The second option is a normal distribution with mean of 0 and a standard deviation of .6, i.e., $\pi^*(\theta) = N(0, 0.6^2)$, which implies that small effects are more likely than large effects. Note that 90% of a normal distribution $N(0, 0.6^2)$ lies within $[-1, 1]$. We focus on the interval $[-1, 1]$ because in the social sciences effect sizes larger than 1 are not realistic. Note that other choices for $\pi^*(\theta)$ could also be used.

In the case of the DBF, the optimal tuning parameter b_{π^*} can be computed using the following formula:

$$b_{\pi^*} = E_{\pi^*(\theta)}[b^*(\theta, n)] = \int b^*(\theta, n)\pi^*(\theta)d\theta \approx T^{-1}\sum_{t=1}^{T} b^*(\theta^{(t)}, n), \tag{4.10}$$

where $\theta^{(t)}$ is the tth draw from $\pi^*(\theta)$ and the total number of draws from $\pi^*(\theta)$, T, must be large enough, e.g., $T = 1000$. For each $b^*(\theta^{(t)}, n)$ an efficient algorithm for the computation of the optimal tuning parameter is provided in Appendix 4.B. This procedure can also be used for determining g_{π^*} and r_{π^*}.

Note that the use of the new default choices g_{π^*}, r_{π^*}, and b_{π^*} results in default Bayes factors that work as a classical test statistic. Thus, when selecting H_0 or H_1 depending on whether B_{01} is larger or smaller than 1, respectively, the type I and the type II error probabilities are equal on average when the standardized effect under H_1 is sampled from $\pi^*(\theta)$. Consequently, the outcome of the new default Bayes factor no longer reflects the relative evidence in the data between the two hypotheses of a researcher. Note however that a similar issue also arises when using the default Bayes

factors based on default choices for g, r, and b because the underlying default priors also do not directly reflect substantive prior beliefs of a researcher.

4.6 Consistency of tuned Bayes factors

Consistency is a crucial property in Bayesian hypothesis testing. This property implies that the Bayes factor will always select the true hypothesis when the sample size is large enough, i.e., it requires that as the sample size n goes to infinity, the Bayes factor for the null hypothesis approaches infinity if H_0 is true, and approaches 0 if H_1 is true. In this section we show that the tuned Bayes factors that were proposed in the previous section are consistent.

It is important to note that the Bayes factors for the null hypothesis based on default choices $b = 2/n$, $b = 1/\sqrt{n}$, $g = n$, $r = \sqrt{n}/2$ and $r = \sqrt{n}$ are consistent (O'Hagan, 1995; Liang et al., 2008). The following shows that the tuned Bayes factor based on b_{π^*}, g_{π^*}, and r_{π^*} are consistent as well, if the tuning parameters are always constrained in $b \in [\frac{2}{n}, 1]$, $g \in (0, n]$, and $r \in (0, \sqrt{n}]$ which have been specified in Section 4.3. Here we use the fact that as n goes to infinity, the observed effect converges to the true effect in the population.

First we consider the case that H_0 is true so that the observed standardized effect, $\bar{x}/s$, goes to 0 as n goes to infinity. In this case, DBF_{01} in (4.9) is a decreasing function of b, and ZBF_{01} in (4.6) and MBF_{01} in (4.8) are increasing functions of g and r. Then as shown in Figure 4.3, when the sample size increases, b^* decreases and g^* and r^* increase for a given effect size unequal to 0. Along this line of reasoning, it holds that b_{π^*} goes to 0, and g_{π^*} and r_{π^*} go to infinity as the sample size goes to infinity for a calibration distribution under H_1 for which $\Pr(\theta = 0) = 0$ holds under $\pi^*(\theta)$. Consequently, if the observed effect converges to 0, as the sample size goes to infinity, the tuned Bayes factors for the null hypothesis go to infinity. For example, $ZBF_{01} \to (1 + g_{\pi^*})^{\frac{1}{2}}$ when $\bar{x}/s \to 0$, which goes to infinity as n goes to infinity because g_{π^*} goes to infinity with n.

Second we consider the case that H_1 is true so that the observed standardized effect, $\bar{x}/s$, converges to a value unequal to 0 as n goes to infinity. As stated earlier the tuning parameters are constrained by $b \geq \frac{2}{n}$, $g \leq n$, and $r \leq \sqrt{n}$. Therefore, based on equation (4.10) b_{π^*}, g_{π^*}, and r_{π^*} under the distribution of standardized effect sizes $\pi^*(\theta)$ are constrained as well, i.e., $b_{\pi^*} \geq \frac{2}{n}$, $g_{\pi^*} \leq n$, and $r_{\pi^*} \leq \sqrt{n}$. Thus, the tuned Bayes factors BF_{01} are always smaller than the Bayes factors under $b = 2/n$, $g = n$, and $r = \sqrt{n}$ (as can be seen from (4.9), for example, DBF decreases as b increases). Consequently, if the observed effect converges to a value unequal to zero, as the sample size goes to infinity, the tuned Bayes factors for the null hypothesis go to 0, since the Bayes factors under $b = 2/n$, $g = n$, and $r = \sqrt{n}$ go to 0. For example, the DBF under b_{π^*} is smaller than the DBF under $b = 2/n$ which goes to 0 as n goes to infinity.

It should be noted that the tuned Bayes factors can be inconsistent when the tuning parameters are without constraints. Morey et al. (in press) elaborates that the tuning parameter specified for equal error probabilities may result in inconsistent Bayes factors. They show that for a specific standardized effect θ under H_1, the tuned Bayes factor for the null is always larger than 1 as long as the observed standardized effect $\bar{x}/s$ is less than half of θ. This implies that as n goes to infinity the tuned Bayes factors for the null hypothesis do not converge to 0 under the observed standardized effects that are less than $\theta/2$ but unequal to 0. For example, for a specific standardized effect of $\theta = 0.5$ under H_1 and sample size of $n = 100$, the tuning parameter in DBF for equal error probabilities is $b^* = 0.0134$ which can be roughly seen from Figure 4.2 (h). For an observed standardized effect $\bar{x}/s = 0.24$ that is less than half of $\theta = 0.5$, the tuned Bayes factor under $b^* = 0.0134$ is about 1.73 which supports Morey et al. (in press)'s finding. In our paper, however, we constrain the tuning parameters in reasonable ranges, i.e., $b \in [\frac{2}{n}, 1]$, $g \in (0, n]$, and $r \in (0, \sqrt{n}]$. This implies $b^* = 0.0134$ in the example above should be abandoned because it is smaller than $2/n$ in the case of $n = 100$. In fact, our method specifies $b^* = 0.02$ according to Figure 4.3 (a), which is equal to the default choice $b = 2/n$. This avoids the inconsistency issue of the tuned Bayes factors, since the previous studies have shown the consistency of Bayes factors under default choices of tuning parameters. Although constraining tuning parameters may lose the property of equal error probabilities for some specific standardized effects, we suggest using distributions $\pi^*(\theta)$ of the standardized effects under H_1 to reduce the influence of these constraints. This specification addresses the consistency issue of the tuned Bayes factors which will be illustrated below, and still results in approximately equal error probabilities which will be shown in the next section.

We illustrate the consistency of the tuned DBF based on b_{π^*} using two distributions of standardized effects, i.e., $\pi^*(\theta) = U(-1, 1)$ and $\pi^*(\theta) = N(0, 0.6^2)$. The sample size n increases from 10 to 500. For each n and a distribution of standardized effects, we compute b_{π^*} using (4.10) and DBF_{01} using (4.9) with different observed effects $\bar{x}/s = 0$, 0.1, 0.2 and 0.5. For an observed effect $\bar{x}/s = 0$, the DBF_{01} should go to infinity, and for $\bar{x}/s = 0.1$, 0.2 and 0.5, the DBF_{01} should go to 0 as sample size n goes to infinity.

Figure 4.4 illustrates the logarithm of the DBF with respect to n under different observed effect sizes. First, the logarithms of DBF_{01} with two default choices $b = 2/n$ and $b = 1/\sqrt{n}$ are shown in Figure 4.4 (a) and (b), respectively, which illustrates consistency since its logarithm goes to positive infinity under H_0 and goes to minus infinity under H_1. Second, Figure 4.4 (c) and (d) show the logarithms of DBF_{01} with b_{π^*} obtained using $\theta \sim U(-1, 1)$ and $\theta \sim N(0, 0.6^2)$ under H_1. As can be seen, the logarithm of DBF_{01} with b_{π^*} goes to positive infinity under observed effect $\bar{x}/s = 0$ and goes to negative infinity under other observed effects which correspond to the fact that H_1 is true. This suggests consistency of the DBF based on b_{π^*} under $\theta \sim U(-1, 1)$ and $\theta \sim N(0, 0.6^2)$. Furthermore, it is interesting to find that

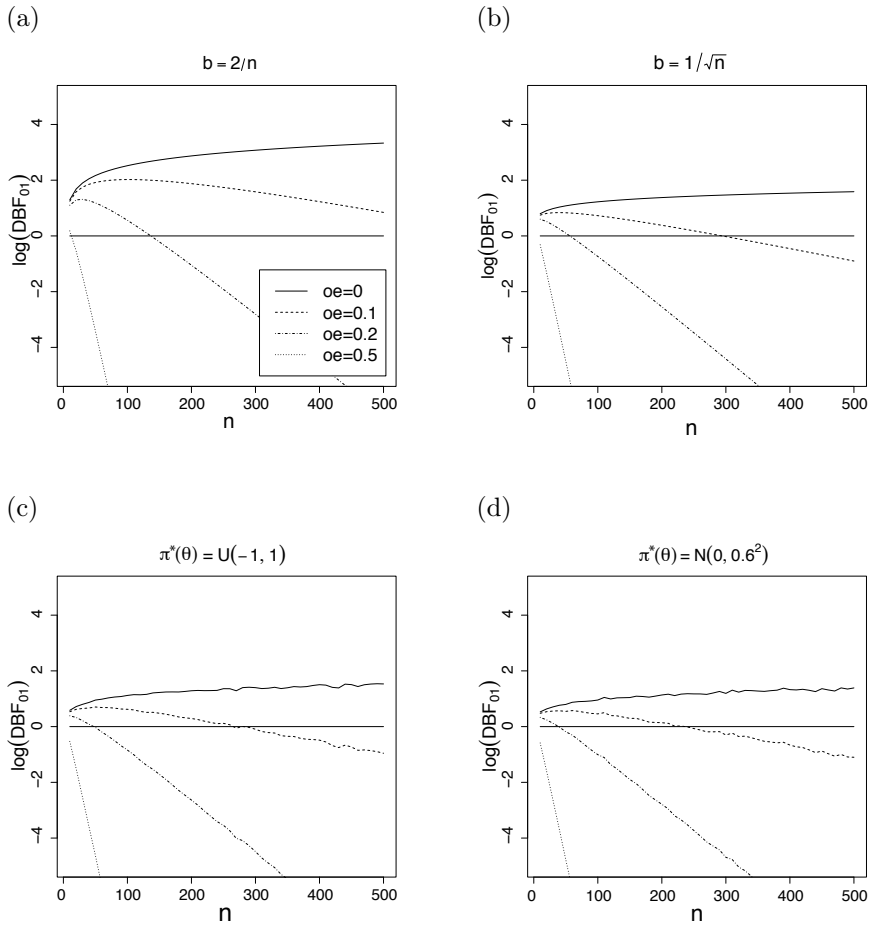

Figure 4.4: The logarithm of the DBF against n with default choices of b and b_{π^*} under two distributions of standardized effect under H_1. Note that oe denotes the observed effect $\bar{x}/s$.

the logarithms of tuned DBFs against n under $\pi^*(\theta) = U(-1, 1)$ in Figure 4.4 (c) and $\pi^*(\theta) = N(0, 0.6^2)$ in Figure 4.4 (d) are similar with the logarithm of the DBF against n under default choice $b = 2/n$ in Figure 4.4 (b). This illustrates that the tuned default Bayes factors also result in reasonable default outcomes of the relative evidence between two hypothesizes (under the condition that the default choices of the tuning parameters also considered to be "reasonable").

4.7 Numerical simulations

We investigated the error probabilities based on the new proposals discussed in the previous section and the default choices in different settings for various conditions that are typical in psychological research. We also conducted a sensitivity analysis of our new proposal by considering distributions of effects under H_1 that do not correspond to the distribution $\pi^*(\theta)$ that was used to tune the default Bayes factors.

4.7.1 Study 1

First, we consider the case where the distribution of the standardized effect under H_1 corresponds with the actual distribution of the standardized effect, for the case of $\pi^*(\theta) = U(-1, 1)$, and $\pi^*(\theta) = N(0, 0.6^2)$. The tuned parameters and error probabilities are obtained using (4.10) with $T = 1000$ for both $\pi^*(\theta) = U(-1, 1)$ and $\pi^*(\theta) = N(0, 0.6^2)$. The results in Tables 4.2 ($n = 20$), 4.3 ($n = 50$), and 4.4 ($n = 100$) display all the error probabilities for default choices and new choices for g, r, and b. The last two rows in each of the fragments for b, g, and r in the tables show the median logarithm of the Bayes factor for true hypothesis, which should be larger than 0. The logarithm of the Bayes factors are reported to check whether the tuned Bayes factors still render reasonable default outcomes (i.e., outcomes that are close to the "accepted" default outcomes).

Several conclusions can be drawn from these tables. If the anticipated distribution of the standardized effect is identical to the distribution used to generate the data in the simulation under H_1, Bayes factors based on optimal tuning parameters are approximately symmetric in information in most cases. Note that optimal tuning parameters can also render slightly unequal error probabilities. The reason is that when either a very large or small effect is sampled from its distribution, the optimal tuning parameters may attain their boundaries as was illustrated in Figure 4.3. For example, if the observed effect equals 0.9 Figure 4.3 (c) shows that the optimal r^* under $n = 20, 50$, and 100 is equal to the default choice $r = \sqrt{n}$, and if the observed effect equals 0.2 Figure 4.3 (b) shows that the optimal g^* is very close to 0.

Furthermore, the tables show that the Bayes factors based on default tuning parameters are always asymmetric in information. This asymmetry can be quite severe. For example, if default choice $r = \sqrt{n} = 4.47$ (Default 2) is used in the MBF

Table 4.2: Choices of b, g, r and error probabilities under two distributions of standardized effect θ and a sample size of $n = 20$. Note that for b, g, and r, Default 1 indicates $b = 2/n$, $g = 1$, and $r = \sqrt{n}/2$, respectively, and Default 2 indicates $b = 1/\sqrt{n}$, $g = n$, and $r = \sqrt{n}$, respectively. For b, g, and r, $mL(BF_{01})(H_0)$ denotes the median logarithm of BF_{01} under H_0 and $mL(BF_{10})(H_1)$ denotes the median logarithm of BF_{10} under H_1.

	$\theta \sim U(-1, 1)$			$\theta \sim N(0, 0.6^2)$		
$n = 20$	Tuned	Default 1	Default 2	Tuned	Default 1	Default 2
b	0.292	0.100	0.224	0.336	0.100	0.224
p_0^D	0.235	0.077	0.195	0.260	0.077	0.195
p_1^D	0.223	0.409	0.294	0.257	0.479	0.358
$p_0^D + p_1^D$	0.458	0.486	0.489	0.517	0.556	0.553
$mL(BF_{01})(H_0)$	0.472	1.435	0.656	0.391	1.445	0.666
$mL(BF_{10})(H_1)$	1.730	0.767	1.546	1.328	0.274	0.105
g	7.84	1.00	20.0	6.77	1.00	20.0
p_0^Z	0.203	0.253	0.085	0.219	0.253	0.085
p_1^Z	0.245	0.225	0.398	0.266	0.300	0.449
$p_0^Z + p_1^Z$	0.448	0.478	0.483	0.485	0.553	0.534
$mL(BF_{01})(H_0)$	0.882	0.230	1.299	0.822	0.231	1.300
$mL(BF_{10})(H_1)$	1.161	0.859	0.916	0.495	0.498	0.152
r	1.46	2.24	4.47	1.22	2.24	4.47
p_0^M	0.209	0.116	0.065	0.231	0.117	0.066
p_1^M	0.252	0.393	0.466	0.292	0.459	0.530
$p_0^M + p_1^M$	0.461	0.509	0.531	0.523	0.576	0.596
$mL(BF_{01})(H_0)$	0.548	0.866	1.510	0.427	0.860	1.518
$mL(BF_{10})(H_1)$	0.730	0.679	0.258	0.383	0.268	0.208

when $n = 20$, Table 4.2 shows that in the case of $\theta \sim N(0, 0.6^2)$ the error probability $p_1^M = 0.530$ under H_1 is 8 times larger than the error probability $p_0^M = 0.066$ under H_0. This was also observed for another case in Table 4.1.

It is also interesting that the optimal tuning parameters also result in a smaller sum of error probabilities under $\theta \sim U(-1, 1)$ and $\theta \sim N(0, 0.6^2)$. This is an interesting finding because it implies that the tuned default Bayes factors are not only symmetric in information when testing H_0 versus H_1, they are also most likely to select the true hypothesis on average if the distribution of standardized effect sizes specified in (4.10) is the same as the distribution used in the simulation under H_1.

The result of the median logarithm of Bayes factors renders the following findings. First, the median logarithms of the default and tuned Bayes factors result in positive evidence for the true hypothesis in all cases. Second, the median logarithms of the Bayes factors increase with the sample size, which supports the consistency of the

Table 4.3: Choices of b, g, r and error probabilities under two distributions of standardized effect θ and a sample size of $n = 50$. Note that for b, g, and r, Default 1 indicates $b = 2/n$, $g = 1$, and $r = \sqrt{n}/2$, respectively, and Default 2 indicates $b = 1/\sqrt{n}$, $g = n$, and $r = \sqrt{n}$, respectively. For b, g, and r, $mL(BF_{01})(H_0)$ denotes the median logarithm of BF_{01} under H_0 and $mL(BF_{10})(H_1)$ denotes the median logarithm of BF_{10} under H_1.

	$\theta \sim U(-1, 1)$			$\theta \sim N(0, 0.6^2)$		
$n = 50$	Tuned	Default 1	Default 2	Tuned	Default 1	Default 2
b	0.174	0.040	0.141	0.227	0.040	0.141
p_0^D	0.151	0.041	0.148	0.194	0.040	0.148
p_1^D	0.134	0.280	0.194	0.181	0.379	0.271
$p_0^D + p_1^D$	0.285	0.321	0.342	0.375	0.419	0.419
$mL(BF_{01})(H_0)$	0.734	1.936	0.852	0.521	1.891	0.806
$mL(BF_{10})(H_1)$	4.611	3.397	4.482	3.049	1.679	2.763
g	28.1	1.00	50.0	21.7	1.00	50.0
p_0^Z	0.136	0.245	0.048	0.188	0.245	0.049
p_1^Z	0.151	0.154	0.279	0.205	0.194	0.373
$p_0^Z + p_1^Z$	0.287	0.399	0.327	0.393	0.439	0.422
$mL(BF_{01})(H_0)$	1.464	0.232	1.741	1.358	0.241	1.758
$mL(BF_{10})(H_1)$	3.886	2.383	3.701	3.224	2.042	2.954
r	3.61	3.54	7.07	2.99	3.54	7.07
p_0^M	0.145	0.075	0.038	0.165	0.075	0.038
p_1^M	0.161	0.290	0.335	0.214	0.343	0.395
$p_0^M + p_1^M$	0.306	0.365	0.373	0.379	0.418	0.433
$mL(BF_{01})(H_0)$	1.332	1.306	1.987	1.138	1.293	1.964
$mL(BF_{10})(H_1)$	3.346	3.352	2.974	1.981	1.882	1.466

Bayes factors under both tuned and default choices. Third, the choices of b_{π^*}, g_{π^*}, and r_{π^*} under $\pi^*(\theta) = U(-1, 1)$ and $\pi^*(\theta) = N(0, 0.6^2)$ always result in a larger median logarithm of the Bayes factor under H_1 than under H_0. This implies our method is symmetric in two error probabilities, but not symmetric in terms of the magnitude of the support for the true hypothesis. This is a common property of the Bayes factor caused by the fact that it is easier to find support against H_0 instead of finding support for H_0 because H_0 is a precise hypothesis while H_1 is a composite hypothesis. Interested readers are referred to Johnson and Rossell (2010) who proposed a method where the evidence for the true hypothesis accumulates with the same rate under H_0 and H_1.

Table 4.4: Choices of b, g, r and error probabilities under two distributions of standardized effect θ and a sample size of $n = 100$. Note that for b, g, and r, Default 1 indicates $b = 2/n$, $g = 1$, and $r = \sqrt{n}/2$, respectively, and Default 2 indicates $b = 1/\sqrt{n}$, $g = n$, and $r = \sqrt{n}$, respectively. For b, g, and r, $mL(BF_{01})(H_0)$ denotes the median logarithm of BF_{01} under H_0 and $mL(BF_{10})(H_1)$ denotes the median logarithm of BF_{10} under H_1.

	$\theta \sim U(-1, 1)$			$\theta \sim N(0, 0.6^2)$		
$n = 100$	Tuned	Default 1	Default 2	Tuned	Default 1	Default 2
b	0.120	0.020	0.100	0.161	0.020	0.100
p_0^D	0.110	0.026	0.121	0.144	0.026	0.120
p_1^D	0.096	0.217	0.148	0.132	0.298	0.210
$p_0^D + p_1^D$	0.206	0.243	0.269	0.276	0.324	0.330
$mL(BF_{01})(H_0)$	0.880	2.282	0.985	0.719	2.285	0.989
$mL(BF_{10})(H_1)$	9.801	8.398	9.695	7.344	5.777	7.074
g	64.5	1.00	100	56.3	1.00	100
p_0^Z	0.105	0.242	0.032	0.122	0.242	0.032
p_1^Z	0.112	0.126	0.226	0.154	0.149	0.281
$p_0^Z + p_1^Z$	0.217	0.368	0.258	0.276	0.391	0.313
$mL(BF_{01})(H_0)$	1.885	0.242	2.101	1.783	0.224	2.064
$mL(BF_{10})(H_1)$	7.930	4.491	7.774	5.315	3.254	5.092
r	6.24	5.00	10.0	5.15	5.00	10.0
p_0^M	0.105	0.053	0.026	0.129	0.053	0.026
p_1^M	0.129	0.220	0.251	0.167	0.284	0.323
$p_0^M + p_1^M$	0.234	0.273	0.277	0.296	0.337	0.349
$mL(BF_{01})(H_0)$	1.842	1.633	2.306	1.683	1.642	2.322
$mL(BF_{10})(H_1)$	8.308	8.396	7.983	5.170	5.212	4.817

4.7.2 Study 2

In the second study the error probabilities were obtained when the distribution of the standardized effects used for determining the optimal tuning parameters differs from the actual distribution of the standardized effects under H_1 (Table 4.5). It can be seen from the top panel of Table 4.5 that a misspecified distribution of standardized effect sizes and the resulting tuning parameters render unequal error probabilities, which implies that the default Bayes factors are asymmetric in information.

Furthermore, the sum of error probabilities with respect to tuning parameters from misspecified distribution of standardized effect sizes is larger than those from true distribution. Based on these findings it can be concluded that the performance of the tuned default Bayes factors depends on whether the true sampling distribution of effects under H_1 corresponds with the chosen distribution π^* that is used for

Table 4.5: Choices of b, g, r under two distributions of standardized effect θ and error probabilities under true distributions of standardized effects and a sample size of $n = 50$.

