Inequality Constrained Normal Linear Models

ISBN: 90-393-3908-2
Printed by: Febodruk BV, Enschede

Inequality Constrained Normal Linear Models

Ongelijkheidsrestricties in Normale Lineaire Modellen

(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. W. H. Gispen,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen op

vrijdag 11 februari 2005 des middags te 16.15 uur

door

Irene Gera Klugkist

geboren op 28 december 1970, te Delfzijl

Promotor: Prof. dr. H. J. A. Hoijtink
Faculteit der Sociale Wetenschappen
Universiteit Utrecht

Preface

It surely is always a good idea to thank family and friends for making my life more pleasurable. It is great and very important to have a social back-up for sharing your 'troubles' from work or even more important to forget everything related to work for a while. Thank you all for that.

However, no matter how demanding and challenging writing a PhD-thesis is, I still consider it mainly 'just a job'. Not an ordinary or nine-to-five job, but certainly one that I enjoyed a lot most of the time. There are several reasons for that.

First of all, I worked in the best environment possible. M&S is a great department!

Secondly, it is wonderful to have a few fellow researchers with the same interests and working on about the same topic. The weekly Bayesian meetings with Herbert, Olav and Bernet were most inspiring and instructive.

But to make a PhD-research really successful and in my opinion the best job you can ever have, most of all you need an interesting and challenging research-project as well as a capable, available and enthusiastic supervisor.

Herbert, thank you for making my job the best job ever!

Contents

Chapter 1

Introduction

Researchers often have one or more theories or expectations with respect to the outcome of their empirical research. To evaluate these theories they have to be translated into statistical models. When researchers talk about the expected relations between variables if a certain theory is correct, their statements are often in terms of one or more parameters expected to be larger or smaller than one or more other parameters. Stated otherwise, their statements are often formulated using inequality constraints. This dissertation deals with normal linear models with inequality constraints among model parameters.

To give an example, consider the following experiment where children with ADD (Attention Deficit Disorder) are randomly assigned to one of four conditions: no treatment (1), behavioral therapy (2), medication (3), and, behavioral therapy plus medication (4). Let the outcome variable of interest be the score on an attention test, and μ_j the average in group j ($j = 1, \ldots, 4$). The researcher that initiates this experiment, probably already has ideas or expectations about the expected outcomes *a priori*. For instance, the researcher may expect that a combined treatment, with both therapy and medication, is the best way to improve the attention span of children with ADD. Consider he does also expect that the effect of one treatment, therapy or medication, is better than no treatment at all. Formally stated, he expects a higher mean attention span for group 4 than for the groups 2 and 3, and the lowest mean attention span for group 1.

Denoting this model by M_1 we have

$$M_1 : \mu_1 < (\mu_2, \mu_3) < \mu_4,$$

that is, M_1 is the statistical model the researcher wants to evaluate. The first question the researcher likes to answer, is: *Do the data of the experiment support this theory?*

Traditional null hypothesis testing
To try to answer this question using traditional null hypothesis testing, the following hypotheses are formulated:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_A : \text{not all } \mu_j \text{ equal.}$$

A p-value resulting from this test indicates whether or not data are likely given the null hypothesis. Note that, the researcher's theory (i.e. M_1) is not one of the hypotheses. So, what can be learned from testing this pair of hypotheses? If the null hypothesis is not rejected testing against H_A, this does *not* mean that it would also not be rejected testing against the hypothesis of interest, i.e. M_1. If the null hypothesis is rejected testing against H_A, this does not provide any information about the fit of the hypothesis of interest, i.e. M_1. In the latter situation, post-hoc pairwise comparisons between the means provide additional information about M_1. However, several test-results have to be combined to evaluate one theory, and the pairwise comparisons may lead to inconsistent results. Both Chapter 3 and Chapter 5 of this thesis provide a more elaborate discussion of the limitations of evaluating inequality constrained models with traditional null hypothesis testing.

Null hypothesis testing with inequality constrained alternatives
It is also possible to test H_0 against an inequality constrained alternative (e.g. M_1). This is more powerful and the null hypothesis will only be rejected if the data point in the direction of the constrained alternative (i.e. post-hoc pairwise comparisons are not needed to draw conclusions about the model of interest). See Barlow, Bartholomew, Bremner and Brunk (1972) and its successor Robertson, Wright and Dykstra (1988) for comprehensive overviews. More recently, the Journal of Statistical Planning and Inference presented a special issue, titled Statistical Inference Under Inequality Constraints (2002). These works and the references contained therein give an overview of the possibilities and limitations of inequality constrained hypothesis testing in a frequentist framework. For the univariate normal linear model, tests for inequality constrained alternatives are available (see the references above). However, constrained multivariate testing problems have not yet received a lot of attention in the literature. Tests for multivariate problems that are discussed are based on asymptotic distributions of the test-statistic (Silvapulle, 1995a, 1995b; Perlman and Wu, 2002; Sasabuchi, Tanaka and Tsukamoto, 2003).

In Chapter 2 of this thesis a method to obtain the similar null distribution of both univariate and multivariate normal linear data is presented. With the similar null distribution of the data, the null distribution of any test-statistic of interest is available. In this chapter, several illustrations dealing with inequality constrained alternatives are provided, for both univariate and multivariate testing problems.

Competing theories

The researcher of the example, is probably aware of the fact that some of his fellow researchers have other expectations or theories. Some colleagues are convinced that medication is the only effective treatment, whereas other colleagues expect only a positive effect of behavioral therapy. Now we have three competing theories, or stated otherwise, three statistical models. Denoting the three models by M_1, M_2 and M_3, these are:

$$M_1 : \mu_1 < (\mu_2, \mu_3) < \mu_4,$$
$$M_2 : (\mu_1, \mu_2) < (\mu_3, \mu_4)$$
$$M_3 : (\mu_1, \mu_3) < (\mu_2, \mu_4)$$

Now, the question of interest is: *Which of the theories is the best according to the data arising from the experiment?*

It is possible to test the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ against each of the theories M_1, M_2 and M_3. This does provide information about the support in the data for each of the theories, but may not give a conclusive answer to the question stated above. It is well possible that for all three tests the null hypothesis is rejected in favor of the ordered alternative. In that scenario no information is obtained about the relative quality of the theories. To answer the research question, it is therefore better to compare the three theories directly with each other.

Developments in Bayesian statistics have rendered a framework for model selection that can be used as an alternative for the testing of hypotheses. An important difference between null hypothesis testing and Bayesian model selection is that the latter computes so-called posterior model probabilities for all models under consideration.

Chapters 3, 4 and 5 of this thesis discuss several aspects of a Bayesian approach to deal with this kind of model selection problems. In Chapter 3, also Bayesian estimation of constrained parameters is discussed.

Bayesian Estimation and Model selection

The Bayesian approach is based on the *posterior* distribution of the model parameters, i.e. the distribution of the parameters *after* observing the data. The posterior

is formed by a combination of the density of the data and the *prior* distribution of the model parameters, i.e. the distribution of the parameters *before* observing the data. With the prior distribution the knowledge (or uncertainty) about the parameters before observing data can be reflected. Gelfand, Smith and Lee (1992) showed that inequality constraints among parameters can straightforwardly be built into the prior distribution.

Estimation of constrained parameters is based on a sample from the posterior. The sample can be obtained using Markov Chain Monte Carlo sampling (see for instance Chen, Shao and Ibrahim, 2000). For the models discussed in this thesis Gibbs sampling is applied. Sampling constrained parameters using the Gibbs sampler is introduced by Gelfand, Smith and Lee (1992) and Smith and Roberts (1993). From the posterior sample, all desired parameter estimates can be computed, including standard errors and credibility intervals. Estimation of constrained parameters is discussed in Chapter 3.

A Bayesian criterion for model selection is the Bayes factor (see for instance Kass and Raftery, 1995). From Bayes factors, posterior model probabilities for each model in a finite set of models can be obtained. In Chapter 3, model selection using Bayes factors and posterior model probabilities is introduced.

The formalization of prior information in prior distributions is a difficult part of Bayesian model selection. The use of subjective prior distributions is often criticized because the resulting inferences are also subjective. Noninformative or vague priors, in general lead to arbitrary Bayes factors (Berger and Pericchi, 1996, 2001). However, competing models that are specified using different inequality constraints have a special feature. They are all nested in one unconstrained, so-called encompassing model.

A novel idea denoted by the encompassing prior approach is presented in this thesis. Encompassing priors are the main topic of Chapter 4, although they are also applied in the Chapters 3 and 5. The basic idea of encompassing priors is to specify a distribution that is noninformative with respect to the parameters for the encompassing model. Subsequently, the priors for the nested models are derived from the encompassing prior. Models M_1, M_2, and M_3 are examples of models that are nested in the corresponding encompassing model $M_e : \mu_1, \mu_2, \mu_3, \mu_4$. In Chapter 4 it is explained how the encompassing prior approach can be used to obtain Bayes factors for this set of models, that are virtually independent of the encompassing prior distribution.

Outline

This dissertation consists of four chapters that are papers submitted for publication and that can be read independently and in arbitrary order.

In Chapter 2, titled 'Obtaining Similar Null Distributions in the Normal Linear Model Using Computational Methods', a procedure is presented, that obtains the null distribution for many test-statistics in the context of normal linear data. With this procedure null hypothesis testing against inequality constrained alternatives is made available, both for univariate and multivariate constrained problems. Although the illustrations in Chapter 2 deal with inequality constrained hypotheses, the procedure is more general and can for example also be used for order statistics, or, statistics used to test the equality of variances.

The Bayesian approach to both estimation and model selection for inequality constrained parameters is introduced in Chapter 3. It is titled 'Inequality Constrained Analysis of Variance: A Bayesian Approach'. Firstly, a motivation for Bayesian model selection as opposed to null hypothesis testing is provided. Subsequently, the Bayesian approach is explained in a non-technical way (the technical details can however be found in the two appendices). The software used in the two illustrations of this chapter is made available through the web-site: www.fss.uu.nl/ms/ik. Also included are a tutorial and illustrations consisting of datasets and accompanying input files.

The fourth chapter is titled 'Encompassing Priors'. In this chapter, sensitivity to encompassing priors is examined and it is shown that for specific classes of models the selection is virtually objective, that is, independent of the encompassing prior. The encompassing prior also leads to a nice interpretation and straightforward estimation of Bayes factors, without the requirement to compute marginal likelihoods (which can be burdensome). This chapter is general, in the sense that the encompassing prior approach can also be used for models other than the normal linear model. An example is provided that deals with a contingency table with inequality constrained odds-ratios.

The goal of the final chapter, 'Comparison of Hypothesis Testing and Bayesian Model Selection', is to show the potential of posterior model probabilities and model selection as an alternative for the use of p-values in traditional hypothesis testing.

Chapter 2

Obtaining Similar Null Distributions in the Normal Linear Model Using Computational Methods

Abstract

This paper discusses hypothesis testing when the null hypothesis belongs to the univariate or multivariate normal linear model. More specifically it discusses how data can be replicated from the null distribution conditional on the sufficient statistics for the parameters of the null hypothesis at hand. It will be shown how these data replications can be used to obtain similar tests for any test statistic that is of interest. Examples of the procedure proposed will be given with respect to (robust) test statistics that have been proposed to test against inequality constrained hypotheses in the univariate normal model. For the multivariate normal model examples will be given with respect to test statistics that test against a multivariate mean vector that is coordinate-wise larger than zero, and two multivariate mean vectors that are coordinate-wise ordered.

2.1 Introduction

This paper presents a method that obtains the similar null distribution of the data, for null hypotheses that have the form of a normal linear model. Note that the expression similar null distribution is not standard terminology. What we mean is that the expected proportion of rejections of the null hypothesis, computed with respect to this distribution is level alpha for any test (Lehmann, 1997, p. 140, formula (7)). The aim of the paper is therefore not to present a new test or test statistic, but to show that the similar null distribution of the data renders similar tests for any test statistic. Subsequently the univariate (2.1), and multivariate normal linear model (2.7) will be considered.

Let y_i $(i = 1, \ldots, N)$ denote the criterion variable and x_{pi} the p-th continuous predictor $(p = 1, \ldots, P)$. In the null hypothesis $k = 1, \ldots, K$ groups can be distinguished. Group membership is denoted by d_{ki}, where $d_{ki} = 1$ indicates that the i-th unit is a member of group k, and $d_{ki} = 0$ that the i-th unit is not a member of group k. The general univariate normal null hypothesis is

$$y_i = \sum_{k=1}^{K} \beta_k d_{ki} + \sum_{p=1}^{P} \alpha_p x_{pi} + \varepsilon_i, \qquad \text{with } \varepsilon_i \overset{iid}{\sim} N(0,\ \sigma^2). \tag{2.1}$$

Note that β_k denotes the (adjusted) mean in group k, that α_p is the regression coefficient relating the p-th continuous predictor to the dependent variable, and that σ^2 is the residual variance.

Specific null hypotheses are obtained by choosing appropriate values for K, and P. Consider the following three examples:

1. A simple null hypothesis

$$y_i = \beta_1 + \varepsilon_i, \qquad \text{with } \varepsilon_i \overset{iid}{\sim} N(0,\ \sigma^2), \tag{2.2}$$

is obtained for (2.1) with $K = 1$, $d_{1i} = 1$ for $i = 1, \ldots, N$, and $P = 0$. Testing this null against an order restricted alternative of the form

$$y_i = \sum_{j=1}^{J} \beta_j^* d_{ji} + \varepsilon_i, \qquad \text{with } \varepsilon_i \overset{iid}{\sim} N(0,\ \sigma^2*), \tag{2.3}$$

with $\beta_1^* \leq ... \leq \beta_J^*$ can be interesting. Note that the * will be used to denote parameters from the alternative hypothesis. These kind of hypotheses have received a lot of attention in the literature. See, Barlow, Bartholomew, Bremner and Brunk (1972), Robertson, Wright and Dykstra (1988), and the

Journal of Statistical Planning and Inference (2002, issue 107) for comprehensive overviews. Invariably, so called 'level-probabilities' are needed in order to be able to derive the null distribution of likelihood ratio statistics with which (2.2) can be tested against inequality constrained or order restricted alternatives. Our approach provides an alternative for the use of level probabilities.

2. Continuing the previous example, Silvapulle (1992a, 1992b) presents test statistics that are robust against 'long-tailed error distributions which may not be symmetric'. For these tests only the asymptotic distribution is known, whereas our approach will render the finite sample or similar distribution.

3. The null hypothesis with $P = 0$, $K = 4$ and unknown but common variance σ^2 could be tested against an alternative of the form

$$y_i = \sum_{k=1}^{K} \beta_k^* d_{ki} + \varepsilon_i, \qquad \text{with } \varepsilon_i \overset{iid}{\sim} N(0, \sum_{k=1}^{K} d_{ki}\sigma_k^2*), \tag{2.4}$$

that is,

$$H_0 : \sigma_1^2 = \ldots = \sigma_K^2 = \sigma^2 \quad \text{versus} \quad H_1 : \text{"not} \quad H_0\text{"}. \tag{2.5}$$

It is well-known that the classical tests for homogeneity of variances, e.g. F-test, Bartlett's test (Bartlett and Kendall, 1946), are highly sensitive to the assumption that population distributions are normal. Therefore, more robust test statistics have been developed, e.g. the Levene/Brown-Forsythe test statistic (Brown and Forsythe, 1974), which is approximately F-distributed. Recently Shoemaker (2003) proposed two other statistics for testing the homogeneity for two or more variances, that further improve the robustness properties. However, for these test statistics, like for the Levene/Brown-Forsythe test, only the asymptotic distribution is known. Our approach will render the finite sample distribution.

To compute a p-value for null hypotheses that are a specification of (2.1), two ingredients are needed in addition to the observed data $\boldsymbol{y}$: the distribution $g(\boldsymbol{z}|H_0)$, where $\boldsymbol{z}$ denotes a data-matrix from the null population, and, a test statistic T. Combination of the two ingredients renders the distribution of the test statistic: $g(T(\boldsymbol{z})|H_0)$. Evaluation of $T(\boldsymbol{y})$ using $g(T(\boldsymbol{z})|H_0)$ renders a p-value. However, since (2.1) contains two or more nuisance parameters, the derivation of $g(T(\boldsymbol{z})|H_0)$ can be complicated. Therefore, for this kind of composite null hypotheses, often null distributions are approximated, leading to asymptotic p-values (Bayarri and Berger, 2000; Robins, Van der Vaart and Ventura, 2000).

In the normal linear model $g(\boldsymbol{z}|\ H_0)$ and $g(T(\boldsymbol{z})\ |\ H_0)$ can be replaced by the distribution of $\boldsymbol{z}$ and $T(\boldsymbol{z})$, respectively, conditional on the observed sufficient statistics for the nuisance parameters of the null hypothesis. Denote the observed sufficient statistics for the parameters of (2.1) by $S_r(\boldsymbol{y})$ $(r = 1, \ldots, R)$. The p-value $P(T(\boldsymbol{z}) \geq T(\boldsymbol{y})|S_1(\boldsymbol{z}) = S_1(\boldsymbol{y}), \cdots, S_R(\boldsymbol{z}) = S_R(\boldsymbol{y}))$ is computed with respect to the distribution

$$g(\boldsymbol{z}|S_1(\boldsymbol{z}) = S_1(\boldsymbol{y}), \cdots, S_R(\boldsymbol{z}) = S_R(\boldsymbol{y})) = c, \tag{2.6}$$

where $g(\boldsymbol{z}|\cdot) = c$ implies that conditional on the sufficient statistics each data vector $\boldsymbol{z}$ has the same density (Lindgren, 1993, p. 229). This procedure renders similar tests (Lehmann, 1997, p. 140), that is, $P(T(\boldsymbol{z}) \geq T(\boldsymbol{y})|S_1(\boldsymbol{z}) = S_1(\boldsymbol{y}), \cdots, S_R(\boldsymbol{z}) = S_R(\boldsymbol{y})) = P(T(\boldsymbol{z}) \geq T(\boldsymbol{y})|H_0)$.

Although the concept of similar tests is well known, we have (for normal linear models) not been able to find references to procedures with which data can be generated from (2.6). In Sections 2 and 3, for the simple null hypothesis (2.2) and the general univariate normal null hypothesis (2.1), it will be shown how the null distribution of the data can be obtained by repeatedly sampling vectors $\boldsymbol{z}$ from (2.6). Subsequent computation of $T(\boldsymbol{z})$ for each $\boldsymbol{z}$ renders the null distribution of the test statistic.

For the simple univariate null hypothesis the procedure will be illustrated using (2.2) and (2.3). To highlight the flexibility of our procedure, two different likelihood ratio tests will be used. The traditional test statistic is based on the mean of the observations for each group that is distinguished under the null and alternative hypotheses. In the second test statistic the means are replaced by medians, resulting in a test which is robust with respect to outliers, and is related to the tests described by Silvapulle (1992a, 1992b).

For the general univariate null hypothesis the procedure will be illustrated using the null hypothesis $y_i = \beta_1 + \alpha_1 x_{1i} + \varepsilon_i$, and the alternative hypothesis $y_i = \sum_{j=1}^{J} \beta_j^* d_{ji} + \alpha_1^* x_{1i} + \varepsilon_i$, with $\beta_1^* < ... < \beta_J^*$. This amounts to testing in an analysis of covariance whether the adjusted means are equal, or ordered. As far as we know, this specific example has not yet received any attention in the literature, although it fits in the general framework presented by Silvapulle (1995a, 1995b). To illustrate the flexibility of our approach, expected a posteriori estimates instead of maximum likelihood estimates will be used for the parameters of both the null and alternative hypothesis. For inequality constrained models, these estimates are easily obtained using the Gibbs sampler. They are an alternative if it is difficult to obtain maximum likelihood estimates.

In Section 4, the multivariate counterpart of (2.1) will be discussed:

$$\boldsymbol{y}_i = \sum_{k=1}^{K} \boldsymbol{\beta}_k d_{ki} + \sum_{p=1}^{P} \boldsymbol{\alpha}_p x_{pi} + \boldsymbol{\varepsilon}_i, \qquad \text{with } \boldsymbol{\varepsilon}_i \overset{iid}{\sim} N(0, \boldsymbol{\Sigma}) \tag{2.7}$$

where, $\boldsymbol{y}_i$, $\boldsymbol{\alpha}_p$, $\boldsymbol{\beta}_k$, and $\boldsymbol{\varepsilon}_i$ are vectors of length Q, and $\boldsymbol{\Sigma}$ is a $Q \times Q$ covariance matrix. It will be described how the multivariate generalization of $\boldsymbol{z}$, that is, the $N \times Q$ matrix $\boldsymbol{Z}$ can be sampled from the multivariate counterpart of (2.6). A Gibbs sampler will be used that is based on a repeated application of the univariate sampling procedure to each of the Q dependent variables.

Specific multivariate null hypotheses are obtained by choosing appropriate values for K, P and Q. Consider the following three examples:

1. The null hypothesis with $K = 1$, $d_{1i} = 1$ for $i = 1, \ldots, N$, $P = 0$ and $Q \geq 2$ can be tested against the alternative hypothesis with $J \geq 2$

$$\boldsymbol{y}_i = \sum_{j=1}^{J} \boldsymbol{\beta}_j^* d_{ji} + \boldsymbol{\varepsilon}_i, \qquad \text{with } \boldsymbol{\varepsilon}_i \overset{iid}{\sim} N(0, \boldsymbol{\Sigma}^*). \tag{2.8}$$

   This is a classic multivariate testing problem for which a number of test-statistics have been developed: Wilk's lambda, Roy's largest root, the Hotelling-Lawley trace and Pillai-Bartlett trace (see, for example, Tatsuoka, 1988, pp. 88-95 and 285-288; Stevens, 1996, pp. 225-226). Many references to F-approximations of the null distributions of these statistics can be found in the literature. However, the finite sample null distribution is not known for each of these statistics (see, for example, Muller, 1998). Our approach can be used to obtain the finite sample or similar null distribution.

2. Constrained multivariate testing problems have not received a lot of attention in the literature. The papers written by Silvapulle (1995a, 1995b), Follmann (1996), Chongcharoen, Singh and Wright (2002), and Perlman and Wu (2002), and the references contained therein give an overview of the state of the art with respect to multivariate one-sided alternatives. Their null hypothesis is (2.7) with $K = 0$, $P = 0$ and $Q \geq 2$. Their alternative hypothesis is

$$\boldsymbol{y}_i = \boldsymbol{\beta}^* + \boldsymbol{\varepsilon}_i, \qquad \text{with } \boldsymbol{\varepsilon}_i \overset{iid}{\sim} N(0, \boldsymbol{\Sigma}^*), \tag{2.9}$$

   with $\boldsymbol{\beta}^* \geq 0$, where the latter is interpreted coordinate-wise. For most of the test-statistics discussed the finite sample null distribution is unknown, and

otherwise the test-statistics have undesirable properties. See Perlman and Wu (2002) for a discussion of Tang's test (Tang, 1994) and Chongcharoen, Singh and Wright (2002) for an evaluation of Follman's test (Follmann, 1996). In Section 2.4.2 it will be shown how our approach will render the similar null distribution for any test statistic.

3. Consider the null hypothesis (2.7) with $K = 1$, $d_{1i} = 1$ for $i = 1, \ldots, N$, $P = 0$ and $Q \geq 2$, and the alternative

$$\boldsymbol{y}_i = \boldsymbol{\beta}_1^* d_{1i} + \boldsymbol{\beta}_2^* d_{2i} + \boldsymbol{\varepsilon}_i, \qquad \text{with } \boldsymbol{\varepsilon}_i \overset{iid}{\sim} N(0, \boldsymbol{\Sigma}^*), \tag{2.10}$$

with $\boldsymbol{\beta}_1^* \geq \boldsymbol{\beta}_2^*$, where the latter has to be interpreted coordinate-wise. This null and alternative hypothesis are interesting when evaluating the effect of an intervention (represented by two or more outcome variables) using a control and an experimental group. Note that, the covariance matrix $\boldsymbol{\Sigma}$ is assumed to be common, but unknown. We found just one reference with respect to the evaluation of these hypotheses in the literature. Sasabuchi, Tanaka and Tsukamoto (2003) present a likelihood ratio test and a method to compute the supremum of its upper tail probability under the null hypothesis, i.e. the finite sample null distribution of the test statistic is not known. With our approach the finite sample null distribution of any test statistic, that can be used to handle this situation, is made available. In Section 2.4.3 an illustration of testing against (2.10) will be presented.

This paper will be concluded with a discussion in which the possibilities of the procedure proposed will be summarized.

2.2 The Simple Null Hypothesis

2.2.1 Replication of Data under the Simple Null Hypothesis

Model (2.1) with $K = 1$ and $P = 0$ leads to the simple null hypothesis (2.2). The sufficient statistics for the nuisance parameters (β_1 and σ^2) of this null hypothesis are $S_1(\boldsymbol{y}) = \sum_{i=1}^{N} y_i$ and $S_2(\boldsymbol{y}) = \sum_{i=1}^{N} y_i^2$. The null distribution of the data conditional on the observed values of the sufficient statistics is not available in closed form, but data $\boldsymbol{z} = [z_1, \ldots, z_N]$ can be replicated from

$$g(\boldsymbol{z} | \sum_{i=1}^{N} z_i = \sum_{i=1}^{N} y_i, \sum_{i=1}^{N} z_i^2 = \sum_{i=1}^{N} y_i^2) = c, \tag{2.11}$$

where the constant c implies that conditional on the sufficient statistics each data vector $\boldsymbol{z}$ has the same density (Lindgren, 1993, p. 229). In this section replication of a vector $\boldsymbol{z}$ from (2.11) will be explained. The replicated vectors $\boldsymbol{z}$ give a discrete representation of the desired distribution (2.11). The precision of this representation increases with the number of replications and can therefore be controlled by the user. As will be shown, for numbers of replications that are easily obtained with modern personal computers, the obtained degree of precision is very high. Subsequently, in Section 2.2.2, it will be shown how the null distribution of the data, i.e. the set of replicated vectors $\boldsymbol{z}$, can be used for testing the null (2.2) against the inequality constrained hypothesis (2.3).

To replicate a vector $\boldsymbol{z}$ from (2.11) amounts to randomly choosing any vector for which $\sum_{i=1}^{N} z_i = \sum_{i=1}^{N} y_i$ and $\sum_{i=1}^{N} z_i^2 = \sum_{i=1}^{N} y_i^2$. By solving $\sum_{i=1}^{N} z_i = \sum_{i=1}^{N} y_i$ for z_N and substituting $z_N = \sum_{i=1}^{N} y_i - \sum_{i=1}^{N-1} z_i$ in $\sum_{i=1}^{N} z_i^2 = \sum_{i=1}^{N} y_i^2$ the following condition is obtained:

$$\sum_{i=1}^{N-1} z_i^2 + \sum_{i=1}^{N-2} \sum_{l=i+1}^{N-1} z_i z_l - \sum_{i=1}^{N} y_i \sum_{i=1}^{N-1} z_i = \frac{1}{2} \sum_{i=1}^{N} y_i^2 - \frac{1}{2} \left(\sum_{i=1}^{N} y_i \right)^2, \tag{2.12}$$

which is the equation of an ellipsoid in $\mathbb{R}^{N-1}$. Note that (2.12) is a function of $\boldsymbol{z}^- = [z_1, \ldots, z_{N-1}]$. To replicate a data vector $\boldsymbol{z}^-$, a point on this ellipsoid is needed, where each point has an equal probability of being sampled. Note also, that when a vector $\boldsymbol{z}^-$ is obtained, the value for z_N can subsequently be computed using $z_N = \sum_{i=1}^{N} y_i - \sum_{i=1}^{N-1} z_i$.

It will be shown that (2.12) is equal to the formula for isodensity ellipsoids of a multivariate normal distribution of dimension $N-1$ and this property will be used in the sampling scheme. To obtain an arbitrary point $\boldsymbol{z}^-$ on the ellipsoid, the multivariate normal distribution corresponding to (2.12) must be defined. The general formula for isodensity ellipsoids of $N_{N-1}(\boldsymbol{\mu^-}, \boldsymbol{\Sigma^-})$ is

$$(\boldsymbol{z}^- - \boldsymbol{\mu}^-)'(\boldsymbol{\Sigma}^-)^{-1}(\boldsymbol{z}^- - \boldsymbol{\mu}^-) = d, \tag{2.13}$$

where each different value of the constant d specifies a different ellipsoid (Tatsuoka, 1988, p. 70). The notation $\boldsymbol{\mu}^-$ and $\boldsymbol{\Sigma}^-$ is used to emphasize that the dimension of the multivariate normal equals the dimension of $\boldsymbol{z}^-$.

Denote $(\boldsymbol{\Sigma}^-)^{-1} = \boldsymbol{\Omega}^-$, with elements ω_{il}. Dropping the matrix notation, (2.13) can be written as

$$\sum_{i=1}^{N-1}\omega_{ii}z_i^2+\sum_{i=1}^{N-2}\sum_{l=i+1}^{N-1}2\omega_{il}z_iz_l-\sum_{i=1}^{N-1}\sum_{l=1}^{N-1}2\omega_{il}\mu_lz_i$$
$$=d-\sum_{i=1}^{N-1}\omega_{ii}\mu_i^2-\sum_{i=1}^{N-2}\sum_{l=i+1}^{N-1}2\omega_{il}\mu_i\mu_l. \tag{2.14}$$

Note that (2.14) is also a function of $\boldsymbol{z}^-$.

Comparing (2.14) with (2.12) it can be seen that for the simple null hypothesis

$$\boldsymbol{\mu}^-=\left[\frac{1}{N-1}\sum_{i=1}^{N}y_i,\quad\ldots,\quad\frac{1}{N-1}\sum_{i=1}^{N}y_i\right], \tag{2.15}$$

$$\boldsymbol{\Omega}^-=\begin{bmatrix}1&\frac{1}{2}&\cdots&&\frac{1}{2}\\\frac{1}{2}&1&&&\\\vdots&&\ddots&&\vdots\\&&&1&\frac{1}{2}\\\frac{1}{2}&&\cdots&\frac{1}{2}&1\end{bmatrix}, \tag{2.16}$$

and

$$d=\frac{1}{2}\sum_{i=1}^{N}y_i^2+\frac{1}{2N-2}\left(\sum_{i=1}^{N}y_i\right)^2. \tag{2.17}$$

As will be explained hereafter, to replicate a vector $\boldsymbol{z}^-$ on the ellipsoid that is specified by (2.12), the value of d, that is (2.17), is not required. Therefore, in the sequel the derivation of d will be omitted.

