Model Selection

How to Evaluate Order Restrictions

Model Selectie

Het Evalueren van Ongelijksheidsrestricties

(met een samenvatting in het Nederlands)

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G. J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op vrijdag 27 januari 2012 des middags te 12.45 uur

door

Rebecca Margaretha Kuiper

geboren op 11 januari 1982
te Groningen

Promotor: Prof.dr. H. J. A. Hoijtink

The research in this dissertation has been funded by the Netherlands Organization for Scientific Research (NWO-VICI-453-05-002).

Beoordelingscommissie: Prof. Dr. Paul Boelen
Prof. Dr. Cees Glas
Prof. Dr. Ludwig A. Hothorn
Prof. Dr. Klaas Sijtsma

Kuiper, Rebecca Margaretha
Model Selection Criteria: How to Evaluate Order Restrictions
Proefschrift Utrecht University. - Met lit. opg. - Met samenvatting in het Nederlands.
ISBN: 978-90-393-5715-6
Printed by ZuidamUithof Drukkerijen, Utrecht, the Netherlands. First print 2011.
Cover designed by R. M. Kuiper.

Copyright © 2011, R. M. Kuiper. All Rights Reserved.

Neither this book nor any part may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, microfilming, and recording, or by any information storage and retrieval system, without written permission of the author. Alle rechten voorbehouden. Niets uit deze uitgave mag worden verveelvuldigd, in enige vorm of op enige wijze, zonder voorafgaande schriftelijke toestemming van de auteur.

Contents

CHAPTER 1

Introduction

Researchers often have ideas about the ordering of model parameters. They frequently have one or more theories about the ordering of the group means, in analysis of variance (ANOVA) models, or about the ordering of coefficients corresponding to the predictors, in regression models. A researcher might have the expectation that the parameters exhibit an increasing trend: $\theta_1 \leq \ldots \leq \theta_k$, where θ_j is, for example, the mean of group j or the regression coefficient corresponding to predictor j, for $j = 1, \ldots, k$. These types of restrictions are called order restrictions or inequality constraints.

Although researchers have directional expectations about the parameters, they usually evaluate it in an exploratory manner. That is, all possible configurations of groups of parameters being equal are examined (or a subset of these possibilities often based on the ordering of the sample parameters). For example, in case of five parameters, five out of the 52 possible configurations are:

$$\begin{aligned}
&\theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5, \\
&\theta_1 = \theta_2 = \theta_3 = \theta_4,\ \theta_5 \\
&\theta_1 = \theta_2,\ \theta_3 = \theta_4 = \theta_5 \\
&\theta_1 = \theta_2,\ \theta_3 = \theta_4,\ \theta_5 \\
&\theta_1,\ \theta_2,\ \theta_3,\ \theta_4,\ \theta_5.
\end{aligned}$$

Thus, it is inspected which groups of parameters are equal and which are not, while the ordering of the (groups of) parameters is not investigated. Regardless the resulting configuration, it generally does not give insight into the hypotheses of interest, that is, the directional hypothesis. There exist, however, methods that can be used to evaluate order restrictions directly, the so-called confirmatory methods. Why are they not used then? Probably because little is known about these methods, most of them can only be applied to a limited set of models, and there is no software available to employ them. This dissertation provides insight in evaluating order restrictions with (confirmatory) model selection techniques: It compares exploratory and confirmatory methods, extends some of the (confirmatory) techniques, and offers software for each of the discussed methods.

This dissertation starts with providing examples of informative hypotheses, that is, hypotheses containing order restrictions, and gaining insight in the properties

of exploratory and confirmatory techniques applicable for ANOVA models. The confirmatory methods are compared to their explorative counterparts for both hypothesis testing and model selection techniques. This is done based on one data set and on multiple data sets (i.e., on a simulation study) in case the three ANOVA assumptions (i.e., normality, independence, and equal variances) are met. Little is known regarding the robustness of the confirmatory methods for violations of one of the assumptions. Therefore, the performance of the confirmatory methods is also inspected when the equal variance assumption, also known as the homogeneity of variance assumption, is not met.

One of the two inspected confirmatory model selection techniques is the order-restricted information criterion (ORIC) of Anraku (1999). The ORIC can only be used for simple order restrictions (i.e., $\theta_1 \leq \ldots \leq \theta_k$, where "$\leq$" may be replaced by "$=$") in ANOVA models. Hence, it is extended in this dissertation to an information criterion that can be applied to a more general form of order restrictions in ANOVA models, the so-called generalized order-restricted information criterion (GORIC). In addition, the GORIC is modified such that it can be employed in multivariate normal linear models. These derivations assume a large sample size. Therefore, a small-sample version of the GORIC is constructed as well. Subsequently, some remaining issues, like the GORIC weights, are discussed.

Another problem researchers often face is missing data. There are methods and software available for handling missing data when estimating parameter values. But, what should be done in model selection using information criteria (ICs)? Since little to nothing is known about handling missing data in both exploratory and confirmatory model selection employing ICs, this dissertation starts with exploring it for the first type. The key issue is the model that is assumed to be the underlying data model. Subsequently, we discuss remaining issues, like *how do today's software programs calculate ICs in the presence of missing data* and *how should missing data be dealt with in confirmatory model selection using ICs*.

Besides using model selection based on ICs, one can use Bayesian model selection (BMS) to evaluate order restrictions. Two advantages of BMS are its ability to quantify evidence for the hypothesis of interest and to consider prior knowledge regarding the hypothesis of interest. These two feature can be helpful in combining the results from several studies concerning the same research question. This dissertation proposes a Bayesian updating method which combines statistical evidence for the hypotheses of interest from multiple studies. It should be stressed that there exist methods to combine the results of several studies (e.g., meta-analysis and prior updating). Nevertheless, these require (among others) that all the studies contain the same variables, whereas the proposed Bayesian updating method only requires that the variables of interest measure the same concept in all studies.

Additionally, software applications are provided for each of the discussed techniques, which are available from my web page:

`http://staff.fss.uu.nl/RMKuiper`.

In this dissertation, the software for the three confirmatory techniques applicable to ANOVA models and the software regarding the GORIC are discussed extensively.

As might be clear, a variety of subjects regarding model selection is discussed in this dissertation. The outline is given next.

Outline

Part I compares the exploratory and confirmatory methods in ANOVA models. In both exploration and confirmation, three types of methods are distinguished: hypothesis testing, model selection using information criteria, and Bayesian model selection. The properties of these methods are examined based on one data set in Chapter 2 and based on multiple data sets in Chapter 3. Moreover, Chapter 3 inspects the performance of the confirmatory methods for the violation of the homogeneity of variance assumption.

Part II focusses on confirmatory model selection criteria. Here, the ORIC is extended to the GORIC, a model selection criterion that can be used to evaluate a more general form of restrictions. In Chapter 4, this is done for ANOVA models and, in Chapter 5, for univariate and multivariate normal linear models. In Chapter 6, the small-sample version of the GORIC is derived. Some remaining issues are discussed in Chapter 7.

Part III again concentrates on model selection criteria, but now in the presence of missing data. Chapter 8 reports on how information criteria, like the Akaike information criterion (AIC) of Akaike (1973), should be calculated in the presence of missing data. Some remaining issues are discussed in Chapter 9.

Part IV concerns BMS. Chapter 10 discusses how the evidence from multiple studies regarding one concept can be combined.

Although software is made for all mentioned techniques, Part V covers the description of some of them. Chapter 12 reports on the software for the three confirmatory methods in ANOVA models and Chapter 13 on that for the GORIC in multivariate regression models.

Part I

Comparisons of Means: Exploration versus Confirmation

Drei wichter by Marga Klungel

CHAPTER 2

Comparisons of Means Using Exploratory and Confirmatory Approaches

Kuiper, R. M., and Hoijtink, H.

Published in Psychological Methods, 15(1), pp. 69-86.

This chapter discusses comparisons of means using exploratory and confirmatory approaches. Three methods are discussed: hypothesis testing, model selection based on information criteria, and Bayesian model selection. Throughout the chapter, an example is used to illustrate and evaluate the two approaches and the three methods. We demonstrate that confirmatory hypothesis testing techniques have more power, that is, have a higher probability of rejecting a false null hypothesis, and confirmatory model selection techniques have a higher probability of choosing the correct or the best hypothesis than their exploratory counterparts. Furthermore, we show that, if more than one hypothesis has to be evaluated, model selection has advantages over hypothesis testing. Another, more elaborate example is used to further illustrate confirmatory model selection. This chapter concludes with recommendations: When a researcher is able to specify reasonable expectations and hypotheses, confirmatory model selection should be used; otherwise, exploratory model selection should be used.

2.1 Introduction

Researchers are often confronted with the question *Do any of the mean responses differ from the others, and if so, which pairs of means are different from each other?*. For example, Palmer and Gough (2007), examined whether there is a difference in the attribution of importance of defective education as an explanation for criminal behavior between three types of "offenders": person offenders, property offenders, and non-offenders. The higher the rating, the lower one rates the importance of defective education and the less one considers defective education to be an explanation for criminal behavior. Table 2.1 presents the descriptive statistics for this example. we use this simple example to describe and illustrate the exploratory and confirmatory methods for comparisons of means. A more elaborate example is discussed at the end of this chapter.

Table 2.1: *Number of Observations (n_i), Sample Means ($\overline{y}_i$), and Sample Standard Deviations (sd_i) of the Importance of Defective Education in Explaining Criminal Behavior for Group i*

i offender type	n_i	$\overline{y}_i$	sd_i
1 person	20	11.95	4.42
2 property	20	9.75	3.78
3 non	31	8.77	3.07

The model used to answer questions about differences in means is the analysis of variance (ANOVA) model:

$$y_{ij} = \mu_i + \epsilon_{ij}, \tag{2.1}$$

where y_{ij} is the jth observation ($j = 1, \ldots, n_i$) of the dependent variable for Group i ($i = 1, \ldots, k$), μ_i is the mean of Group i, and ϵ_{ij} is the error term. The error terms are independently and normally distributed, with expected value 0 and variance σ^2, that is, $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

There are two different approaches for comparisons of means, namely exploration and confirmation. In exploration, the researcher has no theory or expectation about which pairs or subsets of means are equal. Therefore, all the possible configurations of means are inspected. In the example, where $k = 3$, five hypotheses are under investigation:

$$\begin{aligned} H_{E0}&: \ \mu_1 = \mu_2 = \mu_3, \\ H_{E1}&: \ \mu_1 = \mu_2, \ \mu_3, \\ H_{E2}&: \ \mu_1, \ \mu_2 = \mu_3, \\ H_{E3}&: \ \mu_1 = \mu_3, \ \mu_2, \\ H_{EA}&: \ \mu_1, \ \mu_2, \ \mu_3, \end{aligned} \tag{2.2}$$

where "$\mu_i = \mu_{i'}$", respectively,"$\mu_i, \ \mu_{i'}$" denotes that μ_i and $\mu_{i'}$ (for $i, i' = 1, 2, 3$) are equal, respectively, not restricted. In confirmation, the researcher has one or more theories or expectations with respect to the ordering of the means. In the example, the hypotheses of interest with respect to defective education are:

$$\begin{aligned} H_{C0}&: \ \mu_1 = \mu_2 = \mu_3 \\ H_{C1}&: \ \mu_1 > \mu_2 > \mu_3 \\ H_{CA}&: \ \mu_1, \ \mu_2, \ \mu_3, \end{aligned} \tag{2.3}$$

where "$\mu_i > \mu_{i'}$" denotes that μ_i is larger than $\mu_{i'}$ (for $i, i' = 1, 2, 3$). Hypothesis H_{C1} is based on the expectations of Palmer and Gough (2007) and on previous research. Palmer and Gough expected that the person and property offenders would be less likely to rate defective education as being important for explaining crime (against person or property) than the non-offenders (i.e., $\mu_1 > \mu_3$ and $\mu_2 > \mu_3$). According to previous research, crime against property (e.g., burglary) is likely to be attributed to defective education. Therefore, it is expected that person offenders attribute

less importance to defective education than property offenders (i.e., $\mu_1 > \mu_2$). When these expectations are combined, this leads to the hypothesis $H_{C1} : \mu_1 > \mu_2 > \mu_3$. The evaluation of the order-restricted hypothesis H_{C1} requires competing hypotheses. Here, the traditional null (H_{C0}) and alternative (H_{CA}) hypotheses are the competitors.

Within exploration and confirmation, three different types of methods can be distinguished: hypothesis testing, model selection, and Bayesian model selection. Therefore, six different types of techniques for comparisons of means (see Table 2.2) are introduced and evaluated in this chapter.

There are several exploratory hypothesis testing techniques. We look at the Shaffer-Welch Fq (SWFq) test, since it is the most powerful test in exploratory hypothesis testing, when controlling α for all comparisons (Ramsey, 2002; Toothaker, 1993, pp. 42–43, 48). Thus, the SWFq test has the highest probability of rejecting a null hypothesis H_0 that is false while controlling the familywise error rate which is defined as $P(\text{at least one false rejection of } H_0 : \mu_i = \mu_{i'} \text{ (for all } i, i' = 1, \ldots, k)) = \alpha$. However, the SWFq test assumes an equal number of observations per group (i.e., $n_i = n$ for $i = 1, \ldots, k$). According to Toothaker (1993, pp. 60), the Tukey-Kramer (TK) test is a popular procedure for unequal group observations. The TK test is a modification of the t statistic, a simple technique, and it maintains the α control for all comparisons. Therefore, we look at these two exploratory hypothesis testing techniques, although they might not be as familiar as some other techniques. Note that other techniques, for example, the t test, the Scheffé test, and Fisher's least significant difference (Toothaker, 1993, pp. 12–13 and 49, 34 and 51, and 41, respectively), are expected to perform similarly. The two most familiar information criteria are probably the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Another is Risanen's information criterion (RIC). When these are used in a classical way, all possible configurations of means are inspected. For $k = 3$, these are summed up in (2.2). According to Dayton (1998, 2003); Neath and Cavanaugh (2006), the performance of the information criteria can be improved by looking at all the possible configurations of the ordered means. When $k = 3$ and $\overline{y}_1 > \overline{y}_2 > \overline{y}_3$, as in the example (see Table 2.1), one then does not look at H_{E3} in (2.2). That is why we look at this exploratory model selection technique, which is called the paired-comparison information criterion (PCIC). Note that the PCIC can be based on any information criterion. We use the AIC, BIC, and RIC, because Dayton (2003) uses these three. In Bayesian model selection (BMS), one can use the posterior model probabilities. Note that BMS is confirmative in essence. It can, however, also be used in an exploratory way. Furthermore, we use the $\bar{F}$ test as confirmatory hypothesis testing technique and the order-restricted information criterion (ORIC) as a confirmatory model selection technique based on an information criterion. As far as we know, these are the only options.

All methods are based on the ANOVA model. The assumptions of the ANOVA model are (a) the dependent variable must be normally distributed for each group, (b) the population variances are equal for each group, and (c) the observations are independent. We assume that these assumptions are met in the examples used in this chapter. We return to this in the Discussion. In the next section, examples of order-restricted hypotheses encountered in the psychological literature are presented.

.

Table 2.2: *The Six Techniques for Testing or Comparing Hypotheses (With References)*

Technique	Exploration	Confirmation
Hypothesis testing	Equal n_i: the Shaffer-Welch Fq test (Toothaker, 1993, pp. 42–43) Unequal n_i: the Tukey-Kramer test (Toothaker, 1993, pp. 60–61)	The F test (Silvapulle & Sen, 2005, pp. 25–42)
Model selection	Paired-comparison information criterion (Dayton, 1998, 2003)	Order-restricted information criterion (Anraku, 1999)
Bayesian model selection	Posterior model probabilities (Klugkist, Laudy, & Hoijtink, 2005)	Posterior model probabilities (Klugkist, Laudy, & Hoijtink, 2005)

Note. n_i is the number of observations for Group i.

2.2 Examples of Order-Restricted Hypotheses

In addition to Palmer and Gough (2007), this section presents four other examples of order-restricted hypotheses. At the end of this chapter another example is given, based upon Gupta, Turban, and Bhawe (2008).

Lievens and Sanchez (2007) investigated the effect of training on the quality of ratings made by consultants. One variable of interest is the signal detection accuracy index, which "refers to the extent to which individuals were accurate in discerning essential from nonessential competencies for a given job" (Lievens & Sanchez, 2007, p. 817). They distinguish three groups of consultants, namely the expert (1), the training (2), and the control (3) group. The authors expected that accuracy of competency ratings would be higher among experts and trained raters than among raters in the control group (i.e., $\mu_1 > \mu_3$ and $\mu_2 > \mu_3$) and furthermore, that it would be highest among raters who already had competency modeling experience (i.e., $\mu_1 > \mu_2$). These expectations can be represented by the hypothesis $H_1 : \mu_1 > \mu_2 > \mu_3$.

Wiener, Holtje, Winter, Cantone, and Gross (2007) examined the purchasing decisions made during a simulated online shopping trip. One of the variables of interest is the likelihood-to-buy. Their design contained two factors. The factor disclosure has two levels: the enhanced (e) disclosure condition and the unenhanced (u) disclosure condition. In the enhanced disclosure condition, details were given about "credit loan agreements, including interest rates, payment amounts, and repayment time" (Wiener et al., 2007, p. 35); in the unenhanced disclosure condition, no details were given. The factor anticipated emotion has four levels that were manipulated by the researchers: in the pleasant purchase condition (p) persons were manipulated to expect a pleasant feeling after a buy and an unpleasant feeling after a nonbuy; in the unpleasant purchase condition (u) persons were manipulated to expect an unpleasant feeling after a buy and a pleasant feeling after a nonbuy; in the neutral purchase condition (n) persons were manipulated to feel neutral after a buy or a nonbuy; and in the control purchase condition (c) the participants simply purchased products without anticipated emotion manipulations. Wiener et al. expected that in both disclosure conditions the pleasant-to-buy group would be more likely to make purchases than the unpleasant-to-buy group (i.e., $\mu_{ep} > \mu_{eu}$ and $\mu_{up} > \mu_{uu}$). Furthermore, they expected that disclosure would be influential only for the control and neutral purchase conditions, where people with "enhanced disclosure information should buy less than would those without enhanced disclosure" (Wiener et al., 2007, p. 35); $\mu_{ec} > \mu_{uc}$ and $\mu_{en} > \mu_{un}$. The latter implies that no difference is expected in the likelihood-to-buy between the people in the enhanced and unenhanced disclosure groups for both the pleasant-to-buy and unpleasant-to-buy conditions (i.e., $\mu_{ep} = \mu_{up}$ and $\mu_{eu} = \mu_{uu}$). These expectations can be represented by one hypothesis, namely

$$\begin{aligned} H_1 : \ & \mu_{ec} > \mu_{uc}, \\ & \mu_{en} > \mu_{un}, \\ & \mu_{ep} = \mu_{up} > \mu_{eu} = \mu_{uu} \end{aligned}$$

Hasel and Kassin (2009) examined whether a confession alters the identification decisions of eyewitnesses and their confidence in those decisions. The variable of interest is the confidence rating for their identification. The participants were asked

to give a confidence rating after identification (Phase a). Then the participants were randomly assigned to four conditions: The participant was told that the identified suspect had confessed (Condition 1), all suspects had denied involvement (Condition 2), the identified suspect had denied involvement (Condition 3), or another person confessed (Condition 4). Hereafter, the participants were asked to give another confidence rating for their identification (Phase b). The authors expected that in Condition 1, where the identified suspect confessed, their confidence would increase (i.e., $\mu_{b1} > \mu_{a1}$); that in the other three conditions their confidence would decrease (i.e, $\mu_{b2} < \mu_{a2}$ and $\mu_{b3} < \mu_{a3}$ and $\mu_{b4} < \mu_{a4}$); and that the confidence ratings in Phase b would be in decreasing order from Condition 1 to 4 (i.e, $\mu_{b1} < \mu_{b2} < \mu_{b3} < \mu_{b4}$). These expectations can be represented by one hypothesis:

$$\begin{aligned} H_1 : \ & \mu_{b1} > \mu_{a1}, \\ & \mu_{b2} < \mu_{a2}, \\ & \mu_{b3} < \mu_{a3}, \\ & \mu_{b4} < \mu_{a4}, \\ & \mu_{b1} < \mu_{b2} < \mu_{b3} < \mu_{b4} \end{aligned}$$

Lucas (2003) investigated the difference between female and male leadership. The variable of interest is the influence of the leader, which is measured by the number of times (out of a total of 10) the subject switches to the leader's answer. Five experimental groups were distinguished: a group with a randomly selected male leader (Group 1), a group with a randomly selected female leader (Group 2), a group where the male team member who scored highest on a previous task is selected as leader (Group 3), a group where the female team member who scores highest on a previous task is selected as leader (Group 4) and a group in which female leadership is institutionalized and the female team member who scored highest on a previous task is selected as leader (Group 5). The institutionalization was done by showing the participants a film in which it was normal to have female leadership and women did well as leaders. The hypotheses of Lucas (2003) are that male leaders (Groups 1 and 3) exert more influence over participants than female leaders in the same leader selection method (Groups 2 and 4, respectively); that is, $\mu_1 > \mu_2$ and $\mu_3 > \mu_4$. Leaders appointed on the basis of their ability (Groups 3 and 4) exerted more influence over participants than leaders of the same sex appointed randomly (Groups 1 and 2, respectively); that is, $\mu_3 > \mu_1$ and $\mu_4 > \mu_2$. Institutionalized female leaders selected on the basis of their ability (Group 5) exerted more influence over participants than "normal" female leaders selected on the basis of their ability (Group 4) and randomly selected female leaders (Group 1); that is, $\mu_3 > \mu_1$ and $\mu_3 > \mu_4$, which can be written (for ease of notation in the resulting H_1) as $\mu_3 > \{\mu_1, \mu_4\}$. Institutionalized female leaders selected on the basis of their ability (Group 5) exerted the same amount of influence over participants as male leaders appointed on the basis of their ability (Group 3); that is, $\mu_5 = \mu_3$. These expectations can be represented by the hypothesis $H_1 : \mu_5 = \mu_3 > \{\mu_1, \mu_4\} > \mu_2$.

To investigate the hypotheses of interest, software can be helpful. For PCIC only the source code for the programming language GAUSS is made available in an .rtf file (Dayton, 2001). Software for BMS is available, see Klugkist, Laudy, and

Hoijtink (2005). The $\bar{F}$ test and the ORIC are not available in any software program. Therefore, these techniques, the SWFq test, and the TK test are implemented in *ComparisonOfMeans.exe*. A zip file with the software *ComparisonOfMeans.exe*, the appropriate text files, and a tutorial can be found at `http://staff.fss.uu.nl/RMKuiper` and `www.fss.uu.nl/ms/informativehypotheses`. This zip file also includes data, input, and output files of the examples of Palmer and Gough (2007), Gupta et al. (2008), and Lucas (2003).

In the next sections, we describe the three exploratory and three confirmatory methods and illustrate them using the example concerning the importance of defective education in explaining criminal behavior, along with input and output of *ComparisonOfMeans.exe*. Subsequently, we make two comparisons: hypothesis testing versus model selection using information criteria and Bayesian model selection, and exploration versus confirmation. The Appendix contains technical notes for the interested reader who wants to learn more about the technical nature of some of the aspects of each technique.

2.3 Exploration

2.3.1 Hypothesis Testing Using the SWFq Test

Goal: Find significant pairwise differences.

Procedure: Do an overall F test followed by the Welsch (1977) step-down procedure (Ramsey, 2002; Toothaker, 1993, pp. 42–43). According to Shaffer (1979), the addition of the overall F test increases the power of a step-down procedure. This procedure is elaborated and illustrated below. Note that in Toothaker (1993, pp. 42–43) the technique is called Shaffer-Ryan.

Illustration and Interpretation of the Results: The decision steps of the SWFq test are illustrated using the defective education data of Palmer and Gough (2007). The SWFq test requires equal group sizes. To be able to apply it to the data in Table 2.1, we generated data with 20 observations for the third group (i.e., the non-offenders) with the same mean and standard deviation as the original 31 cases.

Let n_i, the number of observations in Group i, equal n for all i and let N be the total number of observations.

1. **Overall F test:** Do an overall F test. If F is significant, proceed with SWFq; otherwise, stop and fail to reject $H_0 : \mu_1 = \ldots = \mu_k$.
   In the example, an $F(k-1=2, N-k=57)$ of 3.68 renders a significant p value of 0.03.
2. **Order means (if F is significant) and determine pairwise differences:** Order the means in ascending order. Calculate the pairwise differences (i.e., $\bar{y}_i - \bar{y}_{i'}$ for $i, i' = 1, \ldots, k$, where $\bar{y}_i$ is the sample mean of Group i) of the ordered means (see Table 2.3).
3. **Select the largest pairwise difference and determine the stretch size:** Select from all the pairwise differences that have not yet been evaluated the largest

pairwise difference $\bar{y}_i - \bar{y}_{i'}$. In case of a tie, select the one with the largest stretch size.

Determine the stretch size s of $\bar{y}_i - \bar{y}_{i'}$. This is the number of means located between $\bar{y}_i$ and $\bar{y}_{i'}$ plus two (for $\bar{y}_i$ and $\bar{y}_{i'}$). One may obtain s by first assigning the numbers 1 to k to the k ordered means such that the numbers indicate the ranking in the ordering and then by calculating

$$s = \text{ ranking of } \bar{y}_i - \text{ ranking of } y_{i'} + 1.$$

In the comparison of the largest and smallest mean, $s = k$. When comparing two adjacent mean, $s = 2$.

In the example, the largest pairwise difference is 3.18 (see Table 2.3), with a stretch size equal to 3.

4. **Test $H_0 : \mu_i = \mu_{i'}$ (accounting for multiple comparisons):** Reject $H_0 : \mu_i = \mu_{i'}$ when

$$|t_{\bar{y}_i - \bar{y}_{i'}}| = \left| \frac{\bar{y}_i - \bar{y}_{i'}}{\sqrt{MS_W \frac{2}{n}}} \right| \geq t_{crit}^{SWFq} = \frac{q_{df_s, df_W}^{\alpha_s}}{\sqrt{2}}, \tag{2.4}$$

were MS_W is the within-group mean square of the whole design; that is,

$$MS_W = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{N - k},$$

and $q_{df_s, df_W}^{\alpha_s}$ is the critical value of the studentized range distribution with a significance level of

$$\alpha_s = \begin{cases} \alpha, & \text{for } s = k, k-1; \\ 1 - (1-\alpha)^{s/k}, & \text{for } s \leq k-2, \end{cases}$$

where α is the nominal significance level, often set to .05, the error degrees of freedom $df_W = N - k$, and the "stretch size" degrees of freedom

$$df_s = \begin{cases} k-1, & \text{for } s = k, k-1; \\ s, & \text{for } s \leq k-2. \end{cases}$$

In the example, $|t_{\bar{y}_3 - \bar{y}_1}|$ is 2.65 (see Table 2.4). The critical value t_{crit}^{SWFq} is 2.00. This implies that $H_0 : \mu_3 = \mu_1$ is rejected. Note that only for $k = 2$ and $k = 3$ the critical values are the same for all stretch sizes (i.e., when $k = 3$, for $s = 2$ and $s = 3$). For $k > 3$, they are decreasing with stretch size.

5. **Stopping rule:** When $\bar{y}_i - \bar{y}_{i'}$ is not significant, the subsets of this nonsignificant set are also nonsignificant. For example, when $k = 3$ and $\bar{y}_1 - \bar{y}_3$ is not significant, the nonsignificant set consists of the Groups 3, 2 and 1, denoted by $\{3, 2, 1\}$. Then, the subsets are $\{3, 2\}$ and $\{2, 1\}$. In other words, when made into a table like Table 2.3, the pairwise differences below and/or to the left of this nonsignificant pairwise difference are nonsignificant too. Note that these subsets / pairwise differences do not have to be evaluated any more. Return to Step 3 and continue with the largest pairwise difference that has not yet been tested (directly or indirectly). Proceed this way until all pairwise differences are evaluated (directly or indirectly). Table 2.4 shows that for the example at hand, only $\mu_1 = \mu_3$ is rejected.

Table 2.3: *Ordered Means and Pairwise Differences ($\bar{y}_i - \bar{y}_{i'}$ for $i, i' = 1, 2, 3$) Between the Ordered Means for Defective Education*

Group no. (i)	Ordered means $\bar{y}_i$	$\bar{y}_3 - \bar{y}_i$	$\bar{y}_2 - \bar{y}_i$	$\bar{y}_1 - \bar{y}_i$
3	8.77		0.98	3.18
2	9.75			2.20
1	11.95			

Table 2.4: *SWFq Test for Defective Education When $n_3 = 20$ (Nominal $\alpha = .05$)*

Pairwise diff. ($\bar{y}_i - \bar{y}_{i'}$)	i & i'	s	$\lvert t_{\bar{y}_i - \bar{y}_{i'}} \rvert$	t_{crit}^{SWFq}	Significant
3.18	3 & 1	3	2.65	2.00	yes
2.20	2 & 1	2	1.83	2.00	no
0.98	3 & 2	2	0.82	2.00	no

Note. s = stretch size.

Input and Output of ComparisonOfMeans: Note that, in this case, we have three groups with each 20 observations. The data, consisting of two columns, must be given in *Data.txt*. The group number (i) must be given in the first column, and the corresponding data (y_{ij}) must be given in the second column. (More details can be found in the tutorial of ComparisonOfMeans.exe.) When performing the SWFq test, lines 1 to 3 can be ignored, and lines 4 to 9 in *Input.txt* should look as follows:

```
Number of groups
3
Number of observations in the groups
20 20 20
Perform: SWFandTK, PCIC, ExplBMS, Fbar, ORIC, ConfBMS (1 = yes, 0 = no)
1 0 0 0 0 0
```

All the other lines are not of interest here, but should not be deleted.

In the first lines of the output file, *Output.txt*, there are some remarks regarding the software, followed by the summary of the data. In case of the above mentioned input, the output of the overall F test and the SWFq test is as follows. (Bold numbers change with the data set.)

```
– Overall F test –

The probability that an F( 2, 57) variate is greater than 3.68 is 0.03,
thus reject H0 for all comparisons (at a nominal alpha level of 0.05)

– SWFq –

The following pairs of means are, according to Shaffer-Welch-test, significant
 different (with 3 groups and 57 degrees of freedom):
```

Pairwise difference between mean **3** and **1** is significant
(significance level = **0.01**, alpha level = 0.05)

Strengths and Weaknesses: A drawback of the SWFq test is that the SWFq test can render inconsistent results, as can be seen in Table 2.4. It is logically impossible that $\mu_1 \neq \mu_3$, $\mu_1 = \mu_2$, and $\mu_2 = \mu_3$, because the latter two imply $\mu_1 = \mu_3$. Another weakness is that when the number of groups (k) and therefore the number of hypotheses increases, the probability of choosing an incorrect hypothesis increases.

2.3.2 Hypothesis Testing Using the TK Test

Goal: Find significant pairwise differences.

Procedure: In the TK test (Tukey, 1953; Kramer, 1956, 1957; Toothaker, 1993, pp. 60–61) a test statistic is calculated for every pairwise difference $\bar{y}_i - \bar{y}_{i'}$ $(i, i' = 1, \ldots, k)$. Accounting for multiple comparisons, they are evaluated using the studentized range distribution:

$H_0 : \mu_i = \mu_{i'}$ is rejected if $|t_{\bar{y}_i - \bar{y}_{i'}}| = \left| \frac{\bar{y}_i - \bar{y}_{i'}}{\sqrt{MS_W(\frac{1}{n_i} + \frac{1}{n_{i'}})}} \right| \geq t^{TK}_{crit} = \frac{q^{\alpha}_{k, df_W}}{\sqrt{2}}$.

Illustration and Interpretation of the Results: In Table 2.5, the significant and nonsignificant differences for defective education are shown. The same conclusions are obtained as for the SWFq test. Note that the SWFq test is based on a data set with equal ns and the TK test on the original data set.

Table 2.5: *TK Test for Defective Education (Nominal α = .05)*

i & i'	Pairwise diff. $(\bar{y}_i - \bar{y}_{i'})$	$\lvert t_{\bar{y}_i - \bar{y}_{i'}} \rvert$	t^{TK}_{crit}	Significant
1 & 2	2.20	1.89	2.40	no
1 & 3	3.18	3.01	2.40	yes
2 & 3	0.98	0.93	2.40	no

Input and Output of ComparisonOfMeans: The input for the TK test is the same as for the SWFq test, only in this case the number of observation for Group 3 is 31 and not 20:

Number of observations in the groups
20 20 31

In the example, the output of the TK test looks like this (bold numbers change with the data set):

– TK –

The following pairs of means are, according to Tukey-Kramer-test, significant different (with **3** groups and **68** degrees of freedom):

Pairwise difference between mean **1** and **3** is significant (significance level = **0.01**)

Strengths and Weaknesses: The TK test has the same drawbacks as the SWFq test.

2.3.3 Model Selection Using the PCIC

Goal: Select the best fitting model/hypothesis of a set of hypotheses. The set of hypotheses consists of 2^{k-1} distinct patterns of subsets of ordered means.

Procedure: Calculate $\log L_m$ and PT_m:

1. $\log L_m$, that is, the log likelihood for hypothesis H_m, is calculated by:

$$\log L_m(\hat{\mu}_{m1}, \ldots, \hat{\mu}_{mk}, \hat{\sigma}^2_m | y) = \\ -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\hat{\sigma}^2_m) - \frac{1}{2\hat{\sigma}^2_m}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \hat{\mu}_{mi})^2, \tag{2.5}$$

   where $\hat{\sigma}^2_m = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \hat{\mu}_{mi})^2$, $N = \sum_i n_i$, and $\hat{\mu}_{m1}, \ldots, \hat{\mu}_{mk}$ are the values for which (2.5) is maximized subject to the restrictions in hypothesis H_m. So, the restricted means, that is, $\hat{\mu}_{m1}, \ldots, \hat{\mu}_{mk}$, are in accordance with the hypotheses of interest. For example, for hypothesis H_{E0} in (2.2), it holds that $\hat{\mu}_{01} = \hat{\mu}_{02} = \hat{\mu}_{03} = \overline{\mu}_0$, where $\overline{\mu}_0 = \frac{\sum_{i=1}^{3} n_i \overline{y}_i}{\sum_{i=1}^{3} n_i}$ is the overall sample mean. For hypothesis H_{E1} in (2.2), it holds that $\hat{\mu}_{01} = \hat{\mu}_{02} = \frac{n_1\overline{y}_1 + n_2\overline{y}_2}{n_1+n_2}$ and $\hat{\mu}_{03} = \overline{y}_3$. Furthermore, for hypothesis H_{EA} in (2.2), it holds that $\hat{\mu}_{01} = \overline{y}_1$, $\hat{\mu}_{03} = \overline{y}_3$ and $\hat{\mu}_{02} = \overline{y}_2$.
   If the restricted mean values of H_m are equal to the sample mean values, like in hypothesis H_{EA}, H_m has the highest log likelihood value (see Table 2.6). The larger the difference between the restricted means and the sample means, the lower the value of the log likelihood. For example, the restrictions in hypothesis H_{E2} are not in agreement with the sample means and the restrictions in H_{E0} even less, therefore, $\log L_2 = -190.13 > \log L_0 = -194.09$ (see Table 2.6).
2. PT_m, that is, the penalty term for hypothesis H_m, is the number of distinct means in H_m (a_m) plus 1 for the residual variance σ^2, that is, $PT_m = a_m + 1$.

The (log) likelihood represents the fit of the hypothesis with respect to the data at hand. However, the hypothesis with the highest (log) likelihood is not necessarily the best hypothesis. In determining the best hypothesis, the size of the hypothesis also needs to be taken into account. This is represented by the penalty term. As in Ockham's razor, the penalty term can be seen as an implementation of a simple-hypothesis-is-preferred principle. Note that the size of the hypothesis is

Table 2.6: *PCIC of the* $2^2 = 4$ *Exploratory Hypotheses* (H_{Em}) *for Defective Education*

Ordered means		8.77	9.75	11.95					
Group nr. (i)		3	2	1					
		Restricted means					PCIC based on		
m	H_{Em}	$\hat{\mu}_{m3}$	$\hat{\mu}_{m2}$	$\hat{\mu}_{m1}$	PT_m	$\log L_m$	AIC_m	BIC_m	RIC_m
0	$\mu_1 = \mu_2 = \mu_3$	9.94	9.94	9.94	2	-196.36	396.71	401.24	394.94
1	$\mu_1 = \mu_2, \mu_3$	8.77	10.85	10.85	3	-193.70	393.41	400.19	390.74
2	$\mu_1, \mu_2 = \mu_3$	9.15	9.15	11.95	3	-192.34	**390.68**	**397.46**	**388.01**
A	μ_1, μ_2, μ_3	8.77	9.75	11.95	4	-191.89	391.79	400.84	388.01

Note. PCIC = paired-comparison information criterion; AIC = Akaike information criterion; BIC = Bayesian information criterion; RIC = Risanen's information criterion. Bolding indicates the lowest value in each column.

sometimes referred to as the complexity of the hypothesis. Therefore, the information criteria are determined by the log likelihood and the penalty. As mentioned, we will look at the PCIC (Dayton, 1998, 2003) based on the AIC, the BIC, and the RIC:

$$
\begin{aligned}
AIC_m &= -2 \log L_m + 2(a_m + 1), \\
BIC_m &= -2 \log L_m + \log(N)(a_m + 1), \\
RIC_m &= -2 \log L_m + \log\left(\frac{N+2}{24}\right)(a_m + 1).
\end{aligned}
$$

The hypothesis with the smallest value for an information criterion is the preferred hypothesis. This hypothesis is preferred because it best uses the information in the data. In other words, it has the optimal balance between the fit (i.e., $\log L_m$) and the size (i.e., $PT_m = a_m + 1$) of the hypothesis.

Illustration and Interpretation of the Results: There are four hypotheses that can be constructed using the three ordered means of the defective education data: namely H_{E0}, H_{E1}, H_{E2}, and H_{EA} from (2.2). The AIC, BIC, and RIC values for these hypotheses among others are given in Table 2.6.

The log likelihood of hypothesis H_{EA} (i.e., $m = A$) is the highest (see Table 2.6), because for this hypothesis the restricted means (i.e., the $\hat{\mu}_{Ai}$) are equal to the sample means. However, H_{EA} has a complexity factor of four, because it contains three distinct means plus one for the residual variance σ^2. In this illustration, hypothesis $H_{E2} : \mu_1,\ \mu_2 = \mu_3$, is the preferred hypothesis (according to all three information criteria).

Input and Output of ComparisonOfMeans: The input for the PCIC is analogous to the input of the TK test. For the example, the output of the PCIC is given below. In the lines concerning group structure, a number represents a group index. So, "1 2 3" represents $\mu_1,\ \mu_2,\ \mu_3$, "1 1 1" represents $\mu_1 = \mu_2 = \mu_3$, and "1 2 2" represents $\mu_1,\ \mu_2 = \mu_3$. (Bold numbers change with the data set.)

– PCIC –

The group-structure and the values of the information criteria for the three best hypotheses, according to the Paired-Comparisons Information Criterion.

AIC = -2 * log likelihood + 2 * penalty
BIC = -2 * log likelihood + log(N) * penalty
RIC = -2 * log likelihood + log((N+2)/24) * penalty

According to the AIC:

Group-structure of the preferred hypothesis: **1 2 2**
Group-structure of the second best hypothesis: **1 2 3**
Group-structure of the third best hypothesis: **1 1 2**

AIC of the preferred hypothesis = -2 * **-192.34** + 2 * **3** = **390.68**
AIC of the second best hypothesis = -2 * **-191.89** + 2 * **4** = **391.79**
AIC of the third best hypothesis = -2 * **-193.70** + 2 * **3** = **393.41**

According to the BIC:

Group-structure of the preferred hypothesis: **1 2 2**
Group-structure of the second best hypothesis: **1 1 2**
Group-structure of the third best hypothesis: **1 2 3**

BIC of the preferred hypothesis = -2 * **-192.34** + log(N) * **3** = **397.46**
BIC of the second best hypothesis = -2 * **-193.70** + log(N) * **3** = **393.41**
BIC of the third best hypothesis = -2 * **-191.89** + log(N) * **4** = **391.79**

According to the RIC:

Group-structure of the preferred hypothesis: **1 2 2**
Group-structure of the second best hypothesis: **1 2 3**
Group-structure of the third best hypothesis: **1 1 2**

RIC of the preferred hypothesis = -2 * **-192.34** + log((N+2)/24) * **3** = **388.01**
RIC of the second best hypothesis = -2 * **-191.89** + log((N+2)/24) * **4** = **388.24**
RIC of the third best hypothesis = -2 * **-193.70** + log((N+2)/24) * **3** = **390.74**

Note that the results of the PCIC is consistent with the results of the TK test.

Strengths and Weaknesses: The strength of the PCIC is that it, like all model selection techniques, always gives consistent results. A weakness of the PCIC is that when the number of groups (k) and therefore the number of hypotheses (2^{k-1}) increases, the probability of selecting the best hypothesis decreases. Because in model selection there is no null hypothesis, we do not call this probability power but the probability of choosing the best hypothesis.

2.3.4 Model Selection Using Exploratory BMS

Goal: Select the best fitting model/hypothesis of a set of hypotheses. In exploratory BMS, like when using the PCIC, the set of hypotheses consists of 2^{k-1} distinct patterns of subsets based on means ordered from smallest to largest.

Procedure: Calculate the marginal likelihood of a hypothesis. In order to do this, the likelihood with respect to the parameters $\mu_1, \ldots, \mu_k$, σ^2, and the prior distribution of these parameters are needed. The log likelihood is given in (2.5).

Prior Distribution. The prior distribution reflects prior knowledge with respect to the means $(\mu_1, \ldots, \mu_k)$ and the residual variance (σ^2). So, the (order) restrictions of H_m are taken into account in the prior (see the technical note "The Prior" in the Appendix for an elaboration). Klugkist, Laudy, and Hoijtink (2005) used the same and mutually independent prior distribution for each mean. They and others (Chen & Sungduk, 2008; Johnson, 2005; Rossell, Baladandayuthapani, & Johnson, 2008) showed that such a prior does not favor any of the hypotheses under investigation. Furthermore, they showed that this prior has good properties when the goal is to select the best of a set of order-restricted hypotheses.

The prior knowledge of each mean is represented by a normal distribution, with a data-based mean β_0 and data-based variance τ_0^2. The prior distribution of σ^2 is a scaled inverse chi-squared distribution, with hyperparameters ν_0 and κ_0^2, where ν_0 is the degrees of freedom parameter and κ_0^2 is the scale parameter. (For an elaboration, see the technical note "Data-based Hyperparameters" in the Appendix.) The prior is chosen such that it has a minimal impact on the results. This is done by choosing a prior which is not only vague (such that it has low impact), but also compatible with the data (such that it is not too vague (Klugkist, Laudy, & Hoijtink, 2005)).

The mean β_0 and the variance τ_0^2 depend not only on the data but also on a user-specified term that reflects the vagueness of the prior (PV). Three degrees of vagueness are used here: PV equal to 1, 2, and 3, where $PV = 3$ renders the vaguest prior.

Marginal Likelihood. The marginal likelihood for hypothesis H_m is a measure of the degree of support for hypothesis H_m provided by the data (Klugkist, Laudy, & Hoijtink, 2005). It is equal to the integral of the likelihood over the prior distribution for the hypothesis at hand, H_m (see the technical note "The Marginal Likelihood" in the Appendix for an elaboration). For an elaboration of the interpretation and calculation of the marginal likelihood, we refer the reader to Klugkist (2008) and Klugkist, Laudy, and Hoijtink (2005).

The marginal likelihood quantifies the support in the data for the hypothesis at hand (H_m) accounting for the fit and complexity/size of H_m. In this way, the marginal likelihood resembles information criteria, like the PCIC based on the AIC, which can be written as -2 log likelihood + 2 penalty. That is, -2 times the log of the marginal likelihood can be written as -2 log likelihood + 2 penalty (see the technical note "The Marginal Likelihood" in the Appendix), where the penalty equals -1 times the log of the ratio of the prior density and the posterior density. Note that the marginal likelihood is higher for a hypothesis with a better fit and/or a lower complexity.

To interpret several marginal likelihoods at once, it can be helpful to transform them into posterior model probabilities (PMPs). A PMP is the probability that, given the data, the corresponding hypothesis is the best of the set of hypotheses (assuming a priori that all the hypotheses have equal probabilities of being the best; the interested reader is referred to the technical note "Posterior Model Probability" in the Appendix). The hypothesis with the highest PMP is the preferred hypothesis, that is, the best hypothesis of the set of hypotheses.

Illustration and Interpretation of the Results: In exploratory BMS, the same hypotheses are compared as in PCIC. In Table 2.7, the PMPs (for each hypothesis and each prior vagueness) are shown for the defective education data. The preferred hypothesis (for each prior vagueness) is hypothesis $H_{E2} : \mu_1,\ \mu_2 = \mu_3$.

Note that when PV increases, the support for hypotheses with equality constraints ("="), that is, H_{E0}, H_{E1}, and H_{E2}, relative to the support for the unconstrained model (i.e., H_{EA}) increases. However, the differences in PMP values are not that large for the different PV values. This implies that the results are robust with respect to the specification of the prior distribution. The interested reader is referred to Hoijtink, Huntjens, Reijntjes, Kuiper, and Boelen (2008), Klugkist, Laudy, and Hoijtink (2005), and Klugkist, Kato, and Hoijtink (2005) for further elaboration.

Table 2.7: *Posterior Model Probabilities (PMP) of the* $2^2 = 4$ *Exploratory Hypotheses* (H_{Em}) *for Defective Education for Three Types of Prior Vagueness (PV)*

		PMP		
m	H_{Em}	$PV = 1$	$PV = 2$	$PV = 3$
0	$\mu_1 = \mu_2 = \mu_3$	.04	.06	.06
1	$\mu_1 = \mu_2,\ \mu_3$	.13	.16	.17
2	$\mu_1,\ \mu_2 = \mu_3$	**.58**	**.55**	**.62**
A	$\mu_1,\ \mu_2,\ \mu_3$	.25	.23	.16

Note. Bolding indicates the highest value in each column.

Input and Output of ComparisonOfMeans: The input for exploratory BMS is analogous to the input for the TK test. However, in case of BMS, two additional specifications are needed: the desired δ and the prior vagueness PV. Specify $\delta = 0$ for an "exact equality" (i.e., $\mu_1 = \ldots = \mu_k$, that is, $|\mu_i - \mu_{i'}| = \delta = 0$ for all $i, i' = 1, \ldots, k$) and any positive number (i.e., $\delta > 0$) for an "about equality" (i.e., $\mu_1 \approx \ldots \approx \mu_k$, that is, $|\mu_i - \mu_{i'}| < \delta$ for all $i, i' = 1, \ldots, k$). In the latter case, one must specify δ in a reasonable way, for example, by looking at previous studies and/or asking experts in the field. The default recommendation for PV is $PV = 2$, but any positive number may be specified. When setting $\delta = 0$ and $PV = 3$, the first lines of *Input.txt* should look as follows:

```
When BMS is performed, an interval for equality relations (delta) is needed and a
 parameter for prior vagueness (pv).
When BMS is not performed, nothing needs to be filled in (but do not delete these
 lines).
0.0 3.0
```

In the example, the output of exploratory BMS looks like the following. (Bold numbers change with the data set.)

```
– exploratory BMS –

The preferred hypothesis, according to exploratory Bayesian model selection, has
 the following group-structure:
 1 2 2
The resulting Bayes factor (BF) value (of this hypothesis versus the unconstrained
 hypothesis) and the posterior model probabilities (PMP) of this hypothesis with
 respect to the whole set of hypotheses:
PMP
0.62
```

As in PCIC, "1 2 2" represents $\mu_1,\ \mu_2 = \mu_3$.

Strengths and Weaknesses: The strength of BMS is that it, like all model selection techniques, always gives consistent results. A weakness of BMS is that it is a time consuming technique. In addition, if a hypothesis contains an equality constraint (i.e., "="), BMS can be sensitive to the choice of the prior. However, as shown in Table 2.7, for these choices of PV this sensitivity usually does not lead to a different evaluation of the hypotheses under investigation.

2.4 Confirmation

2.4.1 Hypothesis Testing Using the $\bar{F}$ Statistic

In standard statistical testing, the hypothesis *all means are equal* (i.e, $H_0 : \mu_1 = \ldots = \mu_k$) is tested against the alternative *not all means are equal* (i.e., $H_A : \mu_1, \ldots, \mu_k$). This is usually tested with an F test using a one-way ANOVA. However, researchers often want to test a certain order restriction because of a theory or expectation with respect to the order of the means in the experiment. For example, it is expected that non-offenders are more likely to rate defective education as being important for explaining crime than the other two offenders, and property offenders are more likely to rate it as being important than person offenders (see H_{C1} in (2.3)). Such an order-restricted hypothesis can be tested with the $\bar{F}$ test.

Goal: Evaluate the null hypothesis. When testing an *ordered alternative* (i.e., test *all means are equal* against an order-restricted hypothesis), the traditional null hypothesis is evaluated. When testing an *ordered null* (i.e., test an order-restricted

hypothesis against *all parameters are free*), the order-restricted hypothesis is evaluated. In summary, in the ordered alternative, $H_0 : \mu_1 = \ldots = \mu_k$ is tested against $H_m : \mu_i - \mu_{i'} \geq 0$ (for some $i, i' = 1, \ldots, k$) and, in the ordered null, $H_m : \mu_i - \mu_{i'} \geq 0$ (for some $i, i' = 1, \ldots, k$) is tested against $H_A : \mu_1, \ldots, \mu_k$.

Procedure: Calculate the value of the $\bar{F}$ statistic (Silvapulle & Sen, 2005, pp. 25–42). Like the classical F test, the $\bar{F}$ test is based on the residual sum of squares (RSS) for the tested null distribution (i.e., the classical null or an order-restricted hypothesis) and the tested alternative (i.e., an order-restricted hypothesis or the classical alternative, respectively). For the classical null (H_0) and the classical alternative (H_A), the RSS are determined with respect to the overall mean ($\bar{y}$) and sample means ($\bar{y}_i$), respectively. Note that these are the values for which the RSS is minimized given that the values are in accordance with the hypothesis at hand (i.e., H_0 and H_A, respectively). The analogue is done for an (order-restricted) hypothesis, H_m. The values that minimize the RSS given that these are in accordance with H_m are called the restricted means. (For an elaboration see the technical note "The $\bar{F}$ Statistic" in the Appendix.) So, the order restrictions are taken into account in one of the RSS (depending on the type of $\bar{F}$ test).

As with classical hypothesis testing, p values must be determined. Because of the order restrictions, this is done via simulation (Silvapulle & Sen, 2005, pp. 32–33 and 40; for an elaborate see the technical note "Calculation of the p Value of the $\bar{F}$ Statistic" in the Appendix).

Illustration and Interpretation of the Results: Table 2.8 shows that $H_{C0} : \mu_1 = \mu_2 = \mu_3$ is rejected (for $\alpha = 0.05$) when it is tested against both $H_{CA} : \mu_1,\ \mu_2,\ \mu_3$ and $H_{C1} : \mu_1 > \mu_2 > \mu_3$ and that H_{C1} is not rejected when it is tested against H_{CA}. So, H_{CA} is preferred over H_{C0} and H_{C1} is preferred over both H_{C0} and H_{CA}. Therefore, H_{C1} is the preferred hypotheses.

Table 2.8: $\bar{F}$ *Test of the Specified Hypotheses for Defective Education*

Hypotheses tested	$\bar{F}$	p
H_{C0} against H_{CA}	9.11	.01
H_{C0} against H_{C1}	9.11	.00
H_{C1} against H_{CA}	0.00	1.00

Input and Output of ComparisonOfMeans: For all three confirmatory techniques (i.e., the $\bar{F}$ test, ORIC, and confirmatory BMS), all hypotheses of interest must be given explicitly. This must be done in *Input.txt* from line 10 on. In the example, the hypotheses of interest are the set of hypotheses specified in (2.3).

In order to write down the hypotheses of interest in the input file, some notation must be introduced. The way the hypotheses are written down in the input file has to do with the hypothesized group numbers and the inequality constraints ">" and "<".

For ease, the first mean is said to be in the first group. When looking at $H_{C0} : \mu_1 = \mu_2 = \mu_3$, all means are hypothesized to be in the same group. Therefore, H_{C0} is written down as "1 1 1". When looking at $H_{CA} : \mu_1,\ \mu_2,\ \mu_3$, all means are hypothesized to be in another group. Because there are no restrictions on the means, H_{CA} is written down as "1 2 3". When looking at $H_{C1} : \mu_1 > \mu_2 > \mu_3$, all means are also hypothesized to be in another group, but now there are restrictions on the means. The sign ">" is represented by a "-3". Therefore, $H_{C1} : \mu_1 > \mu_2 > \mu_3$ is represented by "1 -3 -3", meaning that the first mean is bigger than the second, which is bigger than the third one. In the example, the ordering of the group numbers for all three hypotheses is "1 2 3" and all three hypotheses are written down in one line/restriction. (More details on this can be found in the tutorial of *ComparisonOfMeans.exe*.)

When performing the $\bar{F}$ test, the first half of *Input.txt* is analogous to the input of the TK test, and the second half, from line 10 on, should look as follows:

```
In case of Fbar test and/or ORIC and/or ConfBMS:
Number of hypotheses to be compared
3
Number of restrictions per hypothesis
1
1
1
Ordering of means in restriction
1 2 3
1 2 3
1 2 3
(Order) Restrictions
1  1  1
1 -3 -3
0  0  0
```

The output of the $\bar{F}$ test in the example looks like this (bold numbers change with the data set and/or set of hypotheses):

```
– Fbar test –

Results of the Fbar test for the null hypothesis 1 and the unconstrained hypothesis
 3
Hypotheses numbers              Fbar value      p-value
        1 versus  3               9.11           0.01

Results of the "ordered alternative" Fbar test
Ordered-hypothesis number       Fbar value      p-value
        H0 versus  2              9.11           0.00

Results of the "ordered null" Fbar test
Ordered-hypothesis number       Fbar value      p-value
        2 versus Ha               0.00           1.00
```

Strengths and Weaknesses: A disadvantage of the $\bar{F}$ test is that it can only test one order-restricted hypothesis at a time. In this case, the $\bar{F}$ test results in three p values (see Table 2.8), which can be evaluated straightforwardly and will give a consistent result. When a researcher wants to test more than one order-restricted hypothesis, say H_1 and H_2, problems arise. Then the $\bar{F}$ tests results in five p values: one resulting from the classical F test, two from the ordered alternative for both H_1 and H_2, and two from the ordered null for both H_1 and H_2. No straightforward, nonarbitrary rules exist for the evaluation of these five p values. Furthermore, these p values cannot be used for a direct comparison of the two order-restricted hypotheses H_1 and H_2. Therefore, the $\bar{F}$ test is most useful when one order-restricted hypothesis has to be evaluated.

2.5 Model Selection Using the ORIC

Goal: Select the best fitting model/hypothesis of a set of hypotheses. The ORIC inspects a limited set of well defined hypotheses constructed using one or more restrictions of the form $\mu_i - \mu_{i'} \geq 0$ for $i, i' = 1, \ldots, k$.

Procedure: Calculate the ORIC (Anraku, 1999) for hypothesis H_m by

$$ORIC_m = -2 \log L_m + 2\, PT_m, \tag{2.6}$$

where

1. $\log L_m$ is calculated by (2.5).
   Note that in the ORIC the order restrictions are taken into account in the (order-restricted) likelihood. Namely, the (order-restricted) likelihood is based on means that maximize the likelihood (like the maximum likelihood estimates), only theses estimates are in accordance with the restrictions in H_m. We will refer to these means as the restricted means. Note that the restricted means $(\hat{\mu}_{m1}, \ldots, \hat{\mu}_{mk})$ are the same as the restricted means in the $\bar{F}$ test (see the technical note "The Restricted Means" in the Appendix).
2. PT_m is calculated by:

$$PT_m = 1 + \sum_{l=1}^{a_m} LP_{ml} \cdot l, \tag{2.7}$$

   where LP_{ml} is the level probability for hypothesis H_m, that is, the a priori probability that there are l distinct mean values among $\hat{\mu}_{m1}, \ldots, \hat{\mu}_{mk}$ (which are in agreement with H_m), and a_m equals the number of distinct μ_is in the restrictions of H_m. For example, in $H_1 : \mu_1 > \mu_2 > \mu_3$, $a_1 = 3$ and, in $H_2 : \mu_1 > \mu_2 = \mu_3$, $a_2 = 2$. The computation of the level probabilities can be done via simulation (Silvapulle & Sen, 2005, pp. 78–81). For a description of this simulation, see the technical note "Calculation of the Level Probabilities" in the Appendix.

The order restrictions are not only taken into account in the (order-restricted) likelihood, but also in the penalty term. The penalty term can be seen as the expected number of distinct mean values plus 1 (because of the unknown variance term). When there are no order restrictions, this penalty reduces to the penalty of the PCIC, that is, $PT_m = a_m + 1$, and ORIC reduces to the AIC.

The hypothesis with the smallest value for the ORIC is the preferred hypothesis (as in the PCIC).

Illustration and Interpretation of the Results of the ORIC: In Table 2.9, the restricted means ($\hat{\mu}_{mi}$), the log likelihood values ($\log L_m$), the penalty terms (PT_m), and the ORIC values are given for the three hypotheses for defective education.

Since the sample means are in accordance with the restrictions in both H_{C1} and H_{CA}, the restricted means (i.e., the $\hat{\mu}_{mi}$'s for $m = 1$ and $m = A$) are equal to the sample means. Therefore, hypotheses H_{C1} and H_{CA} have the highest log likelihood. However, H_{C1} is less complex than H_{CA} (i.e, $PT_1 < PT_A$), because the means in H_{CA} are not restricted. When looking at the optimal combination of fit (i.e., $\log L_m$) and size/complexity (i.e., PT_m) of the hypotheses, H_{C1} is the preferred hypothesis.

Table 2.9: *ORIC of the Three Specified Hypotheses (H_{Cm}) for Defective Education*

		Restricted means					
m	H_{Cm}	$\hat{\mu}_{m1}$	$\hat{\mu}_{m2}$	$\hat{\mu}_{m3}$	PT_m	$\log L_m$	$ORIC_m$
0	$\mu_1 = \mu_2 = \mu_3$	9.94	9.94	9.94	2.00	-196.36	396.71
1	$\mu_1 > \mu_2 > \mu_3$	11.95	9.75	8.77	2.82	-191.89	**389.42**
A	$\mu_1,\ \mu_2,\ \mu_3$	11.95	9.75	8.77	4.00	-191.89	391.79

Note. ORIC = order-restricted information criterion.
Bolding indicates the lowest value.

Input and Output of ComparisonOfMeans: The input for the ORIC is analogous to the input for the $\bar{F}$ test. In the example, the output of the ORIC looks like this (bold numbers change with the data set and/or set of hypotheses):

– ORIC –

The value of the Order-Restricted Information Criterion (ORIC) =
-2 * log likelihood + 2 * penalty:

for Hypothesis 1, ORIC = -2 * **-196.36** + 2 * **2.00** = **396.71**
for Hypothesis 2, ORIC = -2 * **-191.89** + 2 * **2.82** = **389.42**
for Hypothesis 3, ORIC = -2 * **-191.89** + 2 * **4.00** = **391.79**

The preferred hypothesis, according to the Order-Restricted Information Criterion,
of the hypotheses to be compared is hypothesis number **2**,
with the following ordering(s) of means:

1 2 3
and corresponding restriction(s):
1 -3 -3

Strengths and Weaknesses of the ORIC: As all model selection techniques, the ORIC always gives consistent results.

Most researchers are able to specify reasonable hypotheses, since they are expert in their research field. However, it is possible that the set of specified hypotheses does not contain a reasonable or good hypothesis. In that case, a model selection technique, like ORIC, will choose the best hypothesis of a set of weak hypotheses. To ensure that a weak hypothesis is not chosen, one can include the unconstrained hypothesis $H_A : \mu_1, \ \ldots, \ \mu_k$ in the set of hypotheses. H_A will always be preferred over a weak order-restricted hypothesis.

2.6 Model Selection Using Confirmatory BMS

Goal: Select the best fitting model/hypothesis of a set of hypotheses. Confirmatory BMS, like the ORIC, examines a limited set of well defined hypotheses constructed using one or more restrictions of the form $\mu_i - \mu_{i'} \geq 0$ for $i, i' = 1, \ldots, k$.

Procedure: Calculate the marginal likelihood of a hypothesis as is done in exploratory BMS. Note that, in BMS, the order restrictions are taken into account via the admissible space of the prior for hypothesis H_m (more details can be found in the technical note "The Prior" in the Appendix).

Illustration and Interpretation of the Results: In Table 2.10, the PMPs (for each prior vagueness) are given for the three hypotheses for defective education. The preferred hypothesis (for each prior vagueness) is $H_{C1} : \mu_1 > \mu_2 > \mu_3$. Note again the robustness of the inferences with respect to the vagueness of the prior distribution.

Table 2.10: *Posterior Model Probabilities (PMP) of the Three Specified Hypotheses (H_{Cm}) for Defective Education for Three Types of Prior Vagueness (PV)*

		PMP		
m	H_{Cm}	$PV = 1$	$PV = 2$	$PV = 3$
0	$H_{C0} : \mu_1 = \mu_2 = \mu_3$	.03	.03	.06
1	$H_{C1} : \mu_1 > \mu_2 > \mu_3$	**.80**	**.80**	**.78**
A	$H_{CA} : \mu_1, \ \mu_2, \ \mu_3$	.17	.17	.17

Note. Bolding indicates the highest value in each column.

Input and Output of ComparisonOfMeans: The input for confirmatory BMS is analogous to the input for the $\bar{F}$ test. Note that, in this case, as when using exploratory

BMS, δ and PV must be specified. In the example, with $\delta = 0$ and $PV = 3$, the output of confirmatory BMS looks like this (bold numbers change with the data set and/or set of hypotheses):

– confirmatory BMS –

The resulting posterior model probabilities (PMP) of the order-restricted hypotheses with respect to the whole set of hypotheses:

	PMP
Hypothesis 1	**0.07**
Hypothesis 2	**0.77**
Hypothesis 3	**0.16**

The preferred hypothesis, according to confirmatory Bayesian model selection, of the hypotheses to be compared is hypothesis number **2**,
with the following ordering(s) of means:
1 2 3
and corresponding restriction(s):
1 -3 -3

Strengths and Weaknesses: Confirmatory BMS gains and suffers from the same things as exploratory BMS. Also here, the illustration (see Table 2.10) shows that for reasonable choices of PV, the prior sensitivity does usually not lead to a different evaluation of the hypotheses. Furthermore, if a hypothesis contains only inequality constraints (i.e., "<" and/or ">"), the relative support of this hypothesis with respect to the unconstrained hypothesis shows that BMS is not sensitive to the choice of the prior.

As is the case when using the ORIC, the unconstrained hypothesis should be included in the set of hypotheses to protect against choosing a weak order-restricted hypothesis.

2.6.1 Conclusions with Respect to Defective Education

From the SWFq test (in case of equal group sizes) and the TK test, it follows that only the pairwise difference between Groups 3 and 1 is significant. Palmer and Gough (2007) concluded the same when performing an ANOVA F test and a Scheffé post hoc test. As mentioned before, this is a logically impossible result, because $\mu_1 = \mu_2$ and $\mu_2 = \mu_3$ would imply $\mu_1 = \mu_3$. When using the PCIC and exploratory BMS, it can be concluded that hypothesis $H_{E2} : \mu_1, \mu_2 = \mu_3$ is the preferred hypothesis (according to all three criteria and for each prior vagueness, respectively). When looking at the sample means, more can be said about the ordering of the means. However, when using an exploratory method, the hypotheses of interest, $\mu_1 > \mu_2 > \mu_3$, is not evaluated. From the three $\bar{F}$ tests (i.e., the classical F test, the ordered alternative, and the ordered null), we concluded that $H_{C1} : \mu_1 > \mu_2 > \mu_3$ is the preferred hypotheses. The same is concluded when using the ORIC or doing confirmatory BMS (for each prior vagueness).