	$\theta \sim U(-1,1)$ is true		$\theta \sim N(0, 0.6^2)$ is true	
$n = 50$	$\theta \sim U(-1,1)$	$\theta \sim N(0, 0.6^2)$	$\theta \sim U(-1,1)$	$\theta \sim N(0, 0.6^2)$
b	0.174	0.227	0.174	0.227
p_0^D	0.151	0.214	0.175	0.194
p_1^D	0.134	0.142	0.237	0.181
$p_0^D + p_1^D$	0.285	0.366	0.412	0.375
g	28.1	21.7	28.1	21.7
p_0^Z	0.136	0.075	0.066	0.188
p_1^Z	0.151	0.264	0.333	0.205
$p_0^Z + p_1^Z$	0.287	0.339	0.399	0.393
r	3.61	2.99	3.61	2.99
p_0^M	0.145	0.087	0.064	0.165
p_1^M	0.161	0.274	0.347	0.214
$p_0^M + p_1^M$	0.306	0.361	0.411	0.379
	$\theta = 0.2$ is true		$\theta = 0.5$ is true	
$n = 50$	$\theta \sim U(-1,1)$	$\theta \sim N(0, 0.6^2)$	$\theta \sim U(-1,1)$	$\theta \sim N(0, 0.6^2)$
b	0.174	0.227	0.174	0.227
p_0^D	0.175	0.205	0.172	0.214
p_1^D	0.483	0.440	0.022	0.018
$p_0^D + p_1^D$	0.658	0.645	0.194	0.232
g	28.1	21.7	28.1	21.7
p_0^Z	0.065	0.082	0.073	0.087
p_1^Z	0.667	0.632	0.048	0.048
$p_0^Z + p_1^Z$	0.732	0.714	0.121	0.132
r	3.61	2.99	3.61	2.99
p_0^M	0.062	0.075	0.077	0.093
p_1^M	0.682	0.638	0.087	0.060
$p_0^M + p_1^M$	0.744	0.713	0.164	0.153

calibration.

However, if we compare the results obtained using the wrong distribution in the top panel of Table 4.5 with the results obtained using the default choices in Table 4.3, the sum of error probabilities from the former is usually smaller than and otherwise about equal to the latter. For example, when $\theta \sim N(0, 0.6^2)$ is true under H_1 and $\pi^*(\theta) = U(-1, 1)$ is used for tuning a default Bayes factor, the resulting sum of the error probabilities equals .399 for the ZBF (see the fourth column of the top panel in

Table 4.6: Bayes factors obtained using different choices of b, g, and r. Note that for b, g, and r, Default 1 indicates $b = 2/n$, $g = 1$, and $r = \sqrt{n}/2$, respectively, and Default 2 indicates $b = 1/\sqrt{n}$, $g = n$, and $r = \sqrt{n}$, respectively. Tuned 1 and Tuned 2 indicate the optimal tuning parameters are obtained using $\theta \sim U(-1, 1)$ and $\theta \sim N(0, 0.6^2)$, respectively.

$n = 28$	Tuned 1	Tuned 2	Default 1	Default 2
b	0.248	0.294	0.071	0.189
DBF_{01}	0.382	0.344	1.12	0.457
g	12.4	10.1	1.00	28.0
ZBF_{01}	0.730	0.682	0.606	1.00
r	2.20	1.76	2.65	5.29
MBF_{01}	0.769	0.723	0.83	1.26

Table 4.5). This sum is smaller than .439 and .422 that can be found in the last two columns of Table 4.3 for Default 1 and Default 2 for the ZBF.

Furthermore, when specifying a wrong distribution, the difference between two error probabilities in the DBF is smaller than in the ZBF and the MBF. For example, if $\theta \sim N(0, 0.6^2)$ is true under H_1, and $\pi^*(\theta) = U(-1, 1)$ is used for tuning a default Bayes factor, this renders $p_0^D = 0.175$ and $p_1^D = 0.237$, $p_0^Z = 0.066$ and $p_1^Z = 0.333$, and $p_0^M = 0.064$ and $p_1^M = 0.347$. This implies that for distributions of effect sizes that cover most commonly used effect in the social sciences, the difference between two error probabilities of the tuned DBF is less sensitive to the wrong specification of distributions of effect sizes than the ZBF and the MBF. As can be seen in the bottom panel of Table 4.5, for each specified effect size the difference between error probabilities may be large. However, returning to the top panel of Table 4.5, averaged over a reasonable distribution of effect sizes for the DBF results in error probabilities that are similar.

4.8 Empirical example revisited

The empirical data of Howell (2012, p. 196) is re-analyzed using default Bayes factors with the new choices for the tuning parameters. We are interested in testing whether there is difference between the observed score and chance score, i.e., $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ using default Bayes factors.

We use the two tuned Bayes factors which are approximately symmetric in information under the distributions: $\pi^*(\theta) = U(-1, 1)$ and $\pi^*(\theta) = N(0, 0.6^2)$. The optimal choice for the tuning parameters for ZBF, MBF, and DBF are obtained given a sample size of $n = 28$, using the algorithm in Appendix 4.B. The resulting Bayes factors are displayed in Table 4.6 for the two tuned choices and the two default choices. The table shows that the ZBF, MBF, and DBF based on the tuned choices

all favor the alternative hypothesis which assumes there is a difference between the observed score and chance score, whereas they favor different hypotheses under two default tuning parameters. This suggests that different default Bayes factors render similar results as long as the tuning parameters are chosen based on the calibration scheme discussed in Section 4.5. The relative evidence for H_1 however is quite small. In order to draw more decisive conclusions more data need to be collected.

4.9 Discussion

In this paper we investigated the type I and type II error probabilities of default Bayes factors in Bayesian hypothesis testing of a population effect with unknown variance. It was shown that the error probabilities are unequal in situations that are typically encountered in psychological research (i.e., for sample sizes of 20, 50 and 100, and standardized effects between .2 and .8). In certain situations the asymmetry was quite severe. For example the Bayes factor based on a mixture of g priors with default tuning parameter $r = \sqrt{n}$ (Default 2), which corresponds to a standardized effect with a Cauchy prior with a scale parameter of 1, has a strong tendency to prefer H_0. Thus if one is interested in default Bayes factors with approximately equal error probabilities, this default Bayes factor is not recommended. The asymmetry in information was less severe for the other default Bayes factors.

It was also shown how a default Bayes factor can be tuned such that the error probabilities for a given sample size and a given distribution of standardized effect under the alternative hypothesis are approximately equal. Two choices for the distribution of effects were chosen (namely, a uniform distribution on [-1,1] and a normal distribution with mean 0 and standard deviation .6) which seem reasonable in psychological research. It was shown that the resulting 'tuned' Bayes factor is consistent in the sense that the evidence for the true hypothesis goes to infinity as sample size goes to infinity.

Furthermore, two numerical simulation studies showed that the tuned default Bayes factors also resulted in smaller sums of the error probabilities (which plays an important role in Bayesian hypothesis testing) than when using default choices. Therefore if the true distribution of standardized effect sizes under H_1 corresponds with the distribution that is used for tuning the default Bayes factors, we are more likely to select the true hypothesis and the error probabilities will also be close to each other.

When the true distribution of standardized effect sizes under H_1 does not correspond to the calibration distribution that is used for the tuned default Bayes factors, the error probabilities will also be unequal. However, the simulation study shows that the error probabilities are still closer to each other than when using the default choices. This simulation study also showed that the DBF seemed to be more robust to misspecification of the calibration distribution in comparison to the ZBF and the

MBF. For this reason, the tuned DBF may be preferred over the other default Bayes factors.

Furthermore, it was interesting to observe that the "tuned" default prior depends on the expected standardized effect size under the alternative in the same way as if we would have specified the prior based on subjective beliefs: when a large (small) standardized effect is expected, the prior variance of the effect under H_1 is relatively large (small). For example, Figure 4.3 (a) shows that a large standardized effect size corresponds to a small b^* which suggests a large prior variance in DBF. This implies that a well-specified subjective prior also results in Bayes factors with good frequency properties in the range of effects that are anticipated.

As was elaborated in the paper, the tuned Bayes factor acts as a classical test statistic. In fact, the test procedure for equal error probabilities can also be developed in a frequentist test. For example, we can conduct a classical t test, where the critical value of t statistic is determined by the sample size and standardized effect size under H_1 such that the type I and type II error probabilities are equal. For the distribution of the standardized effect size, an average of critical t values can be obtained using each standardized effect size from the distribution. The frequentist t test based on the average of these critical values would have similar outcomes of selecting H_0 or H_1 as the Bayesian t test using tuned Bayes factors.

Finally we want to mention two properties of the tuned Bayes factors which are not favorable from a Bayesian point of view. First the method is not coherent when sequentially observing data. For example assume that we observe a sample, say $\mathbf{x}_1$, and we would compute a tuned default Bayes factor. If we would observe a second sample, say $\mathbf{x}_2$, and we would update our tuned Bayes factor according to Bayes' theorem, the resulting Bayes factor would differ from the tuned Bayes factor based on the complete data set $(\mathbf{x}_1', \mathbf{x}_2')'$. Note that this issue is also present in other default Bayes factors such as the fractional Bayes factor (O'Hagan, 1995) and certain intrinsic Bayes factors (Berger & Pericchi, 1996).

Another issue lies in the interpretation of the tuned default Bayes factor. The problem is that the outcome is not the relative evidence between the hypotheses of the researcher because the underlying prior is not based on substantive beliefs but instead it is constructed using frequentist error probabilities. Note however that this is also an issue with default Bayes factors where default priors are chosen, not based on substantive expectations but on theoretical or computational simplicity.

A possible advantage of using a criterion that results in equal error probabilities, such as the tuned default Bayes factors, is that there is no tendency to either select the null or the alternative hypothesis. Furthermore, it selects the true hypothesis on average more often than default Bayes factors in certain scenario's. Consequently when a tuned default Bayes factor results in a preference towards H_0 or H_1 for a given fixed data set, this preference cannot be caused by an a priori tendency to prefer H_0 or H_1, respectively.

Acknowledge We would like to thank the non-anonymous reviewer Richard D. Morey for pointing out an error and an alternative manner of presentation of the material in Appendix 4.A.

4.A The existence of b^*, g^*, and r^*

Theorem 1: There exists a unique $b^* \in (1/n, 1]$ so that $p_0 = p_1$.
Proof: First note that $p_0^D = P(DBF_{01} < 1|H_0)$ and $p_1^D = P(DBF_{01} > 1|H_1) = 1 - P(DBF_{01} < 1|H_1)$, and therefore,

$$p_0^D = p_1^D \Leftrightarrow P(DBF_{01} < 1|H_0) + P(DBF_{01} < 1|H_1) = 1.$$

Furthermore, we set

$$\begin{aligned} c_D(b) &= p_0^D + 1 - p_1^D \\ &= P(DBF_{01} < 1|H_0) + P(DBF_{01} < 1|H_1) \\ &= P(h_D(\bar{x}/s) < A_D(b)|H_0) + P(h_D(\bar{x}/s) < A_D(b)|H_1), \end{aligned}$$

where $DBF_{01} = A_D(b)^{-1} h_D(\bar{x}/s)$, with $h_D(\bar{x}/s) = (1 + (\bar{x}/s)^2)^{-n/2}$ and $A_D(b) = \frac{\Gamma(nb/2)}{\Gamma((nb-1)/2)} / \frac{\Gamma(n/2)}{\Gamma((n-1)/2)}$. Based on the characteristic of the gamma function, it holds that $A_D(b)$ is a strictly increasing function of b. For this reason, $c_D(b)$ is an increasing function of $b \in (1/n, 1]$. If $b \to 1/n$, then $A_D(b) \to 0$ which implies that $\lim_{b \to 1/n} c_D(b) = 0$. On the other hand, if $b = 1$, then $A_D(b) = 1$ which implies that $c_D(1) = 2$ because $h_D(\bar{x}/s) < 1$. Therefore, a unique $b^* \in (1/n, 1]$ exists such that $c_D(b^*) = 1$, which implies that $p_0^D = p_1^D$. □

The $g^* \in (0, \infty)$ and $r^* \in (0, \infty)$ need not exist. For the ZBF, first note that

$$\begin{aligned} & ZBF_{01} < 1 \\ \Leftrightarrow \quad & (1+g)^{-\frac{n-1}{2}} (1 + g/(1 + (\bar{x}/s)^2))^{\frac{n}{2}} < 1 \\ \Leftrightarrow \quad & (1 + g/(1 + (\bar{x}/s)^2))^{\frac{n}{2}} < (1+g)^{\frac{n-1}{2}} \\ \Leftrightarrow \quad & 1 + g/(1 + (\bar{x}/s)^2) < (1+g)^{\frac{n-1}{n}} \\ \Leftrightarrow \quad & (1 + (\bar{x}/s)^2)^{-1} < ((1+g)^{\frac{n-1}{n}} - 1)/g. \end{aligned}$$

This implies $p_0^Z = P(h_Z(\bar{x}/s) < A_Z(g)|H_0)$ and $p_1^Z = P(h_Z(\bar{x}/s) > A_Z(g)|H_1)$, where $h_Z(\bar{x}/s) = (1 + (\bar{x}/s)^2)^{-1}$ and $A_Z(g) = ((1+g)^{\frac{n-1}{n}} - 1)/g$. The first derivative of $A_Z(g)$ is

$$\frac{d}{dg} A_Z(g) = g^{-2}(1+g)^{-1/n}[(1+g)^{1/n} - 1 - g/n]. \tag{4.11}$$

The last term is strictly negative, i.e., $(1+g)^{1/n} - 1 - g/n < 0$ if $n > 1$ and $g > 0$. The reason is that $(1+g)^{1/n} - 1 - g/n$ is a decreasing function of g because of its

derivative $\frac{1}{n}((1+g)^{1/n-1}-1)<0$, and $(1+g)^{1/n}-1-g/n$ goes to 0 when $g\to 0$. Therefore, $\frac{d}{dg}A_Z(g)<0$ which indicates that $A_Z(g)$ is a strictly decreasing function of g. Thus, as g decreases, p_0 increases and p_1 decreases. If $g\to\infty$ then $A_Z(g)\to 0$ and therefore $p_0^Z<p_1^Z$ in the limit. If $g\to 0$ then $A_Z(g)\to\frac{n-1}{n}$. In the case of $g\to 0$, $h_Z(\bar{x}/s)<A_Z(g)$ implies that $(1+(\bar{x}/s)^2)^{-1}<\frac{n-1}{n}$ which implies that $p_0^Z=P(|\bar{x}/s|>1/\sqrt{n-1}|H_0)$ and $p_1^Z=P(|\bar{x}/s|<1/\sqrt{n-1}|H_1)$.

If the effect size is larger than approximately $\frac{2}{\sqrt{n-1}}$, then $P(|\bar{x}/s|<\frac{1}{\sqrt{n-1}}|H_0)>P(|\bar{x}/s|>\frac{1}{\sqrt{n-1}}|H_1)$ because the distributions of $\bar{x}/s$ under H_0 and H_1 are approximately symmetric on $\theta/2$ which is larger than $\frac{1}{\sqrt{n-1}}$. As $g\to 0$, therefore, $p_0^Z>p_1^Z$ under $\theta>\frac{2}{\sqrt{n-1}}$. This renders the result that a unique g^* exists such that $p_0^Z=p_1^Z$ for $\theta>\frac{2}{\sqrt{n-1}}$.

If the effect size is smaller than approximately $\frac{2}{\sqrt{n-1}}$, then $P(|\bar{x}/s|<\frac{1}{\sqrt{n-1}}|H_0)<P(|\bar{x}/s|>\frac{1}{\sqrt{n-1}}|H_1)$ such that $p_0^Z<p_1^Z$ as $g\to 0$. Thus, there is no solution of g for $p_0^Z=p_1^Z$ because $p_0^Z<p_1^Z$ for any $g>0$. In this case, however, it is still applicable that the smaller the g, the less the difference between p_0^Z and p_1^Z. This forces us to select a g^* that is approaching 0 such that p_0^Z is as close as possible to p_1^Z. In the case of $p_0^Z<p_1^Z$ for any $g>0$, the computation algorithm in Appendix 4.B decreases g^* and stops until $g^*<0.001$. Note that $g^*=0.001$ is such a small number that will not much influence $g_{\pi *}$ that is an average of $g^*(\theta,n)$ obtained using different effects from their distribution. The above discussion of existence for g^* can also be applied to r^*.

4.B Computation of b^*, g^* and r^* given standardized effect size and sample size

The computation of the optimal tuning parameter b^*, g^*, and r^* resulting in equal error probabilities can be carried out using a dichotomy algorithm. The basic principle is to gradually adjust the tuning parameter by first computing the error probabilities p_0 and p_1 for a certain value of the tuning parameter b, and then let b decrease (or increase in the case of g or r) if $p_0>p_1$ or let b increase (or decrease in the case of g or r) if $p_0<p_1$. Furthermore, if $p_0>p_1$ when b is smaller than $2/n$, then $b^*=2/n$, the allowed lower bound of b in the DBF. Similarly, $g^*=n$ and $r^*=\sqrt{n}$ are chosen if the resulting $p_0>p_1$ in ZBF and MBF, respectively. The method is described for determining b^* in the DBF.

Computing the error probability. We compute the error probabilities for a given sample size n, standardized effect θ under H_1, and value for b. The error probabilities can then be obtained as follows.

(a) Randomly draw K samples of size n under H_0 and H_1, i.e., $\mathbf{x}_0^{(k)} = (x_{01}^{(k)}, \ldots, x_{0n}^{(k)})'$, $\mathbf{x}_1^{(k)} = (x_{11}^{(k)}, \ldots, x_{1n}^{(k)})'$, where $x_{0i}^{(k)} \sim N(0,1)$ and $x_{1i}^{(k)} \sim N(\theta, 1)$, respectively. Note that variances are set to 1 which is allowed because the ZBF, MBF, and DBF only depend on the standardized effect $\bar{x}/s$.

(b) Estimate the error probabilities as $\hat{p}_0 = \frac{1}{K}\sum_k I(DBF_{01}^k(\mathbf{x}_0, b) < 1)$, and $\hat{p}_1 = \frac{1}{K}\sum_k I(DBF_{01}^k(\mathbf{x}_1, b) > 1)$, where $I(\cdot)$ is the indicator function.

Obtaining the optimal b^*. LB and UB denote the (dynamic) lower and upper bound of b^* in this procedure, respectively.

1. Initialize $b' = 2/n$, $LB = 2/n$ and $UB = 1$.
2. Compute p_0^D and p_1^D using step (a) and (b).
3. If $p_0^D > p_1^D$, then set $b^* = 2/n$, and exit algorithm. Else, set $b' = (LB + UB)/2$.
4. Compute p_0^D and p_1^D based on b' using step (a) and (b).
5. If $|p_0^D - p_1^D|/[(p_0^D + p_1^D)/2] < e_b$, given an acceptable approximation bound $e_b = .01$, then set $b^* = b'$ and exit algorithm. Else,
   (i) let $LB = b'$ and $b' = (LB + UB)/2$ if $p_0^D < p_1^D$.
   (ii) let $UB = b'$ and $b' = (LB + UB)/2$ if $p_0^D > p_1^D$.
   (iii) Go to step 4.

As was discussed in Appendix 4.A, g^* and r^* need not exist for $p_0 = p_1$. To obtain the optimal g^* and r^*, we add an extra step between step 4 and 5: If $g' < 0.001$ or $r' < 0.001$ which is a replacement of b', then set $g^* = g'$ or $r^* = r'$ and exit algorithm.

Chapter 5

Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses[1]

5.1 Introduction

An informative hypothesis explicitly expresses a researcher's expectation with respect to the structure of the model parameters. It consists of equality and/or inequality constraints among the parameters of interest in a statistical model. For example, three equal parameters can be represented by an equality constrained hypothesis $H_1 : \theta_1 = \theta_2 = \theta_3$, and three ordered parameters can be represented by an inequality constrained hypothesis $H_2 : \theta_1 < \theta_2 < \theta_3$. Testing informative hypotheses is more flexible than the traditional null hypothesis testing of a null hypothesis with only equality constraints against an unconstrained alternative with no constraints on the parameters of interest.

The informative hypothesis has drawn a lot of attention both in frequentist hypothesis testing (Barlow, Bartholomew, Bremner, & Brunk, 1972; Silvapulle & Sen, 2004) and in Bayesian hypothesis testing (Hoijtink, 2012). In the frequentist framework, hypothesis testing with inequality constraints has been studied over fifty years starting with (Bartholomew, 1959). Some recent contributions can be found in van de Schoot et al. (2010), and Klugkist, Bullens, and Postma (2012). Bayesian evaluation of informative hypotheses by means of the Bayes factor is relatively new. A decade

[1]This chapter will be submitted as Gu, X., Mulder, J., & Hoijtink, H. Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses.
Author contributions: XG, JM, and HH designed the research. XG developed the software package, performed the data analyses and simulation study, and wrote the paper. JM and HH gave feedback on software development. JM and HH provided extensive feedback on constructing and writing the paper.

ago, Klugkist et al. (2005) started using Bayes factors to evaluate inequality constrained hypotheses in ANOVA models. Follow-up research appears in Klugkist and Hoijtink (2007) for Bayesian testing of inequality and about equality constrained hypotheses, in Mulder et al. (2009) for Bayesian informative hypothesis testing in repeated measures models, in Klugkist et al. (2010) for Bayesian evaluation of equality and inequality constrained hypotheses in contingency tables, and in Mulder et al. (2010) for Bayesian model selection of equality and inequality constrained hypotheses in the context of multivariate normal linear models. The developments on the use of Bayes factors for informative hypothesis testing are summarized in Hoijtink (2012). However, these studies are limited to assess informative hypotheses in specific models and cannot yet be applied in other models, e.g., confirmatory factor analysis or logistic regression. More recently, van de Schoot et al. (2012) enables researchers to test inequality constrained hypotheses in structural equation models, while Gu et al. (2014) allows to evaluate inequality constrained hypothesis in general statistical models. Although these studies enable inequality constrained hypothesis testing in a large number of statistical models using the Bayes factor, these methods cannot be used for testing hypotheses with equality constraints possibly in addition to inequalities.

The incessant debate between frequentist hypothesis testing and Bayesian hypothesis testing (Wagenmakers, 2007) has highlighted an advantage of the Bayes factor: it quantifies the relative support in the data for one hypothesis against another (Kass & Raftery, 1995). This cannot be done using classical p-values. However, the popularity of the Bayes factor is limited because of two reasons: the specification of the prior can be a difficult task, especially when prior information is weak or completely unavailable, and the computation can be very intensive when the statistical model is complex. To break these barriers, Bayesian statisticians have presented several default Bayes factors based on default priors, for example, JZS priors (Jeffreys, 1961; Zellner & Siow, 1980; Rouder et al., 2009), intrinsic priors (Berger & Pericchi, 1996), expected posterior priors (Pérez & Berger, 2002), and fractional priors (O'Hagan, 1995). Default priors usually do not reflect subjective prior beliefs and have distributional forms chosen such that the Bayes factor can easily be computed. The fractional prior stands out for its convenience of evaluating informative hypotheses (Mulder, 2014b). The fractional prior is implicitly specified by a noninformative prior updated with a fraction of the likelihood (O'Hagan, 1995). The remaining fraction of the likelihood is used for testing the hypotheses of interest. The resulting Bayes factor is called the fractional Bayes factor. Recently, Mulder (2014b) proposed an adjustment of the fractional Bayes factor where the fractional prior was shifted around the null value. This approach resulted in an adjusted fractional Bayes factor that converges faster to a true inequality constrained hypothesis. However, the current applications of (adjusted) fractional Bayes factors in informative hypothesis testing are still within the class of multivariate normal linear models.

This paper proposes an approximation of a fractional Bayes factor to extend its applicability for testing informative hypotheses for more general models. These models

can be generalized linear (mixed) models (McCullogh & Searle, 2001) such as logistic regression models and multilevel models, and structural equation models (Kline, 2011) such as path models, confirmatory factor analysis models and latent class models. Due to large sample theory (Gelman et al., 2004, p.101-107), the posterior distribution of the parameters in each model can be approximated by a (multivariate) normal distribution. This paper also approximates the implicit fractional prior with a (multivariate) normal distribution as a general methodology to ensure a fast computation of the (adjusted) fractional Bayesian factor. Based on these approximations we can approximate a fractional Bayes factor to evaluate informative hypotheses in general statistical models. In addition, we discuss different choices of the fraction (O'Hagan, 1995; Gu, Mulder, & Hoijtink, in press), which is a tuning parameter in the fractional prior, and provide a guideline for choosing this fraction.