The procedure to obtain a point from exactly the ellipsoid that is specified by (2.12), consists of the following steps:

1. A point $\boldsymbol{p}^-=[p_1,\ldots,p_{N-1}]$ is sampled from $N(\boldsymbol{\mu}^-,\boldsymbol{\Sigma}^-)$ with $\boldsymbol{\mu}^-$ as specified in (2.15), and $\boldsymbol{\Sigma}^-$ via $\boldsymbol{\Omega}^-$ as specified in (2.16). This point $\boldsymbol{p}^-$ belongs to one of the many isodensity ellipsoids, but typically not the required one, as specified in (2.12). This is illustrated in Figure 2.1, for $N=4$. The inner ellipsoid in this figure represents (2.12) but the sampled point $\boldsymbol{p}^-$ belongs to the outer ellipsoid.

2. A line through the sampled point $\boldsymbol{p}^-$ and the centroid of the ellipsoids $\boldsymbol{\mu}^-$ is drawn (see Figure 2.1). This line has two intersections with the required

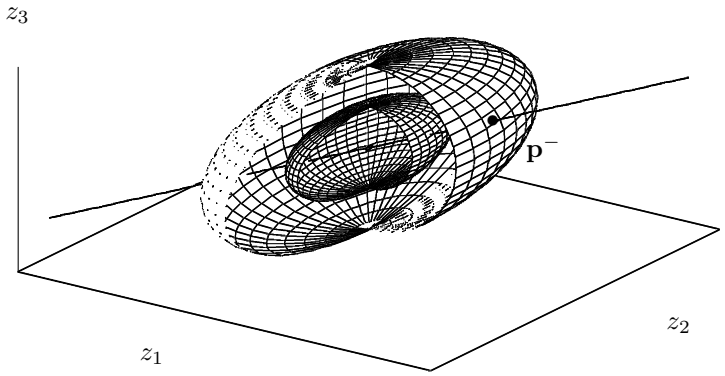

Figure 2.1: Isodensity Ellipsoids of a Three Dimensional Normal Distribution.

ellipsoid, i.e. the inner ellipsoid in Figure 2.1. One of the intersections (each with a probability .5) renders a vector $\boldsymbol{z}^-$. Since $\boldsymbol{z}^-$ must be on the line

$$\boldsymbol{z}^- = \boldsymbol{\mu}^- + \lambda(\boldsymbol{p}^- - \boldsymbol{\mu}^-). \tag{2.18}$$

Note that (2.18) consists of (N–1) equations and N unknowns ($\boldsymbol{z}^-$ and λ) and that (2.12) adds the N-th equation. Solving (2.12) and (2.18) for λ renders two solutions, from which one random choice is inserted in (2.18) to compute $\boldsymbol{z}^-$.

3. The last element is computed using $z_N = \sum_{i=1}^{N} y_i - \sum_{i=1}^{N-1} z_i$.

Going through steps 1 to 3 renders one replicated data vector $\boldsymbol{z}$ from (2.11). The number of replications determines the precision with which (2.11) is represented. How large the number of replications must be to obtain a high precision will be shown in the applications in this paper (Sections 2.2.3, 2.4.2 and 2.4.3).

2.2.2 Testing the Simple Null Hypothesis against an Ordered Alternative

After sampling the null distribution of the data it can be used for testing H_0 : $y_i \overset{iid}{\sim} N(\beta_1, \sigma^2)$ against any alternative hypothesis. An appropriate test statistic against (2.3), i.e. the mean scores of J subgroups are increasing, is $T(\boldsymbol{y})= \overline{E}^2(\boldsymbol{y})$, which is based on the likelihood ratio principle (Barlow et al., 1972, p. 121). Under the null hypothesis β_1 is estimated using maximum likelihood. The maximum likelihood estimators of the means of the subgroups under the ordered alternative, β_j^* $(j = 1, \ldots, J)$, are obtained using the pool adjacent violator algorithm and are called the isotonic means (Barlow et al., 1972, pp. 13-18). Equation (2.19) shows the test statistic:

$$\overline{E}^2(\boldsymbol{y}) = \sum_{j=1}^{J} N_j(\hat{\beta}_j^* - \hat{\beta}_1)^2 / \sum_{i=1}^{N}(y_i - \hat{\beta}_1)^2, \tag{2.19}$$

where $j = 1, \ldots, J$ denotes the subgroups under the alternative hypothesis and N_j the number of persons in subgroup j. In this equation $\hat{\beta}_1$ is the estimated overall mean (H_0) and $\hat{\beta}_j^*$ are the estimated isotonic means (H_1).

To illustrate the flexibility of our approach, a modification of (2.19), that provides a robust test statistic (Silvapulle, 1992a, 1992b) is also included and will be called the $\overline{E}_M^2$-test. The test statistic $\overline{E}_M^2(\boldsymbol{y})$ is computed using (2.19) but instead of the sample *mean*, the sample *median* is used to estimate $\hat{\beta}_1$. Similarly, the isotonic means $\hat{\beta}_j^*$ are estimated via an application of the pool adjacent violator algorithm to the sample *medians* per group.

Observed data will provide a value for $\overline{E}^2(\boldsymbol{y})$ and $\overline{E}_M^2(\boldsymbol{y})$. The probability $P(T(\boldsymbol{z}) \geq T(\boldsymbol{y})|H_0)$ computed with respect to (2.11) renders a similar p-value for each of the two test statistics. A sample from (2.11) is obtained using the procedure presented in the previous section.

2.2.3 Illustration

The data for this illustration were taken from Gelman, Carlin, Stern and Rubin (1995, pp. 213-215) and give the mean daily milk fat produced (y_i) and the level of an additive (methionine hydroxy analog) to the diet of N=50 cows. Four treatment groups with increasing levels of the additive are compared. The null hypothesis of this model is $H_0 : y_i \overset{iid}{\sim} N(\beta_1, \sigma^2), \quad i = 1, ..., N$. To obtain the null distribution of the data, vectors $\boldsymbol{z}$ are replicated conditional on the sufficient statistics of the observed data. Two alternative hypotheses are tested against the null:

Table 2.1: P-values for Testing Against H_1 and H_2 Using Regular and Robust Test Statistics.

	H_0/H_1	H_0/H_2
LR	.174	
LR_M	.029	
$\overline{E}^2$		.039
$\overline{E}^2_M$		.005

$H_1 : \beta_1^*, \beta_2^*, \beta_3^*, \beta_4^*$ and $H_2 : \beta_1^* \leq \beta_2^* \leq \beta_3^* \leq \beta_4^*$ (with at least one strict inequality). To test against H_1 two test statistics are used, a standard likelihood ratio test based on sample means (denoted by LR) and one based on sample medians (denoted by LR_M). To test against H_2 the test statistics $\overline{E}^2$ and $\overline{E}^2_M$ are used. The results can be found in Table 2.2.3.

Testing H_0 against H_1 using the LR test statistic results in $p = .174$, which is equal to the p-value rendered by a traditional analysis of variance ($F(3, 46) = 1.73, p = .174$). Testing the same hypotheses with the robust LR_M-test renders $p = .029$. The difference between the two test results can be explained by the differences between the observed means (under H_1 respectively 194.6, 209.5, 213.0, 235.4) and medians (182.2, 207.2, 205.4, 248.8). Note that both tests are similar and that the different results are a consequence of the statistic used to test the hypotheses. Here, using the traditional likelihood ratio test the null hypothesis is not rejected, but using the robust test LR_M it is. We will not discuss robust testing here (it would go beyond the topic of this paper), but apparently this is an issue that deserves further examination (see also, for example, Silvapulle, 1992a, 1992b). However, we do like to stress that comparing the effect of several test statistics becomes very easy, when our method to obtain the similar null distribution of the data is applied. With this distribution the finite sample distribution of any test statistic is made available.

Testing H_0 against H_2 results in $p = .039$ for the $\overline{E}^2$-test and $p = .005$ for the $\overline{E}^2_M$-test. Again, a considerable difference between the results of the two test statistics can be seen. For both test statistics, $\overline{E}^2$ and $\overline{E}^2_M$, the conclusion is that increasing the level of the additive to the diet of the cows increases the milk fat production.

The precision of the p-value as a function of the number of replicated vectors $\boldsymbol{z}$ from (2.11) will be examined. In Figure 2.2 this is illustrated for just one of the

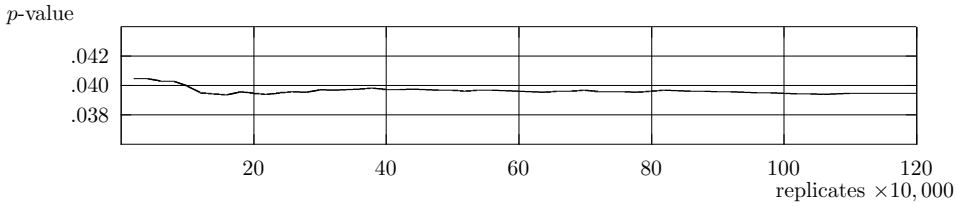

Figure 2.2: Testing H_0 Against H_2 Using the $\overline{E}^2$-test.

four tests. For testing H_0 against H_2 using the $\overline{E}^2$-test, 1,200,000 vectors $\boldsymbol{z}$ are replicated and the p-value is plotted against the number of replications. Note that the scale is chosen such that very small (third decimal) differences in successive p-values are visible. As can be seen from the plot, to get a stable third decimal in this example about 200,000 replications are sufficient. Furthermore (although not visible in the plot) a precision of two decimals is obtained using about 20,000 replications.

2.3 The General Univariate Null Hypothesis

2.3.1 Replication of Data under the General Univariate Null Hypothesis

In this section the method to obtain the null distribution for the general univariate null hypothesis (2.1) will be explained. For the nuisance parameters of (2.1) $R = K + P + 1$ sufficient statistics exist. The density of the data conditional on the sufficient statistics $S_1(\boldsymbol{y}), \ldots, S_R(\boldsymbol{y})$ is the similar null distribution of the data. To obtain this distribution, data $\boldsymbol{z}= [z_1, \ldots, z_N]$ is replicated from

$$g(\boldsymbol{z}|S_1(\boldsymbol{z}) = S_1(\boldsymbol{y}),\ \ldots,\ S_R(\boldsymbol{z}) = S_R(\boldsymbol{y})) = c, \tag{2.20}$$

by execution of the following steps:

1. (2.20) is specified using

   (a) For $r = 1, \ldots, K$

$$S_r(\boldsymbol{z}) = \sum_{i=1}^{N} z_i d_{ri} = \sum_{i=1}^{N} y_i d_{ri}. \tag{2.21}$$

(b) For $r = K+1, \ldots, K+P$

$$S_r(\boldsymbol{z}) = \sum_{i=1}^{N} z_i x_{(r-K),i} = \sum_{i=1}^{N} y_i x_{(r-K),i}. \tag{2.22}$$

(c) For $r = K+P+1$

$$S_r(\boldsymbol{z}) = \sum_{i=1}^{N} z_i^2 = \sum_{i=1}^{N} y_i^2. \tag{2.23}$$

Equations (2.21), (2.22) and (2.23) together consist of $K+P+1$ functions of $\boldsymbol{z}$, where $\boldsymbol{z}$ consists of N elements (z_i), i.e. N unknowns. Note, that the $K+P$ equations of (2.21) and (2.22) are all linear functions of $\boldsymbol{z}$. Each of these $K+P$ linear equations is solved for one specific z_i as will be explained hereafter. The observations are assumed to be ordered by group, i.e if group $k = 1$ consists of N_1 observations, they have indices $i = 1, \ldots, N_1$ and the first observation of group $k = 2$ has index $i = N_1+1$. To denote the number of observations in group k, the notation N_k is used. To denote the last observation of group k, the notation N_{k+} is used. Stated otherwise, $N_{k+} = \sum_{k'=1}^{k} N_{k'}$ is the cumulative sample size of the first k groups.

2. Using the notation introduced in step 1:

(a) (2.21) implies that, for $k = 1, \ldots, K$

$$z_{N_{k+}} = \sum_{i=1}^{N} y_i d_{ki} - \sum_{i \in \mathcal{A}_k} z_i d_{ki}, \tag{2.24}$$

with $\mathcal{A}_k = \{1, \ldots, N_{k+}-1, N_{k+}+1, \ldots, N\}$ for $k = 1, \ldots, K-1$ and $\mathcal{A}_k = \{1, \ldots, N-1\}$ for $k = K$.

(b) (2.22) implies that, for $p = 1, \ldots, P$

$$z_{N-p} = (\sum_{i=1}^{N} y_i x_{pi} - \sum_{i \in \mathcal{B}_p} z_i x_{pi})/x_{p,(N-p)}, \tag{2.25}$$

with $\mathcal{B}_p = \{1, \ldots, N-p-1, N-p+1, \ldots, N\}$ for $p = 2, \ldots, P$ and $\mathcal{B}_p = \{1, \ldots, N-2, N\}$ for $p = 1$.

3. Using a stepwise elimination by substitution procedure, the $K+P$ linear equations from the previous step, render solutions for $K+P$ elements of $\boldsymbol{z}$. Note, that in principle it is arbitrary for which elements z_i the equations are solved. However to avoid divisions by zero in the substitution process, the ordering of the observations in the last group, where P equations are solved for the elements z_{N-p} $(p = 1, \ldots, P)$, must be such that $x_{p,(N-p)} \neq x_{p,N}$, $\forall\ p$.

   Substitution of the linear solutions from (2.24) and (2.25) in the quadratic form (2.23) always renders an equation with only quadratic terms z_i^2, products $z_i z_l$ $(i \neq l)$, single terms z_i and terms that are not a function of $\boldsymbol{z}$. The resulting equation is a function of $\boldsymbol{z}^-$, where $\boldsymbol{z}^-$ has dimension $N-K-P$, since $K+P$ elements are eliminated.

   It can be written as

$$\sum_{i\in\mathcal{C}} c_{ii} z_i^2 + \sum_{i\in\mathcal{C}} \sum_{l>i\in\mathcal{C}} c_{il} z_i z_l - \sum_{i\in\mathcal{C}} b_i z_i = a, \tag{2.26}$$

   with $\mathcal{C} = \{\mathcal{A}_1 \bigcap \ldots \bigcap \mathcal{A}_K \bigcap \mathcal{B}_1 \bigcap \ldots \bigcap \mathcal{B}_P\}$. Note that (2.26) is the general form of (2.12). The constants c_{ii}, c_{il}, b_i and a are based on the observed values of the sufficient statistics, and on the observed values of x_{pi}. Furthermore, (2.26) is an ellipsoid in $\mathbb{R}^{N^-}$, with dimension $N^- = N-K-P$.

4. The equation of an isodensity ellipsoid of a multivariate normal distribution $N_{N-K-P}(\boldsymbol{\mu}^-, \boldsymbol{\Sigma}^-)$, with $(\boldsymbol{\Sigma}^-)^{-1} = \boldsymbol{\Omega}^-$ can, analogous to (2.14), be written as

$$\sum_{i\in\mathcal{C}} \omega_{ii} z_i^2 + \sum_{i\in\mathcal{C}} \sum_{l>i\in\mathcal{C}} 2\omega_{il} z_i z_l - \sum_{i\in\mathcal{C}} \sum_{l\in\mathcal{C}} 2\omega_{il}\mu_l z_i =$$

$$d - \sum_{i\in\mathcal{C}} \omega_{ii}\mu_i^2 - \sum_{i\in\mathcal{C}} \sum_{l>i\in\mathcal{C}} 2\omega_{il}\mu_i\mu_l. \tag{2.27}$$

   Comparing (2.26) with (2.27) it can be seen that for the general null hypothesis $\omega_{ii} = c_{ii}$, $\omega_{il} = \frac{1}{2}c_{il}$ and the values for μ_l $(l \in \mathcal{C})$ can be derived from the set of equations

$$b_i = \sum_{l\in\mathcal{C}} 2\omega_{il}\mu_l \qquad \text{for} \quad i \in \mathcal{C}, \tag{2.28}$$

   Note that there are $N-K-P$ equations in (2.28), and an equal number of parameters to solve, i.e. μ_l for $l \in \mathcal{C}$.

5. Sampling $\boldsymbol{p}^- = [p_1, \ldots, p_{N^-}]$ from $N(\boldsymbol{\mu}^-, \boldsymbol{\Sigma}^-)$, with $\boldsymbol{\mu}^-$ and $\boldsymbol{\Sigma}^-$ as specified in step 4, and computation of the intersections of the line through $\boldsymbol{p}^-$ and $\boldsymbol{\mu}^-$ with the required ellipsoid using (2.18) and (2.26), renders a vector $\boldsymbol{z}^- = [z_1, \ldots, z_{N^-}]$. The remaining $K + P$ observations can be computed using the $K + P$ equations contained in (2.24) and (2.25).

6. The previous step renders one vector $\boldsymbol{z}$ from (2.20). Repeating it many times renders (2.20).

2.3.2 Illustration: Analysis of Covariance with Inequality Constraints on the Adjusted Means

In this illustration the general null hypothesis (2.1), with $K=1$, $P=1$ and $d_{1i} = 1$ for $i = 1, \ldots, N$

$$y_i = \hat{y}_i + \varepsilon_i = \beta_1 + \alpha_1 x_{1i} + \varepsilon_i, \qquad \text{with } \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \tag{2.29}$$

is tested against the alternative with $J = 3$ and $Q = 1$

$$y_i = \hat{y}_i^* + \varepsilon_i = \sum_{j=1}^{3} \beta_j^* d_{ji} + \alpha_1^* x_{1i} + \varepsilon_i, \qquad \text{with } \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2 *), \tag{2.30}$$

with the constraint $\beta_1^* \leq \beta_2^* \leq \beta_3^*$ (at least one strict inequality), i.e. the adjusted means, after controlling for the covariate x_1 (in the sequel the subscript 1 is omitted for notational convenience), are increasing. We found no references in the literature to ordered tests of this kind.

For the K+P+$1 = 3$ nuisance parameters β_1, α_1 and σ^2, three sufficient statistics exist. Vectors $\boldsymbol{z}$ are replicated from $g(\boldsymbol{z}|S_1(\boldsymbol{z}) = S_1(\boldsymbol{y}), S_2(\boldsymbol{z}) = S_2(\boldsymbol{y}), S_3(\boldsymbol{z}) = S_3(\boldsymbol{y})) = c$, using:

$$S_1(\boldsymbol{z}) = \sum_{i=1}^{N} z_i = \sum_{i=1}^{N} y_i, \tag{2.31}$$

$$S_2(\boldsymbol{z}) = \sum_{i=1}^{N} z_i x_i = \sum_{i=1}^{N} y_i x_i, \tag{2.32}$$

and

$$S_3(\boldsymbol{z}) = \sum_{i=1}^{N} z_i^2 = \sum_{i=1}^{N} y_i^2. \tag{2.33}$$

The implication of (2.31) is

$$z_N = \sum_{i=1}^{N} y_i - \sum_{i \in \mathcal{A}_1} z_i, \tag{2.34}$$

with $\mathcal{A}_1 = \{1, \ldots, N-1\}$.
The implication of (2.32) is

$$z_{N-1} = (\sum_{i=1}^{N} y_i x_i - \sum_{i \in \mathcal{B}_1} z_i x_i)/x_{N-1}, \tag{2.35}$$

with $\mathcal{B}_1 = \{1, \ldots, N-2, N\}$.
Solving (2.34) and (2.35) for z_N and z_{N-1}, and substituting the solution in (2.33), renders an ellipsoid in $\mathbb{R}^{N-2}$:

$$\begin{aligned}
&\sum_{i \in \mathcal{C}} (1 + {h_i}^2 - h_i)\ z_i^2 \ + \ \sum_{i \in \mathcal{C}} \sum_{l > i \in \mathcal{C}} (1 + 2h_i h_l - h_i - h_l)\ z_i z_l \\
&+ \sum_{i \in \mathcal{C}} \ (\sum_{i'=1}^{N} y_{i'} - m_1 + m_2 + (2m_1 - 2m_2 - \sum_{i'=1}^{N} y_{i'})\ h_i)\ z_i \\
&= \frac{1}{2}(\sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2) - (m_1 - m_2)^2 + (m_1 - m_2) \sum_{i=1}^{N} y_i,
\end{aligned} \tag{2.36}$$

with $h_i = \frac{x_i - x_N}{x_{N-1} - x_N}$, $h_l = \frac{x_l - x_N}{x_{N-1} - x_N}$, $m_1 = \frac{\sum_{i=1}^{N} x_i y_i}{x_{N-1} - x_N}$, $m_2 = \frac{x_N \sum_{i=1}^{N} y_i}{x_{N-1} - x_N}$ and $\mathcal{C} = \{\mathcal{A}_1 \bigcap \mathcal{B}_1\}$. Note, that (2.36) is a specification of (2.26) with the constants c_{ii}, c_{il}, b_i and a explicitly defined. For this example it can easily be seen that the subjects must be ordered such that $x_{N-1} \neq x_N$, because otherwise the coefficients h_i, h_l, m_1 and m_2 in (2.36) contain divisions through zero (see step 3 in Section 2.3.1).

The ellipsoid described by (2.36) is equal to one of the isodensity ellipsoids from the multivariate normal $N_{N-2}(\boldsymbol{\mu}^-,\ \boldsymbol{\Sigma}^-)$ with $(\boldsymbol{\Sigma}^-)^{-1} = \boldsymbol{\Omega}^-$, with elements $\omega_{ii} = 1 + h_i^2 - h_i$, for $i \in \mathcal{C}$ and $\omega_{il} = \frac{1}{2}(1 + 2h_i h_l - h_i - h_l)$, for $i \neq l;\ i, l \in \mathcal{C}$. The values for μ_l $(l \in \mathcal{C})$ can be derived from the set of equations

$$\sum_{i'=1}^{N} y_{i'} - m_1 + m_2 + (2m_1 - 2m_2 - \sum_{i'=1}^{N} y_{i'}) h_i = \sum_{l \in \mathcal{C}} 2\omega_{il} \mu_l, \qquad \text{for } i \in \mathcal{C}. \tag{2.37}$$

Similarly to (2.28), the number of equations in (2.37) is equal to the number of parameters to solve, i.e. μ_l for $l \in \mathcal{C}$.

Again, sampling from this multivariate normal, deriving $\boldsymbol{z}^-$ on the required ellipsoid, and computation of z_N and z_{N-1} using (2.34) and (2.35), renders a vector $\boldsymbol{z}$. Repetition of the procedure renders the null distribution of the data.

The data used to illustrate testing of (2.29) against (2.30) concern 32 patients undergoing an elective herniorrhaphy. Three groups are distinguished, based on the pre-operative physical status of the patient (discounting that associated with the operation). Group 1 consists of patients of perfect health, group 2 with a rather good health score and the third group contains patients with a somewhat lower health score (Mosteller and Tukey, 1977, pp. 567-568). The alternative hypothesis (2.30) reflects the idea that the lower the general health before the operation the longer the stay in hospital after the operation will be. Age is included as a covariate to control for a possible effect of age on length of stay in the hospital.

The statistic used to test (2.29) against (2.30) is based on a general test statistic introduced by Doveh, Shapiro and Feigin (2002), that can in principle be used in many situations:

$$T(\boldsymbol{y}) = S(\hat{\boldsymbol{\theta}}) - S(\hat{\boldsymbol{\theta}}^*), \tag{2.38}$$

where $\hat{\boldsymbol{\theta}}$, with $\boldsymbol{\theta} = (\beta_1, \alpha_1, \sigma^2)$, denotes the parameter estimates under the null hypothesis, $\hat{\boldsymbol{\theta}}^*$, with $\boldsymbol{\theta}^* = (\beta_1^*, \beta_2^*, \beta_3^*, \alpha_1^*, \sigma^2*)$, denotes the parameter estimates under the alternative hypothesis, $S(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$ and $S(\hat{\boldsymbol{\theta}}^*) = \sum_{i=1}^{N} (y_i - \hat{y}_i^*)^2$.

Doveh et al. (2002) use a quadratic programming approach to obtain the estimates for the constrained parameters $\hat{\boldsymbol{\theta}}^*$, and evaluate $T(\boldsymbol{y})$ deriving its asymptotic distribution. Our approach renders the finite sample distribution for this test.

We used a Bayesian estimation procedure for $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}^*$, which obtains expected a posteriori (EAP) estimates. For H_0 these are based on a posterior distribution that is proportional to the likelihood of the null model. For H_1 these are based on a posterior distribution that is proportional to the likelihood of the alternative with the restriction that only $\boldsymbol{\theta}^*$'s in accordance with (2.30) have a nonzero posterior probability. For both models EAP's are easily computed using a sample from the posterior (see e.g. Gelfand, Smith and Lee, 1992). The resulting $p = P(T(\boldsymbol{z}) \geq T(\boldsymbol{y})|H_0) = .176$, i.e. after controlling for age it can not be concluded that there is an increase in the mean length of hospital stay with decreasing pre-operative physical status of patients.

Note again that we do not advocate this specific (or any particular) test statistic. Taking the results of Section 2.2.3 into consideration, also for this illustration a

robust test might be a good alternative. To accomplish this, the posterior means (EAP) in (2.38) can for example be replaced by posterior medians. Traditionally, defining a new test statistic implies the (often burdensome) task of deriving its null distribution. However, our approach renders the similar null distribution of the data, and therefore the similar distribution of any test statistic.

2.4 The General Multivariate Null Hypothesis

2.4.1 Replication of Data under the General Multivariate Null Hypothesis

To obtain the null distribution of the data for the general multivariate null hypothesis (2.7), data $\boldsymbol{Z}' = [\boldsymbol{z}_1 \ldots, \boldsymbol{z}_Q]$ is replicated from

$$g(\boldsymbol{Z}|S_1(\boldsymbol{Z}) = S_1(\boldsymbol{Y}), \ldots, S_R(\boldsymbol{Z}) = S_R(\boldsymbol{Y})) = c, \tag{2.39}$$

where $\boldsymbol{Y}$ denotes the observed data, and R the total number of sufficient statistics. The sufficient statistics for (2.39) are:

1. For $q = 1, \ldots, Q$ and $k = 1, \ldots, K$ (i.e. Q times K sufficient statistics)

$$\sum_{i=1}^{N} z_{qi} d_{ki} = \sum_{i=1}^{N} y_{qi} d_{ki}. \tag{2.40}$$

2. For $q = 1, \ldots, Q$ and $p = 1, \ldots, P$ (i.e. Q times P sufficient statistics)

$$\sum_{i=1}^{N} z_{qi} x_{pi} = \sum_{i=1}^{N} y_{qi} x_{pi}. \tag{2.41}$$

3. For $q = 1, \ldots, Q$ (i.e. Q sufficient statistics)

$$\sum_{i=1}^{N} z_{qi}^2 = \sum_{i=1}^{N} y_{qi}^2. \tag{2.42}$$

4. For $q = 1, \ldots, Q-1$ and $q' = q+1, \ldots, Q$ (i.e. $\frac{1}{2}(Q^2 - Q)$ sufficient statistics)

$$\sum_{i=1}^{N} z_{qi} z_{q'i} = \sum_{i=1}^{N} y_{qi} y_{q'i}. \tag{2.43}$$

Note that, $R = QK + QP + Q + \frac{1}{2}(Q^2 - Q)$, which is equal to Q times the number of sufficient statistics for the general univariate null hypothesis ($K + P + 1$) plus the $\frac{1}{2}(Q^2 - Q)$ sufficient statistics in (2.43), that concern the relations between the Q dependent variables. Both the equations in (2.43) and the fact that (2.42) consists of Q quadratic equations makes the sampling of replicated data $\boldsymbol{Z}$ complicated. However, an iterative procedure, the Gibbs sampler, enables sampling from (complex) multivariate distributions like (2.39) (Tierney, 1994; Smith and Roberts, 1993). In our application the Gibbs sampler consists of Q steps. In each iteration of the Gibbs sampler one of $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_Q$ is sampled conditional on the current values of the other vectors: i.e. for $q = 1, \ldots, Q$, $\boldsymbol{z}_q$ is sampled from:

$$g(\boldsymbol{z}_q | \boldsymbol{z}_1, \ldots, \boldsymbol{z}_{q-1}, \boldsymbol{z}_{q+1}, \ldots, \boldsymbol{z}_Q, \boldsymbol{S}^{(q)}) = c, \tag{2.44}$$

where $\boldsymbol{S}^{(q)}$ refers to those sufficient statistics that contain $\boldsymbol{z}_q$. Note, that in (2.44) only one of the quadratic equations in (2.42) is active, and, that the $K+P+(Q-1)$ equations in (2.40), (2.41), and, (2.43) that contain $\boldsymbol{z}_q$ are all linear functions of $\boldsymbol{z}_q$. This implies that each step of the Gibbs sampler reduces to a straightforward application of the general univariate sampling procedure introduced in Section 2.3.1, with K equal to the number of groups under the null hypothesis, and P consisting of two parts: the continuous predictors under the null hypothesis plus the set of dependent variables that is conditioned upon (i.e. $Q - 1$).

Burn-in and convergence of the Gibbs sampler
Sampling a multivariate distribution using the Gibbs sampler requires careful monitoring of burn-in and convergence (Cowles and Carlin, 1996). A burn-in period is important because usually the first iterations depend largely on arbitrary starting values. However, in this application the target distribution (2.39) is uniform. The implication is that there is no burn-in period. We used $\boldsymbol{Z} = \boldsymbol{Y}$ as the starting values for the Gibs sampler.

Convergence of the Gibbs sampler implies that the sample obtained is a proportional representation of (2.39). A practical approach to monitor the convergence is plotting one or several chains of the output produced by the algorithm, here $\boldsymbol{Z}$. Since we are interested in the p-value, which is a function of $\boldsymbol{Z}$, in the applications (Sections 2.4.2 and 2.4.3) convergence is monitored by plotting the p-value against increasing numbers of replications.