2.7 Comparison of the Two Approaches and the Three Methods

2.7.1 The TK Test is Less Powerful than the SWFq Test

In case of equal group sizes, the TK test is less powerful than the SWFq test. Thus, the SWFq test has a higher probability of rejecting a false null hypothesis than the TK test has. This can be seen from the critical values for both tests (Toothaker, 1993, pp. 42–43, 48), because both tests are based on the same calculated value $|t_{\bar{y}_i - \bar{y}_{i'}}|$. In the example, in case of equal group sizes in defective education (i.e., $n_1 = n_2 = n_3 = 20$), the SWFq test has a critical value of 2.00 (see Table 2.4) which is lower than 2.41 (see Table 2.11), the critical value of the TK test. So, when $2.41 > |t_{\bar{y}_i - \bar{y}_{i'}}| > 2.00$ (for $i, i' = 1, 2, 3$), the SWFq test does reject the (false) null hypothesis and the TK test does not.

Table 2.11: *TK Test for Defective Education when* $n_3 = 20$ *(Nominal Alpha = 0.05)*

i & i'	Pairwise diff. $(\bar{y}_i - \bar{y}_{i'})$	$\lvert t_{\bar{y}_i - \bar{y}_{i'}} \rvert$	t^{TK}_{crit}	Sign.
3 & 1	3.18	2.65	2.41	yes
2 & 1	2.20	1.83	2.41	no
1 & 2	0.98	0.82	2.41	no

2.7.2 The $\bar{F}$ Test is More Powerful than the SWFq Test or the F Test

The advantage of the $\bar{F}$ test is that it has more power than the SWFq test, since it tests fewer hypotheses and it can test order-restricted hypotheses. The increase in power due to testing order-restricted hypotheses can be illustrated by comparing the ordered alternative $\bar{F}$ test with the classical F test.

For example, look at a generated data set with $k = 3$ groups, effect size ES, and n observations per group. When $ES = 0$, the sample means are the same and, when $ES = 0.1, 0.2, 0.3, 0.4$, or 0.5, the sample means are decreasing, that is, $\overline{y}_1 > \overline{y}_2 > \overline{y}_3$. Note that $ES = \frac{1}{\hat{\sigma}}\sqrt{\frac{1}{3}\sum_{i=1}^{3}(\overline{y}_i - \overline{y})^2}$ (Cohen, 1992), where $\overline{y} = \frac{1}{3}\sum_{i=1}^{3}\overline{y}_i$ and $\hat{\sigma}$ is set to 1. An effect size of $ES = 0.1$ is called small, $ES = 0.25$ medium, and $ES = 0.4$ large. For $ES = 0$, a sample is constructed that is perfect in agreement with $H_0 : \mu_1 = \mu_2 = \mu_3$, and for $ES > 0$, one that is perfect in agreement with $H_1 : \mu_1 > \mu_2 > \mu_3$. In the ordered alternative $\bar{F}$ test, $H_0 : \mu_1 = \mu_2 = \mu_3$ is tested against $H_1 : \mu_1 > \mu_2 > \mu_3$. Note that, in the F test, $H_0 : \mu_1 = \mu_2 = \mu_3$ is tested against $H_A : \mu_1, \mu_2, \mu_3$.

In Table 2.12, the p values of both tests are given for each generated data set. For $ES > 0$, the p values of the ordered alternative $\bar{F}$ are always lower than the p values of the F test. Thus, (the false) H_0 is rejected for an lower effect size when tested against H_1 than when tested against H_A. Therefore, the ordered alternative $\bar{F}$ test has more power than the F test (when the order-restricted hypothesis is true). Hence, testing order-restricted hypotheses increases power.

Table 2.12: *p values of the Ordered Alternative $\bar{F}$ Test ($p_{\bar{F}}$) and the F Test (p_F) for One Generated Data Set With $k = 3$ Groups, Effect Size ES, and n Observations per Group*

	n			
p value	10	20	50	100
		$ES = 0$		
p_F	1.00	1.00	1.00	1.00
$p_{\bar{F}}$	1.00	1.00	1.00	1.00
		$ES = 0.1$		
p_F	0.91	0.82	0.61	0.37
$p_{\bar{F}}$	0.48	0.40	0.26	0.14
		$ES = 0.2$		
p_F	0.6742	0.4543	0.1390	0.0193
$p_{\bar{F}}$	0.3022	0.1802	0.0470	0.0059
		$ES = 0.3$		
p_F	0.4184	0.1746	0.0127	0.0002
$p_{\bar{F}}$	0.1648	0.0599	0.0039	0.0000
		$ES = 0.4$		
p_F	0.220454	0.048182	0.000501	0.000006
$p_{\bar{F}}$	0.080770	0.015600	0.000110	0.000000
		$ES = 0.5$		
p_F	0.100899	0.009983	0.000011	0.000001
$p_{\bar{F}}$	0.034460	0.002880	0.000000	0.000000

What if another order-restricted hypothesis, say H_2, is true? In that case, if H_0 is tested against H_1, H_0 may or may not be rejected. However, if H_1 is tested against H_A, H_1 will be rejected (if the sample is large enough).

2.7.3 The ORIC is More "Powerful" than the PCIC

The ORIC has an higher probability of choosing the best hypothesis than the PCIC, because it evaluates fewer hypotheses and it can evaluate order-restricted hypotheses. We illustrate the latter by comparing the ORIC with the AIC, using the same data sets used to illustrate the gain in power when the $\bar{F}$ test is used instead of the F test.

The three hypotheses to be evaluated are:

$$\begin{aligned} H_0 &: \ \mu_1 = \mu_2 = \mu_3, \\ H_1 &: \ \mu_1 > \mu_2 > \mu_3, \\ H_A &: \ \mu_1, \mu_2, \mu_3. \end{aligned}$$

In Figure 2.1, the ORIC values are given for these three hypotheses (for $n = 10$ and $n = 100$). As mentioned before, the ORIC reduces to the AIC when there are no order restrictions, that is, for hypotheses H_0 and H_A. The ORIC values for H_0 increase with ES, because the differences between the sample means $\overline{y}_i$ and the restricted means

$\hat{\mu}_{0i} = \overline{\mu}_0$ (for all $i = 1, \ldots, k$) increase with increasing effect size. This leads to a decrease in the log likelihood (see (2.5)) and, consequently, an increase in ORIC value (see (2.6)). The ORIC values for H_1 and H_A do not depend on effect size, because the sample means are in accordance with H_1 and, logically, H_A. Thus, the difference between $\overline{y}_i$ and $\hat{\mu}_{mi}$ (for $i = 1, \ldots, k$) is zero for both H_1 and H_A for each effect size. This implies that the likelihood values for H_1 and H_A are equal; therefore, the difference in ORIC values equals two times the difference in the penalty terms, that is, $2(PT_2 - PT_1) = 2(4 - 2\frac{5}{6}) = 2\frac{1}{3}$. So, when the sample means are in accordance with H_1, as in the example, H_1 is always preferred over H_A. As can be seen in Figure 2.1, compared to H_A, H_1 will be preferred over H_0 for smaller effect sizes. This implies that the probability of choosing the correct/best hypothesis is higher if H_0 is compared to H_1 than if H_0 is compared to H_A (when H_1 is true).

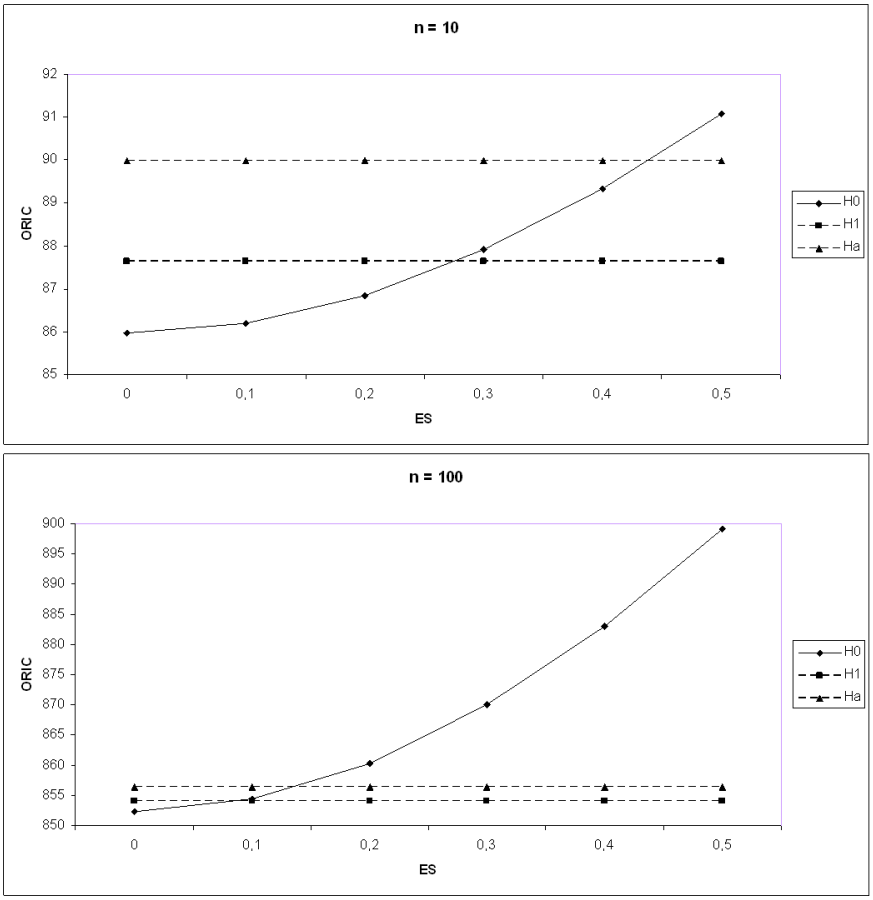


Fig. 2.1: Order-restricted information criterion (ORIC) values for one generated data set with $k = 3$ groups, effect size ES, and n observations per group.

In this example, the ideal case is represented, namely H_1 is true and H_1 is included in the set of hypotheses. What if another order-restricted hypothesis, say H_2, is true? In that case, H_A will be preferred over H_1 if the sample size is large enough to distinguish H_1 from H_2. Therefore, the unconstrained model H_A should always be included in the set of hypotheses. So, if the unconstrained hypothesis is included in the set, model selection techniques select the correct hypothesis or a similar one (i.e., a hypothesis that resembles the true hypothesis, that is, only differs in a few constraints) or otherwise the unconstrained hypothesis.

2.7.4 Exploratory BMS is more "Powerful" than Confirmatory BMS

Confirmatory BMS has the same advantages over exploratory BMS as the ORIC has over the AIC.

2.7.5 Conclusion of Comparisons

In confirmation there will be more power (i.e., a higher probability of rejecting a false null hypothesis) or a higher probability of choosing the best hypothesis than in exploration. This is due to the smaller number of hypotheses that have to evaluated and to the inclusion of expectations, in the form of order restrictions, in the hypotheses of interest (see Table 2.12 and Figure 2.1). But, the most important advantage is that the expectations of researchers (see the examples given in the beginning of this chapter) can be evaluated using a confirmatory approach. The weakness of confirmation is that a researcher must be able to specify his expectations. Note that the inclusion of the unconstrained hypothesis $H_A : \mu_1, \ldots, \mu_k$ is a safeguard against choosing a hypothesis that is weakly supported by the data.

Model selection, in contrast to hypothesis testing, always gives consistent results and can handle more than one (order-restricted) hypothesis at once.

2.8 Elaborate Example to Illustrate the Confirmatory Model Selection Techniques

The final example will be used to further illustrate the (preferred) confirmatory model selection techniques, that is, the ORIC and (confirmatory) BMS.

As mentioned in the introduction, the evaluation of the order-restricted hypothesis (like H_{C1}) requires competing hypotheses. Until now, the traditional null ($H_0 : \mu_1 = \ldots = \mu_k$) and alternative ($H_A : \mu_1, \ldots, \mu_k$) hypotheses are used. However, it is also possible to specify a set of hypotheses without the classical null and even include another order-restricted hypothesis. We suggest that the set of hypotheses should consist of one or two order-restricted hypotheses and the unconstrained hypotheses (as a safeguard). Namely, H_0 is often not of real interest, and sometimes there is more than one theories/expectations, as in the example described next.

Gupta et al. (2008) studied the impact of implicit and explicit activation of gender stereotypes on men's and women's intentions to pursue an entrepreneurship, a traditionally masculine career. The variable of interest is entrepreneurial intensions

Table 2.13: *Number of Observations (n_i), Sample Means ($\overline{y}_i$), and Sample Standard Deviations (sd_i) of Entrepreneurial Intentions for Group i*

i	n_i	$\overline{y}_i$	sd_i
1	38	3.44	1.01
2	46	2.94	1.07
3	36	3.48	1.08
4	37	2.66	1.09
5	33	2.93	1.05
6	39	2.43	1.00

(see Table 2.13), which is measured by the average of four items on a 5-point scale. Here, we look at 3 conditions (control, explicit masculine stereotype, and implicit masculine stereotype) for both men and women. In the control group, "participants read an article about entrepreneurship education that made no mention of gender or gender differences in entrepreneurship" (Gupta et al., 2008, p. 1055). In the other two groups, the participants read an article in which three masculine characteristics (i.e., aggressive, risk taking, and autonomous) are mentioned. In the implicit condition, the article simply described the three characteristics, whereas in the explicit condition, there was more emphasis on the characteristics. For example, the participants in the explicit masculine stereotype group were told that it pays to have masculine characteristics. Thus, we look at six groups (see Table 2.14).

According to Gupta et al. (2008), previous studies show that people tend to behave in a way similar to that predicted by the stereotype they are made aware of (Theory 1). However, Gupta et al. also stated that other recent evidence suggests that under certain circumstances, people may not assimilate with the stereotype, but respond in a way opposite to that predicted by the stereotype. The response depends on whether the stereotype is activated implicitly or explicitly: Implicit stereotype activation leads to behavior consistent with the stereotype, whereas explicit activation leads to behavior opposite to the stereotype (Theory 2).

Based on the above theories, two (main) expectations are distinguished. The first one is based on Theory 1 and the second one is based on the expectations of Gupta et al. (2008), which are based on Theory 2. Based on these two expectations, two hypotheses are formulated, namely H_1 and H_2, respectively. In H_1, it is hypothesized that men who are made aware of an explicit prevalent masculine stereotype (Group

Table 2.14: *The Six Groups in Entrepreneurial Intentions*

	Masculine stereotype		
Gender	Control	Explicit	Implicit
Male	1	2	3
Female	4	5	6

2) have the strongest entrepreneurial intentions with respect to the other same-sex groups (Groups 1 and 3). It is also hypothesized that women who are made aware of an explicit prevalent masculine stereotype (Group 5) have the weakest entrepreneurial intentions with respect to the other same-sex groups (Groups 4 and 6, respectively). So, it is hypothesized that $\mu_2 > \mu_1$, $\mu_2 > \mu_3$, $\mu_5 < \mu_4$, and $\mu_5 < \mu_6$, that is, $\mu_1 < \mu_2 > \mu_3$ and $\mu_4 > \mu_5 < \mu_6$. Furthermore, it is expected that, in the control condition, men (Group 1) have stronger entrepreneurial intentions than women (Group 4), that is, $\mu_1 > \mu_4$ (Gupta et al., 2008). Note that the expectations $\mu_2 > \mu_1$, $\mu_1 > \mu_4$, and $\mu_4 > \mu_5$ imply that $\mu_2 > \mu_5$. Therefore, it is also (indirectly) expected that in the explicit condition, men (Group 2) have stronger entrepreneurial intentions than women (Group 5), that is, $\mu_2 > \mu_5$. In the implicit condition (Groups 3 and 6), however, it is unclear aforehand whose entrepreneurial intentions will be stronger. In H_2, it is expected that "gender and stereotype activation will interact such that men will report stronger entrepreneurial intentions when presented with an implicit" (Group 3) versus an explicit (Group 2) "masculine stereotype whereas women will report stronger entrepreneurial intentions when presented with an explicit" (Group 5) versus an implicit (Group 6) masculine stereotype (Gupta et al., 2008, p. 1055), that is, $\mu_3 > \mu_2$ and $\mu_5 > \mu_6$, respectively. Furthermore, it is expected that, in the control condition, men (Group 1) will report stronger entrepreneurial intentions than women (Group 4), that is, $\mu_1 > \mu_4$. Thus, the set of hypotheses is:

$$\begin{array}{lccc} H_1: & \mu_1 & < \mu_2 & > \mu_3 \\ & \vee & \vee & , \\ & \mu_4 & > \mu_5 & < \mu_6 \end{array}$$

$$\begin{array}{lccc} H_2: & \mu_1 & , \quad \mu_2 & < \mu_3 \\ & \vee & , & , \\ & \mu_4 & , \quad \mu_5 & > \mu_6 \end{array}$$

$$H_A: \mu_1,\ \mu_2,\ \mu_3, \mu_4,\ \mu_5,\ \mu_6,$$

where the comma denotes that there is no restriction between the corresponding means. For example, in H_1 there is no restriction between the means of Group 3 and 6. Note that the unconstrained hypothesis (H_A) is included to protect against choosing the best of two weak order-restricted hypotheses. As mentioned before, when evaluating hypotheses H_1 and/or H_2, the pairwise relations will be evaluated simultaneously.

For the three hypotheses, the ORIC values are given in Table 2.15 and the PMPs in Table 2.16. These tables show that both ORIC and BMS prefer H_2, the hypothesis based on the expectations of Gupta et al. (2008). Furthermore, Table 2.16 shows that, if a hypothesis contains only inequality constraints (i.e., "<" and/or ">"), BMS is not sensitive to the choice of the prior.

To conclude, a direct evaluation of two competing hypotheses renders strong support for H_2. Since H_A was also included in the evaluation, H_2 is more than only the best of two weak hypotheses. This example further illustrates how ORIC and BMS can straightforwardly be used to evaluate the expectations of a researcher. All a researcher has to do is to specify expectations, collect data, and compute the ORIC or do BMS.

Table 2.15: *ORIC of the Three Specified Hypotheses (H_m) for Entrepreneurial Intentions*

m	H_m	*ORIC*
1	H_1	686.31
2	H_2	**677.14**
A	H_A	680.13

Note. ORIC = order-restricted information criterion. Bolding indicates the lowest value.

Table 2.16: *Posterior Model Probabilities (PMP) of the Three Specified Hypotheses (H_m) for Entrepreneurial Intentions (for Three Types of Prior Vagueness (PV))*

		PMP		
m	H_m	$PV = 1$	$PV = 2$	$PV = 3$
1	H_1	.00	.00	.00
2	H_2	**.89**	**.89**	**.89**
A	H_A	.11	.11	.11

Note. Bolding indicates the highest value in each column.

2.9 Discussion

2.9.1 Violations of the Model Assumptions

As mentioned in the introduction, all techniques are based on the ANOVA model, which has three assumptions. We first assumed that these assumptions were met. But, what if there are some violations of these assumptions? It is known that the ANOVA F test is robust with respect to nonnormality (Stevens, 1999, pp. 74–80; Toothaker, 1993, pp. 57–66) It is also robust against violations of equal variances, as long as the ratio of the largest group size versus the smallest group size is smaller than 1.5 (Stevens, 1999, pp. 75–76). If the large variances are associated with the small group sizes, the F tests rejects the null too often; that is, the actual α is bigger than the nominal α, which is often set to .05. In contrast, when the large variances are associated with the large group sizes, the actual α is smaller than the nominal α. The ANOVA model is, however, affected by the violation of independent observations. Even a small violation results in a substantial effect on the actual significance level and power of the F test. The same is expected for the other techniques, although currently no results are available on this topic. In case of dependent observations, one could use a multilevel model (Hox, 2002). The robustness of the techniques with respect to nonnormality and unequal variances are described next and are summarized in Table 2.17.

In general, most multiple comparison procedures seem to be robust to moderate violations of the normality assumption (Toothaker, 1993, pp. 64–66). Miller (1986) states that the studentized range test is a bit more sensitive to nonnormality than the

.

Table 2.17: *Robustness of Techniques Against Violations of Two ANOVA Assumptions*

	Violations of	
	Normality	Equal variances
ANOVA F	Robust to moderate violations.	Robust as long as $\frac{largest\ group\ size}{smallest\ group\ size} < 1.5$.
stud. range (q)	More sensitive to nonnormality than the F test.	Robust as long as $\frac{largest\ group\ size}{smallest\ group\ size} < 1.5$.
SWFq	Probably resembles q, but further research needed.	Probably resembles q, but further research needed.
TK	Probably resembles q, but further research needed.	Probably resembles q, but further research needed.
$\bar{F}$	Probably resembles F, but research needed.	Probably resembles F, but research needed.
PCIC	Seems to be quite robust; further research needed.	Negative impact on the correct identification rates.
ORIC	Might resemble PCIC, but research needed.	Might resemble PCIC, but research needed.
BMS	Might resemble PCIC, but research needed.	Might resemble PCIC, but research needed.

Note. ANOVA = analysis of variance; SWFq = Shaffer-Welch Fq test; TK = Tukey-Kramer test; PCIC = paired-comparison information criterion; ORIC = order-restricted information criterion; BMS = Bayesian model selection.

F test. However, as long as the number of observations per group are large enough, the studentized range statistic q should be approximately correct (according to the central limit theorem). Among others, Martin, Toothaker, and Nixon (1989) examined multiple comparison procedures with unequal variances. Their findings are the same as the findings for the ANOVA F test. Because the SWFq and TK test are multiple comparison procedures, which are based on the studentized range statistic, we expect that the above also holds for these two tests. However, further research is needed.

One could expect that the same holds for the $\bar{F}$ test as for the F test. However, as far as we know, there is no research about the violations of the ANOVA assumptions for the $\bar{F}$ test yet.

Some study has been done with respect to the robustness of the PCIC against nonnormality and heterogeneity (Dayton, 2003), but the PCIC has not been evaluated extensively. The PCIC seems to be quite robust for nonnormality. However, heterogeneity of variance had a negative impact on the correct identification rates in a simple simulation study. Note that there is a modification of the PCIC for heterogeneity of variance (see Dayton, 1998, 2003). Further research is needed for the PCIC.

One could expect that the same holds for the ORIC as for the PCIC. However, as far as we know, this is not been studied yet. The same holds for BMS. So, for the ORIC and BMS, research with respect to the violations of the assumptions is needed.

Check on Normality and Equal Variances in the Two Examples

The check on the assumption of normality and equal variances can be done by looking at the Shapiro-Wilk test and the Levene's test, respectively. For all data examples, the assumptions of normality (per group) and of equal variances are not violated, since the p values of the corresponding tests are larger than 0.05 (see Table 2.18). Thus, the results in this chapter are not influenced by violations of the assumptions.

Table 2.18: *p Values of the Tests Used for the Check on Normality (i.e., Shapiro-Wilk) and Equal Variances (i.e., Levene's) of the Two Data Examples (Defective Education and Entrepreneurial Intentions)*

		Defective education		
		$n_3 = 31$	$n_3 = 20$	Entrepreneurial intentions
Shapiro-Wilk test	Group 1	.255	.255	.210
	Group 2	.446	.446	.903
	Group 3	.423	.994	.729
	Group 4			.554
	Group 5			.989
	Group 6			.839
Levene's		.135	.186	.991

2.9.2 Final Conclusions and Recommendations

Hypothesis testing techniques can give inconsistent results. Some exploratory methods cannot handle unequal group sizes, and the confirmatory approach cannot handle multiple order-restricted hypotheses. Therefore, model selection is preferred over hypothesis testing.

Confirmatory model selection techniques have a higher probability of choosing the best hypothesis than their exploratory counterparts. Furthermore, in confirmatory model selection, a direct evaluation of competing hypotheses, that is, the expectations of a researcher, is possible. However, confirmatory model selection techniques can only be used when a researcher is able to formalize his expectations in order-restricted hypotheses. To ensure that confirmatory model selection does not result in preferring a weak order-restricted hypothesis, the unconstrained hypothesis $H_A : \mu_1, \ \ldots, \ \mu_k$ must be included in the set of hypotheses. Often the classical null hypothesis is not of interest. When that is the case, the classical null hypothesis should be left out of the set of hypotheses.

Because BMS needs a specification of δ and PV (see Section 2.3.4) and it is a time-consuming technique, we prefer the PCIC in exploration and the ORIC in confirmation.

2.A Technical Notes

2.A.1 The Prior

The prior for the unconstrained hypothesis (i.e., the traditional alternative hypotheses) $H_A : \mu_1, \ \ldots, \ \mu_k$ is defined as

$$p_A(\mu_1, \ldots, \mu_k, \sigma^2) = p(\mu_1) \times \ldots \times p(\mu_k) \times p(\sigma^2),$$

where $p(\mu_i) = p(\mu)$ for all i $(i = 1, \ldots, k)$. $p(\mu)$ is a data-based normal distribution with hyperparameters β_0 and τ_0^2 and $p(\sigma^2)$ an scaled inverse chi-squared distribution with hyperparameters ν_0 and κ_0^2.

The prior $p_m(.)$ for hypothesis H_m is (up to a normalizing constant) equal to this prior on the admissible space of H_m (and to zero outside this space). For example, in exploration, the admissible space of $H_E : \mu_1 = \mu_2$ are those combinations of means for which it holds that $|\mu_1 - \mu_2| = 0$. Note that with $|\mu_1 - \mu_2| < \delta$ an "about equality constraint" is inspected, if δ is set to a positive number. For example, in confirmation, the admissible space of $H_C : \mu_1 > \mu_2$ are those combinations of means for which it holds that $\mu_1 > \mu_2$. Here we see that, in BMS, the (order) restrictions are taken into account via the admissible space of the prior for hypothesis H_m. The admissible space covers a certain proportion of the total space, say $b\%$ (for $b \in [0, 100]$); then the prior density of H_m equals $\frac{100}{b}$ times the prior density of $H_A : \mu_1, \ \mu_2$ on this admissible space and equals zero outside this space. For H_C, $b = 50$, because it covers half of the total space.

2.A.2 Data-based Hyperparameters

For each μ_i $(i = 1, \ldots, k)$, a credibility interval is computed:

$$\overline{y}_i \pm \ PV \ \times \hat{\sigma}.$$

Three vague priors are used, namely the priors where PV is set to 1, 2, and 3. Note that if $PV = 3$, 2, and 1, the 99.7%, 95%, and 68% credibility intervals are computed, respectively. The credibility interval for μ_i has a lower bound LB_i and an upper bound UB_i. Let the lowest lower bound be LB_{min} (i.e., $LB_{min} = \min\{LB_1, \ldots, LB_k\}$) and the highest upper bound UB_{max}. Then, the data-based hyperparameters β_0 and τ_0^2 are calculated by:

$$\beta_0 = \frac{LB_{min} + UB_{max}}{2}, \text{ and}$$
$$\tau_0^2 = \left[\frac{UB_{max} - LB_{min}}{2}\right]^2,$$

respectively. The hyperparameter ν_0 is set to 1 and κ_0^2 to the estimate of the error variance.

2.A.3 The Marginal Likelihood (ML)

The marginal likelihood of hypothesis H_m is defined by

$$\mathrm{ML}(y|H_m) = \int \ldots \int \int L_m(\mu_1, \ldots, \mu_k, \sigma^2|y)\ p_m(\mu_1, \ldots, \mu_k, \sigma^2)\ \ d\mu_1 \ldots d\mu_k\ d\sigma^2.$$

The marginal likelihood can be rewritten as

$$\mathrm{ML}(y|H_m) = L_m(\mu_1, \ldots, \mu_k, \sigma^2|y)\frac{p_m(\mu_1, \ldots, \mu_k, \sigma^2)}{post_m(\mu_1, \ldots, \mu_k, \sigma^2|y)},$$

where $post_m(\mu_1, \ldots, \mu_k, \sigma^2|y)$ is the posterior density, for which it holds that

$$post_m(\mu_1, \ldots, \mu_k, \sigma^2|y) \propto L_m(\mu_1, \ldots, \mu_k, \sigma^2|y)p_m(\mu_1, \ldots, \mu_k, \sigma^2).$$

Therefore,

$$-2 \log \mathrm{ML}(y|H_m) = -2 \log\ L_m(\mu_1, \ldots, \mu_k, \sigma^2|y) + 2\ \left[-\log\left(\frac{p_m(\mu_1, \ldots, \mu_k, \sigma^2)}{post_m(\mu_1, \ldots, \mu_k, \sigma^2|y)}\right)\right].$$

2.A.4 Posterior Model Probability (PMP)

Assuming a priori that all the hypotheses have equal probabilities of being the best (i.e., $app_m = app_{m'}$ for $m, m' = 1, \ldots, M$, with M the total number of hypotheses), the posterior model probability of hypothesis $H_{m'}$ is defined as

$$PMP_{m'} = \frac{app_{m'}\mathrm{ML}(y|H_{m'})}{\sum_{m=1}^{M} app_m \mathrm{ML}(y|H_m)} = \frac{\mathrm{ML}(y|H_{m'})}{\sum_{m=1}^{M} \mathrm{ML}(y|H_m)}.$$

2.A.5 The $\bar{F}$ Statistic

The $\bar{F}$ statistic for testing H_{null} against H_{alt} is calculated by

$$\bar{F} = \frac{RSS(H_{null}) - RSS(H_{alt})}{S^2},$$

where S^2 is the mean square error:

$$S^2 = (n_1 + \cdots + n_k - k)^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

and $RSS(H)$ is the residual sum of squares with respect to hypothesis H,

in an ordered alternative (i.e., $H_{null} = H_0$ and $H_{alt} = H_m$)	in an ordered null (i.e., $H_{null} = H_m$ and $H_{alt} = H_A$)
$RSS(H_{null}) = \sum_i \sum_j (y_{ij} - \bar{y})^2$	$RSS(H_{null}) = \sum_i \sum_j (y_{ij} - \hat{\mu}_{mi})^2$
$RSS(H_{alt}) = \sum_i \sum_j (y_{ij} - \hat{\mu}_{mi})^2$	$RSS(H_{alt}) = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$,

with $\bar{y}$ as the overall mean, $\hat{\mu}_{m1}, \ldots, \hat{\mu}_{mk}$ as the restricted mean values, that is, the values of $\mu_1, \ldots, \mu_k$ that minimize $\sum_i \sum_j (y_{ij} - \mu_i)^2$, subject to $H_m : \mu_i - \mu_{i'} \geq 0$ for some $i, i' = 1, \ldots, k$, and $\bar{y}_i$ as the ith group mean. More details about the restricted means can be found in the next section. Note that the classical F test is based on $RSS(H_{null}) = \sum_i \sum_j (y_{ij} - \bar{y})^2$ and $RSS(H_{alt}) = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$.

2.A.6 The Restricted Means

As mentioned, the restricted means (i.e., $\hat{\mu}_{m1}, \ldots, \hat{\mu}_{mk}$) are in accordance with the hypotheses of interest, H_m. This was also the case with the restricted means used in the PCIC, only now H_m can contain order restrictions.

For $H_1 : \mu_1 > \cdots > \mu_k$, it holds that $\hat{\mu}_{11} \geq \ldots \geq \hat{\mu}_{1k}$ and their values depends on the (weighted) sample means. An example is given in Table 2.19. To elaborate on this, numerical examples for $k = 3$ and equal group sizes are given in Table 2.20.

Table 2.19: *The Restricted Mean Values for Different Sample Mean Values When* $H_1 : \mu_1 > \cdots > \mu_k$

	Corresponding restricted means				
Ordering of the sample means	$\hat{\mu}_{11}$	$\ldots$	$\hat{\mu}_{1k-2}$	$\hat{\mu}_{1k-1}$	$\hat{\mu}_{1k}$
$\overline{y}_1 \geq \cdots \geq \overline{y}_k$	$\overline{y}_1$	$\ldots$	$\overline{y}_{k-2}$	$\overline{y}_{k-1}$	$\overline{y}_k$
$\overline{y}_1 \geq \cdots \geq \overline{y}_{k-2}$ and $\overline{y}_{k-1} \leq \overline{y}_k$	$\overline{y}_1$	$\ldots$	$\overline{y}_{k-2}$	$\frac{n_{k-1}\overline{y}_{k-1} + n_k \overline{y}_k}{n_{k-1} + n_k}$	$\frac{n_{k-1}\overline{y}_{k-1} + n_k \overline{y}_k}{n_{k-1} + n_k}$

When the restrictions are simple order restrictions (i.e., the restrictions can be written as $\mu_1 > \cdots > \mu_{k'}$ for $k' \leq k$), the restricted means can be calculated by

Table 2.20: *Numerical Examples of the Restricted Mean Values for $k = 3$ and for Equal Group Sizes When $H_1 : \mu_1 > \cdots > \mu_k$*

Sample means			Corresponding restricted means		
$\overline{y}_1$	$\overline{y}_2$	$\overline{y}_3$	$\hat{\mu}_{11}$	$\hat{\mu}_{12}$	$\hat{\mu}_{13}$
2	1	2	2	1.5	1.5
1	2	1.4	1.5	1.5	1.4
1	2	2	$1\frac{2}{3}$	$1\frac{2}{3}$	$1\frac{2}{3}$

the pool adjacent violators algorithm (PAVA). For more details see Silvapulle and Sen (2005, pp. 47–50). Another method, which can be applied to all types of order restrictions, is quadratic programming. For more details see Silvapulle and Sen (2005, pp. 36–37).

2.A.7 Calculation of the p Value of the $\bar{F}$ Statistic

1. Generate independent observations z_{ij} $(i = 1, \ldots, k$ and $j = 1, \ldots, n_i)$ from the standard normal distribution $\mathcal{N}(0, 1)$.
2. Compute the value of the $\bar{F}$ statistic for the generated data.
3. Repeat the previous two steps R times. R is set to $R = 100,000$, because then we obtain stable estimates for the p values, that is, when we calculated the p value again, the difference was rarely larger than 0.001.
4. Determine the number of times the $\bar{F}$ statistic, calculated in Step 2, exceeds the sample value of the $\bar{F}$. Denoted this by C.
5. The p value is estimated by C/R.

2.A.8 Calculation of the Level Probabilities

1. Generate $z_1, \ldots, z_k$ from $\mathcal{N}_k(0, V)$, that is, the multivariate normal distribution, with mean 0 and the covariance matrix V, where V is a diagonal matrix with the elements $1/n_1, \ldots, 1/n_k$ on the diagonal.
2. Compute the restricted means $\hat{z}_{m1}, \ldots, \hat{z}_{mk}$ analogously to $\hat{\mu}_{m1}$, …, $\hat{\mu}_{mk}$. So, $\hat{z}_{m1}, \ldots, \hat{z}_{mk}$ are the values for which $L_m(\hat{z}_{m1}, \ldots, \hat{z}_{mk}, \hat{\sigma}^2_m |\ z_1, \ldots, z_k)$ (see (2.5)) is maximized subject to hypothesis H_m.
3. Determine the number of distinct values in $\hat{z}_{m1}, \ldots, \hat{z}_{mk}$, called levels. Denote this by $\tilde{l}_m$.
4. Repeat the previous steps R times. R is set to $R = 100,000$, because in that case we obtain stable estimates for LP_{ml}. That is, when we calculated it again, the difference was most of the time less than 0.02.
5. Estimate the level probability LP_{ml} by the proportion of times $\tilde{l}_m$ is equal to l.

Because the restricted means $\hat{z}_{m1}, \ldots, \hat{z}_{mk}$ are in accordance with H_m, the maximum value $\tilde{l}_m$ can take on is a_m. For example, for $H_2 : \mu_1 > \mu_2 = \mu_3$, it holds that $a_2 = 2$ and $\hat{z}_{21} \geq \hat{z}_{22} = \hat{z}_{23}$. When $\hat{z}_{21} = \hat{z}_{22} = \hat{z}_{23}$, there is one level, and when $\hat{z}_{21} > \hat{z}_{22} = \hat{z}_{23}$, there are two levels. Therefore, for $l = a_m + 1, \ldots, k$, $LP_{ml} = 0$.

CHAPTER *3*

Performance and Robustness of Confirmatory Approaches

Kuiper, R. M., Nederhoff, T., and Klugkist, I.

Manuscript submitted

In this chapter, the performance of three confirmatory comparisons of means methods is inspected. The three types are hypothesis testing, model selection using information criteria, and Bayesian model selection. A simulation study is conducted to evaluate the performance of the three methods. For comparison, the performance of their exploratory counterparts are also determined. We demonstrate that confirmatory analyses have more power than exploratory analyses and that model selection has advantages over hypothesis testing.

Little is known about the robustness of the different methods for violations of the assumptions, especially for confirmatory techniques. Therefore, we do another simulation study where we study the performance of the confirmatory methods when the homogeneity of variance assumption is violated. From this study, it can be concluded that the techniques are robust to heterogeneity when the sample sizes are equal. When the sample sizes are unequal, the performance is substantially affected by heterogeneity. However, the deviations from the baseline, where there is no heterogeneity, are not pronounced.

3.1 Introduction

A central issue in most research is to evaluate the researcher's theory. When comparing group means, the researcher often would like to know whether group means differ and, if so, which group means are different from each other. There are two approaches that can be used to address this question, namely exploration and confirmation. In exploration, all the possible configurations of subsets of means are inspected. The number of possible configurations increases rapidly with an increase in the number of groups k. For example, when $k = 3$ and $k = 5$, there are 5 and 52 possible configurations, respectively. In confirmation, researchers solely evaluate their theories or expectations, given that they can specify reasonable ones. This mostly results in a limited set of hypotheses. Besides the classical null *all means are equal* (H_0)

and the classical alternative *there are no restrictions* (H_A), one can include so-called order-restricted hypotheses which represent a certain ordering of the means (e.g., the group means are increasing with k). Both in exploration and confirmation, different methods can be used: hypothesis testing, model selection using information criteria, and Bayesian model selection. These two distinctions lead to six different types of techniques that can be used to evaluate group means (see Table 3.1). The descriptions of the six techniques are summarized in the next section accompanied with an illustration. For each type, a detailed description is given in Kuiper and Hoijtink (2010).

Kuiper and Hoijtink (2010) show for one data set that confirmatory techniques perform better than their exploratory counterparts. In this chapter, we will quantify the performance of the three exploratory and three confirmatory approaches by means of simulation. Besides making the comparison between exploratory and confirmatory techniques, we compare the performance of the three types of methods: hypothesis testing, model selection using information criteria, and Bayesian model selection. Here, the performance of a method is measured by the percentage of times the correct hypothesis is chosen by this method. Moreover, little is known about the performance of confirmatory techniques when model assumptions are violated. To the authors knowledge, only Wesel, Hoijtink, and Klugkist (in press) examined this for Bayesian model selection. To gain more insight into the performance of all three confirmatory methods, we also study their robustness for the violation of the homogeneity of variance assumption.

In the next section, we describe the model used for comparing means and the assumptions of the model. Furthermore, we introduce an example based on Lucas (2003). In addition, we briefly demonstrate the six techniques that can be applied to comparing group means based on this example. Subsequently, the design and results of the two simulation study are described. We end with a discussion.

3.2 Preliminaries

3.2.1 Analysis of Variance Model

All methods which will be described next are based on the analysis of variance (ANOVA) model:

$$y_{ij} = \mu_i + \epsilon_{ij}, \tag{3.1}$$

where y_{ij} is the jth observation ($j = 1, \ldots, n_i$) of the dependent variable for group i ($i = 1, \ldots, k$), μ_i is the mean of group i, and ϵ_{ij} is the error term. The error terms are independent and normally distributed random variables, each with the expected value 0 and variance σ^2, that is, $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

The assumptions of the ANOVA model are (a) the dependent variable must be normally distributed conditional on each group, (b) the observations are independent, and (c) the population variances are equal for each group. The latter is also known as the homogeneity of variance assumption.

The effect of violating the homogeneity of variance assumption on the performance of the traditional ANOVA F test is studied several times (Schumacker & Akers, 2001;

.

Table 3.1: *The Six Techniques for Testing or Evaluating Hypotheses*

Technique	Exploration	Confirmation
Hypothesis testing	Equal n_i: the Shaffer-Welch Fq test (SWFq test)	The $\bar{F}$ test
	Unequal n_i: the Tukey-Kramer test (TK test)	
Model selection	Paired-comparison information criterion (PCIC)	Order-restricted information criterion (ORIC)
based on information criteria	For instance, PCIC-AIC	
Bayesian model selection	Posterior model probabilities (PMPs)	Posterior model probabilities (PMPs)

Note. n_i is the number of observations for group i.

Box, 1954). According to Box (1954), there is no profound effect on the Type I error (i.e., the error that the null hypothesis is incorrectly rejected) when the sample sizes are equal. When the groups sizes exhibit the same trend as the variances, the Type I error is lower than stated (usually .05); in case the groups sizes and the variances have an opposite trend, the Type I error is higher than stated. One can use the ratio of the largest and smallest group variance, called the F_{max} statistic, to test for heterogeneity (Hartley, 1950). Tabachnick and Fidell (2001) conclude that an F_{max} value of 10 is acceptable for analyses with equal group sizes and $F_{max} = 3$ for unequal group sizes. However, Box (1954) shows that the F test is severely affected when $F_{max} = 3$ and the groups sizes and the variances exhibit opposite trends.

Notably, the robustness of exploratory methods has been studied. In the second part of this chapter, we will study the effects of homogeneity of variance violations on the performance of the confirmatory methods. To define/control the severity of (population) heterogeneity of variance in the second simulation study, we employ the following measure (for ease also called F_{max}):

$$F_{max} = \frac{\sigma^2_{max}}{\sigma^2_{min}}. \tag{3.2}$$

Example

The methods for testing hypotheses and selecting models are introduced using Lucas (2003). His experiment contains five experimental groups: 1) a group with a randomly selected male leader, 2) a group with a randomly selected female leader, 3) a group where the male team member who scores highest on the first task is selected as leader, 4) a group where the female team member who scores highest on the first task is selected as leader, and 5) a group in which female leadership is institutionalized and the female team member who scores highest on the first task is selected as leader. The institutionalization is done by showing the participants a film in which female leadership is normal and females do well as leaders. The dependent variable is the influence of the leader, obtained by a second task. The model of interest is (3.1) with $k = 5$ and $n_i = n = 30$. The group means and standard deviations are shown in Table 3.2.

Table 3.2: *Group Means and Standard Deviations of Influence (Lucas, 2003)*

Group	Mean	s.d.	n
1: randomly selected male leader	2.33	1.86	30
2: randomly selected female leader	1.33	1.15	30
3: male leader highest score	3.20	1.79	30
4: female leader highest score	2.23	1.45	30
5: female leader highest score and female leadership is institutionalized	3.23	1.50	30

Note. s.d. = standard deviation and n denotes the number of observations.

In this example, we evaluate the set of four hypotheses in (3.3): the traditional null H_0, the traditional alternative H_A, and two order-restricted ones H_1 and H_2. Both H_1 and H_2 are based on the expectation that leaders appointed on the basis of their ability (Groups 3 and 4) are expected to exert more influence over participants than leaders of the same sex appointed randomly (Groups 1 and 2, respectively); that is, $\mu_1 < \mu_3$ and $\mu_2 < \mu_4$. H_1 is based on two additional theories: *Women, according to status characteristics theory, will be disadvantaged relative to men in social interactions, all other things being equal* and *Institutionalizing women as leaders may overcome the influence gap between women and men.* In the context of the experiment, the first theory leads to the expectation that female leaders (Groups 2 and 4) will exert less influence over the members of a group they lead than male leaders selected in the same manner (Groups 1 and 3, respectively); that is, $\mu_2 < \mu_1$ and $\mu_4 < \mu_3$. This yields $\mu_2 < \mu_4 < \mu_3$ and $\mu_2 < \mu_1 < \mu_3$, which will be written as $\mu_2 < \{\mu_1, \mu_4\} < \mu_3$ for ease of notation. The second theory can be interpret in a few ways. Our interpretation is that "overcome" means that the gap is closed. Hence, based on the second theory, it is expected that institutionalized female leaders selected on the basis of their ability (Group 5) exerted the same amount of influence over participants as male leaders appointed on the basis of their ability (Group 3); that is, $\mu_5 = \mu_3$. These expectations lead to H_1 in (3.3). H_2 is additionally based on two competing theories: *Female leaders selected on the basis of their competence (Group 4) have less influence than male leaders selected at random (Group 1)* and *Institutionalizing women as leaders has no effect.* The first theory is represented by $\mu_4 < \mu_1$. Based on the second theory, it is expected that there is no difference between the influence of female leaders selected on the basis of their competence in the case of institutionalization (Group 5) or in the normal case (Group 4), that is, $\mu_5 = \mu_4$. These expectations are represented by H_2 in (3.3).

$$\begin{aligned} H_0 &: \ \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5, \\ H_1 &: \ \mu_2 < \{\mu_1, \mu_4\} < \mu_3 \ = \mu_5, \\ H_2 &: \ \mu_2 < \mu_5 \ = \mu_4 < \mu_1 < \mu_3, \\ H_A &: \ \mu_1, \ \mu_2, \ \mu_3, \ \mu_4, \ \mu_5. \end{aligned} \tag{3.3}$$

3.2.2 Exploratory and Confirmatory Techniques

Hypothesis Testing

There are several exploratory hypothesis testing techniques. The most common procedure is conducting an ANOVA F test followed by post hoc tests / pairwise multiple comparisons procedures when F is significant. A pairwise multiple comparisons procedure renders pairs of means significantly or insignificantly different from each other. When testing more than one pair, the α level, that is, the Type I error, inflates. There are several corrections to adjust for this, like the Bonferroni correction. Although this correction is perhaps the most familiar, it is not the most powerful. More powerful alternatives are the Shaffer-Welch Fq procedure (SWFq) in case of equal group sizes (Ramsey, 2002; Toothaker, 1993, pp. 42–43, 48; note that the technique

is called Shaffer-Ryan here) or the Tukey-Kramer (TK) method in case of unequal group sizes (Toothaker, 1993, pp. 60-61) Since we solely examine the performance of exploratory methods for equal group sizes, we will employ the SWFq test.

The SWFq method starts with an overall F test and is followed by testing only a selection of pairs of means which is based on the order of the sample means. In the example, $F(k-1=4, N-k=nk-k=145)=7.57$ renders a p value of 0.00. Hence, the group means are not equal, and we proceed with testing a certain selection of pairs of means. We will not give the details of the procedure, this can be found in Kuiper and Hoijtink (2010); Ramsey (2002); Toothaker (1993, pp. 42–43, 48). Using a nominal α level of .05, it is concluded that the means of group 2 and 5 and those of group 2 and 3 are significantly different; all the other pairs of means are not significantly different. From these results, it is hard to conclude anything with respect to the hypotheses H_1 and H_2 in (3.3). Moreover, the results of exploratory hypothesis testing techniques may be hard to interpreted simultaneously. For example, when $k=3$, it is logically impossible that $H_0:\mu_1=\mu_2$ and $H_0:\mu_2=\mu_3$ are not rejected and $H_0:\mu_1\neq\mu_3$ is. Although it might not be directly clear, in the example, there are also inconsistencies: The difference between mean 2 and 1 is not significant, neither is the difference between mean 5 and 1. However, the difference between mean 5 and 2 is significant, which seems to be in conflict with the other two results. Both problems are avoided by testing the hypotheses directly with a confirmatory hypothesis testing technique.

The $\bar{F}$ test (Silvapulle & Sen, 2005, pp. 25-42) is a confirmatory hypothesis testing technique, which is a modification of the F test such that it can test order-restricted hypotheses like $H_m:\mu_1\geq\mu_2\geq\mu_3$. One can test H_0 against an order-restricted alternative and one can test an order-restricted null against H_A. Since it is possible not to reject H_0 in the first (more or less, favoring H_0) and to reject the order-restricted null in favor of H_A, we also test H_0 against H_A. This leads to in $1+2*M$ tests, with M the number of order-restricted hypotheses. Note that no pairwise tests are required and, therefore, there are no inconsistencies. In the case of Lucas, H_0 is tested against H_A, H_1, and H_2 and both H_1 and H_2 are tested against the unconstrained hypothesis H_A. The results are presented in Table 3.3 and show, for $\alpha=.05$ (without multiple testing corrections), that H_A is preferred over H_0 and that both H_1 and H_2 are preferred over H_0 and H_A.

Table 3.3: *The $\bar{F}$ tests of the four specified hypotheses*

Hypotheses tested	$\bar{F}$	p value
H_0 against $\boldsymbol{H_A}$	30.27	< 0.001
H_0 against $\boldsymbol{H_1}$	30.26	< 0.001
$\boldsymbol{H_1}$ against H_A	0.01	0.995
H_0 against $\boldsymbol{H_2}$	22.91	< 0.001
$\boldsymbol{H_2}$ against H_A	7.36	0.070

Note. Bolding indicates the preferred hypothesis.

Although we obtain clearer information regarding our expectations (H_1 and H_2), there is still a drawback of this confirmatory method. The conclusions from the five tests must be combined and, thus, the results do not always lead to one overall preferred hypothesis. This is also the case in Table 3.3. Since no direct comparison between order-restricted hypotheses is possible with the $\bar{F}$ test, nothing can be concluded with respect to H_1 versus H_2. Hence, the $\bar{F}$ test is best used in case of one order-restricted hypothesis.

Model Selection using Information Criteria

Familiar exploratory information criteria are the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). They consist of a fit/likelihood part and a complexity/penalty part and select a best of a set of hypotheses. In classical exploration, all possible configurations of means are inspected. Dayton (2003) introduces a modification, called the Paired-Comparison Information Criterion (PCIC), which does not examine all possibilities. Like in the SWFq test, where comparisons are made based on the order of the sample means. Here, only the possible configurations based on the ordered sample means are examined. This avoids inconsistencies (Dayton, 1998, 2003) and it renders higher true-hypothesis rates when all population means are unequal (Dayton, 2003; Cribbie & Keselman, 2000). In case of $k = 5$ groups, there are 52 possible configurations and $2^{k-1} = 2^{5-1} = 16$ subsets based on ordered sample means. For the example, the order of the sample means and the corresponding group numbers are given in the upper panel in Table 3.4. Based on the ordering (i.e., 2, 4, 1, 3, and 5), 16 subsets can be determined, see Table 3.4 under 'Model', where a number represents the group number and a comma separates two subsets. For example, $\{24135\}$ represents H_0, $\{2, 4, 1, 3, 5\}$ equals H_A, and $\{2, 41, 35\}$ denotes $\mu_2,\ \mu_4 = \mu_1, \mu_3 = \mu_5$. Since model selection criteria evaluate configurations of means and not pairwise differences, they do not provide inconsistencies like exploratory hypothesis testing techniques can do. In summary, the PCIC is a modification that restricts the number of hypotheses to be evaluated and can be applied with the AIC (PCIC-AIC) or BIC (PCIC-BIC). Burnham and Anderson (2002, §6.4) argue that AIC has theoretical advantages over the BIC. In addition, the confirmatory model selection criterion discussed next is another modification of the AIC. Consequently, we will only evaluate the PCIC-AIC.

Table 3.4 displays the number of distinct model parameters (i.e., the number of model means plus one for the unknown σ^2) which equals the penalty of the AIC, denoted by q_m, the log likelihood $\log L_m$, and the PCIC-AIC values for all 16 hypotheses (for now, ignore the last column). The hypothesis with the lowest PCIC-AIC value is the preferred one. Here, this is Hypothesis 7 with group structure $\{2, 41, 35\}$ (i.e., $\mu_2,\ \mu_4 = \mu_1, \mu_3 = \mu_5$). Although PCIC-AIC does not render inconsistencies, it still does not give clear information about H_1 and H_2. This problem is solved by evaluating the set of (order-restricted) hypotheses directly, which can be done with a confirmatory model selection criterion.

The Order-Restricted Information Criterion (ORIC; Anraku, 1999) is a modification of the AIC such that it can evaluate a set of order-restricted hypotheses. Thus, as opposed to the $\bar{F}$ test, the ORIC can evaluate multiple order-restricted

Table 3.4: *PCIC-AIC and PMP Values for the* 16 *Hypotheses based on Ordered Sample Means*

Ordered sample means:		1.33	2.23	2.33	3.20	3.23	
Group nr. (i):		2	4	1	3	5	
Model nr. (m)	*Model*			q_m	$\log L_m$	PCIC-AIC	PMP
1	{24135}	$= H_0$		2	-292.27	588.54	0.00
2	{2,4135}			3	-283.38	572.76	0.03
3	{24,135}			3	-283.68	573.36	0.03
4	{241,35}			3	-281.79	569.57	0.11
5	{2413,5}			3	-288.36	582.71	0.00
6	{2,4,135}			4	-281.27	570.53	0.04
7	**{2,41,35}**			4	-278.08	**564.16**	**0.45**
8	{2,413,5}			4	-281.54	571.09	0.03
9	{24,1,35}			4	-280.57	569.14	0.06
10	{24,13,5}			4	-282.84	573.67	0.01
11	{241,3,5}			4	-281.78	571.57	0.02
12	{2,4,1,35}			5	-278.05	566.10	0.09
13	{2,4,13,5}			5	-280.39	570.79	0.02
14	{2,41,3,5}			5	-278.08	566.16	0.01
15	{24,1,3,5}			5	-280.57	571.13	0.09
16	{2,4,1,3,5}	$= H_A$		6	-278.05	568.10	0.02

Note. Bolding indicates the preferred hypothesis, the lowest PCIC-AIC value, and the highest PMP value.

Table 3.5: *ORIC and PMP values of the Four Specified Hypotheses*

Model	q_m	$\log L_m$	*ORIC*	*PMP*
$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$	2.00	-292.27	588.54	0.00
$\boldsymbol{H_1 : \mu_2 < \{\mu_1, \mu_4\} < \mu_3 = \mu_5}$	3.19	-278.05	**562.48**	**0.96**
$H_2 : \mu_2 < \mu_5 = \mu_4 < \mu_1 < \mu_3$	3.14	-281.76	569.80	0.02
$H_A : \mu_1,\ \mu_2,\ \mu_3,\ \mu_4,\ \mu_5$	6.00	-278.05	568.10	0.01

Bolding indicates the preferred hypothesis, the lowest *ORIC* value, and the highest PMP value.

hypotheses at once. Note that H_A should be included as a safeguard for weak hypotheses (Kuiper & Hoijtink, 2010), that is, hypotheses not supported by the data. Namely, when all hypotheses are weak, H_A will receive the most support. In Table 3.5, the penalty terms q_m, the log likelihood values $\log L_m$, and the ORIC values are given for the four hypotheses of interest denoted in (3.3) (again, ignore the last column). Note that for hypotheses with no order restrictions, like H_0 and H_A, the ORIC reduces to the (PCIC-)AIC. In that case, the penalty equals the number of distinct parameters. Otherwise, the penalty of the ORIC equals the expected number of distinct parameters. This is hard to determine by hand, but can easily be simulated

(Silvapulle & Sen, 2005, pp. 78–81). For the example, hypothesis H_1 is preferred, since it has the smallest ORIC value.

Bayesian Model Selection

Klugkist, Laudy, and Hoijtink (2005) present a Bayesian model selection (BMS) method in which the selection of the best configuration of means is not based on an information criterion but on the marginal likelihood. The marginal likelihood for a specific hypothesis is a measure of the degree of support for the hypothesis provided by the data. To interpret several marginal likelihoods at once, it can be helpful to transform them into the so-called posterior model probabilities (PMPs). A PMP is the probability that, given the data, the corresponding hypothesis is the best of the set of hypotheses (assuming a priori that all the hypotheses have equal probabilities). The marginal likelihood depends on the likelihood and a so-called prior. A prior reflects prior knowledge with respect to the means. In this chapter, we use the normal distribution with a data-based mean and a large variance for every μ_i $(i = 1, \ldots, k)$. The prior is not only chosen vague (such that it has minimal impact on the results), but also compatible with the data (such that it is not too vague), see Klugkist, Laudy, and Hoijtink (2005). The prior mean and variance not only depend on the data, but also on a user-specified term (PV) that reflects the vagueness of the prior (see Kuiper & Hoijtink, 2010; Kuiper, Klugkist, & Hoijtink, 2010), where a higher PV value corresponds to an increasing prior vagueness. Klugkist and Hoijtink (2007) show that for reasonable choices of PV, the prior sensitivity does usually not lead to a different evaluation of the hypotheses. Furthermore, if a hypothesis contains only inequality constraints (i.e., "<" and/or ">"), the relative support of this hypothesis with respect to the unconstrained hypothesis is not sensitive to the choice of the prior. Moreover, simulations with PV values of 1, 2, and 3 showed that with $PV = 1$ the false rejection rate of H_0 is best controlled, with $PV = 3$ the power to find non-null hypotheses is largest, and that $PV = 2$ provides a compromise. As a consequence, we will solely display the results for $PV = 2$ for both exploration and confirmation.

BMS can be used in an exploratory way as well as in a confirmatory one. In exploration, we can for instance evaluate only the 2^{k-1} subsets based on ordered sample means (comparable with PCIC). In confirmation, a limited set of well-defined hypotheses is evaluated. In BMS, the hypothesis with the highest PMP value is the preferred one. The last column in Table 3.4 shows that Hypothesis 7 with group structure $\{2, 41, 35\}$ (i.e., $\mu_2,\ \mu_4 = \mu_1, \mu_3 = \mu_5$) is the preferred hypothesis in exploration. From the last column in Table 3.5 it can be concluded that H_1 is the preferred hypothesis in confirmation. Note that the same conclusions were obtained with the PCIC-AIC and ORIC, respectively.

Comparison of the Six Methods

It should be stressed that hypothesis testing serves another purpose than model selection does. The goal of the first is to reject the null hypothesis, whereas the goal of the latter is to select the best out of a set of hypotheses. Hence, in the first, the null hypothesis is of more importance and, in the latter, all hypotheses are equally

important. Nevertheless, we do compare them to examine the true hypothesis-rates of the null hypothesis and other (order-restricted) hypotheses. Due to their purposes, we expect that true hypothesis-rate of the null hypothesis is the highest for hypothesis testing and that of non-null hypotheses for model selection.

3.3 Performance of Confirmatory and Exploratory Methods under Heterogeneity

The performance of the three exploratory and three confirmatory methods (summarized in Table 3.1) is evaluated by conducting a simulation study. The performance of hypothesis testing can be measured by power, that is, the probability that the test will reject a false null hypothesis. Stated otherwise, power is the probability that the test will favor the alternative hypothesis when it is true. In model selection, one can employ an equivalent of power, namely the probability that the method will render the most support for the correct or best hypothesis. In the simulation study, the performance is quantified by the number of times the method prefers the correct or best hypothesis.

In this section, two comparisons are made: one between the performance of hypothesis testing, model selection using information criteria, and Bayesian model selection; the other between the performance of exploratory and confirmatory approaches. Before describing the results, we discuss the values of k and n_i, the hypotheses, and the population parameters employed in the simulation.

3.3.1 The Number of Groups and Observations

Bear in mind that confirmatory methods have an added value when comparing three means or more, since with two means one can do a one-sided test. To obtain insight in the performance of the methods, we start with a simulation with $k = 3$ groups. We additionally inspect the ANOVA model in (3.1) with $k = 5$ groups (based on the Lucas example discussed before). From these two simulations a pattern becomes clear with respect to the performance of exploratory and confirmatory techniques. Therefore, we only inspect ANOVA models with $k = 3$ and $k = 5$.

Note that the performance of a method increases when the number of observations per group increases. As a consequence, it is more interesting to examine data sets with low to medium group sizes. Based on the findings of Cohen (1992) (to show a medium and large effect with the ANOVA F-test), we will inspect group sizes between 20 and 50 observations. Since our second simulation is based on Lucas (2003), we choose to employ an equal number of observations per group in the first simulation. Due to the reasonings above, we will examine the ANOVA model with $k = 3$ for both $n = 20$ and $n = 50$. The results for $n = 20$ are not shown here, since the patterns are the same as for $n = 50$, the only difference is that the performance itself is lower. For the ANOVA model with $k = 5$ we will employ $n = 30$, since this was the group size in the study of Lucas (2003).

3.3.2 Hypotheses

Table 3.6 depicts the hypotheses of interest in the simulation study for $k = 3$ and $k = 5$. As explained in the previous section, the three exploratory techniques (SWFq, PCIC-AIC, and BMS) do not evaluate all possible configurations of means (5 for $k = 3$ and 52 for $k = 5$) in the observed data set, but a subset based on the ordering of the sample means of the data set at hand (in case of PCIC-AIC and BMS, 4 for $k = 3$ and 16 for $k = 5$). Nevertheless, more configurations of means can be examined, since the ordering of the sample means can differ per data set in the simulation. In the exploratory approaches, the hypotheses to be examined are certain group structures represented by pairwise equality and non-equality relations. Combining the significant and insignificant pairs of means resulting from the SWFq test can lead to favoring one of the hypotheses in the first column of Table 3.6 or can give inconsistencies. PCIC-AIC and BMS always result in preferring one of the hypotheses in Table 3.6.

In the confirmatory approaches ($\bar{F}$, ORIC, BMS), the hypotheses to be tested or selected need to be specified by the researcher. This can be based on previous research and/or on existing theories. One can also inspect competing theories. Table 3.6 displays the hypotheses that are evaluated in this chapter. Bear in mind that the type and the number of hypotheses are an example. For $k = 3$, we choose to evaluate five hypotheses. Note that this number equals the maximum number of possible configurations of means in exploration, but the structure is different. The set for $k = 5$ is based on Lucas (2003) and is the same as presented in (3.3). It should be stressed that 16 hypotheses are evaluated for one data set in exploration for $k = 5$, whereas the researcher often has a limited number of theories/hypotheses; for instance, 4 in the Lucas example and simulation.

The $\bar{F}$ test is designed for testing one order-restricted hypothesis, like H_{C1}. One can choose to test both H_0 against H_{C1} and H_{C1} against H_A in addition to H_0 against H_A. The decision rules for these three test are rather straightforward. However, if M order-restricted hypotheses are evaluated by $1 + 2M$ tests, the decision rules become very ad hoc and more than one plausible set of decision rules exist. Therefore, we will only examine the performance of the $\bar{F}$ test for one order-restricted hypothesis, namely H_{C1}.

3.3.3 Populations

Several populations based on the general ANOVA model presented in (3.1) are inspected. In all populations, the population standard deviation σ is set equal to 1. Sets of population means are given in Table 3.7. The values are based on the number of groups k, the true hypothesis, and the effect size denoted by ES, where

$$ES = \frac{1}{\sigma}\sqrt{\frac{1}{k}\sum_{i=1}^{k}(\mu_i - \overline{\mu})^2}, \tag{3.4}$$

with $\overline{\mu} = \frac{1}{k}\sum_{i=1}^{k}\mu_i$. According to Cohen (1992), the effect size in (3.4) is low for $ES = 0.1$, medium for $ES = .25$, and high for $ES = 0.4$. Two types of populations can be distinguished: one where all the population means are identical ($ES = 0$)

Table 3.6: Hypotheses tested for $k = 3$ and $k = 5$ in the exploratory ($H_{E.}$) and confirmatory approach ($H_{C.}$).

	Exploration	Confirmation
$k = 3$	$H_0: \quad \mu_1 = \mu_2 = \mu_3$	$H_0: \quad \mu_1 = \mu_2 = \mu_3$
	$H_{E1}: \quad \mu_1 = \mu_2, \mu_3$	$H_{C1}: \mu_1 < \mu_2 < \mu_3$
	$H_{E2}: \quad \mu_1, \mu_2 = \mu_3$	$H_{C2}: \mu_1 = \mu_2 < \mu_3$
	$H_{E3}: \quad \mu_1 = \mu_3, \mu_2$	$H_{C3}: \mu_1 < \mu_2 < \mu_3$
	$H_A: \quad \mu_1, \mu_2, \mu_3$	$H_A: \quad \mu_1, \mu_2, \mu_3$
$k = 5$	$H_0: \quad \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$	$H_0: \quad \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
	$H_{E1}: \quad \mu_1 = \mu_2 = \mu_3 = \mu_4, \mu_5$	$H_{C1}: \mu_5 = \mu_3 > \{\mu_1, \mu_4\} > \mu_2$
	$H_{E2}: \quad \mu_1 = \mu_2 = \mu_3 = \mu_5, \mu_4$	$H_{C2}: \mu_3 > \mu_1 > \mu_4 = \mu_5 > \mu_2$
	$\vdots$	$H_A: \quad \mu_1, \mu_2, \mu_3, \mu_4, \mu_5$
	$H_{E49}: \mu_1, \mu_2, \mu_4, \mu_3 = \mu_5$	
	$H_{E50}: \mu_1, \mu_3, \mu_2, \mu_4 = \mu_5$	
	$H_A: \quad \mu_1, \mu_2, \mu_3, \mu_4, \mu_5$	

and one where they are in accordance with H_{C1} ($ES = 0.1, 0.25$, and 0.4). Based on each population in Table 3.7, 1,000 data sets are simulated. Subsequently, the corresponding hypotheses in Table 3.6 are evaluated in each of these data sets. Note that H_A and H_{E49} are the correct hypotheses in exploration when $ES > 0$ for $k = 3$ and $k = 5$, respectively.