This paper is organized as follows. Section 5.2 introduces the informative hypothesis in general statistical models, and illustrates how the informative hypothesis is constructed based on researchers' expectation by means of two empirical examples. Thereafter, Section 5.3 elaborates the specification of the adjusted fractional prior and the posterior distribution using normal approximations. Based on the specified prior and posterior distributions, the approximated adjusted fractional Bayes factor is derived and a software package is presented for the evaluation of informative hypotheses in general statistical models. In Section 5.4 we discuss different choices of the fraction, and conduct a sensitivity study for the fractional Bayes factors with those choices. Subsequently, Section 5.5 revisits the two empirical examples to show how to evaluate informative hypotheses using the proposed fractional Bayes factors. This paper ends with a short conclusion.

5.2 Informative hypotheses in general statistical models

A statistical model is described by the likelihood function $f(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{\zeta})$, where $\boldsymbol{X}$ denotes the data, $\boldsymbol{\theta}$ contains the parameters that are used to specify informative hypotheses, and $\boldsymbol{\zeta}$ contains the nuisance parameters. Informative hypotheses are constructed using equality and/or inequality constraints based on the theories or expectations of researchers. The general form of the informative hypothesis is given by

$$H_i : \boldsymbol{R}_{i_0}\boldsymbol{\theta} = \boldsymbol{r}_{i_0}, \boldsymbol{R}_{i_1}\boldsymbol{\theta} > \boldsymbol{r}_{i_1}, \tag{5.1}$$

where $\boldsymbol{R}_{i_0}$ and $\boldsymbol{R}_{i_1}$ are the restriction matrices for equality and inequality constraints in H_i, respectively, and $\boldsymbol{r}_{i_0}$ and $\boldsymbol{r}_{i_1}$ contain constants. Note that the number of rows in $\boldsymbol{R}_{i_0}$ equals the number of equality constraints, the number of rows in $\boldsymbol{R}_{i_1}$ equals the number of inequality constraints, and the numbers of columns in $\boldsymbol{R}_{i_0}$ and $\boldsymbol{R}_{i_1}$ equal the length of $\boldsymbol{\theta}$. For example, hypothesis $H_1 : \theta_1 = 2\theta_2 = 3\theta_3 > 4\theta_4 < 5$

corresponds to

$$\boldsymbol{R}_{1_0}\boldsymbol{\theta} = \begin{bmatrix} 1 & -2 & 0 & 0 \\ 0 & 2 & -3 & 0 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \boldsymbol{r}_{1_0},$$

$$\boldsymbol{R}_{1_1}\boldsymbol{\theta} = \begin{bmatrix} 0 & 0 & 3 & -4 \\ 0 & 0 & 0 & -4 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{bmatrix} > \begin{bmatrix} 0 \\ -5 \end{bmatrix} = \boldsymbol{r}_{1_1}.$$

Note that informative hypotheses with range constraints, for example, $H_2 : 0 < \theta < 1$ are not considered in this paper.

An informative hypothesis H_i can be tested against the unconstrained hypothesis

$$H_u : \boldsymbol{\theta} \text{ is unconstrained,} \tag{5.2}$$

against its complement

$$H_{i_c} : \text{not } H_i, \tag{5.3}$$

which expresses what a researchers does not expect, or against another informative hypothesis

$$H_{i'} : \boldsymbol{R}_{i'_0}\boldsymbol{\theta} = \boldsymbol{r}_{i'_0}, \boldsymbol{R}_{i'_1}\boldsymbol{\theta} > \boldsymbol{r}_{i'_1}. \tag{5.4}$$

It should be noted that when an informative hypothesis H_i contains at least one equality constraint, the complement of H_i is the same as the unconstrained hypothesis H_u.

Before evaluating the informative hypotheses, the parameters of interest may need to be standardized in some situations. The need of standardization depends on the statistical model and informative hypothesis under evaluation. On the one hand, the parameters have to be standardized when comparing, e.g., coefficients in regression models and factor loadings in confirmatory factor analysis. For example, testing whether the regression coefficient θ_1 is larger than θ_2 requires the standardization of θ_1 and θ_2, because a large coefficient can also result from a large scale of the corresponding predictor. On the other hand, it may not be necessary to standard the parameters $\boldsymbol{\theta}$ if they are compared to constants, and it is undesirable to standardize the parameters $\boldsymbol{\theta}$ if they represent the means. For instance, testing whether a regression coefficient is larger than 0 or testing whether the mean of group 1 is smaller than the mean of group 2 does not require standardization. If standardization is required, Gu et al. (2014) discussed two ways to do this: (1) standardize all observed and latent variables, or (2) use standardized parameters. In the situation considered by Gu et al. (2014), there was little difference between the performances of the two methods. Therefore, researchers can use either of them if necessary.

In what follows, we will use two empirical examples to illustrate how researchers' expectations can be expressed by informative hypotheses.

Table 5.1: Data descriptive for variables in regression model

	y_i	x_{1i}	x_{2i}	x_{3i}
mean	965.92	5.91	35.24	16.86
standard deviation	74.82	1.36	26.76	2.27

5.2.1 Example 1 : multiple regression

The first example concerns a multiple regression model used in Guber (1999) to investigate the relation between the educational costs of the school and the academic performance of the students. The data were collected in 50 U.S. states (available at `www.amstat.org/publications/jse/secure/v7n2/datasets.guber.cfm`). The performance of the students is measured by the average total SAT score y_i ranging from 400 to 1600. Its predictors are the average public school expenditure x_{1i}, the percentage of students taking the SAT exams x_{2i}, and the average pupil/teacher ratio x_{3i}. The descriptives of the dependent variable y_i and independent variables x_{1i}, x_{2i} and x_{3i} are shown in Table 5.1. The relationship between the student performance and its predictors is given in a regression model.

$$y_i = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \theta_3 x_{3i} + \epsilon_i, \tag{5.5}$$

where θ_0 is the intercept, θ_1, θ_2 and θ_3 are the regression coefficients, and $\epsilon_i \sim N(0, \sigma^2)$ denotes the residuals with σ^2 being their residual variance. For this regression model, the likelihood is

$$f(\boldsymbol{X}|\boldsymbol{\theta}, \boldsymbol{\zeta}) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{-\frac{1}{2\sigma^2}(y_i - \theta_0 - \theta_1 x_1 - \theta_2 x_2 - \theta_3 x_3)^2\}, \tag{5.6}$$

where $n = 50$ denotes the sample size, and $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^T$ and $\boldsymbol{\zeta} = (\theta_0, \sigma^2)$.

Guber (1999) theorized that higher education expenditures results in better performance of the students in SAT exams, which implies that the coefficient θ_1 of the predictor x_{1i} is positive. In addition, in those states with a small percentage of the students taking SATs, the students are expected to do well because they have self-selected themselves into the SAT exam which is only required by universities with a high prestige. This implies that the coefficient θ_2 of the predictor x_{2i} is negative. Furthermore, although a lower pupil/teacher ratio would be associated with better performance, a school needs to spend more money on education and therefore this predictor overlaps with the expenditures. This suggests that the coefficient θ_3 of predictor x_{3i} is zero. Consequently, we specify the informative hypothesis:

$$H_1 : \theta_1 > 0, \theta_2 < 0, \theta_3 = 0 \tag{5.7}$$

with $\boldsymbol{R}_{1_0} = (0, 0, 1)$, $\boldsymbol{R}_{1_1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix}$, $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^T$, $r_{1_0} = 0$, and $\boldsymbol{r}_{1_1} = (0, 0)^T$ in $H_1 : \boldsymbol{R}_{1_0}\boldsymbol{\theta} = r_{1_0}, \boldsymbol{R}_{1_1}\boldsymbol{\theta} > \boldsymbol{r}_{1_1}$. Hypothesis H_1 can be tested against its

Table 5.2: Data in repeated measures ANOVA

	Baseline		Training	
Subject	week 1	week 2	week 3	week 4
1	21	22	6	6
2	20	19	4	4
3	17	15	4	5
4	25	30	12	17
5	30	27	8	6
6	19	27	7	4
7	26	16	2	5
8	17	18	1	5
9	26	24	8	9

complement

$$H_{1_c} : \text{not } H_1. \tag{5.8}$$

5.2.2 Example 2: repeated measures ANOVA

We reanalyze the example of the repeated measures ANOVA used in Howell (2012, p.462) based on an experiment with relaxation therapy. The experiment investigated the duration of nine patients's migraine headaches before and after relaxation training. The duration of headaches is measured by the number of hours per week. Our example uses the data for the last two weeks of the baseline where patients received no training and the last two weeks of training. Therefore, the data shown in Table 5.2 consists of four dependent variables, i.e., the number of hours with a headache per week for nine patients in four weeks. The random effects model for these dependent variables is (Hox, 2010, p.83):

$$y_{ij} = \mu + \eta_i + \tau_j + \epsilon_{ij}, \tag{5.9}$$

where y_{ij} for $i = 1, \ldots, 9$ and $j = 1, \ldots, 4$ denotes the four dependent variables, μ denotes the grand mean, $\eta_i \sim N(0, \sigma_\eta^2)$ denotes the random difference for person i which is constant for different j, τ_j denotes the fixed measurement difference for week j which is constant for different i, and $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ is the measurement error with respect to person i and week j. To investigate the effect of relaxation training, we specify the individual differences with a random effect and the treatment differences with a fixed effect. Thus, the mean for each measurement is

$$\theta_j = \mu + \tau_j \tag{5.10}$$

and $\Sigma_{j=1}^4 \tau_j = 0$.

The researchers expected a reduction of the duration of headaches after relaxation training. Furthermore, it is reasonable to expect that the mean durations are equal

in the first two weeks of baseline and in the last two weeks of training to ensure that other factors do not influence the duration of headaches. These expectations can be expressed by the following informative hypothesis:

$$H_2 : \theta_1 = \theta_2 > \theta_3 = \theta_4 \tag{5.11}$$

with $\boldsymbol{R}_{2_0} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}$, $\boldsymbol{R}_{2_1} = [0, 1, -1, 0]$, $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)^T$, $\boldsymbol{r}_{2_0} = (0, 0)^T$, and $r_{2_1} = 0$ in $H_2 : \boldsymbol{R}_{2_0}\boldsymbol{\theta} = r_{2_0}, \boldsymbol{R}_{2_1}\boldsymbol{\theta} > \boldsymbol{r}_{2_1}$. We compare this hypothesis to another informative hypothesis representing that the mean number of headache hours continually declines in the four weeks:

$$H_{2'} : \theta_1 > \theta_2 > \theta_3 > \theta_4, \tag{5.12}$$

which only contains inequality constraints $\boldsymbol{R}_{2'_1}\boldsymbol{\theta} > \boldsymbol{r}_{2'_1}$ with $\boldsymbol{r}_{2'_1} = (0, 0, 0)^T$ and

$$\boldsymbol{R}_{2'_1} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}.$$

The informative hypotheses constructed in these examples can be evaluated using Bayes factors, which will be elaborated in the next section. We will revisit these examples in Section 5.5 to display the results of the evaluation of these informative hypotheses.

5.3 Approximated adjusted fractional Bayes factors

The Bayes factor is the corner-stone of Bayesian hypothesis testing. It quantifies the relative evidence in the data for one hypothesis against another. The Bayes factor of an informative hypothesis H_i against another informative hypothesis $H_{i'}$ is defined by their marginal likelihood ratio (Jeffreys, 1961; Kass & Raftery, 1995):

$$BF_{ii'} = \frac{m(\boldsymbol{X}|H_i)}{m(\boldsymbol{X}|H_{i'})}. \tag{5.13}$$

In Bayesian hypothesis testing, the Bayes factor has a direct interpretation as the relative evidence from the data for one hypothesis against another. If $BF_{ii'} > 1$ ($BF_{ii'} < 1$), this implies that hypothesis H_i ($H_{i'}$) receives more support from the data. Specifically, if $BF_{ii'} = 5$, then the support for H_i is 5 times larger than for $H_{i'}$. For researchers who are new to Bayes factors we recommend using the guidelines for the interpretation of Bayes factors as provided by Kass and Raftery (1995). The degree of evidence in favor of H_i can be classified as unconvincing for $1 < BF_{ii'} < 3$, positive for $BF_{ii'} > 3$, strong for $BF_{ii'} > 20$, and very strong for $BF_{ii'} > 150$.

However, these rules for interpreting Bayes factors are not strict and can differ in different contexts.

The informative hypothesis H_i is nested in the unconstrained hypothesis H_u which does not contain any constraints on $\boldsymbol{\theta}$. When comparing H_i to H_u we can use the encompassing prior approach of Klugkist et al. (2005) where a prior is constructed under H_i via a truncation of the unconstrained (or encompassing) prior under H_u. Consequently, the Bayes factor for the informative hypothesis against the unconstrained hypothesis can be expressed as:

$$BF_{iu} = \frac{\int\int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_i} \pi_u(\boldsymbol{\theta},\boldsymbol{\zeta}|\boldsymbol{X})d\boldsymbol{\theta}d\boldsymbol{\zeta}}{\int\int_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_i} \pi_u(\boldsymbol{\theta},\boldsymbol{\zeta})d\boldsymbol{\theta}d\boldsymbol{\zeta}}, \tag{5.14}$$

where $\pi_u(\boldsymbol{\theta},\boldsymbol{\zeta})$ and $\pi_u(\boldsymbol{\theta},\boldsymbol{\zeta}|\boldsymbol{X})$ are the prior and posterior distributions of $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ under H_u, and $\boldsymbol{\Theta}_i = \{\boldsymbol{\theta}|\boldsymbol{R}_{i_0}\boldsymbol{\theta} = \boldsymbol{r}_{i_0}, \boldsymbol{R}_{i_1}\boldsymbol{\theta} > \boldsymbol{r}_{i_1}\}$ is the parameter space of $\boldsymbol{\theta}$ in agreement with the informative hypothesis H_i. Thus, in order to compute the Bayes factor the unconstrained prior and corresponding unconstrained posterior need to be determined, and subsequently the unconstrained prior and posterior need to be integrated over the constrained region under the informative hypothesis. In this section we propose a novel and general approach by using normal distributions to approximate the unconstrained posterior and the unconstrained fractional prior to compute default Bayes factors.

5.3.1 Fractional prior and posterior

To avoid ad hoc or subjective specification of the unconstrained prior, we consider the default approach of O'Hagan (1995) which is referred to as the fractional Bayes factor. In this approach the prior is automatically generated using a fraction of the likelihood. The resulting unconstrained *fractional prior* is specified using a noninformative prior and a proportion of the likelihood:

$$\pi_u(\boldsymbol{\theta},\boldsymbol{\zeta}|\boldsymbol{X}^b) \propto \pi_u^N(\boldsymbol{\theta},\boldsymbol{\zeta})f(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{\zeta})^b, \tag{5.15}$$

where $\pi_u^N(\boldsymbol{\theta},\boldsymbol{\zeta})$ is the noninformative prior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$, and b is the fraction on the likelihood $f(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{\zeta})$. The posterior distribution can then be obtained using the fractional prior distribution $\pi_u(\boldsymbol{\theta},\boldsymbol{\zeta}|\boldsymbol{X}^b)$ and the remaining likelihood $f(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{\zeta})^{1-b}$:

$$\pi_u(\boldsymbol{\theta},\boldsymbol{\zeta}|\boldsymbol{X}) \propto \pi_u(\boldsymbol{\theta},\boldsymbol{\zeta}|\boldsymbol{X}^b)f(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{\zeta})^{1-b} \propto \pi_u^N(\boldsymbol{\theta},\boldsymbol{\zeta})f(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{\zeta}). \tag{5.16}$$

Note that the unconstrained posterior is identical to the posterior based on a noninformative improper prior updated with the complete data.

Finally note that the nuisance parameters can be integrated out, which results in the following marginal prior and marginal posterior for $\boldsymbol{\theta}$:

$$\pi_u(\boldsymbol{\theta}|\boldsymbol{X}^b) = \int \pi_u(\boldsymbol{\theta},\boldsymbol{\zeta}|\boldsymbol{X}^b)d\boldsymbol{\zeta} \tag{5.17}$$

and

$$\pi_u(\boldsymbol{\theta}|\boldsymbol{X}) = \int \pi_u(\boldsymbol{\theta}, \boldsymbol{\zeta}|\boldsymbol{X}) d\boldsymbol{\zeta}. \tag{5.18}$$

5.3.2 Normal approximations to fractional prior and posterior distributions

Due to large sample theory (e.g., Gelman et al., 2004, p. 101), the posterior in (5.18) can be approximated using a normal distribution where the mean is equal to the maximum likelihood estimate and the covariance matrix is equal to the inverse of the Fisher information matrix, i.e.,

$$\pi_u(\boldsymbol{\theta}|\boldsymbol{X}) \approx N(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}_\theta), \tag{5.19}$$

where $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}_\theta$ denote the maximum likelihood estimate and covariance matrix of $\boldsymbol{\theta}$, respectively. $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}_\theta$ can be obtained using statistical software, such as, Mplus (Muthén & Muthén, 2010) or the R-package lavaan (Rosseel, 2012). This will be further elaborated when we come back to the empirical examples in Section 5.5.

The fractional prior in (5.17) is also centered around the maximum likelihood estimate. However, it is based on a fraction b of the data which implies an approximated covariance matrix of $\hat{\boldsymbol{\Sigma}}_\theta/b$. Consider, for example, a normally distributed data set $x_i \sim N(\theta, \sigma^2)$ with known σ^2. The posterior of θ is given by $\pi_u(\theta|X) = N(\hat{\theta}, \hat{\sigma}^2_\theta)$ where $\hat{\theta}$ equals the sample mean $\bar{x}$ and $\hat{\sigma}^2_\theta = \sigma^2/n$. In this setting the fractional prior of θ would be $\pi_u(\theta|X^b) = N(\hat{\theta}, \hat{\sigma}^2_\theta/b) = N(\bar{x}, \sigma^2/nb)$. For this reason we propose to approximate the fractional prior according to

$$\pi_u(\boldsymbol{\theta}|\boldsymbol{X}^b) \approx N(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}_\theta/b). \tag{5.20}$$

5.3.3 Adjusting the prior mean

It has been suggested to center the prior distribution of $\boldsymbol{\theta}$ around the focal point of interest. This heuristic argument was first proposed by Jeffreys (1961) when evaluating $H_1 : \theta \leq 0$ against its complement $H_2 : \theta > 0$. By constructing the priors for θ under H_1 and H_2 as a truncation of an unconstrained prior that is centered around the focal point 0, the prior distribution for θ under both hypotheses are essentially equivalent; the only difference is the sign. A more detailed discussion on centering prior means can be found in Mulder (2014b). In this paper, we adjust the prior in (5.20) as follows:

$$\pi^*_u(\boldsymbol{\theta}|\boldsymbol{X}^b) = N(\boldsymbol{\theta}^*, \hat{\boldsymbol{\Sigma}}_\theta/b), \tag{5.21}$$

where the adjusted prior mean is given by $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}^*_i = \{\boldsymbol{\theta}|\boldsymbol{R}_{i_0}\boldsymbol{\theta} = \boldsymbol{r}_{i_0}, \boldsymbol{R}_{i_1}\boldsymbol{\theta} = \boldsymbol{r}_{i_1}\}$. For each informative hypothesis, one can define a parameter space $\boldsymbol{\Theta}^*_i$ which contains one or more $\boldsymbol{\theta}^*$. For example, $H_1 : \theta_1 > 2\theta_2 > 4$ results in $\boldsymbol{\theta}^* = (4, 2)^T$, and

$H_2 : \theta_1 = \theta_2$ results in $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}_i^* = \{\theta_1, \theta_2 | \theta_1 = \theta_2\}$ in which $\theta_1^* = \theta_2^*$ can be any value. Below we will deal with the choice of $\boldsymbol{\theta}^*$.

Adjusting the prior mean from $\hat{\boldsymbol{\theta}}$ to $\boldsymbol{\theta}^*$ results in a slight change of the posterior for $\boldsymbol{\theta}$. In particular, the posterior mean of $\hat{\boldsymbol{\theta}}$ would be slightly shifted towards the prior mean $\boldsymbol{\theta}^*$. Large sample theory however dictates that the prior has a negligible effect on the posterior for large samples. Therefore, we leave the approximated posterior for $\boldsymbol{\theta}$, given by $N(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}_\theta)$, unaltered. Note that a similar argument is used in the BIC approximation of the Bayes factor (Schwarz, 1978; Kass & Raftery, 1995).

Based on the adjusted fractional prior distribution (5.21) and the posterior distribution (5.19), the approximated adjusted fractional Bayes factor (AAFBF) for an informative hypothesis versus the unconstrained hypothesis can be defined as:

$$AAFBF_{iu} = \frac{\int_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_i} \pi_u(\boldsymbol{\theta}|\boldsymbol{X}) d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_i} \pi_u^*(\boldsymbol{\theta}|\boldsymbol{X}^b) d\boldsymbol{\theta}}, \tag{5.22}$$

where the parameter space $\boldsymbol{\Theta}_i = \{\boldsymbol{\theta} | \boldsymbol{R}_{i_0}\boldsymbol{\theta} = \boldsymbol{r}_{i_0}, \boldsymbol{R}_{i_1}\boldsymbol{\theta} > \boldsymbol{r}_{i_1}\}$ is in agreement with the informative hypothesis H_i. The computation of the AAFBF will be elaborated in Section 5.3.5.

5.3.4 Comparable informative hypotheses

The prior distribution proposed in (5.21) depends on the informative hypothesis under evaluation, because the prior mean $\boldsymbol{\theta}^*$ is located on the boundary of the constrained region of the informative hypothesis. When two or more informative hypotheses are under comparison, the intersection of their constrained regions must be nonempty so that a common unconstrained prior mean $\boldsymbol{\theta}^*$ exists to evaluate all informative hypotheses against the unconstrained hypothesis. A set of informative hypotheses H_i for $i = 1, \ldots, I$ are comparable if there exists at least one solution of $\boldsymbol{\theta}$ to the set of equations (Mulder et al., 2010):

$$\begin{bmatrix} \boldsymbol{R}_{1_0} \\ \boldsymbol{R}_{1_1} \end{bmatrix} \boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{r}_{1_0} \\ \boldsymbol{r}_{1_1} \end{bmatrix}, \ldots, \begin{bmatrix} \boldsymbol{R}_{I_0} \\ \boldsymbol{R}_{I_1} \end{bmatrix} \boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{r}_{I_0} \\ \boldsymbol{r}_{I_1} \end{bmatrix}. \tag{5.23}$$

The solution of $\boldsymbol{\theta}$ for these equations defines the parameter space $\boldsymbol{\Theta}^*$. Examples for comparable hypotheses are $H_1 : \theta = 0$ versus $H_2 : \theta > 0$ and $H_3 : \theta_1 > \theta_2 > \theta_3$ versus $H_4 : \theta_3 > \theta_2 > \theta_1$. Hypotheses $H_5 : \theta_1 = \theta_2$ versus $H_6 : \theta_1 > \theta_2 + 1$ are not comparable because there is no solution of θ_1 and θ_2 for equations $\theta_1 = \theta_2$ and $\theta_1 = \theta_2 + 1$. It should be noted that the hypothesis $H_7 : \theta_1 > 0, \theta_2 > 0, \theta_2 > \theta_1 - 1$ cannot be properly evaluated yet because a solution does not exist for equations $\theta_1 = 0$, $\theta_2 = 0$, and $\theta_2 = \theta_1 - 1$.