2.4.2 Multivariate One-sided Testing

The data ($N = 9$) used for this illustration are part of a larger research project on the efficacy of cognitive behavioral therapy for traumatic grief (Boelen, van den Bout,

Keijser, 2002). Before and after six cognitive therapy sessions the four criterion variables *grief* (Texas Revised Inventory of Grief), *traumatic grief* (Inventory of Traumatic Grief), *anxiety* (subscale of SCL-90) and *depression* (subscale of SCL-90) are measured. For each criterion variable the decrease in symptoms is measured by the difference between the scores before and after the treatment. The research question investigated in this illustration is whether there is a decrease in the grief symptoms.

The null hypothesis is (2.7) with $Q = 4$, $K = 0$, $P = 0$, i.e

$$H_0 : \boldsymbol{\beta} = 0, \qquad \boldsymbol{\varepsilon}_i \overset{iid}{\sim} N(0, \boldsymbol{\Sigma}). \tag{2.45}$$

It is tested against the alternative $\boldsymbol{y}_i = \boldsymbol{\beta}_1^* + \boldsymbol{\varepsilon}_i$ with the restriction that the mean vector must be positive, i.e.

$$H_1 : \boldsymbol{\beta}_1^* \geq 0, \qquad \boldsymbol{\varepsilon}_i \overset{iid}{\sim} N(0, \boldsymbol{\Sigma}^*) \tag{2.46}$$

Although we found some references in the literature with respect to multivariate one-sided testing, for most of the test statistics discussed, no finite sample distribution is available (Silvapulle, 1995b; Perlman and Wu, 2002). An exception is Follmann's test (Follmamm, 1996), although according to Chongcharoen et al. (2002) this test is not very powerful. There are other references that do present tests with their finite sample distribution, but for these tests it is assumed that the population covariance matrix is known, or known up to a multiplicative constant (Chongcharoen et al., 2002; Silvapulle, 1995a). Our method provides the finite sample distribution for any statistic that can be applied to test $\boldsymbol{\beta} = 0$ against $\boldsymbol{\beta} \geq 0$, when the covariance matrix is unknown.

The nuisance parameters under the null hypothesis are the 10 upper diagonal elements of the 4×4 covariance matrix $\boldsymbol{\Sigma}$. The sufficient statistics for these nuisance parameters are:

1. For $q = 1, \ldots, 4$
$$\sum_{i=1}^{N} z_{qi}^2 = \sum_{i=1}^{N} y_{qi}^2. \tag{2.47}$$

2. For $q = 1, \ldots, 3$ and $q' = q + 1, \ldots, 4$
$$\sum_{i=1}^{N} z_{qi} z_{q'i} = \sum_{i=1}^{N} y_{qi} y_{q'i}. \tag{2.48}$$

The Gibbs sampler applied to sample from $g(\boldsymbol{Z}|S_1(\boldsymbol{Z}) = S_1(\boldsymbol{Y}), \ldots, S_{10}(\boldsymbol{Z}) = S_{10}(\boldsymbol{Y})) = c$, consists of four steps. In each step one vector $\boldsymbol{z}_q$ is sampled conditional on the current values of the other vectors and $\boldsymbol{S}^{(q)}$, i.e. the sufficient statistics that contain $\boldsymbol{z}_q$ (see (2.44)). Note that in each step, $\boldsymbol{S}^{(q)}$ consists of three linear and one quadratic function of $\boldsymbol{z}_q$, and this implies that in each step the general univariate approach as introduced in Section 2.3.1 can be applied, i.e. (2.20) with $K = 0$ and $P = 3$ (the number of dependent variables that is conditioned upon in each step, i.e. $Q - 1$).

The test statistic used in this illustration is

$$T(\boldsymbol{Y}) = \sum_{q=1}^{Q} \frac{\bar{y}_q}{\mathrm{sd}(y_q)} \tag{2.49}$$

where $\bar{y}_q$ denotes the sample mean and $\mathrm{sd}(y_q)$ the sample standard deviation of the q-th dependent variable. The larger the value of $T(\boldsymbol{Y})$ the more the sample means are located in the positive orthant, that is, the larger the evidence in favor of the alternative hypothesis. For each replicated data matrix $\boldsymbol{Z}$, $T(\boldsymbol{Z}) = \sum_{q=1}^{Q} \frac{\bar{z}_q}{\mathrm{sd}(z_q)}$ is computed. The p-value is obtained by $P(T(\boldsymbol{Z}) \geq T(\boldsymbol{Y})|S_1(\boldsymbol{Z}) = S_1(\boldsymbol{Y}), \ldots, S_{10}(\boldsymbol{Z}) = S_{10}(\boldsymbol{Y}))$.

Note that there may be superior tests for the hypotheses at hand. However, the aim of this illustration is not to find the best test statistic, but to demonstrate that our method obtains the similar or finite sample distribution of any test statistic. The simplicity of (2.49) makes it attractive for this illustrative purpose, but our procedure can be applied to any statistic of interest.

Applying our approach to obtain the similar null distribution and the test statistic in (2.49) to the traumatic grief data, results in a significance $p = .0017$. It can be concluded that after six cognitive therapy sessions, the amount of grief has decreased.

In Figure 2.3 the convergence of the Gibbs sampler is monitored, and the precision of the p-value as a function of the number of replications can be seen. It shows a stable third decimal after about 500,000 replications. With this example we illustrated how the similar null distribution of the data is used to obtain a similar p-value for an arbitrary test statistic.

2.4.3 Multivariate Ordered Means

The data used for this illustration come from a large study concerning the effect of Sesame Street on six cognitive ability measures for children in the age range

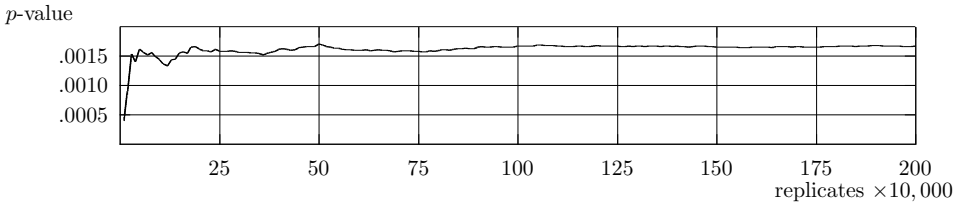

Figure 2.3: Testing H_0 Against H_1.

between three to five (Stevens, 1996, pp. 578-585). The cognitive abilities (e.g. knowledge of numbers and letters) were measured before Sesame Street was broadcasted (pretests), and after the first year of the program (post-tests). Gain scores are obtained by taking the difference between the post-test and pretest scores. Furthermore, two groups are distinguished, based on whether children watched the show frequently (group 1), or, rarely or never (group 2). The studied sample contains children from different social and environmental backgrounds, but in this illustration only one group is included: the disadvantaged rural children ($N = 43$). The research question of interest is whether watching the show regularly compared to watching rarely or never leads to a higher gain in the six measures of cognitive ability.

This leads to testing the null hypothesis

$$\boldsymbol{y}_i = \boldsymbol{\beta}_1 d_{1i} + \boldsymbol{\varepsilon}_i, \qquad \text{with } \boldsymbol{\varepsilon}_i \overset{iid}{\sim} N(0, \boldsymbol{\Sigma}), \tag{2.50}$$

i.e. (2.7) with $K = 1$, $d_{1i} = 1$ for $i = 1, \ldots, N$, $P = 0$, and $Q = 6$, against the alternative

$$\boldsymbol{y}_i = \boldsymbol{\beta}_1^* d_{1i} + \boldsymbol{\beta}_2^* d_{2i} + \boldsymbol{\varepsilon}_i, \qquad \text{with } \boldsymbol{\varepsilon}_i \overset{iid}{\sim} N(0, \boldsymbol{\Sigma}^*), \tag{2.51}$$

with $\boldsymbol{\beta}_1^* \geq \boldsymbol{\beta}_2^*$, where the latter is interpreted coordinate-wise.

We found only one reference with respect to testing this kind of hypotheses (Sasabuchi et al., 2003). However, the finite sample distribution of the test statistic they present, is not known. With our approach the finite sample distribution of any test statistic that can be used to test (2.50) against (2.51) is available.

Under the null hypothesis there are 27 nuisance parameters, i.e. six elements of $\boldsymbol{\beta}_1$ and 21 elements of $\boldsymbol{\Sigma}$. The sufficient statistics for the nuisance parameters, are:

1. For $q = 1, \ldots, 6$

$$\sum_{i=1}^{N} z_{qi} = \sum_{i=1}^{N} y_{qi}. \tag{2.52}$$

2. For $q = 1, \ldots, 6$

$$\sum_{i=1}^{N} z_{qi}^2 = \sum_{i=1}^{N} y_{qi}^2. \tag{2.53}$$

3. For $q = 1, \ldots, 5$ and $q' = q + 1, \ldots, 6$

$$\sum_{i=1}^{N} z_{qi} z_{q'i} = \sum_{i=1}^{N} y_{qi} y_{q'i}. \tag{2.54}$$

The Gibbs sampler that is applied to sample from $g(\boldsymbol{Z}|S_1(\boldsymbol{Z}) = S_1(\boldsymbol{Y}), \ldots, S_{27}(\boldsymbol{Z}) = S_{27}(\boldsymbol{Y})) = c$, consists of six steps. Again, each step of the Gibbs sampler involves the application of the general univariate approach introduced in Section 2.3.1, i.e. (2.20) with $K = 1$ and $P = 5$.

The test statistic used in this illustration is

$$T(\boldsymbol{Y}) = \sum_{q=1}^{Q} \frac{\bar{y}_{q1} - \bar{y}_{q2}}{\mathrm{sd}(y_{q1} - y_{q2})}, \tag{2.55}$$

where $\bar{y}_{q1}$ (respectively $\bar{y}_{q2}$) denotes the sample mean of the q-th dependent variable of group 1 (respectively group 2), and $\mathrm{sd}(y_{q1} - y_{q2})$ denotes the sample standard deviation of the difference scores of the q-th dependent variable. The larger the value of this test statistic the more the pairwise differences between the sample means are located in the positive orthant, that is, the larger the evidence in favor of the alternative hypothesis. For each replicated data matrix $\boldsymbol{Z}$, $T(\boldsymbol{Z}) = \sum_{q=1}^{Q} \frac{\bar{z}_{q1} - \bar{z}_{q2}}{\mathrm{sd}(z_{q1} - z_{q2})}$ is computed. The p-value is obtained by $P(T(\boldsymbol{Z}) \geq T(\boldsymbol{Y})|S_1(\boldsymbol{Z}) = S_1(\boldsymbol{Y}), \ldots, S_{27}(\boldsymbol{Z}) = S_{27}(\boldsymbol{Y}))$.

For comparison, also the regular (two-sided) multivariate analysis is performed using Wilk's Lambda as the test statistic. With a sample of 2,000,000 iterations, for the two-sided test our procedure renders $p = .053$, which is equal to the p-value obtained, evaluating Wilk's lambda using the $F(6, 36)$-distribution (Tatsuoka, 1988, p. 94). Testing against (2.51), i.e. the inequality constrained alternative, using (2.55) renders $p = .0064$. It can be concluded that the disadvantaged rural children that watched Sesame Street regularly have learned more than the disadvantaged rural children that did not watch Sesame Street.

In Figure 2.4 the precision of the p-value as a function of Gibbs iterations is monitored. It shows a stable third decimal after about 500,000 replications.

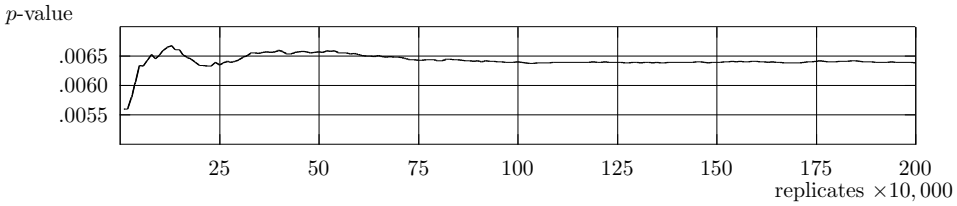

Figure 2.4: Testing (2.50) Against (2.51).

Also in this last example, a similar test is obtained for a multivariate, inequality constrained hypothesis. We did not find any references in the literature to such a test. Note that we used a relatively ad hoc test statistic. However, note that the focus of this paper is not on the test statistic but on obtaining the similar null distribution of the data, which enables similar tests.

2.5 Discussion

This paper showed how to obtain similar or finite sample null distributions of data, if the null hypothesis is a member of the univariate or multivariate normal model. Using relatively simple test statistics, we demonstrated that with this distribution similar p-values can be obtained for any test statistic of interest.

The examples given in this paper concerned both univariate and multivariate null hypotheses with order restricted alternatives. For some of these examples test statistics with known distributions can be found in the literature. For other examples, only asymptotic distributions are known. Our approach renders the finite sample distribution for any test statistic. Furthermore, where test procedures in order constrained inference are available, they are almost invariably based on the use of level probabilities. Our approach offers an alternative which does not use level probabilities.

Finally, order constrained inference is not the only domain in which our method can be useful. As indicated in the introduction it can also be used for testing the equality of variances. Furthermore it can be used to obtain similar distributions of order statistics (Mood, Graybill and Boes, 1974, pp. 251-265).

Chapter 3

Inequality Constrained Analysis of Variance: A Bayesian Approach

Abstract[†]

Theories can be translated into statistical models via restrictions on the model parameters. In this paper a Bayesian approach to evaluate ANOVA or ANCOVA models with inequality constraints on the (adjusted) means will be presented. This evaluation contains two issues: estimation of the parameters given the restrictions using the Gibbs sampler and model selection using Bayes factors in the case of competing theories. The paper will be concluded with two examples: an one-way analysis of covariance; and, an analysis of a three-way table of ordered means.

[†]The software that is referred to in this chapter can be downloaded from the following website: www.fss.uu.nl/ms/ik

3.1 Introduction

This paper deals with analysis of (co)variance models (ANOVA, ANCOVA), with inequality constraints on the (adjusted) means. In the context of an ANOVA an example of such a restricted model is found in an experiment about pain resistance. In the experiment the criterion variable is the length of time that a person resists pain (voluntarily holding a hand in ice-cold water). To each of four groups of respondents different types of photos are displayed during the experiment: photos with negative, pain related pictures (group 1), negative, but not pain related pictures (2), neutral pictures (3) or positive pictures (4). There are different theories in psychological research on this topic. To analyze data using these theories they need to be translated into statistical models and inequality constraints on the model parameters can play an important role in specifying the model. In the pain resistance experiment, it is expected that for the two types of negative pictures, those that are pain related will lead to a shorter resistance time. At the same time it is expected that positive pictures will increase resistance time as compared to the neutral pictures. Denoting the mean resistance time with β_j, where $j = 1, \ldots, 4$ denotes the condition, the statistical model representing this theory is $M_1 : \beta_1 < \beta_2;\ \beta_3 < \beta_4$. Note, that M_1 contains no information about the relation between the first two and the last two conditions. In psychological research there exist two competing theories that are specifications of M_1. The first theory states that the four groups of pictures are ordered from very negative (1) to positive (4) and that the pain resistance will be ordered in the same way (the more positive the photos the longer the resistance time). Translation of this expected outcome into a statistical model leads to $M_2 : \beta_1 < \beta_2 < \beta_3 < \beta_4$. The second theory states: negative, pain related pictures will decrease the pain resistance but negative, not pain related pictures are distracters and will therefore lead to longer resistance times than neutral pictures (but not as long as positive pictures that are even better distracters). This leads to a competing model $M_3 : \beta_1 < \beta_3 < \beta_2 < \beta_4$. Note that M_1, M_2 and M_3 are competing theories, that M_2 and M_3 are specifications of M_1 and that M_2 and M_3 are not nested. To determine which theory best fits the data, a confrontation of each model with the data is needed as well as a criterion for model selection.

As far as known to the authors, there is only one non-Bayesian model selection criterion that can be used in the context of inequality constrained modelling (Anraku, 1999). He presents a criterion that can be used to select the best of a number of models that differ in the simple order restriction imposed on its means.

In this paper a Bayesian criterion will be introduced. Kass and Raftery (1995) describe an information criterion called Bayes factors. As explained and illustrated by Hoijtink (2001), Bayes factors can be used to select the best of a number of models

that differ with respect to the inequality constraints imposed on their parameters. Bayes factors is rather general in the sense that its application is not limited to a specific kind of model or a specific kind of restrictions. There are several methods to estimate the Bayes factor (see for example Newton & Raftery, 1994; Carlin & Chib, 1995; Chib, 1995). However, in this paper a new method, that is particularly suited for the type of models that is considered in this paper, is introduced.

Note that, since the various competing models are not necessarily nested models, and since it is usually not clear which model is the null model, the testing of hypotheses (null versus alternative model) will not be considered in this paper. See also Marden (2000), who discusses the advantages of using Bayes factors as compared to classical hypothesis testing. According to Marden problems can arise in the interpretation of p-values in model selection procedures, especially if non-nested models are compared. Additionally, in a paper focussing on p-values, Bayarri and Berger (2000) recommend using Bayes factors instead of any p-value if clearly formulated competing theories are available.

Estimation of constrained parameters is also of interest. Incorporation of the order information not only leads to estimators that are in accordance with the constraints, but also to a higher efficiency. This paper will use a Bayesian computational approach to obtain estimates of the parameters of constrained models. Estimation of the inequality constrained parameters has also been considered in a non-Bayesian context. See Barlow, Bartholomew, Bremner and Brunk (1972) and its successor Robertson, Wright and Dykstra (1988) for comprehensive overviews. With respect to constrained normal linear models, only relatively simple models have been considered. A few examples are: the simple order $\mu_1 < ... < \mu_5$ (see, McDermott and Mudholkar (1993) for a more recent discussion and extensions of this model); a two-way ANOVA with ordered marginals; and, linear regression with constraints on the regression parameters (see, Mukerjee and Tu (1995) and Robert and Hwang (1996) for a more recent discussion of this model). In these papers methods to obtain estimates of constrained parameters are presented, but standard errors and/or confidence intervals are not included. Furthermore, most estimation methods are tailored to a specific model and generalization to other or more complex models is usually not straightforward. A possible exception is the approach taken by Robert and Hwang (1996) which looks promising although as to yet standard errors and confidence intervals are not provided.

Gelfand, Smith and Lee (1992) and Smith and Roberts (1993) note (and illustrate using simple models) that it is easy to obtain estimates, posterior standard deviations and credibility intervals for the parameters of constrained models, using a Bayesian approach. The latter was confirmed by Hoijtink (2001) who used the Gibbs sampler to perform constrained latent class analysis.

Besides the example presented in this section, many other statistical models and types of constraints are possible, e.g. constraints on adjusted means in models with one or more covariates, or constraints on simple effects in multi-way designs. Examples of theories that are translated into statistical models of this kind (constrained adjusted means and constrained simple effects) are provided in the sections 3.2, 3.5, and 3.6. The general model and its possibilities will be described more extensively in Section 3.2. In this section also comparisons with traditional approaches (null hypothesis testing) will be made. Section 3.3.1 will introduce the concept of prior and posterior distributions which are central in Bayesian statistics. In Section 3.3.2 it will be explained how the Gibbs sampler can be used to sample from the posterior distribution. Based on the output of the Gibbs sampler it is easy to compute posterior means, posterior standard deviations and credibility intervals (Section 3.3.3). In Section 3.3.4 the incorporation of inequality constraints in the Bayesian approach is discussed. Model selection based on Bayes factors is addressed in Section 3.4. Finally, in Sections 3.5 and 3.6 the proposed approaches will be illustrated by two specific cases of the general model. In Section 3.5 an analysis of covariance with constraints on the adjusted means will be executed. In Section 3.6 an analysis of a three-way table of ordered means will be discussed.

3.2 The Model and its Possibilities

3.2.1 The General Model

In the case of a design with one dependent variable (y), one or more categorical variables that together define $j = 1, ..., J$ groups and $k = 1, ..., K$ continuous predictors, the regression equation for person $i = 1, ..., N$ can be written as follows:

$$y_i = \sum_{j=1}^{J} \beta_j d_{ji} + \sum_{k=1}^{K} \alpha_k x_{ki} + \varepsilon_i, \tag{3.1}$$

with the usual assumption that $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. The score of the i'th person on the j'th dummy variable is denoted by d_{ji}. If this person is a member of group j the value $d_{ji} = 1$, if not $d_{ji} = 0$. Therefore each β_j represents the (adjusted) mean of group j. The score of the i'th person on the k'th continuous predictor is denoted by x_{ki} and the regression coefficient of the k'th predictor is denoted by α_k. This paper will limit itself to the situation where the inequality constraints are imposed on the β's, i.e. on the (adjusted) means of independent groups.

3.2.2 Simple Analysis of Variance with Ordered Means

A specification of the general model is (3.1) with one categorical independent variable containing seven levels ($J = 7$) and no continuous predictors ($K = 0$). This is the model of a classical experiment on conformity to group pressure (Asch, 1955). In this experiment, a participant (the subject) had to indicate which one of three lines was equal in length to a fourth line, *after* hearing other participants (confederates of the experimenter) offering an incorrect answer. Eighteen tests of this kind were given to each subject and the number of incorrect answers (i.e. the number of times the subject conforms to the dissenting majority) serves as the dependent variable. Each subject was randomly assigned to one of seven conditions, formed by the size of the dissenting majority (ranging from zero (control condition) to sixteen).

The central hypothesis is: will conformity increase with increasing size of the dissenting (incorrect) majority? Denoting the mean number of incorrect answers per condition with β_j $(j = 1, \ldots, 7)$, translation of this research question into a statistical model leads to:

$$\beta_1 < \beta_2 < \beta_3 < \beta_4 < \beta_5 < \beta_6 < \beta_7. \tag{3.2}$$

Using the Bayesian approach described in this paper this model can be compared with for example the model that reflects the 'no effect' theory ($\beta_1 = \ldots = \beta_7$), the model that allows for group differences but does not impose any ordering $(\beta_1, \ldots, \beta_7)$, or any other model that reflects a theory or expected outcome. For the resulting set of models, the probabilities of each of these models after observing the data (the posterior model probabilities) reflect the support for each of the models after confronting them with the data. Comparing models based on posterior model probabilities will be addressed in Section 3.4.

In a standard classical approach, firstly an analysis of variance (ANOVA) will be applied. In this omnibus test the statistical hypotheses are $H_0: \beta_1 = \ldots = \beta_7$, and $H_1:$ not H_0. Note that (3.2), the research question of interest, is not yet included. Only if the outcome of the ANOVA is significant, it will be followed by planned pairwise comparisons to examine the actual research question. In this example, this implies six directional tests, with the hypotheses:

$$\begin{aligned}
&H_0 : \beta_1 = \beta_2, \quad \text{and} \quad H_1 : \beta_1 < \beta_2, \\
&H_0 : \beta_2 = \beta_3, \quad \text{and} \quad H_1 : \beta_2 < \beta_3, \\
&\vdots \\
&H_0 : \beta_6 = \beta_7, \quad \text{and} \quad H_1 : \beta_6 < \beta_7.
\end{aligned}$$

A possible non-significant outcome of the ANOVA indicates no effect in the population, but may well be due to a lack of power. This is especially problematic, because the actual research question is more specific than the alternative hypothesis of the ANOVA. Applying the non-directional ANOVA for inequality constrained hypotheses leads to a loss of power. However, to avoid capitalization on chance, conventionally one stops after finding a non-significant result for the omnibus test. This means that the analysis stops even before the actual research question is included in the analysis.

Now, consider the scenario where a significant outcome of the ANOVA is obtained. The theory is reflected by the six pairwise tests. What should a researcher conclude if five p-values smaller than .05 were found, but for one test $p = .40$? Or five times $p < .05$, and once $p = .06$? Or six times $p = .051$? Either a researcher should be very strict and might reject the theory in cases where the data is rather convincingly in accordance with this theory, or the researcher could be less strict in evaluating the results leading to subjective and questionable conclusions.

The Bayesian approach addresses the research question of interest directly, by translating it into a statistical model with inequality constraints. As was pointed out in the introduction, there are also non-Bayesian solutions for inequality constrained hypotheses. An advantage of the method described in this paper, as compared to the frequentist approach (see Barlow et al., 1972; and Robertson et al., 1988) is, that it is simple, straightforward and general in the sense that basically the same procedure can be applied for any model.

3.2.3 Two-way Analysis of Variance with Ordered Simple Effects

Consider the experiment where four and six years old children are observed in one of three conditions: alone, with another child or with a parent. The two factors age and condition form a 2×3 design, represented by the general model with $J = 6$, $K = 0$. The criterion variable is the time spent playing with unfamiliar toys. The expectation is that for both ages the playing time is increasing over the three conditions (from 'alone' to 'with a parent'). Denoting the mean playing time with β_j, where j denotes the subgroups formed by the two categorical variables age and condition, the expectation is translated into the statistical model:

$$\beta_{A4} < \beta_{C4} < \beta_{P4}; \quad \beta_{A6} < \beta_{C6} < \beta_{P6}, \tag{3.3}$$

where the letter in the subscript denotes the condition (A=alone, C= with child, P=with parent) and the number in the subscript denotes the child's age. A competing theory states that the ordered effect of condition on time playing with unfamiliar

Table 3.1: Hypothetical Outcomes of Simple Effects Tests

$H_0 : \beta_{A4} = \beta_{C4}$ $H_1 : \beta_{A4} < \beta_{C4}$	$H_0 : \beta_{C4} = \beta_{P4}$ $H_1 : \beta_{C4} < \beta_{P4}$	$H_0 : \beta_{A6} = \beta_{C6}$ $H_1 : \beta_{A6} < \beta_{C6}$	$H_0 : \beta_{C6} = \beta_{P6}$ $H_1 : \beta_{C6} < \beta_{P6}$
$p = .01$	$p = .01$	$p = .01$	$p = .12$

toys will be seen for children of four years but not for children that are six years old. This is translated into the constrained statistical model:

$$\beta_{A4} < \beta_{C4} < \beta_{P4}; \quad \beta_{A6} = \beta_{C6} = \beta_{P6}. \tag{3.4}$$

The two competing theories create a set of two models and posterior model probabilities provide the information about the fit of the data to each of these models.

Applying a traditional frequentist approach, a two-way analysis of variance will be performed and interaction and main effects will be evaluated. Again, the hypotheses of these tests are not directional meaning a loss of power, and therefore the analysis might stop before the models of interest, (3.3) and (3.4), are addressed. Now, let's consider only the expectation reflected in (3.3) and the situation that a significant main effect of condition is found. If the interaction between condition and age is not significant two directional tests on the marginal means of the three conditions are needed to evaluate (3.3). However, if the interaction is significant, (3.3) can still be correct (the increasing effect of condition can be larger for one of the age groups), but to confirm the theory four directional simple effects tests are required to be significant. In interpreting the results of several individual tests to evaluate one theory, the same problems as described in the previous example occur.

Finally, consider the question: Which of the two competing theories (3.3) and (3.4) is supported best by the observed data? This requires a confirmative analysis to determine whether one of the models fits better than the other. However, comparing different inequality constrained models using a classical (null hypotheses testing) approach, can be troublesome. Consider for example the hypothetical results of the simple effects tests in Table 3.2.3. With these outcomes it is hard, if not impossible, to make a decision on which of the two theories fits best.

3.2.4 Other Designs

The two examples in Sections 3.2.2 and 3.2.3 are just a selection of the possibilities. The techniques are applicable to a wide variety of models with a wide variety in

sets of inequality constraints. The first example could, for instance, be extended with the inclusion of age as a covariate. It could be expected that the age of a subject is related to the tendency to conform to the majority. By inclusion of age in the model, possible age differences between the conditions are controlled for and the residual variance is decreased. Note however, that the central research question of interest remains the same: will conformity increase with increasing size of the dissenting (incorrect) majority? The statistical model that is evaluated after correcting for possible age differences between the groups, is: does conformity increase with increasing size of the incorrect majority *for subjects of the same age*?

The example in the introduction showed that not only fully ordered models can be included but also partly ordered models (M_1 of the pain resistance experiment). An example of a model that is partly inequality constrained and partly equality constrained, is provided in the second example of this section. Extensions of the second example are larger factorial designs and again the inclusion of one or more covariates. All kinds of expected outcomes (e.g. specific interaction patterns) can be translated into statistical models such that they can be evaluated with the Bayesian approach described in this paper.

Note, that although several models are suggested throughout the examples, the model selection approach of this paper is not exploratory, nor designed to select the best of all possible models. The goal is to select the best fitting model of a limited set of well defined models that are based on a theoretical framework.

3.3 Estimation Using the Bayesian Approach

In the sections 3.3.1, 3.3.2 and 3.3.3 Bayesian estimation using the Gibbs sampler will be explained for the simple model: $y_i = \beta_1 d_{1i} + \beta_2 d_{2i} + \varepsilon_i$, for $i = 1, \ldots, N$, with $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ and with no inequality constraints on the parameters. Note, that each subject is either a member of the first group ($d_{1i} = 1, d_{2i} = 0$) or a member of the second group ($d_{1i} = 0, d_{2i} = 1$), so β_1 and β_2 denote independent means. In Section 3.3.4 it will be discussed how inequality constraints can be included in the model.

The sampling scheme for the general model (3.1), with inequality constraints on the β's is provided in Appendix I.

3.3.1 Prior and Posterior Distributions

Bayesian parameter estimation is based on the *posterior* distribution of the model parameters, i.e. the distribution of the parameters *after* observing the data. Stated

otherwise, the posterior distribution reflects what is known about the model parameters after observing the data. The posterior is formed by a combination of the density of the data and the *prior* distribution of the model parameters, i.e. the distribution of the parameters *before* observing the data. With the prior distribution the knowledge (or uncertainty) about the parameters before observing data can be reflected.

Prior distribution

Specification of prior distributions is an important part of the Bayesian analysis, because the results of both estimation and model selection depend on the choice of the prior distribution, as will be seen in (3.5) and (3.12), respectively.

Usually, when several models are compared, for each model a prior distribution for the model specific parameters must be defined (Gelman, Carlin, Stern and Rubin, 1995, pp 175-177). When the fit of each of these models after observing the data is compared (how this is done will be explained in Section 3.4) the results depend on the choice of the prior distributions (Lee, 1997, p.120). However, the models discussed in this paper have a special feature. All (in)equality constrained models are nested in an unconstrained, also called encompassing model. An important consequence of this nesting is that only for the unconstrained, encompassing model a prior distribution for the model parameters needs to be specified. The prior distributions for the nested models like (3.2), (3.3) and (3.4) follow directly from this so-called encompassing prior (see Section 3.4).