Table 3.7: Population means for $k = 3$ and $k = 5$ for zero, small, medium, and large effect size (ES).

k	true H_m	ES	μ_1	μ_2	μ_3	μ_4	μ_5
3	H_0	0	0.000	0.000	0.000		
	H_{C1}	0.10	-0.122	0.000	0.122		
	H_{C1}	0.25	-0.306	0.000	0.306		
	H_{C1}	0.40	-0.490	0.000	0.490		
5	H_0	0	0.000	0.000	0.000	0.000	0.000
	H_{C1}	0.10	0.000	-0.122	0.130	0.122	0.130
	H_{C1}	0.25	0.000	-0.306	0.321	0.306	0.321
	H_{C1}	0.40	0.000	-0.490	0.516	0.490	0.516

3.3.4 Results

Exploration for $k = 3$

Table 3.8 displays the proportions of times the hypotheses are preferred for each method and each population. The SWFq test chooses, as expected, H_0 as the preferred hypothesis about 95% of the times when it is indeed true. Notably, the SWFq test is designed such that $\alpha = .05$. The other two methods do not choose H_0 as often. For $ES = 0.10$, all three method lack power to prefer the correct hypothesis, as is to be expected for a small effect size. When $ES = 0.25$ or $ES = 0.40$, the performance is still questionable, but the PCIC-AIC clearly outperforms the other two. Thus, the SWFq test performs well under H_0 and the PCIC-AIC has most power to find the effect specified.

Table 3.8: *The Proportion of Times a Hypothesis is Preferred in Exploration*

$k = 3$ and $n = 50$							
ES	Method	H_0	H_{E1}	H_{E2}	H_{E3}	H_A	'Inconsistent'
0	SWFq	**.944**	.012	.009	.007	.000	.028
0	PCIC-AIC	**.653**	.109	.116	.117	.005	-
0	BMS	**.807**	.060	.063	.059	.011	-
0.1	SWFq	.845	.042	.037	.004	**.000**	.072
0.1	PCIC-AIC	.445	.252	.231	.060	**.012**	-
0.1	BMS	.631	.017	.181	.116	**.055**	-
0.25	SWFq	.227	.268	.262	.000	**.043**	.200
0.25	PCIC-AIC	.040	.335	.337	.002	**.226**	-
0.25	BMS	.066	.431	.433	.000	**.070**	-
0.4	SWFq	.006	.260	.281	.000	**.416**	.037
0.4	PCIC-AIC	.000	.139	.149	.000	**.712**	-
0.4	BMS	.001	.256	.279	.000	**.464**	-

Note. SWFq = Shaffer-Welch Fq test; PCIC = paired-comparison information criterion; AIC = Akaike information criterion; BMS = Bayesian model selection; Bolding indicates the correct or best hypothesis in each row.

Confirmation for $k = 3$

In Table 3.9, the performance of the confirmatory methods is presented for one order-restricted hypothesis in addition to the classical null and alternative hypothesis. The $\bar{F}$ test chooses H_0 as the preferred hypothesis about 90% of the times when it is indeed true. This is to be expected, since we perform two tests with respect to H_0 with $\alpha = .05$ and do not correct for multiple testing. The performance of BMS resembles the one of the $\bar{F}$ test, while the ORIC performs less well under H_0. For $ES = 0.10$, the three techniques do not perform very well, as is to be expected at

Table 3.9: *The Proportion of Times a Hypothesis is Preferred in Confirmation (H_0 vs H_{C1} vs H_A)*

$k = 3$ and $n = 50$					
ES	Method	H_0	H_{C1}	H_A	'Inconsistent'
0	$\bar{F}$	**.912**	.052	.033	.003
0	ORIC	**.724**	.187	.089	-
0	BMS	**.881**	.075	.044	-
0.1	$\bar{F}$	.682	**.310**	.007	.001
0.1	ORIC	.413	**.563**	.025	-
0.1	BMS	.615	**.354**	.031	-
0.25	$\bar{F}$	.114	**.886**	.000	.000
0.25	ORIC	.017	**.981**	.002	-
0.25	BMS	.072	**.917**	.011	-
0.4	F	.004	**.995**	.001	.000
0.4	ORIC	.000	**.999**	.001	-
0.4	BMS	.000	**.998**	.002	-

Note. ORIC = order-restricted information criterion; BMS = Bayesian model selection; Bolding indicates the correct or best hypothesis in each row.

a low effect size. For $ES = 0.25$ and $ES = 0.4$, all three method perform very well, they all prefer H_{C1} more than 88% of the times. For all $ES > 0$, the ORIC performs (somewhat) better than the other two.

Table 3.10 shows the results for evaluating multiple order-restricted hypotheses. Since, the $\bar{F}$ test is hard to use if more than one order-restricted hypothesis is evaluated, the $\bar{F}$ test is excluded. Here one can see that including more hypotheses decreases the proportion of times the correct hypothesis is chosen, especially when the effect size is not large. When H_0 is true, BMS performs better than the ORIC and when H_{C1} is true the ORIC outperforms BMS.

When comparing Table 3.10 to Table 3.8, it is evident that confirmative methods have more power than explorative methods. For instance, for medium effect size, the ORIC selects H_{C1} in about 75% of the simulated data sets and the PCIC-AIC selects H_A in about 23% of the cases. Bear in mind that H_A is the correct hypothesis in exploration when H_{C1} is the true hypothesis.

Exploration for $k = 5$

In exploration, there are 52 possible hypotheses when $k = 5$ (when inspecting more than one data set). Because of the large number of hypotheses, only the results of three of these are given and the results of the other hypotheses are combined (column "other" in Table 3.11). We display the results for the null hypothesis H_0, the alternative hypothesis H_A, and the correct hypothesis H_{E49}. Moreover, we did not include BMS, since evaluating 16 hypotheses per data set with BMS is very

Table 3.10: *The Proportion of Times a Hypothesis is Preferred in Confirmation (H_0 vs H_{C1} vs H_{C2} vs H_{C3} vs H_A)*

		$k = 3$ and $n = 50$				
ES	Method	H_0	H_{C1}	H_{C2}	H_{C3}	H_A
0	ORIC	**.672**	.059	.124	.110	.035
0	BMS	**.777**	.025	.114	.063	.021
0.1	ORIC	.358	**.296**	.298	.044	.004
0.1	BMS	.461	**.153**	.333	.052	.001
0.25	ORIC	.016	**.751**	.219	.014	.000
0.25	BMS	.039	**.634**	.297	.030	.000
0.4	ORIC	.000	**.938**	.062	.000	.000
0.4	BMS	.000	**.888**	.109	.003	.000

Note. ORIC = order-restricted information criterion; BMS = Bayesian model selection; Bolding indicates the correct or best hypothesis in each row.

time-consuming. Furthermore, given the results of the other two methods, we do not expect that examining BMS renders additional information.

The proportions of times the hypotheses are selected are displayed in Table 3.11. It shows that in exploration, if H_0 is true, H_0 is frequently preferred when using the SWFq test, namely about 95% of the times. In contrast, if H_{E49} is true, the true hypothesis is not chosen by the SWFq test. The PCIC-AIC only gives 35% of the times the most support to H_0 when it is true and less than 2% to H_{E49} when it is true. Hence, explorative methods perform poorly under H_{E49}. Note that H_A is not preferred at all.

Table 3.11 shows that the power to test any specific configuration of means is very low whereas the power to detect at least one effect (1 minus first column) is not. Hence, employing these methods will usually give significant results, but they tend to vary across data sets. This was also discussed by Maxwell (2004), where he defined power for a specific comparison, any-pairs power, and all-pairs power in the context of multiple testing.

Confirmation for $k = 5$

Table 3.12 shows the results of the three methods for evaluating one order-restricted hypothesis. This table exhibits the same patterns as the one for $k = 3$, that is, $\bar{F}$ and BMS outperform the ORIC under H_0, whereas the ORIC has more power to detect small or medium effect sizes. For large effect sizes, all three methods perform very well and approximately equal.

Table 3.13 depicts the performance of the ORIC and BMS for examining two order-restricted hypotheses. It can be seen that adding a hypothesis lowers the performance of the methods, but the trend remains the same. BMS performs well under H_0 and under H_{C1} for a large effect size, while the ORIC has more power to detect the correct hypothesis (H_{C1}) for small and medium effect sizes.

Table 3.11: *The Proportion of Times a Hypothesis is Preferred in Exploration*

$k = 5$ and $n = 30$					
ES	Method	H_0	H_{E49}	H_A	other
0	SWFq	**.947**	.000	.000	.053
0	PCIC-AIC	**.349**	.000	.000	.651
0.1	SWFq	.878	**.000**	.000	.122
0.1	PCIC-AIC	.201	**.000**	.000	.799
0.25	SWFq	.371	**.000**	.000	.629
0.25	PCIC-AIC	.015	**.002**	.000	.983
0.4	SWFq	.007	**.000**	.000	.993
0.4	PCIC-AIC	.000	**.016**	.000	.984

Note. SWFq = Shaffer-Welch Fq test; PCIC = paired-comparison information criterion; AIC = Akaike information criterion; Bolding indicates the correct or best hypothesis in each row.

Table 3.12: *The Proportion of Times a Hypothesis is Preferred in Confirmation (H_0 vs H_{C1} vs H_A)*

$k = 5$ and $n = 30$					
ES	Method	H_0	H_{C1}	H_A	'Inconsistent'
0	$\bar{F}$	**.920**	.045	.031	.004
0	ORIC	**.752**	.162	.086	-
0	BMS	**.965**	.027	.008	-
0.1	$\bar{F}$	.715	**.264**	.019	.002
0.1	ORIC	.421	**.536**	.043	-
0.1	BMS	.779	**.200**	.021	-
0.25	$\bar{F}$	.130	**.860**	.007	.003
0.25	ORIC	.026	**.935**	.039	-
0.25	BMS	.201	**.768**	.031	-
0.4	$\bar{F}$	.002	**.990**	.008	.000
0.4	ORIC	.000	**.965**	.035	-
0.4	BMS	.005	**.970**	.025	-

Note. ORIC = order-restricted information criterion; BMS = Bayesian model selection; Bolding indicates the correct or best hypothesis in each row.

3.3.5 Conclusion

When the interest lies in one or more order-restricted hypotheses, exploration has some disadvantages. First, the hypotheses of interest cannot be evaluated directly. Moreover, exploratory hypothesis techniques can render inconsistencies or difficult to interpret results. Last, exploratory methods exhibit low power to detect specific configurations of means, especially when the number of groups (k) increase.

From the simulations, it can be concluded that the confirmatory methods outperform their exploratory counterparts in case the interest lies in one or more order-restricted hypotheses. The $\bar{F}$ test performs very well under H_0. In contrast, it performs not as good when another hypothesis is true. Note that when you have a theory, one could chose to protect H_0 less by testing at a lower α level, which increases the performance when a non-null hypothesis is true. A disadvantage of the $\bar{F}$ test is that it can evaluate only one order-restricted hypothesis (in a straightforward manner). Hence, we recommended the use of confirmatory model selection, that is, the Order-Restricted Information Criterion (ORIC) or Bayesian model selection (BMS). The ORIC performs best when H_0 is not true, whereas BMS performs less well when H_0 is not true (with an exception of large effect sizes where it is about equally good as the ORIC) but better when H_0 is true.

It should be stressed that in this simulation study the true hypothesis is included in the set for the confirmatory methods. What if a hypothesis not contained in the set is true? In that case, H_A will be preferred over the other hypotheses, if the sample size is large enough to distinguish between those in the set from the correct one. Therefore, we recommend to always include the unconstrained hypothesis H_A in the analysis. In this case, model selection techniques select

- the correct hypothesis (if it is included), or
- a similar one (i.e., a hypothesis which resembles the true hypothesis, that is, only differs in a few constraints), or
- the unconstrained hypothesis.

This is also supported by simulation (not shown here).

Table 3.13: *The Proportion of Times a Hypothesis is Preferred in Confirmation (H_0 vs H_{C1} vs H_{C2} vs H_A)*

$k = 5$ and $n = 30$					
ES	Method	H_0	H_{C1}	H_{C2}	H_A
0	ORIC	**.708**	.114	.113	.065
0	BMS	**.948**	.024	.018	.010
0.1	ORIC	.378	**.364**	.221	.038
0.1	BMS	.761	**.152**	.074	.013
0.25	ORIC	.033	**.757**	.175	.035
0.25	BMS	.181	**.690**	.101	.028
0.4	ORIC	.000	**.888**	.076	.036
0.4	BMS	.002	**.918**	.035	.045

Note. ORIC = order-restricted information criterion; BMS = Bayesian model selection; Bolding indicates the correct or best hypothesis in each row.

3.4 Robustness of Performance in Confirmation under Heterogeneity

As mentioned before, little is known about the influence of violations of assumptions of the ANOVA model, especially for the confirmatory methods. In this simulation study, we will investigate their robustness of performance in the presence of heterogeneity.

3.4.1 Populations and Hypotheses

As in the previous section, we use (3.1) in the analyses. We will employ $k = 3$. But, the populations differ, namely now there are group specific variances: σ_i^2 for $i = 1, 2, 3$. Although the population standard deviations are divergent, they are assumed to be equal in the model. As mentioned before, the F test is not as robust for heterogeneity when the groups with the largest sample sizes have the highest variances and the ones with the smallest sample sizes have the lowest variances than when the group sizes are equal. Therefore, we will examine both equal group sizes ($n_1 = n_2 = n_3 = n$, with $n = 20$ and $n = 50$) and unequal group sizes ($n_1 = 20$, $n_2 = 50$, and $n_3 = 100$). In this study, the following set of hypotheses is evaluated:

$$\begin{aligned} H_0 &: \ \mu_1 = \mu_2 = \mu_3, \\ H_1 &: \ \mu_1 < \mu_2 < \mu_3, \\ H_A &: \ \mu_1, \ \mu_2, \ \mu_3. \end{aligned} \tag{3.5}$$

The population means of this study equal the population means in the previous one, see the upper panel in Table 3.7. Thus, there is one set where all the population means are identical ($ES = 0$) and H_0 is true and three where they exhibit an upward trend ($ES > 0$) and H_1 is true. Due to heterogeneity, the effect size is calculated by (3.4), where σ is replaced by the pooled standard deviation σ_p, with

$$\sigma_p = \sqrt{\frac{\sum N_i \sigma_i^2}{\sum N_i}},$$

with N_i the size of group i in the population. In the simulation, we will use n_i instead of N_i, because we assume that the relative sizes of the groups in the samples equals those in the population. By setting σ_p to 1, the same effect sizes as in the previous simulation study (upper panel in Table 3.7) are obtained.

The standard deviations (σ_i) are with $\sigma_p = 1$ solely based on the F_{max} value in (3.2). Evidently, a higher F_{max} value indicates a larger difference in group variances and vice versa. Importantly, $F_{max} = 1$ implies that there is no difference in the group variances, that is, the homogeneity of variance assumption is not violated. Consequently, we will use $F_{max} = 1$ as the baseline. Based on the findings of Tabachnick and Fidell (2001) and Box (1954), we additionally set F_{max} to 3 and 10. To examine the effect of a large violation, we inspect $F_{max} = 100$ as well. For $F_{max} > 1$, different orderings of the σ_i values in relation to the ordering of group sizes exist. Note that the ranking of the σ_i values is arbitrary in case of equal group sizes. Hence, for equal group sizes, we only examine σ_i values with an upward trend, that

is, the σ_i^2 values increase with i. If the group sizes are unequal, we will investigate samples with sizes increasing with i with two rankings of σ_i values based on the results of Box (1954), namely one with an upward trend and one with a downward trend. Due to the two types of group sizes (equal and unequal), the four F_{max} values, and the two types of trends (upward or downward), eleven sets of population standard deviations are investigated, which are given in Table 3.14.

Table 3.14: The Population Standard Deviations

		Equal Group Sizes*			Unequal Group Sizes**		
Type of trend	F_{max}	σ_1	σ_2	σ_3	σ_1	σ_2	σ_3
Baseline	1	1.000	1.000	1.000	1.000	1.000	1.000
Upward	3	0.707	1.000	1.225	0.612	1.000	1.061
Downward	3				1.500	1.000	0.866
Upward	10	0.426	1.000	1.348	0.343	1.000	1.085
Downward	10				2.000	1.000	0.632
Upward	100	0.141	1.000	1.407	0.109	1.000	1.094
Downward	100				2.390	1.000	0.239

* $n_1 = n_2 = n_3 = n$, with $n = 20$ and $n = 50$
** $n_1 = 20, n_2 = 50, n_3 = 100$

For each combination of n_i in relation with σ_i, F_{max}, and ES, 1,000 data sets are simulated. Subsequently, the hypotheses in (3.5) are evaluated in each of these data sets.

3.4.2 Results and Conclusions

Figure 3.1 displays the proportion of times H_1 is preferred by the three confirmatory techniques ($\bar{F}$ on top, ORIC in the middle, and BMS on the bottom) for effect size ES (represented by the different lines in each plot) and heterogeneity level F_{max} (depicted at the x-axis of each plot) in case of unequal group sizes. The results for equal group sizes are not plotted, since they are very robust, but we will briefly elaborate on this below. The performance is measured by the proportion of times the correct hypothesis is preferred (displayed on the y-axis of each plot). Observe that complete robustness for heterogeneity would imply only horizontal lines. The figure shows that the effect of heterogeneity on the performance of the three methods when H_1 is true depends on the effect size. For medium to large effect sizes (i.e., $ES = 0.25$ and 0.4), the proportion of times H_1 is preferred increases with F_{max}, when the σs show an upward trend (see the two top lines in the panels on the left hand side in Figure 3.1). In addition, the proportion of times H_1 is preferred decreases with F_{max}, when the σs show an downward trend (see the two top lines the panels on the right hand side in Figure 3.1). The opposite holds true for small effect sizes (i.e., $ES = 0.1$ and 0). That is, for $ES = 0.1$ and 0, the proportion of times H_1 is preferred decreases (increases) with F_{max}, when the σs show an upward (downward) trend

(see the two bottom lines in the panels on the left (right) hand side in Figure 3.1). Furthermore, the difference in performance due to heterogeneity is larger when the σs have a downward trend compared to when they exhibit an upward trend. It should be stressed that the difference in performance due to heterogeneity is not profound for $F_{max} = 3$ in all cases and for $F_{max} = 10$ when the σs exhibit an upward trend. Moreover, note that an F_{max} value of 100 can be considered an extreme violation, compared to the benchmarks of Tabachnick and Fidell (2001) discussed earlier, where an F_{max} value of 10 and 3 is concluded to be acceptable for analyses with equal group sizes and unequal group sizes, respectively.

Although the general trend is clear from Figure 3.1, the magnitude of the deviations from the baseline is less clear. In addition, it does not report on the performance when the group sizes are equal or on the proportion of times H_0 is selected. Note that, in this simulation, the interest lies in robustness of the techniques and not in the performance itself. Therefore, we included two tables in the appendix which present the difference in performance for an $F_{max} > 1$ compared with $F_{max} = 1$. These differences give an indication of the robustness of heterogeneity on the performance under both H_0 and H_1, where a difference of zero indicates full robustness. As a consequence, a higher absolute difference reflects a poorer robustness.

Table 3.15 depicts the difference in performance for $ES = 0$ when H_0 is correct. In case of equal group sizes, no eminent trend can be seen across methods and/or F_{max} values. Additionally, the difference in proportion of times H_0 is preferred are extremely small for the different F_{max} values. In the unequal sample size condition, the proportion of times H_0 is preferred increases (decreases) with F_{max}, when the σs exhibit an upward (downward) trend. Furthermore, the effect of heterogeneity on performance are more severe for the downward trend than for the upward trend. It should be stressed that these results resemble the effect of heterogeneity on the ANOVA F test.

Table 3.16 shows, for all effect sizes, the difference in the proportion of times H_1 is preferred. It shows that the difference in proportion of times H_1 is preferred do not differ more than .045 in absolute sense for the three F_{max} values for all effect sizes for both sample sizes. Also here, no profound trend arises. The patterns in robustness for unequal group sizes are already discussed with Figure 3.1. In addition, Table 3.16 shows that the absolute difference in performance is less than .10 in case of the upward trend and .16 in case of the downward trend.

We conclude that for all three techniques (i.e., the $\bar{F}$ test, the ORIC, and BMS) the performance under both H_0 and H_1 is robust for heterogeneity when the group sizes are equal. In case of unequal group sizes and when the group standard deviations exhibit the same trend, the performance under both H_0 and H_1 is still quite robust. However, when the group standard deviations exhibit an opposite trend, there are larger deviations, especially for F_{max} values larger than three.

3.5 Discussion

Several populations have been examined, however, for brevity only a summary of the simulation study is given in this chapter. As we have advocated, there is much

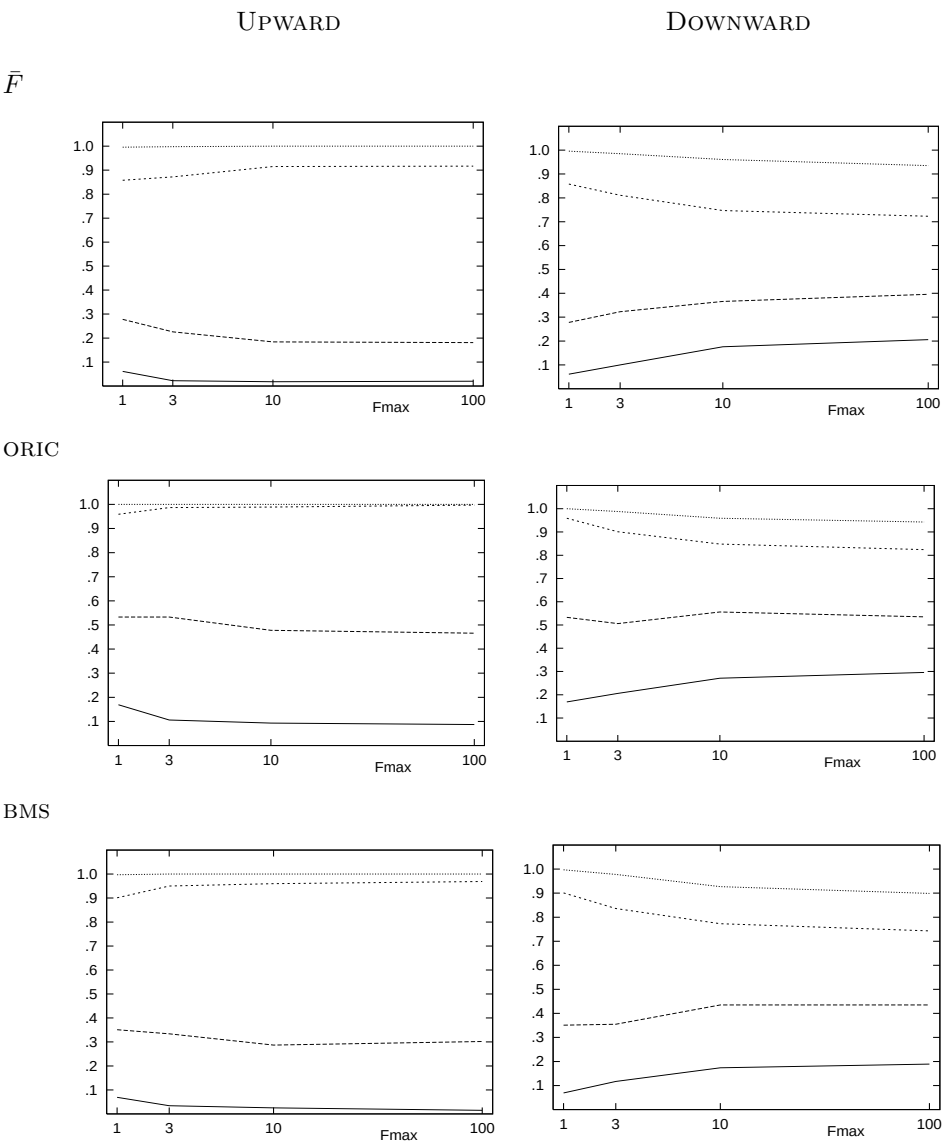


Fig. 3.1: Proportion of times H_1 is selected as best hypothesis for increasing heterogeneity (F_{max}) for unbalanced designs with either larger variances in larger subgroups (Upward) or smaller variances in larger subgroups (Downward). In each plot, the four lines from bottom to top show results for $ES = 0, 0.1, 0.25$, and 0.4, respectively.

to be gained from using confirmatory model selection techniques. Furthermore, we have attempted to provide some insight into the robustness of confirmatory methods. Although more research is required to state that the confirmatory methods are robust

for heterogeneity, our results should encourage the use of confirmatory techniques, primarily confirmatory model selection. Software (with a tutorial) is available (at `http://staff.fss.uu.nl/RMKuiper`) to perform the techniques described on your own data set. More details can be found in Kuiper and Hoijtink (2010) and Kuiper et al. (2010).

3.A Tables with Results of Robustness of Heterogeneity in Confirmation

Table 3.15: Difference in proportion of times H_0 is selected for $F_{max} = 3, 10$, and 100 compared to $F_{max} = 1$ for $ES = 0$

Method	$n_i = 20$			$n_i = 50$			$n_1 = 20, n_2 = 50, n_3 = 100$					
							upward trend			downward trend		
	3	10	100	3	10	100	3	10	100	3	10	100
$\bar{F}$	-.012	-.017	-.006	.005	-.004	-.051	.043	.057	.054	-.086	-.213	-.293
ORIC	-.005	-.016	.006	.040	.039	-.002	.085	.122	.125	-.114	-.210	-.291
BMS	-.016	-.017	-.013	.011	.009	-.048	.039	.056	.059	-.096	-.207	-.286

Note. ES = effect size; ORIC = order-restricted information criterion; BMS = Bayesian model selection.

Table 3.16: Difference in proportion of times H_1 is selected for $F_{max} = 3, 10$, and 100 compared to $F_{max} = 1$

Method		$n_i = 20$			$n_i = 50$			$n_1 = 20, n_2 = 50, n_3 = 100$					
								upward trend			downward trend		
		3	10	100	3	10	100	3	10	100	3	10	100
$\bar{F}$	0	-.002	.007	-.003	.001	.006	.042	-.039	-.043	-.041	.039	.115	.145
	0.1	.031	.012	.026	-.017	.008	.008	-.052	-.094	-.097	.045	.088	.118
	0.25	.039	.025	-.020	.014	.013	.026	.014	.057	.059	-.047	-.111	-.135
	0.4	-.009	-.002	-.002	.005	.004	.005	.002	.004	.004	-.011	-.035	-.061
ORIC	0	-.015	.004	-.010	-.028	-.037	.012	-.063	-.076	-.082	.037	.102	.127
	0.1	.045	.017	-.017	-.004	.002	-.013	.000	-.055	-.067	-.027	.023	.002
	0.25	.012	.009	-.020	.006	-.001	-.003	.028	.030	.038	-.058	-.111	-.135
	0.4	-.009	.003	-.041	.000	.000	.001	.000	.000	.000	-.012	-.041	-.057
BMS	0	-.003	.006	.001	-.011	-.017	.043	-.035	-.044	-.054	.048	.105	.120
	0.1	.030	.012	.011	.012	.021	.021	-.017	-.064	-.049	.004	.084	.084
	0.25	.038	.038	-.004	.025	.008	.019	.049	.059	.068	-.065	-.128	-.158
	0.4	-.014	-.015	-.007	.000	.000	.002	.003	.003	.003	-.019	-.070	-.098

Note. ES = effect size; ORIC = order-restricted information criterion; BMS = Bayesian model selection.

Part II

Generalizations in Confirmatory Model Selection

Buiten is het vrijdag by Marga Klungel

CHAPTER *4*

An Akaike-Type Information Criterion for Model Selection under Inequality Constraints

Kuiper, R. M., Hoijtink, H., and Silvapulle, M. J.

Published in *Biometrika*, 98(2), pp. 495-501.

The Akaike information criterion for model selection presupposes that the parameter space is not subject to order restrictions or inequality constraints. Anraku (1999) proposed a modified version of this criterion, called the order-restricted information criterion, for model selection in the one-way analysis of variance model when the population means are monotonic. We propose a generalization of this to the case when the population means may be restricted by a mixture of linear equality and inequality constraints. If the model has no inequality constraints, then the generalized order-restricted information criterion coincides with the Akaike information criterion. Thus, the former extends the applicability of the latter to model selection in multi-way analysis of variance models when some models may have inequality constraints while others may not. Simulation shows that the information criterion proposed in this chapter performs well in selecting the correct model.

The Akaike information criterion (Akaike, 1973) is among the most widely used criteria for model selection. This criterion assumes that the parameter space of the model is not restricted by inequality constraints of the form $\theta_1 \leq \theta_2$, where θ_1 and θ_2 are unknown parameters. In this note, we propose an Akaike-type information criterion for the analysis variance model when the treatment means $\{\theta_1, \ldots, \theta_p\}$ are assumed to satisfy a mixture of linear equality and inequality constraints.

To illustrate the essentials, let us consider the analysis of variance model

$$y_{ij} \sim N(\theta_i, \sigma^2) \quad (i = 1, \ldots, p, j = 1, \ldots, n_i), \tag{4.1}$$

with independent observations from p normal populations, and let $\theta = (\theta_1, \ldots, \theta_p)^\top$. This setting is general enough to incorporate factorial designs as well. For model (4.1), AIC $= -2\{$maximum log likelihood $-$ number of parameters$\}$ when θ is not subject to inequality constraints. However, in many studies, prior information such as that the new treatment is at least as good as the old treatment, which may take the form $\theta_1 \leq \theta_2$, is available. In such cases, the Akaike information criterion is not suitable

for model selection. When θ satisfies the simple order $\theta_1 \leq \cdots \leq \theta_p$, Anraku (1999) proposed the order-restricted information criterion

$$\text{ORIC} = -2\left\{\text{maximum log likelihood} - 1 - \sum_{i=1}^{p} iw_i\right\}, \tag{4.2}$$

where the constants $\{w_0, \ldots, w_p\}$ are the so-called level probabilities for the simple order $\theta_1 \leq \cdots \leq \theta_p$. In this note, we propose an extension of (4.2), called the generalized order-restricted information criterion, to the case when the parameter θ may be restricted by $R\theta \geq 0$ where R is any matrix of known constants.

The form $R\theta \geq 0$ is general enough to accommodate practically any linear inequality constraints encountered in practice. Some examples to which the generalized order-restricted information criterion is applicable include the simple order, the tree order $\theta_1 \leq \theta_2, \ldots, \theta_1 \leq \theta_p$, and the matrix order (Silvapulle & Sen, 2005, pp. 43, 296). By contrast, (4.2) is applicable only when $\theta_1 \leq \cdots \leq \theta_p$; thus, for example, it is not applicable for the tree order or the matrix order, even after transformation of the model.

4.1 The generalized order-restricted information criterion

4.1.1 Preliminaries

Consider the analysis of variance model (4.1) with θ subject to $R\theta \geq 0$, where R is a $r \times p$ matrix. Let $n = n_1 + \cdots + n_p$ be the total number of observations. It follows from (4.1) that the log likelihood is

$$\ell(\theta, \sigma) = -\frac{1}{2} n \log(2\pi\sigma^2) - \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left\{\frac{y_{ij} - \theta_i}{\sigma}\right\}^2.$$

Let $\ell_0(\theta^*, \sigma^*)$ denote the expected log likelihood function $E\{\ell(\theta^*, \sigma^*) \mid \theta, \sigma\}$ at an arbitrary point (θ^*, σ^*) in the parameter space, where the expectation is evaluated at the true value (θ, σ). Then

$$\ell_0(\theta^*, \sigma^*) = -\frac{1}{2} n\{\log(2\pi) + \log \sigma^{*2}\} - \frac{1}{2}\left[n\left(\frac{\sigma}{\sigma^*}\right)^2 + \sum_{i=1}^{p} n_i \left\{\frac{\theta_i - \theta_i^*}{\sigma^*}\right\}^2\right].$$

Let $(\tilde{\theta}, \tilde{\sigma})$ denote the maximum likelihood estimator of (θ, σ) under equality and/or inequality constraints, if there are any. The objective of an information criterion-based approach is to choose the model for which $\ell_0(\tilde{\theta}, \tilde{\sigma})$, the expected log likelihood at the maximum likelihood estimator $(\tilde{\theta}, \tilde{\sigma})$, is maximized. However, $\ell_0(\cdot)$ depends on the unknown population distribution, and therefore, the standard method is to use an estimator of $\ell_0(\tilde{\theta}, \tilde{\sigma})$. A natural estimator of this is the maximum log likelihood, $\ell(\tilde{\theta}, \tilde{\sigma})$. However, this is not a good estimator because its bias $B(\theta, \sigma) = E\{\ell(\tilde{\theta}, \tilde{\sigma}) - \ell_0(\tilde{\theta}, \tilde{\sigma}) \mid \theta, \sigma\}$ does not reduce to zero as $n \to \infty$. Details of these derivations for the case when there are no inequality constraints are well known (for example, see Claeskens &

Hjort, 2008; Hurvich & Tsai, 1989; McQuarrie & Tsai, 1998; Gourieroux & Monfort, 1995, §22.3.2). If θ is restricted by $R\theta \geq 0$, then

$$B(\theta,\sigma) = -\frac{n}{2} + \frac{n}{2}E\left\{\left(\frac{\sigma}{\tilde{\sigma}}\right)^2 \middle| \theta,\sigma\right\} + \frac{1}{2}E\left\{\sum_{i=1}^{p} n_i \frac{(\tilde{\theta}_i-\theta_i)^2}{\tilde{\sigma}^2} \middle| \theta,\sigma\right\}.$$

Let us temporarily suppose that there are no constraints on θ, and let $(\hat{\theta},\hat{\sigma})$ denote the unconstrained maximum likelihood estimator of (θ,σ). Then the bias $E\{\ell(\hat{\theta},\hat{\sigma}) - \ell_0(\hat{\theta},\hat{\sigma}) \mid \theta,\sigma\}$ in estimating $\ell_0(\hat{\theta},\hat{\sigma})$ by $\ell(\hat{\theta},\hat{\sigma})$ is $p+o(1)$. Therefore, an asymptotically unbiased estimator of $\ell_0(\hat{\theta},\hat{\sigma})$ is $\{\ell(\hat{\theta},\hat{\sigma}) - p\}$, which is proportional to AIC.

In the inequality constrained case, suppose that θ is subject to $R\theta \geq 0$. Now, $B(\theta,\sigma)$ is no longer constant up to a term of order $o(1)$, and therefore, the bias cannot be removed by subtracting a constant. For this setting, Anraku (1999) proposed the order-restricted information criterion $-2\{\ell(\tilde{\theta},\tilde{\sigma}) - \inf_{\theta,\sigma} B(\theta,\sigma)\}$, which resembles the AIC and is most favourable to the parametric model. Because $B(\theta,\sigma)$ depends on the particular inequality constraints, a challenge is to find simple and practical ways of computing $\inf_{\theta,\sigma} B(\theta,\sigma)$ for different inequality constraints.

For the simple order restriction $\theta_1 \leq \cdots \leq \theta_p$, Anraku (1999) showed that $\inf_{\theta,\sigma} B(\theta,\sigma)$ has the closed form $(1 + \sum_{i=1}^{p} i w_i)$, which in turn led to (4.2). A main contribution of this note is to provide a similar simple closed form for $\inf_{\theta,\sigma} B(\theta,\sigma)$ when θ is restricted by $R\theta \geq 0$.

4.1.2 A closed form for the penalty term $\inf_{\theta,\sigma} B(\theta,\sigma)$

Let $W = \mathrm{diag}\{n_1^{-1},\ldots,n_p^{-1}\}$, the diagonal matrix with the ith diagonal being n_i^{-1} $(i = 1,\ldots,p)$. Let $\mathcal{C} = \{\theta^* : R\theta^* \geq 0\}$, $X \sim N(0,W)$ and $\tilde{X} = \arg\min_x\{(X-x)^\top W^{-1}(X-x) : x \in \mathcal{C}\}$. Now, let $\{w_i(p,W,\mathcal{C}), i = 0,\ldots,p\}$ be the nonnegative constants that are uniquely defined and appear in the chi-bar square distribution, $\mathrm{pr}(\tilde{X}^\top W^{-1}\tilde{X} \leq c) = \sum_{i=0}^{p} w_i(p,W,\mathcal{C})\,\mathrm{pr}(\chi_i^2 \leq c)$. These constants, also known as chi-bar square weights, arise naturally in constrained statistical inference where their computation has been studied in detail. For details and references, see §3.5 in Silvapulle and Sen (2005) and Silvapulle (1996).

The crucial result to extend order-restricted information criterion to accommodate more general order restrictions is the following.

Proposition 1. *Consider the normal theory analysis of variance model (4.1). Let $\mathcal{C}$ be a closed convex cone in R^p or $\mathcal{C} = R^p$. Let $\theta \in \mathcal{C}$ and $\sigma > 0$. Then $1 + \sum_{i=1}^{p} i w_i(p,W,\mathcal{C}) + O(n^{-1}) \leq B(\theta,\sigma) \leq (1+p) + O(n^{-1})$, where the lower bound is reached if and only if θ is in the largest linear space contained in $\mathcal{C}$.*

This result is applicable when the constraints are of the form $R\theta \geq 0$ because $\mathcal{C}$ can then be taken to be $\{\theta \in R^p : R\theta \geq 0\}$. In view of the greatest lower bound for $B(\theta,\sigma)$ in Proposition 1, and the form $-2\{\ell(\tilde{\theta},\tilde{\sigma}) - \inf_{\theta,\sigma} B(\theta,\sigma)\}$ for the information criterion, we define the generalized order-restricted information criterion

$$\mathrm{GORIC} = -2\left\{\ell(\tilde{\theta},\tilde{\sigma}) - 1 - \sum_{i=1}^{p} i w_i(p,W,\mathcal{C})\right\}. \tag{4.3}$$

We propose that the model for which this is the minimum to be chosen. For the special case of simple order, (4.3) reduces to (4.2). Suppose that there are no inequality constraints on θ. Then, Proposition 1 is applicable with $\mathcal{C} = R^p$. With this choice, we have $w_p(p, W, \mathcal{C}) = 1$, $w_i(p, W, \mathcal{C}) = 0$ for $i < p$, and thus, generalized order-restricted information criterion reduces to AIC.

The approach proposed in this chapter shares a consistency property with the traditional Akaike information criterion approach. To establish this, let us consider the two models $H_a : \theta \in \mathcal{C}_a$ and $H_b : \theta \in \mathcal{C}_b$, where $\mathcal{C}_a$ and $\mathcal{C}_b$ are closed convex cones and are not equal. Suppose that the true parameter θ lies in $\mathcal{C}_a$ and not in $\mathcal{C}_b$. Then, by mimicking the arguments in Anraku (1999) for his Theorem 4, we have that $n^{-1}(\text{GORIC}^a - \text{GORIC}^b)/(-2) = c + o_p(1)$, where $c = E[\log\{f(y; \theta, \sigma)\} \mid \theta, \sigma] - \log\{f(y; \theta^b, \sigma^b)\} > 0$, and (θ^b, σ^b) is the probability limit of the maximum likelihood estimator of (θ, σ) under model $\mathcal{C}_b$. Hence, the correct model will be chosen with probability going to 1 for $n \to \infty$.

The only computer program required to compute generalized order-restricted information criterion is a quadratic program. Since such programs are available in many mathematical and statistical computer software, computation of generalized order-restricted information criterion does not encounter any serious difficulties. The computer time required to compute the penalty term $\{1 + \sum_{i=1}^{p} i w_i(p, W, \mathcal{C})\}$ in generalized order-restricted information criterion does not depend on the number of observations in the sample, but only on the dimension of θ and the nature of inequality constraints on θ. In most practical settings, the computation of generalized order-restricted information criterion would take only a matter of seconds.

4.2 An example

Zelano, Zelano, and Kolb (1972) conducted an experiment to evaluate the effect of exercise on the age y at which a child starts to walk. The data are available in Silvapulle and Sen (2005, p. 34). Each of the four groups received a different walking exercise. The first group received a seven-week walking exercise for twelve minutes a day beginning at the age of one week. The second group received a daily exercise, but not a daily walking exercise. The third group did not receive any exercises, and serves as control group. The fourth group also did not receive any exercise, but they were checked weekly for progress. The model used here is (4.1), with $p = 4$, $n_1 = n_2 = n_4 = 6, n_3 = 5$, and θ_i the mean age in months at which a child starts to walk $(i = 1, \ldots, 4)$.

Because the effect of the exercises is not completely understood, several different possible competing hypotheses are of interest. One possible hypothesis is H_1 in Table 4.1, that the mean age decreases with increasing intensity of exercise. Another is H_2 in Table 4.1, that Treatments 1 and 2 are at least as good as Treatments 3 and 4, but no ordering is suggested between Treatments 1 and 2, or between Treatments 3 and 4.

The hypotheses H_0, H_1, H_2 and H_u in Table 4.1 have different inequality constraints on θ. Consequently, generalized order-restricted information criterion has different values for these hypotheses. Table 4.1 suggests that, in terms of generalized

Table 4.1: *Estimates of the penalty term* $\inf_{(\theta,\sigma)} B(\theta,\sigma)$*, the log likelihood* $\ell(\tilde{\theta},\tilde{\sigma}^2)$ *and generalized order-restricted information criterion*

Hypothesis	$\inf_{(\theta,\sigma)} B(\theta,\sigma)$	$\ell(\tilde{\theta},\tilde{\sigma}^2)$	GORIC
$H_0 : \theta_1 = \theta_2 = \theta_3 = \theta_4$	2·00	−43·36	90·73
$H_1 : \theta_1 \leq \theta_2 \leq \theta_4 \leq \theta_3$	3·10	−40·01	86·23
$H_2 : \theta_1 \leq \theta_3, \theta_2 \leq \theta_3, \theta_1 \leq \theta_4, \theta_2 \leq \theta_4$	3·61	−40·01	87·25
H_u : No restrictions on the parameters	5·00	−40·01	90·03

GORIC, generalized order-restricted information criterion.

order-restricted information criterion, model H_1 fits better than the other three. The traditional approach based on AIC is unable to provide such a discrimination between these models. Because the order restriction in H_2 cannot be expressed as a simple order, the method in Anraku (1999) is inadequate to compare H_2 with the other models in Table 4.1. In this sense, generalized order-restricted information criterion extends the applicability of generalized order-restricted information criterion to more general linear order restrictions.

4.3 Simulation

A simulation study was carried out to evaluate the performance of the generalized order-restricted information criterion, using the design of a real data example. Berzonsky, Kleven, and Leach (2003) studied the effects of parthenogenesis on wheat embryo formation in the presence and in the absence of maize pollination. This experiment was conducted as a balanced 4×2 factorial design in a glass house. The response variable y is a measure of embryo formation. Factor A is genotype with four levels, and Factor B is maize pollination with two levels. Berzonsky et al. (2003) studied the two-way analysis of variance model, $Y_{ijk} = \mu_{ij} + \eta_{ijk}$ $(i = 1, \ldots, 4,$ $j = 1, 2,$ $k = 1, \ldots, 20)$. They also discussed possible orderings of the cell mean parameters based on prior knowledge about the relationship among the cell means. The main one is stated below as H_1. One use of a well-fitting model in the context of this study is prediction of embryo formation under different experimental conditions.

To apply the results of this chapter, let us rewrite the foregoing model as $y_{ij} = \theta_i + \varepsilon_{ij}$ $(i = 1, \ldots, 8,$ $j = 1, \ldots, 20)$. Now, let us define H_u as the model with no restrictions on θ, $H_0 : \theta_1 = \theta_2 = \theta_3 = \theta_4, \theta_5 = \theta_6 = \theta_7 = \theta_8$, and

$$H_1 : \ \theta_1 \geq \{\theta_2, \theta_3, \theta_4\}, \theta_5 \geq \{\theta_6, \theta_7, \theta_8\}, \theta_1 \geq \theta_5, \theta_2 \geq \theta_6, \theta_3 \geq \theta_7, \theta_4 \geq \theta_8,$$
$$\theta_1 - \theta_5 \geq \{\theta_2 - \theta_6, \theta_3 - \theta_7, \theta_4 - \theta_8\}.$$

To choose suitable parameter values for the simulation, we used the effect size (Cohen, 1992) and the true hypothesis to guide us. Nine different values for the vector of population means were studied. For each value of θ, estimates were obtained using 1,000 independent samples. Based on these, we computed the percentage of times that the correct model was chosen. Table 4.2 shows that the method introduced in this chapter, selected the correct model at least 90% of the times, when the effect size was

Table 4.2: *Percentage of times that different models were chosen by the generalized order-restricted information criterion*

	Case 1: H_0 is true			Case 2: H_1 is true			Case 3: H_u is 'true'		
ES	H_0	H_1	H_u	H_0	H_1	H_u	H_0	H_1	H_u
$0 \cdot 1$	84	9	7	48	49	3	60	30	10
$0 \cdot 25$	91	1	9	7	92	1	18	35	47
$0 \cdot 4$	91	0	9	0	99	1	1	9	91

greater or equal to 0.25. When the effect size was equal to 0.1, the method selected H_0 more often, as expected. More simulation results are given in Appendix 4.B and the computer program for computing generalized order-restricted information criterion is available from `http://staff.fss.uu.nl/RMKuiper`.

4.A Proofs

Let W be a positive definite matrix of order $p \times p$. Let $\|x\|^2$ denote the squared length $x^\top W^{-1} x$, $\langle x, y \rangle$ denote the inner product $x^\top W^{-1} y$, and $P(x \mid \mathcal{C})$ denote the projection of x onto $\mathcal{C}$ defined by $\arg\min_{c \in \mathcal{C}} \|x - c\|$. Thus, $P(x \mid \mathcal{C})$ is the point in $\mathcal{C}$ closest to x with respect to the distance $\|\cdot\|$. For any set $A \subset R^p$, let $\rho(x, A)$ denote the distance $\inf_{a \in A} \|x - a\|$ between the point x and the set A. Let $\mathcal{M}$ be the largest linear space contained in $\mathcal{C}$.

Lemma 1. *Let $X \in R^p$, $\theta_0 \in \mathcal{M}, \theta_1 \in \mathcal{C}$, and $\lambda = \theta_1 - \theta_0$. Then (i) $\|X - P(X \mid \mathcal{C})\| \geq \|(X + \theta_1) - P(X + \theta_1 \mid \mathcal{C})\|$, and (ii) $\|P(X \mid \mathcal{C}) - \theta_0\| \leq \|P(X + \lambda \mid \mathcal{C}) - (\lambda + \theta_0)\|$.*

The first part of the lemma follows from $\mathcal{C} \subset \mathcal{C} - \theta_1$. The second part follows from Nomakuchi (2002, Thm 2.1).

Lemma 2. *Let $R\theta_b \geq 0$ and $R\theta_a = 0$. For a given vector of error terms E, let $Y_{aij} = \theta_{ai} + E_{ij}$ and $Y_{bij} = \theta_{bi} + E_{ij}$. Let the suffices a and b correspond to θ_a and θ_b respectively. Then, (i) $\tilde{\sigma}_b^2 \leq \tilde{\sigma}_a^2$, (ii) $\|\tilde{\theta}_b - \theta_b\|^2 \geq \|\tilde{\theta}_a - \theta_a\|^2$ and (iii) $B_2(\theta_b, \sigma) \geq B_2(\theta_a, \sigma)$.*

Proof. Let $\lambda = \theta_b - \theta_a$. If $R\gamma \geq 0$, then there exists a θ^* such that $R\theta^* \geq 0$ and $\gamma = \theta^* - \lambda$. Therefore, $n\tilde{\sigma}_b^2 = \min_{R\theta^* \geq 0} \sum_{ij} (Y_{bij} - \theta_i^*)^2 = \min_{R\theta^* \geq 0} \sum_{ij} \{\theta_{ai} + E_{ij} - (\theta_i^* - \lambda_i)\}^2 \leq \min_{R\gamma \geq 0} \sum_{ij} (\theta_{ai} + E_{ij} - \gamma_i)^2 = n\tilde{\sigma}_a^2$. By Lemma 1(ii), we have $\|\tilde{\theta}_b - \theta_b\|^2 = \|P(\hat{\theta}_b \mid \mathcal{C}) - \theta_b\|^2 = \|P(\hat{\theta}_a + \lambda \mid \mathcal{C}) - (\theta_a + \lambda)\|^2 \geq \|P(\hat{\theta}_a \mid \mathcal{C}) - \theta_a\|^2 = \|\tilde{\theta}_a - \theta_a\|^2$. Part (iii) follows from (i) and (ii) in Lemma 2.

Lemma 3. *Suppose that $R\theta = 0$. Then $E\{\sigma^2/\tilde{\sigma}^2 \mid (\theta, \sigma)\} = 1 + n^{-1} \sum_{i=0}^{p} i w_i(p, W, \mathcal{C}) + 2n^{-1} + O(n^{-2})$.*

Proof. We use the lemmas from Silvapulle and Sen (2005, pp. 125–132) without further comment. Let $\{F_1, \ldots F_m\}$ be the partition of $\mathcal{C} = \{x \in R^p : Rx \geq 0\}$, where each F_s is the relative interior of a face of $\mathcal{C}$ $(s = 1, \ldots, m)$; for a definition

of face see Silvapulle and Sen (2005, p. 124). Let $S_s = \{x \in R^p : P(x \mid \mathcal{C}) \in F_s\}$ $(s = 1, \ldots, m)$. Then $\{S_1, \ldots, S_m\}$ is a partition of R^p, except for a set of measure zero. Let L_s denote the linear space spanned by F_s for $s = 1, \ldots, m$. By arguments similar to the proof of Theorem 3.4.2 in Silvapulle and Sen (2005), we have

$$\begin{aligned}\operatorname{pr}\left(\frac{\tilde{\sigma}^2}{\sigma^2} \le t\right) &= \sum_{i=0}^{p} \sum_{\text{over all } s \text{ with } \dim(L_s)=p-i} \operatorname{pr}(\hat{\theta} \in S_s) \operatorname{pr}\left(\frac{\tilde{\sigma}^2}{\sigma^2} \le t \,\middle|\, \hat{\theta} \in S_s\right) \\ &= \sum_{i=0}^{p} w_{p-i}(p, W, \mathcal{C}) \operatorname{pr}(\chi^2_{n-p+i} \le nt).\end{aligned}$$

Now, with $N_i = (n - p + i)/2$, we have

$$\begin{aligned}E\left(\frac{\sigma^2}{\tilde{\sigma}^2}\right) &= \int_0^\infty t^{-1} d\left\{\operatorname{pr}\left(\frac{\tilde{\sigma}^2}{\sigma^2}\right) \le t\right\} = \int_0^\infty t^{-1} d\left\{\sum_{i=0}^{p} w_{p-i}(p, W, \mathcal{C}) \operatorname{pr}(\chi^2_{n-p+i} \le nt)\right\} \\ &= \sum_{i=0}^{p} w_{p-i}(p, W, \mathcal{C}) \int_0^\infty t^{-1} \{\Gamma(N_i) 2^{N_i}\}^{-1} \exp\left(\frac{-nt}{2}\right) (nt)^{N_i - 1} n \, dt \\ &= \frac{n}{2} \sum_{i=0}^{p} w_{p-i}(p, W, \mathcal{C})(N_i - 1)^{-1} \\ &= 1 + n^{-1}\left\{2 + \sum_{i=0}^{p} i w_i(p, W, \mathcal{C})\right\} + O(n^{-2}).\end{aligned}$$

Lemma 4. *If $R\theta = 0$, then, $E(\tilde{\sigma}^{-2}\|\tilde{\theta} - \theta\|^2) = \sum_{i=0}^{p} i w_i(p, W, \mathcal{C}) + O(n^{-1})$.*

Proof. Let F_s, S_s, and L_s, $s = 1, \ldots, m$, be the same as those in the proof of the previous lemma. Conditional on $\{\hat{\theta} \in S_s\}$, $\|\tilde{\theta} - \theta\|^2$ and $n\tilde{\sigma}^2$ are independent and are distributed as χ^2_i and χ^2_{n-i} respectively, where $i = \dim(L_s)$. Now,

$$\begin{aligned}\operatorname{pr}(\tilde{\sigma}^{-2}\|\tilde{\theta} - \theta\|^2 \le c) &= \sum_{i=0}^{p} \sum_{\text{over all } s \text{ with } \dim(L_s)=i} \operatorname{pr}(\hat{\theta} \in S_s) \operatorname{pr}(\tilde{\sigma}^{-2}\|\tilde{\theta} - \theta\|^2 \le c \mid \hat{\theta} \in S_s) \\ &= \sum_{i=0}^{p} w_i(p, W, \mathcal{C}) \operatorname{pr}\left(\frac{ni}{n-i} F_{i,n-i} \le c\right).\end{aligned}$$

Hence, $E(\tilde{\sigma}^{-2}\|\tilde{\theta} - \theta\|^2) = \sum_{i=0}^{p} i w_i(p, W, \mathcal{C})\{1 + O(n^{-1})\} = \sum_{i=0}^{p} i w_i(p, W, \mathcal{C}) + O(n^{-1})$.

4.B Additional information on the simulation study

A simulation study was conducted to evaluate the performance of GORIC. For this simulation study, we chose the design of a real data example for which GORIC would be useful, namely the one of Berzonsky et al. (2003). The model used in this study

can be written as $y_{ij} = \theta_i + \varepsilon_{ij}$ where $i = 1, \ldots, 8$ and $j = 1, \ldots, 20$. The following three models/hypotheses were studied:

$$
\begin{aligned}
H_0 &: \ \theta_1 = \theta_2 = \theta_3 = \theta_4, \theta_5 = \theta_6 = \theta_7 = \theta_8, \\
H_1 &: \ \theta_1 \geq \{\theta_2, \theta_3, \theta_4\}, \theta_5 \geq \{\theta_6, \theta_7, \theta_8\}, \theta_1 \geq \theta_5, \theta_2 \geq \theta_6, \theta_3 \geq \theta_7, \theta_4 \geq \theta_8, \text{ and} \\
&\quad\ \ \theta_1 - \theta_5 \geq \{\theta_2 - \theta_6, \theta_3 - \theta_7, \theta_4 - \theta_8\}, \\
H_u &: \ \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8 \text{ are unrestricted.}
\end{aligned}
$$

To acquire a suitable set of parameter values for the simulation, we used the effect size (Cohen, 1992) to guide us. The effect size of an analysis of variance model is measured by $ES = \sigma_m/\sigma$, with σ_m the standard deviation of the $k = 8$ population means and σ the standard deviation of the population, here set to 1. An effect size (ES) of $0 \cdot 1$ is called low, $ES = 0 \cdot 25$ medium, and $ES = 0 \cdot 4$ high. Based on these three values and on the true hypothesis, we obtained nine different values for the vector of population means (see Table 4.5).

For each value of θ, we generated 1,000 independent samples. For each simulated data set we calculated the GORIC for the three models of interest. The percentage of times the correct model was chosen (with $n_i = 20$ for all i) are given in Table 4.2. This table shows that the GORIC selects the correct model at least 90% of the times, when the effect size is greater or equal to $0 \cdot 25$. When the effect size was small, that is, lower than $0 \cdot 25$, the GORIC selects H_0 more often, as expected. Namely, with small effect sizes, large samples would be required to be able to distinguish between the two models. In fact, Table 4.4 displays that the proportion of times the GORIC selects the correct model increases with increasing sample size.

Table 4.3: Percentage of times that the models H_0, H_1, and H_u were chosen by the GORIC for $n_i = 20$

	Case 1: H_0 is true			Case 2: H_1 is true			Case 3: H_u is 'true'		
ES	H_0	H_1	H_u	H_0	H_1	H_u	H_0	H_1	H_u
$0 \cdot 1$	84	9	7	48	49	3	60	30	10
$0 \cdot 25$	91	1	9	7	92	1	18	35	47
$0 \cdot 4$	91	0	9	0	99	1	1	9	91

Table 4.4: Percentage of times that the models H_0, H_1, and H_u were chosen by the GORIC for $n_i = 50$

	Case 1: H_0 is true			Case 2: H_1 is true			Case 3: H_u is 'true'		
ES	H_0	H_1	H_u	H_0	H_1	H_u	H_0	H_1	H_u
$0 \cdot 1$	91	4	5	37	63	1	49	33	19
$0 \cdot 25$	94	0	6	0	99	1	1	10	89
$0 \cdot 4$	94	0	6	0	99	1	0	0	100

Table 4.5: Values of Population Means ($\theta_1 \ldots, \theta_8$)

ES	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8
				Case 1: H_0 is true				
$0 \cdot 1$	$0 \cdot 00$	$0 \cdot 00$	$0 \cdot 00$	$0 \cdot 00$	$0 \cdot 20$	$0 \cdot 20$	$0 \cdot 20$	$0 \cdot 20$
$0 \cdot 25$	$0 \cdot 00$	$0 \cdot 00$	$0 \cdot 00$	$0 \cdot 00$	$0 \cdot 50$	$0 \cdot 50$	$0 \cdot 50$	$0 \cdot 50$
$0 \cdot 4$	$0 \cdot 00$	$0 \cdot 00$	$0 \cdot 00$	$0 \cdot 00$	$0 \cdot 80$	$0 \cdot 80$	$0 \cdot 80$	$0 \cdot 80$
				Case 2: H_1 is true				
$0 \cdot 1$	$0 \cdot 31675$	$0 \cdot 125$	$0 \cdot 125$	$0 \cdot 125$	$0 \cdot 0625$	$0 \cdot 00$	$0 \cdot 00$	$0 \cdot 00$
$0 \cdot 25$	$0 \cdot 80050$	$0 \cdot 125$	$0 \cdot 125$	$0 \cdot 125$	$0 \cdot 0625$	$0 \cdot 00$	$0 \cdot 00$	$0 \cdot 00$
$0 \cdot 4$	$1 \cdot 26086$	$0 \cdot 125$	$0 \cdot 125$	$0 \cdot 125$	$0 \cdot 0625$	$0 \cdot 00$	$0 \cdot 00$	$0 \cdot 00$
				Case 3: $\theta_1 > \ldots > \theta_4, \theta_5 < \ldots < \theta_8$, so, H_u is 'true'				
$0 \cdot 1$	$0 \cdot 34915$	$0 \cdot 26186$	$0 \cdot 17457$	$0 \cdot 087287156$	$0 \cdot 04364$	$0 \cdot 130930734$	$0 \cdot 21822$	$0 \cdot 30551$
$0 \cdot 25$	$0 \cdot 87287$	$0 \cdot 65465$	$0 \cdot 43644$	$0 \cdot 218217890$	$0 \cdot 10911$	$0 \cdot 327326835$	$0 \cdot 54554$	$0 \cdot 76376$
$0 \cdot 4$	$1 \cdot 39659$	$1 \cdot 04745$	$0 \cdot 69830$	$0 \cdot 349148624$	$0 \cdot 17457$	$0 \cdot 523722937$	$0 \cdot 87287$	$1 \cdot 22202$

CHAPTER 5

Generalization of the Order-Restricted Information Criterion for Multivariate Normal Linear Models

Kuiper, R. M., Hoijtink, H., and Silvapulle, M. J.

Manuscript submitted

The order-restricted information criterion (ORIC) of Anraku (1999) is a model selection criterion for analysis of variance models that can be applied to simple order restrictions, which are of the form $\theta_1 \leq \ldots \leq \theta_k$, where θ_j is the mean of group j. We generalize the ORIC to an information criterion, called the GORIC, which can be applied to univariate and multivariate normal linear models for a more general form of order restrictions: $R\theta \leq 0$, where θ is a vector of length tk and R a $c_m \times tk$ matrix. Subsequently, we show that the GORIC works for restrictions of the form $R\theta \leq r$ as well, with r is a vector of length c_m and R a $c_m \times tk$ matrix of full rank.

Like the ORIC, the GORIC selects the best of a set of hypotheses and consists of a likelihood part and a penalty part. We describe how the penalty of the GORIC is assessed by means of simulation. At the end of this chapter, we illustrate that the GORIC is easy to implement in a multivariate regression model.

5.1 Introduction

Anraku (1999) proposes the order-restricted information criterion (ORIC), which is used for models of the form $y_{ij} = \theta_j + \epsilon_{ij}$, where y_{ij} is observation i (with $i = 1, \ldots, N_j$) for group j (with $j = 1, \ldots, k$), θ_j is the mean for group j, and ϵ_{ij} is the error term, which follows a normal distribution with mean 0 and variance σ^2. This model selection criterion can only be applied to hypotheses which can be written as simple order restrictions: $H_m : \theta_1 \leq \ldots \leq \theta_k$, where "$\leq$" may be replaced by "$=$". Furthermore, Hughes and King (2003) derive the so-called one-sided AIC which is "applicable in problems where the signs of some or all the parameters are known or can be inferred on basis of a priori information". In addition, Kuiper, Hoijtink, and Silvapulle (2011) propose a generalization of the order-restricted information criterion, GORIC. This model selection criterion can be used for a more general form of order restrictions:

$R\theta \leq 0$, where θ is a vector of length k and R a $c_m \times k$ matrix. However, the derivation is only done for analysis of variance models.

In this chapter, we derive a generalization of the ORIC, called the GORIC, that can be used for t-variate normal linear models to evaluate commonly encountered hypotheses of the type $H_m : \theta \in \mathcal{C}_m$ for $m \in \mathcal{M}$, where θ is a vector of length tk, $\mathcal{C}_m$ is a closed convex cone, and $\mathcal{M}$ the set of hypothesis indices. Special cases of closed convex cones are simple order restrictions and matrix order restrictions (Silvapulle & Sen, 2005, pp. 82). The latter are of the type $H_m : R\theta \leq 0$, with R a $c_m \times tk$ matrix. The set of restrictions can contain equality restrictions as well: $H_m : R_1\theta \leq 0, R_2\theta = 0$, where R_1 is a $c_{m1} \times tk$ matrix and R_2 a $c_{m2} \times tk$ matrix. Namely, $H_m : R_1\theta \leq 0, R_2\theta = 0$ equals $H_m : R_1\theta \leq 0, R_2\theta \leq 0, -R_2\theta \leq 0$ which can be rewritten as $H_m : R\theta \leq 0$. In this chapter, we prove that the GORIC can also be applied to $H_m : R\theta \leq r$ when R is of full rank, with r a vector of length c_m. Namely, $H_m : R\theta \leq r$ is a relocated closed convex cone or, stated otherwise, it is a close convex cone for shifted data, when R is of full rank. Note that the matrix of full rank may be obtained by discarding redundant restrictions. As before, equality restrictions can be a part of these type of restrictions, that is, $H_m : R_1\theta \leq r_1, R_2\theta = r_2$ is also a relocated closed convex cone when $[R_1', R_2']'$ is of full rank (after discarding redundant restrictions), with r_1 a vector of length c_{m1} and r_2 of c_{m2}. Notably, a hypothesis including the restrictions $\theta_l \geq r_{11}$ and $\theta_l \leq r_{12}$ is not a (closed) convex cone for $r_{11} \neq r_{12}$, with θ_l the lth element of θ for $l = 1, \ldots, tk$, since in that case R is not of full rank and there are no redundant restrictions. The same remains valid for $\theta_l \geq r_{11}, \theta_{l'} \geq r_{12}, \theta_l + \theta_{l'} \leq r_{13}$ for $l \neq l'$. For $H_m : \theta_l \leq r_{11}, \theta_{l'} \leq r_{12}, \theta_l + \theta_{l'} \leq r_{13}$, R is also not of full rank. However, when $r_{11} + r_{12} \geq r_{13}$, there is a redundant restriction, namely $\theta_l + \theta_{l'} \leq r_{13}$. When discarding it, R is of full rank. Hence, in that case $H_m : \theta_l \leq r_{11}, \theta_{l'} \leq r_{12}, \theta_l + \theta_{l'} \leq r_{13}$ equals $H_m : \theta_l \leq r_{11}, \theta_{l'} \leq r_{12}$ which is a relocated closed convex cone.