5.3.5 Bayes factor computation

This section presents the computation of the AAFBF. First of all, we need to determine the adjusted prior mean $\boldsymbol{\theta}^*$ in (5.21). Finding the parameter space $\boldsymbol{\Theta}_i^*$ can be difficult for complicated informative hypotheses (Mulder et al., 2012). However, if we transform the parameters of interest using $\boldsymbol{\beta}_0 = \boldsymbol{R}_{i_0}\boldsymbol{\theta} - \boldsymbol{r}_{i_0}$ and $\boldsymbol{\beta}_1 = \boldsymbol{R}_{i_1}\boldsymbol{\theta} - \boldsymbol{r}_{i_1}$, then the informative hypothesis under consideration becomes $H_i : \boldsymbol{\beta}_0 = 0, \boldsymbol{\beta}_1 > 0$ such that we can simply specify the prior mean vector equal to zero for the new parameter vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^T, \boldsymbol{\beta}_1^T)^T$. This parameter transformation was elaborated in Chapter 3 of this dissertation, and was also used in Mulder (in press) for hypotheses with only inequality constraints. Note that the parameter transformation of $\boldsymbol{\theta}$ to $\boldsymbol{\beta}$ simplifies the form of the hypothesis without changing the expectation of researchers. For instance, testing whether two parameters are equal $\theta_1 = \theta_2$ is identical to testing whether their difference is 0, i.e., $\beta_0 = \theta_1 - \theta_2 = 0$. Consequently, the adjusted fractional prior distribution and posterior distribution for the new parameter $\boldsymbol{\beta}$ are given by:

$$\pi_u^*(\boldsymbol{\beta}|\boldsymbol{X}^b) = N(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_\beta/b) \tag{5.24}$$

and

$$\pi_u(\boldsymbol{\beta}|\boldsymbol{X}) = N(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}_\beta), \tag{5.25}$$

respectively, where $\hat{\boldsymbol{\beta}} = \boldsymbol{R}\hat{\boldsymbol{\theta}} - \boldsymbol{r}$ and $\hat{\boldsymbol{\Sigma}}_\beta = \boldsymbol{R}\hat{\boldsymbol{\Sigma}}_\theta\boldsymbol{R}^T$ with $\boldsymbol{R} = (\boldsymbol{R}_{i_0}^T, \boldsymbol{R}_{i_1}^T)^T$ and $\boldsymbol{r} = (\boldsymbol{r}_{i_0}^T, \boldsymbol{r}_{i_1}^T)^T$. Specifically, $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_0^T, \hat{\boldsymbol{\beta}}_1^T)^T$ where $\hat{\boldsymbol{\beta}}_0 = \boldsymbol{R}_{i_0}\hat{\boldsymbol{\theta}} - \boldsymbol{r}_{i_0}$ and $\hat{\boldsymbol{\beta}}_1 = \boldsymbol{R}_{i_1}\hat{\boldsymbol{\theta}} - \boldsymbol{r}_{i_1}$, and $\hat{\boldsymbol{\Sigma}}_\beta = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{\beta_0} & \hat{\boldsymbol{\Sigma}}_{01} \\ \hat{\boldsymbol{\Sigma}}_{10} & \hat{\boldsymbol{\Sigma}}_{\beta_1} \end{bmatrix}$ where $\hat{\boldsymbol{\Sigma}}_{\beta_0} = \boldsymbol{R}_{i_0}\hat{\boldsymbol{\Sigma}}_\theta\boldsymbol{R}_{i_0}^T$ and $\hat{\boldsymbol{\Sigma}}_{\beta_1} = \boldsymbol{R}_{i_1}\hat{\boldsymbol{\Sigma}}_\theta\boldsymbol{R}_{i_1}^T$.

This parameter transformation from $\boldsymbol{\theta}$ to $\boldsymbol{\beta}$ simplifies the computation of the AAFBF. First, the AAFBF for an informative hypothesis with only equality constraints, i.e., $H_i : \boldsymbol{\beta}_0 = \boldsymbol{0}$, compared to the unconstrained hypothesis can be obtained using the Savage-Dickey density ratio (Dickey, 1971; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010; Mulder, 2014b):

$$AAFBF_{iu}^0 = \frac{\pi_u(\boldsymbol{\beta}_0 = \boldsymbol{0}|\boldsymbol{X})}{\pi_u^*(\boldsymbol{\beta}_0 = \boldsymbol{0}|\boldsymbol{X}^b)}, \tag{5.26}$$

where $\pi_u^*(\boldsymbol{\beta}_0 = \boldsymbol{0}|\boldsymbol{X}^b)$ and $\pi_u(\boldsymbol{\beta}_0 = \boldsymbol{0}|\boldsymbol{X})$ are the densities of the prior (5.24) and posterior (5.25), respectively, for $\boldsymbol{\beta}_0$ at the point $\boldsymbol{\beta}_0 = \boldsymbol{0}$ under H_u. Second, the AAFBF for an informative hypothesis with only inequality constraints, i.e., $H_i : \boldsymbol{\beta}_1 > \boldsymbol{0}$, compared to the unconstrained hypothesis is given by (Hoijtink, 2012; Mulder, 2014b):

$$AAFBF_{iu}^1 = \frac{\int_{\boldsymbol{\beta}_1 > \boldsymbol{0}} \pi_u(\boldsymbol{\beta}_1|\boldsymbol{X}) d\boldsymbol{\beta}_1}{\int_{\boldsymbol{\beta}_1 > \boldsymbol{0}} \pi_u^*(\boldsymbol{\beta}_1|\boldsymbol{X}^b) d\boldsymbol{\beta}_1}, \tag{5.27}$$

where $\pi_u^*(\boldsymbol{\beta}_1|\boldsymbol{X}^b)$ and $\pi_u(\boldsymbol{\beta}_1|\boldsymbol{X})$ are the prior (5.24) and posterior (5.25), respectively, for $\boldsymbol{\beta}_1$. Finally, the AAFBF for an informative hypothesis with both equality and inequality constraints, i.e., $H_i : \boldsymbol{\beta}_0 = \mathbf{0}, \boldsymbol{\beta}_1 > \mathbf{0}$, compared to the unconstrained hypothesis can be obtained via:

$$AAFBF_{iu} = \frac{\pi_u(\boldsymbol{\beta}_0 = \mathbf{0}|\boldsymbol{X})}{\pi_u^*(\boldsymbol{\beta}_0 = \mathbf{0}|\boldsymbol{X}^b)} \cdot \frac{\int_{\boldsymbol{\beta}_1 > \mathbf{0}} \pi_u(\boldsymbol{\beta}_1|\boldsymbol{\beta}_0 = \mathbf{0}, \boldsymbol{X}) d\boldsymbol{\beta}_1}{\int_{\boldsymbol{\beta}_1 > \mathbf{0}} \pi_u^*(\boldsymbol{\beta}_1|\boldsymbol{\beta}_0 = \mathbf{0}, \boldsymbol{X}^b) d\boldsymbol{\beta}_1}, \tag{5.28}$$

where $\pi_u^*(\boldsymbol{\beta}_1|\boldsymbol{\beta}_0 = \mathbf{0}, \boldsymbol{X}^b)$ and $\pi_u(\boldsymbol{\beta}_1|\boldsymbol{\beta}_0 = \mathbf{0}, \boldsymbol{X})$ are the prior and posterior distributions of $\boldsymbol{\beta}_1$ given $\boldsymbol{\beta_0} = \mathbf{0}$, respectively. Note that $\pi_u^*(\boldsymbol{\beta}_1|\boldsymbol{\beta}_0 = \mathbf{0}, \boldsymbol{X}^b) = N(\mathbf{0}, (\hat{\boldsymbol{\Sigma}}_{\beta_1} - \hat{\boldsymbol{\Sigma}}_{10}\hat{\boldsymbol{\Sigma}}_{\beta_0}^{-1}\hat{\boldsymbol{\Sigma}}_{01})/b)$ and $\pi_u(\boldsymbol{\beta}_1|\boldsymbol{\beta}_0 = \mathbf{0}, \boldsymbol{X}) = N(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\Sigma}}_{10}\hat{\boldsymbol{\Sigma}}_{\beta_0}^{-1}\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\Sigma}}_{\beta_1} - \hat{\boldsymbol{\Sigma}}_{10}\hat{\boldsymbol{\Sigma}}_{\beta_0}^{-1}\hat{\boldsymbol{\Sigma}}_{01})$.

We let $c_i^0 = \pi_u^*(\boldsymbol{\beta}_0 = \mathbf{0}|\boldsymbol{X}^b)$ and $c_i^1 = \int_{\boldsymbol{\beta}_1 > \mathbf{0}} \pi_u^*(\boldsymbol{\beta}_1|\boldsymbol{X}^b) d\boldsymbol{\beta}_1$, which can be interpreted as the relative complexities of equality constrained hypothesis and inequality constrained hypothesis, respectively, compared to H_u under prior (5.24). Then, in general

$$c_i = \pi_u^*(\boldsymbol{\beta}_0 = \mathbf{0}|\boldsymbol{X}^b) \cdot \int_{\boldsymbol{\beta}_1 > \mathbf{0}} \pi_u^*(\boldsymbol{\beta}_1|\boldsymbol{\beta}_0 = \mathbf{0}, \boldsymbol{X}^b) d\boldsymbol{\beta}_1 \tag{5.29}$$

represents the relative complexity of informative hypothesis H_i (Hoijtink, 2012; Mulder, 2014a), which is a relative measure of the size of the parameter space under an informative hypothesis in comparison to the unconstrained parameter space. For example, the relative complexity of "$\theta_1 > \theta_2$, and θ_3 unconstrained" is larger than the relative complexity of "$\theta_1 > \theta_2 > \theta_3$". This can be understood from the fact that the parameter space of the latter is a subset of the parameter space of the first. Similarly, the relative complexity of "$\theta_1 = 0$, θ_2 unconstrained" is larger than the relative complexity of "$\theta_1 = 0, \theta_2 = 0$". It is interesting to note that the relative complexity c_i^0 of an equality constrained hypothesis $H_i : \boldsymbol{\beta} = \mathbf{0}$ becomes smaller when the prior variance of $\boldsymbol{\beta}$ under H_u becomes larger. The reason is that a larger variance of the unconstrained prior implies that a larger region of the unconstrained parameter space is likely a priori, which means that H_i is simpler relative to the unconstrained hypothesis. Furthermore, we let $f_i^0 = \pi_u(\boldsymbol{\beta}_0 = \mathbf{0}|\boldsymbol{X})$ and $f_i^1 = \int_{\boldsymbol{\beta}_1 > \mathbf{0}} \pi_u(\boldsymbol{\beta}_1|\boldsymbol{X}) d\boldsymbol{\beta}_1$, which can be interpreted as the measures of relative fit of the equality constrained hypothesis and inequality constrained hypothesis, respectively, compared to H_u. Then,

$$f_i = \pi_u(\boldsymbol{\beta}_0 = \mathbf{0}|\boldsymbol{X}) \cdot \int_{\boldsymbol{\beta}_1 > \mathbf{0}} \pi_u(\boldsymbol{\beta}_1|\boldsymbol{\beta}_0 = \mathbf{0}, \boldsymbol{X}) d\boldsymbol{\beta}_1 \tag{5.30}$$

expresses the relative fit of H_i (Hoijtink, 2012; Mulder, 2014a), which implies how well a hypothesis is supported by the data compared to the unconstrained hypothesis. The relative complexity and fit in the AAFBF can be estimated based on a similar procedure presented in Chapter 3 of this dissertation. The computation of the AAFBF

is implemented in the software package `BaIn` (Bayesian evaluation of informative hypotheses) available at `http://informative-hypotheses.sites.uu.nl/software/`. A user manual for `BaIn` is given in Appendix 5.A. The input of `BaIn` needs the maximum likelihood estimate and covariance matrix of the parameters of interest, which can be obtained using other software packages such as Mplus (Muthén & Muthén, 2010) or the free R-package lavaan (Rosseel, 2012). Executing `BaIn` renders the AAFBF for each informative hypothesis H_i under evaluation.

The Bayes factor of an informative hypothesis H_i against its complement H_{i_c} is

$$AAFBF_{ii_c} = \frac{f_i}{c_i} / \frac{1 - f_i}{1 - c_i}, \tag{5.31}$$

if H_i does not contain equality constraints. Otherwise $AAFBF_{ii_c} = AAFBF_{iu}$ because the marginal likelihood of the complement of a hypothesis which contains equality constraints is equal to the marginal likelihood of the unconstrained hypothesis. For the comparison of two informative hypotheses H_i and $H_{i'}$, the AAFBF for H_i against $H_{i'}$ can be obtained by

$$AAFBF_{ii'} = AAFBF_{iu}/AAFBF_{i'u}. \tag{5.32}$$

Running `BaIn` for H_i and $H_{i'}$ renders $AAFBF_{iu}$ and $AAFBF_{i'u}$ such that $AAFBF_{ii'}$ can be computed using (5.32).

5.4 Choices for b

This section discusses the choices of the fraction b for the specification of fractional priors. We first show the influence of the choices of b on the AAFBF when evaluating informative hypotheses. Thereafter, two traditional choices and one novel choice of b are presented. At the end of this section, a sensitivity study is conducted to investigate the approximation error of the AAFBF relative to the actual adjusted fractional Bayes factor. It should be noted that this paper uses one common fraction b of the likelihood for prior specification. For this reason the AAFBF should only be used for testing hypotheses based on data that come from one population or balanced data with equal group sizes in the case of multiple populations, similar as the fractional Bayes factor (de Santis & Spezzaferri, 2001).

5.4.1 The role of b in AAFBF

The influence of fraction b on the AAFBF is different for the evaluation of equality constraints $\boldsymbol{R}_{i_0}\boldsymbol{\theta} = \boldsymbol{r}_{i_0}$ and for the evaluation of inequality constraints $\boldsymbol{R}_{i_1}\boldsymbol{\theta} > \boldsymbol{r}_{i_1}$. First of all, the fraction b is a very influential parameter when evaluating equality constraints $\boldsymbol{R}_{i_0}\boldsymbol{\theta} = \boldsymbol{r}_{i_0}$. The underlying reason is that a small (large) b implies a prior with large (small) variance such that the prior density evaluated at $\boldsymbol{R}_{i_0}\boldsymbol{\theta} = \boldsymbol{r}_{i_0}$ or

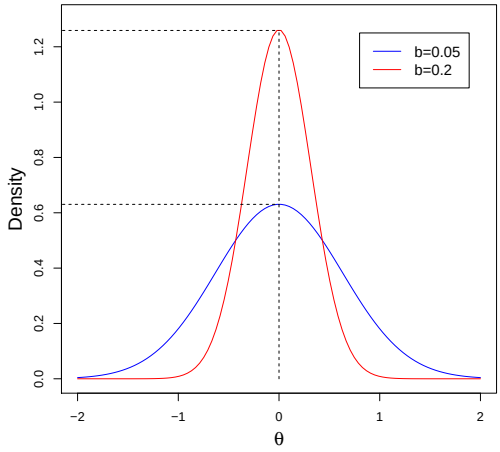

Figure 5.1: Relative complexities under different b

$\boldsymbol{\beta}_0 = \mathbf{0}$ in (5.26) is small (large). This can be illustrated in Figure 5.1 in which the blue line and the red line represent the densities of prior distribution $\pi_u^*(\theta|x^b) = N(0, \sigma_\theta^2/b)$ with $\sigma_\theta^2 = 0.02$ under $b = 0.05$ and $b = 0.2$, respectively. As can be seen, when testing hypothesis $H_1 : \theta = 0$ vs H_u, the prior density at $\theta = 0$ is 0.63 under $b = 0.05$, which is two times smaller than 1.26, the prior density at $\theta = 0$ under $b = 0.2$. Given an estimate of $\hat{\theta} = 0.2$ the resulting AAFBF for H_1 against H_u under $b = 0.05$ is $AAFBF_{1u} = 1.64$, whereas the AAFBF under $b = 0.2$ is $AAFBF_{1u} = 0.82$ according to equation (5.26). Secondly, the AAFBF is independent of the choice of b for inequality constrained hypotheses. This property was proven in Mulder (2014b) and can also be seen in Figure 5.1 where the prior probability that the constraint of $H_3 : \theta > 0$ holds under H_u is equal to 0.5 for both choices of b.

The influence of b on AAFBF is illustrated in Figure 5.2 when comparing the equality constrained hypothesis $H_1 : \theta = 0$ to the unconstrained hypothesis H_u, and comparing the inequality constrained hypothesis $H_2 : \theta > 0$ to H_u. Given the estimate $\hat{\theta} = 0.2$ and variance $\hat{\sigma}_\theta^2 = 0.02$ for θ, Figure 5.2 shows the AAFBF for each informative hypothesis under various $b \in (0, 0.5]$. As can be seen, the AAFBF for H_1 decreases as b increases, and the AAFBF for H_2 is stable as b changes. This illustrates that the fraction b has to be carefully specified when equality constrained hypotheses are of interest by the researcher, while any fraction b can be used when only inequality constrained hypotheses are formulated by the user. In what follows

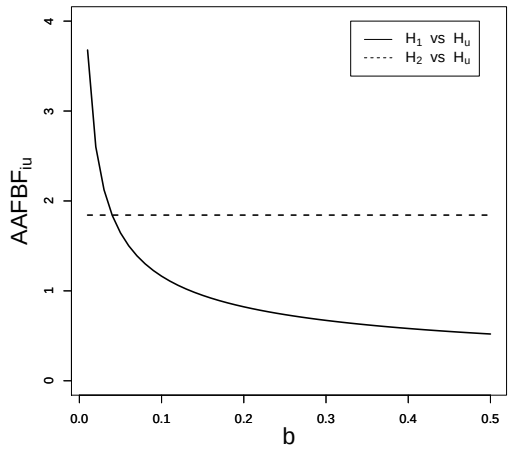

Figure 5.2: Influence of b on AAFBF

we will specify b in three different ways.

5.4.2 Traditional choices for b

Previous studies have recommended two choices for b for the fractional Bayes factor. The first one comes from Berger and Pericchi (1996) and O'Hagan (1995) who suggested using the minimal training sample for prior specification to leave maximal information in the data for hypothesis testing. This corresponds to $b = m/n$ in the fractional prior, where m is the size of the minimal training sample that makes all parameters identifiable. For example, for the one sample t test of $H_0 : \theta = 0$ where data is $x_i \sim N(\theta, \sigma^2)$, the actual adjusted fractional prior distribution for θ is $\pi_u^*(\theta|x^b) = t(0, s^2/(nb-1), nb-1)$, i.e., a Student t density with mean 0, scale parameter $s^2/(nb-1)$, and degree of freedom $nb-1$. In this case, the minimal m is 2 because $m = 1$ results in $b = 1/n$ and a degree of freedom 0, which is not allowed.

For the AAFBF we propose a similar approach to determine our first choice of b. To estimate $\boldsymbol{\beta}$ (with length J) we need at least $J + 1$ observations. Therefore, our first choice of the fraction equals

$$b_{min} = (J + 1)/n. \tag{5.33}$$

Note that J is the number of independent constraints in all the informative hypotheses under investigation, i.e., J equals the rank of $\boldsymbol{R} = (\boldsymbol{R}_{1_0}^T, \boldsymbol{R}_{1_1}^T, \ldots, \boldsymbol{R}_{I_0}^T, \boldsymbol{R}_{I_1}^T)^T$ for

a set of informative hypotheses H_i for $i = 1, \ldots, I$. Thus, if $H_3 : \theta_1 = 0$ and $H_4 : \theta_1 > 0, \theta_2 > 0$ are under evaluation, for example, $J = 2$ when computing the AAFBF for each informative hypothesis against the unconstrained hypothesis because there are two independent constraints.

For the multiple regression model (5.5) in Section 5.2, $J = 3$ because $H_1 : \theta_1 > 0, \theta_2 < 0, \theta_3 = 0$ can be formulated using a vector $\boldsymbol{\beta}$ of length 3. With the sample size of $n = 50$, the first choice of the fraction b can be set as $b_{min} = 4/50$. For the repeated measures model (5.10), $J = 3$ based on a vector $\boldsymbol{\beta}$ of length 3 in $H_2 : \theta_1 = \theta_2 > \theta_3 = \theta_4$ and $H_{2'} : \theta_1 > \theta_2 > \theta_3 > \theta_4$, and therefore $b_{min} = 4/36$ based on sample size $n = 36$.

The second way of choosing b is (O'Hagan, 1995):

$$b_{robust} = \max\{(J+1)/n, 1/\sqrt{n}\}, \tag{5.34}$$

which is in general larger than the first choice. O'Hagan (1995) stated that a larger b can reduce the sensitivity of the fractional Bayes factor to the distributional form of the prior. This choice can also be applied to the AAFBF defined in (5.22). When setting a larger fraction b, the AAFBF becomes more similar to the non-approximated adjusted fractional Bayes factor. We will elaborate more on this topic in Section 5.4.4. Given the sample size of $n = 50$ in the regression model in Section 5.2, $b_{robust} = 1/\sqrt{50}$ is specified to evaluate hypothesis H_1. In the case of the repeated measures model with sample size $n = 36$, one can set $b_{robust} = 1/6$ for the comparison of H_2 and $H_{2'}$.

5.4.3 A frequentist choice for b

Gu et al. (in press) recently proposes another method of specifying b by taking into account the frequentist error probabilities. In Bayesian hypothesis testing, the probability of a Bayes factor favouring H_u when H_i is true is

$$p_1 = P(BF_{iu} < 1|H_i) \tag{5.35}$$

which corresponds to the Type I error probability if H_i would be a traditional null hypothesis, and the probability of a Bayes factor favouring H_i when H_u is true is

$$p_2 = P(BF_{iu} > 1|H_u). \tag{5.36}$$

which then corresponds to the Type II error probability. Gu et al. (in press) found that these probabilities are often quite different when using traditional choices of b in the one sample t test. This may not be preferable from a frequentist point of view where the goal typically is to control the error probabilities. Here we show how to specify b to control the error probabilities under certain conditions. First, we shall use a one sample t test to illustrate the procedure for specifying b based on this method, and then apply it to the AAFBF (5.26) for general statistical models. In the end, a rule of choosing b is proposed.

One sample t test

Consider a one sample t test for which data come from $x_i \sim N(\theta, \sigma^2)$, where θ denotes the population mean and σ^2 denotes the population variance, and the hypotheses under consideration are $H_1 : \theta = 0$ against $H_u : \theta$. The AAFBF for H_1 against H_u can be derived using equation (5.26):

$$AAFBF_{1u} = b^{-1/2} \exp\left(-\frac{1}{2} n(\bar{x}/s)^2\right), \tag{5.37}$$

where $\bar{x} = \sum_{i=1}^{n} x_i/n$ and $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$. For this AAFBF the error probabilities (5.35) and (5.36) become

$$\begin{aligned} p_1 = P(AAFBF_{1u} < 1|H_1) &= P(|\frac{\bar{x}}{s}| > \sqrt{-\log b/n} \mid H_1) \\ &\approx \frac{1}{L}\sum_{l=1}^{L} I(|\frac{\bar{x}^{(1l)}}{s^{(1l)}}| > \sqrt{-\log b/n}) \end{aligned} \tag{5.38}$$

and

$$\begin{aligned} p_2 = P(AAFBF_{1u} > 1|H_u) &= P(|\frac{\bar{x}}{s}| < \sqrt{-\log b/n} \mid H_u) \\ &\approx \frac{1}{L}\sum_{l=1}^{L} I(|\frac{\bar{x}^{(2l)}}{s^{(2l)}}| > \sqrt{-\log b/n}), \end{aligned} \tag{5.39}$$

where $\bar{x}^{(1l)}$ and $s^{(1l)}$ for $l = 1, \ldots, L$ are the mean and standard deviation of data $x_i^{(1l)}$ sampled from H_1, $\bar{x}^{(2l)}$ and $s^{(2l)}$ are the mean and standard deviation of data $x_i^{(2l)}$ sampled from H_u, and $I(\cdot)$ is the indicator function which is 1 if the argument is true and 0 otherwise. When sampling data from H_u, an expected standardized effect size, denoted by β_e, needs to be specified under H_u, i.e., $H_u : \theta = \beta_e\sigma$ so that the scaled data is sampled from $y_i \sim N(\beta_e, 1)$ under H_u where $y_i = x_i/\sigma$. Note that sampling $\frac{\bar{x}^{(2l)}}{s^{(2l)}}$ based on $x_i \sim N(\theta, \sigma^2)$ where $\theta/\sigma = \beta_e$ is identical to sampling the mean $\bar{y}^{(2l)}$ based on $y_i \sim N(\beta_e, 1)$. The specification of the standardized effect size β_e will be discussed in Section 5.4.3.