In this paper a relatively objective prior will be specified for the encompassing model, i.e. no subjective a priori information will be used and the results (both of estimation and model selection) are mainly determined by the data. To accomplish this three criteria are used:

1. The prior distribution must be diffuse (vague) over the whole range of values that are reasonable for the parameters. Specification of the encompassing prior can be based on the response-format, e.g. if data is measured on a scale from zero to ten, the prior distribution for the mean of this data is chosen such that it is vague for the whole range zero through ten. Specification can also be based on the observed data, e.g. based on the mean and variance of the data, the prior is chosen such that it is vague for the range of the parameters where the likelihood is substantial. The latter approach is taken in this paper and will be further discussed in the sequel.

2. The prior distribution should not include too many values that are completely out of range for the data at hand. To continue the previous illustration with

data in the range from zero to ten, the prior should not have a substantial density for values that are much smaller than zero or much larger than ten.

3. The prior distribution for the encompassing model should not 'benefit' one of the models in the set of models that are compared. This is achieved by using the same prior for each (adjusted) mean, that is for each parameter that may be subjected to constraints.

These criteria will be illustrated by applying them to the simple model with parameters β_1, β_2 and σ^2. The models of interest are: $M_1 : \beta_1, \beta_2, \sigma^2$, $M_2 : \beta_1 > \beta_2, \sigma^2$, and, $M_3 : \beta_1 = \beta_2, \sigma^2$. Note, that only the prior for the encompassing model M_1 needs to be specified. A priori the parameters are considered to be independent, i.e. $\pi(\beta_1, \beta_2, \sigma^2) = \pi(\beta_1)\pi(\beta_2)\pi(\sigma^2)$. Furthermore $\pi(\beta_1) = \pi(\beta_2)$, which leads to a prior reflecting no preference for one of the three models (third criterion). For each of the parameters diffuse, conjugate prior distributions are used. For the β's the conjugate prior is the normal distribution with mean μ_0 and variance τ_0^2. For the error variance (σ^2) the conjugate prior is a scaled inverse χ^2-distribution with the parameters ν_0 and κ_0^2 (Lee, 1997, pp 50-53). Specification of these prior distributions, i.e. choosing the values for μ_0, τ_0^2, ν_0 and κ_0^2, must ensure that the first two criteria (diffuse, but not totally out of range) are met. To accomplish this, in this paper (and the accompanying software) the specification of the encompassing prior is data based. Firstly, for β_1 and β_2 the 99.7% (estimated group mean $\pm$ three times the group standard error) credibility intervals are computed. This ensures that per group the range for each β corresponds to the range of values that $\boldsymbol{y}$ can attain, but excludes values that are very unlikely. Subsequently, the two intervals are combined into one. The smallest value of the two lower bounds and the largest value of the two upper bounds provides the range within which the encompassing prior should be vague (to meet the first criterion), and, outside which the encompassing prior should not have a substantial density (to meet the second criterion). The normal prior distribution for both β_1 and β_2 is chosen such that the mean minus one standard deviation equals the lower bound of the range (denoted by LB), and, the mean plus one standard deviation equals the upper bound of the range (UB). Therefore, the parameters of the normal distribution for each β are: $\mu_0 = \frac{UB+LB}{2}$, and, $\tau_0^2 = (\frac{UB-LB}{2})^2$.

The parameter ν_0 of the scaled inverse χ^2-distribution can be thought of as providing the information equivalent to ν_0 observations with average squared deviation κ_0^2, and is always set to the value one. For κ_0^2 the data based estimation of the error variance is used.

For the general model (3.1) the encompassing prior will be specified in the same way. See Appendix I for the details.

Posterior distribution

The posterior distribution is proportional to the product of the density of the data (in this paper normally distributed) and the prior distribution for the model parameters:

$$p(\beta_1, \beta_2, \sigma^2 | \mathbf{y}, \mathbf{d}_1, \mathbf{d}_2) \propto g(\mathbf{y} | \beta_1, \beta_2, \sigma^2, \mathbf{d}_1, \mathbf{d}_2)\ \pi(\beta_1, \beta_2, \sigma^2). \tag{3.5}$$

Obtaining parameter estimates based on multivariate posterior distributions can be complicated. However, a sampling procedure like the Gibbs sampler (see Section 3.3.2) easily provides a sample from (3.5). Using this sample, parameter estimates can easily be computed (see Section 3.3.3). Note that, in the simple model at hand there are other, straightforward (non-Bayesian) methods to obtain parameter estimates. However, when models are specified using inequality constraints, these traditional methods fail. In Section 3.3.4 the inclusion of inequality constraints in the Gibbs sampler will be explained.

The posterior for the general model (3.1) including inequality constraints is given in (3.16) in Appendix I.

3.3.2 The Gibbs Sampler

The Gibbs sampler is an iterative sampling method. The basic idea is that sampling from a multidimensional posterior distribution can be done via repeatedly sampling from the univariate distribution of each parameter, conditional upon the current values of the other parameters.

The first step of the Gibbs sampler is to assign initial values to each of the parameters. Then, in each iteration $s = 1, ..., S$, new values are sampled from the conditional distributions in a fixed sequence (Gelman et al., 1995, pp 326-327). For (3.5) this means that in iteration s, $\beta_1^{(s)}$ is sampled from the conditional distribution of that parameter given the values for the other parameters sampled in iteration $s-1$, denoted by $\beta_2^{(s-1)}$ and $\sigma^{2(s-1)}$. Subsequently $\beta_2^{(s)}$ is sampled conditional on the values of $\beta_1^{(s)}$ and $\sigma^{2(s-1)}$, and in the third step $\sigma^{2(s)}$ is sampled using $\beta_1^{(s)}$ and $\beta_2^{(s)}$. This scheme is repeated for $s+1$, etcetera.

So, to sample (3.5) using the Gibbs sampler each iteration consists of three steps. The conditional posterior distribution of each of the three parameters is a known and easy to sample distribution.

1. β_1 is sampled from:

$$p(\beta_1 | \beta_2, \sigma^2, \mathbf{y}, \mathbf{d}_1, \mathbf{d}_2) \propto$$

$$\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y_i - [\beta_1 d_{1i} + \beta_2 d_{2i}])^2}{\sigma^2}\right) N(\beta_1|\mu_0, \tau_0^2),$$

i.e. the posterior as in (3.5) but now as a function of only β_1. The posterior for β_1 is equal to a normal distribution with expectation and variance

$$\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}\sum_{i=1}^{N} d_{1i} y_i}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}\sum_{i=1}^{N} d_{1i}}, \tag{3.6}$$

$$\tau_1^2 = \left(\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}\sum_{i=1}^{N} d_{1i}\right)^{-1}, \tag{3.7}$$

where $\sum_{i=1}^{N} d_{1i}$ and $\sum_{i=1}^{N} d_{1i} y_i \;/\; \sum_{i=1}^{N} d_{1i}$ are respectively the sample size and the sample mean of the first group.

2. β_2 is sampled from:

$$p(\beta_2|\beta_1, \sigma^2, \mathbf{y}, \mathbf{d}_1, \mathbf{d}_2) \propto$$

$$\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y_i - [\beta_1 d_{1i} + \beta_2 d_{2i}])^2}{\sigma^2}\right) N(\beta_2|\mu_0, \tau_0^2).$$

The posterior for β_2 is equal to a normal distribution with expectation and variance

$$\mu_2 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}\sum_{i=1}^{N} d_{2i} y_i}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}\sum_{i=1}^{N} d_{2i}}, \tag{3.8}$$

$$\tau_2^2 = \left(\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}\sum_{i=1}^{N} d_{2i}\right)^{-1}. \tag{3.9}$$

3. σ^2 is sampled from $p(\sigma^2|\beta_1, \beta_2, \mathbf{y}, \mathbf{d}_1, \mathbf{d}_2)$, a scaled inverse χ^2-distribution with parameters:

$$\nu = N + \nu_0 \tag{3.10}$$

$$\kappa^2 = \frac{\nu_0 \kappa_0^2 + N v}{\nu_0 + N}, \tag{3.11}$$

with $v = \frac{1}{N} \sum_{i=1}^{N} (y_i - [\beta_1 d_{1i} + \beta_2 d_{2i}])^2$ (Gelman et al., 1995, pp 46-47).

The S iterations lead to S values for each of the model parameters. Since the Gibbs sampler is initiated with arbitrary starting values the first iterations are influenced by these values. This is illustrated in Figure 3.3 on page 50, where it can be seen that the small starting value for β_1 has a strong influence for at least the first 20 iterations. Therefore, a 'burn-in' period is used, meaning that a first set of iterations is discarded. The total number of iterations after burn-in is denoted by Q and must be chosen large enough to make sure that the Gibbs sampler has converged. There are many diagnostics for this matter (see for a summary Cowles and Carlin, 1996). In this paper, graphical plots to monitor multiple chains with different starting values are used. Determining the size of burn-in and checking for convergence will be further discussed in the example of Section 3.5.

3.3.3 Posterior Estimates

The sample of Q iterations, remaining after discarding the burn-in, constitutes a sample from the posterior. Using this sample all desired estimators can be computed. For example, the posterior distribution of each model parameter can be summarized using the average of the Q iterations (providing the posterior mean) and the standard deviation of the Q iterations (providing the posterior standard deviation). The posterior median is obtained by taking the 50-th percentile of the distribution of the Q draws. An advantage of having a full representation is that credibility intervals can easily be obtained even if the shape of the distribution is far from normal. The 95% central credibility interval for each of the model parameters is given by the 2.5-th and 97.5-th percentile of the distribution of the Q values that remain after burn-in.

3.3.4 Inequality Constrained Models

In this section the constraint $\beta_1 > \beta_2$ is added to the model, to illustrate sampling under inequality constraints. Inequality constraints on model parameters can be

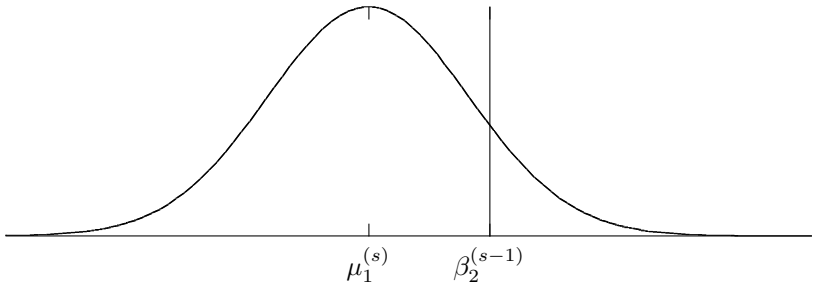

Figure 3.1: Constrained Conditional Posterior Distribution of β_1 in Iteration s with Constraint $\beta_1 > \beta_2$.

built into the specification of the prior distribution (Gelfand et al., 1992). The extra information included in the prior distribution of the model parameters affects also the posterior and thus the conditional distributions of the sampling scheme. Each iteration (consisting of the three steps) should result in a set of values for β_1, β_2 and σ^2 where the constraint $\beta_1 > \beta_2$ is not violated.

The easiest way to accomplish this is to add a checkpoint in the first (respectively second) step of the sampling scheme. A draw of β_1 (respectively β_2) that violates the inequality constraint is discarded. The sampling of $\beta_1^{(s)}$ (see for an illustration Figure 3.1) is repeated until a value that is in accordance with the inequality constraint is sampled, i.e. a value in the right-hand tail of the conditional posterior distribution of $\beta_1^{(s)}$ displayed in Figure 3.1. Keeping only the draws that are in accordance with the constraint does provide a sample from the posterior $p(\beta_1, \beta_2, \sigma^2 | \mathbf{y}, \mathbf{d}_1, \mathbf{d}_2; \beta_1 > \beta_2)$. However, it is not very efficient. Several draws for $\beta_1^{(s)}$ might be necessary before one larger than $\beta_2^{(s-1)}$ is obtained. In both the first and second step of *each* iteration many draws might be discarded. This can slow down the sampler substantially.

A more efficient approach is inverse probability sampling. Via a randomly sampled deviate from the uniform distribution $U(0, 1)$, a matching value for $\beta_1^{(s)}$ that is always in agreement with the constraint is obtained directly. To simplify the explanation (the general procedure is provided in Appendix I) assume that the area below $\beta_2^{(s-1)}$ in Figure 3.1 is .80, so the right-hand tail equals .20. Further assume that the deviate from $U(0, 1)$ in this step appears to be .75. The matching value for $\beta_1^{(s)}$, that is a random draw from the right-hand tail, is the point in the normal distribution with cumulative density: .80 + .75*.20 = .95. This method is applied

in the software accompanying this paper.

3.4 Model Selection

As outlined before there may exist several competing theories. For each theory that is translated into a model using (in)equality constraints, the corresponding set of constraints can be incorporated in the prior distribution of the model parameters. The fit of each of these models after observing the data can be compared using Bayes factors.

Basically, three general types of models can be distinguished, unconstrained, inequality constrained and equality constrained:

$$M_1 : \beta_1 \ , \ \ldots \ , \ \beta_J; \quad \alpha_1, \ldots, \alpha_K, \sigma^2,$$

$$M_2 : \beta_1 < \ldots < \beta_J; \quad \alpha_1, \ldots, \alpha_K, \sigma^2,$$

$$M_3 : \beta_1 = \ldots = \beta_J; \quad \alpha_1, \ldots, \alpha_K, \sigma^2.$$

A special feature of the approach taken in this paper is that all the constrained models are nested in the unconstrained, encompassing model (M_1). This feature is illustrated in Figure 3.2 for $J = 2$, $K = 0$ and, to facilitate the explanation, assuming σ^2 is known. Note, that in the sequel Figure 3.2 will repeatedly be referred to, to illustrate the general approach taken.

In Figure 3.2 three models can be distinguished: M_1: β_1, β_2; M_2: $\beta_1 > \beta_2$; and, M_3: $\beta_1 = \beta_2$. With a diffuse and symmetric prior for the $\beta's$ as specified in the previous section, M_1 implies that a priori each combination of values for β_1 and β_2 is about equally likely. Similarly, M_2 implies that a priori each combination of values for which $\beta_1 > \beta_2$ is about equally likely (the region under the diagonal), and, M_3 that each combination for which $\beta_1 = \beta_2$ is about equally likely. The consequence of this nesting is that once the prior for the unconstrained model is defined, the priors for the constrained models follow directly.

The marginal likelihood (Kass and Raftery, 1995) can be used to quantify the amount of evidence provided by the data in favor of each model/prior. The marginal likelihood for model M_t $(t = 1, \ldots, T)$ is defined as

$$p(\mathbf{y}|\mathbf{X}, \mathbf{D}, M_t) = \int \mathrm{p}(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \mathbf{X}, \mathbf{D}) \pi(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^{\mathbf{2}}|M_t) \mathrm{d}\boldsymbol{\beta} \mathrm{d}\boldsymbol{\alpha} \mathrm{d}\sigma^2, \tag{3.12}$$

where $p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2, \mathbf{X}, \mathbf{D})$ denotes the density of the data, and $\pi(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2|M_t)$ the prior distribution of the model parameters. As can be seen in Figure 3.2 for certain

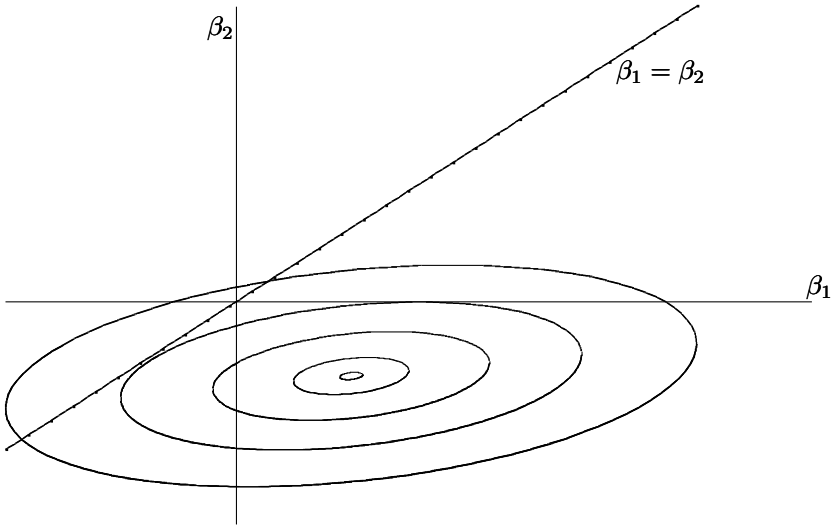

Figure 3.2: Interpretation of the Marginal Likelihood.

parameter values the density of the data is high (the region inside the smallest ellipse). Loosely formulated this implies that the data are quite likely given these parameter values. For other parameter values the density of the data is smaller. In Figure 3.2 this is denoted by ellipses on increasing distances from the central ellipse.

The marginal likelihood is the average of the density of the data over the prior for the model at hand, i.e. over all combinations of parameter values that are admitted by the prior. As can be seen in Figure 3.2, M_1 admits many values of β_1 and β_2 for which the density of the data is rather small (the region outside the largest ellipse). The consequence of the latter is that the marginal likelihood will be relatively small. It can also be seen that M_2 excludes many values of β_1 and β_2 for which the density of the data is rather small. The marginal likelihood of M_2 will be substantially larger than the marginal likelihood of M_1, that is, M_2 is a better model than M_1.

To be able to evaluate models with equality constraints the strict equality constraint ('=') can be replaced by the 'approximately equal' constraint ('$\approx$'), i.e. M_3 is evaluated for a small interval around the line $\beta_1 = \beta_2$. Besides preserving the dimensionality of the encompassing model, $\beta_1 \approx \beta_2$ may also be more realistic than

$\beta_1 = \beta_2$. To quote Berger and Delampady (1987) on this concept: 'It is rare, and perhaps impossible, to have a null hypothesis that can be exactly modeled as $\theta = \theta_0$', and, '... surely there is *some* effect, although perhaps a very miniscule effect'. The researcher decides which interval around the line is meaningful, i.e. how large the difference between the β's must be to be interpreted as relevantly different. From Figure 3.2 (placing an interval around the line $\beta_1 = \beta_2$) it can be seen that M_3 excludes the values of β_1 and β_2 for which the density of the data is relatively large, so the marginal likelihood of M_3 will be relatively small. In conclusion, comparing M_1, M_2 and M_3 in Figure 3.2, M_2 is the best model.

The fit of two models can be compared using Bayes factors (Kass and Raftery, 1995):

$$BF_{tt'} = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{D}, M_t)}{\mathrm{p}(\mathbf{y}|\mathbf{X}, \mathbf{D}, M_{t'})}, \tag{3.13}$$

i.e. the ratio of two marginal likelihoods. If the amount of evidence provided by the data in favor of model t is (much) larger than the amount of evidence in favor of model t', then the value of $\mathrm{BF}_{tt'}$ will be (much) larger than 1. If the results are the opposite, the value for $\mathrm{BF}_{tt'}$ will be smaller than 1.

Estimation of the marginal likelihood can be burdensome (see for example the papers by Newton & Raftery, 1994; Carlin & Chib, 1995; Chib, 1995). In the encompassing setup however it can be avoided. The Bayes factor of each nested model to the encompassing model is estimated without direct computation of the marginal likelihood. This is accomplished by taking a sample from both the prior and the posterior of the unconstrained model, and counting how many of the iterations are in agreement with the constrained model. Denoting the proportion of the prior that is in agreement with the constraints by $\frac{1}{c}$, and the proportion of the posterior that is in agreement with the constraints by $\frac{1}{d}$, the Bayes factor of the constrained to the unconstrained model (BF_{21}) is estimated by $\frac{c}{d}$. The derivation is given in Appendix II.

From the Bayes factors for each of the constrained models with the encompassing model, the posterior model probability for each model in the set of models can be computed (see Appendix II). The posterior model probability is a straightforward and easily interpretable measure of support of the data for each of the models in the set of models under consideration. Two models with posterior probabilities of about .50 and .50 are (after observing the data) equally likely. However, if one model has a posterior probability of .90 and the other of .10, the first is preferred to the latter.

3.5 Analysis of Covariance

The dataset used in this section is part of a large evaluation of the impact of the first year of the Sesame Street television series. Sesame Street was concerned mainly with teaching preschool related skills to children in the 3-5 year age range. Both before and after viewing the series the children were tested on a variety of cognitive variables, including knowledge of body parts, knowledge about letters, numbers, relational terms, etc. (Stevens, 1996, pp 596-603). The sample consists of 240 children and can be divided in four subgroups of the factor Viewing category, based on the intensity of watching the show. Code 1 means 'rarely watched' while children with code 4 watched the show on average more than 5 times a week. The outcome variable (y) of interest is the sum of the six posttest scores and represents cognitive ability after a year of broadcasting Sesame Street. Two pre-treatment measures are added as control variables, the sum of the six pretest scores of cognitive ability (x_1) and a measure of vocabulary maturity (x_2). This leads to model (3.1) with $J = 4$ and $K = 2$. The four β's represent the adjusted means of cognitive ability in the four viewing categories after a year of broadcasting Sesame Street.

Four models are analyzed:

$$\begin{aligned} M_1: &\quad \beta_1 \ , \ \beta_2 \ , \ \beta_3 \ , \ \beta_4, \\ M_2: &\quad \beta_1 \approx \beta_2 \approx \beta_3 \approx \beta_4, \\ M_3: &\quad \beta_1 < \beta_2 < \beta_3 < \beta_4, \\ M_4: &\quad \beta_1 > \beta_2 > \beta_3 > \beta_4. \end{aligned}$$

The last three models represent different theories about the possible outcomes. M_2 represents a 'no effect' theory. After correcting for individual differences on the pretest cognitive ability and vocabulary maturity, no (relevant) posttest differences are expected for the four viewing categories. The second theory does expect differences between the viewing categories: the higher the frequency of viewing the show, the higher the expected posttest score (M_3). In the third theory the opposite effect is expected: the more children still have to learn, the more frequently they will be watching the show (M_4). Not being sure about any of the three theories, the unconstrained model is also included (M_1). This way, the possibility that all three theories are not supported is also examined (in which case the unconstrained model will have a substantially higher posterior probability than M_2, M_3 and M_4).

The prior distributions used for each of the parameters are data based (see Appendix I) and resulted in this example in the normal distribution $N(48.2; 1306.7)$ for each of the β's, the normal distribution $N(.68; .04)$ for α_1 and $N(.30; .14)$ for α_2, and the scaled Inv-$\chi^2(1; 542.0)$ for the error variance. To evaluate model M_2 it

Table 3.2: Posterior Model Probabilities

M_t	$\mathrm{P}(M_t\|\mathbf{y},\mathbf{X},\mathbf{D})$
M_1	.04
M_2	.00
M_3	.96
M_4	.00

Table 3.3: Posterior Means (PM), Posterior Standard Deviations (PSD) and 95% Central Credibility Intervals (CCI) under model M_3

	PM	PSD	95% CCI		(n_j)
β_1	29.7	5.2	21.1	38.1	(54)
β_2	43.6	5.5	34.4	52.6	(60)
β_3	56.2	5.8	46.9	65.6	(64)
β_4	63.5	6.1	53.4	73.3	(62)
α_1	.67	.06	.57	.77	
α_2	.29	.11	.11	.48	
σ^2	542.1	50.5	464.7	629.4	

must be specified when a set of four β's is interpreted as 'approximately equal'. The difference between the smallest and the largest β is compared with a user-specified relevance, in this example 10.0 (which may seem a large value, but is in fact just 5% of the total range (about 200) of the dependent variable).

To compare the models posterior model probabilities are computed. The results are printed in Table 3.2. As can be seen from this table, the posterior model probabilities of the models M_1, M_2 and M_4 are approximately zero, so the models (and accompanying theories) seem very unlikely. Model M_3, with a posterior model probability of .96, fits clearly better. It can be concluded that the higher the frequency of watching Sesame Street, the higher the posttest cognitive ability (controlling for possible differences in pretest cognitive ability and vocabulary maturity). Note that, since this is an observational study no conclusions about a causal relationship can be made.

Estimation under model M_3 leads to the parameter estimates as can be found in Table 3.3. Note that the 95% credibility intervals for α_1 and α_2 are entirely

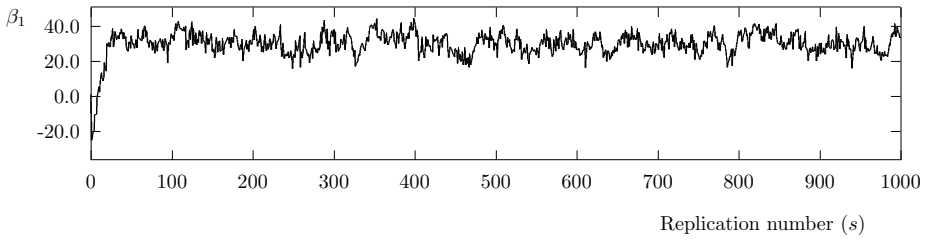

Figure 3.3: One Exemplary Chain for β_1 Against Iteration Number Under M_3.

greater than 0, i.e. there appears to be a positive relationship between both pretest cognitive ability and vocabulary maturity with posttest cognitive ability.

Convergence and burn-in. Convergence and burn-in will be illustrated using the Gibbs sampler that is applied to sample from the posterior of model M_3 (i.e. including inequality constraints on the β's). To choose an appropriate burn-in period and monitor the convergence three chains with different starting values are sampled. The result for just one of the chains of parameter β_1 is plotted in Figure 3.3.

Initially, the influence of the starting value can be seen: the plot between iteration 1 and (say) 100 looks very different from the plot between 100 and 1000. The first part serves as a burn-in period and is discarded. Comparing different intervals of 100 iterations (say 300-400,700-800 and 800-900), it is concluded that they look basically the same. Also the other two chains for β_1 look the same, except for a short burn-in period. It is concluded that the Gibbs sampler has converged. Also for the other parameters convergence was confirmed by monitoring several chains with different starting values.

Note, that the plot (after burn-in) represents a sample from the posterior distribution. With this sample the posterior mean, posterior standard deviation and credibility intervals can be computed.

3.6 Three-way Table of Ordered Means

For this example the same dataset is used, but a three-way design is obtained by the frequency of watching Sesame Street, gender and sampling site. The latter has two categories: disadvantaged children and advantaged children. The dependent variable is the sum of the six posttest scores and represents cognitive ability after

Table 3.4: Posterior Model Probabilities

M_t	$\mathrm{P}(M_t\|\mathbf{y},\mathbf{X},\mathbf{D})$
M_1	.00
M_2	.23
M_3	.77

a year of broadcasting Sesame Street. The two covariates pretest cognitive ability (x_1) and vocabulary maturity (x_2) are again part of the model. This leads to model (3.1) with $J = 16$ and $K = 2$. The 16 β's represent the adjusted means of posttest cognitive ability in the 16 groups formed by the $2 \times 2 \times 4$ design. The number of observations of the groups vary between 7 and 26, but neither this difference nor the cells with only 7 observations causes any trouble to the Bayesian analysis of models with ordered (adjusted) means.

In this example two theories will be compared with the unconstrained model (M_1). The first theory states that there is an ordered effect of frequency of watching within each of the four subgroups formed by the other two factors sampling site and gender (the higher the frequency of watching, the higher the corrected posttest scores). Note, that four ordered simple effects are modeled here simultaneously (M_2). The second theory includes the same effect of frequency of watching, but in addition also an effect of sampling site is modeled (M_3). It is expected that children from advantaged environments have higher corrected posttest scores within each subgroup formed by the frequency of watching and gender, than children from disadvantaged environments.

The prior distribution used for each of the β's is $N(44.1; 2300.1)$. For α_1, α_2 and σ^2 the prior distributions are the same as in the previous example. The posterior model probabilities of each of the models can be found in Table 3.4. Model M_3 has the highest posterior model probability (.77) so it can be concluded that, besides the ordered effect of frequency of watching, children from advantaged environments gain more from watching Sesame Street than children from disadvantaged environments (after controlling for possible differences on the covariates). The parameter estimates of the adjusted means under model M_3 are printed in Table 3.5. It can be seen that the posterior means are ordered in two ways. Within each subgroup formed by site and gender the posterior means increase with frequency of watching, and, for each subgroup formed by gender and frequency of watching the posterior means for the advantaged environment are higher than for the disadvantaged

Table 3.5: Posterior Means (PM) and Posterior Standard Deviations (PSD) for the Adjusted Means Under Model M_3

		Boys:			Girls:		
		PM	PSD	(n_j)	PM	PSD	(n_j)
Advantaged	View: 1	35.1	6.4	(7)	31.7	6.8	(9)
	2	44.6	6.3	(14)	50.6	6.4	(13)
	3	54.7	3.3	(21)	60.7	6.1	(16)
	4	69.7	7.5	(13)	67.2	6.3	(26)
Disadvantaged	View: 1	29.3	6.0	(16)	25.1	6.0	(22)
	2	36.9	6.0	(16)	44.7	6.2	(17)
	3	47.5	6.6	(12)	54.9	6.2	(15)
	4	59.4	7.0	(16)	59.6	6.6	(7)

environment.

3.7 Discussion

In this paper normal linear models with inequality constraints on the model parameters have been discussed. It is shown that a wide variety of models and sets of constraints can be analyzed using Bayesian computational statistics. In the Bayesian approach, theories are translated into statistical models that differ in the constraints that are imposed on the model parameters. Posterior model probabilities provide measures of fit for all models (theories) after observing the data, that are easily interpretable.

Compared to traditional hypothesis testing (e.g. a non-directional ANOVA followed by several directional post-hoc tests) there are several advantages. Firstly, a gain of power is obtained by including the constraints from the start of the analysis. Secondly, combining several test-results to evaluate one theory (model) is no longer necessary. Finally, non-nested models can be compared. Compared to the 'classic' methods that deal with inequality constraints (Barlow et al., 1972, Robertson et al., 1988) the Bayesian approach is more flexible: a large number of groups and covariates can be analyzed, group sizes may differ and sample sizes may be small. Furthermore, posterior standard deviations and credibility intervals can be obtained.

Software. The software for both model selection and parameter estimation are available on the following website: www.fss.uu.nl/ms/ik Besides two executables ('selection.exe' and 'estimate.exe') and a manual, also illustrations consisting of datasets and accompanying input files are provided.