The derivation of the GORIC is done in three steps. In the next section, Section 5.2, we give the main part of the derivation of the GORIC for simple multivariate models of the form $y_i = \theta + \epsilon_i$, where y_i (for $i = 1, \ldots, N$) is the vector of the k dependent variables for observation i (i.e., $y_i = [y_{i1}, \ldots, y_{ik}]'$), θ the vector of the k group means, and ϵ_i the vector of the k error terms for observation i. This forms a good starting point for the derivation of the GORIC for univariate normal linear models, which is discussed in Section 5.3.1. Finally, we extend it to multivariate normal linear models in Section 5.3.2. The demonstration that the GORIC can be applied to relocated close convex cones is given in Section 5.4. Like the ORIC, the GORIC incorporates a likelihood part and a penalty part. In Section 5.5, we proceed by demonstrating (based on Silvapulle & Sen, 2005, pp. 78–81) how the penalty part of the GORIC is obtained by simulation. There, we assume that the residual covariance matrix is equal to an unknown positive constant (σ^2) times a known scale matrix. We discuss briefly what consequences there are if this scale matrix is not known but estimated from the data in Section 5.6. We end, in Section 5.7, with an illustration of the GORIC in a multivariate regression model.

5.2 Derivation of the GORIC

5.2.1 Preliminaries

The GORIC is based on the Kullback–Leibler discrepancy (Kullback & Leibler, 1951):

$$E_{g(y|\xi)}\left\{-2\log f(y|\xi^m)\right\} = -2\int_{-\infty}^{\infty}\log\left\{f(y|\xi^m)\right\}g(y|\xi)\;dy,$$

where $g(y|\xi)$ is the true generating model, with ξ the true parameter, and $f(y|\xi^m)$ a candidate model or hypothesis, that is, a statistical model to approximate $g(y|\xi)$, with ξ^m the parameter corresponding to hypothesis H_m. The preferred hypothesis is the one which renders the lowest Kullback–Leibler discrepancy (parameterized by $\xi^m = \hat{\xi}^m$). Therefore, the GORIC, like the ORIC, is an estimate of the Kullback–Leibler discrepancy or rather of minus two times the expected log-likelihood.

Let the data be $y = [y_1, \ldots, y_N] \in R^{k\times N}$, with $y_i = [y_{i1}, \ldots, y_{ik}]' \in R^{k\times 1}$. It is assumed that the data y_i are normally and independently distributed with means θ and covariance matrix V:

$$y_i \sim \mathcal{N}_k(\theta, V) \text{ for } i = 1, \ldots, N, \tag{5.1}$$

where $\theta \in R^{k\times 1}$ and V a $k \times k$ positive definite matrix. Models of this type are used in multivariate one-sided testing and repeated measures analysis.

The log-likelihood of y is written as

$$\begin{aligned}\log f(y|\theta, V) &= \sum_{i=1}^{N}\left[-\frac{k}{2}\;\log\{2\pi\} - \frac{1}{2}\log|V| - \frac{1}{2}(y_i-\theta)'V^{-1}(y_i-\theta)\right]\\ &= -\frac{Nk}{2}\;\log\{2\pi\} - \frac{N}{2}\log|V| - \frac{1}{2}\sum_{i=1}^{N}\left[(y_i-\theta)'V^{-1}(y_i-\theta)\right].\end{aligned}$$

Based on the premise of normality, we postulate that the true density $g(y|\xi)$ is a normal distribution with means θ and covariance matrix V:

$$g(y|\xi) = f(y|\theta, V).$$

Let the order-restricted maximum likelihood estimators of hypothesis H_m be denoted by $\tilde{\theta}^m$ and $\tilde{V}^m$. They are obtained by

$$\min_{\theta\in H_m, V}\sum_{i=1}^{N}\left[(y_i-\theta)'V^{-1}(y_i-\theta)\right],$$

which leads to

$$\tilde{\theta}^m = \arg\min_{\theta\in H_m}\sum_{i=1}^{N}\left[(y_i-\theta)'\left(\tilde{V}^m\right)^{-1}(y_i-\theta)\right], \tag{5.2}$$

$$\tilde{V}^m = N^{-1}\sum_{i=1}^{N}\left[(y_i-\tilde{\theta}^m)(y_i-\tilde{\theta}^m)'\right].$$

Since $\tilde{\theta}^m$ depends on $\tilde{V}^m$ and $\tilde{V}^m$ on $\tilde{\theta}^m$, iterations are required to calculate them. One could, for example, first set $\tilde{\theta}^m$ equal to the vector of groups means $(\bar{y})$. Based on these values, one can iterate between both components of (5.2) until convergence is reached. To calculate values of $\tilde{\theta}^m$, one could employ a quadratic program algorithm like the IMSL subroutine QPROG (Visual Numerics, 2003, pp. 1307–1310) in Fortran 90.

The statistical model to approximate $f(y|\theta, V)$ corresponding to hypothesis H_m is

$$f(y|\xi^m) = f(y|\tilde{\theta}^m, \tilde{V}^m).$$

For fixed $\tilde{\theta}^m$ and $\tilde{V}^m$, the expected log-likelihood at $(\tilde{\theta}^m, \tilde{V}^m)$, where the expectation is taken with respect to $f(y|\theta, V)$, is

$$\begin{aligned} &E_{f(y|\theta,V)}\left\{\log f(y|\tilde{\theta}^m, \tilde{V}^m)\right\} \\ &\quad = \int_{-\infty}^{\infty} \log\left\{f(y|\tilde{\theta}^m, \tilde{V}^m)\right\} f(y|\theta, V)\ dy \\ &\quad = -\frac{Nk}{2}\ \log\{2\pi\} - \frac{N}{2}\log|\tilde{V}^m| \\ &\qquad -\frac{1}{2}E_{f(y|\theta,V)}\left\{\sum_{i=1}^{N}\left[(y_i - \tilde{\theta}^m)'\left(\tilde{V}^m\right)^{-1}(y_i - \tilde{\theta}^m)\right]\right\}. \end{aligned} \tag{5.3}$$

The GORIC is minus two times an approximation of (5.3). We explain next that, for hypothesis H_m, the GORIC is written as

$$\text{GORIC}_m = -2\ \log f(y|\tilde{\theta}^m, \tilde{V}^m) + 2\ PT_m.$$

Since (5.3) depends on the unknown θ and V, it is ideally estimated by $\log f(y|\tilde{\theta}^m, \tilde{V}^m)$. However, this is not a good estimator, hence a bias results. To adjust for this, the GORIC comprises a likelihood part and a penalty part (denoted by PT_m), where the latter is the infinum of the expectation of the bias. To derive an expression for the penalty term, we need to following definitions. Let

$$\begin{aligned} V &= \sigma^2 U, \\ \tilde{V}^m &= \tilde{\sigma}_m^2 U, \\ \tilde{\sigma}_m^2 &= (Nk)^{-1}\sum_{i=1}^{N}\left[(y_i - \tilde{\theta}^m)'U^{-1}(y_i - \tilde{\theta}^m)\right], \end{aligned} \tag{5.4}$$

where U is known. In Section 5.6, we will return to the issue of U being unknown.

The expectation of the bias between $\log f(y|\tilde{\theta}^m, \tilde{V}^m)$ and (5.3) with respect to $f(\tilde{\theta}^m, \tilde{V}^m|\theta, V)$ (which is for ease of notation denoted by E) is

$$
\begin{aligned}
B^m(\theta, V) &= E\left\{\log f(y|\tilde{\theta}^m, \tilde{V}^m) - E_{f(y|\theta,V)}\left\{\log f(y|\tilde{\theta}^m, \tilde{V}^m)\right\}\right\} \\
&= E\left\{-\frac{1}{2}\sum_{i=1}^{N}\left[(y_i - \tilde{\theta}^m)'\left(\tilde{V}^m\right)^{-1}(y_i - \tilde{\theta}^m)\right] + \right. \\
&\quad \left. \frac{1}{2}\left[Nk\frac{\sigma^2}{\tilde{\sigma}_m^2} + N(\tilde{\theta}^m - \theta)'\left(\tilde{V}^m\right)^{-1}(\tilde{\theta}^m - \theta)\right]\right\} \qquad (5.5) \\
&= -\frac{Nk}{2} + \frac{Nk}{2}E\left\{\frac{\sigma^2}{\tilde{\sigma}_m^2}\right\} + \\
&\quad \frac{1}{2}E\left\{N(\tilde{\theta}^m - \theta)'\left(\tilde{V}^m\right)^{-1}(\tilde{\theta}^m - \theta)\right\}, \qquad (5.6)
\end{aligned}
$$

where to obtain the second part in (5.5) we used

$$
\begin{aligned}
&E_{f(y|\theta,V)}\left\{\sum_{i=1}^{N}\left[(y_i - \tilde{\theta}^m)'\left(\tilde{V}^m\right)^{-1}(y_i - \tilde{\theta}^m)\right]\right\} = \\
&\quad E_{f(y|\theta,V)}\left\{\sum_{i=1}^{N}\left[(y_i - \theta)'\left(\tilde{V}^m\right)^{-1}(y_i - \theta)\right]\right\} + N(\tilde{\theta}^m - \theta)'\left(\tilde{V}^m\right)^{-1}(\tilde{\theta}^m - \theta),
\end{aligned}
$$

and

$$
\begin{aligned}
E_{f(y|\theta,V)}\left\{\sum_{i=1}^{N}\left[(y_i - \theta)'\left(\tilde{V}^m\right)^{-1}(y_i - \theta)\right]\right\} &= \text{trace}\left\{\left(\tilde{V}^m\right)^{-1}VN\right\} \\
&= Nk\frac{\sigma^2}{\tilde{\sigma}_m^2},
\end{aligned}
$$

and to obtain the first part in (5.6) we used

$$
E\left\{\sum_{i=1}^{N}\left[(y_i - \tilde{\theta}^m)'\left(\tilde{V}^m\right)^{-1}(y_i - \tilde{\theta}^m)\right]\right\} = Nk.
$$

It holds true that $B^m(\theta, V) \geq B^m(\theta^0, V)$ for all $\theta^0 \in \mathcal{C}_0 = \{\theta \in R^k | \theta_1 = \ldots = \theta_k\}$ and all $\theta \in \mathcal{C}_m$ and that $B^m(\theta^0, V)$ has the same value for all $\theta^0 \in \mathcal{C}_0$ (Anraku, 1999; Robertson, Wright, & Dykstra, 1988, pp. 101–102). Hence,

$$
PT_m = \inf_{\theta\in\mathcal{C}_m} \; B^m(\theta, V) = \inf_{\theta\in\mathcal{C}_0} \; B^m(\theta, V) = B^m(\theta^0, V). \qquad (5.7)
$$

$B^m(\theta^0, V)$ is calculated by (5.6), where the expectation is now with respect to $f(\tilde{\theta}^m, \tilde{V}^m | \theta \in \mathcal{C}_0, V)$ (for brevity, denoted by E) and θ is replaced by θ_0, which yields

$$
B^m(\theta^0, V) = -\frac{Nk}{2} + \frac{Nk}{2}E\left\{\frac{\sigma^2}{\tilde{\sigma}_m^2}\right\} + \frac{1}{2}E\left\{N(\tilde{\theta}^m - \theta^0)'\left(\tilde{V}^m\right)^{-1}(\tilde{\theta}^m - \theta^0)\right\}. \qquad (5.8)
$$

In the sequel, we use, without loss of generalization, $H_0 : \theta = \theta^0$ in lieu of $H_0 : \theta \in \mathcal{C}_0 = \{\theta \in R^k | \theta_1 = \ldots = \theta_k\}$.

To obtain $B^m(\theta^0, V)$ in (5.8), one requires to determine

$$E\left\{\frac{\sigma^2}{\tilde{\sigma}_m^2}\right\} \approx \frac{1}{E\left\{\frac{\tilde{\sigma}_m^2}{\sigma^2}\right\}} + \frac{\text{var}\left(\frac{\tilde{\sigma}_m^2}{\sigma^2}\right)}{\left[E\left\{\frac{\tilde{\sigma}_m^2}{\sigma^2}\right\}\right]^3} \tag{5.9}$$

and

$$E\left\{N(\tilde{\theta}^m - \theta^0)'\left(\tilde{V}^m\right)^{-1}(\tilde{\theta}^m - \theta^0)\right\} = E\left\{\frac{N(\tilde{\theta}^m - \theta^0)'U^{-1}(\tilde{\theta}^m - \theta^0)}{\sigma^2} \Big/ \frac{\tilde{\sigma}_m^2}{\sigma^2}\right\}, \tag{5.10}$$

where (5.9) is based on a second order Taylor expansion of $\frac{1}{x}$ around $E\{x\}$, with $x = \tilde{\sigma}_m^2/\sigma^2$. We first need to rewrite $\tilde{\sigma}_m^2/\sigma^2$ before we can obtain its null distribution and expectation. It can be shown, using Theorem 1 in Appendix 5.A, that

$$\begin{aligned}\frac{\tilde{\sigma}_m^2}{\sigma^2} &= \frac{1}{Nk}\frac{\sum_{i=1}^N\left[(y_i - \tilde{\theta}^m)'U^{-1}(y_i - \tilde{\theta}^m)\right]}{\sigma^2} \\ &= \frac{1}{Nk}\left[\frac{\sum_{i=1}^N\left[(y_i - \theta^0)'U^{-1}(y_i - \theta^0)\right]}{\sigma^2} - \frac{N(\tilde{\theta}^m - \theta^0)'U^{-1}(\tilde{\theta}^m - \theta^0)}{\sigma^2}\right].\end{aligned} \tag{5.11}$$

The first term in brackets in (5.11) has (assuming that H_0 is true) a chi-square distribution with Nk degrees of freedom (i.e., χ^2_{Nk}) and, therefore, has an expectation of $E\left\{\chi^2_{Nk}\right\} = Nk$. Consequently, we only require the null distribution and expectation of the second term in brackets in (5.11) to compute $E\left\{\tilde{\sigma}_m^2/\sigma^2\right\}$. The expression for $\text{var}\left(\tilde{\sigma}_m^2/\sigma^2\right)$ in (5.9) is written down in Appendix 5.B. In addition, Appendix 5.B demonstrates that (5.9) can be written as

$$E\left\{\frac{\sigma^2}{\tilde{\sigma}_m^2}\right\} = 1 + \frac{2}{Nk} + \frac{1}{Nk}E\left\{\frac{N(\tilde{\theta}^m - \theta^0)'U^{-1}(\tilde{\theta}^m - \theta^0)}{\sigma^2}\right\} + O((Nk)^{-2}). \tag{5.12}$$

Note that the part in the expectation equals the second term in brackets in (5.11). Moreover, the two terms in the second expectation in (5.10) are equal to the second term in brackets in (5.11) and to (5.11), respectively. As a consequence, to calculate (5.12) and (5.10), we need the null distribution of the second term in brackets in (5.11).

5.2.2 The Null Distribution and Expectation of $N(\tilde{\theta}^m - \theta^0)'U^{-1}(\tilde{\theta}^m - \theta^0)/\sigma^2$

To obtain the null distribution of

$$\frac{N(\tilde{\theta}^m - \theta^0)'U^{-1}(\tilde{\theta}^m - \theta^0)}{\sigma^2} = (\tilde{\theta}^m - \theta^0)'\,[V/N]^{-1}\,(\tilde{\theta}^m - \theta^0),$$

we require the following expression:

$$\begin{aligned}
&\overline{\chi}^2(V/N, \mathcal{C}_m) \\
&\quad = \min_{\theta=\theta^0} \sum_{i=1}^{N} \left[(y_i - \theta)'V^{-1}(y_i - \theta)\right] - \min_{\theta\in\mathcal{C}_m} \sum_{i=1}^{N} \left[(y_i - \theta)'V^{-1}(y_i - \theta)\right] \qquad (5.13)\\
&\quad = \sum_{i=1}^{N} \left[(y_i - \theta^0)'V^{-1}(y_i - \theta^0)\right] - \sum_{i=1}^{N} \left[(y_i - \tilde{\theta}^m)'V^{-1}(y_i - \tilde{\theta}^m)\right] \\
&\quad = N(\tilde{\theta}^m - \theta^0)'V^{-1}(\tilde{\theta}^m - \theta^0) \\
&\quad = (\tilde{\theta}^m - \theta^0)'\,[V/N]^{-1}\,(\tilde{\theta}^m - \theta^0),
\end{aligned}$$

where the one but last line is obtained by using Theorem 1 in Appendix 5.A. According to Silvapulle and Sen (2005, pp. 75–77) and Robertson et al. (1988, pp. 70), the null distribution of (5.13) is given by

$$Pr(\overline{\chi}^2(V/N, \mathcal{C}_m) \leq x) = \sum_{j=1}^{k} w_j(k, V/N, \mathcal{C}_m) Pr(\chi_j^2 \leq x),$$

where $w_j(k, V/N, \mathcal{C}_m)$ is the level probability / chi-bar-square weight for hypothesis H_m. A level probability $w_j(k, V/N, \mathcal{C}_m)$ is the probability that $\tilde{\theta}^m$ has j levels or, rather, the probability that the parameter space in accordance with the active constraints in $\mathcal{C}_m$ is of dimension j. An explanation and the calculation of the level probabilities are given in Section 5.5. Employing $E\left\{\chi_j^2\right\} = j$, gives

$$E\left\{\frac{N(\tilde{\theta}^m - \theta^0)'U^{-1}(\tilde{\theta}^m - \theta^0)}{\sigma^2}\right\} = \sum_{j=1}^{k} w_j(k, V/N, \mathcal{C}_m) j. \qquad (5.14)$$

5.2.3 The GORIC

Equations (5.12) and (5.14) amount to

$$E\left\{\frac{\sigma^2}{\tilde{\sigma}_m^2}\right\} = 1 + \frac{2}{Nk} + \frac{1}{Nk}\sum_{j=1}^{k} w_j(k, V/N, \mathcal{C}_m) j + O((Nk)^{-2}). \qquad (5.15)$$

From (5.14) it follows that, if j level sets are given, the second expression in brackets in (5.11) and the term in brackets in (5.11) are conditionally independent and are distributed as chi-square distributions with j and $Nk - j$ degrees of freedom, respectively (see Anraku, 1999; Robertson et al., 1988, pp. 69–74). Therefore,

$$\frac{\frac{N(\tilde{\theta}^m - \theta^0)'U^{-1}(\tilde{\theta}^m - \theta^0)}{\sigma^2}/j}{Nk\frac{\tilde{\sigma}_m^2}{\sigma^2}/(Nk - j)} = \frac{Nk - j}{Nk\ j}\left[\frac{N(\tilde{\theta}^m - \theta^0)'U^{-1}(\tilde{\theta}^m - \theta^0)}{\tilde{\sigma}_m^2}\right] \qquad (5.16)$$

has an F distribution with $(j, Nk - j)$ degrees of freedom for $\theta^0 \in \mathcal{C}_0$. Hence, when j level sets are given, the limit of j times (5.16) has a chi-square distribution with j degrees of freedom for $(Nk - j) \to \infty$. Consequently, (5.10) can be written as

$$E\left\{\frac{N(\tilde{\theta}^m - \theta^0)'U^{-1}(\tilde{\theta}^m - \theta^0)}{\tilde{\sigma}^2_m}\right\} = \sum_{j=1}^{k} w_j(k, V/N, \mathcal{C}_m)j + O((Nk)^{-1}). \quad (5.17)$$

From (5.15) and (5.17), it follows that (5.8) can be rewritten as

$$B^m(\theta^0, V) = 1 + \sum_{j=1}^{k} w_j(k, V/N, \mathcal{C}_m)j + O((Nk)^{-1}).$$

Thus, the GORIC is calculated by

$$\begin{aligned} \text{GORIC}_m &= -2\ \log f(y|\tilde{\theta}^m, \tilde{V}^m) + 2\ PT_m, \text{ where} \\ PT_m &= 1 + \sum_{j=1}^{k} w_j(k, V/N, \mathcal{C}_m)\ j. \end{aligned} \quad (5.18)$$

Note that if V is unknown, the penalty term cannot be determined. We will elaborate on this in Sections 5.5 and 5.6.

Now we have obtained the GORIC for a special type of models and not for normal linear models in general. Examples of this special type of models are models for multivariate one-sided testing and repeated measures analysis without between-subject factors. In the next section, we will derive the GORIC that can be applied to univariate normal linear models and multivariate normal linear models.

5.3 The GORIC for Extended Models

5.3.1 Univariate Normal Linear Models

The derivation of the GORIC in the previous sections elucidates the one that is feasible for univariate normal linear models:

$$y|X \sim \mathcal{N}_N(X\beta, V),$$

where $y \in R^{N\times 1}$, $X \in R^{N\times k}$, $\beta \in R^{k\times 1}$, and $V \in R^{N\times N}$. In Section 5.2, the covariance matrix V is a $k \times k$ matrix, here it is a $N \times N$ matrix.

The order-restricted maximum likelihood estimators, $\tilde{\beta}^m$ and $\tilde{V}^m$, are obtained by

$$\min_{\beta\in H_m, V}(y - X\beta)'V^{-1}(y - X\beta).$$

From this it follows that

$$\begin{aligned} \tilde{\beta}^m &= \arg\ \min_{\beta\in H_m}(y - X\beta)'\left(\tilde{V}^m\right)^{-1}(y - X\beta), \\ \tilde{V}^m &= (y - X\tilde{\beta}^m)(y - X\tilde{\beta}^m)'. \end{aligned} \quad (5.19)$$

Since $\tilde{\beta}^m$ depends on $\tilde{V}^m$ and $\tilde{V}^m$ on $\tilde{\beta}^m$, iterations are needed to calculate them. One could, for example, first set $\tilde{\beta}^m$ equal to $(X'X)^{-1}X'y$. Based on these values one can iterate between both components of (5.19) until convergence is reached. To calculate $\tilde{\beta}^m$ (in software), one could use a quadratic program algorithm like the IMSL subroutine QPROG (Visual Numerics, 2003, pp. 1307–1310) in Fortran 90. It should be stressed that if V is known (up to a positive constant), like in univariate regression models (where $V = \sigma^2 I_N$, with I_N the $N \times N$ identity matrix), no iterations are required. Namely, in univariate regression models, $\tilde{\beta}^m$ does not depend on V at all and $\tilde{V}^m = \tilde{\sigma}^2_m I_N$.

Analogous to Section 5.2, to derive an expression for the penalty term, let

$$\begin{aligned} V &= \sigma^2 U, \\ \tilde{V}^m &= \tilde{\sigma}^2_m U, \\ \tilde{\sigma}^2_m &= \left[(y - X\tilde{\beta}^m)' U^{-1} (y - X\tilde{\beta}^m)\right] \end{aligned}$$

where U is known. Section 5.6 discusses the case where U is not known. Furthermore, let

$$\begin{aligned} H_0 &: \ \beta \in \mathcal{C}_0, \\ H_m &: \ \beta \in \mathcal{C}_m, \end{aligned}$$

with $\mathcal{C}_0 = \{\beta \in R^k | \beta_1 = \ldots = \beta_k\}$. Without loss of generalization, we use $H_0 : \beta = \beta^0$, with $\beta^0 \in \mathcal{C}_0$, instead of $H_0 : \beta \in \mathcal{C}_0$, as done in Section 5.2, in the sequel. The derivation of the GORIC for univariate regression models resembles the derivation in Sections 5.2 up to (5.13), only now we replace θ by $X\beta$, k by N, and N by 1. Thus, the analogue of (5.8), (5.10), (5.12), and (5.13) are

$$\begin{aligned} &B^m(\beta^0, V) = \\ &\qquad -\frac{N}{2} + \frac{N}{2} E\left\{\frac{\sigma^2}{\tilde{\sigma}^2_m}\right\} + \frac{1}{2} E\left\{(X\tilde{\beta}^m - X\beta^0)' \left(\tilde{V}^m\right)^{-1} (X\tilde{\beta}^m - X\beta^0)\right\}, \\ &E\left\{(X\tilde{\beta}^m - X\beta^0)' \left(\tilde{V}^m\right)^{-1} (X\tilde{\beta}^m - X\beta^0)\right\} = \\ &\qquad E\left\{\frac{(X\tilde{\beta}^m - X\beta^0)' U^{-1} (X\tilde{\beta}^m - X\beta^0)}{\sigma^2} \Big/ \frac{\tilde{\sigma}^2_m}{\sigma^2}\right\}, \\ &E\left\{\frac{\sigma^2}{\tilde{\sigma}^2_m}\right\} = \\ &\qquad 1 + \frac{2}{N} + \frac{1}{N} E\left\{\frac{(X\tilde{\beta}^m - X\beta^0)' U^{-1} (X\tilde{\beta}^m - X\beta^0)}{\sigma^2}\right\} + O((N)^{-2}), \end{aligned}$$

and

$$\overline{\chi}^2(V, \mathcal{C}_m) = (X\tilde{\beta}^m - X\beta^0)' V^{-1} (X\tilde{\beta}^m - X\beta^0),$$

respectively. The latter can be written as

$$\overline{\chi}^2(V, \mathcal{C}_m) = (\tilde{\beta}^m - \beta^0)' W^{-1} (\tilde{\beta}^m - \beta^0), \tag{5.20}$$

where

$$W^{-1} = X' V^{-1} X \in R^{k \times k}.$$

Equation (5.20) equals (5.13) with θ replaced by β and V/N by W. Consequently, the remainder of the derivation resembles the derivation in Section 5.2 with these two replacements. Hence, the null distribution of (5.20) is given by

$$Pr(\overline{\chi}^2(V, \mathcal{C}_m) \leq x) = \sum_{j=1}^{k} w_j(k, W, \mathcal{C}_m) Pr(\chi_j^2 \leq x).$$

Thus, for $H_m : \beta \in \mathcal{C}_m$ in univariate normal linear models, it holds true that

$$\begin{aligned} \text{GORIC}_m &= -2 \ \log f(y|X\tilde{\beta}^m, \tilde{V}^m) + 2 \ PT_m, \text{ with} \\ \log f(y|X\tilde{\beta}^m, \tilde{V}^m) &= -\frac{N}{2} \ \log\{2\pi\} - \frac{1}{2} \log |\tilde{V}^m| - \\ &\quad \frac{1}{2} \left[(y - X\tilde{\beta}^m)' (\tilde{V}^m)^{-1} (y - X\tilde{\beta}^m) \right], \\ PT_m &= 1 + \sum_{j=1}^{k} w_j(k, W, \mathcal{C}_m) \ j. \end{aligned}$$

We will discuss the calculation of the penalty term in Section 5.5. Bear in mind that the penalty term cannot be obtained in case W, or rather V, is unknown; it should be estimated then. We will elaborate on this in Section 5.6.

5.3.2 Multivariate Normal Linear Models

A multivariate normal linear model with t dependent variables can be written as

$$\left. \begin{bmatrix} y_1 \\ \vdots \\ y_t \end{bmatrix} \right| X \sim \mathcal{N}_{tN} \left([I_t \otimes X] \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_t \end{bmatrix}, V = \Sigma \otimes U \right), \tag{5.21}$$

where $y = \begin{bmatrix} y_1 \\ \vdots \\ y_t \end{bmatrix} \in R^{tN \times 1}$, $I_t \otimes X = \text{diag}(X, \ldots, X) \in R^{tN \times tk}$, $\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_t \end{bmatrix} \in R^{tk \times 1}$, with $y_h \in R^{N \times 1}$, $X \in R^{N \times k}$, and $\beta_h \in R^{k \times 1}$ for $h = 1, \ldots, t$, $V \in R^{tN \times tN}$, $\Sigma \in R^{t \times t}$, and $U \in R^{N \times N}$. When $t = 1$, so for univariate normal linear models, V reduces to $\sigma^2 U$, as in Section 5.3.1. In addition, for regression models it holds true that $V = \Sigma \otimes I_N$.

The order-restricted maximum likelihood estimators, $\tilde{\beta}^m$ and $\tilde{V}^m$, are obtained by

$$\tilde{\beta}^m = \arg \min_{\beta \in H_m, V} (y - [I_t \otimes X]\beta)' V^{-1} (y - [I_t \otimes X]\beta).$$

From this it follows that

$$\tilde{\beta}^m = \arg \min_{\beta \in H_m} (y - [I_t \otimes X]\beta)' \left(\tilde{V}^m\right)^{-1} (y - [I_t \otimes X]\beta), \tag{5.22}$$

$$\tilde{V}^m = (y - [I_t \otimes X]\tilde{\beta}^m)(y - [I_t \otimes X]\tilde{\beta}^m)'. \tag{5.23}$$

Since $\tilde{\beta}^m$ depends on $\tilde{V}^m$ and $\tilde{V}^m$ on $\tilde{\beta}^m$, iterations are needed to calculate them. The iterations and calculations for multivariate normal linear models are analogous to the ones for univariate normal linear models described in Section 5.3.1. As opposed to univariate regression models, one does require to iterate between $\tilde{\beta}^m$ and $\tilde{V}^m = \tilde{\Sigma}^m \otimes I_N$ in multivariate regression models, since $\tilde{\Sigma}^m$ is a unknown matrix; $\tilde{\Sigma}^m$ is calculated by (5.25) in Section 5.6 with $\hat{B}$ replaced by $\tilde{B}^m$, the $k \times t$ matrix where column h equals $\tilde{\beta}^m_h$.

As in Sections 5.2 and 5.3.1, to derive an expression for the penalty term, let

$$\begin{aligned} V &= \sigma^2 S \otimes U, \\ \tilde{V}^m &= \tilde{\sigma}^2_m S \otimes U, \\ \tilde{\sigma}^2_m &= \left[(y - [I_t \otimes X]\tilde{\beta}^m)'[S^{-1} \otimes U^{-1}](y - [I_t \otimes X]\tilde{\beta}^m)\right], \end{aligned}$$

where in the last line we used that $[S \otimes U]^{-1} = [S^{-1} \otimes U^{-1}]$. Initially, analogous to the previous sections, we assume that $S \otimes U$ is known, or rather both S and U are known. In Section 5.6, we return to the case where S or U is unknown.

Let $H_0 : \beta \in \mathcal{C}_0$ and $H_m : \beta \in \mathcal{C}_m$, with $\mathcal{C}_0 = \{\beta \in R^{tk} | \beta_1 = \ldots = \beta_{tk}\}$. In the sequel, we employ $H_0 : \beta = \beta^0$, with $\beta^0 \in \mathcal{C}_0$, instead of $H_0 : \beta \in \mathcal{C}_0$, analogous to the previous section and Section 5.2. The derivation of the GORIC modified for multivariate normal linear models resembles the one in Section 5.2: We now replace θ by $[I_t \otimes X]\beta$, k by tN, and N by 1 in the formulas up to (5.13) and we use $V = \Sigma \otimes U = \sigma^2[S \otimes U]$. Then, (5.8) and (5.13) yield

$$\begin{aligned} &B^m(\beta^0, V) = \\ &\quad -\frac{tN}{2} + \frac{tN}{2} E\left\{\frac{\sigma^2}{\tilde{\sigma}^2_m}\right\} + \\ &\quad \frac{1}{2} E\left\{([I_t \otimes X]\tilde{\beta}^m - [I_t \otimes X]\beta^0)' \left(\tilde{V}^m\right)^{-1} ([I_t \otimes X]\tilde{\beta}^m - [I_t \otimes X]\beta^0)\right\}, \end{aligned}$$

and

$$\overline{\chi}^2(V, \mathcal{C}_m) = ([I_t \otimes X]\tilde{\beta}^m - [I_t \otimes X]\beta^0)'[\Sigma^{-1} \otimes U^{-1}]([I_t \otimes X]\tilde{\beta}^m - [I_t \otimes X]\beta^0),$$

respectively. The latter can be rewritten as

$$\overline{\chi}^2(V, \mathcal{C}_m) = (\tilde{\beta}^m - \beta^0)' W^{-1} (\tilde{\beta}^m - \beta^0), \tag{5.24}$$

with

$$\begin{aligned} W^{-1} &= [I_t \otimes X]'[\Sigma^{-1} \otimes U^{-1}][I_t \otimes X] \\ &= \Sigma^{-1} \otimes [X' U^{-1} X] \\ &= [\sigma^2 S]^{-1} \otimes [X' U^{-1} X], \end{aligned}$$

since $[I_t \otimes X]' = [I_t \otimes X']$ and $[A \otimes B][C \otimes D] = [AC \otimes BD]$. For univariate regression models (where $V = \sigma^2 I_N$), $W = \sigma^2[X'X]^{-1}$, as can be seen in the previous section. Here we have the same result (for $t = 1$ and $U = I_N$). Moreover, for $U = I_N$, $k = 1$, $t = k$, and X a vector of ones, the multivariate normal linear model simplifies to the model discussed in Section 5.2. In that case, it evidently holds true that $W = \Sigma/N$, where the $k \times k$ matrix Σ in this section equals the $k \times k$ matrix V of Section 5.2.

Equation (5.24) equals (5.13) with θ replaced by β, k by tk and the $k \times k$ matrix V/N by the $tk \times tk$ matrix W. Therefore, the remainder of the derivation resembles the derivation in Section 5.2. As a result, the null distribution of (5.24) is given by

$$Pr(\overline{\chi}^2(V, \mathcal{C}_m) \leq x) = \sum_{l=1}^{tk} w_l(tk, W, \mathcal{C}_m) Pr(\chi_l^2 \leq x).$$

Thus, for $H_m : \beta \in \mathcal{C}_m$ in multivariate normal linear models,

$$\begin{aligned} \text{GORIC}_m &= -2 \ \log f(y|[I_t \otimes X]\tilde{\beta}^m, \tilde{V}^m) + 2 \ PT_m, \text{ with} \\ \log f(y|[I_t \otimes X]\tilde{\beta}^m, \tilde{V}^m) &= -\frac{tN}{2} \ \log\{2\pi\} - \frac{1}{2}\log|\tilde{V}^m| - \\ &\quad \frac{1}{2}\left[(y - [I_t \otimes X]\tilde{\beta}^m)' \left(\tilde{V}^m\right)^{-1} (y - [I_t \otimes X]\tilde{\beta}^m)\right], \\ PT_m &= 1 + \sum_{l=1}^{tk} w_l(tk, W, \mathcal{C}_m) \ l. \end{aligned}$$

When S and/or U are unknown, W and thus the penalty term cannot be calculated. Importantly, for regression models, where $U = I_N$ is known, W cannot be calculated, because S is unknown. In such cases, W needs to be estimated. We will elaborate on this in Sections 5.5 and 5.6.

In Section 5.7, we establish how competing hypotheses for multivariate regression models can be evaluated with the GORIC. In the next section, we demonstrate that the GORIC can be used for restrictions of the form $H_m : R\theta \leq r$, with θ the parameter of interest.

5.4 Restrictions of the form $H_m : R\theta \leq r$

Let θ be the parameter of interest. Until now, we have focussed on $H_m : \theta \in \mathcal{C}_m$. A special case (according to Silvapulle & Sen, 2005) is $H_m : \theta \in \mathcal{C}_m = \{\theta \in R^k : R\theta \leq 0\}$, where R is a $c_m \times k$ matrix. When R is of full rank (after discarding the redundant restrictions),

$$\{\theta \in R^k : R\theta \leq r\} = \{\theta \in R^k : R\theta - r \leq 0\} = \{\theta \in R^k : R\theta^* \leq 0\},$$

where $\theta^* = \theta - q$ and $Rq = r$, with $q \in R^k$.

It should be stressed that q cannot be defined when R is not of full rank (after discarding the redundant restrictions). For example, q cannot be determined for $H_m : \theta_i \geq r_{11}, \theta_i \leq r_{12}$ when $r_{11} \neq r_{12}$. However, for $r_{11} = r_{12}$, $H_m : \theta_i \geq r_{11}, \theta_i \leq r_{12}$

simplifies to $H_m : \theta_i = r_{11}$. Then, q is defined and equals r_{11}. Consequently, the equality restrictions should be handled separately by examining $\mathcal{C}_m = \{\theta \in R^k : R_1\theta \leq r_1, R_2\theta = r_2\}$, where R_1 is a $c_{m1} \times k$ matrix, r_1 a vector of length c_{m1}, R_2 a $c_{m2} \times k$ matrix, and r_2 a vector of length c_{m2}. In that case, $\mathcal{C}_m = \{\theta \in R^k : R_1\theta^* \leq 0, R_2\theta^* = 0\}$ for $[R_1', R_2']'q = [r_1', r_2']'$. Now, q exists when $[R_1', R_2']'$ is of full rank (after discarding the redundant restrictions).

Since $\{\theta \in R^k : R\theta \leq 0\}$ is a closed convex cone, $\{\theta \in R^k : R\theta^* \leq 0\} = \{\theta \in R^k : R\theta \leq r\}$ is too for the shifted data: both y_i and θ are shifted by q. But, only when R is of full rank (when discarding the redundant restrictions). Hence, $\{\theta \in R^k : R\theta \leq r\}$ is a closed convex cone with a relocated base, that is, the origin $(0, 0)$ is no longer the base of the cone. Therefore, we will refer to this cone as a shifted or relocated closed convex cone. Previous result implies that the GORIC for $H_m : \theta \in \mathcal{C}_m = \{\theta \in R^k : R\theta \leq r\}$ has the same expression as the GORIC for $H_m : \theta \in \mathcal{C}_m = \{\theta \in R^k : R\theta \leq 0\}$ denoted in (5.18). Evidently, the same remains valid for $H_m : \theta \in \mathcal{C}_m = \{\theta \in R^k : R_1\theta \leq r_1, R_2\theta = r_2\}$ when $[R_1', R_2']'$ is of full rank (after discarding the redundant restrictions).

The analogue remains true for t-variate normal linear models, where β is the parameter of interest, since

$$\{\beta \in R^{tk} : R\beta \leq r\} = \{\beta \in R^{tk} : R\beta - r \leq 0\} = \{\beta \in R^{tk} : R\beta^* \leq 0\},$$

where $\beta^* = \beta - q$ and $Rq = r$, with $q \in R^{tk}$. Here, y is shifted by $[I_t \otimes X]q$ and β by q.

In the next section, we elaborate on the level probabilities for all discussed types of models. We show how they can easily be calculated by simulation. Furthermore, we give an interpretation of the penalty term.

5.5 Level Probabilities

Let ν be the parameter of interest of length L, with

$$\nu = \begin{cases} \theta \text{ for the model described in Section 5.2} \\ \beta \text{ for the models described in Section 5.3} \end{cases}$$

$$L = \begin{cases} k \text{ for the model described in Section 5.2} \\ tk \text{ for the models described in Section 5.3.} \end{cases}$$

Furthermore, let $W = \sigma^2 W^*$, with W^* known and equal to

$$W^* = \begin{cases} U/N & \text{for the model described in Section 5.2} \\ S \otimes [X'U^{-1}X]^{-1} & \text{for the models described in Section 5.3.} \end{cases}$$

A level probability $w_l(L, W, \mathcal{C}_m)$ is the probability that there are l levels among the L order-restricted maximum likelihood estimators (see also Anraku (1999); Silvapulle and Sen (2005, pp. 77–83); Robertson et al. (1988, p. 69)). In other words, it is the probability that the parameter space in accordance with the active constraints in $\mathcal{C}_m$ is of dimension j. Bear in mind that the parameters ν emanate from the null distribution which is the normal distribution with mean ν^0 and covariance matrix W.

Anraku (1999) laid out on page 149 that in general the calculation of the level probabilities is difficult (see also Robertson et al., 1988, pp. 74–86). However, Silvapulle and Sen (2005, pp. 78–81) point out that it can be done computationally convenient via simulation. The simulation comprises 5 steps:

1. Generate z (of length L) from $\mathcal{N}_L(\nu^0, \sigma^2 W^*)$, with W^* a known matrix. Silvapulle and Sen (2005, pp. 86) and Robertson et al. (1988, pp. 69) prove that the calculation of the level probabilities does not depend on the value of ν^0. Furthermore, Robertson et al. (1988, pp. 69) show that the calculation of the level probabilities is invariant for positive constants, like σ^2 (and N). Consequently, one can generate z from $\mathcal{N}_L(0, W^*)$ as well. One exception is discussed below.
2. Compute $\tilde{z}_m = \arg\min_{\nu \in \mathcal{C}_m} (z-\nu)' W^{-1} (z-\nu) = \arg\min_{\nu \in \mathcal{C}_m} (z-\nu)' \left(W^*\right)^{-1} (z-\nu)$, where $\mathcal{C}_m$ is the set of parameters which is in accordance with the restrictions in H_m, the hypothesis of interest.
3. Determine the the number of levels in $\tilde{z}_m$, denote this by L_m. For $\mathcal{C}_m = \{\nu \in R^L : R\nu \leq 0\}$, where restriction a is denoted by $R_a \nu \leq 0$, this is done as follows: Let $A = \{a : R_a \tilde{z}_m = 0\}$ and $\phi = \{\nu : R_a \nu = 0 \ \forall \ a \in A\}$, then L_m is the dimension of ϕ.
4. Repeat the previous steps T (e.g., $T = 100,000$) times.
5. Estimate the level probability $w_l(L, W^*, \mathcal{C}_m)$ by the proportion of times L_m is equal to l ($l = 1, \ldots, L$).

To implement this in software, one requires a quadratic program algorithm. For example, one can use the IMSL subroutine QPROG (Visual Numerics, 2003, pp. 1307–1310) in Fortran 90.

In case the restrictions are of the type $H_m : R\nu \leq r$, with $r \neq 0$, the data should be shifted accordingly, as explained in Section 5.4. Notably, the calculation of the level probabilities does not involve simulation of the data y, but only the parameters ν. Therefore, we can just simulate the shifted parameters ν^*; denoted by z in the simulation steps. All steps remain valid. Note that q does not need to be determined.

The level probabilities are invariant for the values of ν^0 and σ^2. However, there is one exception, namely restrictions of the type $\nu \leq r$, including $r = 0$. When the hypothesis of interest contains this type of restriction, one must use $\nu^0 = 0$. This results in level probabilities that are invariant for the value of σ^2. Observe that setting ν^0 equal to 0 yields the same result as for $\nu^0 \neq 0$ with $\sigma^2 \to \infty$.

The penalty term

$$PT_m = 1 + \sum_{l=1}^{L} w_l(L, W^*, \mathcal{C}_m) \ l$$

can be seen as the expected dimension of the parameters. In other words, the expected dimension of ν plus 1 because of the unknown variance term σ^2. When there are no restrictions, the penalty is equal to the number of distinct parameter values. Hence, in that case, the GORIC reduces to the AIC. In case of analysis of variance models with simple order restrictions, the penalty equals the expected number of distinct parameter values and the GORIC simplifies to the ORIC. In the most general case, the penalty does not only reflect the number of distinct parameter values, but it takes

the order restrictions into account as well. When there are restrictions, the value of the penalty is lower than the number of distinct parameter values and the penalty becomes increasingly smaller when the restrictions become stricter.

The level probabilities can only be obtained when W^* is known or estimated, or rather when U and S are known or estimated. In the next section, we will establish how U and S can be estimated from the data. Furthermore, we briefly describe the consequences of U or S being unknown.

5.6 U or S unknown

Until now, we have assumed that $W = \sigma^2 W^*$ and that W^* is known in the calculation of the penalty term. As mentioned in Section 5.5, the level probabilities are invariant of positive constants (like σ^2), which implies that $w_l(L, W^*, \mathcal{C}_m) = w_l(L, W, \mathcal{C}_m)$. Consequently, if W^* is known, we will use W^* in the simulation steps discussed in Section 5.5; otherwise, we will use an estimate of W instead of W^*. For the model described in Section 5.2 in (5.1) where $W = V/N = \sigma^2 U/N$, this implies that when W^* or rather U is unknown, W or rather V should be estimated from the data. We will use the maximum likelihood estimator of V:

$$\hat{V} = N^{-1} \sum_{i=1}^{N} [(y_i - \bar{y})(y_i - \bar{y})'] ,$$

with $\bar{y}$ a vector of the sample means of y. It should be stressed that $E\left\{\hat{V}\right\} = \frac{N-1}{N} V$. Because the level probabilities are invariant of positive constants, $w_l(L, \hat{V}, \mathcal{C}_m) \to w_l(L, V, \mathcal{C}_m)$ for $N \to \infty$.

If V is estimated from the data, the dimension of V, which is the number of unknown distinct variance terms, is $(k+1)k/2$ instead of 1. Since the restrictions are always on the θ parameters and never on the variance terms, the number of unknown variance terms is equal for all hypotheses of interest. Thus, although the penalty should then be corrected, the correction is equal for all H_m, with $m \in \mathcal{M}$.

The analogue remains valid for t-variate normal linear models, which are described in Section 5.3. To date, we have assumed that $W^* = S \otimes [X'U^{-1}X]^{-1}$ is known or rather that S and U are known in the calculation of the penalty term. In this section, we will only discuss t-variate regression models for which $U = I_N$ is known. For univariate regression models, $\Sigma = \sigma^2 S$ equals the positive scalar σ^2, for which the level probabilities are invariant. In contrast, S needs to be estimated from the data in case of multivariate regression models. To calculate the maximum likelihood estimator of Σ ($\hat{\Sigma}$), we need to rewrite the multivariate regression model. Let Y be the $N \times t$ matrix where column h equals y_h for $h = 1, \ldots, t$, B the $k \times t$ matrix where column h equals β_h, and D the $N \times t$ matrix where column h equals ϵ_h. One could write the multivariate regression model as

$$Y = XB + D.$$

Let d_i' be row i of D (for $i = 1, \ldots, N$). The dependence of the dependent variables is as follows: $\mathrm{E}\{d_i\} = 0$, $\mathrm{E}\{d_i d_i'\} = \Sigma \in R^{t \times t}$, and $\mathrm{E}\{d_i d_{i'}'\} = 0$ for all $i \neq i'$ for

$i, i' = 1, \ldots, N$. Consequently, $\hat{\Sigma}$ is determined by

$$\hat{\Sigma} = N^{-1}(Y - X\hat{B})'(Y - X\hat{B}), \tag{5.25}$$

where $\hat{B}$ is the unrestricted maximum likelihood estimator of B.

Analogous to the previous case, $w_l(L, \hat{\Sigma}, \mathcal{C}_m) \to w_l(L, \Sigma, \mathcal{C}_m)$ for $N \to \infty$. In addition, if Σ is estimated from the data, the number of unknown distinct variance terms is $(t+1)t/2$ instead of 1. Again, although the penalty should then be corrected, the correction is equal for all H_m, since the restrictions are always on the β parameters and never on the variance terms.

In the next section, we present the evaluation of hypotheses with the GORIC in a multivariate regression model.

5.7 The GORIC Illustrated

In this section, we will illustrate the GORIC based on real data. This example shows the data of Rencher (1995), originally presented by Box and Youle (1955). The descriptive statistics of the data are given in Table 5.1.

Table 5.1: *The Descriptive Statistics of the Dependent Variables (y_h) and the Predictors (x_j)*

Means and standard deviations of y_h			Means and standard deviations of x_j			Sample covariance matrix of the y_hs			
h	$\bar{y}_h$	$s.d.(y_h)$	j	$\bar{x}_j$	$s.d.(x_j)$	h	1	2	3
1	20.18	9.70	1	167.32	6.05	1	0.06	-0.03	-0.07
2	56.34	4.59	2	27.18	4.12	2	-0.03	0.79	-0.40
3	20.78	6.55	3	6.50	1.59	3	-0.07	-0.40	0.36

Let there be $t = 3$ dependent variables, namely 1) y_1, the percentage of unchanged starting material, 2) y_2, the percentage converted to the desired product, and 3) y_3, the percentage of unwanted by-product. These dependent variables are measured in experiments involving a chemical reaction in which various combinations of the temperature (x_1), the concentration (x_2), and the time (x_3) were used. These three are used as the predictors in a multivariate regression model. The data resulted from $N = 19$ designed experiments. The multivariate regression model including the constant can be written as

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \Bigg| x_1, x_2, x_3 \sim \mathcal{N}_{57} \left([I_3 \otimes [\iota,\ x_1,\ x_2,\ x_3]] \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, V = \Sigma \otimes I_{19} \right),$$

with ι a column vector of ones of size 19 and $\beta_h = [\beta_{h0}, \beta_{h1}, \beta_{h2}, \beta_{h3}]'$ the parameter assigned to y_h for $h = 1, 2, 3$. The covariance matrix Σ is estimated from the data by

$$\hat{\Sigma} = \begin{bmatrix} 0.04 & -0.03 & -0.05 \\ -0.03 & 0.62 & -0.31 \\ -0.05 & -0.31 & 0.28 \end{bmatrix}.$$

This estimate is used in determining the level probabilities $w_l(tk, W, \mathcal{C}_m)$. Namely, W is estimated by $\hat{\Sigma} \otimes [X'X]^{-1}$.

We expect that the three predictors each have a negative impact on y_1, the percentage of unchanged starting material, and a positive effect on the changed material measured by y_2 and y_3. Stated differently, we expect that $\beta_{1j} \leq 0$ and $\beta_{hj} \geq 0$ for $h = 2, 3$ and $j = 1, 2, 3$. One theory could be that the effects of both temperature (x_1) and concentration (x_2) are lower on the percentage converted to the desired product (y_2) than on that of unwanted by-product (y_3), which yields $\beta_{2j} \leq \beta_{3j}$ for $j = 1, 2$. Based on this, we can formulate two competing hypotheses, namely H_1 and H_2 which are stated below. Moreover, we would like to know whether H_1 or H_2 is the preferred hypothesis. Since both can be bad/weak, it is informative to include the unconstrained hypothesis H_u in which there are no restrictions on the parameters. For illustration purposes, we also include H_0, in which the β parameters regarding the predictors are restricted to zero, in the set of hypotheses:

$$\begin{aligned}
H_0 &: \ \beta_{h0},\ \beta_{hj} = 0,\ \text{for } h = 1, 2, 3 \text{ and } j = 1, 2, 3, \\
H_1 &: \ \beta_{h0},\ \beta_{1j} \leq 0,\ \beta_{hj} \geq 0,\ \text{for } h = 2, 3 \text{ and } j = 1, 2, 3, \\
H_2 &: \ \beta_{h0},\ \beta_{1j} \leq 0,\ \beta_{hj} \geq 0,\ \text{for } h = 2, 3 \text{ and } j = 1, 2, 3,\ \text{and} \\
&\quad\ \beta_{2j} \leq \beta_{3j},\ \text{for } j = 1, 2, \\
H_u &: \ \beta_{h0},\ \beta_{h1},\ \beta_{h2},\ \beta_{h3},\ \text{for } h = 1, 2, 3.
\end{aligned}$$

To compare the parameters β we have to standardize the dependent variables and the predictors. In that case, the intercepts β_{h0} are zero.

In Table 5.2, the order-restricted maximum likelihood estimators of β ($\tilde{\beta}^m$), the log likelihood values ($\log f(y|[I_t \otimes X]\tilde{\beta}^m, \tilde{\Sigma}^m \otimes I_{19})$, the penalty terms ($PT_m$), and the GORIC values are given for the four hypotheses of interest. The hypothesis with the lowest GORIC value is the preferred one. Hence, it is concluded that H_2 is the preferred hypothesis.

This example illustrates that the evaluation of a set of competing hypotheses is done easily.

5.8 Discussion

In this chapter, we derived (based on the ORIC of Anraku (1999)) the GORIC, an information criterion that evaluates competing hypotheses in univariate and multivariate normal linear models, where the restrictions are of the type $\theta \in \mathcal{C}_m$, with θ is a vector of length tk and $\mathcal{C}_m$ a closed convex cone. For univariate regression models, evaluating hypotheses with the GORIC is simple and straightforward. In case of multivariate regression models, the same remains valid when Σ is known (up to a positive constant). When Σ is unknown, one has to estimate Σ to compute the penalty term of the GORIC, but the calculation remains easy.

Table 5.2: GORIC *of the Four Specified Hypotheses* (H_m)

		Restricted β's ($\tilde{\beta}^m_{hj}$)					
m	j	$\tilde{\beta}^m_{1j}$	$\tilde{\beta}^m_{2j}$	$\tilde{\beta}^m_{3j}$	$\log f(.)$	PT_m	GORIC$_m$
0	0	0.00	0.00	0.00	-42.33	5.00	94.66
	1	0.00	0.00	0.00			
	2	0.00	0.00	0.00			
	3	0.00	0.00	0.00			
1	0	0.00	0.00	0.00	-9.97	10.08	40.09
	1	-0.96	0.53	0.84			
	2	-0.61	0.26	0.57			
	3	-0.37	0.36	0.28			
2	0	0.00	0.00	0.00	-9.97	9.72	**39.37**
	1	-0.96	0.53	0.84			
	2	-0.61	0.26	0.57			
	3	-0.37	0.36	0.28			
u	0	0.00	0.00	0.00	-9.97	13.00	45.93
	1	-0.96	0.53	0.84			
	2	-0.61	0.26	0.57			
	3	-0.37	0.36	0.28			

Note. GORIC = generalized order-restricted information criterion.
Bolding indicates the lowest value.

5.A Theorem 1

According to Silvapulle and Sen (2005, pp. 75), where $\theta^0 = 0$,

$$\sum_{i=1}^{N}[y_i - \tilde{\theta}^m]' V^{-1} \tilde{\theta}^m = 0.$$

In other words, $y_i - \tilde{\theta}^m$ and $\tilde{\theta}^m$ are V-orthogonal. Since θ^0 is a constant, this property remains valid when y and $\tilde{\theta}^m$ are both shifted by θ^0, that is

$$\sum_{i=1}^{N}[(y_i - \theta^0) - (\tilde{\theta}^m - \theta^0)]' V^{-1} (\tilde{\theta}^m - \theta^0) = 0.$$

Using this property yields

$$\begin{aligned}
&\sum_{i=1}^{N}(y_i - \theta^0)' V^{-1} (y_i - \theta^0) \\
&= \sum_{i=1}^{N}([y_i - \tilde{\theta}^m] + [\tilde{\theta}^m - \theta^0])' V^{-1} ([(y_i - \tilde{\theta}^m] + [\tilde{\theta}^m - \theta^0]) \\
&= \sum_{i=1}^{N}(y_i - \tilde{\theta}^m)' V^{-1} (y_i - \tilde{\theta}^m) + N(\tilde{\theta}^m - \theta^0)' V^{-1} (\tilde{\theta}^m - \theta^0).
\end{aligned}$$

The analogue remains true for t-variate normal linear models, where θ should be replaced by $[I_t \otimes X]\beta$ and N by 1. Notably, in this case, y shifted by $[I_t \otimes X]\beta^0$ and $\tilde{\beta}^m$ by β^0

5.B The Expression for var $\left(\tilde{\sigma}_m^2/\sigma^2\right)$

Let $Q = -(\tilde{\theta}^m - \theta^0)' \left(U/N\right)^{-1} (\tilde{\theta}^m - \theta^0)/\sigma^2$. From (5.11), it follows that

$$E\left\{\tilde{\sigma}_m^2/\sigma^2\right\} = 1 - \frac{1}{Nk} E\left\{-Q\right\}, \text{ and}$$
$$\text{var}\left(\frac{\tilde{\sigma}_m^2}{\sigma^2}\right) = \frac{1}{(Nk)^2}\left[\text{var}\left(\chi^2_{Nk}\right) + \text{var}\left(Q\right)\right],$$

with var $\left(\chi^2_{Nk}\right) = 2Nk$. Because $Q = O(1)$,

$$\text{var}\left(Q\right) = E\left\{Q^2\right\} - \left(E\left\{Q\right\}\right)^2 = O(1), \text{and}$$
$$\text{var}\left(\frac{\tilde{\sigma}_m^2}{\sigma^2}\right) = \frac{1}{(Nk)^2}\left[2Nk + \text{var}\left(Q\right)\right] = \frac{1}{(Nk)^2}\left[2Nk + O(1)\right]$$
$$= \frac{2}{Nk} + O((Nk)^{-2}).$$

Using the power series expansion, it holds true that

$$\frac{1}{1 - \frac{1}{Nk} E\left\{-Q\right\}} = 1 + \frac{1}{Nk} E\left\{-Q\right\} + O((Nk)^{-2}).$$

Now, it follows from (5.9) that

$$E\left\{\frac{\sigma^2}{\tilde{\sigma}_m^2}\right\} = \left[1 + \frac{1}{Nk} E\left\{-Q\right\} + O((Nk)^{-2})\right] + \frac{\frac{2}{Nk} + O((Nk)^{-2})}{\left[1 + \frac{1}{Nk} E\left\{-Q\right\} + O((Nk)^{-2})\right]^3}$$
$$= 1 + \frac{1}{Nk} E\left\{-Q\right\} + \frac{2}{Nk} + O((Nk)^{-2}),$$

which equals the expression in (5.12).

The analogue remains true for t-variate normal linear models, where θ should be replaced by β, Nk by tN, and U/N by $S \otimes [X'U^{-1}X]^{-1}$.

CHAPTER 6

Model Selection under Inequality Constraints in Small Samples

Kuiper, R. M.

Manuscript submitted.

The generalized order-restricted information-criterion is a modification of the Akaike information criterion such that it can be applied to order restrictions in univariate or multivariate regression models. However, a bias can occur in case of small samples or when there are many parameters in comparison with the sample size. A bias correction to the generalized order-restricted information-criterion is derived for univariate and multivariate regression models. Simulation shows that the corrected criterion has good frequency properties and that it outperforms the generalized order-restricted information-criterion in regression models in case of small sample sizes or rather when the number of parameters is moderate to large in comparison with the sample size.

6.1 Introduction

An often used information criterion is the Akaike information criterion (Akaike, 1973). Sugiura (1978) shows that the Akaike information criterion tends to overfit when the number of parameters is moderate to large in relation to the sample size or when the sample sizes are small. This bias is corrected in the small-sample Akaike information criterion (Hurvich & Tsai, 1989; Sugiura, 1978) and is, like the Akaike information criterion, of the form AICC $= -2\{$maximum log likelihood $-$ penalty$\}$. For univariate regression models, the penalty is calculated by

$$\frac{N(p+1)}{N-p-2}, \tag{6.1}$$

where N is the sample size and p the number of distinct regression parameters. The same remains valid in analysis of variance models with p distinct group means, k

groups, and $N = \sum_{j=1}^{k} n_j$ number of observations. For multivariate regression models, it is calculated by

$$\frac{tN \dim(\chi)}{tN - \dim(\chi) - 1}, \tag{6.2}$$

where $\dim(\chi)$ is the number of distinct parameters, that is, the number of distinct regression parameters and the number of distinct covariance elements. Bear in mind that, in univariate regression models and analysis of variance models, $t = 1$ and $\dim(\chi) = p+1$ in which case (6.2) reduces to (6.1). If the ratio $N/\dim(\chi)$ is sufficiently large, the Akaike information criterion and the corrected version will choose the same model. Burnham and Anderson (2002) prefer the use of the AICC when $N/\dim(\chi) < 40$.

In this note, we derive the corrected penalty for the generalized order-restricted information-criterion for univariate and multivariate regression models. We end with a simulation study in which the generalized order-restricted information-criterion and the corrected criterion are compared.

6.2 The small-sample generalized order-restricted information criterion

A t-variate normal linear model can be written as

$$y|\, X \sim \mathcal{N}_{tN}\left([I_t \otimes X] \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_t \end{bmatrix}, V = \Sigma \otimes U \right),$$

where $y = \begin{bmatrix} y_1 \\ \vdots \\ y_t \end{bmatrix} \in R^{tN\times 1}$, $I_t \otimes X = \text{diag}(X, \ldots, X) \in R^{tN\times tk}$, $\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_t \end{bmatrix} \in R^{tk\times 1}$, $V \in R^{tN\times tN}$, $\Sigma \in R^{t\times t}$, and $U \in R^{N\times N}$, with $y_h \in R^{N\times 1}$, $X \in R^{N\times k}$, and $\beta_h \in R^{k\times 1}$ for $h = 1, \ldots, t$. Notably, the analysis of variance models and univariate regression models are special cases of multivariate regression models.

The generalized order-restricted information-criterion is an estimate of the Kullback–Leibler discrepancy or rather of minus two times the expected log-likelihood. The difference with the Akaike information criterion or the order-restricted information-criterion of Anraku (1999) is that it can evaluate more general restrictions, namely all linear (order) restrictions except for range restrictions. It is of the form GORIC $= -2\{$maximum order-restricted log likelihood $-$ penalty$\}$ (Kuiper et al., 2011; Kuiper, Hoijtink, & Silvapulle, unpublished). The penalty is based on

$$B^m(\beta^0, V) = -\frac{tN}{2} + \frac{tN}{2} E\left\{ \frac{\sigma^2}{\tilde{\sigma}^2_m} \right\} + \frac{1}{2} E\left\{ ([I_t \otimes X]\tilde{\beta}^m - [I_t \otimes X]\beta^0)' \left(\tilde{V}^m\right)^{-1} ([I_t \otimes X]\tilde{\beta}^m - [I_t \otimes X]\beta^0) \right\}.$$

According to Kuiper et al. (unpublished), the first and second expectations equal

$$1 + \frac{2}{tN} + \frac{1}{tN}\sum_{l=1}^{tk} w_l(tk, W, \mathcal{C}_m)l + O((tN)^{-2}), \tag{6.3}$$

$$\sum_{l=1}^{tk} w_l(tk, W, \mathcal{C}_m)l + O((tN)^{-1}), \tag{6.4}$$

respectively, which results in

$$B^m(\beta^0, V) = 1 + \sum_{l=1}^{tk} w_l(tk, W, \mathcal{C}_m)l + O((tN)^{-1}), \tag{6.5}$$

where $w_l(.)$ is a level probability, also called chi square weight (Silvapulle & Sen, 2005). This yields, for model / hypothesis H_m,

$$\begin{aligned} \text{GORIC}_m &= -2\ \log f(y|[I_t \otimes X]\tilde{\beta}^m, \tilde{V}^m) + 2\ PT_m, \text{ with} \\ \log f(y|[I_t \otimes X]\tilde{\beta}^m, \tilde{V}^m) &= -\frac{tN}{2}\ \log\{2\pi\} - \frac{1}{2}\log|\tilde{V}^m| - \\ &\quad \frac{1}{2}\left[(y - [I_t \otimes X]\tilde{\beta}^m)' \left(\tilde{V}^m\right)^{-1} (y - [I_t \otimes X]\tilde{\beta}^m)\right], \\ PT_m &= 1 + \sum_{l=1}^{tk} w_l(tk, W, \mathcal{C}_m)\ l. \end{aligned}$$

The small-sample bias correction is equal to the term $O((tN)^{-1})$ in (6.5). Hence, when the two expectations are not compressed to terms of order 1 and higher, we obtain the small-sample generalized order-restricted information-criterion:

$$\text{GORICC}_m = -2\ \log f(y|[I_t \otimes X]\tilde{\beta}^m, \tilde{V}^m) + 2\ B^m(\beta^0, V).$$

Equation (6.4) equals

$$\sum_{l=1}^{tk} w_l(tk, W, \mathcal{C}_m)l \left(\frac{tN}{tN - l - 2}\right)$$

and (6.3) can be derived in two ways, leading to GORICC_m^a (based on $B_a^m(\beta^0, V)$) and GORICC_m^b (based on $B_b^m(\beta^0, V)$). It can be grounded on a second order Taylor expansion of $\frac{1}{x}$ around $E\{x\}$, with $x = \tilde{\sigma}_m^2/\sigma^2$. This leads to

$$E\left\{\frac{\sigma^2}{\tilde{\sigma}_m^2}\right\} = \frac{tN\left\{(tN - p(tk))^2 + 2tN + q(tk)\right\}}{(tN - p(tk))^3},$$

with

$$\begin{aligned} p(tk) &= \sum_{l=1}^{tk} w_l(tk, W, \mathcal{C}_m)l, \\ q(tk) &= -2p(tk) + \sum_{l=1}^{tk} w_l(tk, W, \mathcal{C}_m)l^2 - \{p(tk)\}^2, \end{aligned}$$

resulting in

$$B_a^m(\beta^0, V) = -\frac{tN}{2} + \frac{tN}{2}\frac{tN\left\{(tN - p(tk))^2 + 2tN + q(tk)\right\}}{(tN - p(tk))^3} + \frac{1}{2}\sum_{l=1}^{tk} w_l(tk, W, \mathcal{C}_m) l\left(\frac{tN}{tN - l - 2}\right).$$

The other derivation is analogously to the one of Sugiura (1978) and Hurvich and Tsai (1989). For a given level l,

$$\frac{\sigma^2}{\tilde{\sigma}_m^2} = \frac{tN}{tN - l}\frac{1}{(tN\frac{\tilde{\sigma}_m^2}{\sigma^2})^2/(tN - l)}.$$

Since the last ratio follows an $F(1, tN - l)$ distribution,

$$E\left\{\frac{\sigma^2}{\tilde{\sigma}_m^2}\right\} = \sum_{l=0}^{tk} w_l(tk, W, \mathcal{C}_m)\frac{tN}{tN - l - 2}.$$

This yields

$$\begin{aligned} B_b^m(\beta^0, V) &= -\frac{tN}{2} + \frac{tN}{2}\sum_{l=0}^{tk} w_l(tk, W, \mathcal{C}_m)\frac{tN}{tN - l - 2} + \\ &\quad \frac{1}{2}\sum_{l=1}^{tk} w_l(tk, W, \mathcal{C}_m) l\left(\frac{tN}{tN - l - 2}\right) \\ &= \sum_{l=0}^{tk} w_l(tk, W, \mathcal{C}_m)\frac{tN(l+1)}{tN - l - 2}. \end{aligned}$$

For univariate regression models the same remains true, but evidently for $t = 1$. For analysis of variance models, the same remains valid, but for $t = k$, $k = 1$, and $N = \sum_{j=1}^{k} n_j$.