In the one sample t test, $\frac{\bar{x}}{s}$ is the observed standardized effect size known as Cohen's d (Cohen, 1992). It has sampling distributions under H_1 and H_u which can be obtained using $\frac{\bar{x}^{(1l)}}{s^{(1l)}}$ and $\frac{\bar{x}^{(2l)}}{s^{(2l)}}$, respectively. Figure 5.3 shows the distributions of $\frac{\bar{x}}{s}$ under $H_1 : \theta = 0$ (solid line) and $H_u : \theta = \beta_e$ (dashed line) given $\sigma^2 = 1$ and $n = 20$, where $\beta_e = 0.5$ is the pre-specified standardized effect size under H_u. Note that according to Cohen (1992) $\beta_e = .2$, $.5$, and $.8$ correspond to the small, medium, and large effects, respectively. If we use $b_{min} = 2/n$, the error probabilities in (5.38) and (5.39) become $p_1 = P(|\frac{\bar{x}}{s}| > 0.34|H_1) = 0.073$ and $p_2 = P(|\frac{\bar{x}}{s}| <$

$0.34|H_u) = 0.241$, whereas if we specify $b_{robust} = 1/\sqrt{n}$, the error probabilities are $p_1 = P(|\frac{\bar{x}}{s}| > 0.27|H_1) = 0.122$ and $p_2 = P(|\frac{\bar{x}}{s}| < 0.27|H_u) = 0.159$. These error probabilities are marked in Figure 5.3 (a) for b_{min} and (b) for b_{robust}, where the dark grey area represents p_1 and the light grey area represents p_2. As can be seen, $p_1 < p_2$ under both b_{min} and b_{robust}, which means that we are more likely to incorrectly prefer H_1 when H_u is true than incorrectly prefer H_u when H_1 is true.

In order to correct for this, Gu et al. (in press) showed how to choose b such that $p_1 = p_2$ given sample size n and effect size β_e under H_u. A direct way of obtaining such a b is proposed by Morey et al. (in press) and illustrated in Figure 5.3 (c). As can be seen in Figure 5.3 (c), the distributions of $\frac{\bar{x}}{s}$ under $H_1 : \theta = 0$ and $H_u : \theta = \beta_e$ are symmetric on $\beta_e/2$. This implies that we can simply specify $\sqrt{-\log b/n} = \beta_e/2$ or equivalently $b = \exp(-n\beta_e^2/4)$ to attain equal error probabilities, because $p_1 = P(|\frac{\bar{x}}{s}| > \beta_e/2|H_1)$ is equal to $p_2 = P(|\frac{\bar{x}}{s}| < \beta_e/2|H_u)$. For example, given $n = 20$ and $\beta_e = 0.5$ under H_u in Figure 5.3 (c), the dark grey area for p_1 has the same size as the light grey area for p_2 when setting $b = \exp(-n\beta_e^2/4) = 0.287$. The error probabilities under this setting are $p_1 = p_2 = 0.139$.

General case

The method of choosing b based on equal error probabilities can be generalized to the AAFBF of any $H_i : \boldsymbol{\beta}_0 = 0$ against $H_u : \boldsymbol{\beta}_0 \neq \mathbf{0}$. Based on the adjusted fractional prior (5.24) and approximated posterior (5.25), the AAFBF in (5.26) is

$$AAFBF^0_{iu} = b^{-1/2} \exp\left(-\frac{1}{2}\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\Sigma}}_\beta^{-1}\hat{\boldsymbol{\beta}}^T\right). \tag{5.40}$$

It is interesting to note that $\sqrt{\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\Sigma}}_\beta^{-1}\hat{\boldsymbol{\beta}}^T}$ in (5.40) is the test statistic in Wald test (Engle, 1984) which assumes that $\boldsymbol{\beta}$ is approximately normally distributed. The test statistic is not only the cornerstone in frequentist hypothesis testing, but it is also important in default Bayes factors. For example, the Bayes factor proposed by Rouder et al. (2009) for the t test is a function of t statistic, and the Bayes factor based on Zellner's g prior (Zellner & Siow, 1980) in regression models is a function of F statistic. The standardized effect size is often defined as a test statistic divided by $\sqrt{n}$ to offset the influence of the sample size (Cohen, 1992), because the effect size should not be affected by the sample size as it expresses the degree to which H_u differs from H_i. Thus, the observed standardized effect size in this case can be defined as

$$\hat{\beta}_e = \sqrt{\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\Sigma}}_\beta^{-1}\hat{\boldsymbol{\beta}}^T/n}. \tag{5.41}$$

Then using the steps as in (5.38) and (5.39) for the one sample t test, the error probabilities of AAFBFs are defined as

$$p_1 = P(AAFBF^0_{iu} < 1|H_i) = P(\hat{\beta}_e > \sqrt{-\log b/n}|H_i) \tag{5.42}$$

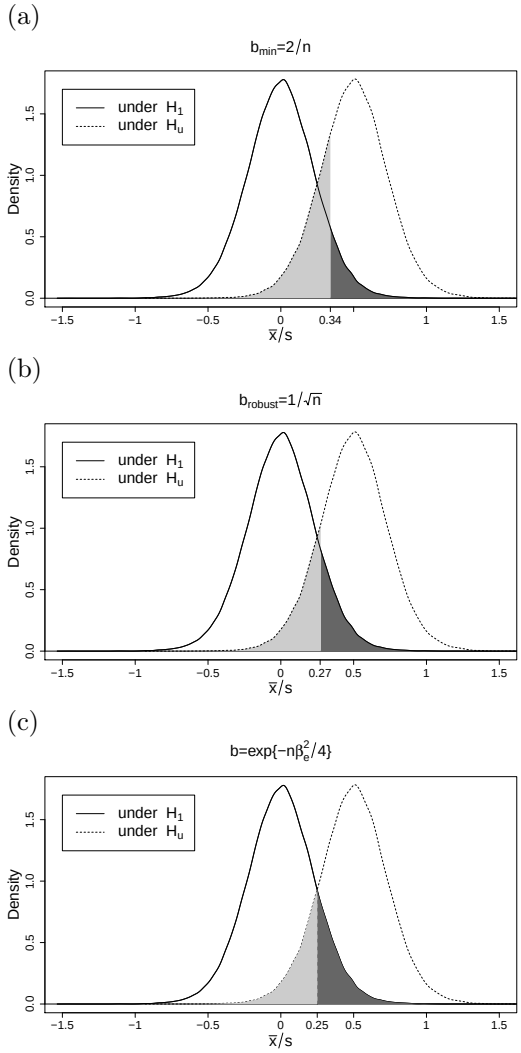

Figure 5.3: Sampling distributions of observed effect size $\bar{x}/s$ in one sample t test for $n = 20$ and $\beta_e = 0.5$ under H_u.

and

$$p_2 = P(AAFBF_{iu}^0 > 1|H_u) = P(\hat{\beta}_e < \sqrt{-\log b/n}|H_u). \tag{5.43}$$

The observed standardized effect size $\hat{\beta}_e$ is usually within the interval $[0, 1]$ for equality constrained hypothesis testing, because $\hat{\beta}_e$ can be interpreted analogously as the Cohen's d or Cohen's f^2 (Cohen, 1992), which rarely exceeds 1. First, for a one sample t test $x_i \sim N(\theta, \sigma^2)$, and $H_1 : \theta = 0$ versus $H_u : \theta$, the maximum likelihood estimate of $\beta = \theta$ is $\hat{\beta} = \bar{x}$ and the standard deviation is $\hat{\sigma}_\beta = s/\sqrt{n}$. Then the observed standardized effect size (5.41) becomes $\hat{\beta}_e = \frac{\hat{\beta}}{\hat{\sigma}_\beta}/\sqrt{n} = \frac{\bar{x}}{s}$ which is the same as Cohen's d. Second, we consider the F test of $H_2 : \theta_1 = 0$ versus $H_u : \theta_1$ in a simple linear regression model $y_i = \theta_0 + \theta_1 x_i + \epsilon_i$, where θ_0 is the intercept, θ_1 is the regression coefficient, and $\epsilon_i \sim N(0, \sigma^2)$ is the residual. The maximum likelihood estimate of $\beta = \theta_1$ is $\hat{\beta} = r_{xy}\frac{s_y}{s_x}$ and the standard deviation is $\hat{\sigma}_\beta = \frac{\sigma}{s_x}/\sqrt{n}$, where s_x and s_y are the standard derivations of x_i and y_i, and r_{xy} is the correlation coefficient between x_i and y_i. Note that r_{xy}^2 is equal to the coefficient of determination R^2 in the case of the simple linear regression model. Thus, because the coefficient of determination is equal to $R^2 = 1 - \sigma^2/s_y^2$, the observed standardized effect size in (5.41) becomes $\hat{\beta}_e = \frac{\hat{\beta}}{\hat{\sigma}_\beta}/\sqrt{n} = r_{xy}\frac{s_y}{\sigma} = \sqrt{\frac{R^2}{1-R^2}}$, which is the square root of Cohen's $f^2 = \frac{R^2}{1-R^2}$.

Analogous to the effect size $\bar{x}/s$ in the one sample t test, the observed standardized effect size $\hat{\beta}_e$ also has sampling distributions under H_i and H_u, which are symmetric around half of the pre-specified standardized effect size β_e under H_u. Therefore, by setting $\sqrt{-\log b/n} = \beta_e/2$ or equally

$$b = \exp(-n\beta_e^2/4), \tag{5.44}$$

the test for H_i against H_u using AAFBF has equal error probability:

$$p_1 = P(\hat{\beta}_e > \beta_e/2|H_i) = P(\hat{\beta}_e < \beta_e/2|H_u) = p_2. \tag{5.45}$$

How to specify β_e in (5.44) will be discussed in the next subsection.

A new rule of choosing b

Before presenting the new choice of b based on equal error probabilities, we need to deal with two issues: the range of b for consistent Bayes factors and the specification of standardized effect size β_e under H_u. The consistency of the Bayes factor is an important property in Bayesian hypothesis testing. The Bayes factor for $H_i : \boldsymbol{\beta} = 0$ against $H_u : \boldsymbol{\beta} \neq \mathbf{0}$ is consistent if it goes to infinity as sample size goes to infinity when H_i is true, and goes to 0 when H_u is true. Morey et al. (in press) found that the prior specification based on frequentist error probabilities may result in inconsistent Bayes factors. Gu et al. (in press) showed how to resolve this by restricting the

fraction b according to $b \geq (J+1)/n$ in the one sample t test. As stated earlier in Section 5.4.2, $b = (J+1)/n$ is based on the minimal number of observations to specify proper priors, and therefore we will always constrain $b \geq (J+1)/n$ in the AAFBF. Furthermore, we also suggest constraining $b \leq 1/2$ because $b > 1/2$ implies that more than half of the likelihood is used for prior specification, which is undesirable in Bayesian tests (Berger & Pericchi, 1996). Consequently, the range of the fraction b is set as $b \in [(J+1)/n, 1/2]$.

To obtain the fraction b in (5.44) for equal error probabilities, the standardized effect size β_e under H_u has to be specified. Given any specific β_e, a fraction b in (5.44) can be obtained such that $p_1 = p_2$. However, in practice β_e is unknown. Therefore, a distribution for β_e is specified that covers a range of realistic effect sizes, i.e., $\beta_e \in [0, 1]$ as elaborated before. Here we consider a uniform distribution $\pi^*(\beta_e) = U(0, 1)$ in which every effect size from small to large is equally likely within the interval $[0, 1]$ (Gu et al., in press). Note that this choice for b would be the same as when using $\pi^*(\beta_e) = U(-1, 1)$ because the choice of b is independent of the sign of the effect.

Based on the distribution of effects $\pi^*(\beta_e) = U(0, 1)$, the third choice of fraction b for equal error probabilities is given by:

$$b_{freq} = E_{\pi^*(\beta_e)}[\exp(-n\beta_e^2/4)] = \int_0^1 \exp(-n\beta_e^2/4) d\beta_e. \tag{5.46}$$

The integration in (5.46) can be numerically calculated (see Gu et al., in press). Although b_{freq} cannot always achieve equal error probabilities as we constrain $b \in [(J+1)/n, 1/2]$ and specify $\pi^*(\beta_e) = U(0, 1)$, Gu et al. (in press) show that this choice results in error probabilities that are often about equal for the one sample t test. It was shown that the difference between the type I and type II error probabilities was typically smaller for this choice than when using the more traditional choices for b. We recommend the choice b_{freq} when the sample size is small, because in this case the error probabilities p_1 and p_2 are relatively large and difference between p_1 and p_2 can be quite severe. In the following subsection, we will discuss the sensitivity of AAFBF based on different choices of b.

5.4.4 Sensitivity to prior distributions

In Section 5.3, we specified the normal prior (5.24) for $\boldsymbol{\beta}$ in general statistical models. However, the adjusted fractional prior for the parameters in a specific model is often not normally distributed. Thus, when using a normal approximation of the fractional prior, as in the case of the AAFBF, we may misspecify the prior distribution for the parameters of interest. For example, if the parameter is a probability which is bounded in $[0, 1]$ in a binomial model, the (implicit) fractional prior would have a beta distribution. Therefore the use of the AAFBF, where the fractional prior is approximated using a normal distribution, may be different from the non-approximated

adjusted fractional Bayes factor. Thus, it is useful to investigate the sensitivity of the AAFBF when the fractional prior is far from normally distributed.

O'Hagan (1995) argued that the sensitivity of the fractional Bayes factor depends on the magnitude of the fraction b. Increasing b reduces the sensitivity to the distributional form of the fractional prior. This is also the case for the adjusted fractional Bayes factor (AFBF) of Mulder (2014b), because a larger fraction b implies that more information in the data is used for prior specification, which makes the distribution of the adjusted fractional prior in the AFBF more similar to a normal distribution. This section will use two simple examples to illustrate how much difference there is between the AAFBF using the normal prior and the AFBF using the actual fractional prior. In these examples, we will only focus on equality constrained hypotheses because, as elaborated earlier, the AFBF for inequality constrained hypotheses is independent of the fraction b.

The first example again concerns the one sample t test, where data come from $x_i \sim N(\theta, \sigma^2)$ with unknown mean and variance, and the hypotheses under consideration are $H_1 : \theta = 0$ against $H_u : \theta$. In the AAFBF, the default prior (5.24) for $\beta = \theta$ is $\pi_u^*(\beta|X^b) = N(0, s^2/nb)$, while the actual adjusted fractional prior for a normal mean has a t distribution $\pi_u^*(\beta|X^b) = t(0, s^2/(nb-1), nb-1)$ with mean of 0, variance of $s^2/(nb-1)$, and degrees of freedom of $nb-1$. It is well known that the t distribution has heavier tail than the normal distribution such that the density at the mode $\beta = 0$ from the normal distribution is larger than the density from the t distribution. Furthermore, as the fraction b increases, the degrees of freedom $nb-1$ increase such that the t distribution $t(0, s^2/(nb-1), nb-1)$ becomes more similar to the normal distribution $N(0, s^2/nb)$. This implies that for a larger b the AAFBF where the default prior has a normal distribution performs more similarly as the AFBF under the actual fractional prior. This is illustrated in Figure 5.4.

Figure 5.4 shows the logarithms of AFBFs and AAFBFs for H_1 versus H_u under different observed effect sizes $\bar{x}/s = 0, 0.1, 0.2$, and different fractions b_{min}, b_{robust}, and b_{freq}. The sample size n varies from 10 to 500. First, as can be seen in Figure 5.4 (a), based on b_{min} the logarithms of AAFBFs under the normal prior distribution (dashed line) differ substantially from the logarithms of AFBFs under the t prior distribution (solid line). This difference does not decrease as n increases because when setting $b_{min} = 2/n$ the degree of freedom in the t distribution is $nb - 1 = 1$, which is independent of n. This suggests high sensitivity to the functional form of the prior distribution. Second, Figure 5.4 (b) shows that based on b_{robust} there is not much difference between the logarithms of AAFBFs and AFBFs. This implies that the choice of b_{robust} results in less sensitivity to the functional form of the prior distribution than b_{min}. Third, Figure 5.4 (c) demonstrates the logarithms of AAFBFs and AFBFs under b_{freq}. As can be seen, with b_{freq} there is no sensitivity either.

It is interesting to note that Figure 5.4 also illustrates the consistency of AAFBFs. The consistency in this example requires that as sample size goes to infinity the AAFBF for H_1 against H_u approaches to infinity when the observed effect size is

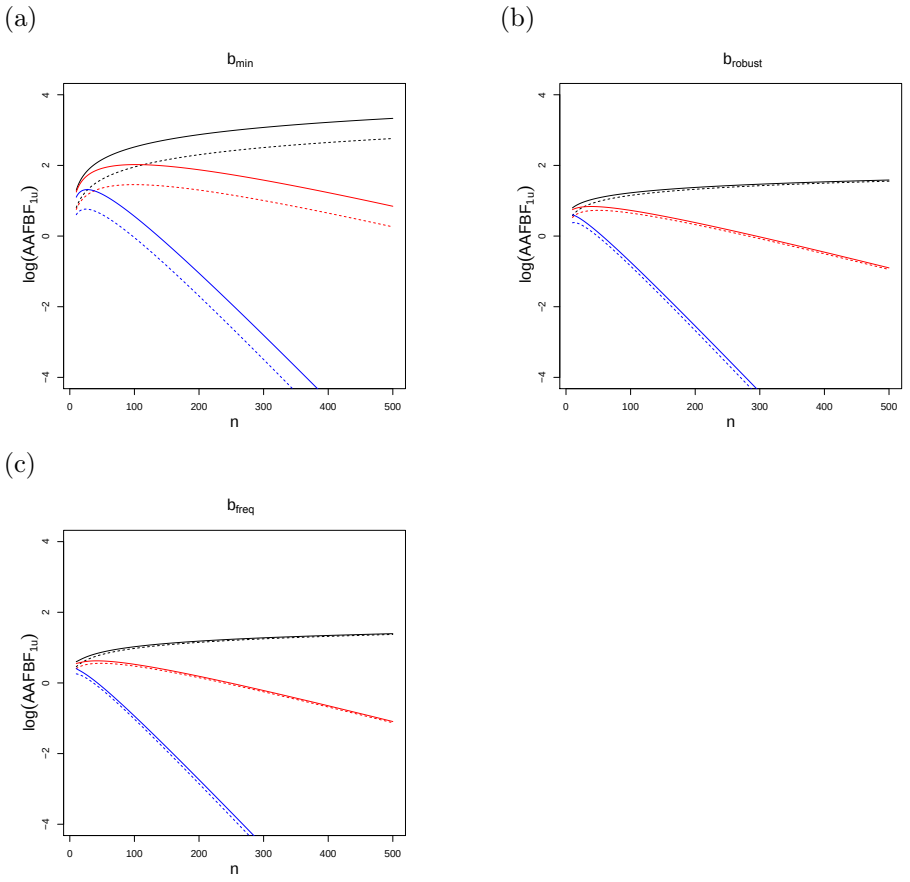

Figure 5.4: The logarithms of the AFBF with a Student t prior (solid line) and the AAFBF with a normal prior (dashed line). The dark, red, and green lines correspond to the logarithms of Bayes factors under observed effect sizes $\bar{x}/s = 0$, 0.1, and 0.2, respectively.

equal to 0 and the AAFBF goes to zero when the observed effect size is unequal to 0. As can be seen in Figure 5.4, for an observed effect size $\bar{x}/s = 0$ the logarithm of the AAFBF (black lines) in each figure goes to infinity as sample size n increases. Conversely, the logarithms of the AAFBF based on an observed effect size of $\bar{x}/s = 0.1$ (red lines) and $\bar{x}/s = 0.2$ (blue lines) diverge to minus infinity, which implies decisive evidence for the true unconstrained hypothesis as the sample size goes to infinity.

Next, we consider a binomial model, where data come from $x \sim Bin(n, p)$. The hypotheses under evaluation are $H_2 : p = 0.4$ against $H_u : 0 \leq p \leq 1$. Since H_2 is nested in H_u, we can use the AAFBF (5.26) to evaluate H_2 against H_u. Given data $x \sim Bin(n, p)$, the estimate of $\beta = p - 0.4$ is $\hat{\beta} = x/n - 0.4$ and the variance is $\hat{\sigma}^2_\beta = \frac{x(n-x)}{n^2(n+1)}$, and therefore the normal adjusted fractional prior (5.24) is $\pi^*_u(\beta|X^b) = N(0, \frac{bx(n-x)}{n^2(n+1)})$. On the other hand, following the idea of adjusted fractional Bayes factors the fractional prior has a beta distribution, i.e., $p = \beta + 0.4 \sim Beta(0.4nb, 0.6nb)$ which has a mean of 0.4 and thus β has a prior mean of 0. Note that this prior is centered on the focal point of 0.4 in H_2.

Figure 5.5 draws the lines of the logarithms of the AFBFs and AAFBFs for H_2 against H_u as the sample size n increases from 10 to 500. The observed data are $x = 0.4n, 0.5n, 0.6n$. As can be seen in Figure 5.5 there is a considerable smaller approximation error of the AAFBF with respect to the AFBF in comparison to the first example in Figure 5.4. Again, the difference is largest for b_{min} because this fraction is always smaller than b_{robust} and b_{freq}.

These two examples include the evaluation of equality constrained hypotheses in both continuous data and discrete data. Although the models used are simple, the results of the sensitivity study of adjusted fractional Bayes factors can be applied in the multivariate normal model where the parameters (e.g., the group means in ANOVA model, the coefficients in regression model) have a multivariate t distribution, and in multinomial model where the parameters (e.g., the probabilities in Contingency Tables) have a Dirichlet distribution which is the multivariate generalization of the Beta distribution. Furthermore, in more complicated settings such as structural equation models and generalized linear models, it can be anticipated that the larger b will result in less sensitive AFBFs because this implies that more data are used to specify the fractional prior such that the normal approximation to the prior has better performance based on the large sample theory.

Based on the discussion in this section, we propose the following scheme for specifying the fraction b in the AAFBF.

- Choose $b_{min} = (J + 1)/n$ to have a default prior that is based on the idea of a minimal training sample.

- Choose $b_{robust} = \max\{(J + 1)/n, \sqrt{n}/n\}$ to ensure that the default prior is close to normal.

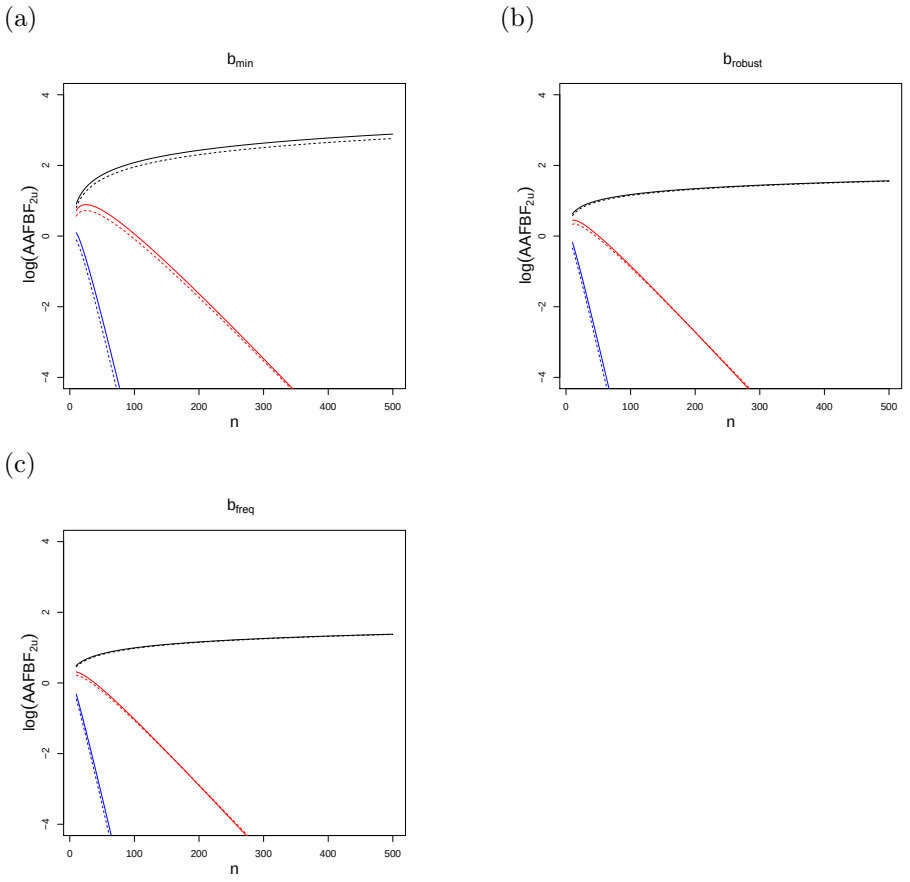

Figure 5.5: The logarithms of the AFBF with a Beta prior (solid line) and the AAFBF with a normal prior (dashed line). The dark, red, and green lines correspond to the logarithms of Bayes factors under observed effect sizes $\bar{x}/s = 0.4n$, $0.5n$, and $0.6n$, respectively.