Appendix I

Prior Distribution
In this paper, the general joint prior distribution for model M_t, including inequality constraints, has the form:

$$\pi(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2 | M_t) \propto$$
$$I_{\boldsymbol{\beta} \in M_t} \prod_{j=1}^{J} N(\beta_j | \mu_0, \tau_0^2) \prod_{k=1}^{K} N(\alpha_k | \eta_{k0}, \omega_{k0}^2) \text{ Inv-}\chi^2(\sigma^2 | \nu_0, \kappa_0^2), \tag{3.14}$$

where the indicator function I equals 1 if $\boldsymbol{\beta}$ in M_t and 0 if $\boldsymbol{\beta}$ not in M_t. The indicator function shows that a constrained model is obtained through a truncation of the parameter space of the unconstrained model.

Note that only the encompassing prior must be defined (i.e. (3.14) where the indicator function is not active). The specification of the parameters of the prior distributions is data based. For the β's (note that $\pi(\beta_1) = \ldots = \pi(\beta_J)$) the same procedure as described in Section 3.3.1 is applied. The 99.7% credibility intervals for each of the β's are the basis for an overall interval that represents the range of the data where the prior should be vague. From the lower bound (LB) and upper bound (UB) of this overall interval the parameters $\mu_0 = \frac{UB+LB}{2}$, and, $\tau_0^2 = (\frac{UB-LB}{2})^2$ are derived. For the model selection procedures the prior distributions for the other parameters α_k $(k = 1, \ldots, K)$ and σ^2 are the same for each of the models (the models differ only with respect to constraints imposed on β's) and are therefore not very influential. For each of the α's the 99.7% credibility interval is used to determine the normal prior distribution. The mean minus one standard deviation of the prior normal equals the lower bound of this credibility interval, and, the mean plus one standard deviation equals the upper bound. The parameters for the scaled inverse χ^2-distribution are one (ν_0), and, the data based estimate of the error variance (κ_0^2).

Posterior distribution
The product of the joint prior distribution (3.14) and the density of the data

$$g(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\alpha},\sigma^2,\mathbf{X},\mathbf{D}) = \prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{(y_i-[\sum_{j=1}^{J}\beta_j d_{ji}+\sum_{k=1}^{K}\alpha_k x_{ki}])^2}{\sigma^2}\right), \tag{3.15}$$

leads to the general posterior distribution for model M_t:

$$p(\boldsymbol{\beta},\boldsymbol{\alpha},\sigma^2|\mathbf{y},\mathbf{X},\mathbf{D},M_t) \propto\ g(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\alpha},\sigma^2,\mathbf{X},\mathbf{D})$$
$$\prod_j N(\beta_j|\mu_0,\tau_0^2)\ \prod_k N(\alpha_k|\eta_{k0},\omega_{k0}^2)\ \text{Inv-}\chi^2(\sigma^2|\nu_0,\kappa_0^2)\ \ I_{\boldsymbol{\beta}\in M_t}. \tag{3.16}$$

Gibbs sampler
To sample from (3.16) the Gibbs sampler is applied. Each iteration of the Gibbs sampler consists of the following steps:

1. For $j = 1, ..., J$ sample β_j from

$$p(\beta_j \mid \beta_1,...,\beta_{j-1},\beta_{j+1},...,\beta_J,\boldsymbol{\alpha},\sigma^2,\mathbf{y},\mathbf{X},\mathbf{d}_j,M_t) \propto$$
$$\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2}\frac{(y_i-[\sum_{j'=1}^{J}\beta_{j'}d_{j'i}+\sum_{k=1}^{K}\alpha_k x_{ki}])^2}{\sigma^2})\ N(\beta_j|\mu_0,\tau_0^2)I_{\boldsymbol{\beta}\in M_t}, \tag{3.17}$$

   i.e. the posterior as in (3.16) but now as a function of only β_j. The posterior for β_j is equal to a truncated normal distribution with expectation and variance:

$$\mu_j = \frac{\frac{1}{\tau_0^2}\mu_0+\frac{1}{\sigma^2}(\sum_{i=1}^{N}d_{ji}y_i-\sum_{k=1}^{K}(\alpha_k\sum_{i=1}^{N}d_{ji}x_{ki}))}{\frac{1}{\tau_0^2}+\frac{1}{\sigma^2}\sum_{i=1}^{N}d_{ji}}, \tag{3.18}$$

$$\tau_j^2 = \left(\frac{1}{\tau_0^2} + \frac{1}{\sigma^2} \sum_{i=1}^{N} d_{ji} \right)^{-1}. \tag{3.19}$$

2. For $k = 1, ..., K$ sample α_k from

$$p(\alpha_k \mid \boldsymbol{\beta}, \alpha_1, ..., \alpha_{k-1}, \alpha_{k+1}, ..., \alpha_K, \sigma^2, \mathbf{y}, \mathbf{X}, \mathbf{D}, M_t),$$

i.e. a normal distribution with expectation and variance:

$$\eta_k = \frac{\frac{1}{\omega_{k0}^2}\eta_{k0} + \frac{1}{\sigma^2}\left(\sum_{i=1}^{N} y_i x_{ki} - \sum_{i=1}^{N}\sum_{j=1}^{J} x_{ki}\beta_j d_{ji} - \sum_{i=1}^{N}\sum_{k' \in \mathcal{A}} \alpha_{k'} x_{ki} x_{k'i} \right)}{\frac{1}{\omega_{k0}^2} + \frac{1}{\sigma^2}\sum_{i=1}^{N} x_{ki}^2}, \tag{3.20}$$

with $\mathcal{A} = \{1, \ldots, k-1, k+1, \ldots, K\}$,

$$\omega_k^2 = \omega_{k0}^2 + \frac{\sigma^2}{\sum_{i=1}^{N} x_{ki}^2}. \tag{3.21}$$

3. Sample σ^2 from $p(\sigma^2 \mid \boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{y}, \mathbf{X}, \mathbf{D}, M_t)$, i.e. a scaled inverse χ^2-distribution with parameters:

$$\nu = N + \nu_0, \tag{3.22}$$

$$\kappa^2 = \frac{\nu_0 \kappa_0^2 + N v}{\nu_0 + N}, \tag{3.23}$$

with $v = \frac{1}{N}\sum_{i=1}^{N}(y_i - [\sum_{j=1}^{J} \beta_j d_{ji} + \sum_{k=1}^{K} \alpha_k x_{ki}])^2$ (Gelman et al., 1995, pp 46-47).

Inverse probability sampling
To sample from truncated normal distributions inverse probability sampling is used. To sample β_j from the truncated normal (3.17), the procedure introduced by Gelfand et al. (1992, p. 525) is applied. The largest $\beta_{j'}$ $(j \neq j')$ that must be smaller than β_j is the lower-bound a. The smallest $\beta_{j'}$ $(j \neq j')$ that must be greater than β_j is the upper-bound b. The sampling is achieved using an uniform (0,1) deviate U and the computation of $\beta_j = \Phi^{-1}_{\beta_j}[\Phi_{\beta_j}(a) + U(\Phi_{\beta_j}(b) - \Phi_{\beta_j}(a))]$, where $\Phi_{\beta_j}(a)$ is the density of (3.17) below a and $\Phi_{\beta_j}(b)$ is the density of (3.17) below b. $\Phi^{-1}_{\beta_j}[.]$ denotes the inverse cumulative density evaluated at the argument. This procedure always renders a deviate from (3.17) within the bounds a and b.

Appendix II

Estimation of Bayes Factors
Using a general notation, the marginal likelihood $p(\boldsymbol{y}|M_t)$ for data $\boldsymbol{y}$ and parameter vector $\boldsymbol{\theta}$ (in this paper $\boldsymbol{\theta}$=$\{\boldsymbol{\beta},\boldsymbol{\alpha},\sigma^2\}$) can be written as

$$p(\boldsymbol{y}|M_t) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta}, M_t)\pi(\boldsymbol{\theta}|M_t)}{\pi(\boldsymbol{\theta}|\boldsymbol{y}, M_t)}, \tag{3.24}$$

where the numerator is the product of the sampling density and the prior, with all integrating constants included, and the denominator is the posterior density (Chib, 1995). Consequently, the Bayes factor for two models M_1 and M_2 can be written as follows:

$$\mathrm{BF}_{21} = \frac{f(\boldsymbol{y}|\boldsymbol{\theta}, M_2)\pi(\boldsymbol{\theta}|M_2)/\pi(\boldsymbol{\theta}|\boldsymbol{y}, M_2)}{f(\boldsymbol{y}|\boldsymbol{\theta}, M_1)\pi(\boldsymbol{\theta}|M_1)/\pi(\boldsymbol{\theta}|\boldsymbol{y}, M_1)}. \tag{3.25}$$

Note that, M_1 is the encompassing (unconstrained) model and that the constrained model M_2 is nested in M_1. The models contain the same parameters, so for a value of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}'$, that exists in both models M_1 and M_2 the Bayes factor as given in (3.25) reduces to

$$\mathrm{BF}_{21} = \frac{\pi(\boldsymbol{\theta}'|M_2)/\pi(\boldsymbol{\theta}'|\boldsymbol{y}, M_2)}{\pi(\boldsymbol{\theta}'|M_1)/\pi(\boldsymbol{\theta}'|\boldsymbol{y}, M_1)}, \tag{3.26}$$

because $f(\boldsymbol{y}|\boldsymbol{\theta}',\ M_1) = f(\boldsymbol{y}|\boldsymbol{\theta}',\ M_2)$.

Furthermore, the model parameters have the same prior distributions, except for the truncation of the prior for M_2 by the constraints. Therefore, $\pi(\boldsymbol{\theta}'|M_2)$ can be written as $c \cdot \pi(\boldsymbol{\theta}'|M_1)$, as well as $\pi(\boldsymbol{\theta}'|\boldsymbol{y}, M_2) = d \cdot \pi(\boldsymbol{\theta}'|\boldsymbol{y}, M_1)$. This leads to

$$\mathrm{BF}_{21} = \frac{c \cdot \pi(\boldsymbol{\theta}'|M_1)/d \cdot \pi(\boldsymbol{\theta}'|\boldsymbol{y}, M_1)}{\pi(\boldsymbol{\theta}'|M_1)/\pi(\boldsymbol{\theta}'|\boldsymbol{y}, M_1)} = \frac{c}{d}. \tag{3.27}$$

The proportion of the encompassing prior that is in agreement with the constraints of model M_2 provides the value $\frac{1}{c}$, the proportion of the encompassing posterior in agreement with the constraints of M_2 is $\frac{1}{d}$. The values c and d are estimated via samples from the encompassing prior and posterior respectively.

Posterior model probabilities

The posterior odds of two models (M_1 and M_2) is the product of the prior odds and the Bayes factor:

$$\frac{p(M_2|\boldsymbol{y})}{p(M_1|\boldsymbol{y})} = \frac{p(M_2)}{p(M_1)}\ \mathrm{BF}_{21}, \tag{3.28}$$

where $p(M_t)$ is the prior probability and $p(M_t|\boldsymbol{y})$ the posterior probability of model M_t $(t = 1, 2)$. Assuming that a priori each model is equally likely, i.e. $p(M_t) = 1/T$, for $t = 1, \ldots, T$, Bayes factors can straightforwardly be interpreted as the posterior odds of the models involved. From the posterior odds of each of the constrained models $(M_2, \ldots M_T)$ to the unconstrained model (M_1) the posterior model probabilities for each model in a set of T models can be computed, using:

$$p(M_t|\boldsymbol{y}) = \frac{\mathrm{BF}_{t1}}{\mathrm{BF}_{11} + \mathrm{BF}_{21} + \ldots + \mathrm{BF}_{T1}}, \quad \text{for } \ t = 1, \ldots, T, \tag{3.29}$$

with $\mathrm{BF}_{11} = 1$.

Chapter 4

Encompassing Priors

Abstract

In this paper Bayesian model selection using encompassing priors is introduced. The focus will be on models that are specified using inequality constraints among the model parameters. A nice feature of the encompassing set-up is that only a prior for the parameters of the unconstrained model has to be specified. It will be shown that, the prior distribution of each constrained model follows directly from this encompassing prior. The issue of sensitivity to model specific prior distributions now reduces to sensitivity to one prior distribution. It will be shown that for specific classes of models, selection is virtually independent of the encompassing prior, that is virtually objective.

The encompassing prior leads to a nice interpretation of Bayes factors. The Bayes factor for any constrained model with the encompassing model reduces to the ratio of two proportions, namely the proportion of respectively the encompassing prior and posterior, in agreement with the constraints. This enables easy and straightforward estimation of the Bayes factor and its standard error. It is also rather efficient because with only one sample from the encompassing prior and one sample from the encompassing posterior, Bayes factors for all pairs of models are obtained.

4.1 Introduction

In this paper the use of encompassing priors in Bayesian model selection is introduced. The focus will be on models that are specified using inequality constraints among the model parameters. The encompassing prior is the prior specified for the model without constraints on the model parameters. Consider a parameter vector $\boldsymbol{\theta}$ of dimension J, for a finite set $\mathcal{M} = \{M_1, \ldots, M_T\}$ of T models:

$$\begin{array}{ll} M_1: & \boldsymbol{\theta} \in \boldsymbol{\Theta}_1 \\ M_2: & \boldsymbol{\theta} \in \boldsymbol{\Theta}_2 \subset \boldsymbol{\Theta}_1 \\ \ldots & \\ M_T: & \boldsymbol{\theta} \in \boldsymbol{\Theta}_T \subset \boldsymbol{\Theta}_1. \end{array} \tag{4.1}$$

In (4.1), M_1 is the encompassing, unconstrained model wherein M_2 through M_T are nested. Examples of nested model are $M_2 : \theta_1 < \theta_2 < \theta_3$ and $M_3 : \theta_1 \approx \theta_2 \approx \theta_3$, with corresponding encompassing model $M_1 : \theta_1, \theta_2, \theta_3$.

Estimation and testing of inequality constrained models has received a lot of attention in the frequentist framework (see Barlow, Bartholomew, Bremner and Brunk, 1972; Robertson, Wright, and Dykstra, 1988; and a special issue of the Journal of Statistical Planning and Inference: Statistical Inference under Inequality Constraints, 2002, Volume 107, Numbers 1-2). In the Bayesian framework inequality constrained models have not yet received a lot of attention. However, Gelfand, Smith and Lee (1992) showed that inequality constraints can straightforwardly be built into the prior distribution.

Besides models based on inequality constraints, it often is also of interest to include a model reflecting a 'no-effect' theory. In this paper the latter is incorporated using approximate equality constraints between parameters. For instance, parameters can be assumed to be approximately equal if the largest difference between any pair of parameters is smaller than a specified value, that is, $|\theta_v - \theta_w| < \epsilon \; \forall v, w \; | v \neq w$ and a small value for ϵ. This will be elaborated in Section 4.4.2.

A nice feature of the encompassing set-up is that just *one* prior for the parameter vector $\boldsymbol{\theta}$ has to be specified: $\pi(\boldsymbol{\theta}|M_1)$, the so-called *encompassing prior*. The prior distribution of M_2, M_3 and in general of any nested model M_t, follows directly from this encompassing prior, using

$$\pi(\boldsymbol{\theta}|M_t) = \frac{\pi(\boldsymbol{\theta}|M_1) I_{M_t}}{\int \pi(\boldsymbol{\theta})|M_1) I_{M_t} \, \mathrm{d}\boldsymbol{\theta}}, \tag{4.2}$$

where the indicator function I_{M_t} has the value one if the argument is true, that is, if the parameter values are in accordance with the constraints imposed by model M_t, and zero otherwise.

The evaluation of models after observing data $\boldsymbol{y}$ is based on Bayes factors. The Bayes factor for M_1 and M_t is the ratio of the two corresponding marginal likelihoods:

$$\mathrm{BF}_{t1} = \frac{p(\boldsymbol{y}|M_t)}{p(\boldsymbol{y}|M_1)} = \frac{\int f(\boldsymbol{y}|\boldsymbol{\theta}, M_t)\pi(\boldsymbol{\theta}|M_t)d\boldsymbol{\theta}}{\int f(\boldsymbol{y}|\boldsymbol{\theta}, M_1)\pi(\boldsymbol{\theta}|M_1)d\boldsymbol{\theta}}, \tag{4.3}$$

for $t = 2, \ldots, T$. Note that from the Bayes factors for each of the constrained models $(M_2, \ldots M_T)$ with the encompassing model (M_1), the Bayes factor for two nested models is obtained by

$$\mathrm{BF}_{t't} = \frac{\mathrm{BF}_{t'1}}{\mathrm{BF}_{t1}}. \tag{4.4}$$

The integrands in (4.3) are often complicated to compute analytically. Therefore, Bayes factors or marginal likelihoods are often estimated using Monte Carlo methods. (e.g. Newton and Raftery, 1994; Gelfand and Dey, 1994; Kass and Raftery, 1995; Chib, 1995; Green, 1995; Carlin and Chib, 1995; Chen and Shao, 1997). Some general concerns of these methods are computational complexity, convergence, and, efficiency (see for a recent comparative review Han and Carlin, 2001). However, the encompassing prior approach leads to a straightforward estimate of the Bayes factor, without the requirement to compute marginal likelihoods.

Another issue is the sensitivity of Bayes factors to the prior. In (4.3), it can be seen that the Bayes factor does not only depend on the density of the data, $f(\boldsymbol{y}|\boldsymbol{\theta}, M_t)$, but also on the choice of the prior distributions for the model parameters for each model, $\pi(\boldsymbol{\theta}|M_t)$ for $M_t \in \mathcal{M}$. Defining these model specific prior distributions is not straightforward. This has led to a search for 'noninformative' or 'objective' priors. Using such priors (with slightly different interpretations also denoted by reference, default or conventional priors) works well for estimation problems (as long as the data dominate the prior, the precise form of the latter is unimportant) but are usually problematic in model selection procedures. Improper noninformative priors almost always lead to indeterminate Bayes factors (Berger and Pericchi, 1996). Proper 'noninformative' priors (also called vague or diffuse priors) do not solve this problem. Berger and Pericchi (2001) conclude that using vague priors often result in arbitrary values for Bayes factors.

Our approach is based on the use of an encompassing prior, that is specified using the general properties described in the next section. It has the following advantages:

- In the encompassing prior approach only *one* prior needs to be specified, namely the encompassing prior. The prior distributions for the model parameters of the nested models follow from the encompassing prior via (4.2).
- For specific classes of models (discussed in Section 4.4) vague encompassing priors can be used and lead to sensible Bayes factors. For those models, the model selection procedure is virtually objective as long as the prior is dominated by the data, that is the actual specification of the (vague) encompassing prior does almost not affect the resulting Bayes factor.

The rest of the paper is organized as follows. In Section 4.2, the general properties of the encompassing prior are introduced. Subsequently, our estimate of Bayes factors in the context of encompassing priors is presented in Section 4.3. In Section 4.4, it will be shown that for specific classes of models, Bayes factors are virtually independent of the encompassing prior. For other classes of models, the sensitivity of Bayes factors to encompassing priors is discussed. Section 4.5 presents some general examples of models where the encompassing prior approach can be used. The paper will be concluded with two examples, one dealing with normal data (Section 4.6) and one with multinomial data (Section 4.7), in which a number of the features of our approach will be illustrated. The paper will be concluded with a discussion in Section 4.8.

4.2 Specification of Encompassing Priors

In this section it will be outlined how the encompassing prior is specified. Applying the criteria of this section, leads to: i) a nice interpretation of the Bayes factor of any nested model with the encompassing model, as will be shown in Section 4.3; and ii) virtually objective model selection for specific classes of models, as will be discussed in Section 4.4.

Let $\boldsymbol{\theta}$= $\{\boldsymbol{\theta}^c, \boldsymbol{\theta}^u\}$, where $\boldsymbol{\theta}^c$ contains model parameters that are involved in constraints, and, $\boldsymbol{\theta}^u$ contains one or more parameters that are unconstrained in all models. A priori $\boldsymbol{\theta}^c$ and $\boldsymbol{\theta}^u$ are considered to be independent. For notational convenience, in the sequel the encompassing prior $\pi(\boldsymbol{\theta}|M_1) = \pi(\boldsymbol{\theta}^c|M_1)\pi(\boldsymbol{\theta}^u|M_1)$ is denoted by $\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}^c)\pi(\boldsymbol{\theta}^u)$.

In this paper the encompassing prior is defined using the following properties

1. The encompassing prior $\pi(\boldsymbol{\theta})$ is continuous on the real line or a subsection of the real line.

2. The encompassing prior, $\pi(\boldsymbol{\theta})$, is diffuse. Using diffuse or vague priors leads to an encompassing posterior that is virtually independent of the prior. Note that vague priors rarely work well for testing or model selection problems (Berger and Pericchi, 2001). However, it will be shown in Section 4.4.1 that for specific classes of models the encompassing prior approach can be used with diffuse, vague priors, and, that the results of the model selection procedure are not sensitive to the actual specification of the encompassing prior, as long as it is dominated by the data. Model $M_2 : \theta_1 < \theta_2 < \theta_3$ from the introduction, is an example of such a model.

   On the contrary, comparing the encompassing model with $M_3 : \theta_1 \approx \theta_2 \approx \theta_3$ from the introduction, provides an example of a class of models for which the Bayes factor does depend on the encompassing prior distribution. The Bayes factor for the nested with the encompassing model is increasing in favor of the nested model (M_3) when the prior gets more diffuse, irrespective of the data. This is a phenomenon well-known as Bartlett's or Lindley's paradox (e.g. Lindley, 1957; Bernardo and Smith, 1994). A subjective choice with respect to the degree of diffuseness of the encompassing prior must be made. This will be elaborated upon in the fourth criterion.

3. The encompassing prior does not favor any of the models, i.e. the prior is similar for those elements that are subjected to constraints in one or more of the nested models.

   - When the parameters contained in $\boldsymbol{\theta}^c$ are a priori considered to be independent, this means that for all elements the same prior distribution is specified, i.e. $\pi(\boldsymbol{\theta}^c) = \pi(\theta_1) \cdots \pi(\theta_k)$, with $\pi(\theta_k) = \pi(\theta)$ for $\theta_k \in \boldsymbol{\theta}^c$ and $k = 1, \ldots, K$. If $\boldsymbol{\theta}^c = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_B\}$ for $b = 1, \ldots, B$, i.e. $\boldsymbol{\theta}^c$ falls apart into B groups, and the constraints apply only to parameters from the same group, then $\pi(\theta_k) = \pi^b(\theta)$ for $\theta_k \in \boldsymbol{\theta}_b$.
   - When the parameters contained in $\boldsymbol{\theta}^c$ are a priori considered to be correlated, this means that also the correlations between all constrained parameters are a priori considered to be equal. For example, when the encompassing prior is a multivariate normal distribution, it is specified using equal prior means, equal prior variances and equal prior covariances. However, the influence of any prior correlation between parameters is dominated by the large prior variances (property 2), and therefore not very relevant.

4. For the class of models where the model selection is sensitive to the encompassing prior (e.g. for $M_3 : \theta_1 \approx \theta_2 \approx \theta_3$) a subjective choice for the specification of the prior distribution must be made. The resulting Bayes factor depends on the encompassing prior distribution.

   In the context of precise hypotheses (e.g. $\theta = \theta_0$ versus $\theta \neq \theta_0$) the same phenomenon occurs. In this context many suggestions for 'automatic' or default priors can be found in the literature (for instance, Jeffreys, 1961; Zellner and Siow, 1980; Smith and Spiegelharter, 1980). Although the exact functional form of the prior is rather irrelevant, the scale factor (i.e. the amount of 'vagueness' of the prior) has a large effect on the answer. Berger and Delampady (1987) conclude that there is no prior distribution for θ that can claim to be objective. In Section 4.4.2, it will be shown that the same holds for models using approximate equality constraints (e.g. $\theta_1 \approx \theta_2$) and for models including effect sizes (e.g. $\theta_1 > \theta_2 + \epsilon$).

   For the class of models where the results are strongly affected by the amount of vagueness of the encompassing prior, the following additional criterion is used: the prior should not exclude regions of the parameter space with substantial posterior probability, but it should also not include too many values that are completely out of range for the data at hand, i.e. the prior is restricted to the region of interest, aiming for a balance between diffuse ('noninformative') but not too diffuse (to avoid Lindley's paradox).

   When the parameter space is bounded, a restricted prior range follows automatically. For example, if θ is a probability, the parameter space is 'naturally' bounded by zero and one. Specification of the encompassing prior can also be based on the response-format of the data. For instance, if data are measured on a scale from zero to ten, the prior distribution for the mean of this data is chosen such that it is vague for the whole range zero through ten, and zero otherwise. In other situations, the range of the observed data can be used to specify the encompassing prior. An example is given in Section 4.6.1.

4.3 The Bayes Factor for a Nested and the Encompassing Model

In this section it will be shown that the use of an encompassing prior, specified as outlined in the previous section, renders a nice interpretation of the Bayes factor for a constrained (M_t) with the encompassing model (M_1).

The marginal likelihood $p(\boldsymbol{y}|M_t)$, for $t = 1, \ldots, T$, is written as follows:

$$p(\boldsymbol{y}|M_t) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta}, M_t)\pi(\boldsymbol{\theta}|M_t)}{\pi(\boldsymbol{\theta}|\boldsymbol{y}, M_t)}, \tag{4.5}$$

where the numerator is the product of the density and the prior, with all integrating constants included, and the denominator is the posterior density (Chib, 1995). Consequently, the Bayes factor for M_t with M_1 is:

$$\text{BF}_{t1} = \frac{f(\boldsymbol{y}|\boldsymbol{\theta}, M_t)\pi(\boldsymbol{\theta}|M_t)/\pi(\boldsymbol{\theta}|\boldsymbol{y}, M_t)}{f(\boldsymbol{y}|\boldsymbol{\theta}, M_1)\pi(\boldsymbol{\theta}|M_1)/\pi(\boldsymbol{\theta}|\boldsymbol{y}, M_1)}. \tag{4.6}$$

For a value of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}'$, that is allowed in both models M_1 and M_t (4.6) reduces to

$$\text{BF}_{t1} = \frac{\pi(\boldsymbol{\theta}'|M_t)/\pi(\boldsymbol{\theta}'|\boldsymbol{y}, M_t)}{\pi(\boldsymbol{\theta}'|M_1)/\pi(\boldsymbol{\theta}'|\boldsymbol{y}, M_1)}, \tag{4.7}$$

because $f(\boldsymbol{y}|\boldsymbol{\theta}', M_1) = f(\boldsymbol{y}|\boldsymbol{\theta}', M_t)$.
Note that, $\boldsymbol{\theta}'$ can include parameters that are involved in constraints, $\boldsymbol{\theta}^c$, and, parameters that are unconstrained in all models, $\boldsymbol{\theta}^u$. Because $\pi(\boldsymbol{\theta}^c)$ and $\pi(\boldsymbol{\theta}^u)$ are a priori assumed to be independent, (4.7) is equal to

$$\text{BF}_{t1} = \frac{\pi(\boldsymbol{\theta}^c|M_t)\pi(\boldsymbol{\theta}^u|M_t)/\pi(\boldsymbol{\theta}'|\boldsymbol{y}, M_t)}{\pi(\boldsymbol{\theta}^c|M_1)\pi(\boldsymbol{\theta}^u|M_1)/\pi(\boldsymbol{\theta}'|\boldsymbol{y}, M_1)}, \tag{4.8}$$

where $\pi(\boldsymbol{\theta}^u|M_1)$ and $\pi(\boldsymbol{\theta}^u|M_t)$ are equal and therefore cancel. Furthermore, $\boldsymbol{\Theta}_t \subset \boldsymbol{\Theta}_1$, as was defined in (4.1) in the introduction. Therefore, $\pi(\boldsymbol{\theta}^c|M_t) = c_t \cdot \pi(\boldsymbol{\theta}^c|M_1)$ and $\pi(\boldsymbol{\theta}'|\boldsymbol{y}, M_t) = d_t \cdot \pi(\boldsymbol{\theta}'|\boldsymbol{y}, M_1)$, where c_t and d_t are constants. This leads to

$$\text{BF}_{t1} = \frac{c_t \cdot \pi(\boldsymbol{\theta}^c|M_1)/d_t \cdot \pi(\boldsymbol{\theta}'|\boldsymbol{y}, M_1)}{\pi(\boldsymbol{\theta}^c|M_1)/\pi(\boldsymbol{\theta}'|\boldsymbol{y}, M_1)} = \frac{c_t}{d_t}. \tag{4.9}$$

The proportion of the encompassing *prior* that is in agreement with the constraints of model M_t provides the value $1/c_t$, the proportion of the encompassing *posterior* in agreement with the constraints of M_t is $1/d_t$. The ratio $\frac{1/d_t}{1/c_t} = \frac{c_t}{d_t}$, nicely reflects that the Bayes factor for two models (here M_t and M_1) actually measures the change in the odds in favor of M_t when going from the prior to the posterior (Lavine and Schervish, 1999). In the next section, the quantities $1/c_t$ and $1/d_t$ will be used to examine the sensitivity to the encompassing prior.

4.4 Sensitivity to the Encompassing Prior

For a large class of models Bayes factors are virtually objective, that is, independent of the actual specification of a vague encompassing prior. This will be illustrated using $1/c_t$ and $1/d_t$, the proportion of the encompassing prior and posterior in agreement with a constrained model, respectively. The sensitivity with respect to $\pi(\boldsymbol{\theta}^c)$ and $\pi(\boldsymbol{\theta}^u)$ will be discussed separately.

In the previous section, it was already seen that $\pi(\boldsymbol{\theta}^u)$ does not influence $1/c_t$. It is also clear that $\pi(\boldsymbol{\theta}^u)$ does influence the posterior distribution, and thus $1/d_t$. However, it is well-known (see, for example, Gelman, Carlin, Stern and Rubin, 1995, p.101) that this influence is small when vague priors are used, i.e. if the data dominate the prior distribution. Stated otherwise, the proportion of the encompassing posterior in agreement with the constraints of model M_t is almost independent of the encompassing prior.

Similarly, if $\pi(\boldsymbol{\theta}^c)$ is vague, its influence on the posterior distribution is negligible, that is, its influence on $1/d_t$ is negligible. The influence of $\pi(\boldsymbol{\theta}^c)$ on $1/c_t$ is discussed for different classes of models in the next two subsections. Both models where $1/c_t$ is independent and dependent of $\pi(\boldsymbol{\theta}^c)$ will be discussed.