6.3 Simulation

A simulation study was conducted to evaluate the performance of the generalized order-restricted information criterion and the two small-sample versions. To obtain a first insight in the difference between these two criteria, we will examine the three criteria in an univariate regression model. We employ the simulation design used by Hurvich and Tsai (1989), that is, the regression parameter vector equals $(1, 2, 3, 0, 0, 0, 0)$ and the variance is set to 1. Also here, the seven predictors each come from a standard normal distribution. Logically, Hurvich and Tsai (1989) investigated equality constrained models. We feel that the researcher is often not interested in examining which predictors contribute, but in which contribute more (i.e., directional effects). Therefore, we examined the following set of order-restricted models.

$$
\begin{aligned}
H_1 &: \ \theta_1 > 0, \theta_2 > 0, \theta_3 = \theta_4 = \theta_5 = \theta_6 = \theta_7 = 0,\\
H_2 &: \ \theta_1 > 0, \theta_2 > 0, \theta_3 > 0, \theta_4 = \theta_5 = \theta_6 = \theta_7 = 0,\\
H_3 &: \ \theta_1 > 0, \theta_2 > 0, \theta_3 > 0, \theta_4 > 0, \theta_5 = \theta_6 = \theta_7 = 0,\\
H_4 &: \ \theta_1 > 0, \theta_2 > 0, \theta_3 > 0, \theta_4 > 0, \theta_5 > 0, \theta_6 = \theta_7 = 0,\\
H_5 &: \ \theta_1 > 0, \theta_2 > 0, \theta_3 > 0, \theta_4 > 0, \theta_5 > 0, \theta_6 > 0, \theta_7 = 0,\\
H_6 &: \ \theta_1 > 0, \theta_2 > 0, \theta_3 > 0, \theta_4 > 0, \theta_5 > 0, \theta_6 > 0, \theta_7 > 0.
\end{aligned}
$$

Observe that H_2 is the correct model. Table 6.1 displays the percentage of times, out of 1,000 times, each model was chosen for $N = 10$ and 20. This table shows that the two small-sample corrected versions (GORICC^a_m and GORICC^b_m) outperform the GORIC and that the difference are compelling. Comparably to the AIC, the GORIC tends to overfit the model and the small-sample corrected versions perform best.

Table 6.1: Percentage of times that the models H_1, H_2, H_3, H_4, H_5 and H_6 were chosen by the generalized order-restricted information criterion and the two small-sample versions for $N = 10$ and 20

N	Method	H_1	H_2	H_3	H_4	H_5	H_6
10	GORIC$_m$	0	71	10	6	5	8
	GORICC^b_m	1	90	5	2	1	1
	GORICC^a_m	1	88	6	3	1	1
20	GORIC$_m$	0	74	10	6	5	4
	GORICC^b_m	0	85	8	4	3	1
	GORICC^a_m	0	84	8	4	3	2

The computer program for computing the generalized order-restricted information criterion and its two small-sample versions is available from `http://staff.fss.uu.nl/RMKuiper`.

CHAPTER 7

Remaining Issues

regarding the Generalized Order-Restricted Information Criterion

Kuiper, R. M.

7.1 Normal Distributions with Known Variance Ratios

One of the assumptions of the analysis of variance model is that the group variances are equal. When there is evidence for heterogeneity (using Levene's test), one can employ (analogously to Section 5.6) the sample estimates of the variance ratios as the known variance ratios in calculating the GORIC.

Let the data y_{ij} be normally and independently distributed with means θ_i and variances $\sigma_i^2 = \tau_i\sigma^2$, that is, $y_{ij} \sim \mathcal{N}(\theta_i, \tau_i\sigma^2)$ for $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$, where τ_i is known and σ^2 unknown. Notably, this is a univariate normal linear model and can be written as the model in Section 5.3.1, with $y = (y_1', \ldots, y_k')'$, $N = \sum_{i=1}^{k} n_i$, $\beta = \theta$, $X = (d_1, \ldots, d_k)$, where $d_{ih} = 1$ if observation h $(h = 1, \ldots, N)$ belongs to group i and zero otherwise, hence d_i represents group membership and consists of n_i ones and $N - n_i$ zeros, and $U = \text{diag}(\tau_1 I_{n_1}, \ldots, \tau_k I_{n_k})$, where I_{n_i} is the $n_i \times n_i$ identity matrix. Thus, in case of heterogeneity, one can use the expression of the GORIC on page 88, with $W = \sigma^2(X'U^{-1}X)^{-1} = \sigma^2 \text{diag}\left\{\frac{\tau_1}{n_1}, \ldots, \frac{\tau_k}{n_k}\right\}$.

7.2 Generalized Order-Restricted Information Criterion Weights

As can be seen from the four examples in Chapters 4, 5, and 13, the interest does not lie in the GORIC values, but in their differences. Since the likelihood increases with the number of observations, the GORIC values themselves are not interpretable. To improve the interpretation, we introduce GORIC weights (w_m), comparable to the Akaike weights (Burnham & Anderson, 2002, p. 75), with

$$w_m = \frac{exp\{-1/2(GORIC_m - GORIC_{min})\}}{\sum_{m' \in \mathcal{M}} exp\{-1/2(GORIC_{m'} - GORIC_{min})\}},$$

where $\mathcal{M}$ is the set of (say, M) hypothesis indices and $GORIC_{min}$ is the lowest GORIC value, that is, the GORIC value of the preferred model. Because the GORIC can be seen as a likelihood for Hypothesis m, the GORIC weight represents the relative likelihood (or, stated otherwise, the weight of evidence) of Hypothesis m given the data and the set of M hypotheses.

Table 7.1: GORIC *Weights of Four Examples*

Example	m	GORIC$_m$	w_m
Zelano et al. (1972, see Chapter 4)	0	90.73	0.06
$n_1 = n_2 = n_4 = 6, n_3 = 5$	1	86.23	0.54
	2	87.25	0.32
	u	90.03	0.08
Rencher (1995, see Chapter 5)	0	90.66	0.00
$N = 19$	1	40.09	0.40
	2	39.37	0.58
	u	45.93	0.02
Lievens and Sanchez (2007, see Chapter 13)	1	55.38	0.44
$n_1 = 21, n_2 = 25, n_3 = 26$	2	55.50	0.42
	u	57.70	0.14
Silvapulle and Sen (2005, see Chapter 13)	0	821.09	0.00
$n_1 = n_2 = n_3 = n_4 = 10$	1	808.66	0.07
	u	803.61	0.93
Lucas (2003, see Chapters 3 and 12)	0	588.54	0.00
$n_1 = \ldots n_5 = 30$	1	562.49	0.91
	2	569.79	0.02
	u	568.10	0.06

Note. GORIC = generalized order-restricted information criterion and w_m is the GORIC weight for Hypothesis m.

For the five examples in Chapters 3 (and 12), 4, 5, and 13, the GORIC weights are given in Table 7.1. From these weights, one can also determine the relative evidence for Hypothesis m compared to m'. For instance, in the example of Zelano et al. (1972), H_1 is $0.54/0.08 \approx 6.75$ more likely than H_u. Therefore, it is not a weak hypothesis. Furthermore, H_1 is $0.54/0.32 \approx 1.67$ more likely than H_2. Thus, although H_1 is the preferred hypothesis in the set (and not weakly supported by the data), there is no compelling evidence, since H_2 receives quite some support as well. A similar observation can be made for H_2 and H_1, respectively, in the example of Rencher (1995). In the example of Lievens and Sanchez (2007), H_1 and H_2 receive (about) the same amount of support (and are not weak). Therefore, both H_1 and H_2 are preferred in this set. Bear in mind that in all three cases the second preferred hypothesis is contained in the preferred one and that they strongly resemble each other. For instance, in the third example, $H_2 : \beta_1 \geq \beta_2 \geq 2\ \beta_3$ is contained in $H_1 : \beta_1 \geq \beta_2 \geq \beta_3$ and resembles it. In contrast, there is eminent support for one hypothesis in both the examples of Silvapulle and Sen (2005) and Lucas (2003). In the first, H_u

is preferred and it has $0.93/0.07 \approx 12.49$ times more support than H_1. In the latter, H_1 is preferred and it has $0.91/0.06 \approx 16.53$ times more support than H_1. For the example of Lucas (2003), the posterior model probabilities are also calculated (see Tables 3.5 and 12.4). Observe that (in this example) the GORIC weights resemble the posterior model probabilities.

Note that, in the first three examples, the differences in GORIC values for H_1, H_2, and H_u equal the differences in penalty term values, since the data are in accordance with all three hypotheses (rendering the same likelihood). Hence, increasing the number of observations does not affect the relative evidence (assuming that the data are still in agreement with the hypotheses and, additionally for the second example, that the covariance matrix estimate remains the same). One should perhaps take into account the maximum value of the relative evidence for two hypotheses, when the data are in accordance with these two or when their likelihood values are the same. Guidelines for GORIC weights are not available yet. More research is required regarding the performance of these weights, like Burnham and Anderson (2002, p. 75) did for the Akaike weights.

7.3 Simulation Study of the Generalized Order-Restricted Information Criterion and the Two Small-Sample Versions

In this section, we further examine the properties of the GORIC and its small-sample versions GORICC^a and GORICC^b. All three are of the form $IC = -2\ \log f(y|[I_t \otimes X]\tilde{\beta}^m, \tilde{V}^m) + 2\ PT_m^{IC}$, with

$$\begin{aligned}
PT_m^{\text{GORIC}} &= 1 + \sum_{l=1}^{tk} w_l(tk, W, \mathcal{C}_m)\ l, \\
PT_m^{\text{GORICC}^a} &= -\frac{tN}{2} + \frac{tN}{2}\frac{tN\left\{(tN - p(tk))^2 + 2tN + q(tk)\right\}}{(tN - p(tk))^3} + \\
&\quad \frac{1}{2}\sum_{l=1}^{tk} w_l(tk, W, \mathcal{C}_m) l \left(\frac{tN}{tN - l - 2}\right), \\
PT_m^{\text{GORICC}^b} &= \sum_{l=0}^{tk} w_l(tk, W, \mathcal{C}_m)\frac{tN(l+1)}{tN - l - 2},
\end{aligned}$$

where

$$\begin{aligned}
p(tk) &= \sum_{l=1}^{tk} w_l(tk, W, \mathcal{C}_m) l, \\
q(tk) &= -2p(tk) + \sum_{l=1}^{tk} w_l(tk, W, \mathcal{C}_m) l^2 - \left\{p(tk)\right\}^2.
\end{aligned}$$

We inspect ANOVA models where sets contain one, more than one, or none order-restricted models / hypotheses. In ANOVA models, the small-sample versions do not per se outperform the GORIC(since $N = \sum_{i=1}^{k} n_i$ is often not that low

due to multiple observations for multiple groups). Therefore, we also inspect them in regression models (with N observations). There, the small-sample versions do outperform the GORIC.

ANOVA Model - One Order-Restricted Model

For this simulation study, we chose the design of a real data example for which GORIC, GORICC^a, and GORICC^b would be useful, namely the one of Berzonsky et al. (2003) discussed in Chapter 4. We use here the same simulation design, choice of effect size values, population means, and models of interest:

$$
\begin{aligned}
H_0 &: \ \theta_1 = \theta_2 = \theta_3 = \theta_4, \theta_5 = \theta_6 = \theta_7 = \theta_8, \\
H_1 &: \ \theta_1 \geq \{\theta_2, \theta_3, \theta_4\}, \theta_5 \geq \{\theta_6, \theta_7, \theta_8\}, \theta_1 \geq \theta_5, \theta_2 \geq \theta_6, \theta_3 \geq \theta_7, \theta_4 \geq \theta_8, \text{ and} \\
&\quad \ \theta_1 - \theta_5 \geq \{\theta_2 - \theta_6, \theta_3 - \theta_7, \theta_4 - \theta_8\}, \\
H_u &: \ \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8 \text{ are unrestricted.}
\end{aligned}
$$

For each value of θ, we generated 1,000 independent samples. For each simulated data set, we calculated the GORIC and the two small-sample versions for the three models of interest. The percentage of times each model was chosen are given in Table 7.2 to Table 7.6 for $n_i = n = 10$ to 150, respectively. It should be stressed that the key focus is the difference in performance between the three criteria. Tables 7.2 to 7.6 show that the performances of GORICC^a and GORICC^b are equal. For $n_i = 10$, there are four cells for which the difference was $0 \cdot 1$ and one cell with $0 \cdot 2$, and for $n_i = 20$ there was only one cell for which the difference was $0 \cdot 1$. Furthermore, the tables show that the performances of GORICC^a and GORICC^b resemble the one of the GORIC, especially for $n_i \geq 50$. Moreover, the GORICC^a and GORICC^b perform at least as good as the GORIC in Case 1. For Case 2, this only remains valid for $n_i = 10$ and $n_i = 20$ when $ES = 0 \cdot 4$. In contrast, the GORIC outperforms the GORICC^a and GORICC^b in Case 3. It should be stressed that all differences in performance are small.

Table 7.2: Percentage of times that H_0, H_1, and H_u were chosen by the GORIC, GORICC^a, and GORICC^b in an ANOVA model with $n_i = n = 10$

		Case 1: H_0 is true			Case 2: H_1 is true			Case 3: H_u is 'true'		
ES	Method	H_0	H_1	H_u	H_0	H_1	H_u	H_0	H_1	H_u
$0 \cdot 1$	GORIC$_m$	80	12	8	57	39	5	65	27	8
	GORICC^a_m	85	11	4	63	35	2	72	23	5
	GORICC^b_m	85	11	4	63	35	2	72	23	5
$0 \cdot 25$	GORIC$_m$	87	4	9	19	78	3	41	34	25
	GORICC^a_m	92	3	5	25	73	2	48	36	16
	GORICC^b_m	92	3	5	25	73	2	48	36	16
$0 \cdot 4$	GORIC$_m$	90	0	10	2	95	3	12	29	60
	GORICC^a_m	95	0	5	3	96	1	18	35	47
	GORICC^b_m	94	0	5	3	96	1	18	35	48

Table 7.3: Percentage of times that H_0, H_1, and H_u were chosen by the GORIC, GORICC^a, and GORICC^b in an ANOVA model with $n_i = n = 20$

		Case 1: H_0 is true			Case 2: H_1 is true			Case 3: H_u is 'true'		
ES	Method	H_0	H_1	H_u	H_0	H_1	H_u	H_0	H_1	H_u
$0 \cdot 1$	GORIC$_m$	84	9	7	48	49	3	60	30	10
	GORICC^a_m	87	8	5	52	47	2	63	29	8
	GORICC^b_m	87	8	5	52	47	2	63	29	8
$0 \cdot 25$	GORIC$_m$	91	1	9	7	92	1	18	35	47
	GORICC^a_m	94	1	6	8	91	1	23	38	40
	GORICC^b_m	94	1	6	8	91	1	23	38	40
$0 \cdot 4$	GORIC$_m$	91	0	9	0	99	1	1	9	91
	GORICC^a_m	94	0	6	0	99	1	1	11	88
	GORICC^b_m	94	0	6	0	99	1	1	11	88

Table 7.4: Percentage of times that H_0, H_1, and H_u were chosen by the GORIC, GORICC^a, and GORICC^b in an ANOVA model with $n_i = n = 50$

		Case 1: H_0 is true			Case 2: H_1 is true			Case 3: H_u is 'true'		
ES	Method	H_0	H_1	H_u	H_0	H_1	H_u	H_0	H_1	H_u
$0 \cdot 1$	GORIC$_m$	91	4	5	37	63	1	49	33	19
	GORICC^a_m	92	4	4	38	61	1	50	33	17
	GORICC^b_m	92	4	4	38	61	1	50	33	17
$0 \cdot 25$	GORIC$_m$	94	0	6	0	99	1	1	10	89
	GORICC^a_m	95	0	5	0	99	1	1	11	88
	GORICC^b_m	95	0	5	0	99	1	1	11	88
$0 \cdot 4$	GORIC$_m$	94	0	6	0	99	1	0	0	100
	GORICC^a_m	95	0	5	0	99	1	0	0	100
	GORICC^b_m	95	0	5	0	99	1	0	0	100

Table 7.5: Percentage of times that H_0, H_1, and H_u were chosen by the GORIC, GORICC^a, and GORICC^b in an ANOVA model with $n_i = n = 100$

		Case 1: H_0 is true			Case 2: H_1 is true			Case 3: H_u is 'true'		
ES	Method	H_0	H_1	H_u	H_0	H_1	H_u	H_0	H_1	H_u
$0 \cdot 1$	GORIC$_m$	93	2	5	22	78	0	26	36	38
	GORICC^a_m	94	2	4	22	78	0	27	36	37
	GORICC^b_m	94	2	4	22	78	0	27	36	37
$0 \cdot 25$	GORIC$_m$	95	0	5	0	100	0	0	0	100
	GORICC^a_m	95	0	5	0	100	0	0	0	100
	GORICC^b_m	95	0	5	0	100	0	0	0	100
$0 \cdot 4$	GORIC$_m$	95	0	5	0	100	0	0	0	100
	GORICC^a_m	95	0	5	0	100	0	0	0	100
	GORICC^b_m	95	0	5	0	100	0	0	0	100

Table 7.6: Percentage of times that H_0, H_1, and H_u were chosen by the GORIC, GORICC^a, and GORICC^b in an ANOVA model with $n_i = n = 150$

		Case 1: H_0 is true			Case 2: H_1 is true			Case 3: H_u is 'true'		
ES	Method	H_0	H_1	H_u	H_0	H_1	H_u	H_0	H_1	H_u
$0 \cdot 1$	GORIC$_m$	92	0	8	14	86	0	13	29	58
	GORICC^a_m	92	0	7	15	85	0	13	30	57
	GORICC^b_m	92	0	7	15	85	0	13	30	57
$0 \cdot 25$	GORIC$_m$	92	0	8	0	100	0	0	0	100
	GORICC^a_m	93	0	8	0	100	0	0	0	100
	GORICC^b_m	93	0	8	0	100	0	0	0	100
$0 \cdot 4$	GORIC$_m$	92	0	8	0	100	0	0	0	100
	GORICC^a_m	93	0	8	0	100	0	0	0	100
	GORICC^b_m	93	0	8	0	100	0	0	0	100

ANOVA Model - Four Order-Restricted Models

In the simulation above, only one order-restricted model is included in the set. In this section, we examine a set containing four order-restricted models namely the one of the previous study and three subsets. The following six models were studied:

H_0 : $\theta_1 = \theta_2 = \theta_3 = \theta_4, \theta_5 = \theta_6 = \theta_7 = \theta_8$,

H_1 : $\theta_1 \geq \{\theta_2, \theta_3, \theta_4\}, \theta_5 \geq \{\theta_6, \theta_7, \theta_8\}, \theta_1 \geq \theta_5, \theta_2 \geq \theta_6, \theta_3 \geq \theta_7, \theta_4 \geq \theta_8$, and $\theta_1 - \theta_5 \geq \{\theta_2 - \theta_6, \theta_3 - \theta_7, \theta_4 - \theta_8\}$,

H_2 : $\theta_1 \geq \{\theta_2, \theta_3, \theta_4\}, \theta_5 \geq \{\theta_6, \theta_7, \theta_8\}, \theta_1 \geq \theta_5, \theta_2 \geq \theta_6, \theta_3 \geq \theta_7, \theta_4 \geq \theta_8$,

H_3 : $\theta_1 \geq \{\theta_2, \theta_3, \theta_4\}, \theta_5 \geq \{\theta_6, \theta_7, \theta_8\}, \theta_1 - \theta_5 \geq \{\theta_2 - \theta_6, \theta_3 - \theta_7, \theta_4 - \theta_8\}$,

H_4 : $\theta_1 \geq \{\theta_2, \theta_3, \theta_4\}, \theta_5 \geq \{\theta_6, \theta_7, \theta_8\}$,

H_u : $\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8$ are unrestricted.

Table 7.7 shows the percentage of times each model was chosen for $n_i = n = 10$. Bear in mind that for higher group sizes the performance is more alike. Table 7.7 shows that the performances of GORICC^a and GORICC^b are equal. There are only a few minor differences (not shown here). Furthermore, the GORICC^a and GORICC^b outperform the GORIC in Case 1. In contrast, the GORIC outperforms the GORICC^a and GORICC^b in Case 3. For Case 2, the three criteria perform about equal. It should be stressed that all differences in performance are quite small. Notably, the same was observed in the previous simulation.

Table 7.7: Percentage of times that H_0, H_1, H_2, H_3, H_4, and H_u were chosen by the GORIC, GORICC^a, and GORICC^b in an ANOVA model with $n_i = n = 10$

		Case 1: H_0 is true						Case 2: H_1 is true						Case 3: H_u is 'true'					
ES	Method	H_0	H_1	H_2	H_3	H_4	H_u	H_0	H_1	H_2	H_3	H_4	H_u	H_0	H_1	H_2	H_3	H_4	H_u
$0 \cdot 1$	GORIC$_m$	76	6	2	7	5	4	54	28	10	4	2	1	63	19	4	7	2	6
	GORICC^a_m	83	6	1	5	3	2	60	27	8	3	1	1	71	17	3	5	1	3
	GORICC^b_m	83	6	1	5	3	2	60	27	8	3	1	1	71	17	3	5	1	3
$0 \cdot 25$	GORIC$_m$	79	1	0	10	6	4	18	68	5	8	1	1	39	24	3	13	1	20
	GORICC^a_m	85	1	0	7	4	2	23	67	4	5	1	0	49	24	2	11	1	13
	GORICC^b_m	85	1	0	7	4	2	23	67	4	5	1	0	49	24	2	11	1	13
$0 \cdot 4$	GORIC$_m$	80	0	0	10	6	4	3	86	1	10	0	1	10	22	1	14	1	52
	GORICC^a_m	86	0	0	8	4	2	4	88	1	7	0	0	16	27	1	14	1	40
	GORICC^b_m	86	0	0	8	4	2	4	88	1	7	0	0	16	27	1	14	1	40

ANOVA Model - No Order-Restricted Model

In the previous situations, the small-sample versions do not outperform the GORIC in general. Note that the GORIC reduces to the AIC and GORICC^b to the AICC when there are no order restrictions and Hurvich and Tsai (1989) show that the AICC outperforms the AIC for regression models. In this section, we examine whether this only holds true for regression models or also for ANOVA models with the same features: i) (evidently) in each of the models, the parameters are restricted to be zero or non-zero; ii) the correct model is included in the set; iii) almost all models in the set include the correct one and the exception does only exhibit one incorrect constraint out of seven. Observe that our Case 1, where the GORICC^a and GORICC^b perform better than the GORIC, exhibits these three features, namely H_0, containing only equality constrains, is true, is included in the set, and is contained in the other two models. However, there are a few exceptions: there is an order-restricted model, there are fewer competitors (in the first simulation), and there are 8 instead of 7 parameters of interest. Therefore, we also investigated an ANOVA model with 7 parameters of interest, with mean (1, 2, 3, 0, 0, 0, 0) and variance 1, and six models containing solely equality constrains:

$$
\begin{aligned}
H_1 &: \theta_1, \theta_2, \theta_3 = \theta_4 = \theta_5 = \theta_6 = \theta_7 = 0, \\
H_2 &: \theta_1, \theta_2, \theta_3, \theta_4 = \theta_5 = \theta_6 = \theta_7 = 0, \\
H_3 &: \theta_1, \theta_2, \theta_3, \theta_4, \theta_5 = \theta_6 = \theta_7 = 0, \\
H_4 &: \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6 = \theta_7 = 0, \\
H_5 &: \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7 = 0, \\
H_u &: \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7 \text{ are unrestricted.}
\end{aligned} \tag{7.1}
$$

Note that these are the models that Hurvich and Tsai (1989) used (although for a regression model) and that H_2 is the correct model. Table 7.8 shows the percentage of times, out of 100 times (instead of 1,000 in the previous two simulations), each model was chosen for $n_i = n = 10$ and 20. This table shows that the GORICC^a and GORICC^b outperform the GORIC, but all differences in performance are small. Hence, the improvement of the GORICC^b = AICC on the GORIC = AIC is small in ANOVA models. This seems to imply that the biggest improvement of the GORICC^a and GORICC^b is found in regression models and not in ANOVA models, like for the AICC. Note that the ANOVA model (with n_i observations for Group i) is a special case of the regression model (with $N = \sum_{i=1}^{k} n_i$ observations). Thus when n_i is low, N might not be due to the number of groups (k). Notably, in the inspected ANOVA model $N = \sum_{i=1}^{7} n_i = 70$ and 140. As a consequence, we inspect the regression model (with N observations) in the next section.

Table 7.8: Percentage of times that H_1, H_2, H_3, H_4, H_5, and H_u (hypotheses without order restrictions) were chosen by the GORIC, GORICC^a, and GORICC^b in an ANOVA model with $n_i = n = 10$ and 20

$n_i = n$	Method	H_1	H_2	H_3	H_4	H_5	H_u
10	GORIC$_m$ = AIC$_m$	0	72	12	8	4	4
	GORICC^b_m = AICC$_m$	0	81	10	5	3	1
	GORICC^a_m	0	80	11	5	3	1
20	GORIC$_m$	0	69	18	4	4	5
	GORICC^b_m = AICC$_m$	0	74	17	2	4	3
	GORICC^a_m	0	74	17	2	4	3

Simulation Study: Regression Model

In this section, we examine the three criteria in an univariate regression model. We employ the simulation design used by Hurvich and Tsai (1989), that is, the regression parameter vector equals $(1, 2, 3, 0, 0, 0, 0)$ and the variance is set to 1. The seven predictors each come from a standard normal distribution. We first examine the equality constrained models in Equation 7.1 and, subsequently, order-restricted ones.

For the models in Equation 7.1, Table 7.9 depicts the percentage of times, out of 1,000 times, each model was chosen for $N = 10$ and 20. Notably, H_2 is the correct model. Here, one can see that GORICC^a and GORICC^b = AICC outperform the GORIC = AIC and that the difference are now (more) compelling.

Table 7.9: Percentage of times that H_1, H_2, H_3, H_4, H_5, and H_u (hypotheses without order restrictions) were chosen by the GORIC, GORICC^a, and GORICC^b in a regression model with $N = 10$ and 20

N	Method	H_1	H_2	H_3	H_4	H_5	H_u
10	GORIC$_m$ = AIC$_m$	1	68	7	1	7	12
	GORICC^b_m = AICC$_m$	1	89	1	0	0	0
	GORICC^a_m	1	90	1	0	0	0
20	GORIC$_m$	0	78	1	0	4	5
	GORICC^b_m = AICC$_m$	0	94	0	0	0	0
	GORICC^a_m	0	94	0	0	0	0

We feel that the researcher is often not interested in examining which predictors contribute, but in which contribute more (i.e., directional effects). Therefore, we also examined the following set of order-restricted models.

$$
\begin{aligned}
H_1 &: \ \theta_1 > 0, \theta_2 > 0, \theta_3 = \theta_4 = \theta_5 = \theta_6 = \theta_7 = 0,\\
H_2 &: \ \theta_1 > 0, \theta_2 > 0, \theta_3 > 0, \theta_4 = \theta_5 = \theta_6 = \theta_7 = 0,\\
H_3 &: \ \theta_1 > 0, \theta_2 > 0, \theta_3 > 0, \theta_4 > 0, \theta_5 = \theta_6 = \theta_7 = 0,\\
H_4 &: \ \theta_1 > 0, \theta_2 > 0, \theta_3 > 0, \theta_4 > 0, \theta_5 > 0, \theta_6 = \theta_7 = 0,\\
H_5 &: \ \theta_1 > 0, \theta_2 > 0, \theta_3 > 0, \theta_4 > 0, \theta_5 > 0, \theta_6 > 0, \theta_7 = 0,\\
H_6 &: \ \theta_1 > 0, \theta_2 > 0, \theta_3 > 0, \theta_4 > 0, \theta_5 > 0, \theta_6 > 0, \theta_7 > 0.
\end{aligned}
$$

Observe that also here H_2 is the correct model. Table 7.10 displays the percentage of times, out of 1,000 times, each model was chosen for $N = 10$ and 20. This table shows again that the GORICC^a and GORICC^b outperform the GORIC and that the difference are compelling.

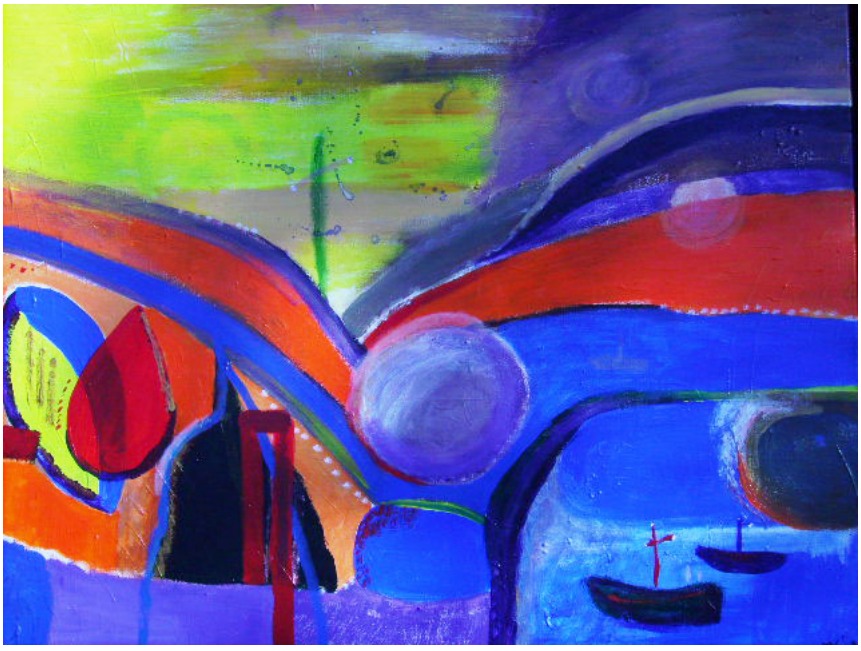
Thus, in case of small samples in regression models, one should employ a corrected version of the GORIC. We recommend the use of GORICC$_b$, since it performs a bit better and has a simpler expression than GORICC$_a$. It should be stressed that this simulation study gives a first insight. More research is needed for better insight into the properties of the small-sample versions of the GORIC. Research should, among others, be done for several effect sizes and under several true models (like the first simulation in this Chapter) for different numbers of observations.

Table 7.10: Percentage of times that H_1, H_2, H_3, H_4, H_5, and H_6 were chosen by the GORIC, GORICC^a, and GORICC^b in a regression model with $N = 10$ and 20

N	Method	H_1	H_2	H_3	H_4	H_5	H_6
10	GORIC$_m$	0	71	10	6	5	8
	GORICC^b_m	1	90	5	2	1	1
	GORICC^a_m	1	88	6	3	1	1
20	GORIC$_m$	0	74	10	6	5	4
	GORICC^b_m	0	85	8	4	3	1
	GORICC^a_m	0	84	8	4	3	2

Part III

Model Selection Criteria in Presence of Missing Data



... by Marga Klungel

CHAPTER 8

How to Handle Missing Data in Regression Models using Information Criteria

Kuiper, R. M., and Hoijtink, H.

Published in Statistica Neerlandica, 65(4), pp. 489-506.

An important application of multiple regression is predictor selection. When there are no missing values in the data, information criteria can be used to select predictors. For example, one could apply the small-sample-size corrected version of the Akaike information criterion (AIC), the AICC. In this chapter, we discuss how information criteria should be calculated when the dependent variable and/or the predictors contain missing values. Therewith, we extensively discuss and evaluate three models that can be employed to deal with the missing data, that is, to predict the missing values. The most complex model, that is, the model with all available predictors, outperforms the other models. These results also apply to more general hypotheses than predictor selection and also to structural equation modeling (SEM) models.

8.1 Introduction

When the goal is to estimate parameters, it is known how to deal with missing data. See, among others, Little and Rubin (1987), Schafer (1997), Jamshidian (2004), Schafer and Graham (2002), Scheffer (2002), Jamshidian and Bentler (1999), Hens, Aerts, and Molenberghs (2006), and Liu, Wei, and Zhang (2006). For example, the expectation-maximization (EM) algorithm or multiple imputation can be used. In contrast, there is not much literature on handling missing data in model selection using information criteria (ICs), like the Akaike information criterion (AIC; Akaike, 1973, 1974). Schafer (1997) and Little and Rubin (1987) discuss how the observed-likelihood should be calculated. Cavanaugh and Shumway (1998) propose a penalty for the AIC in presence of missing data for, among others, ANOVA and regression models. Claeskens and Consentino (2008) developed model selection criteria that can be used in regression models when the dependent variable is completely observed. These authors do not discuss which model should be used to predict the missing values.

This chapter will deal with predictor selection in regression models in the presence of missing values in the dependent variable and/or the predictors. We will focus on which model is assumed to be the underlying data model, since this model is used to predict the missing values. As a consequence, its choice influences the value of an IC and this might affect the result regarding the preferred hypothesis. Our findings also apply to more general hypotheses and to structural equation modeling (SEM) models. We will start with an introduction of the regression model and model selection.

8.1.1 The regression model

It is assumed that

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \text{ with } \boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n), \tag{8.1}$$

where

$$\begin{aligned}
\boldsymbol{y} &= [y_1, \ldots, y_n]', \\
\boldsymbol{X} &= [\boldsymbol{1}, \boldsymbol{x}] = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1k-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{nk-1} \end{bmatrix}, \\
\boldsymbol{x} &= \begin{bmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_i' \\ \vdots \\ \boldsymbol{x}_n' \end{bmatrix} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_j, \ldots, \boldsymbol{x}_{k-1}], \\
\boldsymbol{\beta} &= [\beta_0, \ldots, \beta_{k-1}]',
\end{aligned} \tag{8.2}$$

$\boldsymbol{0}$ an n-vector with zeros, and $\boldsymbol{I}_n$ the $n \times n$ identity matrix. It should be stressed that $\boldsymbol{x}_1$ is now defined twice, namely as $\boldsymbol{x}_1 = [x_{11}, \ldots, x_{1k-1}]'$ and as $\boldsymbol{x}_1 = [x_{11}, \ldots, x_{n1}]'$. From the context and/or subscript it will be clear to which vector we refer to. To deal with missing data in both the dependent variable $\boldsymbol{y}$ and the predictors $\boldsymbol{x}$, it is furthermore assumed that

$$\boldsymbol{x}_i \sim \mathcal{N}_{k-1}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx}), \tag{8.3}$$

with mean $\boldsymbol{\mu}_x = \left[\mu_{x_1}, \ldots, \mu_{x_{k-1}}\right]'$ and covariance matrix $\boldsymbol{\Sigma}_{xx}$.

From Equations (8.1) and (8.3), it follows that

$$\boldsymbol{z}_i = [y_i, \boldsymbol{x}_i']' \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{8.4}$$

with mean $\boldsymbol{\mu} = [\mu_y, \boldsymbol{\mu}_x']'$ and covariance matrix $\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \Sigma_{yy} & \boldsymbol{\Sigma}_{yx} \\ \hline \boldsymbol{\Sigma}_{yx}' & \boldsymbol{\Sigma}_{xx} \end{array}\right]$.

The density of the data $\boldsymbol{z}$ is

$$f(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{z}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}_i - \boldsymbol{\mu})\right\}.$$

In the sequel, $f(.)$ will be used as the notation for the probability density function and the likelihood, since they are proportional to each other. Note that the values of $\boldsymbol{\beta}$, σ^2, $\boldsymbol{\mu}_x$, and $\boldsymbol{\Sigma}_{xx}$ can be determined from the value of $\boldsymbol{\xi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The latter two follow directly from $\boldsymbol{\xi}$. The first two can be determined as follows:

$$\begin{aligned}\boldsymbol{\beta} &= (\boldsymbol{\Sigma}^{+}_{xx})^{-1}\boldsymbol{\Sigma}^{+\prime}_{yx} \\ &= \left[\begin{array}{c|c} 1 & \boldsymbol{\mu}_x' \\ \hline \boldsymbol{\mu}_x & \boldsymbol{\Sigma}_{xx} + \boldsymbol{\mu}_x\boldsymbol{\mu}_x' \end{array}\right]^{-1} \left[\mu_y \middle| \boldsymbol{\Sigma}_{yx} + \mu_y\boldsymbol{\mu}_x'\right]', \qquad (8.5)\\ \sigma^2 &= \Sigma_{yy} + \mu_y\mu_y - \boldsymbol{\Sigma}^{+}_{yx}\boldsymbol{\beta}, \qquad (8.6)\end{aligned}$$

where $\boldsymbol{\Sigma}^{+}$ is used to stress that the intercept is included in the set of parameters. These equations will be used in the next subsection.

8.1.2 Model selection

In model selection, a set of hypotheses is evaluated with an IC. There are several criteria, the two most familiar ones are the Bayesian IC (BIC; Neath & Cavanaugh, 2006) and the AIC. The first is designed to provide an asymptotic approximation to a transformation of the so-called posterior model probability and does not require the specification of priors. The latter originates from the so-called expected Kullback-Leibler (K-L) distance, which measures the expected information loss when the true data generating density is approximated by the density of the model / hypothesis of interest. The interested reader is referred to Burnham and Anderson (2002).

Let $\mathbb{M}$ be the set of hypothesis/model indices and let Hypothesis H_m have the form

$$H_m : \ \boldsymbol{C}_m\boldsymbol{\beta} = \mathbf{0}, \qquad (8.7)$$

for $m \in \mathbb{M}$, where the rows in $\boldsymbol{C}_m \in \mathbb{R}^{c_m \times k}$ are a permutation of $[1, 0, \ldots, 0]$, $\text{rank}(\boldsymbol{C}_m) = c_m \leq k$, and $\mathbf{0}$ is a vector of zeros with length c_m.

An information criterion IC, like the AIC, can be used to select the best of a set of hypotheses. For the model in Equation (8.4), it has the form

$$IC_m = -2 \ \log \ f(\boldsymbol{z}|\hat{\boldsymbol{\xi}}_{H_m}) + 2 \ p_m, \qquad (8.8)$$

with $\log \ f(\boldsymbol{z}|\hat{\boldsymbol{\xi}}_{H_m})$ the log-likelihood, $\hat{\boldsymbol{\xi}}_{H_m}$ the restricted maximum likelihood estimator (restricted MLE) of $\boldsymbol{\xi}$, that is,

$$\hat{\boldsymbol{\xi}}_{H_m} = \arg \max_{\boldsymbol{\xi} \in H_m} f(\boldsymbol{z}|\boldsymbol{\xi}),$$

and p_m the penalty part. For the AIC, p_m equals the number of distinct parameters. When there are no restrictions, $p_m = (k+1)k/2 + k$. It should be stressed that restricting a parameter β to zero comes down to restricting the corresponding element in $\boldsymbol{\Sigma}_{yx}$ and not in $\boldsymbol{\mu}_x$, $\boldsymbol{\mu}_y$ nor $\boldsymbol{\Sigma}_{xx}$. Hence, for the model in Equation (8.4), the penalty is equal to

$$p_m = \frac{(k+1)k}{2} + (k - c_m). \tag{8.9}$$

The preferred hypothesis is the one with the smallest IC value.

Let $\hat{\boldsymbol{\beta}}_{H_m}$ and $\hat{\sigma}^2_{H_m}$ be the restricted MLEs of $\boldsymbol{\beta}$ and σ^2, that is, the MLEs in accordance with the restrictions in H_m. The restricted MLEs $\hat{\boldsymbol{\beta}}_{H_m}$ and $\hat{\sigma}^2_{H_m}$ can, analogously to Equations (8.5) and (8.6), be determined from $\hat{\boldsymbol{\xi}}_{H_m} = (\hat{\boldsymbol{\mu}}_{H_m}, \hat{\boldsymbol{\Sigma}}_{H_m})$:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{H_m} &= (\hat{\boldsymbol{\Sigma}}^{+}_{H_m,xx})^{-1}\hat{\boldsymbol{\Sigma}}^{+\prime}_{H_m,yx}, \\
\hat{\sigma}^2_{H_m} &= \hat{\Sigma}_{H_m,yy} + \hat{\mu}_{H_m,y}\hat{\mu}_{H_m,y} - \hat{\boldsymbol{\Sigma}}^{+}_{H_m,yx}\hat{\boldsymbol{\beta}}_{H_m}.
\end{aligned}$$

Note that the likelihood can be written as

$$\begin{aligned}
f(\boldsymbol{z}|\hat{\boldsymbol{\xi}}_{H_m}) &= f(\boldsymbol{y}|\boldsymbol{x}, \hat{\boldsymbol{\xi}}_{H_m})\ f(\boldsymbol{x}|\hat{\boldsymbol{\xi}}_{H_m}) \\
&= f(\boldsymbol{y}|\boldsymbol{x}, \hat{\boldsymbol{\beta}}_{H_m}, \hat{\sigma}^2_{H_m}) f(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_x, \hat{\boldsymbol{\Sigma}}_{xx}).
\end{aligned} \tag{8.10}$$

The restrictions in $H_m : \boldsymbol{C}_m\boldsymbol{\beta} = \boldsymbol{0}$ come down to restricting the corresponding part in $\boldsymbol{\Sigma}_{yx}$ and not in $\boldsymbol{\Sigma}_{xx}$. For example, when $H_m : \beta_2 = 0$, the second element in $\hat{\boldsymbol{\Sigma}}_{yx}$ is adjusted, but $\hat{\boldsymbol{\mu}}_x$ and $\hat{\boldsymbol{\Sigma}}_{xx}$ are unaffected. Thus, $f(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_x, \hat{\boldsymbol{\Sigma}}_{xx})$ is constant over all hypotheses. Therefore, the IC for Hypothesis $H_m : \boldsymbol{C}_m\boldsymbol{\beta} = \boldsymbol{0}$ can also be calculated by

$$IC_m = -2\ \log\ f(\boldsymbol{y}|\boldsymbol{x}, \hat{\boldsymbol{\beta}}_{H_m}, \hat{\sigma}^2_{H_m}) + 2\ p_m. \tag{8.11}$$

Note that this is the expression of an IC for model (8.1). In this case, the penalty of the AIC equals

$$p_m = 1 + (k - c_m). \tag{8.12}$$

In a regression model (8.1), only the conditional distribution of the response $\boldsymbol{y}$ given the predictors $\boldsymbol{x}$ is relevant. As mentioned before, the model in Equation (8.4) is required in the context of missing values, because it explicitly models the distribution of the predictors. As a consequence, this results in an IC based on the joint density of $\boldsymbol{y}$ and $\boldsymbol{x}$, as in Equation (8.8). It is questionable whether this is appropriate in the context of multiple regression. However, it will be shown that in the presence of missing data the IC of the form (8.8) also reduces to an IC of the form (8.11) when employing the preferred approach.

In the next section, we will show how ICs of the form (8.8) or (8.11) should be calculated in presence of missing data. But first, we give some remarks about ICs used in predictor selection. Burnham and Anderson (2002, §6.4) argue that the AIC has theoretical advantages over the BIC. The AIC is an asymptotically unbiased estimator of the approximation of the K-L distance (Burnham & Anderson, 2002, p. 61). It should be stressed that the AIC does not assume that the true model is subsumed in the set of candidate models (Burnham & Anderson, 2002, p. 65). Both nested and non-nested models can be included in the set (Burnham & Anderson, 2002, p. 88). One drawback is that the AIC tends to overfit, that is, tends to choose the model with more parameters than present in the true model (Burnham & Anderson,

2002, p. 417). Another drawback is that the AIC may perform poorly when there are too many parameters in relation to the sample size (Burnham & Anderson, 2002, p. 66), that is, $\frac{n}{1+k-c_m} > 40$. In such cases in the context of regression models, one should use the so-called AICC / second-order AIC / small sample AIC proposed by Hurvich and Tsai (1989), which is calculated by Equation (8.11) with

$$p_m = \frac{1 + (k - c_m)}{1 - \frac{(k-c_m)+2}{n}}. \tag{8.13}$$

8.2 Missing data

Let $\boldsymbol{z} = [\boldsymbol{y}, \boldsymbol{x}]$ correspond to the complete set of measurements, $\boldsymbol{z}_{\text{obs}} = [\boldsymbol{y}_{\text{obs}}, \boldsymbol{x}_{\text{obs}}]$ to the observed measurements, and $\boldsymbol{z}_{\text{mis}} = [\boldsymbol{y}_{\text{mis}}, \boldsymbol{x}_{\text{mis}}]$ to the missing measurements. Furthermore, let $\boldsymbol{r} \in \mathbb{R}^{n \times k}$ be the missingness indicator, that is,

$$r_{ih} = \begin{cases} 1, \text{ if } z_{ih} \text{ is observed,} \\ 0, \text{ if } z_{ih} \text{ is missing,} \end{cases}$$

where z_{ih} corresponds to element h in $[y_i, x_{i1}, \ldots, x_{i,k-1}]$, for $i = 1, \ldots, n$ and $h = 1, 2, \ldots, k$. Throughout the chapter, it is assumed that the missing data are missing at random (MAR):

$$f(\boldsymbol{r}_i | \boldsymbol{z}_{\text{obs},i}, \boldsymbol{z}_{\text{mis},i}, \boldsymbol{\phi}) = f(\boldsymbol{r}_i | \boldsymbol{z}_{\text{obs},i}, \boldsymbol{\phi}),$$

where $\boldsymbol{\phi}$ is the parameter of the density of $\boldsymbol{r}_i$.

In the presence of missing data, two models should be distinguished: the analytical model and the assumed underlying data model. This is comparable to the analyst's model and the imputation model in multiple imputation discussed in Schafer (1997). The analytical model is the hypothesis of interest, that is, H_m. For example,

$$H_1 : \beta_1 = 0, \text{ that is, } H_1 : y_i = \beta_0 + \sum_{j=2}^{k-1} \beta_j x_{ij} + \epsilon_i \text{ for } i = 1, \ldots, n.$$

To deal with missing values, the so-called assumed underlying data model is needed. In the sequel, this model will be used to predict the values of the missing data. For example, one can assume that the hypothesis of interest is the underlying data model. In that case, only the predictors which parameters are not hypothesized to be zero and the dependent variable will be used in the prediction of the missing data. Let the hypothesis of interest be H_1 (see above) and let, for row i, only x_{ij} $(j \geq 2)$ be missing. The prediction of the missing value x_{ij} comes from the regression of $\boldsymbol{x}_j$ on the other predictors and the dependent variable, that is, $[\boldsymbol{x}_2, \ldots, \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, \ldots, \boldsymbol{x}_{k-1}]$ and $\boldsymbol{y}$. This will be elaborated upon in the next section. A missing value can of course also depend on a variable which parameter is set to zero in the hypothesis of interest or even on a variable which is not included in $\boldsymbol{x}$. Thus, the set of variables used for the prediction of the missing values can consist, beside the dependent variable, of (i) the predictors for $\boldsymbol{y}$ which coefficients are not set to zero in H_m, (ii) all the predictors

for $\boldsymbol{y}$, that is, $\boldsymbol{x}_1$ to $\boldsymbol{x}_{k-1}$, and (iii) all the predictors for $\boldsymbol{y}$ and additional variables, which will be denoted by $\boldsymbol{x}_k$ to $\boldsymbol{x}_{K-1}$. In the next section, we will use $L-1$ as the number of $\boldsymbol{x}$-variables. Depending on the assumed underlying data model, L takes on the value $k-c_m$ or K.

When two variables are correlated, they can be used in the prediction of the missing values of each other. Let $\boldsymbol{x}_j$ and $\boldsymbol{x}_{j'}$ be correlated and let x_{ij} be missing. In that case, we say that $\boldsymbol{x}_{j'}$ is a relevant predictor for the missing value x_{ij}. When all relevant predictors are used to predict a missing value, the prediction of the missing value will be unbiased, that is, the expected value of the prediction equals the true value. Hence, when missing a relevant predictor (like $\boldsymbol{x}_{j'}$ for $\boldsymbol{x}_j$) in the prediction of a missing value x_{ij}, the prediction will be biased.

Using additional variables to predict a missing value x_{ij} has a downside. In case variables are included in the prediction of x_{ij} that are not correlated with $\boldsymbol{x}_j$, the standard deviation of the prediction of x_{ij} increases. In other words, the prediction of x_{ij} will be less efficient. However, when all relevant variables are included, the prediction of x_{ij} is still unbiased.

In summary, the predictions of the missing values and whether the predictions are unbiased depend on the set of variables used for the prediction of the missing values, which in turn depends on the assumed underlying data model. Consequently, the choice of the assumed underlying data model is important. Three types of assumed underlying data models will be distinguished: the analytical model, the unconstrained model, and the restricted unconstrained model. These will be discussed in the next section.

The following questions with respect to predictor selection in the presence of missing data will be covered in this chapter:
1. How should the log-likelihood part of an IC be calculated when there are missing values in $\boldsymbol{y}$ and/or $\boldsymbol{x}$? This is addressed, for example, by Schafer (1997) and Little and Rubin (1987).
2. Which assumed underlying data model should be used?
3. Should the log-likelihood be based on the observed data $\boldsymbol{z}_{\text{obs}}$ or on the observed data $\boldsymbol{z}_{\text{obs}}$ and the observed missingness indicator $\boldsymbol{r}$?

By addressing the three questions above in the next sections, we will discuss calculating ICs, like the AIC and AICC, for predictor selection in regression models in the presence of missing data being MAR. Subsequently, we illustrate the AIC and AICC in the presence of missing data. We end with a discussion in which we also elaborate on the proposed approach.

8.3 ICs in the presence of missing data

8.3.1 Maximizing the observed-data likelihood using the EM algorithm

A data point in $\boldsymbol{z}$ is either observed or missing (as indicated by the missingness indicator $\boldsymbol{r}$). Let the data consist of L variables, namely one dependent variable and $L-1$ predictors. Hence, the maximum number of missing data patterns is 2^L. Note that not all missing data patterns have to exist in the data $\boldsymbol{z}$. Let S be the total

number of missing data patterns in $\boldsymbol{z}$ and let $I(s)$ denote all i for which $\boldsymbol{z}_i$ has missing data pattern $s = 1, \ldots, S$. Furthermore, let $\boldsymbol{z}_i^s$ be the observed part of $\boldsymbol{z}_i$ with missing data pattern s. The likelihood of the observed data $\boldsymbol{z}_{\text{obs}}$ is called the observed-data likelihood:

$$f(\boldsymbol{z}_{\text{obs}}|\boldsymbol{\xi}) = \prod_{s=1}^{S} \prod_{i \in I(s)} \frac{1}{(2\pi)^{L_s/2} |\boldsymbol{\Sigma}^s|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\boldsymbol{z}_i^s - \boldsymbol{\mu}^s)' (\boldsymbol{\Sigma}^s)^{-1} (\boldsymbol{z}_i^s - \boldsymbol{\mu}^s) \right\} \quad (8.14)$$

where $\boldsymbol{\Sigma}^s$ and $\boldsymbol{\mu}^s$ are the submatrix of $\boldsymbol{\Sigma}$ and the subvector of $\boldsymbol{\mu}$, respectively, corresponding to the observed variables in pattern s and L_s is the number of observed variables in pattern s (Schafer, 1997).

To obtain an estimate of $\boldsymbol{\xi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, when accounting for the missing data, the EM algorithm (Dempster, Laird, & Rubin, 1977; Little & Rubin, 1987; Schafer, 1997) can be used. Note that EM maximizes the complete-data likelihood. However, in well-behaved problems like addressed in this chapter, the parameters maximizing the complete-data likelihood equal the parameters maximizing the observed-data likelihood (Schafer, 1997; Little & Rubin, 1987).

Expectation-maximization is an iterative procedure and consists of two distinct steps: the E-step and the M-step. Since the data are normally distributed, the E-step (for iteration $t = 1, \ldots, T$) comes down to determining the expectation of the complete-data sufficient statistics $\sum_{i=1}^{n} z_{ih}$ and $\sum_{i=1}^{n} z_{ih} z_{ig}$ for $h, g = 1, \ldots, L$ with respect to $f(\boldsymbol{z}_{\text{mis}}|\boldsymbol{z}_{\text{obs}}, \boldsymbol{\xi})$ (Schafer, 1997):

$$\begin{aligned} E(z_{ih}) &= z_{ih} I_{\{r_{ih}=1\}} + z_{ih}^t I_{\{r_{ih}=0\}}, \\ E(z_{ih} z_{ig}) &= z_{ih} z_{ig} I_{\{r_{ih}=r_{ig}=1\}} + z_{ih}^t z_{ig} I_{\{r_{ih}=0, r_{ig}=1\}} + \\ &\quad z_{ih} z_{ig}^t I_{\{r_{ih}=1, r_{ig}=0\}} + (\gamma_{shg}^t + z_{ih}^t z_{ig}^t) I_{\{r_{ih}=r_{ig}=0\}}, \end{aligned} \quad (8.15)$$

with $I_{\{.\}}$ the indicator function which is one if the argument is true and zero otherwise and z_{ih}^t the predicted value of z_{ih} in iteration $t = 1, \ldots, T$ of the EM algorithm. For every missing data pattern s, the variable which has a missing value in pattern s is regressed on the variables which are observed in missing data pattern s. For example, when z_{ih} is missing, z_{ih} is predicted using the regression of $\boldsymbol{z}_h$ on $\{\boldsymbol{z}_g : r_{ig} = 1\}$:

$$z_{ih}^t = \gamma_{sh0}^t + \sum_{\{g:\ r_{ig}=1\}} \gamma_{shg}^t\ z_{ig} \quad \text{for all } i \in I(s), \quad (8.16)$$

where γ_{shg}^t denotes the regression coefficient in pattern s in iteration t of the EM algorithm corresponding to $\boldsymbol{z}_g$, when $\boldsymbol{z}_h$ is the dependent variable. The γ_{shg}^ts can be determined in iteration t from $(\boldsymbol{\mu}^{t-1}, \boldsymbol{\Sigma}^{t-1})$, which are the values of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ determined in the M-step in the previous iteration. These γ_{shg}^ts can be calculated using the sweep operator. We will not elaborate on this, the interested reader is referred to Schafer (1997). In the M-step (for iteration $t = 1, \ldots, T$), $\mu_{z_h}^t$ and $\Sigma_{z_h z_g}^t$ for $h, g = 1, \ldots, L$ are calculated by

$$\begin{aligned} \mu_{z_h}^t &= \frac{\sum_{i=1}^{n} E(z_{ih})}{n}, \\ \Sigma_{z_h z_g}^t &= \frac{\sum_{i=1}^{n} E(z_{ih} z_{ig})}{n} - \mu_h^t \mu_g^t. \end{aligned} \quad (8.17)$$

These steps are iterated until the complete-data likelihood reaches a stationary point.

As can be seen in Equation (8.16), the value of the prediction z_{ih}^t depends on what variables are included in z_{ig}. Which in turn depends on what model is assumed to be the underlying data model. Hence, the choice of the underlying data model influences Equations (8.15), (8.16), and (8.17). Three different types of assumed underlying data models are discussed in the next section.

8.3.2 Three types of assumed underlying data models

Three types of underlying data models will be distinguished: the analytical model, the unconstrained model, and the restricted unconstrained model. When the analytical model is assumed to be the underlying data model, the following model is used

$$y_i = \beta_0 + \sum_{\{j:\ \beta_j \neq 0 \text{ in } H_m\}} \beta_j x_{ij} + \epsilon_i \text{ and } \boldsymbol{x}_i^m \sim \mathcal{N}_{k-1-c_m}(\boldsymbol{\mu}_{x^m}, \boldsymbol{\Sigma}_{x^m x^m}),$$

where $\boldsymbol{\mu}_{x^m}$ and $\boldsymbol{\Sigma}_{x^m x^m}$ denote the parameters corresponding to $\boldsymbol{x}^m$, the predictors in H_m which coefficients are not restricted to zero. In this case, $L = k - c_m$ and the z_{ih}s appearing in Equations (8.15), (8.16), and (8.17) stand for $z_{ih} \in \boldsymbol{z}_i^m = (y_i, \boldsymbol{x}_i^m)$. Let $\boldsymbol{\xi}^m$ be the subvector of $\boldsymbol{\xi}$ with respect to $(\boldsymbol{y}, \boldsymbol{x}^m)$. The estimate of $\boldsymbol{\xi}^m$ is determined with the EM algorithm. When using the analytical model, EM results for Hypothesis H_m in

$$\hat{\boldsymbol{\xi}}_{H_m|\text{An}}^m = (\hat{\boldsymbol{\mu}}_{H_m|\text{An}}, \hat{\boldsymbol{\Sigma}}_{H_m|\text{An}}) = \arg\max_{\boldsymbol{\xi}^m} f(\boldsymbol{y}_{\text{obs}}, \boldsymbol{x}_{\text{obs}}^m | \boldsymbol{\xi}^m),$$

where $\boldsymbol{x}_{\text{obs}}^m$ is the observed part of $\boldsymbol{x}^m$. Analogously to Equations (8.5) and (8.6), the restricted MLEs $\hat{\boldsymbol{\beta}}_{H_m|\text{An}}^m$ and $\hat{\sigma}_{H_m|\text{An}}^2$ can be determined from $\hat{\boldsymbol{\xi}}_{H_m|\text{An}}^m$. Since there are only restrictions on $\boldsymbol{\beta}$, the estimates $\hat{\boldsymbol{\mu}}_{H_m|\text{An},x^m}$ and $\hat{\boldsymbol{\Sigma}}_{H_m|\text{An},x^m x^m}$ follow directly from $\hat{\boldsymbol{\xi}}_{H_m|\text{An}}^m$. Note that both the size and the values of $\hat{\boldsymbol{\mu}}_{H_m|\text{An}}$ and $\hat{\boldsymbol{\Sigma}}_{H_m|\text{An}}$ and, therefore, of $\hat{\boldsymbol{\mu}}_{H_m|\text{An},x^m}$ and $\hat{\boldsymbol{\Sigma}}_{H_m|\text{An},x^m x^m}$ differ per hypothesis. When using the analytical model as assumed underlying data model, the IC of Hypothesis H_m is determined by

$$IC_m^{\text{An}} = -2\ \log f(\boldsymbol{z}_{\text{obs}}^m | \hat{\boldsymbol{\xi}}_{H_m|\text{An}}^m) + 2\ p_m, \text{ with} \tag{8.18}$$
$$f(\boldsymbol{z}_{\text{obs}}^m | \hat{\boldsymbol{\xi}}_{H_m|\text{An}}^m) = f(\boldsymbol{y}_{\text{obs}} | \boldsymbol{x}_{\text{obs}}^m, \hat{\boldsymbol{\beta}}_{H_m|\text{An}}^m, \hat{\sigma}_{H_m|\text{An}}^2) f(\boldsymbol{x}_{\text{obs}}^m | \hat{\boldsymbol{\mu}}_{H_m|\text{An},x^m}, \hat{\boldsymbol{\Sigma}}_{H_m|\text{An},x^m x^m}).$$

Since $f(\boldsymbol{x}_{\text{obs}}^m | \hat{\boldsymbol{\mu}}_{H_m|\text{An},x^m}, \hat{\boldsymbol{\Sigma}}_{H_m|\text{An},x^m x^m})$ is not constant over all hypotheses, IC_m^{An} cannot be written analogously to Equation (8.11). Hence, when using the analytical model, the IC cannot be reduced to the form (8.11) which seems to be most appropriate in regression models, because it focusses on the conditional distribution of the response $\boldsymbol{y}$ given the predictors $\boldsymbol{x}$. Consequently, the AICC cannot be calculated when using this model.

When the unconstrained model is assumed to be the underlying data model, the following model is used

$$y_i = \beta_0 + \sum_{j=1}^{K-1} \beta_j x_{ij} + \epsilon_i \text{ and } \boldsymbol{x}_i \sim \mathcal{N}_{K-1}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx}).$$

Here, the z_{ih}s appearing in Equations (8.15), (8.16), and (8.17) stand for $z_{ih} \in \boldsymbol{z}_i = (\boldsymbol{y}_i, \boldsymbol{x}_i)$. In case there are additional variables to predict the missing values, these will be included, that is, $\boldsymbol{x} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{k-1}, \boldsymbol{x}_k, \ldots, \boldsymbol{x}_{K-1}]$ and $L = K$. All the predictors of $\boldsymbol{y}$, including those whose coefficient is restricted to zero, and additional variables are used in the prediction of the missing values. The parameters of the additional variables (i.e., $\beta_k, \ldots, \beta_{K-1}$) will be set to zero in H_m, because these variables are not predictors of $\boldsymbol{y}$. For Hypothesis H_m, EM then results in

$$\hat{\boldsymbol{\xi}}_{\text{Unc}} = (\hat{\boldsymbol{\mu}}_{\text{Unc}}, \hat{\boldsymbol{\Sigma}}_{\text{Unc}}) = \arg\max_{\boldsymbol{\xi}} f(\boldsymbol{y}_{\text{obs}}, \boldsymbol{x}_{\text{obs}} | \boldsymbol{\xi}). \tag{8.19}$$

Note that $\hat{\boldsymbol{\xi}}_{\text{Unc}}$ does not depend on Hypothesis H_m. As a consequence, $\hat{\boldsymbol{\xi}}_{\text{Unc}}$ has the same value for each hypothesis $\hat{\boldsymbol{\xi}}_{\text{Unc}}$. When the unconstrained model is used, the parameter estimates are adjusted afterwards such that they are in accordance with H_m. Since the restrictions on $\boldsymbol{\beta}$ are of the form $\beta_j = 0$, $\hat{\boldsymbol{\beta}}^m_{H_m|\text{Unc}}$ and $\hat{\sigma}^2_{H_m|\text{Unc}}$ can be determined, analogously to Equations (8.5) and (8.6), from $\hat{\boldsymbol{\xi}}^m_{\text{Unc}}$, the subvector of $\hat{\boldsymbol{\xi}}_{\text{Unc}}$ corresponding to $\boldsymbol{x}^m$. Because there are only restrictions on $\boldsymbol{\beta}$, the estimates $\hat{\boldsymbol{\mu}}_{\text{Unc},x}$ and $\hat{\boldsymbol{\Sigma}}_{\text{Unc},xx}$ follow directly from $\hat{\boldsymbol{\xi}}_{\text{Unc}}$. Note that $\hat{\boldsymbol{\mu}}_{\text{Unc},x}$ and $\hat{\boldsymbol{\Sigma}}_{\text{Unc},xx}$ do not differ per hypothesis. Let $\hat{\boldsymbol{\xi}}_{H_m|\text{Unc}}$ be the set of $\hat{\boldsymbol{\beta}}^m_{H_m|\text{Unc}}$, $\hat{\sigma}^2_{H_m|\text{Unc}}$, $\hat{\boldsymbol{\mu}}_{\text{Unc},x}$, and $\hat{\boldsymbol{\Sigma}}_{\text{Unc},xx}$. The IC of Hypothesis H_m is, in this case, determined by

$$IC^{\text{Unc}}_m = -2 \ \log f(\boldsymbol{z}_{\text{obs}} | \hat{\boldsymbol{\xi}}_{H_m|\text{Unc}}) + 2 \ p_m, \text{ with} \tag{8.20}$$
$$f(\boldsymbol{z}_{\text{obs}} | \hat{\boldsymbol{\xi}}_{H_m|\text{Unc}}) = f(\boldsymbol{y}_{\text{obs}} | \boldsymbol{x}^m_{\text{obs}}, \hat{\boldsymbol{\beta}}^m_{H_m|\text{Unc}}, \hat{\sigma}^2_{H_m|\text{Unc}}) f(\boldsymbol{x}_{\text{obs}} | \hat{\boldsymbol{\mu}}_{\text{Unc},x}, \hat{\boldsymbol{\Sigma}}_{\text{Unc},xx}).$$

Since $f(\boldsymbol{x}_{\text{obs}} | \hat{\boldsymbol{\mu}}_{\text{Unc},x}, \hat{\boldsymbol{\Sigma}}_{\text{Unc},xx})$ is constant over all hypotheses, IC^{Unc}_m can, analogously to Equation (8.11), also be calculated by

$$IC^{\text{Unc}}_m = -2 \ \log f(\boldsymbol{y}_{\text{obs}} | \boldsymbol{x}^m_{\text{obs}}, \hat{\boldsymbol{\beta}}^m_{H_m|\text{Unc}}, \hat{\sigma}^2_{H_m|\text{Unc}}) + 2 \ p_m. \tag{8.21}$$

Thus, when using the unconstrained model, the IC based on the joint density is proportional to the IC based on the conditional density, like in regression models with completely observed data. Hence, for this model, the IC does reduce to the form (8.11) which seems to be most pertinent in regression models, since it focusses on the conditional distribution of the response $\boldsymbol{y}$ given the predictors $\boldsymbol{x}$.