Table 5.3: Result for regression model example

	$b_{min} = 0.080$	$b_{robust} = 0.141$	$b_{freq} = 0.216$
$AAFBF_{1_c}$	6.04	4.46	3.55

- Choose $b_{freq} = \int_0^1 \exp\left(-n\beta_e^2/4\right) d\beta_e$ to control the frequentist error probabilities when testing an equality constrained hypothesis against the unconstrained alternative.

Note that n and J denote the sample size and the number of independent constraints for all the informative hypotheses, respectively.

5.5 Results for empirical examples

The examples introduced in Section 5.2 are revisited to illustrate how the AAFBF can be used to evaluate informative hypotheses. In the regression model, three parameters with respect to the regression coefficients are considered in the informative hypothesis $H_1 : \theta_1 > 0, \theta_2 < 0, \theta_3 = 0$. The first step is to specify the prior and posterior distributions in (5.24) and (5.25), which needs the estimates $\hat{\boldsymbol{\theta}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_\theta$ of the parameters. These can be obtained by analyzing the regression model with the data in Table 5.1 using a number of statistical software (packages), such as Mplus (Muthén & Muthén, 2010) and R package lavaan (Rosseel, 2012). Note that we do not need to standardize the three coefficients as they are compared with zero. The analysis of data in lavaan renders the maximum likelihood estimates of the parameters, i.e., $\hat{\theta}_1 = 11.01$, $\hat{\theta}_2 = -2.85$, $\hat{\theta}_3 = -2.03$, and the covariance matrix:

$$\hat{\boldsymbol{\Sigma}}_\theta = \begin{bmatrix} 18.236 & -0.500 & 2.812 \\ -0.500 & 0.043 & -0.004 \\ 2.812 & -0.004 & 4.481 \end{bmatrix}.$$

To obtain the AAFBF for H_1 against H_{1_c}, the fraction b has to be specified. Based on the sample size of $n = 50$ and the length of vector $\boldsymbol{\beta}$ of $J = 3$ in this example, the three choices of fraction are $b_{min} = 0.080$, $b_{robust} = 0.141$, and $b_{freq} = 0.216$. Running `BaIn` with the estimates and covariance matrix of parameters of interest renders the AAFBF displayed in Table 5.3. As can be seen, $AAFBF_{1_c}$ is larger than 3 under each choice of b, which implies positive evidence in the data for H_1 against H_{1_c} according to Kass and Raftery (1995)'s rule.

The hypotheses in the repeated measures ANOVA model consists of four parameters of which the estimates are $\hat{\theta}_1 = 22.33$, $\hat{\theta}_2 = 22$, $\hat{\theta}_3 = 5.78$ and $\hat{\theta}_4 = 6.78$, and

the covariance matrix is

$$\hat{\boldsymbol{\Sigma}}_\theta = \begin{bmatrix} 5.18 & 4.86 & 2.61 & 2.86 \\ 4.86 & 5.13 & 2.90 & 3.03 \\ 2.61 & 2.90 & 1.93 & 1.97 \\ 2.86 & 3.03 & 1.97 & 2.39 \end{bmatrix}.$$

Given sample size $n = 36$ and length of vector $\boldsymbol{\beta}$ of $J = 3$, three choices of b are automatically specified in `BaIn` as $b_{min} = 0.111$, $b_{robust} = 0.167$, and $b_{freq} = 0.255$. Based on these specification `BaIn` renders the AAFBFs $AAFBF_{2u}$ for H_2 versus H_u and $AAFBF_{2'u}$ for $H_{2'}$ versus H_u. The results are shown in Table 5.4. As can be seen, $AAFBF_{2'u}$ is independent of b because the AAFBF for inequality constrained hypotheses is invariant to the choice of the fraction b. Thereafter, the AAFBF $AAFBF_{22'}$ for H_2 versus $H_{2'}$ can be computed by $AAFBF_{2u}/AAFBF_{2'u}$ which is shown in the last row in Table 5.4. The result of $AAFBF_{22'}$ in the last row suggests positive evidence in the data for H_2 against $H_{2'}$.

5.6 Conclusion

This paper presented a new approximate Bayesian procedure for the evaluation of informative hypotheses that can be used for virtually any model. The methodology is based on the prior adjusted default Bayes factor of Mulder (2014b). Furthermore, normal approximations were used to ensure fast computations. Numerical results showed that the approximation is close to the prior adjusted fractional Bayes factor. This implies that the proposed AAFBF provides an accurate quantification of the relative evidence between informative hypotheses. Furthermore, different choices were given for the fraction b, similar as in the fractional Bayes factor of O'Hagan (1995). The first choice relies on the concept of priors containing minimal information. The second choice uses a robustness argument resulting in a default prior distribution that is close to normal. The third choice is based on a frequency argument to control the classical error probabilities. The choice can be made by the user depending on the property which he/she finds most important. By computing the AAFBF for each choice of b we get a complete picture how much support there is in the data between two hypotheses when taking into account different philosophies.

Table 5.4: Result for repeated measures ANOVA example

	$b_{min} = 0.111$	$b_{robust} = 0.167$	$b_{freq} = 0.255$
$AAFBF_{2u}$	4.60	3.07	2.01
$AAFBF_{2'u}$	0.24	0.24	0.24
$AAFBF_{22'}$	19.2	12.8	8.38

We provide a software package `BaIn` with a user manual in the Appendix to evaluate the informative hypotheses which only needs the maximum likelihood estimates and covariance matrix of the parameters of interest, denoted by $\boldsymbol{\theta}$ in this paper. `BaIn` computes the AAFBF for an informative hypothesis against an unconstrained hypothesis. By computing these quantifies for each informative hypothesis against the unconstrained hypothesis we can straightforwardly compute the relative support in the data for pairs of informative hypotheses.

5.A User manual of BaIn

The software package BaIn is developed in Fortran 90 with the IMSL 5.0 numerical library. It computes Bayes factors to evaluate any informative hypotheses (Section 5.2) and compare pairs from a set of informative hypotheses if they are comparable (Section 5.3.4). BaIn can be freely downloaded from the website `http://informative-hypotheses.sites.uu.nl/software/bain/`. The downloaded folder consists of an executable file "BaIn.exe", an input file "Input.txt", and an output file "Output.txt". Running "BaIn.exe" with "Input.txt" located in the same folder renders "Output.txt". This appendix instructs researchers to fill in the "Input.txt" such that "BaIn.exe" can properly read the information. The "Input.txt" mainly contains the estimates and covariance matrix of parameters $\boldsymbol{\theta}$ for prior and posterior specification, and the restriction matrix and constant vector for each informative hypothesis.

The repeated measures ANOVA example in Section 5.2.2 is used to illustrate the valid specification of input file. We will first display and then explain the context below written in the "Input.txt" when evaluating informative hypothesis H_2 (5.11) and $H_{2'}$ (5.12).

```
1    #Number of parameters of interest; Number of informative
      hypotheses; Sample size
2    4 2 36
3    #Estimates of parameters
4    22.33 22 5.78 6.78
5    #Covariance matrix of parameters
6    5.18 4.86 2.61 2.86
7    4.86 5.13 2.90 3.03
8    2.61 2.90 1.93 1.97
9    2.86 3.03 1.97 2.39
10   #Numbers of equality and inequality constraints in H1
11   2 1
12   #Restriction matrix (R0|r0) for equality constraints
13   1 -1  0  0   0
14   0  0  1 -1   0
```

```
15   #Restriction matrix (R1|r1) for inequality constraints
16   0  1 -1  0   0
17   #Numbers of equality and inequality constraints in H2
18   0 3
19   #Restriction matrix (R0|r0) for equality constraints
20   #Restriction matrix (R1|r1) for inequality constraints
21   1 -1  0  0   0
22   0  1 -1  0   0
23   0  0  1 -1   0
```

The input text has strictly fixed structure. There are annotation lines starting with # below which the corresponding information (numbers) has to be given. The first line is the annotation for the number of structural parameters, number of informative hypotheses, and sample size, which means we need to write three numbers in the second line, i.e., 4, 2 and 9. Because the number of structural parameters is 4, four numbers for the estimates of parameters are presented in line 4, and a 4×4 covariance matrix is written in lines 6 to 9. Furthermore, because the number of informative hypotheses is 2, two hypotheses are specified. For the first hypothesis, line 11 specifies 2 and 1 for the numbers of equality and inequality constraints, respectively. Therefore, the augmented restriction matrix with constant vector for equality constraints has two rows shown in lines 13 and 14, and one row for inequality constraints in line 16. For the second hypothesis, the numbers of equality and inequality constraints are 0 and 3 given in line 18, respectively. As can be seen, there is not a line with numbers below the annotation line 19 `#Restriction matrix (R0|r0) for equality constraints` because this hypothesis does not contain any equality constraint. While from lines 21 to 23 the augmented restriction matrix for three inequality constraints are displayed.

The estimates and covariance matrix of structural parameters can be obtained from other statistical software, e.g., Mplus (Muthén & Muthén, 2010) and R package lavaan (Rosseel, 2012), and the augment restriction matrix `(R0|r0)` and `(R1|r1)` can be specified based on the informative hypotheses under evaluation. Executing "BaIn.exe" with these information renders the relative complexities, fits, and Bayes factors for informative hypotheses under different choices of fraction b in the "Output.txt". The results for repeated measures ANOVA example is shown as follows.

```
Result for H1

Equality constraints
   Fit    Complexity (b1)   Complexity (b2)   Complexity (b3)
  0.091        0.049             0.059             0.096

Inequality constraints (conditional on equality constraints)
```

```
   Fit    Complexity (b1)   Complexity (b2)   Complexity (b3)
 1.000         0.500             0.500             0.500
Number of iterations
  3000          3000              3000              3000

BF1u (b1=0.111)    BF1u (b2=0.167)    BF1u (b3=0.255)
     4.603              3.069              2.006

BF1c (b1=0.111)    BF1c (b2=0.167)    BF1c (b3=0.255)
     4.603              3.069              2.006


Result for H2

Equality constraints
   Fit    Complexity (b1)   Complexity (b2)   Complexity (b3)
 1.000         1.000             1.000             1.000

Inequality constraints (conditional on equality constraints)
   Fit    Complexity (b1)   Complexity (b2)   Complexity (b3)
 0.023         0.096             0.098             0.097
Number of iterations
  46000          9000              9000              9000

BF2u (b1=0.111)    BF2u (b2=0.167)    BF2u (b3=0.255)
     0.240              0.237              0.238

BF2c (b1=0.111)    BF2c (b2=0.167)    BF2c (b3=0.255)
     0.223              0.219              0.220
```

The results contain the relative fits and complexities for both equality and inequality constraints, as well as the Bayes factors under different fraction b in each hypothesis. For equality constraint, the relative fit and complexity are the normal posterior and prior densities in (5.26), and thus can be directly computed. However, the computation of relative fit and complexity for inequality constraints is often difficult and needs to sample from the posterior and prior distributions using Monte Carlo Markov Chain methods (Gu et al., 2014). BaIn uses an efficient algorithm presented in Chapter 3, which requires less number of iterations (displayed below fit and complexities) in the Markov chains to accurately estimate the relative fit and complexity. Note that the Bayes factor for informative hypotheses H_1 against H_2 can be computed using (5.32) with BF_{1u} and BF_{2u}.

Chapter 6

An n-of-one RCT for intravenous immunoglobulin G for inflammation in hereditary neuropathy with liability to pressure palsy (HNPP)[1]

6.1 Background

Hereditary neuropathy with liability to pressure palsy (HNPP; tomaculous neuropathy) is an autosomal dominant disorder caused by a loss of function of the gene for peripheral myelin protein 22 (PMP22; OMIM #601097) on chromosome 17.p12. HNPP is a rare disorder, with an estimated prevalence of two to five per 100,000 (Bird, 1998). Symptoms usually start in the second or third decade of life and consist of recurrent painless episodes of focal sensory loss and muscle weakness (palsy) in the distribution of a peripheral nerve. Episodes are often provoked by compression of the nerve and resolve spontaneously within days to months (Dubourg, Mouton, Brice, LeGuernb, & Bouchea, 2000; Stögbauer, Young, Kuhlenbäumer, de Jonghe, & Timmerman, 2000; Mouton et al., 1999). There is no curative treatment; management consists of supportive measures to prevent nerve compression, and bracing to

[1] A short version of this chapter has been published as a letter as Vrinten, C., Gu, X., Weinreich, S., Schipper, M., Wessels, J., Ferrari, M., Hoijtink, H., & Verschuuren, J. (2015). An n-of-one RCT for intravenous immunoglobulin G for inflammation in hereditary neuropathy with liability to pressure palsy (HNPP). *Journal of Neurology, Neurosurgery & Psychiatry.* doi:10.1136/jnnp-2014-309427.
Author contributions: JV, JW and MF conceived the study and carried out the data collection. JV, CV, SW, HH and XG formulated the informative hypotheses which were evaluated by HH and XG. JV and CV drafted the paper, with assistance from MS. All authors provided extensive feedback on writing the paper.

alleviate muscle weakness.

In this report, we describe the case of a female patient with HNPP who initially presented with symptoms of a painful neuropathy which were successfully treated with intravenous immunoglobulin (IVIg), as well as the results of a subsequent placebo-controlled n-of-one randomised controlled trial (RCT) that was conducted to formally assess the effects of IVIg on pain and muscle strength and the need for continued treatment with IVIg.

6.1.1 Case report

In 2002, a 35-year-old female patient presented to the Leiden University Medical Centre Neurology Clinic with a 15-month history of neuropathic pain in the right gluteal region that radiated via the back of the leg to the right foot. Four months before presentation, she had experienced weakness and sensory loss in the lower left leg after a prolonged car journey, but this resolved spontaneously after several weeks. Two months later, she experienced more severe weakness and sensory loss: she was unable to lift her left leg when lying prone and also experienced numbness in her left hand. No triggering events were reported for this episode. Her medical history was unremarkable, and there were no family members with similar symptoms.

Her physical examination at the time of presentation showed mild proximal weakness of the left leg (MRC 4) and severe weakness (MRC 0-2) of the left foot extensor muscles. Hypoalgesia was found in the ulnar side of the left hand and the left lower leg. She had reduced tendon reflexes; Achilles tendon reflexes were completely absent. The following examinations were normal or negative: lumbar MRI, cerebrospinal fluid analysis, serum anti-GM1, and serology for cytomegalovirus, Epstein-Barr virus, mycoplasma, and Borrelia burgdorferi. Faecal tests for Salmonella, Shigella and Campylobacter were also negative. A nerve biopsy was not performed.

Electromyographic studies showed bilateral demyelinating conduction blocks at compression sites of the ulnar nerves, prolonged distal motor latencies of the right and left-sided ulnar, tibial, peroneal, and left median nerves, and absent F-waves in both peroneal and the right tibial nerves, consistent with HNPP, but also with definite electrodiagnostic criteria for chronic inflammatory demyelinating polyneuropathy (CIDP) according to the EFNS/PNS CIDP guidelines (van den Bergh et al., 2010). Based on these results, a preliminary diagnosis of CIDP was made and a DNA test for suspected HNPP was ordered.

She was treated with IVIg (0.4 mg/kg per day) for five days, which resulted in marked improvement: after three weeks she was able to do domestic chores again for the first time in a year. She continued to receive a maintenance dose of IVIg every three weeks, and her muscle strength continued to improve. The pain disappeared completely and she only suffered residual mild weakness of left foot dorsiflexion (MRC4). However, DNA analysis subsequently revealed a deletion of 17p11.2 including the PMP22 gene, and a definite diagnosis of HNPP was made.

6.1.2 Rationale for n-of-one trial

It remains debated whether genetic neuropathies can give rise to superimposed immune-mediated neuropathies (Sagnelli, Piscosquito, & Pareyson, 2013), and the diagnosis of HNPP raised doubts whether continued IVIg was needed, especially given its high cost and limited availability (Kuitwaard & van Doorn, 2009). In light of this ambiguity, the patient consented to a formal assessment of the effects of IVIg in an n-of-one trial. This is a multiple crossover trial in a single patient in which intervention and control treatment periods are randomised over time (e.g. AB-BA-BA). It is suitable to evaluate the effects of relatively fast-acting, symptomatic treatment for chronic and relatively stable disease symptoms in individual patients (J. Nikles et al., 2009; Guyatt et al., 1988).

By means of the n-of-one trial, we aimed to evaluate the effects of IVIg on pain (primary outcome) and muscle strength (secondary outcome) in this patient with HNPP and an associated CIDP-like inflammatory neuropathy. The following hypotheses were tested: a) IVIg infusions reduce pain more than placebo infusions and this reduction is clinically meaningful; b) IVIg infusions increase subjective muscle strength more than placebo infusions and this increase is clinically meaningful. To assess the need for continued use of IVIg, we also tested the following hypotheses: c) following IVIg, pain levels first decrease and then increase again; and finally, d) following IVIg, subjective muscle strength first increases and then decreases again.

6.2 Methods

6.2.1 Trial design

We conducted a double-blind, multiple crossover n-of-one trial of four trial infusions that were given in hospital on an outpatient basis and in a randomised order at three week intervals. The intervention treatment consisted of intravenous immunoglobulin (0.4 mg/kg) and was compared to an inactive placebo infusion of 0.9% saline. A placebo infusion was chosen as comparator, because there is currently no pharmacotherapy for HNPP. The patient consented to participate in this study as a way to optimise her personal long-term clinical treatment.

A week after each trial infusion, an optional "rescue" infusion with the opposite treatment was offered (i.e. placebo if IVIg had been administered most recently and vice versa). The patient could accept or refuse this rescue infusion depending on her subjective assessment of the effects of the trial infusion (see Figure 6.1). The rescue infusion was offered to ensure that the most beneficial treatment was not withheld for more than a week after it was due according to her pre-trial 3-weekly treatment schedule. An open run-in period had shown that a 1-week delay in administering IVIg was not associated with unacceptable muscle weakness or pain. If the patient

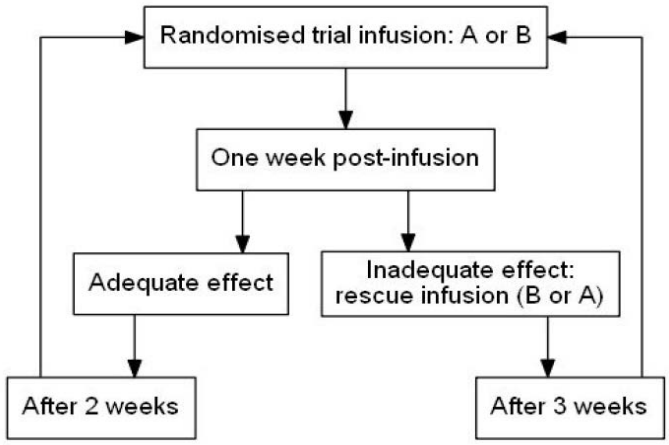

Figure 6.1: Flow diagram of one cycle of the n-of-one trial with optional rescue infusion one week after each trial infusion. Note that the interval between the last infusion (i.e. 'trial' if deemed adequate 1 week post-infusion, or 'rescue' if trial infusion is deemed inadequate 1 week post-infusion) and the next trial infusion is held constant at 3 weeks.

opted to have the rescue infusion, she returned to the randomisation schedule 3 weeks later.

Simple randomisation was carried out by the dispensing hospital pharmacy, which was also responsible for blinding of treatment by delivering all infusion packs to the hospital infusion room wrapped in opaque tin foil. This ensured that the patient and clinician remained blind to the order of the trial infusions, although both were aware that the rescue infusion was always the opposite one to the trial infusion given the week before.

6.2.2 Outcomes and data collection

Pain was chosen as primary outcome measure and muscle strength as secondary outcome measure. Pain scores for the right leg, which had always been most affected by pain, were recorded by the patient three times per week in a patient diary at home. Pain was scored on a 14 cm visual analogue scale (VAS) ranging from 0 (indicating complete absence of pain) to 14 (worst possible pain imaginable). Ratings were converted into scores in millimetres, from which the percentage change from baseline was calculated. A clinically meaningful reduction in pain was defined as a 30% reduction compared to the baseline level of pain at the time of the last infusion.

A reduction of this magnitude was previously found to correspond to 'some' to 'much' change in pain, and is associated with not needing rescue medication for chronic pain (Dworkin et al., 2008).

Analogous to pain, subjective muscle strength was recorded three times a week on a 14 cm VAS scale (0 = complete paralysis to 14 = normal strength for this patient). This was done for the left leg, which was most affected by weakness. Ratings were converted into scores in millimetres, from which the percentage change from baseline was calculated. No reference values were available from the literature and we chose to define a clinically meaningful difference in muscle strength as an increase of at least 30% compared to baseline. Finally, at the time of each infusion, the patient was asked to report any side effects since the last infusion.

6.2.3 Data analysis

The effect of IVIg on pain and subjective muscle strength was assessed for the first 7 days after each infusion only (not longer because rescue infusions were offered 7 days after each randomised infusion). To assess the need for continued administration of IVIg every 3 weeks, the course of pain and subjective muscle strength was evaluated over the course of three weeks following IVIg. Coefficients were first estimated using SPSS version 20.0 (Dworkin et al., 2008), followed by Bayesian evaluation of informative hypotheses using BIG (Gu et al., 2014). Bayesian hypothesis testing allowed us to evaluate the inequality constrained hypotheses we had formulated regarding the magnitude of the increase/decrease following IVIg and placebo (Gu et al., 2014). We compared the inequality constrained hypotheses that IVIg was superior to placebo to an unconstrained hypothesis which did not specify a relationship between the magnitude of the effect following IVIg and placebo infusions. For each comparison, a Bayes factor was calculated, which is a measure of support for two competing hypotheses. A Bayes factor of 1 indicates that the data support both hypotheses equally. In the present study, a Bayes factor of more than 1 indicates that our (inequality constrained) hypotheses are more supported by the data than the unconstrained hypothesis, while a Bayes factor of less than one indicates the reverse. Conventionally, Bayes factors larger than 10 would denote strong support for the inequality constrained hypothesis (Jeffreys, 1961). A detailed description of the analyses is provided in Appendix 6.A, and the data archive is provided on the Web at `http://jnnp.bmj.com/content/suppl/2015/07/17/jnnp-2014-309427.DC1/jnnp-2014-309427supp3.zip`.