4.4.1 Classes of Models that Allow Virtually Objective Model Selection

Let $\boldsymbol{\theta}^c = \{\boldsymbol{\theta}^{(1)},\ \boldsymbol{\theta}^{(2)}\}$. For models using a constraint of the form

$$\sum_{p=1}^{P} m_p\ \theta_p^{(1)} > \sum_{q=1}^{Q} n_q\ \theta_q^{(2)}, \tag{4.10}$$

with the restriction $\sum_{p=1}^{P} m_p = \sum_{q=1}^{Q} n_q$, the value for $1/c_t$ is independent of the encompassing prior $\pi(\boldsymbol{\theta}^c)$. Firstly, consider an encompassing prior assuming independence between the prior elements, that is $\pi(\boldsymbol{\theta}^c) = \pi(\theta_1^{(1)}) \cdots \pi(\theta_P^{(1)})\pi(\theta_1^{(2)}) \cdots \pi(\theta_Q^{(2)})$. According to the third property of Section 4.2, $\theta_p^{(1)}, \theta_q^{(2)} \overset{i.i.d.}{\sim} \pi(\theta)$, for $p = 1, \ldots, P$ and $q = 1, \ldots, Q$. As a consequence $P(\theta_p^{(1)} > \theta_q^{(2)}) = .50\ \forall p, q$. Therefore, also $P(\sum_{p=1}^{P} m_p\ \theta_p^{(1)} > \sum_{q=1}^{Q} n_q\ \theta_q^{(2)})$ with $\sum_{p=1}^{P} m_p = \sum_{q=1}^{Q} n_q$ is .50, independent of the choice of $\pi(\theta)$.

In a similar way, $P(\theta_p^{(1)} > \theta_q^{(2)})$ for

$$\pi(\boldsymbol{\theta}^c) = \pi \begin{pmatrix} \theta_p \\ \theta_q \end{pmatrix} \sim N \left(\begin{bmatrix} \theta_p \\ \theta_q \end{bmatrix} \middle| \middle| \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma^2 & \tau \\ \tau & \sigma^2 \end{bmatrix} \right), \tag{4.11}$$

with μ, σ^2 and τ denoting respectively the prior mean, variance and covariance of the encompassing prior, is also .50 $\forall p, q$, because $\pi \begin{pmatrix} a \\ b \end{pmatrix} = \pi \begin{pmatrix} b \\ a \end{pmatrix}$ $\forall$ a, b.

These results can be elaborated in two ways. Firstly, consider models with more constraints of the form (4.10). For example, let $\boldsymbol{\theta}^c = \{\theta_1, \theta_2, \theta_3, \theta_4\}$. The value of $1/c_2$ for model $M_2 : \theta_1 > \theta_2 > \theta_3 > \theta_4$ equals $1/4! = .04167$, independent of the actual choice of $\pi(\boldsymbol{\theta}^c)$. Secondly, the parameters can be organized in two groups, with as the model of interest $M_3 : \theta_1 > \theta_2 \wedge \theta_3 > \theta_4$. The value of $1/c_3$ equals $.50 \times .50 = .25$ independent of the prior distribution for each group of parameters. Note that, this holds for encompassing priors assuming prior independence as well as priors with homogeneous prior covariances.

Similar considerations hold for models using one or more constraints of the form

$$\prod_{p=1}^{P} \theta_p^{(1)} > \prod_{q=1}^{Q} \theta_q^{(2)}, \tag{4.12}$$

with the restriction $P = Q$. Constraints of the form (4.12) are for instance seen in models for contingency tables, where the θ's represent cell probabilities and constraints are imposed on odds ratios. An illustration will be provided in Section 4.7.

4.4.2 Classes of Models that do NOT Allow Objective Model Selection

In this paper a 'no effect'-model is represented by approximate equality constraints between the model parameters. The motivation is twofold. Firstly, the model 'θ_1 and θ_2 are *approximately* equal' is in our opinion more realistic than 'θ_1 and θ_2 are *exactly* equal'. To quote Berger and Delampady (1987) on this concept: 'It is rare, and perhaps impossible, to have a null hypothesis that can be exactly modeled as $\theta = \theta_0$' and '... surely there is *some* effect, although perhaps a very miniscule effect'. Secondly, from (4.9) it is immediately seen that using exact equality constraints, the Bayes factor for the encompassing with the equality constrained model is not defined, i.e. $1/c_t = 1/d_t = 0$.

Using approximate equality constraints, e.g.

$$M_2 : \theta_1 \approx \theta_2 \quad \Leftrightarrow \quad \mid \theta_1 - \theta_2 \mid < \epsilon, \tag{4.13}$$

with a small value for ϵ, the Bayes factor is defined. However, for this type of constraints the model selection based on encompassing priors is *not* objective. Like before $1/d_2$ is virtually independent of the encompassing prior. However, this does not hold for $1/c_2$. Let $\pi(\theta) \sim U[-z, z]$, than, $1/c_2 \to 0$ for $z \to \infty$. Stated otherwise, the larger z, the smaller the Bayes factor in favor of the unconstrained model. This is a phenomenon similar to Lindley's paradox. For this type of models the conclusions depend on the specification of the encompassing prior.

Another example of a nested model for which the Bayes factor is affected by the encompassing prior, is

$$M_3 : \theta_1 > \theta_2 + \epsilon, \tag{4.14}$$

where ϵ can be chosen such that θ_1 is relevantly larger than θ_2. If the encompassing setup is used, again $1/d_3$ is virtually independent of the encompassing prior if the latter is dominated by the data. However, the same does not hold for $1/c_3$. Again let $\pi(\theta) \sim U[-z, z]$, then $1/c_3 \to .5$ for $z \to \infty$. This is not as problematic as the effect seen in (4.13), but it does imply that for $z \to \infty$, $1/c_3$ is evaluated as if the model is $\theta_1 > \theta_2$ (i.e. the quantity ϵ does not affect $1/c_3$). Stated otherwise, the effect of ϵ depends on the vagueness of the prior and thus vague priors are not appropriate for this situation.

Finally, consider

$$M_4 : \theta_1 > 0. \tag{4.15}$$

Again, the value for $1/c_4$ is not independent of the encompassing prior. When $\pi(\theta) \sim U[-z, z]$, then $1/c_4 \to .5$ for $z \to \infty$.

4.5 Examples of Encompassing and Nested Models

Consider the encompassing model

$$y_i = \sum_{k=1}^{4} \mu_k d_{ki} + \varepsilon_i, \qquad \text{with} \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2). \tag{4.16}$$

with $i = 1, \ldots, N$. In the encompassing model four groups are distinguished. Group membership is denoted by d_{ki}, where $d_{ki} = 1$ if the i-th person is a member of group

k, and zero otherwise. Nested models are formed by imposing constraints on the means of the subgroups (μ_k), so $\boldsymbol{\theta}^c = \{\mu_1, \mu_2, \mu_3, \mu_4\}$ and $\boldsymbol{\theta}^u = \{\sigma^2\}$. Consider for example the following set of models

$$\begin{array}{lll} M_1: & \mu_1,\ \mu_2,\ \mu_3,\ \mu_4, & \sigma^2, \\ M_2: & \mu_1 < \mu_2 < \mu_3 < \mu_4, & \sigma^2, \\ M_3: & (\mu_1 + \mu_2) < (\mu_3 + \mu_4), & \sigma^2, \\ M_4: & \mu_1 \approx \mu_2 \approx \mu_3 \approx \mu_4, & \sigma^2. \end{array} \tag{4.17}$$

Consider another example, where the relation between a continuous predictor ($\boldsymbol{x}$) and a normally distributed criterion variable ($\boldsymbol{y}$) is examined for two independent groups. The encompassing model is

$$y_i = \sum_{k=1,2} \alpha_k d_{ki} + \sum_{k=1,2} \beta_k x_i d_{ki} + \varepsilon_i, \qquad \text{with} \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \tag{4.18}$$

where $d_{ki} = 1$ if the i-th person is a member of group k ($k = 1, 2$), and zero otherwise. When $\boldsymbol{x}$ is centered around zero, the intercepts α_1 and α_2 can be interpreted as the average performance of group 1 and 2 respectively. The slopes, β_1 and β_2 represent the relation between $\boldsymbol{x}$ and $\boldsymbol{y}$ in the two groups. Theories about group effects on slopes and/or intercepts can be translated into constrained models, like:

$$\begin{array}{lll} M_1: & \alpha_1,\ \alpha_2, & \beta_1,\ \beta_2, \sigma^2, \\ M_2: & \alpha_1 < \alpha_2, & \beta_1 < \beta_2, \sigma^2, \\ M_3: & \alpha_1 < \alpha_2, & \beta_1 > \beta_2, \sigma^2, \\ M_4: & \alpha_1 \approx \alpha_2, & \beta_1 \approx \beta_2, \sigma^2. \end{array} \tag{4.19}$$

In this example the constrained parameters fall apart in two groups. Since the constraints of M_2 and M_3 are of the form (4.10), the selection will be virtually objective, as long as the prior is dominated by the data. However, the results for M_4, with constraints of the form (4.13) will depend strongly on the specification of the encompassing prior.

Note that, the models in both examples, (4.16) and (4.18), are based on a non-standard dummy coding. This parameterization is chosen because it leads to constraints of the form (4.10), thus allowing for model selection that is independent of the specification of the encompassing prior. For example, the more common parameterization for (4.18) is

$$y_i = \alpha_1 + \alpha_2 d_i + \beta_1 x_i + \beta_2 x_i d_i + \varepsilon_i, \tag{4.20}$$

with $d_i = 1$ if the i-th person is a member of group two, and zero otherwise. The interpretation of the parameters has changed. Consequently the models representing the same theories as in (4.19) also change, that is

$$\begin{array}{lll} M_1: & \alpha_1,\ \alpha_2, & \beta_1,\ \beta_2, \sigma^2, \\ M_2: & \alpha_2 > 0, & \beta_2 > 0, \sigma^2, \\ M_3: & \alpha_2 > 0, & \beta_2 < 0, \sigma^2, \\ M_4: & \alpha_2 \approx 0, & \beta_2 \approx 0, \sigma^2. \end{array} \tag{4.21}$$

Both M_2 and M_3 are of the form (4.15), i.e. the model selection is not independent of the specification of the encompassing prior. Moreover, none of the nested models in (4.21) is of the form (4.10) or (4.12), and thus virtually objective model selection is not possible.

Choosing a convenient parameterization does not change the likelihood but can enable easier calculation, interpretation, or in this case specification of prior distributions (see also Gelman, 2004). Two more examples are given below.

Consider the simple linear regression model $y_i = \alpha + \beta x_i + \varepsilon_i$, with $i = 1, \ldots, N$ and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, and the models of interest:

$$\begin{array}{lll} M_1: & \beta\ , & \alpha,\ \sigma^2, \\ M_2: & \beta > 0, & \alpha,\ \sigma^2, \\ M_3: & \beta \approx 0, & \alpha,\ \sigma^2, \end{array} \tag{4.22}$$

i.e. $\boldsymbol{\theta}^c = \{\beta\}$ and $\boldsymbol{\theta}^u = \{\alpha, \sigma^2\}$. Specification of prior distributions for a set of models like (4.22) is not straightforward. Note that M_2 is again of the form (4.15). For both BF_{21} and BF_{31}, the results will depend on the specification of the prior distribution for β. Reparameterization however, leads to easier specification of prior distributions for the transformed parameters.

Denote $y_{(x=1)} = \gamma$ and $y_{(x=-1)} = \lambda$. The regression model becomes $y_i = \frac{1}{2}(\gamma + \lambda) + \frac{1}{2}(\gamma - \lambda)x_i + \varepsilon_i$. Instead of (4.22), the parameters γ and λ can be used to define the models:

$$\begin{array}{lll} M_1: & \gamma\ , & \lambda,\ \sigma^2, \\ M_2: & \gamma > \lambda, & \sigma^2, \\ M_3: & \gamma \approx \lambda, & \sigma^2. \end{array} \tag{4.23}$$

Note that M_2 in (4.23) is of the form (4.10) and therefore the model selection is independent of the specification of the encompassing prior. Note also, that M_3 in (4.23) is of the form (4.13) and therefore the model selection does depend on the specification of the prior.

Table 4.1: Weight Gain of Rats

	Sample size	Mean	Standard deviation
Group 1 (beef, high)	10	100.0	15.1
Group 2 (beef, low)	10	79.2	13.9
Group 3 (cereal, high)	10	85.9	15.0
Group 4 (cereal, low)	10	83.9	15.7

The final example concerns an experiment with a pretest, an intervention, and a posttest. Each subject has two scores $y_{pre,i}$ and $y_{post,i}$ and the model parameters are μ_{pre}, μ_{post} and the covariance matrix Σ. To compare the theories 'no effect of intervention' and 'positive effect of intervention' two approaches can be taken. The usual approach would be to compute an effect score for each person ($d_i = y_{post,i} - y_{pre,i}$) and formulate the models in terms of a gain δ. Consider the three models $M_1 : \delta,\ \sigma^2$, $M_2 : \delta > 0,\ \sigma^2$, and, $M_3 : \delta \approx 0,\ \sigma^2$. Again, an objective prior can not easily be found and BF_{21} and BF_{31} will depend on the specification of the encompassing prior. In this example, using the original variables μ_{pre} and μ_{post} leads to the models: $M_1 : \mu_{pre},\ \mu_{post},\ \Sigma$; $M_2 : \mu_{pre} < \mu_{post},\ \Sigma$; and, $M_3 : \mu_{pre} \approx \mu_{post},\ \Sigma$. Again, in this set-up the model selection based on encompassing priors is virtually objective for model M_2, but not for model M_3.

4.6 Example: Analysis of Variance

4.6.1 The Data

The data come from an experiment to study the gain in weight of 40 rats fed on different diets, distinguished by both source and amount of proteins (Snedecor and Cochran, 1967). The rats are randomly assigned to one of four groups receiving a high protein beef diet (group 1), a low protein beef diet (group 2), a high protein cereal diet (group 3), a low protein cereal diet (group 4). Gain in weight, y_i, (measured in grams) is assumed to be normally distributed:

$$y_i = \sum_{k=1}^{4} \mu_k d_{ki} + \varepsilon_i, \qquad \text{with} \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2). \tag{4.24}$$

Group membership is denoted by d_{ki}, where $d_{ki} = 1$ if the i-th rat is a member of group k, and zero otherwise. The data are summarized in Table 4.1.

Three nested models are compared with the encompassing model (M_1). In the first theory (M_2), a large effect of amount of protein is expected (more weight gain for higher amounts), followed by a smaller effect of the source (more gain for beef diets). The second theory (M_3) describes the same directions of effects but expects a larger effect of source than of amount of proteins. Finally, a 'no-effect' (or 'approximately equal means') model is defined using the restriction that the difference between the smallest and highest mean gain in weight is less than 5 grams (M_4).

$$\begin{array}{ll} M_1: & \mu_1, \mu_2, \mu_3, \mu_4,\ \sigma^2, \\ M_2: & \mu_1 > \mu_3 > \mu_2 > \mu_4,\ \sigma^2, \\ M_3: & \mu_1 > \mu_2 > \mu_3 > \mu_4,\ \sigma^2, \\ M_3: & \mu_1 \approx \mu_2 \approx \mu_3 \approx \mu_4,\ \sigma^2. \end{array} \tag{4.25}$$

The encompassing prior that will be used in Section 4.6.2, is the product of independent conjugate prior distributions for each of the parameters, specified using the properties of Section 4.2, i.e.

$$\pi(\boldsymbol{\mu}, \sigma^2|M_1) = \text{Inv-}\chi^2(\sigma^2|1;\ 236.4) \times \prod_{k=1}^{4} N(\mu_k|89.6;\ 123.8). \tag{4.26}$$

The specification of the hyperparameters of $\pi(\mu_k|M_1)$ is based on the 99% credibility intervals of each element μ_k. The smallest lower bound and the largest upper bound of the four intervals provide one interval $l - u$. The prior distribution for each μ_k is chosen such that the mean of the prior minus two standard deviations equals l and the mean plus two standard deviations equals u. The scaled Inv-χ^2 distribution, i.e. $\pi(\sigma^2|M_1)$, has one degree of freedom, which can be thought of as providing the information equivalent to one observation, and a data based estimation of the error variance. Note, that any other method to specify a reasonable and diffuse encompassing prior can also be used, e.g. Jeffreys priors (Jeffreys, 1961), or Zellner-Siow priors (Zellner and Siow, 1980). For specific models (in this example M_2 and M_3) any encompassing prior will obtain the same results as long as the data dominate the prior, that is, the results are independent of the specification of the encompassing prior. For other models, the specification of the encompassing prior does influence the results.

The Bayes factors comparing each nested model with M_1 using the encompassing prior in (4.26), are presented in Section 4.6.2. Sensitivity of Bayes factors to different encompassing priors is examined and discussed in Section 4.6.3.

4.6.2 Estimation and Evaluation of Bayes Factors

As can be seen in (4.9), the Bayes factor for a constrained model with the encompassing model reduces to the ratio c_t/d_t. The proportion of the encompassing *prior* that is in agreement with the constraints of model M_t provides the value $1/c_t$, the proportion of the encompassing *posterior* in agreement with the constraints of M_t is $1/d_t$. As we showed in Section 4.4, for many models $1/c_t$ can be computed exactly. In other cases the encompassing prior is sampled and $1/c_t$ is estimated by the proportion of the sample that is in agreement with the constraints of model M_t. To obtain $1/d_t$, the encompassing posterior is sampled using a Gibbs sampler (see for instance Smith and Roberts, 1993) and the proportion of that sample in agreement with M_t provides an estimate for $1/d_t$. Note that, with one posterior sample and, if $1/c_t$ can not be obtained exactly one prior sample, the values for $\hat{BF}_{t1} = \hat{c}_t/\hat{d}_t$ for *each* nested model ($t = 2, \ldots, T$) with the encompassing model, can be computed.

For models where $1/c_t$ is computed exactly, the standard error (SE) of the estimated Bayes factor, is computed using

$$\mathrm{SE}^2(\hat{BF}_{t1}) = \mathrm{VAR}\left(\frac{1/\hat{d}_t}{1/c_t}\right) = c_t^2\ \mathrm{VAR}(1/\hat{d}_t) \approx c_t^2\ \left(\frac{Q_t}{Q}(1-\frac{Q_t}{Q})Q^{-1}\right), \quad (4.27)$$

where Q is the total number of parameter vectors sampled from the posterior, and Q_t is the number of parameter vectors in agreement with the constraints of model M_t.

When $1/c_t$ is estimated using a sample from the prior and $1/\hat{c}_t$ and $1/\hat{d}_t$ are assumed to be independent, the delta method gives an approximation of the variance of the ratio of $1/\hat{d}_t$ and $1/\hat{c}_t$ (Johnson, Kotz and Kemp, 1993, p. 55):

$$\begin{aligned}\mathrm{VAR}\left(\frac{1/\hat{d}_t}{1/\hat{c}_t}\right) &\approx \left(\frac{\mathrm{E}(1/\hat{d}_t)}{\mathrm{E}(1/\hat{c}_t)}\right)^2 \left(\frac{\mathrm{VAR}(1/\hat{d}_t)}{[\mathrm{E}(1/\hat{d}_t)]^2} + \frac{\mathrm{VAR}(1/\hat{c}_t)}{[\mathrm{E}(1/\hat{c}_t)]^2}\right) \\ &\approx \left(\frac{Q_t/Q}{R_t/R}\right)^2 \left(\frac{\frac{Q_t}{Q}(1-\frac{Q_t}{Q})Q^{-1}}{[Q_t/Q]^2} + \frac{\frac{R_t}{R}(1-\frac{R_t}{R})R^{-1}}{[R_t/R]^2}\right), \end{aligned} \quad (4.28)$$

where R is the total number of parameter vectors sampled from the prior, and R_t is the number of parameter vectors in agreement with the constraints of model M_t.

The Bayes factors and standard errors (SE) for the set of models specified in (4.25) are provided in Table 4.2. For both BF_{21} and BF_{31} the value for $1/c_t$ is easily calculated, i.e. $1/c_2 = 1/c_3 = 1/4! = .04167$. Although the value for $1/c_4$

Table 4.2: Evaluation of the Encompassing Prior Approach

BF_{21}	(SE)	BF_{31}	(SE)	BF_{41}	(SE)
3.70	(.09)	1.24	(.05)	.33	(.06)

can also be obtained analytically, here it is estimated based on a sample from the prior. The number of iterations is large ($R = 1,000,000$), because the proportion of the encompassing prior in agreement with approximate equality constraints is often very small. Note however, that deriving the value for $1/c_4$ is still very fast, since no data and no complex calculations are involved. The estimates of $1/d_t$ for all models are based on one sample of $Q = 10,000$ iterations (after an initial burn-in period of 5000 iterations). As can be seen, the standard errors for the Bayes factors are small. It can be concluded that the numbers of iterations are large enough to give stable estimates.

4.6.3 Sensitivity to Encompassing Priors

The results for several specifications of the encompassing prior are compared in Table 4.3. The first three rows give the results of independent conjugate prior distributions for all model parameters, with different values for the hyperparameters. The fourth and fifth row provide the results of independent bounded uniform prior distributions for all model parameters, with differing bounds.

The number of iterations from prior (for BF_{41}) and posterior are respectively $R = 1,000,000$ and $Q = 10,000$. It can be seen that, as expected, the values for BF_{21} and BF_{31} are virtually independent of the specification of the encompassing prior. However, the value for BF_{41} is strongly affected by the prior. This is reflected in different values for $1/c_4$ for the different encompassing priors. For instance, for the prior on the first row of Table 4.3, $1/c_4 = .01$, whereas for the last prior $1/c_4 = .00002$. The fit of model M_4 can only be interpreted conditional on the specification of the encompassing prior.

The standard errors are mainly small and for the models where only $1/d_t$ is estimated, similar for different encompassing priors. However, for BF_{41} where both $1/d_4$ and $1/c_4$ are estimated the standard errors become larger for more diffuse encompassing priors. The standard errors can be reduced by taking a larger sample from the encompassing prior (apparently and not surprisingly, $R = 1,000,000$ is not a lot when priors as diffuse as $N(50, 2000)$ and $U(0, 300)$ are explored).

Table 4.3: Prior Sensitivity of Bayes Factors

$\pi(\mu_k)$	$\pi(\sigma^2)$	BF_{21}	(SE)	BF_{31}	(SE)	BF_{41}	(SE)
$N(90, 125)$	Inv-$\chi^2(1, 250)$	3.73	(.09)	1.30	(.05)	0.33	(.05)
$N(0, 500)$	Inv-$\chi^2(1, 250)$	3.70	(.09)	1.27	(.05)	1.54	(.33)
$N(50, 2000)$	Inv-$\chi^2(1, 250)$	3.65	(.09)	1.17	(.05)	8.18	(2.14)
$U(60, 120)$	$U(0, 900)$	3.68	(.09)	1.21	(.05)	0.73	(.18)
$U(0, 300)$	$U(0, 900)$	3.69	(.09)	1.20	(.05)	79.81	(27.3)

4.7 Example: Contingency Tables with Ordered Odds Ratios

4.7.1 The Data

The data (obtained from Agresti, 1984) presented in Table 4.4 are an example of a $2 \times 2 \times 2$ contingency table containing counts of death penalties (p) for each of the four combinations of defendant's race (d) and victim's race (v). The 326 subjects cross-classified according to these variables were defendants in homicide indictments in 20 Florida counties during 1976-1977.

Analysis of these data ($\boldsymbol{x}$) is based on a multinomial likelihood

$$L(\boldsymbol{x} \mid \boldsymbol{\theta}) \propto \prod_{d=0,1} \prod_{v=0,1} \prod_{p=0,1} \theta_{dvp}^{x_{dvp}}, \tag{4.29}$$

where 0 denotes 'White' or 'Yes' and 1 denotes 'Black' or 'No'.

Four models are compared. The first model is the encompassing model (M_1). The second model represents the expectation that, both for white and black victims, the death penalty is imposed more often when the defendant is black than when the defendant is white. For white victims this expectation is represented by the restriction that the odds ratio $\frac{\theta_{000}\theta_{101}}{\theta_{100}\theta_{001}}$ should be smaller than one. Likewise, for the black victims the constraint is: $\frac{\theta_{010}\theta_{111}}{\theta_{110}\theta_{011}} < 1$. In M_2, these constraints are written in the form (4.12). The third model represents the expectation that, both for white and black defendants, the death penalty is imposed more often when the victim is white than when the victim is black (M_3). The final model is a combination of both

Table 4.4: Death Penalty Verdict by Defendant's Race and Victim's Race

Defendant's Race	Victim's Race	Death Penalty	
		Yes	No
White	White	19	132
	Black	0	9
Black	White	11	52
	Black	6	97

expectations (M_4).

$$
\begin{aligned}
M_1 &: \theta_{000}, \theta_{001}, \theta_{010}, \theta_{011}, \theta_{100}, \theta_{101}, \theta_{110}, \theta_{111}, \\
M_2 &: \theta_{000}\theta_{101} < \theta_{100}\theta_{001} \wedge \theta_{010}\theta_{111} < \theta_{110}\theta_{011}, \\
M_3 &: \theta_{000}\theta_{011} > \theta_{010}\theta_{001} \wedge \theta_{100}\theta_{111} > \theta_{110}\theta_{101}, \\
M_4 &: \theta_{000}\theta_{101} < \theta_{100}\theta_{001} \wedge \theta_{010}\theta_{111} < \theta_{110}\theta_{011} \wedge \\
&\quad \theta_{000}\theta_{011} > \theta_{010}\theta_{001} \wedge \theta_{100}\theta_{111} > \theta_{110}\theta_{101}.
\end{aligned}
$$

Agresti and Coull (2002) present an overview of methods for the analysis of contingency tables under inequality constraints. Apparently, there are no classical alternatives for the Bayesian analysis executed in the sequel. Since all nested models in the set above are of the form (4.12), the Bayesian model selection is not sensitive to the encompassing prior, if the prior is dominated by the data. This will be illustrated in the next subsection.

4.7.2 Sensitivity to Encompassing Priors

The encompassing prior used, is the conjugate Dirichlet distribution

$$g(\boldsymbol{\theta}) \propto \prod_{d=0,1} \prod_{v=0,1} \prod_{p=0,1} \theta_{dvp}^{x_0-1}$$

where x_0 denotes the parameter of the prior distribution. An uniform density is obtained by $x_0 = 1$; the distribution assigns equal density to any vector $\boldsymbol{\theta}$ satisfying $\sum_{d,v,p} \theta_{dvp} = 1$ (Gelman, et al. 1995, p. 76). However, the amount of information in

Table 4.5: Prior Sensitivity of Bayes Factors

x_0	BF_{21} (SE)	BF_{31} (SE)	BF_{41} (SE)	$1/c_4$	$1/d_4$
1	1.62 (.02)	2.94 (.006)	4.81 (.05)	.083	.401
.5	2.34 (.006)	3.49 (.004)	7.05 (.08)	.082	.581
.1	3.05 (.005)	3.88 (.002)	9.30 (.10)	.082	.761
.01	3.22 (.005)	3.96 (.001)	9.68 (.10)	.083	.805

the data (note that one cell has zero observations) does not dominate the information in the prior. Consequently, we expect that the model selection will to some extent be affected by the encompassing prior. The results are compared with three other encompassing priors, by specifying the values .5, .1 and .01 for x_0.

It is easily seen that $1/c_2 = 1/c_3 = .25$. The value for $1/c_4$ is estimated based on $R = 100,000$ iterations from the encompassing prior. The values for $1/d_t$, for $t = 2, 3, 4$ are obtained by $Q = 100,000$ iterations from the encompassing posterior. In Table 4.5 the values of the Bayes factors and corresponding standard errors (SE) are provided, for each of the four encompassing priors. The last two columns provide the values for $1/c_4$ and $1/d_4$.

The resulting Bayes factors differ for the four encompassing priors, although the order in which the four models are supported is similar (i.e. M_4, M_3, M_2, M_1). For the encompassing prior with $x_0 = 1$, the Bayes factors are closer to one than for priors with smaller values for x_0. In this example where some cells have small observed frequencies, and more specifically where one cell has zero observations, the data do not dominate the prior, and therefore the value for $1/d_t$ varies for the different priors. This is illustrated in the last two columns of Table 4.5, where the estimates for $1/c_4$ and $1/d_4$ are provided. As can be seen $1/c_4$ is about equal for the four priors, but the estimates for $1/d_4$ are increasing with a decreasing amount of information in the encompassing prior.

4.8 Discussion

In this paper the use of encompassing priors in Bayesian model selection is introduced and examined. It can be used when the set of models of interest consists of models that are all nested in one encompassing model. The approach introduced in this paper has several advantages. Firstly, only a prior distribution for the parame-

ters of the encompassing model is specified. The priors for the constrained (nested) models follow from this so-called encompassing prior by restriction of the parameter space. Secondly, the Bayes factor of any nested model with the encompassing model has a nice interpretation, that is, it is the ratio of the proportions of the encompassing posterior and prior respectively, that are in agreement with the constraints of the nested model (respectively $1/d_t$ and $1/c_t$). These proportions can be estimated using samples from encompassing prior and posterior. So, the computation of Bayes factors does not include the estimation of marginal likelihoods, which can be burdensome. Furthermore, with our approach with just one prior and one posterior sample the Bayes factors for all pairs of models are obtained. It is also shown that for some classes of models the value for $1/c_t$ can be derived exactly, making the estimation even more efficient. Since the estimation of Bayes factors reduces to the estimation of one or two proportions, standard errors are easily obtained.

Applying some general properties (described in Section 4.2) for the specification of encompassing priors, allows virtually objective model selection for specific classes of models, that is, the resulting Bayes factors are (almost) independent of the actual specification of the encompassing prior. For other classes of models, the model selection is strongly affected by the encompassing prior. An example is a model with approximate equality constraints between model parameters. The evidence in favor of the constrained model increases with the amount of vagueness of the encompassing prior, a phenomenon known as Bartlett's or Lindley's paradox. Our approach does not solve this well-known problem, but it does give some insight in the process lying underneath. For this kind of models, a subjective choice for the specification of the encompassing prior must be made. The resulting Bayes factor can only be interpreted conditional on this choice.