When the restricted unconstrained model is assumed to be the underlying data model, the following model is used

$$y_i = \beta_0 + \sum_{\{j: \ \beta_j \neq 0 \text{ in } H_m\}} \beta_j x_{ij} + \epsilon_i \text{ and } \boldsymbol{x}_i \sim \mathcal{N}_{K-1}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx}).$$

Like when using the unconstrained model, the z_{ih}s appearing in Equations (8.15), (8.16), and (8.17) stand for $z_{ih} \in \boldsymbol{z}_i = (\boldsymbol{y}_i, \boldsymbol{x}_i)$, where $\boldsymbol{x}$ can contain the additional variables $\boldsymbol{x}_k, \ldots, \boldsymbol{x}_{K-1}$, that is, $L = K$. For this model, EM results in

$$\hat{\boldsymbol{\xi}}_{H_m|\text{Restr}} = (\hat{\boldsymbol{\mu}}_{H_m|\text{Restr}}, \hat{\boldsymbol{\Sigma}}_{H_m|\text{Restr}}) = \arg\max_{\boldsymbol{\xi} \in H_m} f(\boldsymbol{y}_{\text{obs}}, \boldsymbol{x}_{\text{obs}}|\boldsymbol{\xi}), \tag{8.22}$$

for Hypothesis H_m. It should be stressed that an adjustment of the EM procedure described in the previous section is needed. Since

$$f(\boldsymbol{z}_{\text{obs}}|\boldsymbol{\xi}) = f(\boldsymbol{y}_{\text{obs}}|\boldsymbol{x}_{\text{obs}}, \boldsymbol{\beta}, \sigma^2) f(\boldsymbol{x}_{\text{obs}}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx}),$$

maximizing Equation (8.14) comes down to maximizing both $f(\boldsymbol{y}_{\text{obs}}|\boldsymbol{x}_{\text{obs}}, \boldsymbol{\beta}, \sigma^2)$ and $f(\boldsymbol{x}_{\text{obs}}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx})$. In maximizing the observed-data likelihood under H_m in iteration t, the estimates for $\boldsymbol{\beta}_{H_m}$, $\sigma^2_{H_m}$, $\boldsymbol{\mu}_{H_m,x}$, and $\boldsymbol{\Sigma}_{H_m,xx}$ in iteration t are obtained. These parameters can be transformed to $\boldsymbol{\mu}^t_{H_m}$ and $\boldsymbol{\Sigma}^t_{H_m}$. The estimates for $\mu^{t-1}_{H_m,z_h}$ and $\Sigma^{t-1}_{H_m,z_h z_g}$ will be used, as in the previous section, to determine the $\boldsymbol{\gamma}^t_{shg}$s, which are used to determine z^t_{ih}. Let $\hat{\boldsymbol{\xi}}_{H_m|\text{Restr}}$ be the set of the estimates $\hat{\boldsymbol{\beta}}^m_{H_m|\text{Restr}}$, $\hat{\sigma}^2_{H_m|\text{Restr}}$, $\hat{\boldsymbol{\mu}}_{H_m|\text{Restr},x}$, and $\hat{\boldsymbol{\Sigma}}_{H_m|\text{Restr},xx}$. When using the restricted unconstrained model as the assumed underlying data model, the IC of Hypothesis H_m is determined by

$$IC^{\text{Restr}}_m = -2 \ \log f(\boldsymbol{z}_{\text{obs}}|\hat{\boldsymbol{\xi}}_{H_m|\text{Restr}}) + 2 \ p_m. \tag{8.23}$$

Note that the values, but not the size, of $\hat{\boldsymbol{\mu}}_{H_m|\text{Restr}}$ and $\hat{\boldsymbol{\Sigma}}_{H_m|\text{Restr}}$ and, therefore, of $\hat{\boldsymbol{\mu}}_{H_m|\text{Restr},x}$ and $\hat{\boldsymbol{\Sigma}}_{H_m|\text{Restr},xx}$ differ per hypothesis. Therefore, IC^{Restr}_m cannot be written analogously to Equation (8.11). Hence, when using the restricted unconstrained model, the IC cannot be reduced to the form (8.11) which seems to be most relevant in regression models due to its focus on the conditional distribution of the response $\boldsymbol{y}$ given the predictors $\boldsymbol{x}$. As a consequence, the AICC cannot be calculated when using this model.

Note that the three assumed underlying data models result in different predictions for the missing values and, therefore, in a different restricted MLE $\hat{\boldsymbol{\xi}}_{H_m}$. In the next section, we will evaluate the three types of assumed underlying data models.

8.3.3 Evaluation of the analytical model

When the analytical model is assumed to be the underlying data model, the IC is determined by Equation (8.18). Note that, in this case, the IC is based on a different data set, namely $\boldsymbol{z}^m$, for every Hypothesis H_m. Consequently, non-comparable densities $f(.)$ are used to compare different hypotheses. This violates the proper use of ICs, since proper implies using the same data for every hypothesis of interest. Therefore, we will not discuss the analytical model as assumed underlying data model anymore.

We will now elaborate on the difference between using the unconstrained and the restricted unconstrained model as assumed underlying data model.

8.3.4 The unconstrained versus the restricted unconstrained model

To compare the restricted unconstrained model and the unconstrained model, we will compare IC values resulting from both approaches. Here, we will use the

analog of Equation (8.8), since the analog of Equation (8.11) is not defined for the restricted unconstrained model. The maximum value of the observed-data likelihood $f(\boldsymbol{y}_{\text{obs}}, \boldsymbol{x}_{\text{obs}}|\hat{\boldsymbol{\xi}})$ under H_m is per definition attained at $\hat{\boldsymbol{\xi}} = \hat{\boldsymbol{\xi}}_{H_m|\text{Restr}}$ (see Equation (8.22)). Thus, the observed-data likelihood for $\hat{\boldsymbol{\xi}}_{H_m|\text{Unc}}$ (or any other value) will be smaller than or equal to the observed-data likelihood attained at $\hat{\boldsymbol{\xi}}_{H_m|\text{Restr}}$. Notably, the penalty for Hypothesis $H_m : \boldsymbol{C}_m\boldsymbol{\beta} = \boldsymbol{0}$ is p_m for both assumed underlying data models. Therefore, the minimum value of the IC, which is based on $f(\boldsymbol{y}_{\text{obs}}, \boldsymbol{x}_{\text{obs}}|\hat{\boldsymbol{\xi}})$, is also attained at $\hat{\boldsymbol{\xi}} = \hat{\boldsymbol{\xi}}_{H_m|\text{Restr}}$. From this it follows that

$$f(\boldsymbol{y}_{\text{obs}}, \boldsymbol{x}_{\text{obs}}|\hat{\boldsymbol{\xi}}_{H_m|\text{Unc}}) \leq f(\boldsymbol{y}_{\text{obs}}, \boldsymbol{x}_{\text{obs}}|\hat{\boldsymbol{\xi}}_{H_m|\text{Restr}}), \tag{8.24}$$

$$IC_m^{\text{Unc}} \geq IC_m^{\text{Restr}}. \tag{8.25}$$

Let, without loss of generalization, the set of hypotheses consist of three hypotheses

$$\begin{aligned} H_u &: \ \beta_0, \ \beta_1, \ \ldots, \ \beta_{k-1}, \\ H_1 &: \ \text{some } \beta_i\text{s are set to 0}, \\ H_2 &: \ \text{some } \beta_i\text{s are set to 0 } (\neq H_1), \end{aligned}$$

where H_u, the hypotheses with no restrictions on $\boldsymbol{\beta}$, is referred to as the unconstrained hypothesis. When the hypothesis of interest is H_u, both the restricted unconstrained model and the unconstrained model lead to the same assumed underlying data model. Hence, both approaches result in the same IC value (say b) for H_u (Table 8.1). Let the IC value of H_1 and H_2, when using the restricted unconstrained model, be a and d (Table 8.1), respectively. Due to Equation (8.25), the IC value of H_1 and H_2, when using the unconstrained model, is $a + \alpha$ and $d + \delta$, respectively, where $\alpha, \delta \geq 0$ (Table 8.1).

Table 8.1: *The IC values of the three hypotheses (H_u, H_1, and H_2) for the two types of assumed underlying data models (AUDM)*

AUDM:	Unconstrained			Restricted unconstrained		
Hypothesis H_m:	H_u	H_1	H_2	H_u	H_1	H_2
IC_m:	b	$a + \alpha$	$d + \delta$	b	a	d

Note: $\alpha, \delta \geq 0$.

When using the unconstrained model, the assumed underlying data model contains all available reasonable data. Therefore, in case of MAR, it leads to unbiased estimates of the missing values and, therefore, the regression parameters $\boldsymbol{\beta}$. However, other assumed underlying data models, when missing a relevant predictor, will result in biased estimates. Bear in mind that the restricted unconstrained data model depends on the hypothesis of interest. Hence, when the hypothesis of interest is equal to (or embeds) the correct underlying data model, both the unconstrained and the restricted unconstrained model will lead to unbiased estimates. In that case, both types of assumed underlying data models will render the same estimates and IC

values asymptotically. Hence, when H_1, respectively, H_2 (or a model embedded in this hypothesis) is the true underlying data model, α, respectively, δ are asymptotically zero.

When examining the performance of both models, two cases will be distinguished: (i) H_u is the correct underlying data model and (ii) H_1 (or H_2) is the correct model. First we inspect the case where H_u is the correct hypothesis (top panel Table 8.2). When the use of the unconstrained model results in choosing H_u (Scenario 1), it holds true that $b < a+\alpha$ and $b < d+\delta$. This does not *per se* imply $b < a$ and $b < d$, it could also be the case that $a < b < d$ or $d < b < a$, in which case H_1, respectively, H_2 is selected. Hence, employing the restricted unconstrained model does not *per se* result in choosing H_u when the use of the unconstrained model does. On the other hand, when the restricted unconstrained model is used and it selects H_u (Scenario 2), the use of the unconstrained model also results in preferring H_u. Namely, $b < a$ and $b < d$ imply $b < a+\alpha$ and $b < d+\delta$, respectively, for α, $\delta \geq 0$. Thus, when H_u is the correct hypothesis, the unconstrained model outperforms the restricted unconstrained model.

Secondly, we examine the case where a constrained hypothesis, H_1, is the correct data model (bottom panel Table 8.2). It should be stressed that in this case α is asymptotically equal to zero. When H_1 is chosen when employing the unconstrained model (Scenario 3), it holds true that $a + \alpha < b, a + \alpha < d + \delta$. Then, when using the restricted unconstrained model, H_1 is selected if $a < b, a < d$ and H_2 is chosen if $a < b, d < a (< d + \delta)$. In contrast, when H_1 is selected when employing the restricted unconstrained model (Scenario 4), H_1 is also chosen when using the unconstrained model (at least, asymptotically, that is, for large enough observations). Hence, when H_1 is the correct hypothesis, the unconstrained model outperforms the restricted unconstrained model (at least, asymptotically). The analog holds for H_2 or any other constrained hypothesis.

In summary, when employing the restricted unconstrained model results in choosing the correct hypothesis, the use of the unconstrained model does too and, when using the unconstrained model results in preferring the correct hypothesis, the use of the restricted unconstrained model does not *per se*. Namely, when predicting under H_m, the support for H_m increases, while it is not *per se* the correct model. When the proportion of missing values is high enough, this could lead to selecting the wrong model.

Hence, the unconstrained model outperforms the restricted unconstrained model. Moreover, the IC for regression models based on the joint density reduces to an IC based on the conditional density, as when the data are completely observed, solely for the unconstrained model. Therefore, the unconstrained data model performs best.

8.4 The missingness is observed too

Until now, we have examined the log-likelihood of the observed data $\boldsymbol{z}_{\text{obs}}$. In the presence of missing data, the missingness indicator $\boldsymbol{r}$ is observed too. In parameter estimation, $(\boldsymbol{z}_{\text{obs}}, \boldsymbol{r})$ should be the focus of interest instead of $\boldsymbol{z}_{\text{obs}}$, according to Jamshidian (2004) and Little and Rubin (1987). They also state that in many important applications $\boldsymbol{\xi}$ and $\boldsymbol{\phi}$ are disjoint, which implies that

Table 8.2: *Scenarios of preferring the correct hypothesis for the two assumed underlying data models (AUDM)*

H_u is correct			
AUDM:	Unconstrained		Restricted unconstrained
	Scenario 1		
Order of ICs:	$b < a + \alpha, b < d + \delta$	$\not\to$	$b < a, b < d$
Preferred hypothesis:	H_u		could be H_u, or H_1 or H_2
	Scenario 2		
Order of ICs:	$b < a + \alpha, b < d + \delta$	$\leftarrow$	$b < a, b < d$
Preferred hypothesis:	H_u		H_u
H_1 is correct (asymptotically $\alpha = 0$)			
AUDM:	unconstrained		restricted unconstrained
	Scenario 3		
Order of ICs:	$a + \alpha < b, a + \alpha < d + \delta$	$\to$	$a < b, a < d + \delta$ not *per se* $a < d$
Preferred hypothesis:	H_1		could be H_1, but also H_2
	Scenario 4		
Order of ICs:	$a + 0 < b, a + 0 < d + \delta$	$\leftarrow$	$a < b, a < d$
Preferred hypothesis:	H_1		H_1
H_2 is correct (asymptotically $\delta = 0$) – analogously to H_1 is correct			

Note: $a = IC_1^{\text{Restr}}$, $d = IC_2^{\text{Restr}}$, $b = IC_u^{\text{Restr}} = IC_u^{\text{Unc}}$, $a + \alpha = IC_1^{\text{Unc}}$, $d + \delta = IC_2^{\text{Unc}}$, with α, $\delta \geq 0$.

$$f(\boldsymbol{z}_{\text{obs}}, \boldsymbol{r}|\boldsymbol{\xi}, \boldsymbol{\phi}) = f(\boldsymbol{z}_{\text{obs}}|\boldsymbol{\xi}) f(\boldsymbol{r}|\boldsymbol{z}_{\text{obs}}, \boldsymbol{\phi}).$$

Since $f(\boldsymbol{z}_{\text{obs}}|\boldsymbol{\xi})$ and $f(\boldsymbol{r}|\boldsymbol{z}_{\text{obs}}, \boldsymbol{\phi})$ are independent, it does not matter whether one maximizes the log-likelihood of the observed data $\boldsymbol{z}_{\text{obs}}$ or of $(\boldsymbol{z}_{\text{obs}}, r)$, when the aim is the estimation of $\boldsymbol{\xi}$. In model selection, $(\boldsymbol{z}_{\text{obs}}, \boldsymbol{r})$ should also be the focus of interest. Fortunately, as described below, when assuming that the unconstrained model is the true underlying data model the focus of interest is arbitrary, like in parameter estimation.

In case the analytical model is assumed to be the true underlying data model, the IC for Hypothesis $H_m : \boldsymbol{C}_m\boldsymbol{\beta} = \boldsymbol{0}$ should actually be based on $f(\boldsymbol{z}^m_{\text{obs}}, \boldsymbol{r}^m|\boldsymbol{\xi}^m_{H_m|\text{An}}, \boldsymbol{\phi}^m)$ $= f(\boldsymbol{z}^m_{\text{obs}}|\boldsymbol{\xi}^m_{H_m|\text{An}})\ f(\boldsymbol{r}^m|\boldsymbol{z}^m_{\text{obs}}, \boldsymbol{\phi}^m)$. Since $\boldsymbol{r}^m$ and $\boldsymbol{z}^m_{\text{obs}} = (\boldsymbol{y}_{\text{obs}}, \boldsymbol{x}^m_{\text{obs}})$ will differ per hypothesis, $f(\boldsymbol{r}^m|\boldsymbol{z}^m_{\text{obs}}, \boldsymbol{\phi}^m)$ differs per hypothesis. Consequently, the focus of interest is not arbitrary.

In case the unconstrained model is used to be the true underlying data model, the IC for Hypothesis H_m should be based on

$$f(\boldsymbol{z}_{\text{obs}}, \boldsymbol{r}|\boldsymbol{\xi}_{H_m|\text{Unc}}, \boldsymbol{\phi}) = f(\boldsymbol{z}_{\text{obs}}|\boldsymbol{\xi}_{H_m|\text{Unc}}) f(\boldsymbol{r}|\boldsymbol{z}_{\text{obs}}, \boldsymbol{\phi}),$$
$$\propto f(\boldsymbol{z}_{\text{obs}}|\boldsymbol{\xi}_{H_m|\text{Unc}}) \text{ for all } H_m.$$

Note that $\boldsymbol{\xi}_{H_m|\text{Unc}}$ denotes the set of $\boldsymbol{\beta}^m_{H_m|\text{Unc}}$, $\sigma^2_{H_m|\text{Unc}}$, $\boldsymbol{\mu}_{\text{Unc},x}$, and $\boldsymbol{\Sigma}_{\text{Unc},xx}$. In this case, $f(\boldsymbol{r}|\boldsymbol{z}_{\text{obs}}, \boldsymbol{\phi})$ has no hypothesis-dependent components. Therefore, the focus of interest is arbitrary. Analogously, the focus of interest is arbitrary, when using the restricted unconstrained model.

8.5 Illustration

As an example, we use the data reported in Table 7.4 on page 154 of Little and Rubin (1987), see Table 8.3, first printed by Woods, Steiner, and Starke (1932). They examined how "the heat evolved during setting and hardening" of cement depends on its composition. "The compounds comprised in this analysis are tricalcium aluminate [...], tricalcium silicate [...], tertracalcium aluminoferrite [...], and β-dicalcium silicate [...]." The results are analyzed by Equation (8.1) "for the contribution of each percent of each compound to the total heat evolution on the assumption that there exists a linear relationship between the compound composition of a cement and its heat evolution." The dependent variable $\boldsymbol{y}$ is the heat involved in calories per gram of cement. There are four predictors: the amount of tricalcium aluminate ($\boldsymbol{x}_1$), the amount of tricalcium silicate ($\boldsymbol{x}_2$), the amount of tertracalcium alumino ferrite ($\boldsymbol{x}_3$), and the amount of dicalcium silicate ($\boldsymbol{x}_4$). In summary, the data are modeled by $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$ for $i = 1, \ldots, 13$ with $\epsilon_i \sim \mathcal{N}_1(0, \sigma^2)$.

In the data reported in Table 8.3 some predictors have missing values. It should be stressed that the dependent variable may also have missing values. One should predict the missing values based on the other predictors and the dependent variable. To deal with missing data, we employ Equation (8.4), where the estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are assessed by the EM algorithm. When the unconstrained model is assumed to be the true underlying data model, $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are the same for each model. When the restricted unconstrained model is used, $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ differ per model. These equal the estimates under the unconstrained model when evaluating the unconstrained model H_u.

Normally, a researcher has a few competing theories and one should evaluate those (including the unconstrained model to safeguard for weak hypothesis). To give insight into the difference between employing the unconstrained model and the restricted unconstrained model as underlying data model, we will first examine all possible models of the form (8.7). Second, we will demonstrate how one should evaluate a set of hypotheses with the AIC or AICC based on the unconstrained model.

8.5.1 Illustration of the unconstrained versus the restricted unconstrained model

To compare the unconstrained model with the restricted unconstrained model, we will examine all 16 possible models of the form (8.7), depicted in Table 8.4. Bear in

Table 8.3: *Example data*

i	y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	-
8	72.5	1	31	22	-
9	93.1	2	54	18	-
10	115.9	-	-	4	-
11	83.8	-	-	23	-
12	113.3	-	-	9	-
13	109.4	-	-	8	-

Note: The missing values are represented by "-".

mind that the IC based on the joint density does not reduce to an IC of the form (8.11) for the restricted unconstrained model. As a consequence, the AICC which is based on the conditional density cannot be employed when using the restricted unconstrained model. Therefore, we inspect Equation (8.8) with Equation (8.9), that is, the AIC based on the joint density, to examine the difference between the two assumed underlying data models.

Table 8.4 shows, among others, the differences between the AICs for these two types of underlying data models. Here one can see that not only the AIC values differ for most models but also the ordering of the models. When using the restricted unconstrained model, the missing values are predicted under H_m. Then, the fit of H_m increases, as can be seen from the likelihood values in Table 8.4. This induces an increase in the support for H_m, as can be seen from the AIC values in Table 8.4. But, H_m is not *per se* the correct model. As a consequence, using the restricted unconstrained model could lead to selecting the wrong model when the proportion of missing values is high enough.

8.5.2 Illustration of the unconstrained model

A theory could be, for example, based on previous research, that the amount of tricalcium aluminate has no effect on the heat of cement. Another theory could be that the amount of dicalcium silicate does not either. As a safeguard we include the unconstrained model. Namely, when the other two hypotheses are weak, the unconstrained will be selected. Let the hypothesis of interest be

Table 8.4: *The number of distinct parameters p_m, the joint observed-data log-likelihood $f_m(y_{obs}, x_{obs}|\hat{\mu}_{H_m}, \hat{\boldsymbol{\Sigma}}_{H_m})$, the AIC values, and the order of the AIC values for each model of interest for both the unconstrained (Unc) and restricted unconstrained model (Restr)*

			Unc			Restr			Difference	
m	H_m	p_m	$f_m(.)$	AIC	Rank	$f_m(.)$	AIC	Rank	in AICs	in rank
0	$\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$	16	-160.5	353.0	16	-160.5	353.0	16	0.0	0
1	$\beta_1,\ \beta_2 = \beta_3 = \beta_4 = 0$	17	-158.7	351.4	15	-156.7	347.3	13	-4.1	-2
2	$\beta_2,\ \beta_1 = \beta_3 = \beta_4 = 0$	17	-153.9	341.7	10	-153.2	340.3	11	-1.4	1
3	$\beta_3,\ \beta_1 = \beta_2 = \beta_4 = 0$	17	-158.3	350.6	14	-158.3	350.6	15	0.0	1
4	$\beta_4,\ \beta_1 = \beta_2 = \beta_3 = 0$	17	-156.5	347.1	12	-156.2	346.5	12	-0.6	0
5	$\beta_1,\ \beta_2,\ \beta_3 = \beta_4 = 0$	18	-142.9	321.9	3	-141.1	318.2	4	-3.7	1
6	$\beta_1,\ \beta_3,\ \beta_2 = \beta_4 = 0$	18	-156.5	349.0	13	-156.5	349.0	14	0.0	1
7	$\beta_1,\ \beta_4,\ \beta_2 = \beta_3 = 0$	18	-151.5	339.0	9	-146.3	328.5	8	-10.5	-1
8	$\beta_2,\ \beta_3,\ \beta_1 = \beta_4 = 0$	18	-147.6	331.2	6	-147.6	331.2	9	0.0	3
9	$\beta_2,\ \beta_4,\ \beta_1 = \beta_3 = 0$	18	-154.3	344.7	11	-151.9	339.7	10	-4.9	-1
10	$\beta_3,\ \beta_4,\ \beta_1 = \beta_2 = 0$	18	-149.3	334.6	7	-144.5	325.0	6	-9.6	-1
11	$\beta_1 = 0,\ \beta_2,\ \beta_3,\ \beta_4$	19	-144.9	327.8	5	-143.9	325.7	7	-2.0	2
12	$\beta_2 = 0,\ \beta_1,\ \beta_3,\ \beta_4$	19	-148.3	334.6	8	-142.7	323.4	5	-11.2	-3
13	$\beta_3 = 0,\ \beta_1,\ \beta_2,\ \beta_4$	19	-143.0	324.1	4	-141.1	320.2	3	-3.9	-1
14	$\beta_4 = 0,\ \beta_1,\ \beta_2,\ \beta_3$	19	-141.1	320.1	2	-141.1	320.1	2	0.0	0
u	$\beta_1,\ \beta_2,\ \beta_3,\ \beta_4$	20	-132.9	305.9	1	-132.9	305.9	1	0.0	0

$H_{11} : \beta_1 = 0$, that is, $H_{11} : y_i = \beta_0 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$,

$H_9 :\ \beta_1 = \beta_3 = 0$, that is, $H_9 :\ y_i = \beta_0 + \beta_2 x_{i2} + \beta_4 x_{i4} + \epsilon_i$,

$H_u :\ \beta_0, \beta_1, \beta_2, \beta_3, \beta_4$, that is, $H_u :\ \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$.

It should be stressed that the IC based on the joint density does reduce to an IC of the form (8.11) when applying the unconstrained model. Hence, in this case, the AICC can be employed to evaluate the set of hypotheses. As mentioned before, $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are the same for each model when the unconstrained model is assumed to be the true underlying data model. From these estimates, $\hat{\boldsymbol{\beta}}_{H_m}$ and $\hat{\sigma}^2_{H_m}$ are determined, which are the estimates in accordance with the restrictions in H_m; reported in Table 8.5. Based on these values, one can calculate the conditional log-likelihood for $\boldsymbol{y}|\boldsymbol{x}$, denoted by $f_m(\boldsymbol{y}_{\text{obs}}|\boldsymbol{x}_{\text{obs}}, \hat{\boldsymbol{\beta}}_{H_m}, \hat{\sigma}^2_{H_m})$; displayed in Table 8.5. The number of distinct parameters with respect to $\hat{\boldsymbol{\beta}}_{H_m}$ equals $k - c_m$. Note that σ^2 is unknown too. Based on Equations (8.11) and (8.13), the $AICC_m$ can be calculated; these values are depicted in Table 8.5. From the results, it is concluded that H_u is the preferred model.

Table 8.5: *The restricted maximum likelihood estimators of $\boldsymbol{\beta}$ and σ^2, the observed-data log-likelihood $f_m(.)$, the number of distinct parameters $1+k-c_m$, and the AICC values for each model of interest*

	H_1	H_2	H_u
$\hat{\beta}_{H_m,0}$	157.2	5.7	-196.7
$\hat{\beta}_{H_m,1}$	0.0	0.0	4.5
$\hat{\beta}_{H_m,2}$	-0.3	1.5	3.1
$\hat{\beta}_{H_m,3}$	-1.6	0.0	3.6
$\hat{\beta}_{H_m,4}$	-1.0	0.6	2.3
$\hat{\sigma}^2_{H_m}$	7.7	38.1	0.3
$f_m(\boldsymbol{y}_{\text{obs}}\|\boldsymbol{x}_{\text{obs}}, \hat{\boldsymbol{\beta}}_{H_m}, \hat{\sigma}^2_{H_m})$	-37.6	-47.0	-25.6
$1+k-c_m$	5	4	6
$AICC_m$	93.7	107.0	77.2

8.6 Discussion

An important principle has been discussed: In case of missing data in the dependent variable $\boldsymbol{y}$ and/or the predictors $\boldsymbol{x}$ of the regression model, one should carefully choose the assumed underlying data model. The values of the regression parameter estimates and of the employed IC depend on which underlying data model is used. Three choices can be made: the analytical model, the unconstrained model, and the restricted unconstrained model. When using the analytical model as the assumed underlying data model, the ICs for the hypotheses of interest are based on different data sets. In this case, the ICs cannot be compared. Therefore, the analytical model should not be used. When comparing the unconstrained model and the restricted unconstrained model, the performance of the first approach is better than the performance of the second: when assuming that the restricted unconstrained model is the underlying data model results in selecting the correct hypothesis, the same holds true for the unconstrained model and, when employing the unconstrained model results in choosing the correct hypothesis, the use of the restricted unconstrained model does not *per se*. Namely, when predicting under the hypothesis of interest, its support is strengthened and may lead to selecting an incorrect model/hypothesis. Therefore, the unconstrained model should be used as the assumed underlying data model in calculating an IC in the presence of missing data for predictor selection and model selection in regression models.

It should be stressed that these results also apply to structural equation modeling (SEM) models and/or to hypothesis of the form $H_m^G : \boldsymbol{C}_m\boldsymbol{\beta} = \boldsymbol{a}_m$, with $\boldsymbol{C}_m \in \mathbb{R}^{c_m \times k}$, $\text{rank}(\boldsymbol{C}_m) = c_m \leq k$, and $\boldsymbol{a}_m \in \mathbb{R}^{c_m}$. Namely, Equation (8.24) and, therefore, Equation (8.25) still hold true. For restrictions of the type H_m^G, when using the unconstrained model, the EM procedure remains the same, only the calculations of the parameter estimates $\hat{\boldsymbol{\beta}}_{H_m|\text{Unc}}$ and $\hat{\sigma}^2_{H_m|\text{Unc}}$ changes. They should be derived from $\hat{\boldsymbol{\xi}}_{\text{Unc}} = (\hat{\boldsymbol{\mu}}_{\text{Unc}}, \hat{\boldsymbol{\Sigma}}_{\text{Unc}})$ using the method of Lagrange multipliers (see Johnston & DiNardo, 1997):

$$\hat{\boldsymbol{\beta}}_{H_m|\mathrm{Unc}} = \hat{\boldsymbol{\beta}} + (\hat{\boldsymbol{\Sigma}}_{xx}^{+})^{-1}\boldsymbol{C}_m'[\boldsymbol{C}_m(\hat{\boldsymbol{\Sigma}}_{xx}^{+})^{-1}\boldsymbol{C}_m']^{-1}(\boldsymbol{a}_m - \boldsymbol{C}_m\hat{\boldsymbol{\beta}}),$$
$$\hat{\sigma}^2_{H_m|\mathrm{Unc}} = \hat{\sigma}^2 + (\boldsymbol{a}_m - \boldsymbol{C}_m\hat{\boldsymbol{\beta}})'[\boldsymbol{C}_m(\hat{\boldsymbol{\Sigma}}_{xx}^{+})^{-1}\boldsymbol{C}_m']^{-1}(\boldsymbol{a}_m - \boldsymbol{C}_m\hat{\boldsymbol{\beta}}).$$

In conclusion, we strongly recommend the use of an IC based on the unconstrained model, since it outperforms the restricted unconstrained model. Moreover, solely for the unconstrained model, the IC based on the joint density reduces to one appropriate for regression models, that is, the one based on the conditional density. As a consequence, the AICC can only be calculated when using this model. In addition, for this model it holds true that the IC based on the observed data is proportional to the IC based on both the observed data and the missingness indicator.

To deal with missing data in both the dependent variable $\boldsymbol{y}$ and the predictors $\boldsymbol{x}$ of the regression model, one can use *PredictorSelectionInMissingData.exe* (`http://staff.fss.uu.nl/RMKuiper`). This software can only be used for predictor selection in regression models with one dependent variable and renders AIC and AICC values.

CHAPTER 9

Remaining Issues

regarding Information Criteria in the Presence of Missing Data

Kuiper, R. M.

9.1 The Complete-Cases Information Criterion in Case of MCAR

Chapter 8 describes how information criteria (ICs) should be calculated in case of MAR. When the missing data are assumed to be MCAR, that is, $f_R(\boldsymbol{r}|\boldsymbol{z}_{obs}, \boldsymbol{z}_{mis}, \boldsymbol{\phi}) = f_R(\boldsymbol{r}|\boldsymbol{\phi})$, one could calculate an IC based on the completely observed cases, since these cases are representable for all the cases.

Let, without loss of generalization, the first n_{obs} observations be completely observed, $\boldsymbol{z}_{CC} = (\boldsymbol{y}_{CC}, \boldsymbol{x}_{CC})$ denote the set of complete cases in $\boldsymbol{z}$, and $\hat{\boldsymbol{\xi}}_{CC} = (\hat{\boldsymbol{\mu}}_{CC}, \hat{\boldsymbol{\Sigma}}_{CC})$ the maximum likelihood estimator (MLE) of $\boldsymbol{\xi}$ for the complete cases. Due to MCAR, it asymptotically yields that $(\hat{\boldsymbol{\mu}}_{CC}, \hat{\boldsymbol{\Sigma}}_{CC}) = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ and $f_Z(\boldsymbol{z}_i|\hat{\boldsymbol{\mu}}_{CC}, \hat{\boldsymbol{\Sigma}}_{CC}) = f_Z(\boldsymbol{z}_i|\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ for all $\boldsymbol{z}_i$, where $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ is the MLE of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Moreover, since

$$f_{Z_{CC}}(\boldsymbol{z}_{CC}|\hat{\boldsymbol{\mu}}_{CC}, \hat{\boldsymbol{\Sigma}}_{CC}) = \prod_{i=1}^{n_{obs}} f_Z(\boldsymbol{z}_i|\hat{\boldsymbol{\mu}}_{CC}, \hat{\boldsymbol{\Sigma}}_{CC}),$$

the likelihood of the complete cases is asymptotically equal to $\frac{n_{obs}}{n}$ times the likelihood of the complete data (as if all the data were observed).

Because the complete-cases IC (IC^{CC}) is easy to determine, we recommend the use of it in case of MCAR:

$$IC^{CC} = -2 \ \log \ f_{Z_{CC}}(\boldsymbol{z}_{CC}|\hat{\boldsymbol{\xi}}_{CC}) \ + \ 2 \ p_m,$$

where p_m is the penalty, which equals the number of distinct parameters in case of the AIC. In other MAR cases, the IC described in Chapter 8 should be used.

9.2 AIC in Mplus and AMOS

This section evaluates how today's software handles missing data in model selection. Horton and Lipsitz (2001) discuss how several software programs (like SAS and S-plus) handle missing data in regression models. Since it is assumed that both $\boldsymbol{y}$ and $\boldsymbol{x}$ have a distribution, two structural equation software programs are examined: Mplus (Muthén & Muthén, 2007) and AMOS (Arbuckle, 2007). Mplus can use EM (as discussed in Little & Rubin, 1987) to handle missing data and AMOS employs FIML, where the observed-data log-likelihood is maximized.

The following example is used to demonstrate the calculation of the AIC in Mplus and AMOS. We have generated data $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2]$ from a bivariate normal distribution and an error term $\boldsymbol{\epsilon}$ from a standard normal distribution: $\boldsymbol{x}_i' \sim \mathcal{N}_2([2\ 3], \boldsymbol{I}_2)$ and $\epsilon_i \sim \mathcal{N}(0,1)$, for $i = 1, \ldots, 100$. We created $\boldsymbol{y}$ by employing $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$, with $\beta_0 = -16$, $\beta_1 = 1$, and $\beta_2 = 5$. Then we created 30% missing values, by means of the procedure described next. In this example, there are 8 missing data patterns (Table 9.1). First, we assign each row in the data to a missing data pattern with use of "the assigning probability" P_s (Table 9.1) when applying the following rule: row i, that is, $[y_i, x_{i1}, x_{i2}]$, belongs to missing data pattern s', when $U_{i1} < \sum_{s=1}^{s'} P_s$, where U_{i1} is a random variable that is uniformly distributed on the interval [0,1]. Let $\boldsymbol{r}_i^*$ denote the assigned missing data pattern for row i, that is, when $r_{ij}^* = 0$, z_{ij} is missing in the assigned missing data pattern. For example, when row i belongs to missing data pattern $s' = 4$, $\boldsymbol{r}_i^* = [r_{i1}^*, r_{i2}^*, r_{i3}^*] = [1, 1, 0]$. Subsequently, it is determined whether row i actually has missing values (in accordance with missing data patterns s'). This probability of having missing values depends on the weighted sum of the values of the observed variables in pattern s'. In this example, $y_i = z_{i1}$ has a weight of $w_1 = \frac{1}{10}$, $x_{i1} = z_{i2}$ of $w_2 = 1$, and $x_{i2} = z_{i3}$ of $w_3 = 1000$, when they are observed in pattern s' (i.e., $r_{ij}^* = 1$ for $j = 1, 2, 3$). Now, $[y_i, x_{i1}, x_{i2}]$ has missing values (in accordance with missing data patterns s'), when $U_{i2} < \phi((\sum_{j=1}^{3} w_j\ r_{ij}\ z_{ij} / \sum_{j=1}^{3} w_j\ r_{ij}) + c)$, where U_{i2} is a random variable that is uniformly distributed on the interval [0,1], ϕ is the cumulative standard normal distribution function, and c is a parameter that controls the expected proportion of missing data, with $c = \sqrt{2}\phi^{-1}(\pi)$. To generate 30% missing values (with the assigning probabilities of Table 9.1), we set π equal to $\pi = .54$.

Table 9.1: *Missing Data Patterns (0 = missing, 1 = observed) and the Assigning Probability (P_s)*

s	y_i	x_{i1}	x_{i2}	P_s
1	0	0	0	0
2	0	1	0	$\frac{2}{9}$
3	1	0	0	$\frac{2}{9}$
4	1	1	0	$\frac{1}{9}$
5	0	0	1	$\frac{2}{9}$
6	0	1	1	$\frac{1}{9}$
7	1	0	1	$\frac{1}{9}$
8	1	1	1	0

Let the hypotheses of interest be

$$H_1 : \beta_0, \beta_1, \beta_2 = 0 \text{ or, stated otherwise, } H_1 : y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i,$$
$$H_u : \beta_0, \beta_1, \beta_2 \quad \text{ or, stated otherwise, } H_u : y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i.$$

From the literature, it is not clear what type of assumed underlying data model is employed in Mplus or AMOS. Therefore, we have constructed a software program, called "*PredictorSelectionInMissingData.exe*", where the AIC is calculated for predictor selection in regression models when using the unconstrained model as the assumed underlying data model. We have made the software in Fortran 90, therefore, we refer to our software in the tables as "Fortran".

In Table 9.2, the estimates of $\boldsymbol{\xi}_{Unc}$ can be found for our software. Since the unconstrained model is employed, $\hat{\boldsymbol{\xi}}_{Unc}$ is the same for H_1 and H_u. From $\hat{\boldsymbol{\xi}}_{Unc}$, the estimates of $\boldsymbol{\beta}_{H_m|Unc}$ and $\sigma^2_{H_m|Unc}$ for both H_1 and H_u can be determined (Table 9.3). Furthermore, the observed-data log-likelihood and AIC values are reported for both the model of $\boldsymbol{y}|\boldsymbol{x}$ and the model of $(\boldsymbol{y}, \boldsymbol{x})$. The results with respect to $(\boldsymbol{y}, \boldsymbol{x})$ are displayed in Table 9.4. Furthermore, this software uses the number of distinct $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ values as penalty for the model with respect to $(\boldsymbol{y}, \boldsymbol{x})$ (see Table 9.4).

When employing Mplus, one should be aware of which variables you use as input and what hypothesis/model constraints you declare (if any). For example, in Mplus, the analytical model can be employed as the assumed underlying data model for H_1 by specifying:

```
VARIABLE: NAMES ARE y x1;
MODEL:    y ON    x1(b1)
```

But, the analytical model should not be used. It is possible to specify H_1 differently in Mplus:

```
VARIABLE: NAMES ARE y x1 x2;
MODEL:    y ON x1(b1)
```

Table 9.2: *The estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in our Fortran software, Mplus, and AMOS for H_1 and H_u*

		Mplus & Fortran			AMOS					
		H_1 & H_u			H_1			H_u		
		$\boldsymbol{y}$	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{y}$	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{y}$	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$
$\hat{\boldsymbol{\mu}}$		1.911	2.011	3.212	-	2.082	3.207	-	2.011	3.212
$\hat{\boldsymbol{\Sigma}}$	$\boldsymbol{y}$	29.911			-			-		
	$\boldsymbol{x}_1$	1.645	0.859		-	0.915		-	0.859	
	$\boldsymbol{x}_2$	5.965	0.191	1.256	-	0.315	1.256	-	0.191	1.256

Note. In AMOS, the estimated values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ with respect to $\boldsymbol{y}$ are not given. with respect to $\boldsymbol{y}$ are not given.

Table 9.3: *The restricted MLEs of $\boldsymbol{\beta}$ and σ^2 in our Fortran software, Mplus, and AMOS for H_1 and H_u*

	H_1			H_u
	Fortran	Mplus	AMOS	
$\hat{\beta}_{H_m,0}$	-1.937	-3.394	-3.627	-14.699
$\hat{\beta}_{H_m,1}$	-1.914	2.238	2.331	0.888
$\hat{\beta}_{H_m,2}$	0.000	0.000	0.000	4.615
$\hat{\sigma}^2_{H_m}$	26.764	27.039	26.490	0.922

Table 9.4: *The observed-data log-likelihood $f_m(.)$, penalty p_m, and AIC_m values for our Fortran software, Mplus, and AMOS for H_1 and H_u*

	H_1			H_u		
	Fortran	Mplus	AMOS	Fortran	Mplus	AMOS
$f_m(.)$	-404.739	-403.365	$\frac{159.661}{2}$	-322.242	-322.242	0
p_m	8	3	8	9	4	9
AIC_m	825.478	812.731	175.661	662.484	652.484	18

Note. AMOS does not give the observed-data log-likelihood but two times the discrepancy of the observed-data log-likelihood with respect to the observed-data log-likelihood of H_u.

```
                x2(b2);
MODEL CONSTRAINT: b2 = 0;
```

To inspect H_u, the same code can be employed, but the last line should be deleted. Mplus gives the same $\hat{\boldsymbol{\xi}}$ as our Fortran software (Table 9.2). This seems to imply that

the unconstrained model is used as the underlying data model. However, as can be seen from Table 9.3, the restricted MLEs of $\boldsymbol{\beta}$ and σ^2 given by Mplus differ from the restricted MLEs given by our Fortran software (which can also be determined by hand). This implies that Mplus does not employ the unconstrained model. Table 9.4 displays the observed-data log-likelihood, the penalty, and the AIC values reported by Mplus, where Mplus uses the number of distinct $\boldsymbol{\beta}$ and σ^2 values as penalty.

In AMOS, both the analytical and the restricted unconstrained model can be employed as the assumed underlying data model, but not the unconstrained model. Which one is used depends on how a model is declared in AMOS (see Figure 9.1). Figure 9.1a represents the case where H_1, that is, $\beta_2 = 0$, is inspected and the analytical model is employed. However, the analytical model should not be used. To evaluate hypotheses employing the restricted unconstrained model, one should include all available variables (i.e., $(\boldsymbol{y}, \boldsymbol{x})$), add the relations between $\boldsymbol{y}$ and $\boldsymbol{x}^m$, and add all the possible covariances between all the predictors as in Figure 9.1b. The resulting parameter estimate value $\hat{\boldsymbol{\xi}}^{Restr}$ for both H_1 and H_u are displayed in Table 9.2. As expected when using the restricted unconstrained model, the parameter estimate values differ per hypothesis. The restricted MLEs of $\boldsymbol{\beta}$ and σ^2 are displayed in Table 9.3 for each hypothesis. Furthermore, AMOS gives two times the difference between the observed-data log-likelihood of the hypothesis of interest and the observed-data log-likelihood of the unconstrained hypothesis. The AIC is also based on this discrepancy instead of the observed-data log-likelihood itself. Table 9.4 shows the discrepancy of the observed-data log-likelihood and of the AIC based on this discrepancy and the penalty, where AMOS employs the number of distinct $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ values as penalty.

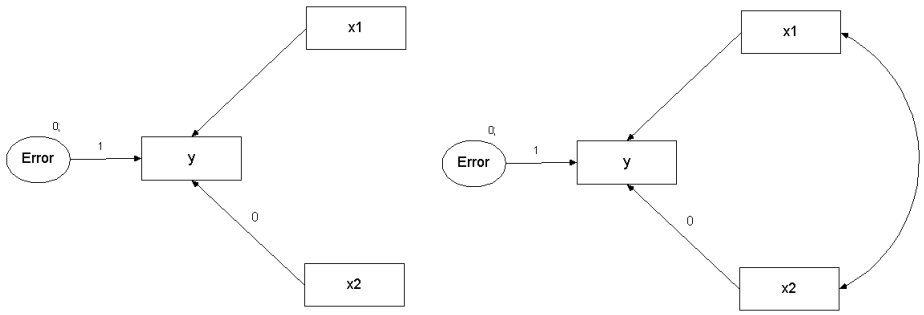


Fig. 9.1: specification of H_1 in AMOS, (a) when using the analytical model as assumed underlying data model and (b) when using the restricted unconstrained model as assumed underlying data model.

In Table 9.3, it can be seen that the parameter estimates of AMOS resemble the parameter estimates of Mplus. Mplus renders results which are more in agreement with the results of AMOS than with the results of our Fortran software. Therefore,

it seems that Mplus employs the restricted unconstrained model as the underlying data model. Note that the results of H_u are the same for the unconstrained model (used in our own software) and the restricted unconstrained model (used in Mplus and AMOS).

Since both AMOS and Mplus do not (yet) employ the unconstrained data model, we recommend to use *PredictorSelectionInMissingData.exe*. However, it should be noted that both Mplus and AMOS could easily make this option available in their software. Our software is available free of use, but can only be used for predictor selection for a regression model with one dependent variable. It calculates the AIC for both the models $\boldsymbol{y}|\boldsymbol{x}$ and $(\boldsymbol{y}, \boldsymbol{x})$ and the AICC for the model $\boldsymbol{y}|\boldsymbol{x}$, when employing the unconstrained model as assumed underlying data model. The software *PredictorSelectionInMissingData.exe* can be found at `http://staff.fss.uu.nl/RMKuiper`.

9.3 Penalty Part

In Chapter 8, we used the conventional penalty term (p_m); for the AIC this equals the number of distinct parameters. But, this is not the correct expression, when minimizing the Kullback-Leibler discrepancy (Kullback & Leibler, 1951) and employing the observed-likelihood.

Cavanaugh and Shumway (1998) derived the AIC from the Kullback-Leibler distance in the presence of missing data for several kind of models, like regression and ANOVA models. However, it is derived for the case where the analytical or the restricted unconstrained is used as the assumed underlying data model. For the unconstrained data model, another penalty has to be derived.

Moreover, this penalty is (most probably) a function of the parameter estimates. It should be stressed that these differ per assumed underlying data model and that, therefore, the penalty values for H_m do as well. Since the comparison of the restricted unconstrained and unconstrained model is based on having the same penalty for each model, one should compare the restricted unconstrained and unconstrained model again. This is beyond the scope of this dissertation. Hence, more research is required.

9.4 Confirmatory Model Selection in Presence of Missing Data

In Chapter 8, we showed that, in the presence of missing data, one should carefully choose the assumed underlying data model, since the values of the regression parameter estimates depend on it. There, the preferred assumed underlying data model is the unconstrained model, that is, the most complex model. In that chapter, we examined model selection without order restrictions. But, what should be done in constrained model selection?

As the other information criteria, the generalized order-restricted information criterion (GORIC) is based on a likelihood and penalty part. Because of the order-restricted hypotheses, the likelihood of the GORIC uses order-restricted MLEs instead of restricted MLEs and the penalty is calculated differently, as can be seen in

(13.4). The penalty depends, among others, on the unrestricted MLE of the covariance matrix with respect to the dependent variables (denoted by $\hat{\Sigma}$ in Equations (5.25) and (13.6)). It should be stressed that this will differ per type of underlying data model due to the different parameter estimates. Hence, not only the likelihood, but also the penalty of the GORIC varies for the different types of assumed underlying data models. Thus, for constrained model selection, we cannot conclude (yet) that the unconstrained model outperforms the restricted unconstrained model. Therefore, more research is required for calculating the GORIC in presence of missing data. Moreover, the GORIC should be derived from the Kullback-Leibler distance in the presence of missing data for the three types of assumed underlying data models (see Section 9.3).

Nevertheless, the restricted unconstrained model has some disadvantages over the unconstrained model. Firstly, when predicting under the hypothesis of interest, its support is strengthened, which may lead to selecting an incorrect model/hypothesis. Secondly, the likelihood part of an information criterion must now be the joint density and cannot be the conditional density. More details can be found in Chapter 8. Therefore, one can choose to employ the unconstrained model as the assumed underlying data model. In that case, the missing values are implied by the covariance structure of the data and not by a specific model (which differs per hypothesis).

Chapter 8 describes how information criteria like the small-sample corrected version of the Akaike information criterion should be calculated in univariate regression models when the dependent variable and/or the predictors contain missing values. This section extends this to the GORIC which can be applied to hypotheses with order restrictions in multivariate regression models. The multivariate regression model with t dependent variables is described in Section 13.2.1. The derivation of the expression is analogously to Chapter 8. One of the differences is (evidently) that sizes of most the vectors and matrices change. For example, $\boldsymbol{z}$ and the missingness indicator are now $n \times (t + k - 1)$ matrices and $L = t + K - 1$. Another difference is (evidently) the form of the hypotheses. The form and more details are given in Chapter 5 or 13. The expression for the GORIC and its penalty term can also be found there, as well as the calculation of the order-restricted maximum likelihood estimators $\tilde{\boldsymbol{B}}_m$ and $\tilde{\boldsymbol{S}}_m$ (when there is no missing data present). Note that in multivariate regression an iteration process is required to obtain the latter (see for example Section 13.3). Analogously to Chapter 8, in presence of missing data, the GORIC can be calculated by the observed-data likelihood ($f(\boldsymbol{z}_{obs}|\tilde{\boldsymbol{\xi}}_m)$) and a penalty term. Also here, the restrictions come down to restricting the corresponding part in $\boldsymbol{\Sigma}_{yx}$ and not in $\boldsymbol{\Sigma}_{xx}$ or $\hat{\boldsymbol{\mu}}_x$. Note that $f(\boldsymbol{x}_{obs}|\hat{\boldsymbol{\mu}}_x, \hat{\boldsymbol{\Sigma}}_{xx})$ is constant over all hypotheses, when using the unconstrained model. Thus, the GORIC (in presence of missing data and when using the unconstrained model) can be written as a likelihood part where the dependent variables are conditional upon the predictors, as when there are no missing data present, and a penalty part analogously to, for example, Section 13.4:

$$GORIC_m^{Unc} = -2 \ \log f(\boldsymbol{y}_{obs}|\boldsymbol{x}_{obs}, \tilde{\boldsymbol{\beta}}_m, \tilde{\boldsymbol{S}}_m) + 2 \ p_m^{Unc}, \tag{9.1}$$

with

$$f(\boldsymbol{y}_{obs}|\boldsymbol{x}_{obs},\tilde{\boldsymbol{\beta}}_m,\tilde{\boldsymbol{S}}_m) =$$
$$\prod_{s=1}^{S}\prod_{i\in I(s)}\frac{1}{(2\pi)^{L_s/2}|\tilde{\boldsymbol{S}}_m^s|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\boldsymbol{y}_i^s-(\tilde{\boldsymbol{B}}_m^s)'\boldsymbol{x}_i^s)'\left(\tilde{\boldsymbol{S}}_m^s\right)^{-1}(\boldsymbol{y}_i^s-(\tilde{\boldsymbol{B}}_m^s)'\boldsymbol{x}_i^s)\right\},$$

$$p_m^{Unc} = 1+\sum_{l=1}^{tk} w_l(tk,\boldsymbol{W}^{Unc},\mathcal{C}_m)\ l,$$
$$\boldsymbol{W}^{Unc} = \hat{\boldsymbol{S}}\otimes(\hat{\boldsymbol{\Sigma}}_{xx}^{+})^{-1}.$$

where $\tilde{\boldsymbol{S}}_m^s$, $\boldsymbol{y}_i^s$, $\tilde{\boldsymbol{B}}_m^s$, and $\boldsymbol{x}_i^s$ are the submatrices of $\tilde{\boldsymbol{S}}_m$, $\boldsymbol{y}_i$, $\tilde{\boldsymbol{B}}_m$, and $\boldsymbol{x}_i$, respectively, corresponding to the observed variables in pattern s, L_s is the number of observed dependent variables in pattern s (since we examine $\boldsymbol{y}_{obs}|\boldsymbol{x}_{obs}$ in lieu of $\boldsymbol{z}_{obs} = (\boldsymbol{y}_{obs},\boldsymbol{x}_{obs})$), and $\hat{\boldsymbol{S}}$ and $(\hat{\boldsymbol{\Sigma}}_{xx}^{+})^{-1}$ are given below.

The observed-data likelihood $f(\boldsymbol{y}_{obs}|\boldsymbol{x}_{obs},\tilde{\boldsymbol{\beta}}_m,\tilde{\boldsymbol{S}}_m)$ is obtained in three steps:

1. The (unrestricted / unconstrained) MLEs of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ (i.e., $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$, respectively) are obtained via EM, where the unconstrained model is employed as assumed underlying data model, see Chapter 8.
2. From these estimates, the (unrestricted / unconstrained) MLEs of $\boldsymbol{B}$ and $\boldsymbol{S}$ are calculated:

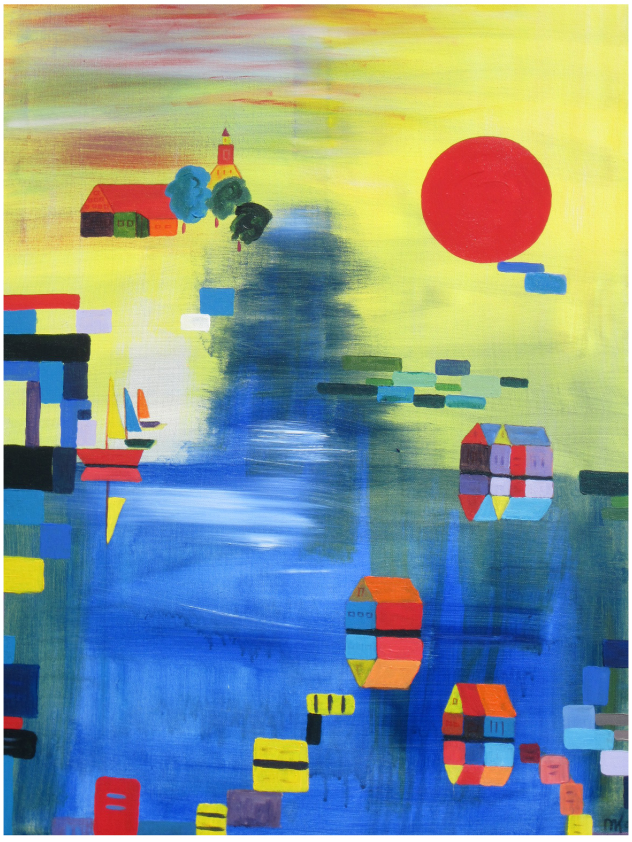
$$\hat{\boldsymbol{B}} = (\hat{\boldsymbol{\Sigma}}_{xx}^{+})^{-1}\hat{\boldsymbol{\Sigma}}_{yx}^{+\prime},$$
$$= \left[\begin{array}{c|c}\frac{1}{\hat{\boldsymbol{\mu}}_x} & \hat{\boldsymbol{\mu}}_x' \\ \hline \hat{\boldsymbol{\mu}}_x & \hat{\boldsymbol{\Sigma}}_{xx}+\hat{\boldsymbol{\mu}}_x\hat{\boldsymbol{\mu}}_x'\end{array}\right]^{-1}\left[\hat{\boldsymbol{\mu}}_y\middle|\hat{\boldsymbol{\Sigma}}_{yx}+\hat{\boldsymbol{\mu}}_y\hat{\boldsymbol{\mu}}_x'\right]',$$
$$\hat{\boldsymbol{S}} = \hat{\boldsymbol{\Sigma}}_{yy}+\hat{\boldsymbol{\mu}}_y\hat{\boldsymbol{\mu}}_y'-\hat{\boldsymbol{\Sigma}}_{yx}^{+}\hat{\boldsymbol{B}}_m.$$

3. These estimates are used in the iteration process in Section 13.3 to determine the constrained / restricted MLEs $\tilde{\boldsymbol{B}}_m^s$ and $\tilde{\boldsymbol{S}}_m^s$. This last step is done for every hypothesis.

Bear in mind that more research is required to conclude which assumed underlying data model (the unconstrained model or the restricted unconstrained model) is the preferred one. Nevertheless, Equation (9.1) presents the expression for the GORIC in the presence of missing data and when using the unconstrained model as assumed underlying data model. Software for the GORIC and the two small-sample corrected versions handling missing data (with use of the unconstrained model) is available from `http://staff.fss.uu.nl/RMKuiper`.

Part IV

Combining Statistical Evidence from Multiple Studies



Blauwe stad by Marga Klungel

CHAPTER *10*

Combining Statistical Evidence from Several Studies:
A Method Using Bayesian Updating and an Example from Research on Trust Problems in Social and Economic Exchange

Kuiper, R. M., Buskens, V., Raub, W., and Hoijtink, H.

Manuscript accepted for publication in Sociological Methods and Research.

The effect of an independent variable on a dependent variable is often evaluated with hypothesis testing. Sometimes, multiple studies are available that test the same hypothesis. In such studies the dependent variable and the main predictors might differ, while they do measure the same theoretical concepts. In this chapter, we present a Bayesian updating method that can be used to quantify the joint evidence in multiple studies regarding the effect of one variable of interest. We apply our method to four studies on how trust in social and economic exchange depends on experience from previous exchange with the same partner. In addition, we examine five hypothetical situations in which the results from the separate studies are less clear cut than in our trust example.

10.1 Introduction

Researchers may have several data sets that can be used to address a research question with respect to the relation between two variables. The main variables of interest may be operationalized in different ways, be measured on different scales, and the statistical model used to relate both variables may differ between studies. This chapter shows how Bayesian updating can be used to summarize the evidence in the data sets for hypotheses on the relation between the two variables. Before introducing our method, we briefly present a case study that we use for illustrative purposes and we sketch the limitations of meta-analysis if studies address the same relation, but employ different variables and models.

Research in economic sociology and social dilemma research on trust problems in exchange relations (see Raub & Buskens, 2008; Buskens & Raub, 2010, for overviews) often studies how trust depends on prior exchange and interactions between the

Table 10.1: Summary of the Four Studies

t Study t	Type of study	Number of observations n	Type of model
1 Batenburg, Raub, and Snijders (2003)	survey	895 transactions	univariate regression
2 Buskens and Raub (2002)	experiment	348 decisions by 40 subjects	univariate regression
3 Buskens and Weesie (2000)	experiment	1249 decisions by 125 subjects	probit regression
4 Buskens, Raub, and van der Veer (2010)	experiment	2160 decisions by 144 subjects	three-level logistic regression
t Study t	y (trust)		scale y
1 Batenburg et al. (2003)	effort invested in management		ratio
2 Buskens and Raub (2002)	effort invested in management		ratio
3 Buskens and Weesie (2000)	choice of vignettes		dummy
4 Buskens et al. (2010)	trustfulness		dummy
t Study t	x_1 (past)		scale x_1
1 Batenburg et al. (2003)	existence relationship with supplier		dummy
2 Buskens and Raub (2002)	type of relationship with supplier		interval
3 Buskens and Weesie (2000)	bought a car from The Autoshop before		dummy
4 Buskens et al. (2010)	number of times a trustee honored trust in the past		ratio
t Study t	some of the other predictors		
1 Batenburg et al. (2003)	transaction characteristics, expected future transactions, network embeddedness*		
2 Buskens and Raub (2002)	transaction characteristics, expected future transactions, network embeddedness*		
3 Buskens and Weesie (2000)	expected future transactions, network embeddedness*		
4 Buskens et al. (2010)	future interactions, network embeddedness*		

* "network embeddedness" means "network of the exchange partners with third parties".

Note. y denotes the dependent variable and x_1 the predictor of interest.

partners. For example, Batenburg, Raub, and Snijders (2003) study the extent to which buyers of IT products trust their sellers using a survey on about 1000 buyers of IT products. One of their hypotheses is that if the buyer has had positive experiences with the seller from transactions in the past, the buyer trusts the seller more in the present transactions. To test this hypothesis, they analyze the effect of the variable "past" (a measure for the amount of positive past experiences) on the dependent variable "lack-of-trust", which is measured by the extent to which the buyer invests in management of the relation such as writing a contract to prevent the seller from untrustworthy behavior. Batenburg et al. (2003) use a linear regression model with additional independent variables to test this hypothesis. Rooks, Raub, Selten, and Tazelaar (2000) and Buskens and Raub (2002) test the same hypothesis with similar variables, using a vignette experiment with hypothetical transactions. In this experiment, purchase managers decide how much time and effort they want to invest to prevent untrustworthy behavior of their seller, while the past experiences of the buyer with the seller are one of the variables describing the hypothetical transactions. They use a linear regression model, too. Two further studies have been used to test the same hypothesis. Buskens and Weesie (2000) use another vignette experiment to test whether past experiences have an effect on trust of students in a second-hand car dealer. Here, trust is measured by the choice between two dealers. Thus, trust is measured as a dichotomous variable. Consequently, Buskens and Weesie use a probit analysis to test the hypothesis. Finally, Buskens, Raub, and van der Veer (2010) test whether past experiences have an effect on trust in a laboratory experiment in which subjects have to decide whether or not to trust another subject. Because subjects play a series of these interactions with the same partner, subjects can make their behavior conditional on the partner's behavior in the past. The choice between trusting or not trusting is again a dichotomous variable and is analyzed via a three-level logistic regression. Table 10.1 provides an overview of the studies.

In each study, the authors are interested in whether past has a positive effect on trust. In null hypothesis testing, one usually tests whether past has no impact on trust versus it has a (positive/negative) effect. Another method to evaluate a positive (or negative) effect is by quantifying evidence for the three effects/hypotheses $H_0 : \beta_1 = 0$, $H_> : \beta_1 > 0$, and/or $H_< : \beta_1 < 0$. Royall (1997) describes how evidence for the hypotheses at hand can be quantified using the likelihood ratio test. In Bayesian model selection, one can also quantify evidence for several hypotheses at hand using the Bayes factor (BF), which is equal to the likelihood ratio test for point hypotheses (like H_0) and can be seen as a generalization of the likelihood ratio test for other hypotheses. The BF gives the relative support for each hypothesis, enabling statements of the type "$H_>$ is ten times as likely as $H_<$".

10.2 Combining Effect Sizes Versus Updating Evidence

To combine multiple studies, one can employ (Bayesian) meta-analysis (among others, Cooper, Hedges, & Valentine, 2009; Lipsey & Wilson, 2000). We briefly discuss meta-analysis and its limitations. Subsequently, we describe our own method. Table 10.2 provides an overview of the differences between the methods.

Table 10.2: Meta-Analysis Versus Bayesian Updating

	meta-analysis	Bayesian updating
Effect size	required	not required
Design	equal across studies	equal or unequal across studies
Main results	estimate of effect size (or parameter) and corresponding p value or confidence interval	evidence; i.e., posterior model probabilities

* Our method uses the parameter estimates and their standard errors of each study, but it does not require that they can be transformed into one effect size, like Cohen's d or R^2.

Meta-analysis can be based on the parameter estimate and its standard error or on a corresponding effect size for each study. When one is interested in the parameter β_1, the hypotheses to be tested are $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$, $H_A : \beta_1 > 0$ or $H_A : \beta_1 < 0$. Meta-analysis results in a parameter estimate or an estimate of effect size based on all studies and a corresponding p value. This estimate is only interpretable when the parameters or effect sizes are comparable. Hence, one cannot use parameters from another type of model. Also, one cannot use effect sizes that cannot be transformed into one type (e.g. the hazard ratio and the odds ratio). In addition, the design of the model should be the same in all studies, that is, the predictors in the model should be the same for all studies. Namely, the parameter estimate or effect size is a conditional one and, therefore, it might change when adding or discarding predictors. Thus, parameter estimates and effect sizes cannot be compared Combining multiple studies with different models can also be based on p values. However, a drawback of this method is, among other things, that p values do not only reflect effect size but also the number of observations (Cooper et al., 2009; Lipsey & Wilson, 2000).