6.3 Results

The total number of infusions given during the trial was eight, but there were reasons to exclude data from one infusion for the analyses (The reason for excluding data from one infusion is that the trial partly took place over the summer and the patient

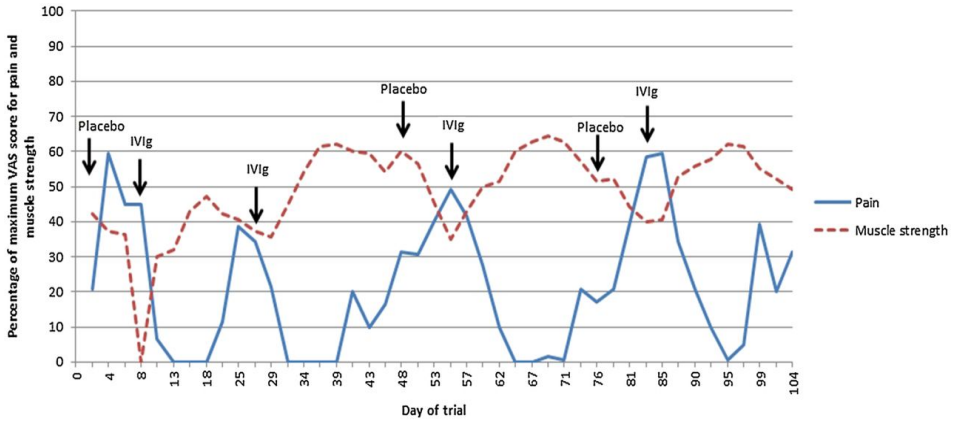

Figure 6.2: Trial timeline, administered infusions and VAS scores for pain and subjective muscle strength (IVIg, intravenous immunoglobulin G; VAS, visual analogue scale).

requested to receive one non-randomised and non-blinded IVIg infusion before her summer holiday. The results of this infusion were not used in the analyses; data for this infusion are not shown in tables or graphs). Four infusions were given according to the randomisation schedule: one IVIg and three placebo infusions. After each placebo infusion, the patient opted for a rescue infusion with the alternative treatment one week later; she did not ask for a rescue infusion after the randomised IVIg infusion. The total duration of the trial was 15 weeks. The timeline of the trial and VAS scores for pain and self-reported muscle strength are shown in Figure 6.2. The results presented below can be found in Appendix 6.A.

6.3.1 Pain

We first tested the expectation that the decrease in pain in the first 7 days after IVIg is greater than after placebo. We obtained a Bayes factor of 33.22 when we compared this hypothesis to the unconstrained hypothesis, providing strong evidence that IVIg reduces pain more than placebo. We obtained a Bayes factor of 13.40 when we compared the hypothesis that IVIg produces a clinically relevant reduction in pain ($\geq 30\%$) to the unconstrained hypothesis, which implies that there is strong support for the hypothesis that IVIg reduces pain.

When these hypotheses were combined in a single hypothesis, i.e. pain decreases more rapidly after IVIg than after placebo in the first week after infusion, and it decreases to a clinically relevant level, and evaluated against the unconstrained hy-

pothesis, we obtained a Bayes factor of 63.74. This strongly supported the hypothesis that IVIg has a clinically meaningful effect on pain, compared to placebo. Estimates and variances of the coefficients are shown in Table 6.1 in Appendix 6.A.

6.3.2 Subjective muscle strength

We assessed the effects on subjective muscle strength in a similar fashion. We obtained a Bayes factor of 36.24 when we compared the hypothesis that the subjective increase in muscle strength in the first 7 days after IVIg infusion would be greater than after placebo to the unconstrained hypothesis. This implies that there is strong evidence that IVIg increases muscle strength more than placebo. We then assessed whether the increase in subjective muscle strength could be considered meaningful, as expressed by a 30% increase in subjective muscle strength compared to the baseline muscle strength score for each infusion. We obtained a Bayes factor of 15.05 for the hypothesis that IVIg produces a clinically relevant increase in muscle strength ($\geq$ 30%) when compared to the unconstrained hypothesis. This implies that there is strong evidence that IVIg increases muscle strength.

When these hypotheses were combined, i.e. muscle strength increases more rapidly and to a clinically relevant level in the first week after IVIg than after placebo, and evaluated against the unconstrained hypothesis, we obtained a Bayes factor of 61.51. This strongly supported the hypothesis that IVIg has a clinically meaningful effect on subjective muscle strength, compared to placebo.

6.3.3 Course of pain and muscle strength

Finally, to assess the need for regular IVIg infusions, we used quadratic models to test the hypotheses that pain first decreases and then increases again, and that muscle strength first increases and then decreases, in the three weeks following IVIg. The Bayes factor for the hypothesis about pain was 13.78, and the Bayes factor for the hypothesis about muscle strength was 15.67. These findings strongly support the notion that IVIg needs to be administered regularly to control pain and improve muscle strength. No adverse effects were reported during the trial.

6.3.4 Follow-up

We have now followed up this patient for 11 years. After the trial, she first continued to receive IVIg infusions every three to four weeks for two years, without any adverse effects. The interval was then successfully increased to five weeks. After a period of symptom stability, we attempted to give infusions every six weeks. However, this was followed by an increase in muscle weakness and pain, and the interval was reduced again to five weeks. Multiple EMGs during follow-up (2003-2014) initially showed signs of demyelination (prolonged distal motor latencies and decreased nerve

conduction times), but over the years became more consistent with stable axonal damage. The patient's quality of life has remained stable: her muscle strength is stable, and she continues to work in the same job.

6.4 Discussion

The results of this trial suggest that IVIg had a clinically meaningful effect on pain and weakness in this patient with HNPP. The positive effects of IVIg diminished after several weeks, necessitating continued treatment with regular IVIg infusions every few weeks for a sustained clinical response.

Our findings lend support to the growing number of case reports suggesting that some patients with hereditary neuropathies such as HNPP, Charcot-Marie-Tooth disease, and hereditary brachial plexus neuropathy may also be affected by inflammation (Mouton et al., 1999; Korn-Lubetzki, Argov, Raas-Rothschild, Wirguin, & Steiner, 2002; Remiche, Abramowicz, & Mavroudakis, 2013; Luigetti, Zollino, Conti, Romano, & Sabatelli, 2013; Pou Serradell et al., 2002; Le Forestier et al., 1997; Ginsberg et al., 2004; Marques et al., 2010; Desurkar et al., 2009; Watanabe et al., 2002; Klein et al., 2002). Like our patient, most of these patients initially presented with clinical and electrophysiological findings suggestive of an acute or chronic inflammatory demyelinating polyneuropathy (AIDP or CIDP), but were later diagnosed with an hereditary neuropathy. Some also responded favourably to immunomodulatory treatment with steroids or intravenous immunoglobulin (IVIg) (Korn-Lubetzki et al., 2002; Remiche et al., 2013; Le Forestier et al., 1997; Ginsberg et al., 2004; Watanabe et al., 2002). Although the co-occurrence of inflammatory and hereditary neuropathies may be purely coincidental, some have suggested that the tissue damage caused by hereditary neuropathies could evoke an immune response leading to superimposed inflammatory neuropathies (Korn-Lubetzki et al., 2002; Remiche et al., 2013).

Regardless of whether their CIDP is idiopathic or not, a diagnosis of inflammation in patients with an hereditary neuropathy may be difficult. Clinical signs and symptoms may overlap, and evaluations such as electrophysiology or nerve biopsies are not helpful to establish a diagnosis of inflammatory demyelinating disease when demyelination is already present due to hereditary disease. Moreover, current diagnostic criteria for CIPD list hereditary demyelinating neuropathies as a diagnostic exclusion criterion (van den Bergh et al., 2010), meaning that inflammatory neuropathies may go unrecognised and untreated in patients with an established diagnosis of an hereditary neuropathy.

However, it is important to recognise possible inflammation in patients with hereditary neuropathies, because of its therapeutic implications: where hereditary neuropathies can usually only be managed with lifestyle changes, bracing, and physical therapy, inflammation may be amenable to pharmacological treatment. The use of IVIG in CIPD, for example, is well established (Eftimov, Winer, Vermeulen, de Haan,

& van Schaik, 2013), and there is a growing body of evidence on the use of IVIg in chronic pain syndromes (Goebel, 2014). Hereditary neuropathies are usually painless, so the presence of pain, like in our patient, may indicate inflammation. Inflammation should also be considered in patients who show signs of a long-term, progressive neuropathy rather than the regular episodic weakness seen in HNPP. An n-of-one trial to test the effect of treatment for this potentially coexisting inflammatory neuropathy, such as the one described for our patient, could be considered in these patients.

Clinical n-of-one trials, such as the one presented here, are a tool that can be used to guide appropriate treatment in rare diseases (Guyatt et al., 1988). N-of-one trials have been used in the past to optimise treatment for individual patients, reduce unnecessary prescribing, and increase treatment compliance (C. Nikles, Clavarino, & Del Mar, 2005; Scuffham et al., 2010). Formal "trials of therapy", such as the one described in this study, can be valuable in guiding clinical practice when there is no evidence available from group-randomised clinical trials (RCTs), when the results of such trials do not necessarily generalise to one's patient in the consultation room, or when there are other pertinent reasons to optimise treatment, for example, because of the high cost of a medicinal product (Guyatt et al., 1988). IVIg to treat inflammation associated with HNPP fulfils these criteria: there are no clinical treatment guidelines, there is no evidence from earlier trials available, and IVIg is costly to produce and its availability is limited. Moreover, many diseases for which IVIg is prescribed require long-term treatment (Donofrio et al., 2009; Kumar, Teuber, & Gershwin, 2006), including when it is used to treat CIDP. The majority of patients require infusions every two to six weeks for a sustained response (Kuitwaard & van Doorn, 2009), and a review suggests that it can be withdrawn in less than 15% without causing a relapse (van Doorn, Dippel, & van Burken, 2003). In our patient, increasing the interval between infusions from five to six weeks led to an unacceptable clinical deterioration. N-of-one trials such as the current one may help to establish whether a particular patient has a true need for this type of treatment, and may thus aid appropriate prescription.

To our knowledge, this is the first randomised controlled trial (RCT) of IVIg to treat symptoms of inflammation in patients with HNPP; thus far, only anecdotal evidence suggested that IVIg may be effective in such patients (Mouton et al., 1999; Korn-Lubetzki et al., 2002; Remiche et al., 2013). The lack of RCTs may partly be due to the challenges associated with conducting RCTs in such small patient populations (Griggs et al., 2009). The n-of-one trial design could greatly facilitate the process of conducting RCTs in this type of patient population, since data from several n-of-one trials can be aggregated to obtain population effect estimates (Zucker, Ruthazer, & Schmid, 2010; Zucker et al., 1997). Furthermore, Bayesian analysis methods, which can make use of prior knowledge, allow for continued updating of treatment effect estimates as new data become available (Zucker et al., 1997). Thus, results from future trials in similar patients can be meaningfully combined with the results from the current trial to obtain an increasingly robust estimate of the population effect of

IVIg to treat inflammation in patients with HNPP. Such personalised and adaptive approaches may also be useful in other situations where only very few patients are available for research (Griggs et al., 2009).

Because this study was done in only one patient, its results may not necessarily generalise to other patients. Other limitations of the design include the need for multiple crossovers between the active and control intervention, which means that the participant burden in n-of-one trials is generally higher than in most other intervention research. Efforts should be made to reduce this burden and to prevent dropout during the trial. We chose to use a patient diary with two separate VAS scales to measure our outcomes and minimise the number of hospital visits for the patient. Although the VAS scale for pain has been extensively validated (Dworkin et al., 2008), this was not the case with the VAS for subjective muscle strength. Furthermore, subjective scores of pain and strength may co-vary. For example, when a limb is painful, it may also be self-reported as being weak, even if bedside strength assessment is normal. Future studies could benefit from using only validated outcome measures and from including more objective outcome measures alongside subjective ones, if this is possible without increasing the participant burden to an unacceptable level. Because of the frequent crossovers between IVIg and placebo, we were unable to assess whether the effect of IVIg is cumulative over several doses. Unblinding of the patient may also be a problem in multiple crossover trials, and may occur more easily when there are clear treatment or adverse effects. Our patient experienced such a clear effect of treatment (but no adverse effects). Although she was blinded to the infusion type at the time of each trial infusion, the clear treatment effect of IVIg meant that she was able to guess the nature of the infusion after several hours to days. This may have introduced some bias in the outcome measures, although Figure 6.1 still displays considerable variation and trend changes in both outcomes over time and regardless of the type of trial infusion. In future studies, bias may be reduced by using objective outcome measures and blinding of the outcome assessor. Finally, readers may not be familiar with Bayesian testing of informative hypotheses, a method which is more common in psychological research than clinical medicine. Therefore, it is noted that conventional statistical analysis of this n-of-1 RCT could not have accommodated consideration of multiple, clinically relevant hypotheses. Furthermore, conventional analyses would have suffered from low power. Despite the limitations of the trial, the results were useful to guide treatment of this patient.

In conclusion, we presented a trial of a patient with HNPP and concomitant symptoms of pain and muscle weakness which improved after continued treatment with IVIg. This suggests that some patients with hereditary neuropathies may have co-existing inflammation, which is important to recognise because adequate treatment can improve their symptoms and quality of life. We also demonstrated the value of n-of-one trials for conducting research in rare conditions.

6.A Description of analyses

6.A.1 Effect of IVIg on pain

The effect of pain after the infusions can be analysed using the regression model:

$$y_{1ti} = \alpha_{1t} + \beta_{1t} x_{1ti} + \epsilon_{1ti}, \tag{6.1}$$

where y_{1ti}, $t = 1, \ldots, 7$ denotes the pain on a 14 point scale after 7 infusions, x_{1ti} denotes the number of days after the treatment, α_{1t} is the intercept, β_{1t} is the coefficient of the day number, and ϵ_{1ti} is the residual, which is normally distributed with mean 0 and unknown variance. We first tested the expectation that the decrease in pain in the first 7 days after IVIg infusions is greater than after placebo. This expectation can be translated into the following hypothesis H_1 among the coefficients of the day numbers:

$$H_1 : \{\beta_{11}, \beta_{12}, \beta_{13}\} > \{\beta_{14}, \beta_{15}, \beta_{16}, \beta_{17}\}, \tag{6.2}$$

which was compared to the unconstrained hypothesis H_a:

$$H_a : \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{15}, \beta_{16}, \beta_{17}, \tag{6.3}$$

where $\beta_{1t}, t = 1, \ldots, 3$ denote the coefficients after three placebo infusions, whereas $\beta_{1t}, t = 4, \ldots, 7$ are those after four IVIg infusions. Note that the day numbers x_{1ti} for each infusion are 1, 4, 6, and 8, except the infusion $t = 6$ for which the day numbers are 0, 2, 5, and 7. Using SPSS, the estimate and squared standard error of each coefficient were obtained, which are displayed in Table 6.1 (see the end of this appendix). Hypothesis H_1 was then evaluated using BIG, which rendered a Bayes factor of 33.22 for H_1 against H_a, which implies that H_1 gains strong support from the data.

We then assessed whether IVIg produced a clinically relevant reduction in pain, which can be expressed by:

$$H_2 : \beta_{14} < l_{14}, \beta_{15} < l_{15}, \beta_{16} < l_{16}, \beta_{17} < l_{17}, \tag{6.4}$$

where $l_{1t} = -30\% * y_{1t1}/7, t = 4, \ldots, 7$ is the clinically relevant level for each treatment, and y_{1t1} is the pain on the day after each IVIg infusion is given. Using the results in Table 6.1 for β_{1t} and $l_{1t}, t = 4, \ldots, 7$, we obtained a Bayes factor of 13.40 for H_2 against H_a, which implies strong evidence that IVIg produces a clinically meaningful reduction in pain in this patient.

We then proceeded to test the combined first and second expectations, i.e., in the first week after infusion, pain decreases more rapidly after IVIg than after placebo, and the pain after IVIg decreases beyond the clinically relevant level. This leads to the following hypothesis:

$$H_3 : \begin{array}{c} \{\beta_{11}, \beta_{12}, \beta_{13}\} > \{\beta_{14}, \beta_{15}, \beta_{16}, \beta_{17}\}; \\ \beta_{14} < l_{14}, \beta_{15} < l_{15}, \beta_{16} < l_{16}, \beta_{17} < l_{17}. \end{array} \tag{6.5}$$

This hypothesis can be expressed as $H_3 : H_1 \& H_2$ because it contains the constraints both in H_1 and H_2. Evaluating this hypothesis rendered a Bayes factor of 63.74, which implies that there is strong evidence that IVIg decreases pain more than placebo, and that it decreases pain to a clinically meaningful extent in this patient.

6.A.2 Subjective muscle strength

A similar model was specified for the effect on subjective muscle strength:

$$y_{2ti} = \alpha_{2t} + \beta_{2t} x_{2ti} + \epsilon_{2ti}, \tag{6.6}$$

where y_{2ti}, x_{2ti}, α_{2t}, β_{2t}, and ϵ_{2ti} are the same notations but for muscle strength. Analogous to the effect on pain, we assessed whether the subjective increase in muscle strength in the first 7 days after IVIg infusion is greater than after placebo. This results in the following hypothesis:

$$H_4 : \{\beta_{21}, \beta_{22}, \beta_{23}\} < \{\beta_{24}, \beta_{25}, \beta_{26}, \beta_{27}\}, \tag{6.7}$$

where $\beta_{2t}, t = 1, \ldots, 7$ again denote the coefficients after the treatments (see Table 6.1). Evaluating H_4 in BIG rendered a Bayes factor of 36.24, which implies strong evidence that IVIg increases muscle strength more than placebo in this patient.

Secondly, we assessed whether the increase in muscle strength was subjectively meaningful by evaluating the hypothesis:

$$H_5 : \beta_{24} > l_{24}, \beta_{25} > l_{25}, \beta_{26} > l_{26}, \beta_{27} > l_{27}, \tag{6.8}$$

where $l_{2t} = 30\% * y_{2t1}/7, t = 4, \ldots, 7$ is the subjectively relevant level for an increase in muscle strength, and y_{2t1} represents the muscle strength at the time of IVIg infusion. Using the results in Table 6.1 for β_{2t} and $l_{2t}, t = 4, \ldots, 7$, we obtained a Bayes factor of 15.05, which implies strong evidence that IVIg increases muscle strength to a meaningful extent in this patient.

Similar to pain, we then combined these hypotheses and tested whether in the first week after infusion muscle strength increased more rapidly after IVIg than after placebo and whether it increased beyond the clinically meaning level. This hypothesis can be expressed by:

$$H_6 : \begin{array}{c} \{\beta_{21}, \beta_{22}, \beta_{23}\} < \{\beta_{24}, \beta_{25}, \beta_{26}, \beta_{27}\}; \\ \beta_{24} > l_{24}, \beta_{25} > l_{25}, \beta_{26} > l_{26}, \beta_{27} > l_{27}, \end{array} \tag{6.9}$$

where $H_6 : H_4 \& H_5$. Evaluating this hypothesis resulted in a Bayes factor of 61.51, meaning that there is strong evidence that IVIg increases subjective muscle strength more than placebo, and that it increases it to a clinically meaningful extent.

6.A.3 Course of pain and muscle strength following IVIg infusions

Finally, to assess the need for regular IVIg infusions, we tested the hypothesis that pain first decreases and then increases again in the three weeks following IVIg infusion. To investigate this expectation, a quadratic regression model was used:

$$y_{1ti} = \alpha_{1t} + \beta_{1t}x_{1ti} + \gamma_{1t}x_{1ti}^2 + \epsilon_{1ti}, \tag{6.10}$$

where $\gamma_{1t}, t = 1, \ldots, 4$ is the coefficient of the squared day number. If $\gamma_{1t} > 0$, this means the pain y_{1ti} decreased during the first several days after infusion and then increased again. For this reason we constructed the hypothesis:

$$H_7 : \gamma_{11} > 0, \gamma_{12} > 0, \gamma_{13} > 0, \gamma_{14} > 0. \tag{6.11}$$

Running BIG with the estimates and variances of γ_{1t} shown in Table 6.1 rendered a Bayes factor of 13.78, which implies strong evidence that pain first decreases following IVIg, and then increases again as the effects of IVIg start to wear off. A similar quadratic model was used for subjective muscle strength:

$$y_{2ti} = \alpha_{2t} + \beta_{2t}x_{2ti} + \gamma_{2t}x_{2ti}^2 + \epsilon_{2ti}. \tag{6.12}$$

A negative γ_{2t} indicates that in the beginning days muscle strength y_{2ti} increases and thereafter it decreases again. Thus, hypothesis H_8 is as follows:

$$H_8 : \gamma_{21} < 0, \gamma_{22} < 0, \gamma_{23} < 0, \gamma_{24} < 0. \tag{6.13}$$

The Bayes factor for this hypothesis was 15.67, indicating that there is strong support that subjective muscle strength increases in the first days after IVIg infusion, and then decreases again over the following weeks.

Table 6.1: Estimates and variances of the coefficients (l_{1t} and l_{2t} denote the relevant levels for the decrease of pain and increase of muscle strength, respectively, in the first week after IVIg. Note that n denotes the number of measurements upon which the estimates are based.

		Pain				Muscle strength				
		n	estimates	variance	l_{1t}		n	estimates	variance	l_{2t}
placebo	β_{11}	4	0.416	0.187		β_{21}	4	-0.734	0.142	
	β_{12}	4	0.253	2.79E-2		β_{22}	4	-0.319	3.17E-2	
	β_{13}	4	0.824	2.76E-2		β_{23}	4	-0.234	5.04E-3	
IVIg	β_{14}	4	-0.907	0.106	-0.270	β_{24}	4	0.823	3.39E-2	0
(1 week)	β_{15}	4	-0.412	3.61E-2	-0.126	β_{25}	4	0.508	8.41E-4	0.216
	β_{16}	4	-0.753	1.04E-2	-0.294	β_{26}	4	0.321	4.10E-3	0.210
	β_{17}	4	-0.984	4.62E-3	-0.354	β_{27}	4	0.340	5.93E-3	0.246
IVIg	γ_{11}	8	0.063	1.96E-4		γ_{21}	8	-0.041	4.9E-5	
(3 week)	γ_{12}	9	0.016	1.69E-4		γ_{22}	9	-0.029	9.0E-6	
	γ_{13}	10	0.044	3.60E-5		γ_{23}	10	-0.022	9.0E-6	
	γ_{14}	9	0.056	2.89E-4		γ_{24}	9	-0.024	9.0E-6	

References

Agresti, A. (2007). *An introduction to categorical data analysis.* Hoboken, NJ: John Wiley and Sons.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáaki (Eds.), *Proc. 2nd int. symp. information theory* (p. 267-281). Budapest: Akademiai kiado.

Azen, R., & Walker, C. M. (2010). *Categorical data analysis for the behavioral and social sciences.* New York, NY: Routledge.

Barlow, R., Bartholomew, D., Bremner, J., & Brunk, H. (1972). *Statistical inference under order restrictions: Theory and application of isotonic regression.* New York, NY: John Wiley and Sons.

Bartholomew, D. (1959). A test of homogeneity for ordered alternatives. *Biometrika*, *46*, 36-48.

Berger, J. (1985). *Statistical decision theory and Bayesian analysis.* New York: Springer-Verlag.

Berger, J., Boukai, B., & Wang, Y. (1997). Unified frequentist and Bayesian testing of a precise hypothesis. *Statistical Science*, *12*, 133-160.

Berger, J., Brown, L., & Wolpert, R. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *The Annals of Statistics*, *22*, 1787-1807.

Berger, J., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317-335.

Berger, J., & Pericchi, L. (1996). The intrinsic Bayes factor for model selection and

prediction. *Journal of the American Statistical Association*, *91*, 109-122.

Bird, T. (1998). Hereditary neuropathy with liability to pressure palsies. In R. Pagon, M. Adam, & H. Ardinger (Eds.), *Genereviews [internet].* Seattle, WA: University of Washington.

Bollen, K. (1989). *Structural equations with latent variables.* New York, NY: John Wiley and Sons.

Burnham, K., & Anderson, D. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York, NY: Springer-Verlag.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

Cohen, J. (1994). The earth is round, $p < .05$. *American Psychologist*, *49*, 997-1003.

Deković, M., Wissink, I. B., & Meijer, A. M. (2004). The role of family and peer relations in adolescent antisocial behaviour: Comparison of four ethnic groups. *Journal of Adolescence*, *27*, 497-514.