Chapter 5

Comparison of Hypothesis Testing and Bayesian Model Selection

Abstract

The main goal of both Bayesian model selection and classical hypotheses testing is to make inferences with respect to the state of affairs in a population of interest. The main differences between both approaches are the explicit use of prior information by Bayesians, and the explicit use of null distributions by the classicists. Formalization of prior information in prior distributions is often difficult. In this paper two practical approaches (encompassing priors and training data) to specify prior distributions will be presented. The computation of null distributions is relatively easy. However, as will be illustrated, a straightforward interpretation of the resulting p-values is not always easy.

Bayesian model selection can be used to compute posterior probabilities for each of a number of competing models. This provides an alternative for the currently prevalent testing of hypotheses using p-values. Both approaches will be compared and illustrated using case studies. Each case study fits in the framework of the normal linear model, that is, analysis of variance and multiple regression.

5.1 An Introduction to Hypothesis Testing and Bayesian Model Selection

Testing of hypotheses using p-values (see, Bayarri and Berger, 2000, for a comprehensive overview) is a statistical tool that is used quite often in psychological research. The p-value can formally be defined as:

$$P(T(\boldsymbol{Y}) \geq t(\boldsymbol{y}) \mid H_0), \tag{5.1}$$

where $\boldsymbol{y}$ denotes the observed data, and $t(\boldsymbol{y})$ a test statistic (e.g. the student t-test or Pearson chi-square statistic) computed for the observed data. Furthermore, $\boldsymbol{Y}$ denotes a data matrix from the null-population described by the null-hypothesis H_0, and $T(\boldsymbol{Y})$ the test statistic computed for this data matrix. The probability in (5.1) is computed over the distribution of $T(\boldsymbol{Y})$ under H_0. This implies that the p-value represents the proportion of $T(\boldsymbol{Y})$ resulting from H_0 that is larger than $t(\boldsymbol{y})$. According to a popular rule, a p-value smaller than .05 is a strong indication that H_0 should be rejected. The motivation for the latter (although the choice of the reference value .05 is arbitrary) is that small p-values imply that $t(\boldsymbol{y})$ is rarely observed if H_0 is true. Usually, the specification of the alternative hypothesis is rather vague, that is, not H_0. However, as will be illustrated later, H_0 can also be tested against a specific H_1. The latter is accomplished using a test statistic that has power against the alternative hypothesis of interest.

Developments in Bayesian statistics (see, Kass and Raftery, 1995, for a comprehensive overview and references) have rendered a framework for model selection that can be used as an alternative for the testing of hypotheses. Let M_m for $m \in \{0, 1\}$ denote the null and alternative model, respectively. Bayes theorem states that

$$P(M_m \mid \boldsymbol{y}) = \frac{P(\boldsymbol{y} \mid M_m)P(M_m)}{P(\boldsymbol{y})}, \tag{5.2}$$

where,

$$P(\boldsymbol{y}) = \sum_m P(\boldsymbol{y} \mid M_m)P(M_m), \tag{5.3}$$

$P(M_m \mid \boldsymbol{y})$ is the posterior probability, that is after observing the data, of model m, $P(M_m)$ is the prior probability, that is before observing the data, of model m, and $P(\boldsymbol{y} \mid M_m)$ is the density of the data for model M_m. Two models with (about) the same posterior probability, are (after observing the data) about equally likely. However, if one model has a posterior probability of .90 and the other of .10, the first is preferred to the latter.

The posterior probabilities of the null and alternative model constitute an alternative for the p-value resulting from testing a hypothesis. A difference between p-values and posterior probabilities is that the latter are computed for all models under consideration, and not only the null-model. This leads to a straightforward interpretation of posterior probabilities. 'Level alpha' and 'level beta' are classical terms used to denote the probability of an error of the first and second kind, respectively, that is, the error probabilities *before* observing the data. Posterior probabilities can be seen as conditional error probabilities (Cohen, 1994; Sellke, Bayarri and Berger, 2001), that is, *after* observing the data. The conditional error of the first kind is the probability that M_0 is incorrectly rejected after observing the data. The latter is equal to the probability of M_0 after observing the data. Similarly $P(M_1 \mid \boldsymbol{y})$ can be interpreted as the conditional error of the second kind.

As will be illustrated in the sequel, the testing of hypotheses using p-values has its limitations. The goal of this paper is to introduce Bayesian model selection, to compare it to hypothesis testing, and to give illustrations of its use in psychological research. In the next section two simple hypotheses and models are used to provide a first comparison of hypothesis testing and model selection. In the two sections that follow, both approaches will be illustrated and compared in the context of one-way analysis of variance followed by pairwise comparisons of means, and, selection of the best of a number of non-nested regression models. The paper will be concluded with a discussion.

5.2 A First Comparison of Hypothesis Testing and Bayesian Model Selection

Consider the data $\boldsymbol{y} = [0, .5, 1.5, 2]$ i.e. the sample size N is four. Assume that the data come from a normal population with variance $\sigma^2 = 1$ and unknown mean μ. The common null hypothesis is $H_0 : \mu_0 = 0$ the corresponding alternative hypothesis is $H_1 : \mu_1 \neq 0$. The null and alternative models are $M_0 : \mu_0 = 0$ and $M_1 : \mu_1 \neq 0$. The null hypothesis can be tested using the Z-test with $t(\boldsymbol{y}) = \frac{\overline{y}-\mu_0}{\sigma/\sqrt{N}}$, where $\overline{y}$ denotes the sample average of $\boldsymbol{y}$. The resulting (two-sided) p-value is .048, and indicates that the observed value of the test statistic is not very common if H_0 is true. Frick (1996) and Wainer (1999) present examples where it is sufficient to determine whether data are consistent with H_0 or not, and, where the specification of H_1 beyond 'not H_0' is irrelevant. If this is all a researcher wants to know, p-values are an excellent tool to provide answers to the research questions. In fact, in this situation posterior probabilities are *not* an alternative for the use of p-values, because specification of H_1 beyond 'not H_0' is necessary. Stated otherwise,

the (prior) distribution of μ_1 under H_1 has to be specified in order to be able to compute posterior probabilities (this topic will return in Sections 5.3 and 5.4).

Often researchers are not satisfied with the p-value based conclusion that the observed data are not in agreement with the null-population. They also want to know which of the many alternatives for which $\mu_1 \neq 0$ are in the ball park. Implicitly, they want to know whether the distance between μ_1 and μ_0 is so large that it is relevant, that is, interesting from a practical or scientific point of view. Since p-values only indicate whether data are likely given H_0 or not (an illustration follows below), effect size measures are usually used as a tool to translate the p-value based conclusion 'not H_0' into the research conclusion *relevant* H_1. This can be seen as an informal (it involves a subjective evaluation of the effect size) attempt to quantify the evidence that the data come from a relevant H_1. In Section 5.3.4 it will be shown that the set of models can contain formal counterparts of the informal *relevant* H_1.

The following situation illustrates that p-values only indicate whether or not data are likely given H_0, and that posterior probabilities consider both M_0 and M_1. The data and the null model and hypothesis are the same as before. However, $H_1 : \mu_1 = 2$ and $M_1 : \mu_1 = 2$. A test statistic that has power against the alternative hypothesis is the likelihood ratio statistic

$$t(\boldsymbol{y}) = -2 \log \frac{P(\boldsymbol{y} \mid H_0)}{P(\boldsymbol{y} \mid H_1)}, \tag{5.4}$$

where $P(\boldsymbol{y} \mid H_m) = P(\boldsymbol{y} \mid M_m)$ as given in (5.5). Under H_0 $T(\boldsymbol{Y})$ has a $N(-16, 64)$ null distribution. For the data at hand $t(\boldsymbol{y}) = 0$, the corresponding one-sided p-value (5.1) is .024. The choice for a one-sided test is motivated by the specification of the alternative hypothesis: it would be awkward to choose the alternative if $\overline{y}$ is, for example, -2, which in a two-sided setup would lead to a small p-value.

Using (5.2) the posterior model probabilities can be computed. For the setup at hand

$$P(\boldsymbol{y} \mid M_m) = \prod_{i=1}^{4} \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}(y_i - \mu_m)^2. \tag{5.5}$$

The necessity to specify prior model probabilities is a step typical for Bayesian model selection. One option is to use a noninformative prior: $P(M_1) = P(M_2) = .5$. The implication is that differences in posterior model probabilities are caused by differences in the degree to which the data support the model, that is, $P(\boldsymbol{y} \mid M_m)$ for $m \in \{0, 1\}$. The resulting posterior probabilities are .5 for both the null and the alternative model i.e. after observing the data both models are equally likely.

The conclusions obtained using p-values and posterior probabilities are strikingly different. Loosely speaking the first rejects H_0, while the latter does not prefer M_0 over M_1. Since $\overline{y} = 1$ is nicely centered between the population means of the null and alternative hypothesis, the conclusion based on the posterior probabilities is correct. This simple example illustrates a drawback of p-values: the size of a p-value may point to H_1, not because the data are likely given H_1, but because the data are unlikely given H_0. This pitfall can to some extend be avoided if the evaluation of a p-value includes an evaluation of the effect size (here $\overline{y} = 1$) in relation to what is specified by H_0 and H_1. However, although the resulting approach is useful and valuable, it is not longer a formal testing procedure since it includes a subjective evaluation of the effect size. Posterior probabilities are computed for each model in the set of models under consideration. Using posterior probabilities the pitfall that 'not H_0' does not necessarily imply H_1 is avoided.

The computation of posterior probabilities is not completely unproblematic. The bottle-neck is the specification of prior distributions (for each model) for the parameters that are not fixed at a specific value. The main criticism is that these specifications are subjective, researcher dependent, and thus, that the conclusions are subjective and researcher dependent (Sober, 2002, pp. 22-24; Howson, 2002, pp. 53-61). In the next two sections two approaches to obtain prior distributions that are not based on (subjective) prior knowledge will be outlined. The first approach will be illustrated using a one-way analysis of variance followed by pairwise comparisons of means. Encompassing priors that are noninformative with respect to the untransformed parameters of the one-way model will be formulated. The second approach will be illustrated using the selection of the best of a number of non-nested regression models. Here priors will be formulated using training data. Using these examples it will be shown that Bayesian model selection is a viable alternative for hypotheses testing.

5.3 Encompassing priors

5.3.1 Analysis of Variance: Data, Hypotheses and P-values

One way analysis of variance is often used in psychological research. A common design is the situation where a researcher has a control group (C), and two experimental groups (E1 and E2) that differ in the 'treatment' that has been received. If the persons participating in the experiment have been randomized over the three groups, the dependent variable can be the outcome of a test given to the persons after the treatment, or, in a design with a pre-test and a post-test, the gain-score.

In this section a simple one way analysis of variance will be executed. Hypotheses

Table 5.1: The One Way Analysis of Variance Data

Data	C	E1	E2
	3	3	4
	2	5	3
	1	4	5
	4	3	5
	3	5	3
	2	3	3
	3	4	4
	4	3	4
	4	3	4
	2	5	4
	2	5	5
	3	4	2
	1	1	3
	2	3	1
Mean	2.57	3.64	3.57
SD	1.02	1.15	1.15

will be formulated and tested, and the resulting p-values will be discussed. In the next section, models will be formulated, and posterior probabilities computed and discussed. The data that will be analyzed come from Toothaker (1993, p. 3). He describes an experiment from developmental psychology: "A total of 42 first-grade students were randomly assigned to one of three groups, n=14 per group. Subjects in the first group were informed that there was another child alone in the next room who had been warned not to climb on a chair. This group was called indirect responsibility (C). In addition to the story told to the subjects in (C), subjects in the second group were informed that when the adult left, they were in charge and to take care of anything that happened. This group was called direct responsibility one (E1). Subjects started a simple task, and the adult left the room. Next, there was a loud crash from the next room and a short time of crying and sobbing. In the third group, direct responsibility two (E2), subjects had the same instructions as (E1), but there were also calls for help. From behind a one-way mirror, two raters gave ratings on helping behavior. The scale was from 1 (no help) to 5 (went to the next room)." The data from this experiment and per group the sample mean and standard deviation are presented in Table 5.1.

Hypotheses testing usually consists of two steps. In the first step $H_0 : \mu_C = \mu_{E1} = \mu_{E2}$ is tested. In the second step three pairwise comparisons are tested: $H_{0a} : \mu_C = \mu_{E1}$, $H_{0b} : \mu_C = \mu_{E2}$ and $H_{0c} : \mu_{E1} = \mu_{E2}$. Since in the second step three hypotheses are tested, the risk of an error of the first kind (incorrectly rejecting one or more of the pairwise null hypotheses) is larger than level alpha. A wide variety of procedures has been developed to control the 'family wise error rate' that is, the probability of one or more errors of the first kind. The interested reader is referred to Toothaker (1993) and Ramsey (2002) for a comprehensive overview. Here the Hayter modification of Fisher's LSD procedure (Ramsey, 2002) will be used. This is a protected procedure i.e. the pairwise comparisons are only tested if H_0 is significant.

The p-value of H_0 (obtained using a F-test for the equality of independent means) is smaller than the commonly used alpha-level of .05 (see, Table 5.2). Stated otherwise, it is unlikely that the data result from a population described by H_0. As illustrated in Section 5.2, the conclusion 'not H_0' does not necessarily imply H_1 or '*relevant* H_1'. However, inspection of the means in Table 5.1 shows that μ_C is quite (subjective evaluation of the authors) different from the other means (about 1 on a scale running from 1 to 5). For the example at hand it seems save to conclude that 'not H_0' implies a '*relevant* H_1'.

Usually, a significant H_0 is followed by pairwise comparisons of the means. The p-values obtained using the Hayter modification of Fisher's LSD procedure are displayed in Table 5.2. As can be seen the mean of the control group is significantly different from the means of both experimental groups. The means of the experimental groups are not significantly different. It seems save to conclude that the mean of the control group is smaller (both significantly and relevantly) than the means of both experimental groups.

Pairwise comparisons among means may lead to inconsistent results (Dayton, 2003). Consider, for example, the situation where H_{0a}, H_{0b} and H_{0c}, have p-values of .045, .055 and .70, respectively. This results in a set of conclusions that can not all be true: $\mu_C \neq \mu_{E1}$, $\mu_C = \mu_{E2}$ and $\mu_{E1} = \mu_{E2}$. There are two drawbacks to the use of p-values (that, as will be illustrated in the next section, are not shared by posterior probabilities) that make it hard to properly deal with inconsistencies: (i) testing non-nested hypotheses is an underdeveloped area of statistics; and, (ii) the use of a fixed alpha-level.

To elaborate on (i). Suppose, that the overall F-test indicates that H_0 is not true, and that the pairwise comparisons indicate that not all the means are different (the situation described in the previous paragraph). This leaves the question which pair of means is the same. However, this question cannot be answered using p-values: none of H_{0a}, H_{0b} and H_{0c} is the obvious null hypothesis; and, comparison

Table 5.2: P-values and Posterior Probabilities

Hypothesis	p-value	Model	Post. Prob.	Non Sharp
H_0	.025	M_0	.028	.051
H_{0a}	.015	M_{1a}	.025	.045
H_{0b}	.022	M_{1b}	.037	.073
H_{0c}	.866	M_{1c}	.610	.680
		M_2	.301	.151

of non-nested hypotheses using p-values is not a well-developed area of statistics. As will be illustrated in the next section, comparison of non-nested models using posterior probabilities is straightforward and easy.

To elaborate on (ii). The wish to control the probability of an error of the first kind at a level of 5% leads to the strict rule that only p-values smaller than .05 lead to a rejection of the corresponding null hypothesis. As illustrated above, this may lead to inconsistent results. Common sense dictates that p-values of .045 and .055 (or, if you like, .049 and .051) should be evaluated in the same way. However, in practice many researchers (and journals) are 'happy' with .045, and 'disappointed' with .055. This problem can be avoided if researchers are willing to trade control of the error of the first kind before analysis of the data, for conditional error probabilities, that is, the probability of wrong decisions after observing the data. As will be illustrated in the next section, the latter can be achieved using posterior probabilities without needing an arbitrary criterion like '.05'.

5.3.2 Model selection, Encompassing Priors and Posterior Probabilities

Model selection usually proceeds in three steps. First, all the models that are of interest have to be specified. For the situation at hand there are five models (note, that a comma implies that the preceding or succeeding parameter is not equal to one of the other parameters):

$$M_0 : \mu_C = \mu_{E1} = \mu_{E2}$$

$$M_{1a} : \mu_C = \mu_{E1}, \mu_{E2} \qquad M_{1b} : \mu_C = \mu_{E2}, \mu_{E1} \qquad M_{1c} : \mu_C, \mu_{E1} = \mu_{E2}$$

$$M_2 : \mu_C, \mu_{E1}, \mu_{E2}.$$

The second step in model selection is the specification of prior probabilities for each of the models. As is clear from (5.2) the size of the posterior model probabilities

is influenced by the prior model probabilities. In the examples in this paper, prior probabilities are chosen to be the same for each model. The implication is that differences in posterior model probabilities are caused by differences in the degree to which the data support the model, that is, $P(\boldsymbol{y} \mid M_m)$, for $m \in \{0, 1a, 1b, 1c, 2\}$.

If a model contains unknown parameters the prior distribution of the unknown parameters has to be specified. This is the third step in model selection. It is necessary because if not all parameters are fixed at a specific value, $P(\boldsymbol{y} \mid M_m)$ can only be computed if the prior distribution $g(\cdot)$ of the model parameters is specified:

$$P(\boldsymbol{y} \mid M_m) = \int_{\boldsymbol{\theta}} P(\boldsymbol{y} \mid \boldsymbol{\theta}) g(\boldsymbol{\theta} \mid M_m) d\boldsymbol{\theta}. \tag{5.6}$$

Note that, $\boldsymbol{\theta}$ denotes the parameters of the model at hand, that is, the three population means and the within group variance σ^2, and that

$$P(\boldsymbol{y} \mid \mu_C, \mu_{E1}, \mu_{E2}, \sigma^2) = \prod_{A=C,E1,E2} \prod_{i \in A} \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{1}{2} \frac{(y_i - \mu_A)^2}{\sigma^2}. \tag{5.7}$$

Looking at (5.6), it is clear that the posterior model probabilities (5.2) depend on the prior distributions for each model. In the sequel it will be explained how prior distributions can be chosen such, that differences in posterior model probabilities depend mainly on the support of the data for the model at hand.

Specification of prior distributions is the bottle-neck of Bayesian model selection. The use of subjective prior distributions is often criticized because the resulting inferences are also subjective. See, for example, Sober (2002, pp. 22-24) and Howson (2002, pp. 53-61) for interesting discussions. However, irrespective of whether subjective inference is appreciated or not, the specification of subjective prior distributions is far from easy. Consider, for example, M_2. It is rather common (see, for example, Gelman, Carlin, Stern and Rubin, 1995, pp. 71-76) to use independent normal priors for the population means, and a scaled inverse chi-square prior for the within group variance. However, specifying the means and variances of the normal priors, and the scale and degrees of freedom of the scaled inverse chi-square prior such that these priors represent a researchers subjective prior knowledge, is a task that is beyond most researchers (and statisticians).

Both the criticism and the specification problem motivated a search for prior distributions that are not subjective and easily specified. The basic idea of encompassing priors is to specify a distribution that is noninformative with respect to the untransformed means and variance for the encompassing model, that is, the model within which all other models are nested. Subsequently, the priors for the nested models are derived from the encompassing prior. For the application at hand, M_2 is

the encompassing model. It contains four parameters: three population means and a within group variance. The other models can be obtained via restrictions on the parameter space of M_2 and are thus nested within this encompassing model. Using independent distributions for each model parameter, the prior distribution of M_2 can be denoted by $g(\mu_C \mid M_2)g(\mu_{E1} \mid M_2)g(\mu_{E2} \mid M_2)g(\sigma^2 \mid M_2)$, how to choose it such that it is noninformative will be discussed below. The prior distributions of the nested models follow directly from this encompassing prior:

$$g(\mu_C, \mu_{E1}, \mu_{E2}, \sigma^2 \mid M_m) =$$

$$\frac{g(\mu_C \mid M_2)g(\mu_{E1} \mid M_2)g(\mu_{E2} \mid M_2)g(\sigma^2 \mid M_2)I_{M_m}}{\int g(\mu_C \mid M_2)g(\mu_{E1} \mid M_2)g(\mu_{E2} \mid M_2)g(\sigma^2 \mid M_2)I_{M_m} d\mu_C, \mu_{E1}, \mu_{E2}, \sigma^2}, \qquad (5.8)$$

where the indicator function I has the value 1 if the argument is true, that is, if the parameter values are in accordance with the restrictions imposed by model M_m, and 0 otherwise.

Specification of the encompassing prior is often facilitated if the data are taken into consideration. As can be seen in Table 5.1, the response format is such that all data are in the interval 1-5. This implies that the means of the control and both experimental groups also have to be in this interval. Furthermore, with this response format, σ^2 can not be larger than 4 (if half of the persons in a group responds 1, and the other half 5). Consequently, noninformative priors for the means and variance are obtained if $g(\mu_C \mid M_2) = g(\mu_{E1} \mid M_2) = g(\mu_{E2} \mid M_2) = U[1, 5]$ and $g(\sigma^2 \mid M_2) = U[0, 4]$, where U denotes a uniform distribution with lower and upper bound as indicated. The resulting encompassing prior distribution is proper, the same holds for the prior distributions of the other models that can be derived from the encompassing prior using (5.8). If encompassing priors are used, (5.6) is the likelihood $P(\boldsymbol{y} \mid \boldsymbol{\theta})$ averaged over intervals specified by the prior distributions. Stated otherwise, the posterior probabilities are 'averaged likelihoods' transformed to probabilities using Bayes theorem.

After specification of the set of models, prior probabilities and the encompassing prior distribution, the posterior probability of each model can be computed using (5.2) with $P(\boldsymbol{y} \mid M_m)$ as specified in (5.6). Due to the integration involved, the computation of (5.6) is not always easy. The interested reader is referred to Carlin and Chib (1995) and Kass and Raftery (1995) for an inventory of methods for the computation of (5.6). Newton and Raftery (1994) propose to approximate the integral using importance sampling based on a sample of parameter vectors from both the posterior distribution of the model at hand, and, an imaginary prior distribution. Their method is used for all analyzes executed in this section.

The posterior probabilities for the models under consideration in the example at hand are given in the fourth column of Table 5.2. Models M_0, M_{1a} and M_{1b} have small posterior probabilities, that is, after observing the data it is rather unlikely that these models are true. The posterior probabilities of models M_{1c} and M_2 indicate that there is a posterior probability of about .9 (.61+.30) that the mean of the control group differs from the means of both experimental groups, and, that there is a posterior probability of about .3 that all three means are different.

5.3.3 A Comparison of P-values and Posterior Probabilities

In this section the limitations of p-values discussed previously will be summarized. It will also be elaborated how these limitations can be overcome using posterior probabilities.

1. As indicated in Section 5.2 a small p-value does not necessarily imply that H_0 has to be rejected in favor of the alternative hypothesis. As several authors have indicated before, the p-value usually overstates the amount of evidence against the null hypothesis (Berger and Sellke, 1987; Cohen, 1994). The reason for the latter is that $P(T(\boldsymbol{Y}) > t(\boldsymbol{y}) \mid H_0) \neq P(H_0 \mid \boldsymbol{y})$. This phenomenon can also be observed in the example at hand. The p-value for testing H_0 versus H_1 is .025, which is usually considered to be a substantive amount of evidence *against* H_0. However, if only M_0 and M_2 are considered, the posterior probability of the corresponding model is .085 (.028/(.028+.301)), which does not imply a straightforward rejection of M_0.

2. In contrast to p-values, the comparison of non-nested models is not a problem if posterior probabilities are used. Looking at Table 5.2 it is evident that M_{1c} is superior to M_{1a} and M_{1b}.

3. In contrast to p-values posterior probabilities that differ only slightly in size are not evaluated differently because of the requirement to adhere to a criterion (like an alpha-level of .05) that has to be fixed *before* the analysis. Posterior probabilities can be interpreted as conditional error probabilities (Cohen, 1994; Sellke, Bayarri and Berger, 2001). To elaborate on the latter, consider only M_0 and M_2. The posterior probabilities if only these two models are considered are .085 and .915, respectively. Stated otherwise, the probability *after* observing the data that M_0 is true is .085, that is, the probability of incorrectly rejecting M_0 after observing the data (the conditional error of the first kind) is .085. It is interesting that this probability is larger than the common level alpha of .05. However, since there is no pre-specified criterion

that has to be used to evaluate posterior probabilities, each researcher can make his own decision whether .085 is small enough to discard M_0.

4. It is relatively easy to define and evaluate models that are *relevantly* different from each other using posterior probabilities. In the next section this topic will be elaborated and illustrated using the data at hand.

5.3.4 The Nil Hypothesis and Non Sharp Null Models

As mentioned before, both Frick (1996) and Wainer (1999) present examples where it is sufficient to determine whether data are consistent with H_0 or not. The null hypotheses they consider are of the kind 'an average is zero' and 'the difference between two averages is zero', that is, so called 'nil-hypotheses' (Cohen, 1994), or sharp or point null hypotheses (Berger and Sellke, 1987). The examples given by Frick (1996) and Wainer (1999) are convincing. However, there are also examples where the evaluation of a sharp H_0 does not make sense. Cohen (1994) clearly takes the position that 'the difference between two averages' is never zero, and thus, that it is meaningless to test the corresponding null hypothesis. Apparently many researchers agree with Cohen, because the evaluation of a significant null hypothesis is often supported by the evaluation of an effect size measure in order to be able to replace the conclusion H_1 by *relevant* H_1.

Model M_0 introduced in Section 5.3.1 could be called a sharp model. Models M_{1a}, M_{1b} and M_{1c} contain 'sharp elements'. The criticism on 'nil-hypotheses' also applies to these models. Whether or not M_2 is *relevantly* better than M_0 can only be determined using effect size measures in addition to the posterior probabilities. However, it also possible to reformulate these models such that effect sizes are included. All a researcher has to do is determine which difference between two means is considered to be relevant. Here a difference of .4 (which is 10% of the distance of the scale on which helping behavior is rated) is considered to be relevant. Furthermore, the prior expectation that helping behavior should increase from C via $E1$ to $E2$ is used during the construction of M_{1a}, M_{1c} and M_2. Note that M_{1b} is not in agreement with this prior expectation. This leads to the following counterparts of the models used sofar:

$$M_0 : |\mu_C - \mu_{E1}| \leq .4, \quad |\mu_C - \mu_{E2}| \leq .4, \quad |\mu_{E1} - \mu_{E2}| \leq .4,$$

$$M_{1a} : |\mu_C - \mu_{E1}| \leq .4, \quad \mu_{E2} - \mu_C \geq .4, \quad \mu_{E2} - \mu_{E1} \geq .4,$$

$$M_{1b} : \mu_{E1} - \mu_C \geq .4, \quad |\mu_C - \mu_{E2}| \leq .4, \quad \mu_{E1} - \mu_{E2} \geq .4,$$

$$M_{1c} : \mu_{E1} - \mu_C \geq .4, \quad \mu_{E2} - \mu_C \geq .4, \quad |\mu_{E1} - \mu_{E2}| \leq .4,$$

$$M_2 : \mu_{E1} - \mu_C \geq .4, \quad \mu_{E2} - \mu_C \geq .4, \quad \mu_{E2} - \mu_{E1} \geq .4.$$

Using (5.2), (5.6) and (5.7) the posterior probabilities of these models can be computed. Note that, the prior distributions for the parameters of these models can still be derived using (5.8), even if the encompassing model itself is no longer part of the models under investigation. As can be seen in the last column of Table 5.2, model M_{1c} 'the average in group C is *relevantly* different from the averages in groups $E1$ and $E2$' has a higher posterior probability than the counterpart discussed in Section 5.3.2 (one but last column of Table 5.2). Model M_2 'all the averages are relevantly different' has a substantially lower posterior probability. Stated otherwise, M_{1c} has by far the highest posterior probability, but the other models can not completely be ruled out.

Giving a difference between two means of more than .4 the label *relevant* is of course subjective. However, nothing prevents other researchers from using a different number. As long as the threshold value used is reported, the meaning of the corresponding posterior probabilities is clear. Comparison of hypotheses corresponding to the five models using p-values can, as far as the authors know, not be found in the literature. The main problems here are null models where the parameters of interest are *not* fixed at zero, and, a test-statistic that has power against the alternative hypotheses specified. A point of departure to solve these problems might be developments in testing hypotheses with inequality constraints among the parameters. The interested reader is referred to Robertson, Wright and Dykstra (1988), who give a comprehensive overview of order restricted inference, that is, hypotheses that are more informative than the traditional null and alternative hypotheses.

5.4 Training Data

5.4.1 Multiple Regression

Stevens (1996, pp. 578-585) describes the Sesame street data. These data contain measurements of the knowledge of body-parts, numbers, forms and the like, before and after the first year of the television program, for 240 children in the age range between three to five. The research question of interest in this section is whether knowledge of letters after watching Sesame street for a year is better predicted using knowledge of letters (a topic that receives a lot of attention in the series) before watching Sesame street, using the Peabody picture vocabulary test which is not related to topics presented in Sesame street, or both.

The research question at hand is representative for a class of questions that is

Table 5.3: Analyzing Three Models Using Multiple Regression and Training Data

Predictor	Letters	Peabody	Both
Constant	12.41	5.40	2.52
Regr. Coeff. Letters	.83		.63
Regr. Coeff. Peabody		.46	.29
R^2	.34	.27	.43
p-value	.00	.00	.00
R^2-Cross Validated	.27	.21	.35
$P(M_m \mid \boldsymbol{x})$	.001	.000	.999
$P(M_m \mid \boldsymbol{x})$	.98	.02	

often encountered: is one set of predictors superior to another set, or, should the sets be combined (see, for example, Congdon, 2001, pp. 139-142, and Tabachnick and Fidell, 2001, pp. 131-139). The main results of straightforward multiple regression analyzes are presented in Table 5.3. Not the whole data matrix was used, but $N = 122$ randomly selected children. This part of the data matrix will subsequently be called the calibration sample. Irrespective of the model used, all predictors have a positive relationship with knowledge of numbers after watching Sesame street for a year. The models with one predictor explain, 34% (*Letters*) and 27% (*Peabody*) of the variation of the dependent variable, the model with two predictors explains 43%. For each model the null hypothesis that the predictors can not be used to predict the dependent variable is rejected (each p-value is .00).