Note that the types and designs of the models employed in our case study differ in various ways (see Table 10.1). Each of the studies tests (among other things) whether there is an effect of the variable past (a measure for the amount of positive past experiences) on trust. In every study, trust and past are measured by different variables. Actually, in the first two studies, the effect of past on lack-of-trust is inspected. Consequently, we multiplied the estimates of the first two studies by minus one. Also, trust is measured on a different scale in each model. Therefore, the studies employ different models. Each model also includes different sets of other predictors. Despite all these differences, the predictor past measures the same concept in all studies. Nevertheless, meta-analysis cannot be employed to combine the four studies regarding trust.

To combine multiple studies of different types and designs, but regarding one theoretical concept, we introduce a Bayesian updating method. In this method, as opposed to meta-analysis, the hypotheses do not address the specific parameter β_1, a parameter that is the same in all studies. Instead, it covers an underlying effect and uses the parameters β_1^t (for $t = 1, \ldots, T$) of the T studies, since they may not be comparable. Nevertheless, they are indicative for the same underlying effect. The method can be employed to evaluate the following hypotheses:

$$\begin{aligned} H_0 &: \text{ no effect,} \\ H_> &: \text{ positive effect,} \\ H_< &: \text{ negative effect.} \end{aligned} \tag{10.1}$$

Notably, our method does not combine estimates but the evidence for a positive ($H_>$), negative ($H_<$), and null effect (H_0) of the predictor of interest (which measures one theoretical concept, say, past) on the dependent variable (which measures one theoretical concept, say, trust).

The input for our method is the estimate of the parameter of interest ($\hat{\beta}_1$) and its standard error ($\hat{\sigma}_{\beta_1}$). This input can be obtained in two ways, from the data or by simply using the values of the parameter estimates reported in the studies. Note that all the necessary information in the data with respect to β is adequately summarized by using $\hat{\beta}_1$ and $\hat{\sigma}_{\beta_1}$. For the four studies that we use as an example, the parameter estimates corresponding to past ($\hat{\beta}_1$) and the standard errors ($\hat{\sigma}_{\beta_1}$) are provided in Table 10.3. Thus, this method does not require an effect size, like Cohen's d, R^2 or odds ratio, or comparable parameter estimates, and different types of models as well as different sets of predictors may be used in each study.

Table 10.3: The Parameter Estimates ($\hat{\beta}_1^t$ and $\hat{\sigma}_{\beta_1}^t$ for Study t) and Corresponding One-sided p values of the Four Studies for Trust

t	Study t	$\hat{\beta}_1^t$	$\hat{\sigma}_{\beta_1}^t$	p
1	Batenburg et al. (2003)	0.090	0.029	.001
2	Buskens and Raub (2002)	0.140	0.054	<.001
3	Buskens and Weesie (2000)	1.090	0.093	<.001
4	Buskens et al. (2010)	1.781	0.179	<.001

It should be stressed that the researcher has to make sure that the β_1^ts do reflect comparable relationships between the two key variables. Although adding a control variable usually affects only the magnitude of β_1^t, it sometimes renders a change in the sign of β_1^t. Hence, one should pay attention to the model specification. For instance, if in one study the relation between trust and past is examined in a regression model and in a second study this relation is examined with a logistic regression and is also modeled as being conditional on various predictors that characterize the network of the exchange partners with third parties, one should be careful that both models do inspect the same theoretical relationship between trust and past. It is up to the researcher to decide whether the β_1^ts reflect the same theoretical relationships.

Figure 10.1 shows how all the parameter estimates of all studies are used for updating the evidence/support for the three hypotheses. We now briefly sketch the updating (a more detailed discussion follows in a subsequent section). First, we assume that all three hypotheses are equally likely and initialize the so-called prior model probabilities (π_m^0) of each hypothesis H_m to be 1/3 for $m \in \{0, >, <\}$; that is, m can take on one of the values in the set $\{0, >, <\}$ comprising the subscripts of H_0, $H_>$, and $H_<$. The prior model probability (PrMP) is a number on a scale of zero to one, which quantifies the weight attached to the current hypothesis. Subsequently, we

start with one study and use its parameter estimate and standard error to calculate or approximate the likelihood. Based on the likelihood, the BF for each hypothesis is determined. Bear in mind that the BF quantifies the support of the data of a pair of hypotheses. We employ BF_{mu}, that is, the BF of H_m (i.e., H_0, $H_<$, or $H_>$) versus a hypothesis without constraints on the parameter of interest. If, for instance, BF_{0u} equals 10, then H_0 has 10 times more support than the hypothesis without constraints. Since the unconstrained hypothesis is not of importance in this chapter, BF_{mu} is just a useful technical tool. Based on these BFs and the initial PrMPs, the PMPs ($\pi^1_{1,m}$) for the three hypotheses ($m \in \{0, >, <\}$) can be assessed, which reflects the evidence/support in the data for the three hypotheses when evaluating solely Study 1. Then, these PMPs are used in the calculation of the PMPs for the evaluation of two studies ($\pi^1_{2,m}$). This process is repeated for all T studies (resulting in $\pi^1_{T,m}$).

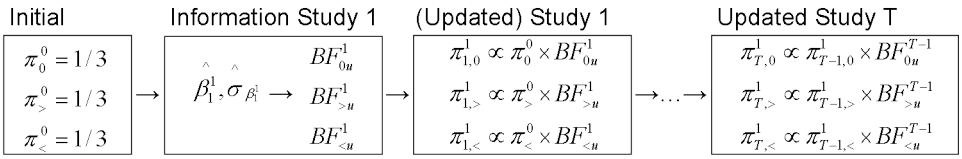


Fig. 10.1: Bayesian updating in case of T studies, where π^0_m represents the prior model probability, $\pi^1_{t,m}$ the posterior model probability after evaluating t studies, $\hat{\beta}^1_1$ and $\hat{\sigma}^1_{\beta_1}$ are the parameter estimates for Study 1, and BF^t_{mu} is the Bayes factor for Hypothesis H_m versus the unconstrained hypothesis in Study t, with $m \in \{0, >, <\}$

In the remainder of this chapter, we firstly elaborate on the concepts likelihood, prior, posterior, BFs, and PMPs. Secondly, we describe our proposed method of using multiple studies to quantify the evidence for H_0, $H_>$, and $H_<$. Thirdly, we illustrate how to apply the method by combining the evidence from the studies on how trust depends on previous experience. Additionally, we inspect five hypothetical situations in which the results from the separate studies are less clear cut than in our example on trust where the results in each study consist of significant positive effects (see the p values in Table 10.3). Moreover, we investigate the sensitivity of the method with respect to the prior distribution that is needed as input for computing the BF and introduce ways to deal with sensitivity.

10.3 Information in the Data

The method proposed here can be applied when different models are used in different studies (e.g., a regression model and a logistic regression model), provided that a function of μ, the expectation of y, can be written as a linear combination of the predictors in all these models (see McCullagh & Nelder, 1989; McCulloch & Searle, 2005)

$$g(\mu_i) = \alpha + \sum_{j=1}^{k} \beta_j x_{ij}, \text{ for } i = 1, \ldots, n \text{ and } j = 1, \ldots, k,$$

where $g(\mu_i)$ is a function of μ_i, α denotes a constant, β_j the parameter that corresponds to x_{ij}, x_{ij} the jth predictor for observation i, n the number of observations, and k the number of predictors. In regression models, $g(\mu_i) = \mu_i$, and in logistic regression models, $g(\mu_i) = \log(\frac{\mu_i}{1-\mu_i})$.

There is one key variable among the k predictors, namely, $x_1 = (x_{11}, \ldots, x_{n1})$. In our example, the key variable is "past". Thus, we are interested in β_1, the parameter corresponding to x_1. Observe that the β_1s of different studies might not be comparable, but they do all reflect the effect of the key variable on the same theoretical concept. For each of the studies, we know or can calculate the estimate $\hat{\beta}_1$ and the standard error $\hat{\sigma}_{\beta_1}$ (see Table 10.3). This enables us to approximate the true likelihood $L^*(.)$ of β_1 by $L(.)$, which is a normal distribution with mean $\hat{\beta}_1$ and variance $\hat{\sigma}^2_{\beta_1}$:

$$L^*(\beta_1|y, x) \approx L(\beta_1|y, x) = \mathcal{N}(\hat{\beta}_1, \hat{\sigma}^2_{\beta_1}). \tag{10.2}$$

More details on large-sample inference and normal approximations can be found in, for example, Chapter 4 of Gelman, Carlin, Stern, and Rubin (2004). Bear in mind that many distributions can be approximated by a normal distribution for a large number of observations and some even for a moderate number of observations. Related to this, a maximum likelihood estimator follows asymptotically a normal distribution (under some regularity conditions and when the conditions for consistency of maximum likelihood estimator are satisfied: Ferguson, 1996). Furthermore, the central limit theorem states that (under some conditions) the mean of a sufficiently large number of independent random variables will be approximately normally distributed (Rice, 1995).

Note that the likelihood quantifies the support in the data for each value of β_1. In the next section, it will be shown how a combination of prior and likelihood renders the posterior distribution, where the prior will be used to quantify the complexity and the posterior the fit of a hypothesis. Hence, as in model selection using information criteria, like the AIC (Akaike, 1973, 1974), Bayesian model selection employs a trade-off between fit and complexity. To simplify notation, the dependence on x is implied throughout below.

10.4 Prior and Posterior

To evaluate the hypotheses of interest in (10.1), one first needs to specify a prior distribution for H_m for $m \in \{0, >, <\}$. We use a so-called conjugate prior (more details can be found in Gelman et al., 2004). This implies that the prior distribution of the parameter β_1 is a normal distribution, since the likelihood of β_1 is approximately a normal distribution, and will result in a normal posterior as discussed below. To determine the prior distribution for H_m ($m \in \{0, >, <\}$), we first need to specify the prior distribution of the parameter β_1 for the case where there are no restrictions. We refer to this as the unconstrained prior, which is defined by

$$p(\beta_1) = \mathcal{N}(\beta_0, \sigma_0^2). \tag{10.3}$$

Subsequently, the prior distribution for H_m ($m \in \{0, >, <\}$) is determined by

$$\begin{aligned} p(\beta_1|H_m) &= p(\beta_1)\ \frac{I_{\beta_1 \in H_m}}{\int_{-\infty}^{\infty} p(\beta_1) I_{\beta_1 \in H_m} d\beta_1} \\ &\propto \mathcal{N}(\beta_0, \sigma_0^2)\ I_{\beta_1 \in H_m}, \end{aligned}$$

where the indicator function $I_{\beta_1 \in H_m}$ equals one if the argument is true, that is, if the parameter value is in accordance with the constraints imposed by H_m, and zero otherwise. Thus, the prior for H_m is proportional to the unconstrained prior when the parameters are in accordance with H_m and otherwise it is zero. Note that the integral in the denominator is a normalizing constant, which is needed to make $p(\beta_1|H_m)$ a density, that is, to let $p(\beta_1|H_m)$ integrate to 1. One needs to specify the parameters of the prior distribution (10.3), that is, β_0 and σ_0^2.

We want the a priori belief in $\beta_1 > 0$ and $\beta_1 < 0$ to be the same. Therefore, we choose $\beta_0 = 0$ such that 50% of the prior is in agreement with $H_<$ and 50% with $H_>$ (more details can be found in Mulder, Hoijtink, & Klugkist, 2010; Jeffreys, 1961; Berger & Mortera, 1999). Bear in mind that this implies that the complexity of both hypotheses is the same, namely 50%. According to the authors mentioned, BFs computed based on such a prior are well calibrated for the selection between $H_<$ and $H_>$. Subsequently, we need to deal with σ_0^2, the variance of the prior. The prior variance should be chosen such that the prior is vague / non-informative, that is, such that it has little influence on the results. But, it should not be too vague, because then H_0 will receive the highest support also when it is not true. This is known as the Lindley paradox (Lindley, 1957). Grounding the prior variance on the data avoids having too vague priors (Berger & Pericchi, 2004, 1996). In our method, a value for σ_0^2 is determined using the confidence intervals of β_1 for all the studies, analogous to the approach proposed in Klugkist, Laudy, and Hoijtink (2005) and Kuiper and Hoijtink (2010). In each of the studies one can compute the 99% confidence interval of β_1. The 99% confidence interval in Study t is

$$\hat{\beta}_1^t \pm 2.576\ \hat{\sigma}_{\beta_1^t},$$

where $\hat{\beta}_1^t$ is the parameter estimate of β_1 in Study t and $\hat{\sigma}_{\beta_1^t}$ the standard error of $\hat{\beta}_1$ in Study t. The 99% prior credibility interval of β_1 is given by

$$0 \pm 2.576\ \sigma_0^t,$$

since $\beta_0^t = 0$ for all t, where β_0^t is the prior parameter estimate of β_1 in Study t and σ_0^t the prior standard error of $\hat{\beta}_1$ in Study t. To let the prior credibility interval include the confidence interval based on the data of one study, the bounds of the prior credibility interval must embed the most extreme bound of the confidence interval of this study. Figure 10.2 depicts the 99% confidence intervals of β_1 for all four studies in our example. Here, one must set $2.576\sigma_0^t$ equal to 0.165, 0.279, 1.330, and 2.242 for $t = 1, 2, 3$, and 4, respectively, which leads to $\sigma_0^{t2} = 0.004, 0.012, 0.266$, and 0.758, respectively. In the section "Example", we will show that this method to determine σ_0^2 has satisfactory properties.

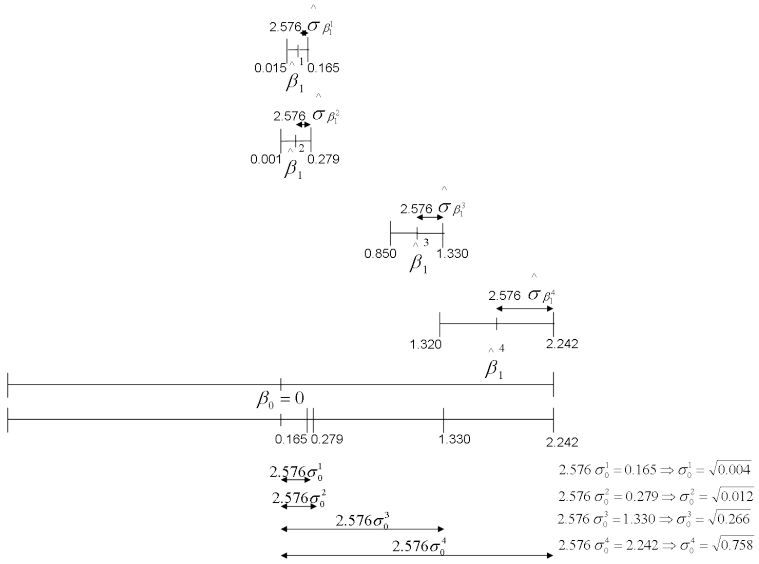


Fig. 10.2: The 99% confidence interval of β_1 for each of the four studies and the four resulting 99% prior credibility interval

The posterior is proportional to the prior times the likelihood. It can be used to quantify the fit of hypotheses. If, for example, the posterior has a mean of 1.5 and a variance of 1, 93% of the posterior is in agreement with $H_>$ and 7% with $H_<$. This implies that the data support $H_>$ more than $H_<$. Stated otherwise, the fit of $H_>$ to the data is better than that of $H_<$. Since both the prior and the likelihood are normal distributions, the posterior is a normal distribution as well (Gelman et al., 2004). To determine the posterior distribution for H_m ($m \in \{0, >, <\}$), we first need to specify the posterior distribution of the parameter β_1 for the case where there are no restrictions. We refer to this as the unconstrained posterior, which is defined by

$$\begin{aligned} p(\beta_1|y) &= L(\beta_1|y, x)p(\beta_1) \\ &\approx \mathcal{N}(\tilde{\beta}, \tilde{\sigma}^2), \end{aligned} \tag{10.4}$$

with

$$\tilde{\beta} = \frac{\frac{1}{\hat{\sigma}^2_{\beta_1}}\hat{\beta}_1 + \frac{1}{\sigma_0^2}\beta_0}{\frac{1}{\hat{\sigma}^2_{\beta_1}} + \frac{1}{\sigma_0^2}},$$

$$\tilde{\sigma}^2 = \frac{1}{\frac{1}{\hat{\sigma}^2_{\beta_1}} + \frac{1}{\sigma_0^2}}.$$

The posterior for H_m ($m \in \{0, >, <\}$) is then given by

$$p(\beta_1|y, H_m) \approx p(\beta_1|y)\ \frac{I_{\beta_1 \in H_m}}{\int_{-\infty}^{\infty} p(\beta_1|y) I_{\beta_1 \in H_m} d\beta_1} \propto \mathcal{N}(\tilde{\beta}, \tilde{\sigma}^2)\ I_{\beta_1 \in H_m}. \tag{10.5}$$

Thus, the posterior for H_m equals the unconstrained posterior when the parameters are in accordance with H_m and otherwise it is zero. Bear in mind that the integral in the denominator is a normalizing constant.

In the next section it will be shown how the prior and posterior distribution can be used to compute BFs and PMPs.

10.5 Bayes Factors and Posterior Model Probabilities

To quantify the evidence for the hypotheses at hand, one can calculate BFs or PMPs. The BF gives the support of a hypothesis relative to another hypothesis. Thus, BF_{mu} quantifies the relative support of H_m with respect to the unconstrained hypothesis. If $BF_{>u}$ equals 1, the data do not support $H_>$ more than the unconstrained hypothesis. If $BF_{>u}$ is larger (smaller) than one, the data support $H_>$ more (less) than the unconstrained hypothesis. Evidently, the more extreme (above or below one) the value of the BF is, the stronger the evidence (for or against, respectively, the hypothesis of interest). The BF is calculated by the ratio of marginal likelihoods of two hypotheses (e.g., Kass & Raftery, 1995; Chib, 1995), where the marginal likelihood for H_m is the normalizing constant in (10.5). From (10.5) it follows that

$$\text{marginal likelihood for } H_m = \frac{L(\beta_1|y)p(\beta_1|H_m)}{p(\beta_1|y, H_m)},$$

for any value of β in agreement with H_m, where $p(\beta_1|H_m)$ and $p(\beta_1|y, H_m)$ are the prior and posterior distribution of the parameter β_1 for hypothesis H_m, respectively, which are described in the previous section. Thus, for H_m and $H_{m'}$ $(m, m' = 0, >, <)$,

$$\begin{aligned} \text{BF}_{mm'} &= \frac{L(\beta_1|y)p(\beta_1|H_m)/p(\beta_1|y, H_m)}{L(\beta_1|y)p(\beta_1|H_{m'})/p(\beta_1|y, H_{m'})} \\ &= \frac{\frac{L(\beta_1|y)p(\beta_1|H_m)/p(\beta_1|y,H_m)}{L(\beta_1|y)p(\beta_1)/p(\beta_1|y)}}{\frac{L(\beta_1|y)p(\beta_1|H_{m'})/p(\beta_1|y,H_{m'})}{L(\beta_1|y)p(\beta_1)/p(\beta_1|y)}} \\ &= \frac{\text{BF}_{mu}}{\text{BF}_{m'u}}, \end{aligned} \tag{10.6}$$

where BF_{mu} and $BF_{m'u}$ are convenient vehicles for the computation of $BF_{mm'}$. As is elaborated in, for example, Klugkist, Laudy, and Hoijtink (2005), BF_{mu} is the BF for a constrained hypothesis (like $H_>$) with respect to the unconstrained hypothesis, that is, the case where there are no restrictions on the parameters. As already mentioned, since the unconstrained hypothesis is not of importance in this chapter, BF_{mu} is just a useful technical tool.

Klugkist, Laudy, and Hoijtink (2005) derive a simple form of BF_{mu} for $m \in \{>, <\}$, namely

$$\mathrm{BF}_{mu} = \frac{f_m}{c_m},$$

where c_m and f_m are the proportions of the unconstrained prior in (10.3) and the unconstrained posterior in (10.4), respectively, in agreement with the constraints of hypothesis H_m for $m \in \{>, <\}$. As elaborated above, since $\beta_0 = 0$, $c_m = 1/2$ for $m \in \{>, <\}$, because half of the prior distribution is in agreement with $H_> : \beta_1 > 0$ and the other half with $H_< : \beta_1 < 0$. If, for instance, the posterior has a mean of 1.5 and a variance of 1, $f_> = .93$.

The Savage-Dickey representation (Dickey, 1971) offers an easy way of calculating BF_{0u} (i.e., $m = 0$), namely

$$BF_{0u} = \frac{p(\beta_1 = 0|y)}{p(\beta_1 = 0)}.$$

One only has to evaluate the unconstrained posterior and prior density at $\beta_1 = 0$ to compute BF_{0u}.

A PMP for hypothesis H_m, π^1_m, gives the relative support for H_m in a finite set of hypotheses (Kass & Raftery, 1995). For $m \in \{0, >, <\}$,

$$\pi^1_m = \frac{\pi^0_m BF_{mu}}{\pi^0_0 BF_{0u} + \pi^0_> BF_{>u} + \pi^0_< BF_{<u}},$$

where π^0_m is the PrMP of hypothesis H_m, which represents the degree of belief of a researcher in each hypothesis before observing the data. An uninformative choice is to set the PrMPs equal for all hypotheses. In the example, π^0_m then equals 1/3 for $m \in \{0, >, <\}$. When $BF_{0u} = 0.5$, $BF_{>u} = 8.5$, and $BF_{<u} = 1$, the PMP values for the three hypotheses (with equal PrMPs) are $\pi^1_0 = \frac{0.5}{0.5+8.5+1} = .05$, $\pi^1_> = .85$, and $\pi^1_< = .10$. Note that it can also be seen from the PMPs, among other things, that the support for $H_>$ is $\frac{.85}{.10} = 8.5$ times higher than the support for $H_<$. Another choice, in case there is previous research, is to set the PrMPs equal to the PMPs of a previous study. We will elaborate on this in the next section, where we explain how one can combine several studies.

10.6 Updating Evidence from Multiple Studies

The results of several studies, that is, the evidence for the hypotheses at hand, can be combined/updated by setting the PrMP of Study t equal to the PMP of Study $t-1$. In the first study, the PrMP will be set equal for all hypotheses (i.e., $\pi^0_m = \frac{1}{3}$ for $m \in \{0, >, <\}$). Let $\pi^0_{t,m}$ and $\pi^1_{t,m}$ represent the PrMPs and PMPs, respectively, for hypothesis H_m in Study t, let BF^t_{mu} be BF_{mu} for Study t, and let T be the total number of studies to combine. Then, for $m \in \{0, >, <\}$,

$$\begin{aligned}
\pi^0_{1,m} &= \frac{1}{3}, \\
\pi^0_{t,m} &= \pi^1_{t-1,m}, \text{ for } t = 2, \ldots, T, \\
\pi^1_{t,m} &= \frac{\pi^0_{t,m} BF^t_{mu}}{\pi^0_{t,0} BF^t_{0u} + \pi^0_{t,1} BF^t_{1u} + \pi^0_{t,2} BF^t_{2u}}, \text{ for } t = 1, \ldots, T.
\end{aligned}$$

When the T studies are combined, this results in an overall PMP for H_m ($\pi^1_{T,m}$). It can be shown for $\pi^0_{1,m} = \frac{1}{3}$ that $\pi^1_{T,m}$ is calculated by

$$\pi^1_{T,m} = \frac{\pi^0_{1,m} \prod_{t=1}^{T} BF^t_{mu}}{\sum_{m' \in \{0,>,<\}} \pi^0_{1,m'} \prod_{t=1}^{T} BF^t_{m'u}} = \frac{\prod_{t=1}^{T} BF^t_{mu}}{\sum_{m' \in \{0,>,<\}} \prod_{t=1}^{T} BF^t_{m'u}}.$$

From this it can be seen that the order of the studies does not influence the outcome of the method.

Having discussed how the method of quantifying evidence for the hypotheses of interest works in general, we illustrate this method in the following section by combining the four studies described in the introduction and summarized in Tables 10.1 and 10.3. In addition, we will study the sensitivity of the method with respect to σ_0^2.

10.7 Example

In the illustration, we use the studies of Batenburg et al. (2003), Buskens and Raub (2002), Buskens and Weesie (2000), and Buskens et al. (2010). Each of these studies tests (among other things) whether there is an effect of the variable past (a measure for the amount of positive past experiences) on trust. As mentioned before, because of the different types and designs of the models (Table 10.1), meta-analysis cannot be employed to combine the four studies regarding trust. But, we can combine these studies via Bayesian updating. Notably, our method does not combine estimates but evidence for null, positive, and negative effects. In all studies, the parameter estimate and the standard error of the coefficient of past is calculated (see Table 10.3). Hence, we can implement our method.

For this example, the optimal values of the prior variance for Studies $t = 1, 2, 3$, and 4 are $\sigma_0^{t2} = 0.004, 0.012, 0.266$, and 0.758, respectively. The panel "σ_0^{t2}" in Table 10.4 shows, for the different steps of the method, the updated PMPs for hypothesis H_m after adding Study t ($\pi^1_{t,m}$) for $m \in \{0, >, <\}$, $t = 1, \ldots, 4$, and the optimal σ_0^{t2} values for Studies 1, 2, 3, and 4. The column for $t = 1$ displays the PMPs for only one study, namely Study 1, and employs equal PrMPs of the three hypotheses. Here, the support for $H_>$ is high and that for H_0 is low. The column for $t = 2$ displays the PMPs for Study 2 and uses the PMPs of $t = 1$ as the PrMPs. Like for $t = 1$, the support for $H_>$ is high. When the third study is added, there is full support for $H_>$ and none for H_0 and $H_<$, namely $\pi^1_{4,>} = 1$ and $\pi^1_{4,0} = \pi^1_{4,<} = 0$. The same remains valid when including the fourth study. From the overall PMP value for $H_>$ ($\pi^1_{4,1}$), it follows that we favor $H_>$ over H_0 and $H_<$. Furthermore, it can be said that support for $H_>$ is $\frac{\pi^1_{4,>}}{\pi^1_{4,m}} \approx \frac{1}{\text{near } 0}$ times higher than the support for H_m for $m \in \{0, <\}$. In the example, the support for $H_>$ is infinitely huge, since in all studies a (large) positive effect of past on trust was found. Note that when one combines studies with mixed effects, that is, not solely positive effects, the value of $\frac{\pi^1_{4,m'}}{\pi^1_{4,m}}$ (for $m', m \in \{0, >, <\}$) is more illustrative.

Table 10.4: $\pi^1_{t,m}$ Values for Hypothesis H_m in Study t for $\frac{1}{2}\sigma_0^{t2}$, σ_0^{t2}, and $2\sigma_0^{t2}$

	$\pi^1_{t,m}$			
m / t	1	2	3	4
		$\frac{1}{2}\sigma_0^{t2}$		
0	0.030	0.003	5.671e-31	1.536e-50
$>$	0.966	0.997	1.000	1.000
$<$	0.004	7.809e-05	0.000	0.000
		σ_0^{t2}		
0	0.022	0.002	6.212e-32	3.659909e-52
$>$	0.976	0.998	1.000	1.000
$<$	0.002	2.418e-05	0.000	0.000
		$2\sigma_0^{t2}$		
0	0.020	0.002	2.787e-32	8.596e-53
$>$	0.978	0.998	1.000	1.000
$<$	0.002	1.139e-05	0.000	0.000

To increase confidence in the conclusions obtained, one could elaborate with a sensitivity study using $\frac{1}{2}\sigma_0^{t2}$ and $2\sigma_0^{t2}$ for Study t. As can be seen in Table 10.4, the results for these values are for all practical purposes the same. The conclusion is that $H_>$ is the preferred hypothesis when combining the four studies and has much more support than H_0 and $H_<$.

In general, the following guidelines will be employed. First, examine the results for σ_0^{t2}. When one of the hypotheses renders the highest overall PMP value, the sensitivity of the results with respect to the prior specification should be checked (see the next step). Otherwise, the studies under investigation cannot distinguish between some or all the hypotheses. Hence, more studies are required. Second, inspect the stability of the results by examining the results for $\frac{1}{2}\sigma_0^{t2}$ and $2\sigma_0^{t2}$. When the results are stable, one can conclude that the hypothesis with the highest PMP value is the preferred one. In case the results are not stable, more studies are needed to draw conclusions. These decision rules will be applied in the next section to situations that are less clear cut than in the example above.

10.8 An Examination of Hypothetical Situations

The example illustrates that combining four studies with persuasive evidence for $H_>$ (namely $p \leq .001$ for all four studies) renders high to full support for $H_>$. Since this is to be expected, we will additionally examine five hypothetical situations, that is, situations based on artificial data. They are depicted in Table 10.5, where p represents the one-sided p value (regarding $H_>$) corresponding to the reported parameter estimates ($\hat{\beta}_1^t$ and $\hat{\sigma}_{\beta_1}^t$ for Study t). Effects for which $\hat{\beta}_1^t$ is larger than zero are called positive effects and those smaller than zero, negative effects. Positive effects with a p value smaller than .05 are called significant positive effects, those with a p value larger than .05 and smaller than .10, small positive effect, and those

Table 10.5: The Parameter Estimates ($\hat{\beta}_1^t$ and $\hat{\sigma}_{\beta_1}^t$ for Study t) and Corresponding One-sided p values for Five Hypothetical Situations

						Mixed effects					
		No effects		Small effects		1		2		3	
t	$\hat{\sigma}_{\beta_1}^t$	$\hat{\beta}_1^t$	p	$\hat{\beta}_1^t$	p	$\hat{\beta}_1^t$	p	$\hat{\beta}_1^t$	p	$\hat{\beta}_1^t$	p
1	0.029	0.007	0.400	0.045	0.060	0.055	0.030	0.068	0.010	0.009	0.378
2	0.054	0.014	0.400	0.084	0.060	-0.102	0.970	-0.084	0.940	-0.084	0.940
3	0.093	0.078	0.400	0.175	0.060	0.153	0.100	0.175	0.060	0.175	0.060
4	0.179	0.151	0.400	0.337	0.060	0.151	0.400	0.337	0.060	0.337	0.060

with a p value larger than .10 and smaller than .50, positive null effects. Negative effects with a p value larger than .95 are referred to as significant negative effects, those with a p value smaller than .95 and larger than .90, small negative effect, and those with a p value smaller than .90 but larger than .50, negative null effects. Five hypothetical situations are distinguished: 1) Positive null effects: all effects, although not significant, are positive, namely, $p = .40$ for each study. 2) Insignificant positive effects: all effects are small and positive with $p = .06$ for each study. 3) - 5) Mixed effects (situations with positive, negative, and null effects). In the first mixed effect situation, there is a significant positive effect ($p = .03$), a significant negative effect ($p = .03$ leading to $p = .97$ for a $H_>$), a small positive effect ($p = .1$), and a positive null effect ($p = .4$). In the second mixed effect situation, there is a significant positive effect (with a lower p value, namely $p = .01$), a small negative effect ($p = .06$ leading to $p = .94$ for $H_>$), and two small positive effects ($p = .06$). The third one resembles the second, the only difference is that the significant positive effect is replaced by a positive null effect with $p = .38$. In all situations, the standard errors of $\hat{\beta}_1^t$ are set equal to the ones of the example. The p values were chosen based on the type of situation, which resulted in the $\hat{\beta}_1^t$ values depicted in Table 10.5.

Table 10.6 displays the overall PMPs for these five situations for σ_0^{t2}, $\frac{1}{2}\sigma_0^{t2}$, and $2\sigma_0^{t2}$. The latter two are inspected to examine the sensitivity of the conclusions obtained due to the choice of the prior. The σ_0^{t2} values are given in Table 10.6.

Table 10.6 shows that combining four studies with a positive null effect lead to support for both H_0 and $H_>$. This is supported by the results for $\frac{1}{2}\sigma_0^{t2}$, but not by that for $2\sigma_0^{t2}$, where H_0 receives most support. Since our results are not supported by the sensitivity analysis, more studies should be collected and added. Moreover, the support for H_0 is 1.27 higher than for $H_>$, which is not compelling evidence.

Combining four studies with a small positive effect (each not significant at $\alpha = 0.05$) renders profound support for $H_>$. The sensitivity analysis shows stability, that is, $H_>$ is preferred over the other two hypotheses for $\frac{1}{2}\sigma_0^{t2}$ and $2\sigma_0^{t2}$ as well. Thus, even though the four studies did not find a significant positive effect, combining them does lead to compelling support for a positive effect, since all four studies comprise small positive effects.

In the first mixed effect situation, where there is a significant positive effect, a significant negative effect, a positive null effect, and a small positive effect, $H_>$ has

the highest support. The sensitivity analysis shows that the support for $H_>$ decreases and that for H_0 increases for increasing prior variances. The support for $H_>$ is about 3, 2, and 1 times higher than for H_0 for $\frac{1}{2}\sigma_0^{t2}, \sigma_0^{t2}$ and $2\sigma_0^{t2}$, respectively. Because of this variability, more studies should be collected and added to be able to draw conclusions.

In the second mixed effect situation, where there is a small negative effect and two small positive effects besides a positive effect, $H_>$ has the highest support for all three σ_0^{t2} values. In this situation, one can conclude that $H_>$ has profoundly more support than H_0 or $H_<$.

In the third mixed effect situation, which equals the second one except that the significant positive effect has a lower p value, $H_>$ renders high(est) support for all the prior variance values. For σ_0^{t2} and $\frac{1}{2}\sigma_0^{t2}$, the support for $H_>$ is 2.45 and 3.37 higher than for H_0, respectively. For $2\sigma_0^{t2}$, this is only 1.26. Nevertheless, we conclude that $H_>$ has highest support and has 2.45 and 50 times more support than H_0 and $H_<$, respectively.

In sum, it evidently depends on the types of effect (positive, negative, null) and their p values which effect / hypothesis receives the highest support. When the results are not sensitive for the prior variance, one can conclude that the support for $H_{m'}$ is $\frac{\pi^1_{4,m'}}{\pi^1_{4,m}}$ as large as for H_m (for $m', m \in \{0, >, <\}$). In other situations, more studies should be collected and added.

Table 10.6: $\pi^1_{4,m}$ Values for Hypothesis H_m for $\frac{1}{2}\sigma_0^{t2}$, σ_0^{t2}, and $2\sigma_0^{t2}$

	No effects	Small effects	Mixed effects 1	Mixed effects 2	Mixed effects 3
σ_0^{12}	0.001	0.002	0.003	0.003	0.001
σ_0^{22}	0.004	0.008	0.009	0.008	0.008
σ_0^{32}	0.015	0.026	0.023	0.026	0.026
σ_0^{42}	0.056	0.096	0.056	0.096	0.096
m	No effects	Small effects	Mixed effects 1	Mixed effects 2	Mixed effects 3
			$\frac{1}{2}\sigma_0^{t2}$		
0	0.427	0.016	0.250	0.038	0.223
$>$	0.524	0.984	0.717	0.960	0.752
$<$	0.050	1.160e-04	0.033	0.000	0.024
			σ_0^{t2}		
0	0.544	0.015	0.320	0.040	0.286
$>$	0.429	0.985	0.661	0.959	0.700
$<$	0.027	3.116e-05	0.019	0.001	0.014
			$2\sigma_0^{t2}$		
0	0.715	0.021	0.482	0.065	0.440
$>$	0.272	0.979	0.507	0.935	0.552
$<$	0.012	1.248e-05	0.010	2.842e-04	0.008

10.9 Conclusion

This chapter introduces a Bayesian updating method to quantify the evidence for the hypotheses of interest (i.e., H_0 : no effect, $H_>$: positive effect, and $H_<$: negative effect) from multiple studies with possibly different types of models and designs. Specifically, one obtains an overall posterior model probability for each hypothesis of interest (i.e., H_0, $H_>$, and $H_<$), which reflects the relative support and allows statements of the type "$H_>$ is ten times more likely than $H_<$".

In terms of the example of the effect of positive past experiences on trust, we see that combining the four studies increases the confidence in the hypothesis that there is indeed a positive effect of past on trust. Although the example is illustrative, the result might not be surprising, given that the hypothesis under consideration is supported in all separate studies. The evaluations of the five hypothetical situations show that also with limited or mixed evidence in individual studies our method helps to distinguish between situations with more and less convincing evidence if different studies are combined.

From the results of the illustration, one can see that our Bayesian updating method is useful for choosing the best of a set of hypotheses in case of multiple studies regarding one theoretical concept. The method is practical because it can be used even when only parameter estimates and standard errors from the different studies are available. It is not necessary that the original data are available and even if the data would be available, this would not lead to better evaluation of the hypotheses because all the information our Bayesian updating needs is incorporated in the parameters and the related standard errors. To facilitate using our Bayesian updating method, we provide software that implements the method in R as described in more detail in the Appendix.

10.A Software

We offer software in R (to be found at `http://staff.fss.uu.nl/RMKuiper`) that implements our method. The software enables the user to combine several studies. The computer package R is open-source software that can be freely downloaded from www.stats.bris.ac.uk/R. To use our software, one needs to make a working directory in which to save the software. The software consists of one .r file and two .txt files. The two .txt files should not be modified.

To employ our software, one should open R and open the .r file in R. This .r file should be modified such that it fits the data of the studies one wants to combine. Running the software renders overall posterior model probabilities for the optimal prior variance (σ_0^{t2}; in the software denoted by sigma^{t2}_0). As described in this chapter, the overall posterior model probabilities for $\frac{1}{2}\sigma_0^{t2}$ and $2\sigma_0^{t2}$ provide insight into the sensitivity of the overall PMPs for σ_0^{t2}.

Part V

Software

Geen zee te hoog by Marga Klungel

CHAPTER *11*

Overview of Software

For every technique mentioned in this dissertation, software is made, which can be found on `http://staff.fss.uu.nl/RMKuiper`. Table 11.1 gives an overview of the software regarding the type of model it can be applied to and the type of restrictions it can analyze. In the next two chapters, Confirmatory ANOVA and GORIC-General are discussed more extensively.

Table 11.1: *Overview of Software*

Software	Model	Restriction
Comparison Of Means *	ANOVA	Exploration: subset of all possible pairs of means based on sample means Confirmation: $\theta_i - \theta_{i'} \geq 0$ (for some $i, i' = 1, \ldots, k$)
Confirmatory ANOVA *	ANOVA	$\theta_i - \theta_{i'} \geq 0$ (for some $i, i' = 1, \ldots, k$)
GORIC-ANOVA	ANOVA	$R\theta \geq r$ (**)
GORIC-General ***	t-variate regression	$R\theta \leq r$ (**)
GORICC ****	t-variate regression	$R\theta \leq r$ (**)
Predictor Selection in Missing Data	regression	$\theta_i = 0$ (for some $i = 1, \ldots, k$)
GORIC(C) handling Missing Data	t-variate regression	$R\theta \leq r$ (**)
Combining Studies	different types *****	$\theta_1 = 0, \theta_1 > 0, \theta_1 < 0$

* with and without interface.

** When $r \neq 0$, R should be of full rank (after discarding redundant restrictions), because the restrictions should be a (relocated) closed convex cone, see Chapter 5.

*** Besides the stand alone program downloadable from the url stated above, there is a goric package available in R (freeware). More information can be found at `http://cran.r-project.org/web/packages/goric`.

**** The GORICC is the small-sample corrected version of the GORIC. Two versions are described in Chapter 6.

***** provided that a function of μ, the expectation of the dependent variable, can be written as a linear combination of the predictors (e.g., a regression model and a logistic regression model), see Chapter 10.

Note. ANOVA = analysis of variance, θ is the parameter of interest, k is the number in groups in ANOVA models and the number of predictors (including the intercept) in other models, R is a $c \times k$ matrix, r a c-vector, and c is the number of restrictions.

CHAPTER *12*

A Fortran 90 Program for Confirmatory Analysis of Variance

Kuiper, R. M., Klugkist, I., and Hoijtink, H.

Published in Journal of Statistical Software, 34(8), pp. 1-31.

There are different confirmatory techniques to compare means, like hypothesis testing and (Bayesian) model selection. However, there is no software package in which these techniques are available. A Fortran 90 program is written, which enables researchers to apply these techniques to their data. Besides traditional hypotheses, like $H_0 : \mu_1 = \mu_2 = \mu_3$ and $H_u : \mu_1,\ \mu_2,\ \mu_3$, order-restricted hypotheses, like $\mu_1 > \mu_2 > \mu_3$ or $\mu_1 > \mu_2 = \mu_3$ or $\mu_1 > \mu_2 < \mu_3$, can be evaluated.

12.1 Introduction

Often researchers have a theory with respect to the ordering of the means in the experiment. These theories can be written as order-restricted hypotheses (e.g., $\mu_1 > \mu_2 > \mu_3$) and can be tested with confirmatory methods. In the context of comparing independent means, that is, analysis of variance (ANOVA), three approaches are distinguished:

- Silvapulle and Sen (2005, pp. 25–42) present the $\bar{F}$ test. There are two types of $\bar{F}$ tests. Namely the *ordered alternative* and the *ordered null*. In the *ordered alternative*, the classical null (H_0) is tested against an order-restricted hypothesis, for example, H_1 in (12.2). In the *ordered null*, an order-restricted hypothesis, like H_1, is tested against the classical alternative (H_u).
- Anraku (1999) introduces the order-restricted information criterion (ORIC). The ORIC can be used to select the best of a set of order-restricted hypotheses, like the hypotheses in (12.2).
- Klugkist, Laudy, and Hoijtink (2005) present a Bayesian model selection (BMS) criterion, which can be used in the same context as the ORIC.

For these approaches, user-friendly software is not available. In this chapter, software is introduced with which the three confirmatory approaches can be executed.

The model used in this chapter and in the software is the ANOVA model:

$$y_{ij} = \mu_i + \epsilon_{ij}, \tag{12.1}$$

where $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, y_{ij} is the jth observation of the dependent variable for Group i, which has n_i observations, μ_i is the mean of Group i, and ϵ_{ij} is an error term. The error terms are independently and normally distributed with expected value 0 and variance σ^2, that is, $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

As an example, consider the simple ANOVA with five groups, presented by Lucas (2003). In Section 12.3, the theoretical background of his research will be elaborated. Lucas expresses clear theories with respect to the ordering of means, leading to the following specific hypotheses:

$$\begin{aligned} H_0 &: \ \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5, \\ H_1 &: \ \mu_5 \ = \mu_3 > \{\mu_1, \mu_4\} > \mu_2, \\ H_2 &: \ \mu_3 > \mu_1 > \mu_4 \ = \mu_5 > \mu_2, \\ H_3 &: \ \mu_1, \ \mu_2, \ \mu_3, \ \mu_4, \ \mu_5. \end{aligned} \tag{12.2}$$

The hypotheses H_0 and H_3 are the classical hypotheses, the other two are order-restricted hypotheses. It is also possible to specify a set of models/hypotheses without the classical null $H_0 : \mu_1 = \ldots = \mu_k$ and/or the classical alternative $H_u : \mu_1, \ldots, \mu_k$. We recommend to include H_u (when doing model selection) as a safeguard for choosing a weak hypotheses (Kuiper & Hoijtink, 2010). Furthermore, one should include H_0 only when there is real interest in H_0.

Although software for exploratory approaches is widely available – e.g., classical hypothesis testing in SPSS (SPSS Inc., 2006) and model selection using information criteria, like AIC, in R with the package nlme (Pinheiro, Bates, DebRoy, Sarkar, & R Development Core Team, 2009) – this is not the case for confirmatory approaches, that is, for evaluation of the four hypotheses in (12.2). The software presented in this chapter can evaluate different types of hypotheses which can be formulated by $A\boldsymbol{\mu} \geq 0$, for some matrix A in which each row is a permutation of the k-vector $(-1, 1, 0, \ldots, 0)$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^\top$.

Note that also factorial ANOVA models fit in the framework presented. For instance, a 2×3-design can be represented by (12.1) with $k = 6$. Specific hypotheses about the six group means can again be formulated by $A\boldsymbol{\mu} \geq 0$ (as explained earlier). In a standard two-way ANOVA, three hypotheses are tested concerning the presence of *a* main effect of the first factor, *a* main effect of the second factor, and the presence of *an* interaction effect. If one or more of these effects are found, further evaluation is required to describe the direction of the effects (making the approach exploratory). In a confirmatory approach, in contrast to a exploratory approach, expected patterns are specified beforehand. For instance, $H : \{\mu_1 = \mu_2 = \mu_3\}, \{\mu_4 < \mu_5 < \mu_6\}$ is a prespecified and specific interaction effect. Competing (interaction) effects can also be specified.

In the next section, subsequently, the $\bar{F}$ test, the ORIC, and Bayesian model selection will be shortly explained. In Section 12.3, a practical example is provided and analyzed using each of the three approaches. The appendix contains a user manual for the software.

12.2 Three confirmatory techniques for comparing means

12.2.1 Hypothesis testing using the F-bar ($\bar{F}$) statistic

In classical statistical testing, the hypothesis *all means are equal* is tested against the alternative *not all means are equal*. This is usually tested with an F test using a one-way ANOVA. However, often researchers want to test a certain order restriction, because of a theory with respect to the order of the means in the experiment. See, for example, H_1 and H_2 in (12.2).

In Silvapulle and Sen (2005, pp. 25–42), the F test is modified such that an order-restricted hypothesis can be tested. This test is called the F-bar ($\bar{F}$) test. It is possible to test the null hypothesis *all means are equal* (H_0) against an *ordered alternative*, like H_1, and it possible to test an *ordered null* (H_1) against the alternative *all parameters are free* (H_u).

The $\bar{F}$ test statistic is calculated by: $\bar{F} = \frac{RSS(H_{null}) - RSS(H_{alt})}{S^2}$, where $RSS(H)$ is the residual sum of squares under hypothesis H and $S^2 = (n_1 + \cdots + n_k - k)^{-1} \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$ is the mean square error, with $n_1 + \cdots + n_k - k$ the error degrees of freedom. This is applied to the two types of test. For each of the two tests, the $RSS(H_{null})$ and $RSS(H_{alt})$ will be elaborated on.

The first test is the *ordered alternative*, in this test $H_0 : \mu_1 = \ldots = \mu_k$ is tested against an order restriction of the form $H_1 : A\boldsymbol{\mu} \geq 0$, for some matrix A in which each row is a permutation of the k-vector $(-1, 1, 0, \ldots, 0)$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^\top$. In this test,

$$RSS(H_{null}) = \sum_i \sum_j (y_{ij} - \bar{y})^2,$$

where $\bar{y}$ is the overall mean, and

$$RSS(H_{alt}) = \sum_i \sum_j (y_{ij} - \tilde{\mu}_i)^2,$$

where

$$\tilde{\boldsymbol{\mu}} = (\tilde{\mu_1}, \ldots, \tilde{\mu_k})^\top = \arg min_{\boldsymbol{\mu} \in H_1} \sum_i \sum_j (y_{ij} - \mu_i)^2.$$

Since $\sum_i \sum_j (y_{ij} - \mu_i)^2$ can be rewritten as

$$\sum_i \sum_j (y_{ij} - \mu_i)^2 = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 + \sum_i n_i (\bar{y}_i - \mu_i)^2 = C(\boldsymbol{y}) + q(\boldsymbol{\mu}),$$

where $\bar{y}_i$ is the ith group/treatment mean, $\boldsymbol{y}$ is the matrix consisting of the elements y_{ij}, and

$$q(\boldsymbol{\mu}) = \sum_i n_i (\bar{y}_i - \mu_i)^2 = (\bar{\boldsymbol{y}} - \boldsymbol{\mu})^\top \text{ diag}\{n_1, \ldots, n_k\} \ (\bar{\boldsymbol{y}} - \boldsymbol{\mu}),$$

it holds that

$$\tilde{\boldsymbol{\mu}} = \arg min_{\boldsymbol{\mu} \in H_1} \ q(\boldsymbol{\mu}). \tag{12.3}$$

This constrained minimization problem, where the objective function $q(\boldsymbol{\mu})$ is quadratic in $\boldsymbol{\mu}$ and the (equality and inequality) constraints are linear in $\boldsymbol{\mu}$, is a *quadratic programming problem*. There are efficient computer algorithms for this minimization problem, here the IMSL subroutine QPROG (Visual Numerics, 2003, pp. 1307–1310) is used in the Fortran 90 program.

The second test is the *ordered null* in which a null of the form $H_1 : A\boldsymbol{\mu} \geq 0$ (as explained earlier) is tested against the alternative *no restrictions on the* μ_i $\forall i$, that is, $H_u : \mu_1, \ldots, \mu_k$. Then,

$$RSS(H_{null}) = \sum_i \sum_j (y_{ij} - \tilde{\mu}_i)^2$$

and

$$RSS(H_{alt}) = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2.$$

As with classical hypothesis testing, p values must be determined. The exact p value, for the $\bar{F}$, can be obtained via simulation. In the ANOVA model, the errors are normally distributed, that is $\epsilon_j \sim \mathcal{N}(0, \sigma^2)$. Therefore, the simulation consists of the following three steps (Silvapulle & Sen, 2005, pp. 32–33 and 40):

1. Generate independent observations z_{ij} ($i = 1, \ldots, k$ and $j = 1, \ldots, n_i$) from the standard normal distribution $\mathcal{N}(0, 1)$.
2. Compute the $\bar{F}$ for the generated data.
3. Repeat the previous two steps R_p times. In the program, the default value of R_p is $R_p = 100,000$. Calculate the number of times the $\bar{F}$ statistic, calculated in Step 2, exceeds the sample value of the $\bar{F}$ statistic, this number is denoted by M. The p value is calculated by M/R_p.

When the p value is smaller than the nominal α-level, often set equal to 0.05, the null hypothesis is rejected. Thus, in the *ordered alternative*, when $p < \alpha$, H_0 is rejected and, in the *ordered null*, when $p < \alpha$, the order-restricted hypothesis is rejected.

12.2.2 Model selection using order-restricted information criterion

Anraku (1999) proposes the order-restricted information-criterion (ORIC). It can be used to select the best of a set ($\mathcal{M}$) of models/hypotheses H_m, $m \in \mathcal{M}$. The set of hypotheses can contain H_0, H_u, and order-restricted hypotheses of the form $A\boldsymbol{\mu} \geq 0$, for some matrix A in which each row is a permutation of the k-vector $(-1, 1, 0, \ldots, 0)$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^\top$.

Like other information criteria, the ORIC is based on the log likelihood ($\log L$) and a penalty term (PT): $ORIC = -2 \log L + 2PT$. The hypothesis with the smallest ORIC value is the preferred hypothesis.

The maximum likelihood estimators (mle's) $\hat{\boldsymbol{\mu}}_m$ and $\hat{\sigma}^2_m$ for H_m with $m \in \mathcal{M}$ are the values of $\boldsymbol{\mu}$ and σ^2, respectively, that maximize the log likelihood for H_m:

$$\log L(\hat{\boldsymbol{\mu}}_m, \hat{\sigma}^2_m | y) = -\frac{N}{2} \log(2\pi) - \frac{N}{2} \log(\hat{\sigma}^2_m) - \frac{1}{2\hat{\sigma}^2_m} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}_{mi})^2, \quad (12.4)$$

where $\hat{\sigma}_m^2 = \frac{1}{N}\sum_i \sum_j (y_{ij} - \hat{\mu}_{mi})^2$ and $N = \sum_i n_i$. Because of the order restrictions, the order-restricted mle $\tilde{\boldsymbol{\mu}}_m$ must be found (Anraku, 1999). Since the term $\hat{\sigma}_m^2$ cancels out the term $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \hat{\mu}_{mi})^2$, the maximization of the likelihood actually comes down to minimizing $\log(\sigma_m^2)$, subject to $\mu_i - \mu_{i'} \geq 0$, for $i, i' = 1, \ldots, k$ (Silvapulle & Sen, 2005, pp. 42–43). Which results in the order-restricted mle $\tilde{\boldsymbol{\mu}}_m$ as defined in (12.3).

The penalty term of H_m (PT_m) is equal to:

$$PT_m = 1 + \sum_{l=1}^{q_m - 1} LP_{ml}(q_m - 1, V_m)\ l,$$

where V_m is explained below, $LP_{ml}(q_m - 1, V_m)$ is a level probability, that is, the probability that there are l distinct mean values / levels among the order-restricted means, and $q_m - 1 \leq k$ the number of distinct μ_i's in Hypothesis m. The distinct number of parameter values equals q_m, because of the unknown variance term. For example, in case $k = 5$ and the hypothesis is $H_1 : \mu_5 = \mu_3 > \{\mu_1, \mu_4\} > \mu_2$, there are 4 distinct mean values, namely μ_5 and μ_3 are a distinct value and μ_1, μ_4 and μ_2 are each a distinct value, and an unknown variance term. Thus, $q_1 = 5$. Note that the number of observations of the four distinct mean values are: $\tilde{n}_1 = n_3 + n_5$, $\tilde{n}_2 = n_1$, $\tilde{n}_3 = n_4$ and $\tilde{n}_4 = n_2$. The computation of the level probabilities (corresponding to the restrictions) can be done via simulation (Silvapulle & Sen, 2005, pp. 78–81), consisting of 5 steps (where for convenience the subscript m is left out):

1. Generate $\boldsymbol{Z}$ (of dimension $q - 1$) from $\mathcal{N}(0, V)$, where $q - 1$ equals the number of distinct μ_i values in the hypothesis. It holds that $V = \text{diag}\{1/\tilde{n}_1, \ldots, 1/\tilde{n}_{q-1}\}$, where $\tilde{n}_l$ is the number of observations of Group l ($l = 1, \ldots, q - 1$).
2. Compute $\tilde{\boldsymbol{Z}}$ via (12.3), that is, $\tilde{\boldsymbol{Z}} = \arg min_{\boldsymbol{\mu} \in H}(\boldsymbol{Z} - \boldsymbol{\mu})^T V^{-1}(\boldsymbol{Z} - \boldsymbol{\mu})$, where H is the order-restricted hypothesis.
3. Determine the number of distinct values in $\tilde{\boldsymbol{Z}}$, called levels, denote this by s.
4. Repeat the previous steps R_{PT} times. In the program, the default value of R_{PT} is $R_{PT} = 100,000$.
5. Estimate the level probability $LP_l(q - 1, V)$ by the proportion of times s is equal to l ($l = 1, \ldots, q - 1$).

The penalty term can thus be seen as the expected number of distinct parameters, that is, the expected number of distinct mean values plus 1 (because of the unknown variance term).

12.2.3 Bayesian model selection

Klugkist, Laudy, and Hoijtink (2005) present the (Bayesian) encompassing prior approach for order-restricted hypotheses in ANOVA. The model selection criterion used is the Bayes factor (Kass & Raftery, 1995; Chib, 1995), which is the ratio of marginal likelihoods of two hypotheses, say H_m and $H_{m'}$:

$$BF_{mm'} = \frac{L(\boldsymbol{\mu}, \sigma^2|y)p(\boldsymbol{\mu}, \sigma^2|H_m)/p(\boldsymbol{\mu}, \sigma^2|\boldsymbol{y}, H_m)}{L(\boldsymbol{\mu}, \sigma^2|y)p(\boldsymbol{\mu}, \sigma^2|H_{m'})/p(\boldsymbol{\mu}, \sigma^2|\boldsymbol{y}, H_{m'})}, \tag{12.5}$$

where $p(\boldsymbol{\mu}, \sigma^2|H_m)$ and $p(\boldsymbol{\mu}, \sigma^2|\boldsymbol{y}, H_m)$ are the prior and posterior distribution of the model parameters, respectively, which will be elaborated upon next.

In the encompassing prior approach, a prior $p(\boldsymbol{\mu}, \sigma^2|H_u)$ is specified for the unconstrained hypothesis $H_u : \mu_1, \ldots, \mu_k$. The prior distribution of any hypothesis H_m nested in H_u follows from the encompassing prior, using:

$$p(\boldsymbol{\mu}, \sigma^2|H_m) = p(\boldsymbol{\mu}, \sigma^2|H_u) \left(\frac{I_{\boldsymbol{\mu} \in H_m}}{\int p(\boldsymbol{\mu}, \sigma^2|H_u) I_{\boldsymbol{\mu} \in H_m} \mathrm{d}\boldsymbol{\mu} \mathrm{d}\sigma^2} \right), \tag{12.6}$$

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

The encompassing prior is specified as follows (Klugkist, Laudy, & Hoijtink, 2005):

- All model parameters are a priori considered to be independent, that is,

$$p(\boldsymbol{\mu}, \sigma^2) = p(\mu_1) \times \ldots \times p(\mu_k) \times p(\sigma^2).$$

- The prior distributions for all means are equal, that is,

$$p(\mu_1) = \ldots = p(\mu_k).$$

- As will be shown in the sequel, for each parameter a relatively uninformative, conjugate prior will be specified, that is, $p(\mu_i) \sim \mathcal{N}(\mu_0; \tau_0^2)$ for $i = 1, \ldots, k$, and $p(\sigma^2) \sim \text{Inv-}\chi^2(1; \sigma_0^2)$.

In sum, the encompassing prior for the ANOVA model is:

$$p(\boldsymbol{\mu}, \sigma^2|H_u) = \prod_{i=1}^{k} \mathcal{N}(\mu_0; \tau_0^2) \times \text{Inv-}\chi^2(1; \sigma_0^2). \tag{12.7}$$

Combination of (12.6) and (12.7) gives the prior distribution (up to proportionality) of any order-restricted hypothesis H_m:

$$p(\boldsymbol{\mu}, \sigma^2|H_m) \propto \prod_{i=1}^{k} \mathcal{N}(\mu_0; \tau_0^2) I_{\boldsymbol{\mu} \in H_m} \times \text{Inv-}\chi^2(1; \sigma_0^2). \tag{12.8}$$

In a similar way as in (12.6), the posterior of any hypothesis H_m is

$$p(\boldsymbol{\mu}, \sigma^2|\boldsymbol{y}, H_m) = p(\boldsymbol{\mu}, \sigma^2|\boldsymbol{y}, H_u) \left(\frac{I_{\boldsymbol{\mu} \in H_m}}{\int p(\boldsymbol{\mu}, \sigma^2|\boldsymbol{y}, H_u) I_{\boldsymbol{\mu} \in H_m} \mathrm{d}\boldsymbol{\mu} \mathrm{d}\sigma^2} \right). \tag{12.9}$$

The posterior distribution is proportional to the density of the data times the prior distribution, that is

$$p(\boldsymbol{\mu}, \sigma^2|\boldsymbol{y}, H_m) \propto L(\boldsymbol{\mu}, \sigma^2|y) \times \prod_{i=1}^{k} \mathcal{N}(\mu_0; \tau_0^2) I_{\boldsymbol{\mu} \in H_m} \times \text{Inv-}\chi^2(1; \sigma_0^2). \tag{12.10}$$

The encompassing posterior $p(\boldsymbol{\mu}, \sigma^2|\boldsymbol{y}, H_u)$ is (12.10) where the indicator function always equals one.

Klugkist, Laudy, and Hoijtink (2005) have shown that (12.5) when using the prior in (12.6) and subsequent posterior in (12.9) leads to a simple form for the Bayes factor for a nested hypothesis H_m with the unconstrained hypothesis H_u:

$$BF_{mu} = \frac{c_m}{d_m}, \tag{12.11}$$

where c_m and d_m are the last terms (between the large brackets) of (12.6) and (12.9), respectively. The inverse of these constants, that is, c_m^{-1} and d_m^{-1} are the proportions of the encompassing prior and posterior, respectively, in agreement with the constraints of hypothesis H_m. Estimation of these proportions is straightforward using sampling. This means that in the context of order-restricted ANOVA, Bayes factors can be obtained without the – often burdensome – estimation of marginal likelihoods.

Specification of the encompassing prior

To complete the specification of the prior distribution, values must be assigned to the hyper-parameters μ_0, τ_0^2, and σ_0^2. Klugkist and Hoijtink (2007) showed that Bayes factors for hypotheses formulated using inequality constraints among parameters are insensitive to the exact specification, as long as the encompassing prior is relatively vague. However, the results for hypotheses containing equality constraints are sensitive to the choice of τ_0^2. Although we want relatively uninformative priors, that is, a large τ_0^2, too large values result in Bartlett's or Lindley's paradox (cf. Lindley, 1957; Bernardo & Smith, 1994). Hence, to obtain reasonable values for the hyper-parameters, the prior specification is data-based. A Gibbs sample (Smith & Roberts, 1993) is drawn from the unconstrained posterior $p(\boldsymbol{\mu},\sigma^2|\boldsymbol{y}, H_u) = L(\boldsymbol{\mu}, \sigma^2|y) \times p(\boldsymbol{\mu}, \sigma^2|H_u)$, where $p(\boldsymbol{\mu}, \sigma^2|H_u) \propto 1$. Summaries of the posterior sample provide values for the hyper-parameters according to the following choices:

- For σ_0^2, the posterior mean of σ^2 is used. This provides a value that is reasonable for the data at hand and, with 1 degree of freedom, a posterior that is hardly affected by the prior.
- To obtain μ_0 and τ_0^2, the information about each of the μ's in the posterior sample is combined as follows: Based on the posterior sample, a credibility interval for each μ_i $(i = 1, \ldots, k)$ is determined by: $\bar{\mu}_i \pm pv \cdot s_{\mu_i}$, where $\bar{\mu}_i$ and s_{μ_i} are the mean and standard deviation of the sampled values for μ_i, respectively, and pv stands for *prior vagueness*. With the pv value it is specified which interval is used (e.g., $pv = 2$ provides the 95% credibility interval) and allows the user a choice for the amount of vagueness in the encompassing prior. Subsequently, the smallest lower bound (lb) and the largest upper bound (ub) of the k intervals define one broad interval containing all reasonable values for each of the μ's. From this interval, $\mu_0 = (lb + ub)/2$ and $\tau_0 = (ub - lb)/2$ are specified.

For hypotheses containing equality constraints, the value specified for pv will affect the resulting Bayes factors. Larger pv values provide more support for hypotheses containing equality constraints (i.e., Lindley's paradox). Specification of the pv value by the user also provides the option to investigate prior sensitivity by running the program several times with different values. This is recommended if one or more of the hypotheses contain equality constraints among the means.

Estimation of the Bayes factor

The computation of Bayes factors of order-restricted hypotheses versus the unconstrained hypothesis is straightforward by taking a sample from the unconstrained prior (12.7) and a sample from the unconstrained posterior, that is, (12.10) with $I_{\boldsymbol{\mu} \in H_m} = 1$ for all $\boldsymbol{\mu}$. Samples are obtained by application of the Gibbs sampler (Smith & Roberts, 1993). The proportions of prior and posterior iterations in agreement with the order-restricted hypotheses provide estimates for c_m^{-1} and d_m^{-1}, respectively. However, for a hypothesis containing at least one strict equality (e.g., $\mu_1 = \mu_2$), the direct application of this approach would result in the problematic outcome $d_m^{-1} = 0$ and $c_m^{-1} = 0$. Therefore, the program evaluates 'about equality' constraints instead, that is, $|\mu_1 - \mu_2| < \delta$ for a positive small δ. Two options are provided in the software: Researchers have the opportunity to investigate 'relevant differences' between means by specifying a non-zero δ, or strict equality constrained hypotheses can be evaluated, in which case δ approaches zero in a stepwise method. The latter requires an extension of the basic approach and includes constrained sampling. This will be elaborated in the next section.

For hypotheses that only require unconstrained sampling, estimation of Bayes factors in the software is based on a minimum of R_{BMS} iterations from both prior and posterior. The default value of R_{BMS} is $R_{BMS} = 500,000$. The posterior sample is taken after discarding 1,000 iterations that serve as burn-in. Note that, sampling from the unconstrained prior does not require a burn-in period, since all parameters are a priori independent.

For highly constrained hypotheses, the default value of $R_{BMS} = 500,000$ iterations may be insufficient to obtain stable estimates of d_m and c_m. Therefore, some additional rules are incorporated in the program, when the default setting is chosen: For more than 6 groups, the number of iterations from prior and posterior are doubled; for more than 10 groups, the number of iterations from prior and posterior are set at 5 million. Furthermore, another additional rule is incorporated (whether the default setting is used or not): if a minimum of 100 prior 'hits' (iterations in agreement with the constraints) is not reached for each hypothesis in the set, more iterations are added until this is the case. For an elaboration on the investigation of the stability and Monte Carlo errors of Bayes factors computed via (12.11), see Klugkist and Hoijtink (2007).