De Leeuw, J., & Meijer, E. (2008). *Handbook of multilevel analysis.* New York, NY: Springer.

de Santis, F., & Spezzaferri, F. (2001). Consistent fractional Bayes factor for nested normal linear models. *Journal of Statistical Planning and Inference*, *97*, 305-321.

Desurkar, A., Lin, J., Mills, K., Al-Sarraj, S., Jan, W., Jungbluth, H., & Wraige, E. (2009). Charcot-Marie-Tooth (CMT) disease 1A with superimposed inflammatory polyneuropathy in children. *Neuropediatrics*, *40*, 85-88.

DiCiccio, T., Kass, R., Raftery, A., & Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, *92*, 903-915.

Dickey, J. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Statistics*, *42*, 204-223.

Donofrio, P., Berger, A., Brannagan, T. r., Bromberg, M., Howard, J., Latov, N., ... Tandan, R. (2009). Consensus statement: the use of intravenous immunoglobulin in the treatment of neuromuscular conditions report of the AANEM ad hoc

committee. *Muscle Nerve*, *40*, 890-900.

Dubourg, O., Mouton, P., Brice, A., LeGuernb, E., & Bouchea, P. (2000). Guidelines for diagnosis of hereditary neuropathy with liability to pressure palsies. *Neuromuscul Disord*, *10*, 206-208.

Dworkin, R., Turk, D., Wyrwich, K., D., B., Cleeland, C., J.T., F., ... McQuay, H. (2008). Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. *The Journal of Pain*, *9*, 105-121.

Eftimov, F., Winer, J., Vermeulen, M., de Haan, R., & van Schaik, I. (2013). Intravenous immunoglobulin for chronic inflammatory demyelinating polyradiculoneuropathy. *Cochrane Database of Systematic Reviews*, *12*.

Engle, R. F. (1984). Wald, likelihood ratio, and lagrange multiplier tests in econometrics. *Handbook of econometrics*, *2*, 775-826.

Garcia-Donato, G., & Chen, M.-H. (2005). Calibrating Bayes factor under prior predictive distributions. *Statistics Sinica*, *15*, 359-380.

Gelfand, A. E., Smith, A. F. M., & Lee, T. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, *87*, 523-532.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Ginsberg, L., Malik, O., Kenton, A., Sharp, D., Muddle, J., Davis, M., ... King, R. (2004). Coexistent hereditary and inflammatory neuropathy. *Brain*, *127*, 193-202.

Goebel, A. (2014). Cellular and behavioural models to predict responses to immunoglobulin G treatment in complex regional pain syndrome. *Clinical & Experimental Immunology*, *178*, 136-137.

Good, I. (1992). The Bayes/non-Bayes compromise: A brief review. *Journal of the American Statistical Association*, *87*, 597-606.

Gourieroux, C., Holly, A., & Monfort, A. (1982). Likelihood ratio test, Wald test, and Kuhn-Tucker test in linear models with inequality constraints on the regression parameters. *Econometrica*, *50*, 63-80.

Griggs, R., Batshaw, M., Dunkle, M., Gopal-Srivastava, R., Kaye, E., Krischer, J., ... Network, R. D. C. R. (2009). Clinical research for rare disease: opportunities, challenges, and solutions. *Molecular Genetics and Metabolism*, *96*, 20-26.

Gu, X., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, *19*(4), 511-527.

Gu, X., Mulder, J., & Hoijtink, H. (in press). Error probabilities in default bayesian hypothesis testing. *Journal of Mathematical Psychology*.

Guber, D. (1999). Getting what you pay for the debate over equity in public school expenditures. *Journal of Statistics Education*, *7*(8). `www.amstat.org/publications/jse/secure/v7n2/datasets.guber.cfm`. ([Online])

Guyatt, G., Sackett, D., Adachi, J., Roberts, R., Chong, J., Rosenbloom, D., & Keller, J. (1988). A clinician's guide for conducting randomized trials in individual patients. *CMAJ*, *139*, 497-503.

Hoijtink, H. (2012). *Informative hypotheses: Theory and practice for behavioural and social scientists*. Boca Raton, FL: Chapman and Hall/CRC.

Howell, D. (2012). *Statistical methods for psychology* (8th ed.). Belmont, CA: Cengage Learning.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Oxford University Press.

Johnson, V., & Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B*, *72*, 143-170.

Kass, R., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795.

Kass, R., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz Criterion. *Journal of the American Statistical Association*, *90*, 928-934.

Kato, B. S., & Hoijtink, H. (2006). A Bayesian approach to inequality constrained linear mixed models: Estimation and model selection. *Statistical Modelling*, *6*, 231-249.

Klein, C., Dyck, P., Friedenberg, S., Burns, T., Windebank, A., & Dyck, P. (2002). Inflammation and neuropathic attacks in hereditary brachial plexus neuropathy. *Journal of Neurology, Neurosurgery, and Psychiatry*, *73*, 45-50.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: The Guilford Press.

Klugkist, I., Bullens, J., & Postma, A. (2012). Evaluating order constrained hypotheses for circular data using permutation tests. *British Journal of Mathematical and Statistical Psychology*, *65*, 222-236.

Klugkist, I., & Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics & Data Analysis*, *51*, 6367-6379.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Baysian approach. *Psychological Methods*, *10*, 447-493.

Klugkist, I., Laudy, O., & Hoijtink, H. (2010). Bayesian evaluation of inequality and equality constrained hypotheses for contingency tables. *Psychological Methods*, *15*, 281-299.

Korn-Lubetzki, I., Argov, Z., Raas-Rothschild, A., Wirguin, I., & Steiner, I. (2002). Family with inflammatory demyelinating polyneuropathy and the HNPP 17p12 deletion. *American Journal of Medical Genetics*, *113*, 275-278.

Kuiper, R. M., & Hoijtink, H. (2010). Comparisons of means using exploratory and confirmatory approaches. *Psychological Methods*, *15*, 69-86.

Kuiper, R. M., Klugkist, I., & Hoijtink, H. (2010). A Fortran 90 program for confirmatory analysis of variance. *Journal of Statistical Software*, *34*(8), 1-31.

Kuitwaard, K., & van Doorn, P. (2009). Newer therapeutic options for chronic inflammatory demyelinating polyradiculoneuropathy. *Drugs*, *69*, 987-1001.

Kumar, A., Teuber, S., & Gershwin, M. (2006). Intravenous immunoglobulin: striving for appropriate use. *International Archives of Allergy and Immunology*, *140*, 185-198.

Le Forestier, N., LeGuern, E., Coullin, P., Birouk, N., Maisonobe, T., Brice, A., ... P., B. (1997). Recurrent polyradiculoneuropathy with the 17p11.2 deletion. *Muscle Nerve*, *20*, 1184-1186.

Liang, F., Paulo, R., Molina, G., Clyde, M., & Berger, J. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*, 410-423.

Lindley, D. (1957). A statistical paradox. *Biometrika*, *44*, 187-192.

Luigetti, M., Zollino, M., Conti, G., Romano, A., & Sabatelli, M. (2013). Inherited neuropathies and deafness caused by a PMP22 point mutation: a case report and a review of the literature. *Neurological Sciences*, *34*, 1705-1707.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325-337.

Marques, W., Funayama, C., Secchin, J., Lourenço, C., Gouvêa, S., Marques, V., ... Barreira, A. (2010). Coexistence of two chronic neuropathies in a young child: Charcot–Marie–Tooth disease type 1A and chronic inflammatory demyelinating polyneuropathy. *Muscle Nerve*, *42*, 598-600.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

McCullogh, C. E., & Searle, S. R. (2001). *Generalized linear and mixed models*. New York, NY: Wiley.

Morey, R., Rouder, J., Pratte, M., & Speckman, P. (2011). Using MCMC chain outputs to efficiently estimate Bayes factors. *Journal of Mathematical Psychology*, *55*, 368-378.

Morey, R., Wagenmakers, E.-J., & Rouder, J. (in press). "Calibrated" Bayes factors should not be used: A reply to Hoijtink, van Kooten, and Hulsker. *Multivariate Behavioral Research*.

Mouton, P., Tardieu, S., Gouider, R., Birouk, N., Maisonobe, T., Dubourg, O., ... Bouche, P. (1999). Spectrum of clinical and electrophysiologic features in HNPP patients with the 17p11.2 deletion. *Neurology*, *52*, 1440-1446.

Mulder, J. (2014a). Bayes factors for testing inequality constrained hypotheses: Issues with prior specification. *British Journal of Mathematical and Statistical Psychology*, *67*, 153-171.

Mulder, J. (2014b). Prior adjusted default Bayes factors for testing (in)equality

constrained hypotheses. *Computational Statistics & Data analysis*, *71*, 448-463.

Mulder, J. (in press). Bayes factors for testing order-constrained hypotheses on correlations. *Journal of Mathematical Psychology*.

Mulder, J., & Fox, J. P. (2013). Bayesian tests for variance components in a compound symmetry covariance structure. *Statisics and Computing*, *23*, 109-122.

Mulder, J., Hoijtink, H., & de Leeuw, C. (2012). BIEMS: A Fortran90 program for calculating Bayes factors for inequality and equality constrained models. *Journal of Statistical Software*, *46*(2).

Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference*, *140*, 887-906.

Mulder, J., Klugkist, I., van de Schoot, R., Meeus, W., Selfhout, M., & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology*, *53*, 530-546.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Nikles, C., Clavarino, A., & Del Mar, C. (2005). Using n-of-1 trials as a clinical tool to improve prescribing. *British Journal of General Practice*, *55*, 175-180.

Nikles, J., Mitchell, G., Walters, J., Hardy, J., Good, P., Rowett, D., ... Currow, D. (2009). Prioritising drugs for single patient (n-of-1) trials in palliative care. *Palliative Medicine*, *23*, 623-634.

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. Hoboken, NJ: John Wiley and Sons.

O'Hagan, A. (1995). Fractional Bayes factors for model comparisons (with discussion). *Journal of the Royal Statistical Society, Series B*, *57*, 99-138.

Pérez, J., & Berger, J. (2002). Expected-posterior prior distributions for model selection. *Biometrika*, *89*(3), 491-512.

Popper, K. (1959). *The logic of scientific discovery*. London: Hutchinson.

Pou Serradell, A., Monells, J., Téllez, M., Fossas, P., Löfgren, A., Meuleman, J., ... Martin, J. (2002). [hereditary neuropathy with liability to pressure palsies: study of six Spanish families]. *Rev Neurol (Paris)*, *158*, 579-588.

Remiche, G., Abramowicz, M., & Mavroudakis, N. (2013). Chronic inflammatory demyelinating polyradiculoneuropathy (CIDP) associated to hereditary neuropathy with liability to pressure palsies (HNPP) and revealed after influenza AH1N1 vaccination. *Acta Neurologica Belgica*, *113*, 519-522.

Rosseel, Y. (2012). lavaan: an R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t*-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*, 225-237.

Royall, R. (1997). *Statistical evidence: A likelihood paradigm.* New York: NY: Chapman and Hall/CRC.

Sagnelli, A., Piscosquito, G., & Pareyson, D. (2013). Inherited neuropathies: an update. *Journal of Neurology*, *260*, 2684-2690.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461-464.

Scuffham, P., Nikles, J., Mitchell, G., Yelland, M., Vine, N., Poulos, C., ... Glasziou, P. (2010). Using n-of-1 trials to improve patient management and save costs. *Journal of General Internal Medicine*, *25*, 906-913.

Silvapulle, M., & Sen, P. (2004). *Constrained statistical inference; order, inequality, and shape constraints.* New York: NY: Wiley.

Sober, E. (2002). Bayesianism: Its scope and limits. In R. Swinburn (Ed.), *Bayes's theorem* (Vol. 113, p. 21-28). Proceedings of the British Academic Press.

Stögbauer, F., Young, P., Kuhlenbäumer, G., de Jonghe, P., & Timmerman, V. (2000). Hereditary recurrent focal neuropathies: clinical and molecular features. *Neurology*, *54*, 546-551.

van den Bergh, P., Hadden, R., Bouche, P., Cornblath, D., Hahn, A., Illa, I., ... van Schaik I.N. (2010). European Federation of Neurological Societies/Peripheral Nerve Society guideline on management of chronic inflammatory demyelinating

polyradiculoneuropathy: report of a joint task force of the European Federation of Neurological Societies and the Peripheral Nerve Society—first revision. *European Journal of Neurology*, *17*, 356-363.

van de Schoot, R., Hoijtink, H., & Deković, M. (2010). Testing inequality constrained hypotheses in SEM models. *Structural Equation Modeling*, *19*, 1-17.

van de Schoot, R., Hoijtink, H., Hallquist, M., & Boelen, P. (2012). Bayesian evaluation of inequality constrained hypotheses in SEM models using Mplus. *Structural Equation Modeling*, *17*, 443-463.

van de Schoot, R., Hoijtink, H., & Romeijn, J. W. (2011). Moving beyond traditional null hypothesis testing: Evaluating expectations directly. *Frontiers in Psychology*, *2*(24).

van Doorn, P., Dippel, D., & van Burken, M. (2003). Longterm IV Immunoglobulin treatment in CIDP. *Journal of the Peripheral Nervous System*, *8*, 70.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin and Review*, *14*, 779-804.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: a tutorial on the savage dickey method. *Cognitive Psychology*, *60*, 158-189.

Warren, R. D., White, J. K., & Fuller, W. A. (1974). An errors-in-variables analysis of managerial role performance. *Journal of the American Statistical Association*, *69*, 886-893.

Watanabe, M., Yamamoto, N., Ohkoshi, N., Nagata, H., Kohno, Y., Hayashi, A., ... Shoji, S. (2002). Corticosteroid-responsive asymmetric neuropathy with a myelin protein zero gene mutation. *Neurology*, *59*, 767-769.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. Goel & A. Zellner (Eds.), *Bayesian inference and decision techniques: Essays in honor of bruno de finetti* (p. 233-243). Amsterdam: North-Holland/Elsevier.

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics: proceedings of the first international meeting held in valencia* (p. 585-603). Spain: University of Valencia.

Zucker, D., Ruthazer, R., & Schmid, C. (2010). Individual (N-of-1) trials can be combined to give population comparative treatment effect estimates: methodologic considerationst. *Journal of Clinical Epidemiology*, *63*, 1312-1323.

Zucker, D., Schmid, C., McIntosh, M., D'Agostino, R., Selker, H., & Lau, J. (1997). Combining single patient (N-of-1) trials to estimate population treatment effects and to evaluate individual patient responses to treatment. *Journal of Clinical Epidemiology*, *50*, 401-410.

Summary

The evaluation of informative hypotheses has gained in popularity in applied sciences, because it enables researchers to investigate their expectations with respect to the population of interest. In this dissertation, approximate Bayesian approaches are developed to evaluate informative hypotheses by means of the Bayes factor in a very general class of statistical models. The Bayes factor quantifies the support from the data in favor of one hypothesis against another. The computation of the Bayes factor requires the specification of the prior distribution and the derivation of the posterior distribution for model parameters under the unconstrained hypothesis.

In this dissertation several prior specification methods are presented in different situations. On the one hand, for the evaluation of informative hypotheses specified using only inequality constraints, Chapter 2 specifies a noninformative normal prior distribution under the unconstrained hypothesis in which every combination of values is equally likely for the parameters used in the hypotheses. An alternative noninformative prior is proposed in Chapter 3 such that the Bayes factor is invariant to linear one-one transformation of the data. On the other hand, the Bayesian evaluation of informative hypotheses that contain equality constraints requires the specification of default priors under the unconstrained hypothesis when subjective prior information is not available. In Chapter 4 a new method is proposed for specifying default priors in the one sample t test based on the frequentist properties. A follow-up study in Chapter 5 generalizes this method to evaluate informative hypotheses in a general class of statistical models.

The posterior distribution of the parameters integrates the information contained in the prior and the data. The distributional form of the posterior depends on the statistical model at hand. In Chapter 2, Chapter 3, and Chapter 5, however, the posterior distribution in any model is approximated using a (multivariate) normal distribution based on large sample theory, which leads to an easy and straightforward tool for the computation of the Bayes factor for informative hypotheses. This approximation renders a generally applicable Bayesian procedure and therefore substantially extends the class of statistical models to which the evaluation of informative hypotheses can be applied.

The development of software packages for the proposed Bayesian approaches is

another aspect of this research. Chapter 2 provides a software package BIG to compute the Bayes factor for the evaluation of informative hypotheses that only consist of inequality constraints. In Chapter 3 a new efficient algorithm is explored for the computation of Bayes factors, which dramatically reduces the computational time in BIG. Finally, a software package BaIn is developed in Chapter 5 to fill the vacancy of the Bayesian evaluation of informative hypotheses with possible equality constraints. The user manuals for BIG and BaIn are included in the dissertation.

Samenvatting

Het evalueren van informatieve hypotheses heeft aan populariteit gewonnen binnen de toegepaste wetenschappen, omdat het wetenschappers in staat stelt om hun specifieke verwachtingen over de populatie van interesse te onderzoeken. In deze dissertatie worden Bayesiaanse benaderingen gepresenteerd voor het evalueren van informatieve hypotheses middels de Bayes factor, voor een zeer brede klasse van statistische modellen.

Door middel van de Bayes factor wordt gekwantificeerd in hoeverre de data een bepaalde hypothese ondersteunen in verhouding tot een andere hypothese. Het berekenen van de Bayes factor in de context van informatieve hypotheses vereist het specificeren van prior kansverdelingen voor de parameters in het model en het afleiden van de posterior kansverdeling, gegeven een hypothese zonder restricties.

In deze dissertatie worden verschillende methoden voor het specificeren van prior kansverdelingen gepresenteerd die geschikt zijn voor verschillende omstandigheden. In Hoofdstuk 2 en 3 ligt de focus op hypotheses met alleen ongelijkheidsrestricties. In Hoofdstuk 2 worden niet-informatieve normaalverdelingen gespecificeerd, waarin elke combinatie van waardes voor de parameters die betrekking hebben op de hypotheses even waarschijnlijk zijn. Een alternatieve niet-informatieve priorspecificatie wordt voorgesteld in Hoofdstuk 3, waarbij de Bayes factor invariant is voor lineaire één-op-één transformaties van de data.

In Hoofdstuk 4 en 5 ligt de focus op het Bayesiaans evalueren van informatieve hypotheses met gelijkheidsrestricties. Als er geen subjectieve prior informatie beschikbaar is in deze context, is het noodzakelijk om standaardpriors te specificeren. In Hoofdstuk 4 wordt een nieuwe methode voorgesteld om dergelijke standaardpriors te specificeren voor een t-toets voor één steekproef, gebaseerd op frequentistische eigenschappen. Een generalisatie van deze methode voor het evalueren van informatieve hypotheses wordt gepresenteerd in Hoofdstuk 5, zodat de methode gebruikt kan worden voor een meer algemene klasse van statistische modellen.

De posterior kansverdeling van de parameters integreert de informatie in de prior kansverdeling en in de data. De vorm van de posterior kansverdeling hangt af van het statistische model. Echter, in Hoofdstuk 2, 3 en 5 wordt de posterior kansverdeling in ieder model benaderd met een (multivariaat) normale verdeling op basis van de wet

van de grote aantallen, wat resulteert in een eenvoudige, recht-toe-recht-aan-methode voor de berekening van de Bayes factor voor informatieve hypotheses. Deze benadering levert een algemeen toepasbare Bayesiaanse procedure op en breidt daarmee de klasse van statistische modellen waarop de evaluatie van informatieve hypothese kan worden toegepast aanzienlijk uit.

Een ander aspect van dit onderzoek is de ontwikkeling van softwarepaketten voor de voorgestelde Bayesiaanse benaderingen. Voor het onderzoek in Hoofdstuk 2 is het software pakket BIG ontwikkeld dat gebruikt kan worden voor het berekenen van de Bayes factor voor het evalueren van informatieve hypotheses die alleen uit ongelijkheidsrestricties bestaan. In Hoofdstuk 3 wordt een nieuw efficient algoritme besproken dat de tijd die nodig is voor het berekenen van Bayes factors in pakket BIG drastisch verminderd. Ten slotte wordt in Hoofdstuk 5 het softwarepakket BaIn gepresenteerd, dat is ontwikkelt om een gat in de markt — het evalueren van informatieve hypotheses met eventuele gelijkheidsrestricties — te vullen. In deze dissertatie zijn ook de handleidingen voor BIG en BaIn toegevoegd.

Acknowledgement

I would like to express my gratitude to many people for their support during my PhD study. First of all, my deepest gratitude goes to my supervisor Herbert Hoijtink and co-supervisor Joris Mulder. Herbert, you led me to know the advantages of informative hypotheses and Bayesian tests. Thank you for offering me this PhD project and encouraging me during my study. I learned a lot from you about the way of thinking, producing, writing, and presenting. Joris, you always brought bright ideas into my PhD project. I enjoyed our discussions, although we sometimes had different opinions and you always won. Both of you provide substantial support in my PhD research and extensive feedback for each of my papers. Thanks for your constant encouragement and guidance. Without your support, this thesis could not have reached its present form.

The department of Methodology and Statistics offers me an enjoyable working atmosphere. I appreciate the warm welcome from my colleagues to make me not lonely. In particular, I wish to thank Noemi, Maria, Maryam, Susanna, Silvia, who started their PhD at the same time as me. We had pleasure time together during IOPS courses and meetings. Furthermore, I would like to thank my friends Haifang and Yasin. It was nice to meet you here. My thanks also go to my officemates Sanne Nooijens, Rebecca, Floryt, Sanne Smid, and Corine for their support in many aspects. The secretaries in the department provided me kind help through years. Thank you, Kevin, Flip, and Chantal.

My Chinese friends enriched my life in the Netherlands. I give my thankfulness to Wei Chen, Bing Yuan, Bo Liu, Qingyi Feng and many others for all the fun we have had in Utrecht in the last four years. Besides, I would also like to thank Jing Zhao and Weibing Han for their assistance.

There are a number of persons in China to whom I wish to give my gratitude as well. I would like to thank my master supervisor Yimin Shi for his sufficient support when I applied for my PhD scholarship. My thanks also go to my master classmates Wei Tan, Song Mao, and Wantao Jia for their essential help.

Last but not the least, I am grateful to my family for their endless support. Special

thanks go to my wife Meng Ma and my son Tinghe Gu with whom my life becomes more meaningful and colorful.

Xin Gu
Utrecht, April, 2016

About the author

Xin Gu was born on May 7th 1987 in Xi'an, China. He obtained his BSc with specialization in Information and Computing Science at Northwestern Polytechnical University in China in 2009. Thereafter, he became a research master student in Applied Mathematics and obtained his MSc in 2012. In January 2012, he started his PhD research in the Department of Methodology and Statistics at Utrecht University in the Netherlands, supported by China Scholarship Council (CSC).

His PhD work focused on developing approximate Bayesian procedure in hypothesis testing for general statistical models under the supervision of Prof. dr. Herbert Hoijtink and Dr. Joris Mulder. Several papers have been published based on his research as a PhD student. One of the published papers won the best paper award 2014 of the Interuniversity Graduate School of Psychometrics and Sociometrics (IOPS). He also presented the results of his research in some international conferences.

Publications:

Gu, X., Mulder, J., & Hoijtink, H. (2015). Error probabilities in default Bayesian hypothesis testing. *Journal of Mathematical Psychology.* Online First: 26 Sept. doi:10.1016/j.jmp.2015.09.001.

Vrinten, C., **Gu, X.**, Weinreich, S., Schipper, M., Wessels, J., Ferrari, M., Hoijtink, H., & Verschuuren, J. (2015). An n-of-one RCT for intravenous immunoglobulin G for inflammation in hereditary neuropathy with liability to pressure palsy (HNPP). *Journal of Neurology, Neurosurgery & Psychiatry.* Online First: 17 July. doi:10.1136/jnnp-2014-309427.

Gu, X., Mulder, J., Dekovic, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, 19(4), 511-527. doi:10.1037/met0000017.