Two questions remain: are both one predictor models about equally good, or, is one better than the other; and, is the two predictor model better than both one predictor models, or, is the increase in explained variation not that impressive considering the fact that the model contains an extra predictor compared to the other models. For the example at hand, the second question could be investigated via the testing of nested regression models (Tabachnick and Fidell, 2001, pp. 165-170). However, in related but more elaborate situations (more than two sets of predictors, and various combinations of these sets) this does not provide a solution since mostly non-nested models have to be compared. Even for the example at hand (two non-nested models with one predictor) this is not easily done via hypothesis testing. It is not clear which are the null and alternative model, and, the actual testing of non-nested models is an underdeveloped area in statistics.

Nevertheless, both questions can be addressed using non-Bayesian methods via cross-validation (Camstra and Boomsma, 1992; Stevens, 1996, 96-98). There are

several kinds of cross-validation. In the simplest form the data matrix is randomly split into two parts: the calibration sample and the validation sample. The calibration sample is used to estimate the regression coefficients of each model of interest. Subsequently, the validation sample (here $N = 118$) is used to compute for each model the squared correlation between the dependent variable and its predicted value using the regression equations obtained in the calibration sample. The result is called the cross-validated proportion of variance explained. It is a measure of the predictive performance of a model that is not biased by the number of predictors in the model, or the step-wise or other 'trial and error' methods by which models are constructed and the regression coefficients estimated. As can be seen in Table 5.3, the cross-validated R^2 of the model with both predictors is .08 higher than the model containing only Letters, and .14 higher than the model containing only Peabody. It appears (based on our subjective evaluation of the increase in cross-validated R^2) that the model with two predictors gives better predictions than the models with one predictor. The decision whether the model containing only Letters is relevantly better than the model containing only Peabody is left to the reader.

Possibly the only drawback of the cross-validated R^2 is the lack of a formal comparison of the models. Questions like is one model 'significantly better' or 'more probable' than the others, have not yet been addressed. In the next section it will be explained how these questions can be answered using posterior model probabilities.

5.4.2 Posterior Probabilities Computed using Training Data

To compute posterior probabilities, for each model the prior distribution of each parameter that is not fixed at a specific value has to be specified. One of the goals of this paper is to show that prior distributions can be specified such that they do not represent the subjective view of the researcher, but nevertheless are useful and informative. As illustrated in Section 5.3, encompassing priors are one way to achieve this. In this section training data (Berger and Pericchi, 1996) or, if you like, the calibration sample, will be used to specify objective priors.

One of the interpretations given to prior knowledge is that it should reflect the current state of affairs with respect to the research questions and models at hand. It is rare (if it happens at all) that previous research can be used to specify prior distributions for the parameters of the models of interest. However, it does occur that researchers have enough data to randomly assign each person to a training and a validation set. If cross-validation is used, training (calibration) data are used to construct and estimate the parameters of multiple regression models. If the goal is to compute posterior model probabilities, training data are used to construct and

summarize the information with respect to the parameters of multiple regression models. This summary is the *posterior* distribution of the model parameters for the *training data*, which will serve as the *prior* distribution of the model parameters for the *validation* data. Stated otherwise, the training data are used to summarize the knowledge with respect to the current state of affairs for each of the models under consideration, and the validation data are used to compute posterior probabilities.

For the multiple regression model

$$P(\boldsymbol{y} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_P, \beta_0, \ldots, \beta_P, \sigma^2) =$$

$$\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{1}{2} \frac{(y_i - \beta_0 - \sum_{p=1}^{P} \beta_p x_{ip})^2}{\sigma^2}, \tag{5.9}$$

where x_{ip} denotes the response of the i-th person to the p-th predictor, y_i the response to the criterion variable, β_p the regression coefficient of the p-th predictor, and σ^2 the residual variance. Subsequently, the superscripts t and v will be used to denote prior distributions and data for the training and validation data set, respectively.

Different regression models are obtained using different subsets of the total set of predictors. For each model it is known which predictors are included. In the sequel M_m will be used to denote the set of predictors included for model m. More specifically, $M_1 = \{Letters\}$, $M_2 = \{Peabody\}$ and $M_3 = \{Letters, Peabody\}$. The parameters of these models will be denoted by $\boldsymbol{\theta}_m = \{\beta_{0m}, \ldots, \beta_{Pm}, \sigma_m^2\}$, where P denotes the number of predictors in the model at hand.

In the Bayesian approach the information with respect to the parameters of a model is summarized in the posterior distribution (Gelman, Carlin, Stern and Rubin, 1995, pp. 32-33):

$$Post(\boldsymbol{\theta}_m \mid \boldsymbol{y}^t, M_m^t) \propto P(\boldsymbol{y}^t \mid M_m^t, \boldsymbol{\theta}_m) g^t(\boldsymbol{\theta}_m), \tag{5.10}$$

which combines the information provided by the training data with the prior information with respect to the model parameters that is available *before* the *training* data are analyzed. Because it is in accordance with the goal to compute posterior probabilities that are independent of subjective choices, a noninformative prior will be used. For multiple regression models noninformative prior distributions can be chosen in various ways (Gelman, Carlin, Stern and Rubin, 1995, Chapter 8). In this paper the improper noninformative prior $g^t(\boldsymbol{\theta}_m) = 1$ will be used. This renders a proper posterior distribution that is proportional to the likelihood of the data.

The posterior distributions constructed using the training data represent the prior knowledge with respect to the current state of affairs before the *validation*

data are used to compute posterior probabilities for each model in the set of models under consideration, that is, $g^v(\boldsymbol{\theta}_m) = Post(\boldsymbol{\theta}_m \mid \boldsymbol{y}^t, M_m^t)$. These, resulting from 'previous research' and thus objective prior distributions can be used to compute (5.6) for the multiple regression model, that is,

$$P(\boldsymbol{y}^v \mid M_m^v) = \int_{\boldsymbol{\theta}_m} P(\boldsymbol{y}^v \mid M_m^v, \boldsymbol{\theta}_m) g^v(\boldsymbol{\theta}_m). \tag{5.11}$$

Combined with equal prior probabilities for each of the models, this is sufficient to compute objective posterior probabilities using (5.2):

$$P(M_m^v \mid \boldsymbol{y}^v) = \frac{P(\boldsymbol{y}^v \mid M_m^v)}{\sum_m P(\boldsymbol{y}^v \mid M_m^v)}. \tag{5.12}$$

The integral in (5.11) can be evaluated if the integral with respect to $\boldsymbol{\theta}_m$ is replaced by a summation over a sample $\boldsymbol{\theta}_m^q$, for $q = 1, \ldots, Q$ from $g^v(\boldsymbol{\theta}_m)$, that is,

$$P(\boldsymbol{y}^v \mid M_m^v) \approx \frac{1}{Q} \sum_{q=1}^{Q} P(\boldsymbol{y}^v \mid M_m^v, \boldsymbol{\theta}_m^q). \tag{5.13}$$

This sample can be obtained using the Gibbs-sampler. The interested reader is referred to Gelman, Carlin, Stern and Rubin (1995, pp. 235-239,326-329).

As can be seen in Table 5.3, if all three models are compared the model with both predictors has by far the largest posterior probability. Since posterior probabilities implicitly penalize models with more parameters (Ockham's razor, see Kass and Raftery (1995)) the result is not just caused by the fact that an extra predictor usually leads to an increased amount of variance explained in the sample. It is caused by the fact that the extra predictor leads to an increased amount of variance explained in the population. If only both one-predictor models are compared (the last line of Table 5.3), M_1 has a substantially higher posterior probability than M_2. Stated otherwise, after observing the data, the (conditional error) probability that M_2 is the best model is so small, that it is save to conclude that *Letters* is a better predictor than *Peabody*.

In this section it was illustrated that prior distributions can be specified using training data. The interested reader is referred to Berger and Pericchi (1996), who show that this procedure can be refined in various ways. One way is to split a data file in a training and validation sample not once, but many times and combine the results. The other way is to adjust the former such that the training data contain as few persons as possible. In this section it was shown that training data based posterior probabilities can be used in addition to cross-validation, to select the best of a number of non-nested regression models. As far is known to the authors, there are no onsets in the literature to do the same using hypothesis testing.

5.5 Summary and Remaining Issues

The goal of this paper is to show the potential of posterior probabilities and model selection as an alternative for the use of p-values and hypotheses testing. As was illustrated using a simple model for the evaluation of a population mean, a one-way analysis of variance followed by pairwise comparisons of means, and multiple regression, posterior probabilities have a number of advantages over p-values:

1. Posterior probabilities are computed for each of the models in the set of models under consideration. This avoids (possibly incorrect) indirect conclusions of the kind 'not H_0' implies 'H_1'.

2. Posterior probabilities can be used to compare non-nested models, and in doing so incorporate the complexity (number of parameters) of a model (Ockham's razor).

3. Posterior probabilities do not need arbitrary criteria like .05 in order to be evaluated.

4. Posterior probabilities can be interpreted as conditional error probabilities, that is, the error probabilities *after* observing the data.

5. Posterior probabilities can be used to evaluate models that are *relevantly* different from each other.

A difficulty using posterior probabilities is that for each model the prior distribution of the parameters of that model has to be specified. This paper illustrated two methods that are easy to apply, and can be used to derive priors that are not based on subjective a priori information: encompassing priors and training data.

Unlike hypotheses testing and p-values, posterior probabilities have not yet found their way into the toolkit of psychological researchers. Only in the last decade (see, for example, Berger and Pericchi, 1996; Carlin and Chib, 1995; and, Kass and Raftery, 1995) computational developments have enabled the calculation and use of posterior probabilities. Research with respect to application of posterior probabilities is still ongoing, and the standard software packages used by psychological researchers have not yet included modules and models for which posterior probabilities can be computed. Readers interested in using the methods proposed, can write an e-mail to the authors describing their data and research questions.

References

Agresti, A. (1984), *Analysis of ordinal categorical data*, New York: John Wiley & Sons.

Agresti, A. and Coull, B.A. (2002), The analysis of contingency tables under inequality constraints, *Journal of Statistical Planning and Inference, 1-2*, 45-73.

Anraku, K. (1999), An information criterion for parameters under a simple order restriction, *Biometrika, 86*, 141-152.

Asch, S.E. (1955), Opinions and social pressure, *Scientific American, 193*, 31-35.

Barlow, R.E., Bartholomew, D.J., Bremner, J.M. and Brunk, H.D. (1972), *Statistical inference under order restrictions*, New York: John Wiley & Sons.

Bartlett, M.S. and Kendall, D.G. (1946), The statistical analysis of variance-heterogeneity and the logarithmic transformation, *Supplement to the Journal of the Royal Statistical Society, Series B, 1,* 128-138.

Bayarri, M.J. and Berger, J.O. (2000), P values for composite null models, *Journal of the American Statistical Association, 95*, 1127-1142.

Berger, J.O. and Delempady, M. (1987), Testing precise hypotheses, *Statistical Science, 2*, 317-352.

Berger, J. and Pericchi, L. (1996), The intrinsic Bayes factor for model selection and prediction, *Journal of the American Statistical Association, 91*, 109-122.

Berger, J. and Pericchi, L. (2001), Objective Bayesian methods for model selection: introduction and comparison. In *Model Selection* (P.Lahiri, ed.) Monograph Series, 38, 135-207, Beachwood Ohio: Institute of Mathematical Statistics Lecture Notes.

Berger, J.O. and Sellke, T. (1987), Testing a point null hypothesis: the irreconcilability of p-values and evidence, *Journal of the American Statistical Association, 82*, 112-122.

Bernardo, J.M. and Smith, A.F.M. (1994), *Bayesian theory*, Chichester: John Wiley & Sons.

Boelen, P. A., Bout, J. van den, and Keijser, J. de (2002), A controlled study on the efficacy of cognitive behavioural therapy for traumatic grief: first results. Paper presented at the 32nd annual congress of the European Association for Behavioural and Cognitive Therapies, Maastricht, the Netherlands. September 19-21, 2002.

Brown M.B. and Forsythe A.B. (1974), Robust tests for the equality of variances, *Journal of the American Statistical Association, 69*, 364-367.

Camstra, A. and Boomsma, A. (1992), Cross-Validation in regression and covariance structure analysis, *Sociological Methods and Research, 21*, 89-95.

Carlin, B.P. and Chib, S. (1995), Bayesian model choice via Markov chain Monte Carlo methods, *Journal of the Royal Statistical Society, B, 57*, 473-484.

Chen, M.H. and Shao, Q.M. (1997), On Monte Carlo methods for estimating ratios of normalizing constants, *Annals of Statistics, 25*, 1563-1594.

Chen, M.H., Shao, Q.M. and Ibrahim, J.G. (2000), *Monte Carlo methods in Bayesian computation*, New York: Springer-Verlag.

Chib, S. (1995), Marginal likelihood from the Gibbs output, *Journal of the American Statistical Association, 90*, 1313-1321.

Chongcharoen, S., Singh, B. and Wright, F.T. (2002), Powers of some one-sided multivariate tests with the population covariance matrix known up to a multiplicative constant, *Journal of Statistical Planning and Inference, 107,* 103-121.

Cohen, J. (1994), The earth is round ($p < .05$), *American Psychologist, 12*, 997-1003.

Congdon, P. (2001), *Bayesian statistical modelling.* New York: John Wiley & Sons.

Cowles, M.K. and Carlin, B.P. (1996), Markov chain Monte Carlo convergence diagnostics: a comparative review, *Journal of the American Statistical Association, 91*, 883-904.

Dayton, C.M. (2003), Information criteria for pairwise comparisons, *Psychological Methods, 8*,61-71.

Doveh, E., Shapiro, A. and Feigin, P.D. (2002). Testing of monotonicity in parametric regression models, *Journal of Statistical Planning and Inference, 107*, 289-306.

Follmann, D. (1996), A simple multivariate test for one-sided alternatives, *Journal of the American Statistical Association, 91*, 854-861.

Frick, R.W. (1996), The appropriate use of null hypothesis testing, *Psychological Methods, 1*,379-390.

Gelfand, A.E. and Dey, D. (1994), Bayesian model choice: asymptotics and exact calculations, *Journal of the Royal Statistical Society, B, 56*, 501-514.

Gelfand, A.E., Smith, A.F.M. and 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. (2004), Parameterization and Bayesian modeling, *Journal of the American Statistical Association, 99*, 537-545.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995), *Bayesian data analysis*, London: Chapman & Hall.

Green, P.J. (1995), Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika, 82*, 711-732.

Han, C. and Carlin, B.P. (2001), Markov chain Monte Carlo methods for computing Bayes factors: a comparative review, *Journal of the American Statistical Association, 96*, 1122-1132.

Hoijtink, H. (2001), Confirmatory latent class analysis: model selection using Bayes factors and (pseudo) likelihood ratio statistics, *Multivariate Behavioral Research, 36*, 563-588.

Howson, C. (2002), Bayesianism in statistics. In: R. Swinburne (Ed.), *Bayes Theorem*, pp. 39-69, Oxford: Oxford University Press.

Jeffreys, H. (1961), *Theory of probability*, 3rd ed., London: Oxford Univ. Press.

Johnson, N.L., Kotz, S. and Kemp, A.W. (1993), *Univariate discrete distributions*, New York: John Wiley & Sons.

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

Lavine, M. and Schervish, M.J. (1999), Bayes factors: what they are and what they are not, *The American Statistician, 53,* 119-122.

Lee, P.M. (1997), *Bayesian Statistics: An introduction*, London: Arnold.

Lehmann, E.L. (1997), *Testing statistical hypotheses. Second Edition*, New York: Springer-Verlag.

Lindley, D.V. (1957), A statistical paradox, *Biometrika, 44*, 187-192.

Lindgren, B.W. (1993), *Statistical theory, Fourth Edition*, Boca Raton: Chapman & Hall/CRC.

Marden, J.I. (2000), Hypothesis testing: from p values to Bayes factors, *Journal of the American Statistical Association, 95*, 1316-1320.

McDermott, M.P. and Mudholkar, G.S. (1993), A simple approach to testing homogeneity of order-constrained means, *Journal of the American Statistical Association, 88*, 1371-1379.

Mood, A.M., Graybill, F.A. and Boes, D.C. (1974), *Introduction to the theory of statistics, third edition,* Singapore, McGraw-Hill.

Mosteller, F. and Tukey, J.W. (1977), *Data analysis and regression*, Reading, Massachusetts: Addison-Wesley, Exhibit 8.

Mukerjee, H. and Tu, R. (1995), Order-restricted inferences in linear regression, *Journal of the American Statistical Association, 90*, 717-728.

Muller, K.E. (1998), A new F-approximation for the Pillai-Bartlett trace under H_0, *Journal of Computational and Graphical Statistics, 7*, 131-138.

Newton, M.A. and Raftery, A.E. (1994), Approximate Bayesian inference by the weighted likelihood bootstrap, *Journal of the Royal Statistical Society, B, 42,* 213-220.

Perlman, M.D. and Wu. L. (2002), A class of conditional tests for a multivariate one-sided alternative, *Journal of Statistical Planning and Inference, 107,* 155-171.

Ramsey, P.H. (2002). Comparison of closed testing procedures for pairwise testing of means. *Psychological Methods, 7*,504-523.

Robert, C.P. and Hwang, J.T.G. (1996). Maximum likelihood estimation under order restriction by the prior feedback method. *Journal of the American Statistical Association, 91*, 167-172.

Robertson, T., Wright, F.T. and Dykstra, R.L. (1988). *Order restricted statistical inference*. New York: John Wiley & Sons.

Robins, J.M., van der Vaart, A. and Ventura, V. (2000). The asymptotic distribution of p values in composite null models. *Journal of the American Statistical Association, 95*, 1143-1156.

Sasabuchi, S., Tanaka, K. and Tsukamoto, T. (2003), Testing homogeneity of multivariate normal mean vectors under an order restriction when the covariance matrices are common but unknown, *The Annals of Statistics, 31*, 1517-1536.

Sellke, T., Bayarri, M.J. and Berger, J.O. (2001), Calibration of p values for testing precise null hypotheses, *The American Statistician,55*, 62-71.

Shoemaker, L.H. (2003), Fixing the F test for equal variances, *The American Statistician, 57*, 105-114.

Silvapulle, M.J. (1992a), Robust tests of inequality constraints and one-sided hypotheses in the linear model. *Biometrika, 79,* 621-630.

Silvapulle, M.J. (1992b), Robust Wald-type tests of one-sided hypotheses in the linear model, *Journal of the American Statistical Association, 87,* 156-161.

Silvapulle, M.J. (1995a), On an F-type statistic for testing one-sided hypotheses and computation of chi-bar-squared weights, *Statistics & Probability Letters, 28,* 137-141.

Silvapulle, M.J. (1995b), A Hotelling's T^2-type statistic for testing against one-sided hypotheses, *Journal of Multivariate Statistics, 55,* 312-319.

Smith, A.F.M. and Roberts, G.O. (1993), Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods, *Journal of the Royal Statistical Society, B, 55*, 3-23.

Smith, A.F.M. and Spiegelharter, D.J. (1980), Bayes factors and choice criteria for linear models, *Journal of the Royal Statistical Society, B, 56*, 501-514.

Snedecor, G.W. and Cochran, G.C. (1967), *Statistical methods,* 6th ed., Ames, Iowa: Iowa State University Press.

Sober, E. (2002), Bayesianism - its scope and limits. In: R. Swinburne (Ed.), *Bayes Theorem*, pp. 21-38, Oxford: Oxford University Press.

Statistical inference under inequality constraints, 2002, special issue of the *Journal of Statistical Planning and Inference, 107, numbers 1-2.*

Stevens, J. (1996), *Applied multivariate statistics for the social sciences, third edition,* Mahwah, New Jersey: Lawrence Erlbaum.

Tabachnick, B.G. and Fidell, L.S. (2001), *using multivariate statistics*, London: Allyn and Bacon.

Tang, D.I. (1994), Uniformly more powerful tests in a one-sided multivariate problem, *Journal of the American Statistical Association, 89,* 1006-1011.

Tatsuoka, M.M. (1988), *Multivariate analysis. Techniques for educational and psychological research*, New York: Macmillan Publishing Company.

Tierney, L. (1994), Markov chains for exploring posterior distributions, *The annals of statistics, 22,* 1701-1762.

Toothaker, L.E. (1993), *Multiple comparison procedures.* London: SAGE.

Wainer, H. (1999), One cheer for null hypothesis significance testing, *Psychological Methods, 4*, 212-213.

Zellner, A. and Siow, A. (1980), Posterior odds ratios for selected regression hypotheses, In *Bayesian Statistics* (J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds.), 585-603, Valencia: University Press.

Samenvatting

Onderzoekers hebben veelal een of meer theorieën of verwachtingen met betrekking tot de uitkomsten van empirisch onderzoek. Om deze theorieën te toetsen, moeten ze worden omgezet in statistische modellen. Onderzoekers spreken vaak over verwachte relaties en uitkomsten in termen van ongelijkheidsrestricties. Dat wil zeggen dat er verwacht wordt dat een of meer parameters van het statistische model grotere waarden zullen aannemen dan een of meer andere parameters. In dit proefschrift worden normale lineaire modellen besproken waarin ongelijkheidsrestricties zijn opgelegd aan model parameters.

Wanneer de onderzoeker één theorie heeft die met ongelijkheidsrestricties gespecificeerd kan worden, kan een nulhypothese toets met een ongelijkheids-gerestricteerde alternatieve hypothese gebruikt worden. Dit wordt besproken in het tweede hoofdstuk van dit proefschrift. Veelal zijn er echter meerdere, competitieve theorieën, ofwel meerdere statistische modellen met ongelijkheids-gerestricteerde parameters. In dat geval is het beter om de modellen direct met elkaar te vergelijken. In de overige hoofdstukken van dit proefschrift worden de mogelijkheden besproken van Bayesiaanse modelselectie in de context van modellen met gerestricteerde parameters.

Hoofdstuk 2 behandelt hypothese toetsing voor nulhypothesen die behoren tot het univariate of multivariate normale model. Om een p-waarde te bepalen voor een geobserveerde dataset zijn twee ingrediënten nodig: een toetsingsgrootheid en de verdeling van die toetsingsgrootheid onder de nulhypothese. Voor nulhypothesen die *nuisance parameters* bevatten, is het vaak niet eenvoudig om de verdeling van de toetsingsgrootheid onder de nulhypothese vast te stellen. Er worden dan ook vaak benaderende verdelingen gebruikt, die leiden tot asymptotische p-waarden.

In dit hoofdstuk wordt een nieuwe methode gepresenteerd die onder de nulhypothese data repliceert, conditioneel op de geobserveerde waarden voor de *sufficient statistics* voor de *nuisance parameters* van de betreffende nulhypothese. De verdeling van de gerepliceerde data wordt de *similar* nulverdeling van de data genoemd. Dit is analoog aan de term *similar* p-waarde voor toetsen waarbij de verdeling van de toetsingsgrootheid conditioneel op de *sufficient statistics* gebruikt wordt om de (exacte) p-waarde te bepalen. Door de verdeling van de data te repliceren, is voor elke gewenste toetsingsgrootheid de *similar* verdeling en daarmee de *similar* p-waarde bekend.

In dit hoofdstuk worden voorbeelden gepresenteerd voor zowel univariate als multivariate nulhypothesen met ongelijkheids-gerestricteerde alternatieve hypothesen. Ook toepassingen in andere domeinen zijn mogelijk, zoals het toetsen op gelijkheid van varianties.

In Hoofdstuk 3 wordt de Bayesiaanse aanpak uiteengezet. Zowel het schatten van parameters onder ongelijkheidsrestricties (inclusief standaardfouten en schattingsintervallen) als een modelselectie procedure wordt besproken. Er worden voorbeelden gegeven van (psychologische) theorieën die in termen van ongelijkheidsrestricties worden gepresenteerd. De restricties worden als *a priori* informatie opgenomen in de Bayesiaanse analyse. Dit levert een getrunceerde prior verdeling op en als gevolg daarvan ook een getrunceerde posterior verdeling. Er wordt uitgelegd hoe met behulp van een Gibbs sampler en *importance sampling*, een steekproef uit deze posterior kan worden verkregen. Met een steekproef uit de posterior kunnen alle gewenste schatters en schattingsintervallen worden bepaald.

Vaak zijn er competitieve theorieën, dat wil zeggen meerdere modellen met verschillende restricties op de modelparameters. In dat geval is het van belang om via een modelselectie procedure de fit van de data op elk van deze modellen te vergelijken. Het gebruikte Bayesiaanse modelselectie criterium is de Bayes factor. Om deze te berekenen is voor elk model de marginale likelihood van de data nodig. Het schatten van een marginale likelihood is complex maar kan vermeden worden voor het type modellen die in dit proefschrift besproken worden. Voor deze modellen geldt namelijk dat alle gerestricteerde modellen genest zijn in een ongerestricteerd ofwel *encompassing* model. Voor de parameters van het *encompassing* model wordt een prior verdeling gespecificeerd. De prior verdelingen voor de geneste modellen volgen hieruit door het restricteren van de toegestane parameter ruimte. Als gevolg hiervan kan op relatief eenvoudige wijze de Bayes factor worden bepaald voor elk genest model met het *encompassing* model. Tenslotte is het ook mogelijk om voor een eindige set van modellen de relatieve fit van elk van deze modellen uit te drukken in zogenaamde posterior modelkansen.

In het volgende hoofdstuk (Hoofdstuk 4) wordt verder ingegaan op en onderzoek gedaan naar Bayesiaanse modelselectie. Een belangrijk issue in de Bayesiaanse statistiek is de sensitiviteit van de resultaten voor de prior verdeling. Dat wil zeggen dat de uitkomsten van de analyse beïnvloed worden door de keuze die gemaakt is voor de prior verdeling van de parameters. Het is mogelijk om non-informatieve, ofwel vage of diffuse priors te kiezen, zodat de posterior verdeling wordt gedomineerd door de data en dus relatief onafhankelijk is van de prior. In het algemeen kunnen deze priors voor modelselectie echter niet gebruikt worden, omdat Bayes factors dan arbitraire waarden leveren. Dit is echter niet het geval als gebruik wordt gemaakt van de in dit hoofdstuk gepresenteerde *encompassing* prior aanpak.

Er wordt aangetoond dat bij gebruikmaking van een diffuse *encompassing* prior, voor specifieke klassen van modellen de modelselectie vrijwel objectief is. Dit betekent dat de Bayes factors niet sensitief zijn voor de specificatie van de *encompassing* prior. Voor andere klassen van modellen worden de resultaten van de modelselectie sterk beïnvloed door de *encompassing* prior. Een voorbeeld is een model met 'ongeveer gelijk' restricties tussen de parameters. Hoe diffuser de *encompassing* prior, hoe groter de support voor het geneste model. Dit fenomeen staat bekend als Lindley's of Bartlett's paradox.

Tevens wordt in dit hoofdstuk een nieuwe manier om Bayes factors te berekenen gepresenteerd, die in de context van *encompassing* priors gebruikt kan worden. In de *encompassing* prior aanpak is de Bayes factor voor elk genest model met het *encompassing* model gelijk aan de ratio van twee proporties, namelijk de proportie uit de *encompassing* prior die voldoet aan de restricties van het geneste model, en de proportie uit de *encompassing* posterior die voldoet aan de restricties van het geneste model. Kortom, de Bayes factors voor alle geneste modellen met het *encompassing* model kunnen worden geschat met één steekproef uit de ongerestricteerde prior en één steekproef uit de ongerestricteerde posterior. Ook de standaardfout van de Bayes factor kan met deze aanpak eenvoudig worden bepaald.

In het laatste hoofdstuk worden traditionele nulhypothese toetsing en Bayesiaanse modelselectie met elkaar vergeleken. Een belangrijk verschil is het expliciete gebruik van voorkennis door de Bayesianen, en het expliciete gebruik van nulverdelingen door frequentisten. Het specificeren van de nulverdeling is vaak geen probleem. Het is daarentegen niet altijd evident hoe resulterende p-waarden geïnterpreteerd moeten worden. Een aantal nadelen of risico's van het gebruik van p-waarden worden besproken: (1) Een p-waarde geeft alleen informatie over het wel of niet passen van de nulhypothese. Het is verleidelijk maar incorrect om op grond van een verworpen nulhypothese te concluderen dat de alternatieve hypothese dus juist is. Bovendien overschat de p-waarde veelal de hoeveelheid bewijs tegen de

nulhypothese. (2) Het is problematisch om niet-geneste modellen via nulhypothese toetsing met elkaar te vergelijken. (3) De algemene kritieke waarde .05 is arbitrair en kan in het geval van meervoudige toetsingen leiden tot inconsistente conclusies. (4) Veelal wordt ook relevantie betrokken bij het beoordelen van (significante) resultaten. Het is echter niet mogelijk om de relevantie formeel mee te nemen in de nulhypothese toets.

Modelselectie wordt besproken als mogelijk alternatief voor nulhypothese toetsing. In dit hoofdstuk wordt een Bayesiaans modelselectie criterium gepresenteerd, resulterend in posterior modelkansen. Het is typisch voor Bayesiaanse procedures dat de onderzoeker de prior verdelingen moet specificeren. Dit formaliseren van voorkennis is vaak moeilijk. In dit hoofdstuk worden twee praktische benaderingen gepresenteerd voor de specificatie van prior verdelingen, namelijk *encompassing* priors en *training* data. Deze twee benaderingen worden geïllustreerd aan de hand van voorbeelden uit de psychologie, te weten een variantie analyse met daarop volgende paarsgewijze vergelijkingen en een multipele regressie waarbij niet-geneste modellen worden vergeleken.

Curriculum Vitae

Irene Klugkist was born on December 28, 1970 in Delfzijl, The Netherlands. In 1989, she completed pre-university education (VWO) at Buys Ballot College in Goes. She studied Pedagogical Sciences at the University of Utrecht from 1989, until graduation in 1995. In 1996, she started as a part-time lecturer at the Department of Methodology and Statistics at the Faculty of Social Sciences at Utrecht University. From 1997 until 1999 she owned and ran a research company and did several statistical consultations, mainly in the field of evaluation research. In 1999 she started working full-time at the Department of Methodology and Statistics as a part-time lecturer and working the rest of the time on this dissertation. Currently, she is an assistant professor at the Department of Methodology and Statistics at the University of Utrecht.