Stepwise estimation for small δ

For small values of δ, the estimation as just described would be rather inefficient. Furthermore, for $\delta = 0$ it would give the result $c_m^{-1} = d_m^{-1} = 0$. Therefore, a procedure is applied where the unconstrained samples are evaluated with a not too small initial δ value, denoted δ_0, followed by a procedure that decreases δ_0 in a stepwise way, using $\delta_r = \delta_{r-1}/3$, for steps $r = 1, \ldots, R$ (see also Klugkist, 2008).

The stepwise procedure is based on the following product rule for the Bayes factor of H_m with the unconstrained H_u:

$$BF_{mu} \approx BF_{m_{\delta_0} u} \times BF_{m_{\delta_1} m_{\delta_0}} \times \ldots \times BF_{m_{\delta_R} m_{\delta_{R-1}}}$$

$$= \frac{c_{m_{\delta_0}u}}{d_{m_{\delta_0}u}} \times \frac{c_{m_{\delta_1}m_{\delta_0}}}{d_{m_{\delta_1}m_{\delta_0}}} \times \ldots \times \frac{c_{m_{\delta_R}m_{\delta_{R-1}}}}{d_{m_{\delta_R}m_{\delta_{R-1}}}}. \tag{12.12}$$

Consider the case where H_m includes both inequality and strict equality ($\delta = 0$) constraints. The notation m_{δ_r} is used to denote the constraints of H_m, where the desired value $\delta = 0$ is replaced by a larger value δ_r ($r = 0, \ldots, R$).

The first Bayes factor in (12.12), $BF_{m_{\delta_0}u} = \frac{c_{m_{\delta_0}u}}{d_{m_{\delta_0}u}}$, requires sampling from the unconstrained prior and posterior and counting the number of iterations in agreement with all constraints in m_{δ_0}, that is, the order restrictions as well as the equalities evaluated with δ_0. The second and subsequent steps require constrained sampling, that is, sampling from (12.8) and (12.10). Consider, for instance, $BF_{m_{\delta_1}m_{\delta_0}} = \frac{c_{m_{\delta_1}m_{\delta_0}}}{d_{m_{\delta_1}m_{\delta_0}}}$, where $(c_{m_{\delta_1}m_{\delta_0}})^{-1}$ denotes the proportion of iterations, sampled from the prior of H_m with δ replaced by δ_0, that are in agreement with H_m with δ replaced by δ_1 ($\delta_1 < \delta_0$). Similarly, $d_{m_{\delta_1}m_{\delta_0}}$ is based on a sample from the posterior constrained to the area that complies with the constraints of m_{δ_0}. Constrained sampling is also done by application of the Gibbs sampler, with inverse probability sampling to obtain samples in agreement with the constraints (see Gelfand, Smith, & Lee, 1992).

In each step, H_m is evaluated with a smaller δ_r value. The stepwise Bayes factors represent change in the estimated BF_{mu}, as a consequence of the decrease in δ_r. At a certain point, a further decrease of the value for δ_r no longer changes the Bayes factor, that is, for large enough R, $BF_{m_{\delta_R}m_{\delta_{R-1}}} \to 1$ (Berger & Delempady, 1987). This implies that a good approximation of BF_{mu} with *exact* equalities is obtained.

To obtain an efficient estimation procedure, it is important to start with a large enough δ_0. In the software, the starting value is prior-based and equals $\tau_0/2$ (τ_0 for more than 8 groups) unless this value is smaller than the user-specified δ in which case δ is evaluated directly. Sampling from the unconstrained prior and posterior (first step) is as explained in the previous section. In each subsequent step ($r = 1, \ldots, R$), samples are drawn from the constrained priors and posteriors after a burn-in of 100 iterations. The number of iterations from each constrained prior is minimally 500,000. If necessary, up to 1 million iterations are done until the number of hits reaches 500. If after 1 million iterations the number of hits is below 100, more samples are drawn until 100 hits are obtained. From each constrained posterior, samples are drawn until 500 hits are obtained, with a maximum of 1 million iterations.

The number of steps, R, is determined by one of two stopping rules: convergence of the estimate of the final Bayes factor is assumed if two subsequent $BF_{m_{\delta_r}m_{\delta_{r-1}}}$ values deviate less than 0.05 from 1. When $\delta \neq 0$, the last step of the procedure is performed as soon as $\delta_r \leq \delta$ (if $\delta_r < \delta$, it is set at δ).

Interpretation of the results

The software estimates Bayes factors for each order-restricted hypothesis with the unconstrained hypothesis using (12.11) or (12.12). The Bayes factor for the comparison of two order-restricted hypotheses, say H_m and $H_{m'}$ can be computed, using:

$$BF_{mm'} = \frac{BF_{mu}}{BF_{m'u}}.$$

A Bayes factor provides the amount of support of one hypothesis compared to another. If, for instance $BF_{mm'} = 6$, the support for H_m is 6 times as large as for $H_{m'}$. Likewise, $BF_{mm'} = 0.5$ shows that the support for H_m is 2 times as small as the support for $H_{m'}$.

Furthermore, the software provides posterior model probabilities (*pmp*), representing the relative support for each hypothesis in a finite set of hypotheses ($\mathcal{M}$). To obtain *pmp*'s from Bayes factors, prior model probabilities must be specified, representing the degree of belief in each hypothesis before observing the data. A usual – objective – choice are equal prior probabilities for all hypotheses, that is, $p(H_m) = 1/M$, for $m \in \mathcal{M}$, where M denotes the total number of hypotheses. This prior specification, which is also adopted in the software, leads to the following equation for $pmp(H_m)$:

$$pmp(H_m) = \frac{BF_{mu}}{\sum_{m \in \mathcal{M}} BF_{mu}}.$$

Posterior model probabilities can be computed including or excluding the unconstrained hypothesis. As an example, consider H_0, H_1 and H_2 from (12.2). In the software, the user can specify the presence of just these 3 hypotheses (i.e., $M = 3$) and the *pmp*'s are computed excluding the unconstrained hypothesis:

$$pmp(H_m) = BF_{mu} / \left(BF_{0u} + BF_{1u} + BF_{2u}\right).$$

Alternatively, one can also explicitly add the unconstrained hypothesis as a hypothesis of interest ($M = 4$). The resulting *pmp* values, denoting the unconstrained hypothesis by H_3, are:

$$pmp(H_m) = BF_{mu} / \left(BF_{0u} + BF_{1u} + BF_{2u} + BF_{3u}\right),$$

for $m \in \mathcal{M}$ and with $BF_{3u} = 1$.

12.3 Example based on Lucas (2003)

The three approaches for confirmatory ANOVA will be illustrated using data from the research of Lucas (2003). In this study, the interest lies in the amount of influence a leader has on his/her group members. The experiment contained five experimental groups: (1) a group with a randomly selected male leader, (2) a group with a randomly

Group	Mean influence	SD	n
1	2.33	1.86	30
2	1.33	1.15	30
3	3.20	1.79	30
4	2.23	1.45	30
5	3.23	1.50	30

Table 12.1: Group means and standard deviations (SD) of influence (Lucas, 2003).

selected female leader, (3) a group with a male leader selected on ability, (4) a group with a female leader selected on ability, and (5) a group with a female leader selected on ability after institutionalization of female leadership. The institutionalization is done by showing the participants a film in which it is normal to have female leadership and females do well as leaders. The resulting group means and standard deviations of the influence of the leader are shown in Table 12.1.

The research question of Lucas (2003) is: "Can institutionalization of female leadership reduce the influence gap between woman and men by legitimating structures of female leadership?" The expectations of Lucas (2003) are in short:

- Male leaders (Groups 1 and 3) have higher influence over participants than female leaders (Groups 2 and 4, respectively), ceteris paribus.
- Leaders appointed on ability (Groups 3 and 4) have higher influence over participants than leaders appointed randomly (Groups 1 and 2, respectively), ceteris paribus.
- Institutionalized female leaders selected on ability (Group 5) have higher influence over participants than 'normal' female leaders selected on ability (Group 4), or than randomly selected female leaders (Group 1).
- Institutionalized female leaders selected on ability (Group 5) have (almost) the same influence over participants as male leaders appointed on ability (Group 3).

These expectations can be represented by the hypothesis $H_1 : \mu_5 = \mu_3 > \{\mu_1, \mu_4\} > \mu_2$, where μ_i represents the mean influence of the leader in Group i.

Another hypothesis of interest can be:

- Leaders chosen on basis of ability score higher than leaders selected at random (so, Group 3 scores higher than Group 1 and Group 4 scores higher than Group 2).
- Male leaders selected at random (Group 1) have an higher influence than female leaders selected on competence (Group 4).
- There is no difference in influence of female leaders selected on competence in case of institutionalization (Group 5) or in the 'normal' case (Group 4).

This can be represented by $H_2 : \mu_3 > \mu_1 > \mu_4 = \mu_5 > \mu_2$.

In addition, the traditional null and alternative hypothesis, H_0 and H_3 from (12.2), respectively, will be used to illustrate the $\bar{F}$ test, the order-restricted information criterion and Bayesian model selection.

12.3.1 Results using the $\bar{F}$ test

Using the $\bar{F}$ test for the evaluation of the four hypotheses specified in the Lucas' example (see (12.2)), five tests are performed (Table 12.2). First, H_0 is tested against H_3. This test results in a p value smaller than 0.001, rejecting H_0. Second, two tests are done with respect to H_1: H_0 is tested against the order-restricted H_1 and H_1 is tested against the unconstrained hypothesis H_3. These tests result in a p value smaller than 0.001 and a p value of 0.995, respectively, both favoring H_1. Third, two tests are done with respect to H_2: H_0 is tested against H_2 and H_2 is tested against H_3. The resulting p values are a p value smaller than 0.001 and a p value of 0.07,

respectively, both favoring H_2. So, both H_1 and H_2 are preferred. Since no direct comparison between order-restricted hypotheses is possible with the $\bar{F}$ test, nothing can be concluded with respect to an overall preferred hypothesis.

Hypotheses tested	$\bar{F}$	p value
H_0 against H_3	30.27	< 0.001
H_0 against H_1	30.26	< 0.001
H_1 against H_3	0.01	0.995
H_0 against H_2	22.91	< 0.001
H_2 against H_3	7.36	0.070

Table 12.2: The $\bar{F}$ tests of the four specified hypotheses.

12.3.2 Results using ORIC

The ORIC consists of a fit/likelihood part and a complexity/penalty part. In Table 12.3 the likelihood, penalty term, and ORIC values are given for H_0, H_1, H_2 and H_3. The penalty for H_0 equals 2, because there are two distinct parameters: all means are equal, which represents one distinct parameter, and the parameter σ^2. Analogously, the penalty for H_3 is 6: there are five distinct means (since there are no restrictions among the means) and one variance parameter. In hypotheses H_1 and H_2 there are also five means, but these hypotheses contain inequality constraints, as opposed to H_0 and H_3. The penalties of these order-restricted hypotheses are calculated as explained earlier. The values of the penalties are respectively 3.20 and 3.13. The hypothesis with the smallest value for the ORIC is the preferred hypothesis. Thus, in the example, H_1 is the preferred hypothesis according to the ORIC.

Hypothesis	$\log L_m$	Penalty	ORIC
$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$	-292.27	2.00	588.54
$H_1 : \mu_5 \ = \mu_3 > \{\mu_1,\ \mu_4\} > \mu_2$	-278.05	3.19	**562.49**
$H_2 : \mu_3 > \mu_1 > \mu_4 \ = \mu_5 > \mu_2$	-281.76	3.14	569.79
$H_3 : \mu_1,\ \mu_2,\ \mu_3,\ \mu_4,\ \mu_5$	-278.05	6.00	568.10

Table 12.3: The ORIC values of the four specified hypotheses.

12.3.3 Results using BMS

The results using BMS are presented in Table 12.4. For each hypothesis, the Bayes factor comparing that hypothesis with the unconstrained hypothesis (H_3) is presented for three different values of pv. The equality constraints are evaluated as strict

equalities, that is, $\delta = 0$ is approximated using the stepwise approach explained in Section 12.2.3.

	$pv = 1$		$pv = 2$		$pv = 3$	
Model	*BF*	*pmp*	*BF*	*pmp*	*BF*	*pmp*
$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$	0.0	0.00	0.0	0.00	0.0	0.00
$H_1 : \mu_5 = \mu_3 > \{\mu_1, \mu_4\} > \mu_2$	57.8	**0.96**	67.9	**0.96**	80.6	**0.97**
$H_2 : \mu_3 > \mu_1 > \mu_4 = \mu_5 > \mu_2$	1.4	0.02	1.5	0.02	1.8	0.02
$H_3 : \mu_1, \mu_2, \mu_3, \mu_4, \mu_5$	1.0	0.02	1.0	0.01	1.0	0.01

Table 12.4: Bayes factors (*BF*) and posterior model probabilities (*pmp*) for different prior specifications.

As was explained in Section 12.2.3, the encompassing prior is specified to be low informative and is based on the observed data. The resulting encompassing prior for the Lucas data and $pv = 1$ is:

$$p(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \sigma^2 | H_3) = \prod_{i=1}^{5} \mathcal{N}(2.28; 1.53) \times \text{Inv-}\chi^2(1; 2.50).$$

The priors using $pv = 2$ and $pv = 3$ differ only with respect to the variance of the normal distributions, with values 2.32 and 3.27, respectively.

Irrespective of the choice made for the prior, H_1 is clearly the most supported order-restricted hypothesis. The corresponding posterior probabilities (using equal prior model probabilities) for the four hypotheses are (about) 0.00, 0.97, 0.02 and 0.01, respectively (for each *pv* value). These results again lead to the conclusion that H_1 is the best of the four models/hypotheses considered, as was also concluded when using the ORIC. Comparison of the three prior specifications shows that, for the hypotheses at hand, the results are not very sensitive to the specification of the prior.

12.A *ConfirmatoryANOVA.exe* user manual

This user manual will describe and illustrate the options available in *ConfirmatoryANOVA.exe* (published along with this manuscript and also available at `http://staff.fss.uu.nl/RMKuiper`). This program is made in Fortran 90 using the Intel Visual Fortran Compiler 9.1 for Windows. This compiler uses IMSL 5.0.

To make the program more user-friendly, an interface is made with use of C# in Microsoft Visual Studio 2005 (Appendix 12.B). The interface calls the exe file made in Fortran 90 and it makes the appropriate input file needed for the Fortran 90 program. The input for the Fortran 90 program is described in this appendix. Those interested in only the interface are referred to Appendix 12.B.

ConfirmatoryANOVA.exe is free of use. However, when results obtained with this program are published, please refer to Kuiper et al. (2010). In the program, the following methods can be performed:

- the $\bar{F}$ test,
- the order-restricted information criterion (ORIC),
- Bayesian model selection (BMS).

12.A.1 Modification input files

No matter what analysis should be performed two text files have to be modified (such that they apply to your data), namely *Input.txt* and *Data.txt*.

It should be noted that:

- The names of the text files are fixed and cannot be changed. These files have to be text files (also known as ASCII files).
- The format of these files should not be changed, that is, do not add empty lines and do not delete lines containing labels.
- The data in *Data.txt* should be complete, that is, missing data are not allowed.

First half of *Input.txt*

First of all, you must denote which analyses should be performed. This has to be done in *Input.txt*. A certain analysis will be performed if in the line below the name of that analysis a 1 is filled in. It will not be performed, when a 0 is filled in.

When the ORIC and the $\bar{F}$ test should be performed and BMS should not, the first half of *Input.txt* should look as follows (when using the default values for the seed value and number of iterations):

```
Seed value and number of iterations (>0) for Fbar test, ORIC, and BMS
123 100000 100000 500000
Perform F bar test, ORIC, BMS (1 = yes, 0 = no)
1 1 0
```

We will come back to the seed value and number of iterations in Section 12.A.5.

Data.txt

Second of all, group membership and the corresponding data must be given in *Data.txt*, where in the first column the group numbers must be given and in the second column the corresponding data point y_{ij}. The order of the group numbers and data points does not matter as long as the group number corresponds to the data point in the same row. The following are three examples of how *Data.txt* could look:

1	3.58
1	-0.15
...	...
2	1.67
2	1.85
...	...
3	1.39
3	4.53
...	...
4	1.57
4	2.97
...	...
5	1.38
5	4.58
...	...

1	3.58
2	1.67
3	1.39
4	1.57
5	1.38
1	-0.15
2	1.85
3	4.53
4	2.97
5	4.58
...	...
...	...
...	...
...	...
...	...

3	1.39
2	1.85
5	4.58
5	1.38
3	4.53
4	2.97
1	-0.15
2	1.67
4	1.57
1	3.58
...	...
...	...
...	...
...	...
...	...

The group numbers do not need to be sequential. For example, if you have a SPSS or Excel file with data for 10 groups and in the current analysis you want to compare only 5 groups (which are not the first five). In that case, you can just copy the appropriate data from the SPSS or Excel file to a text file without adjusting the group numbers. In the software, the group numbers will be made sequential and the output will be given for these adjusted sequential group numbers. For example, if you have data with group numbers 1, 4, 5, 6, and 8 (whether the data are in order or not), these will become group numbers 1, 2, 3, 4, and 5, respectively. Note that, in specifying the restrictions (see the next sections), you need to use the adjusted sequential group numbers.

From the data, the number of groups and the number of observations per group are determined.

12.A.2 Basic elements of writing constraints

In performing an $\bar{F}$ test, in determining the ORIC or in doing BMS, all the hypotheses of interest, like H_0 to H_3 in (12.2), must be given explicitly. Note that it is also possible to specify a set of hypotheses without the classical null $H_0 : \mu_1 = \ldots = \mu_k$ and/or alternative $H_u : \mu_1, \ldots, \mu_k$. However, we recommend to include the alternative H_u (when doing model selection), since it can be used to protect against choosing a weak hypothesis (Kuiper & Hoijtink, 2010). Note that one should include H_0 only when there is real interest in H_0.

When using the ORIC or doing BMS several models/hypotheses are compared to each other. In the $\bar{F}$ test, the order-restricted hypotheses (like H_1 and H_2 in (12.2)) are tested against H_0 and H_u. If the classical null and/or the alternative are included in the set of hypotheses, the classical null will be tested against the classical alternative.

The basic elements, for writing down the hypotheses of interest are:

1. Representation of an equality sign (=)
   Suppose the hypothesis of interest is $\mu_5 = \mu_3$, that is, $\mu_5 = \mu_3,\ \mu_1,\ \mu_2,\ \mu_4$. The ordering of the group numbers in this restriction is represented by: 5 3 1 2 4. The

restriction is represented by: 1 1 0 0 0, where the 1s indicate that mean 5 and 3 belong to Set 1 and are equal to each other, and where 0 indicates that the corresponding mean is unrestricted. N.B. in a restriction the first set is always labeled as 1, the second as 2 (and so on).

2. Representation of a greater than sign (>)
   Suppose the hypothesis of interest is $\mu_1 > \mu_3,\ \mu_2,\ \mu_4,\ \mu_5$. The ordering of the group numbers in this restriction is represented by: 1 3 2 4 5. The restriction is represented by: 1 -3 0 0 0, where -3 means that mean 3 is smaller than mean 1. Thus, it represents $\mu_3 < \mu_1$, which is equal to $\mu_1 > \mu_3$. Here again 1 indicates that mean 1 belongs to Set 1. Because of the inequality restriction between mean 1 and 3, mean 3 belongs to Set 2 (the importance of this will be made clear in the next section).
3. Representation of a smaller than sign (<)
   Suppose the hypothesis of interest is $\mu_1 < \mu_2,\ \mu_3,\ \mu_4,\ \mu_5$. The ordering of the group numbers in this restriction is represented by: 1 2 3 4 5. The restriction is represented by: 1 -1 0 0 0, where -1 means that mean 2 is greater than mean 1. Thus, it represents $\mu_2 > \mu_1$, which is equal to $\mu_1 < \mu_2$.

Every hypothesis can be represented by these basic elements in one or more restrictions. For example, $\mu_5 > \mu_3 < \mu_1,\ \mu_2 < \mu_4$, can be represented by the restrictions:
$\mu_5 > \mu_3,\ \mu_1,\ \mu_2,\ \mu_4$,
$\mu_5,\ \mu_3 < \mu_1,\ \mu_2,\ \mu_4$,
$\mu_2 < \mu_4,\ \mu_5,\ \mu_3,\ \mu_1$,
which can be represented by:

```
Ordering of means in restriction
5 3 1 2 4
5 3 1 2 4
2 4 5 3 1
(Order) Restrictions
1 -3  0  0  0
0  1 -1  0  0
1 -1  0  0  0
```

12.A.3 Combinations of basic elements

Often the hypothesis of interest can be represented in a smaller number of restrictions than when using only the basic elements. The following shortcuts can be used:

1. $\mu_5 = \mu_3 = \mu_1,\ \mu_2 = \mu_4$
   Because of the equality constraints ("="), means 5, 3 and 1 belong to Set 1. Therefore, means 2 and 4 belong to Set 2. Thus, this hypothesis can be represented by the following ordering of the group numbers and corresponding restriction:

   ```
   Ordering of means in restriction
   5 3 1 2 4
   (Order) Restrictions
   1  1  1  2  2
   ```

2. $\{\mu_5 = \mu_3 = \mu_1\} > \{\mu_2 = \mu_4\}$
   Because of the equality constraints, mean 5, 3 and 1 belong to Set 1, and means 2 and 4 belong to Set 2. Because of the constraint between mean 1 and 2, mean 2 belongs implicitly to Set 2. This hypothesis can be represented by the following ordering of the group numbers and corresponding restriction:

   ```
   Ordering of means in restriction
   5 3 1 2 4
   (Order) Restrictions
   1  1  1 -3  2
   ```

   The 1s indicate the equality constraints between mean 5, 3 and 1, the -3 represents the inequality constraint ">" between mean 1 and 2, and the 2 indicates the equality constraints between mean 2 and 4 (because mean 2 implicitly belongs to Set 2). Notably, the restrictions $\mu_5 = \mu_1,\ \mu_5 > \mu_2,\ \mu_3 > \mu_2,\ \mu_5 > \mu_4,\ \mu_3 > \mu_4,\ \mu_1 > \mu_4$ do not have to be formulated explicitly, these will hold since it holds that $\mu_5 = \mu_3,\ \mu_3 = \mu_1,\ \mu_1 > \mu_2$ and $\mu_2 = \mu_4$.
3. $\mu_5 = \mu_3 > \mu_1 > \mu_2 = \mu_4$
   This hypothesis can be represented by the following ordering of the group numbers and corresponding restriction:

   ```
   Ordering of means in restriction
   5 3 1 2 4
   (Order) Restrictions
   1  1 -3 -3  3
   ```

   The 1s indicate the equality constraint between mean 5 and 3. The -3s represents the inequality constraint ">" between mean 3 and 1 and between 1 and 2. Note that mean 1 implicitly belongs to Set 2 and mean 2 implicitly to Set 3. Therefore, the equality constraint between mean 2 and 4 is represented by the 3, because mean 2 and 4 belong to Set 3.
4. $\mu_5 = \mu_3 > \mu_1 < \mu_2 = \mu_4$
   Likewise, this hypothesis can be represented by the following ordering of the group numbers and corresponding restriction:

   ```
   Ordering of means in restriction
   5 3 1 2 4
   (Order) Restrictions
   1  1 -3 -1  3
   ```

5. $\mu_5 = \mu_3 > \mu_1 > \mu_2,\ \mu_4$
   This hypothesis can be represented by the following ordering of the group numbers and corresponding restriction:

   ```
   Ordering of means in restriction
   5 3 1 2 4
   (Order) Restrictions
   1  1 -3 -3 0
   ```

Note that, mean 4 is a free parameter, that is, a mean which is not restricted at all. The free parameters are denoted by a "0" and no set numbers are assigned (directly or indirectly). Here, as in the previous two examples, mean 2 belongs to Set 3 and mean 4 belongs to another set. However, this is not Set 4. Another example is given next.

6. $\mu_5 = \mu_3,\ \mu_4,\ \mu_1 > \mu_2$
   This hypothesis can be represented by the following ordering of the group numbers and corresponding restriction:

```
Ordering of means in restriction
5 3 4 1 2
(Order) Restrictions
1  1  0  2 -3
```

   The 1s indicate the equality constraint between mean 5 and 3. Note that, as mentioned in the previous example, no set number is assigned to free parameters. So, mean 4 belongs to another set than all the other means, but no set number is assigned. A 0 is filled in. Mean 1 belong to the next set, that is, Set 2. The -3 represents the inequality constraint ">" between mean 1 and 2. Note that mean 2 indirectly belongs to Set 3.

7. $\mu_5 = \mu_3 > \mu_1 < \mu_2 = \mu_4$
   This hypothesis can be represented by the following ordering of the group numbers and corresponding restriction:

```
Ordering of means in restriction
5 3 1 2 4
(Order) Restrictions
1  1 -3 -1  3
```

8. $\mu_5 = \mu_3 > \mu_1 = \mu_2 > \mu_4$
   This hypothesis can be represented by the following ordering of the group numbers and corresponding restriction:

```
Ordering of means in restriction
5 3 1 2 4
(Order) Restrictions
1  1 -3  2 -3
```

   The 1s indicate the equality constraints between mean 5 and 3, the -3s represents the inequality constraint ">" between mean 3 and 1 and between 2 and 4. Note that mean 1 implicitly belongs to Set 2. Therefore, the equality constraint between mean 1 and 2 is represented by the 2.
   In the program *ConfirmatoryANOVA.exe*, error messages are built in to detect erroneously stated hypotheses. But sometimes wrongly reported hypotheses are not detected, because the expressed hypothesis represents another existing hypothesis. When accidently a 3 is given instead of a 2, another existing hypothesis is formulated. Namely, the restriction "1 1 -3 3 -3" represents the hypothesis $\mu_5 = \mu_3 > \mu_1,\ \mu_2 > \mu_4$. So, care must be taken in writing down the hypothesis of interest.

9. $\mu_5 = \mu_3 > \mu_1,\ \mu_2 > \mu_4$
   This hypothesis can be represented by the following ordering of the group numbers and corresponding restriction:

```
Ordering of means in restriction
5 3 1 2 4
(Order) Restrictions
1  1 -3  3 -3
```

12.A.4 Equalities and about equalities in BMS

In the $\bar{F}$-test, in the ORIC and in BMS for strict equalities ($\delta = 0$), the two hypotheses $\mu_1 = \mu_2 = \mu_3$ (specified by 1 restriction, with ordering of means 1 2 3, and (order) restrictions 1 1 1) and $\mu_1 = \mu_2, \mu_2 = \mu_3$ (specified by 2 restrictions, both with ordering of means 1 2 3, and (order) restrictions 1 1 0 and 0 1 1) are equivalent.

However, the BMS approach in the program also provides the option to specify about equality constraints ($\delta > 0$). In that case, the second hypothesis is evaluated using $|\mu_1 - \mu_2| < \delta$ and $|\mu_2 - \mu_3| < \delta$, whereas the first hypothesis adds a third constraint: $|\mu_1 - \mu_3| < \delta$. The results of the first and second hypothesis may differ and therefore careful consideration of the formulation of hypotheses is important.

12.A.5 Set the seed value and number of iterations

The calculation of the p value of the $\bar{F}$, the penalty of the ORIC (i.e., PT), and BMS are sampling based approaches. For example, when generating data from a normal distribution (to determine the p value of the $\bar{F}$ or the penalty of the ORIC), a seed value is needed. When using the same seed value, the same data will be 'sampled'. When looking at another seed value in a rerun of the same problem, one can also see how stable the results are. Thus, the p value of the $\bar{F}$, the penalty of the ORIC (i.e., PT), and the results of BMS can differ for various seed values.

In case a result is not stable, the number of iterations needs to be set higher. In the $\bar{F}$ test, the p value depends on the number of iterations R_p. When using the ORIC, the penalty is dependent on the number of iterations R_{PT}. When doing BMS, the Gibbs sampler is used, which is based on a minimum of R_{BMS} iterations. These values can also be set in the input, namely in the second line of *Input.txt* (see Section 12.A.1). The default values of the number of iterations in each method are: $R_p = 100,000$, $R_{PT} = 100,000$, and $R_{BMS} = 500,000$.

Note that the higher the number of iterations the higher the computing time. If one lowers the number of iterations (to lower the computing time), one must be aware that this probably affects the stability of the results. Furthermore, when the initial number of iterations for BMS (i.e., R_{BMS}) is lowered, the computing time is not necessarily decreased, because of the requirement of a minimum of 100 prior hits (see subsections "Estimation of the Bayes factor" and "Stepwise estimation for small δ" in Section 12.2.3).

12.A.6 Error messages

In the program *ConfirmatoryANOVA.exe*, error messages are built in to detect incorrectly stated hypotheses. However, it does not detect all erroneously formulated hypotheses, since the reported hypothesis can represent another existing hypothesis, as is made clear in Section 12.A.3.

It is also possible to state other input wrongly. For example, an improper number of restrictions is given. When making a mistake, an informative warning will be given.

However, it is possible to make a mistake that we have not foreseen. In that case, check the input and compare it to the data. If you cannot solve the problem, send the input and data file to R.M.Kuiper@uu.nl.

12.A.7 Modification of the second half of *Input.txt*

For all three methods (i.e., the $\bar{F}$ test, the ORIC, and BMS), all hypotheses of interest must be given explicitly. In case BMS is performed, two additional specifications need to be made: The desired δ (for exact equalities specify, $\delta = 0$, and for an about equality, any positive number can be specified) and the prior vagueness pv (default recommendation $pv = 2$; any positive number may be specified). This must be done in the second half of *Input.txt*. If the hypotheses of interest are the set of hypotheses of Lucas stated in (12.2), *Input.txt* has the following format (where "< ... >" is not part of the format, but is used to give remarks):

Number of models to be compared
<Fill in the number of models / hypotheses you want to compare; e.g.,>
4
Number of restrictions per model
<Fill in, for every model / hypothesis, the number of restrictions that represent that model / hypothesis; e.g.,>
1
2
1
1
Ordering of means in restriction
<Fill in the ordering of the means / group numbers for each restriction for every model / hypothesis. The orderings per restrictions are separated by an "enter". The ordering consists of the numbers 1 to "the total number of groups". For more details see "Basic Elements of Writing Constraints" and "Combinations of Basic Elements"; e.g.,>
1 2 3 4 5
5 3 1 2 4
3 4 2 1 5
3 1 4 5 2
1 2 3 4 5

(Order) Restrictions
<Fill in the restrictions. This must be done in a certain manner, which is explained in "Basic Elements of Writing Constraints" and "Combinations of Basic Elements"; e.g.,>
1 1 1 1 1
1 1 -3 -3 0
1 -3 -3 0 0
1 -3 -3 3 -3
0 0 0 0 0
When BMS is performed, an interval for equality relations (delta) is needed and a parameter for prior vagueness (pv)
<Fill in $\delta \geq 0$ *and* $pv > 0$*; e.g.,>*
0.0 2.0

12.A.8 Save and close

When you have modified *Input.txt* (such that it applies to your data), you should save and close it.

12.A.9 Run *ConfirmatoryANOVA.exe*

When *ConfirmatoryANOVA.exe* is run, the output file *Output.txt* will be created in the folder you are working in.

Output.txt

Output.txt gives the results of the requested analyses. In Section 12.B.5, the output is given when using the interface for the Lucas example described in this chapter (provided that all three analyses, that is, the $\bar{F}$ test, the ORIC, and BMS, are performed).

In case the interface is not used, the output will be a bit different, namely in two ways:
1) The hypotheses of interest will be displayed in the way they are filled in *Input.txt*. When using the interface, the hypotheses of interest are formulated in terms of "μ_i", ">", "<", "=", and ",".
2) The numbering of the hypotheses is different.

When using the interface and when H_0 is included in the set of hypotheses, H_0 will become Hypotheses 1 in the output. When H_u is included in the set of hypotheses, H_u will become Hypotheses 2, when H_0 is also included, and Hypothesis 1, when H_0 is not included. The other hypotheses will also be adjusted to the appropriate hypothesis number. So, an order-restricted hypotheses H_1 will become Hypotheses 3 (when both H_0 and H_u are specified) or Hypotheses 2 (when only H_0 or H_u is specified); et cetera.

12.B User manual of *ConfirmatoryANOVA.exe* with interface

12.B.1 Read, write or copy data

First of all, the data should be entered. Press on the *Data* button in the *ConfirmatoryANOVA.exe* form (see Figure 12.1) to go the *DataInput* form (see Figure 12.2). The data can be entered manually, by copying it from a file (say an SPSS or Excel file) or by reading it from a text (i.e., *.txt*) file. See also Section 12.A.1 for a description of the data format.

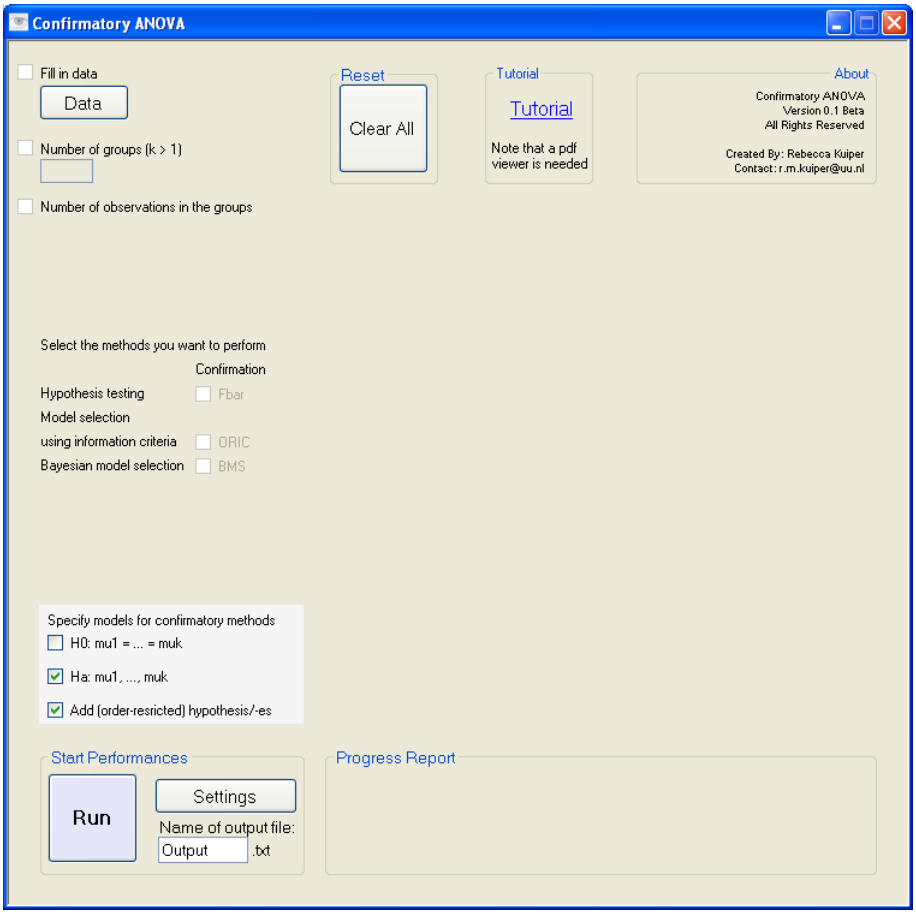


Fig. 12.1: *ConfirmatoryANOVA.exe* form.

When the data are read from a file or, otherwise, after clicking on the *OK* button, the data are validated. In case of invalid data (e.g., in case no number is entered or in case a row has only a group number or has only a data point), the corresponding lines are made red. The invalid data should be corrected (by adjusting the data in the textbox/field or by adjusting the *.txt* file and rereading the adjusted file) or deleted

(e.g., when all lines should be deleted, by clicking on the *Clear Invalid Data* button). After the adjustments, press the *OK* button. The data will be validated again. Note that, in case of deleting data, the number of observations per group (which is shown in the *ConfirmatoryANOVA.exe* form) will be adjusted automatically. In case the data are valid, you return to the *ConfirmatoryANOVA.exe* form (see Figure 12.1).

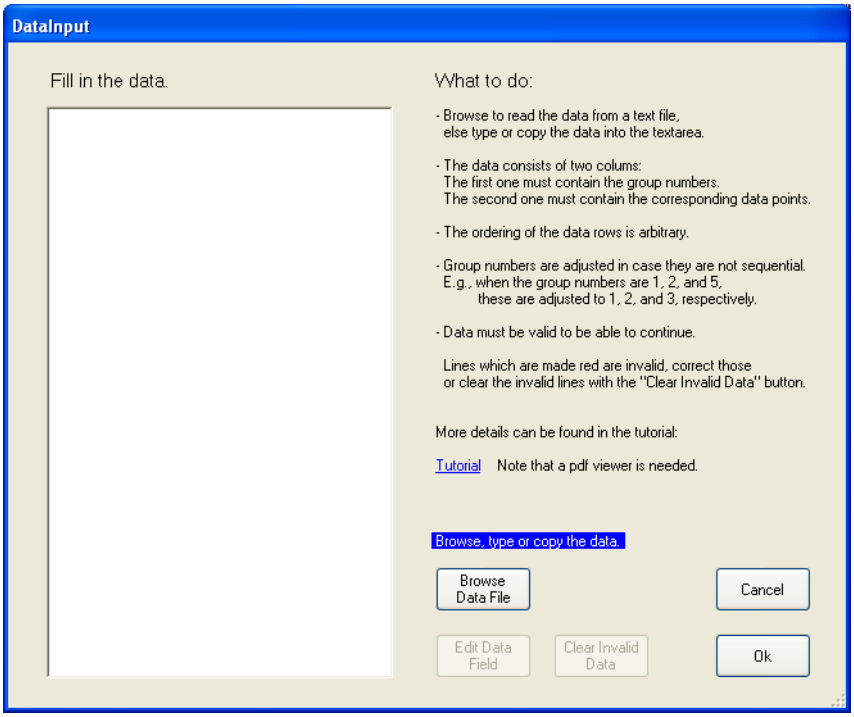


Fig. 12.2: *DataInput* form.

12.B.2 Specify methods

From the data (and group membership), the number of groups and number of observations per group are determined. Then, you must denote which analyses should be performed. A certain analysis will be performed if the corresponding checkbox is checked.

In case BMS is performed, two additional specifications need to be made (in a popup-panel): The desired δ and the prior vagueness pv. It holds that $\delta \geq 0$ and $pv > 0$. When specifying exact equality restrictions in BMS, δ must be set to $\delta = 0$. When specifying about equality restrictions in BMS, δ must be set to $\delta > 0$. In the latter case, one should carefully specify the restrictions (see Section 12.A.4). The default recommendation of pv is $pv = 2$.

For all three methods, the hypotheses of interest must be given explicitly (in the then appearing panel). Specifying the order-restricted hypotheses will be explained in

the next section. When you do not want to specify any order-restricted hypothesis, you should uncheck the *Add (order-restricted) hypothesis/-es* checkbox (see Figure 12.1). One should also specify whether one wants to evaluate the classical null hypothesis (i.e., $H_0 : \mu_1 = \ldots = \mu_k$) and the classical alternative (i.e., $H_a : \mu_1, \ldots, \mu_k$), here also called the unconstrained hypothesis (see Figure 12.1).

As discussed in Section 12.A.5, the values for the seed value and the number of iterations can be specified. This is done in the *Settings* form (see Figure 12.3) appearing when pressing the *Settings* button.

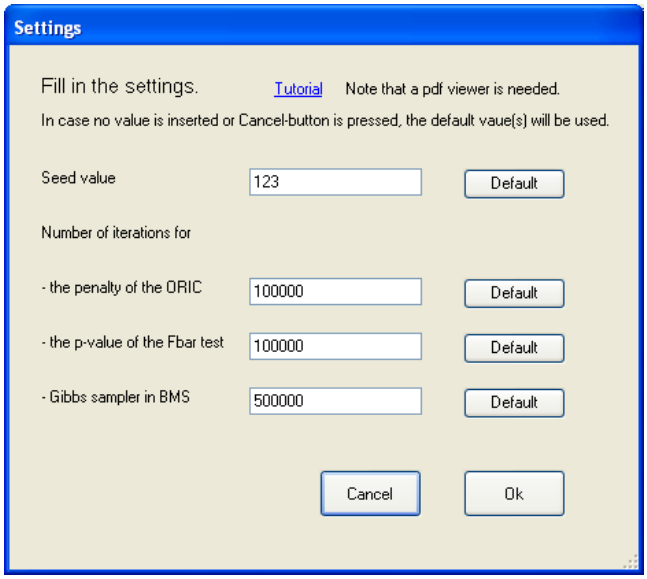


Fig. 12.3: *Settings* form.

More details on specifying the restrictions are given next.

12.B.3 Specifying the order-restricted hypotheses

An (order-restricted) hypothesis, say H_1, can be specified in the panel appearing when clicking on the *Add* button in the *Edit Hypothesis 1* panel (see Figure 12.4). Then fill in the hypothesis, that is, fill in the group numbers and the corresponding constraints between the means of these group. Note that a hypothesis can consist of multiple 'restrictions'. To add another restriction in, say H_1, you must press the *Add* button in the *Edit restriction(s) in H1* panel (see Figure 12.4).

As mentioned before, the numbers in a certain restriction always consists of the numbers 1 to "the number of groups" (in the example, 5) and each number is used precisely once. It should be noted that in case of 5 means and you want to evaluate $H_1 : \mu_3 > \mu_1 > \mu_4$, you should fill in $H_1 : \mu_3 > \mu_1 > \mu_4, \mu_2, \mu_5$ or, better, "$3 > 1 > 4, 2, 5$". If the entry is a number greater than "the number of groups" or the entry is not an integer, then the corresponding textbox will be made red and

an error message is given in the *ConfirmatoryANOVA.exe* form. The check on using every number only once is done after pressing the *Run* button. In that case, an error message will be given in the *Progress Report* panel and in *Error.txt* (see also the next subsection).

One can also specify the classical H_0 and/or the classical H_a as a hypothesis of interest. Note that one should include H_0 only when there is real interest in H_0 (Kuiper & Hoijtink, 2010). We recommend to include H_a (when doing model selection) as a safeguard for choosing a weak hypotheses (Kuiper & Hoijtink, 2010). H_0 and H_a do not need to be specified explicitly, one can just check the corresponding checkboxes (see Figure 12.1 and Figure 12.4).

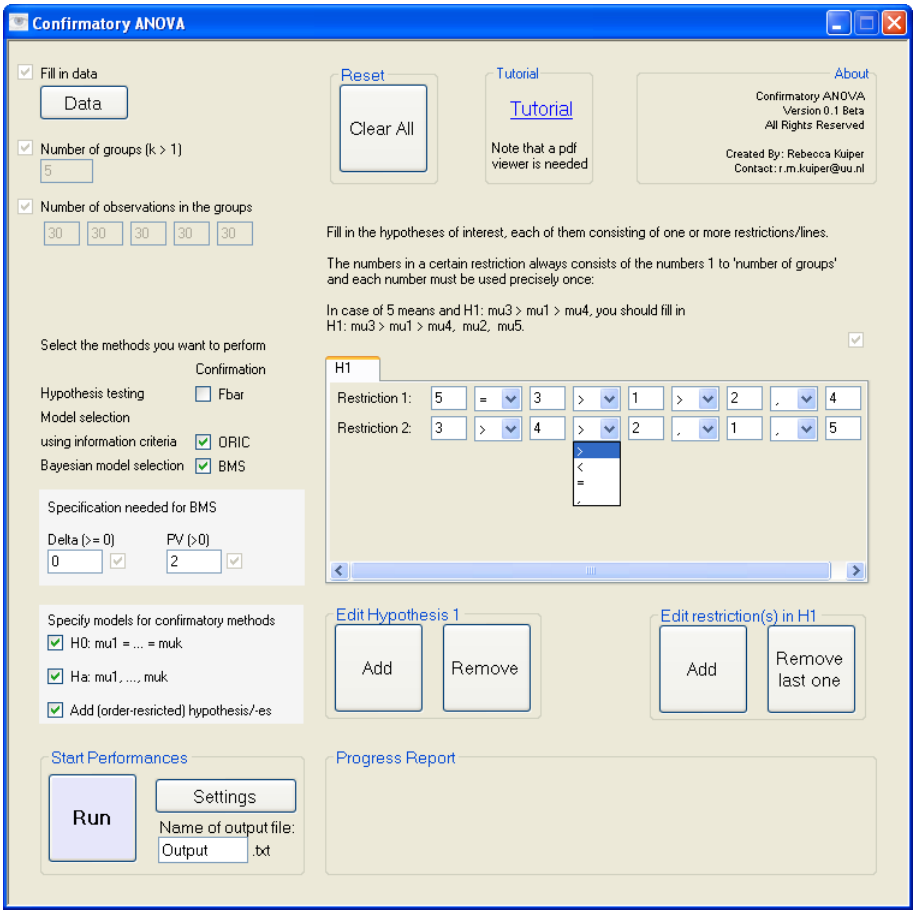


Fig. 12.4: Specifying H_1.

12.B.4 Error messages

Before pressing the *Run* button, one should check whether the requirements are met. When not all requirements are met and you press the *Run* button, a popup appears with the text "Not all requirements are met yet.". The requirements are that the data are valid, the specification of δ and pv (if needed) is correct, and that the hypotheses (if any) are specified correctly. In case a requirement is met, the checkbox is checked; otherwise, it is not checked and an error message is given in the *ConfirmatoryANOVA.exe* form.

Furthermore, error messages are built in to detect other wrongly reported input (e.g., when not using every number precisely once in a (order) restriction). A popup will appear with the text "Error in input. ... This run will be stopped.". These error messages will be given in the *Progress Report* panel in the *ConfirmatoryANOVA.exe* form (see Figure 12.1 and Figure 12.4) and in *Error.txt* (which will be created in the folder where *ConfirmatoryANOVA.exe* is saved in). In most cases, the methods will not be performed and no output will be given.

When, for some reason, the error is not detected, a popup will appear with the text "The program has stopped. No methods are performed. Check input. ...". In that case, no output will be given. One should look at the input again, especially the input needed for the method during which the error occurred. When looking at the *Progress Report* panel one can obtain a better idea of during which method the error has occurred. It should be noted that the methods are performed in a fixed order: the $\bar{F}$ test is performed first, then the ORIC and the program ends with BMS (when all methods are performed). The progress report tells you when the method has started (e.g., "The Fbar test is running...") and when it is ended (e.g., "Fbar is performed").

12.B.5 Output

Let the hypotheses of interest be the set of hypotheses specified in (12.2). As mentioned before, in the output, these hypotheses will be referred to as H_1, H_3, H_4, and H_2 corresponding to H_0, H_1, H_2, and H_3, respectively, in (12.2).

When *ConfirmatoryANOVA.exe* is done, the output file *Output.txt* (or the name you entered in the *Name of output file* textbox) will be created in the folder where *ConfirmatoryANOVA.exe* is saved in. The output file gives the results of the requested analyses. In the example, the output for the three methods (i.e., $\bar{F}$, ORIC, and BMS), with $\delta = 0.3$ and $pv = 2$, is:

This program is free of use. However, when results obtained with this program are published, please refer to:

Rebecca M. Kuiper, Irene Klugkist, and Herbert Hoijtink (2010).
A Fortran 90 Program for Confirmatory Analysis of Variance.
Journal of Statistical Software, 34(8), 1-31.
URL http://www.jstatsoft.org/v34/i08/.

N.B. This paper is available upon request (R.M.Kuiper@uu.nl).

Summary of observed data

Group number, means, standard deviations, and sample sizes per group

1	**2.33**	**1.86**	**30**
2	**1.33**	**1.15**	**30**
3	**3.20**	**1.79**	**30**
4	**2.23**	**1.45**	**30**
5	**3.23**	**1.50**	**30**

Restricted means

Group number: 1 2 3 4 5
Sample means: **2.33 1.33 3.20 2.23 3.23**

Hypothesis 1	**2.46**	**2.46**	**2.46**	**2.46**	**2.46**
Hypothesis 2	**2.33**	**1.33**	**3.20**	**2.23**	**3.23**
Hypothesis 3	**2.33**	**1.33**	**3.21**	**2.23**	**3.21**
Hypothesis 4	**2.60**	**1.33**	**3.20**	**2.60**	**2.60**

The hypotheses of interest are stated below.

– Fbar test –
<*See Section 12.2.1*>

Results of the Fbar test for the null hypothesis 1 and the unconstrained hypothesis 2

Hypotheses numbers	Fbar value	p-value
1 versus 2	**30.27**	**0.00**

Results of the ordered alternative Fbar test

Ordered-hypothesis number	Fbar value	p-value
H0 versus 3	**30.26**	**0.00**
H0 versus 4	**22.91**	**0.00**

Results of the ordered null Fbar test

Ordered-hypothesis number	Fbar value	p-value
3 versus Hu	**0.01**	**1.00**
4 versus Hu	**7.36**	**0.07**

Residual sum of squares

Hypothesis 1	**432.53**
Hypothesis 2	**0.00**
Hypothesis 3	**357.85**
Hypothesis 4	**375.99**

The hypotheses of interest are stated below.

– ORIC –
<*See Section 12.2.2*>

The value of the Order-Restricted Information Criterion (ORIC) =
-2 * log likelihood + 2 * penalty:

for Hypothesis 1, ORIC = -2 * **-292.27** + 2 * **2.00** = **588.54**
for Hypothesis 2, ORIC = -2 * **-278.05** + 2 * **6.00** = **568.10**
for Hypothesis 3, ORIC = -2 * **-278.05** + 2 * **3.19** = **562.49**
for Hypothesis 4, ORIC = -2 * **-281.76** + 2 * **3.14** = **569.79**

The preferred hypothesis, according to the Order-Restricted Information Criterion, of the hypotheses to be compared is hypothesis number **3**.

The hypotheses of interest are stated below.

– BMS –
<*See Section 12.2.3*>

The resulting Bayes factor values (of the order-restricted hypothesis and the posterior model probabilities (with respect to the whole set of models) are:

Hypothesis 1	**0.00**	**0.00**
Hypothesis 2	**1.00**	**0.01**
Hypothesis 3	**67.94**	**0.96**
Hypothesis 4	**1.52**	**0.02**

The preferred hypothesis, according to Bayesian model selection,
of the hypotheses to be compared is hypothesis number **3**.

The hypotheses of interest are stated below.

Specification of the encompassing prior:

For all means, the same normal prior with mean
2.28
and variance
2.32
is used.

For the residual variance, a scaled inverse chi-square with
degrees of freedom
1.00
and scale parameter
2.50
is used.

– The hypotheses of interest –

Hypothesis 1 (= 'H0')

Restriction 1: mu1 = mu2 = mu3 = mu4 = mu5

Hypothesis 2 (= 'Ha')

Restriction 1: mu1 , mu2 , mu3 , mu4 , mu5

Hypothesis 3

Restriction 1: mu5 = mu3 > mu1 > mu2 , mu4
Restriction 2: mu3 > mu4 > mu2 , mu1 , mu5

Hypothesis 4

Restriction 1: mu3 > mu1 > mu4 = mu5 > mu2

CHAPTER *13*

A Fortran 90 Program
for the Generalized Order-Restricted Information Criterion

Kuiper, R. M., and Hoijtink, H.

Manuscript submitted.

The generalized order-restricted information criterion (GORIC) can evaluate hypotheses that are closed convex cones for multivariate normal linear models. It can examine the traditional hypotheses $H_0 : \beta_{1,1} = \cdots = \beta_{t,k}$ and $H_u : \beta_{1,1}, \ldots, \beta_{t,k}$ and hypotheses containing simple order restrictions $H_m : \beta_{1,1} \geq \ldots \geq \beta_{t,k}$, where any "$\geq$" may be replaced by "=", $\beta_{h,j}$ denotes a parameter for the hth dependent variable and the jth predictor in a t-variate regression model with k predictors (which might include the intercept), and m is the model/hypothesis index. But, the GORIC can also be applied to restrictions of the form $H_m : R_1\beta \geq r_1, R_2\beta = r_2$, with β a vector of length tk, R_1 a $c_{m1} \times tk$ matrix, r_1 a vector of length c_{m1}, R_2 a $c_{m2} \times tk$ matrix, and r_2 a vector of length c_{m2}. It should be noted that $[R_1', R_2']'$ should be of full rank when $[r_1', r_2']' \neq 0$. A Fortran 90 program is presented, which enables researchers to compute the GORIC for hypotheses in the context of multivariate regression models.

13.1 Introduction

Anraku (1999) proposes the order-restricted information criterion, ORIC. The ORIC is applied to models of the form $y_{ij} = \beta_j + \epsilon_{ij}$, where y_{ij} is observation i (with $i = 1, \ldots, N_j$) for group j (with $j = 1, \ldots, k$), β_j is the mean for group j, and ϵ_{ij} is the error term, which follows a normal distribution with mean 0 and variance σ^2. This model selection criterion can only be used to select the best of a set of hypotheses that can be written as simple order restrictions (e.g., $H_1 : \beta_1 \geq \ldots \geq \beta_k$ and $H_2 : \beta_1 = \ldots = \beta_{k'} \geq \ldots \geq \beta_k$). Kuiper et al. (2011) propose a generalization of the ORIC, called the GORIC, that can be applied to a more general form of order restrictions, namely $H_m : R\beta \geq 0$ for $m \in \mathcal{M}$, where $\mathcal{M}$ is the set of hypothesis indices, β a vector of length k, and R a $c_m \times k$ matrix. Special cases of these matrix order restrictions are the simple order (i.e., $H_m : \beta_1 \geq \ldots \geq \beta_k$) and the tree order

(i.e., $H_m : \beta_1 \geq \beta_2, \ldots, \beta_1 \geq \beta_k$). Kuiper et al. (unpublished) extend the use of the GORIC to univariate and multivariate normal linear models with hypotheses of the type $H_m : \beta \in \mathcal{C}_m$, where $\mathcal{C}_m$ is a closed convex cone or a relocated one and β is a vector of length tk containing the parameters in a t-variate normal linear model, with k the number of predictors (which can include an intercept) as elaborated below. The hypotheses of interest and therewith the closed convex cones are further discussed in Section 13.2.2.

In the next section, the GORIC will be presented in the context of multivariate regression models. The GORIC comprises a likelihood part and a penalty part. The likelihood is computed using order-restricted maximum likelihood estimators. The iteration process employed to obtain the order-restricted maximum likelihood estimators is described in Section 13.3. In Section 13.4, we will elaborate on the penalty part. Section 13.5 illustrates the application of the GORIC in the context of univariate and multivariate analysis of variance. Appendix 13.A contains a user manual for the software.

13.2 The GORIC

In this section, we provide the GORIC applicable to hypotheses of the form $H_m : \beta \in \mathcal{C}_m$ formulated for a t-variate regression model. The derivation is shown in Kuiper et al. (2011). First, we briefly discuss the t-variate regression model. Then, we give the expression of the GORIC. Finally, we elaborate on the hypotheses that can be evaluated by it.

13.2.1 The t-variate regression model

A multivariate regression model with t dependent variables can be written as

$$\begin{aligned} y_{1i} &= \beta_{1,1}d_{1i} + \ldots + \beta_{1,k'}d_{k'i} + \beta_{1,k'+1}x_{k'+1,i} + \ldots + \beta_{1,k}x_{ki} + \epsilon_{1i} \\ &\vdots \\ y_{ti} &= \beta_{t,1}d_{1i} + \ldots + \beta_{t,k'}d_{k'i} + \beta_{t,k'+1}x_{k'+1,i} + \ldots + \beta_{t,k}x_{ki} + \epsilon_{ti} \end{aligned} \tag{13.1}$$

where y_{hi} denotes the score of the ith person on the hth dependent variable for $i = 1, \ldots, N$ and $h = 1, \ldots, t$. The d variables are the predictors that represents group membership. When $d_{ji} = 1$, person i belongs to group j for $j = 1, \ldots, k'$. The mean of dependent variable h of group j (conditional upon the x variables) is denoted by $\beta_{h,j}$. The x variable are continuous predictors, where x_{ji} reflects the score of the ith person on the jth predictor for $j = k' + 1, \ldots, k$. The relationship between x_{ji} and y_{hi} (controlled for the other predictors) is denoted by $\beta_{h,j}$. Finally, it is assumed that

$$\begin{bmatrix} \epsilon_{1i} \\ \vdots \\ \epsilon_{ti} \end{bmatrix} \sim \mathcal{N}_t \left(\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1t} \\ \vdots & \ddots & \vdots \\ \sigma_{1t} & \cdots & \sigma_t^2 \end{bmatrix} \right).$$

It is noteworthy that the βs associated with x variables regarding the same dependent variable are only comparable when the corresponding x variables are

standardized. Moreover, βs associated with x variables belonging to different dependent variables can solely be examined if both the dependent variables and the x variables are standardized.

13.2.2 The hypotheses of interest

Let $\beta = (\beta_{1,1}, \ldots, \beta_{1,k}, \ldots, \beta_{t,1}, \ldots, \beta_{t,k})$ and β_l the lth element of β for $l = 1, \ldots, tk$. The GORIC can be applied to hypotheses that are closed convex cones or relocated ones; both denoted by $\mathcal{C}_m$. In this chapter, we will focus on

$$H_m: \; R_1\beta \geq r_1, R_2\beta = r_2, \tag{13.2}$$

where R_1 is a $c_{m1} \times tk$ matrix, R_2 a $c_{m2} \times tk$ matrix, r_1 a vector of length c_{m1}, and r_2 a vector of length c_{m2}. For closed convex cones it holds true that $r_1 = r_2 = 0$. Special cases of closed convex cones are the simple order, the tree order, and the matrix order (Silvapulle & Sen, 2005, pp. 82). In case of a relocated closed convex cone, that is, for $[r_1', r_2']' \neq 0$, a requirement is needed (see Kuiper et al. (2011) and Section 13.4): $R = [R_1', R_2']'$ is of full rank. Note that full rank of R may be obtained by discarding redundant restrictions. For example, a set of restrictions containing $\beta_l \geq r_{11}, \beta_l \leq r_{12}$ is not a relocated closed convex cone for $r_{11} \neq r_{12}$, since R is not of full rank and there are no redundant restrictions. For $\beta_l \geq r_{11}, \beta_{l'} \geq r_{12}, \beta_l + \beta_{l'} \geq r_{13}$ for $l \neq l'$, R is not of full rank either. However, when $r_{11} + r_{12} \geq r_{13}$, $\beta_l + \beta_{l'} \geq r_{13}$ is redundant. In case this redundant restriction is discarded, R is of full rank, that is, $H_m : \beta_l \geq r_{11}, \beta_{l'} \geq r_{12}$ is a relocated closed convex cone.

13.2.3 The GORIC

Let

$$\begin{aligned}
Y &= \begin{bmatrix} y_{11}, \ldots, y_{t1} \\ \vdots \quad\quad \vdots \\ y_{1N}, \ldots, y_{tN} \end{bmatrix}, \\
y_i &= [y_{1i}, \ldots, y_{ti}]', \\
X &= \begin{bmatrix} d_{11}, \ldots, d_{k'1}, x_{k'+1,1}, \ldots, x_{k1} \\ \vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \vdots \\ d_{1n}, \ldots, d_{k'n}, x_{k'+1,n}, \ldots, x_{kn} \end{bmatrix}, \\
x_i &= [d_{1i}, \ldots, d_{k'i}, x_{k'+1,i}, \ldots, x_{ki}]', \\
B &= \begin{bmatrix} \beta_{1,1}, \ldots, \beta_{t,1} \\ \vdots \quad\quad \vdots \\ \beta_{1,k}, \ldots, \beta_{t,k} \end{bmatrix}.
\end{aligned} \tag{13.3}$$

According to Kuiper et al. (unpublished), it holds true for t-variate regression models with $H_m : \beta \in \mathcal{C}_m$ that

$$GORIC_m = -2\ \log f(Y|X, \tilde{B}^m, \tilde{\Sigma}^m) + 2\ PT_m, \tag{13.4}$$

with

$$\log f(Y|X, \tilde{B}^m, \tilde{\Sigma}^m) = -\frac{tN}{2}\ \log\{2\pi\} - \frac{N}{2}\log|\tilde{\Sigma}^m| - \frac{1}{2}\sum_{i=1}^{N}\epsilon_i'\left(\tilde{\Sigma}^m\right)^{-1}\epsilon_i,$$

and

$$PT_m = 1 + \sum_{l=1}^{tk} w_l(tk, W, \mathcal{C}_m)\ l,$$

where $\log f(Y|X, \tilde{B}^m, \tilde{\Sigma}^m)$ is the log-likelihood, $\tilde{B}^m$ and $\tilde{\Sigma}^m$ are the order-restricted maximum likelihood estimators of B and Σ, respectively, PT_m is the penalty part, $w_l(tk, W, \mathcal{C}_m)$ denotes the level probability for level l, and

$$\begin{aligned} \epsilon_i &= y_i - (\tilde{B}^m)'x_i, \\ W &= \hat{\Sigma} \otimes [X'X]^{-1}, \end{aligned} \tag{13.5}$$

with

$$\hat{\Sigma} = N^{-1}(Y - X\hat{B})'(Y - X\hat{B}) \tag{13.6}$$

and

$$\hat{B} = (X'X)^{-1}X'Y.$$

Hence, $\hat{\Sigma}$ and $\hat{B}$ are the (unrestricted) maximum likelihood estimators of Σ and B, respectively. The derivation of the penalty can be found in Kuiper et al. (unpublished). In that, Σ is assumed to be known up to a positive constant, that is, $\Sigma = \sigma^2 S$ with S a known $t \times t$ matrix and σ^2 a constant which represents the variance when $t = 1$. Since Σ is often not known, it is estimated by $\hat{\Sigma}$, see Equation (13.6). The GORIC is easily applied, namely the hypothesis/model H_m (see Equation (13.2)) with the lowest GORIC value (see Equation (13.4)) is the preferred one.

In the next two sections, we will subsequently elaborate upon the order-restricted maximum likelihood estimators $\tilde{B}^m$ and $\tilde{\Sigma}^m$ and the penalty term PT_m.

13.3 Order-Restricted maximum likelihood estimators

The order-restricted maximum likelihood estimators, $\tilde{B}^m$ and $\tilde{\Sigma}^m$, are obtained by

$$\min_{\beta \in H_m, \Sigma} \sum_{i=1}^{N} (y_i - (\tilde{B}^m)'x_i)'\Sigma^{-1}(y_i - (\tilde{B}^m)'x_i).$$

From this it follows that

$$\tilde{B}^m = \arg\ \min_{\beta \in H_m} \sum_{i=1}^{N} (y_i - B'x_i)'\left(\tilde{\Sigma}^m\right)^{-1}(y_i - B'x_i), \tag{13.7}$$

$$\tilde{\Sigma}^m = N^{-1}(Y - X\tilde{B}^m)'(Y - X\tilde{B}^m). \tag{13.8}$$

It should be stressed that in univariate regression (i.e., for $t = 1$) the β parameters do not depend on $\tilde{\Sigma}^m = \tilde{\sigma}^2_m$. In multivariate regression (i.i., $t > 1$), $\tilde{B}^m$ depends on

the unknown $\tilde{\Sigma}^m$ and $\tilde{\Sigma}^m$ on the unknown $\tilde{B}^m$. Therefore, iterations are required to calculate them. The iteration process comprises the following steps:

1. Set $\tilde{B}_0^m$ equal to $\hat{B} = (X'X)^{-1}X'Y$, the (unrestricted) maximum likelihood estimator of B. Note that any value for $\tilde{B}_0^m$ can be chosen. We employ $\hat{B}$ to increase the speed of convergence and, therefore, to reduce computing time.
2. Optimize $\tilde{\Sigma}_p^m$ by substituting $\tilde{B}^m$ for $\tilde{B}_{p-1}^m$ in Equation (13.8) for $p = 1, \ldots, P$.
3. Optimize $\tilde{B}_p^m$ by replacing $\tilde{\Sigma}^m$ with $\tilde{\Sigma}_p^m$ in Equation (13.7) for $p = 1, \ldots, P$. For the calculation of $\tilde{B}^m$, one can use a quadratic programming algorithm like the IMSL subroutine QPROG (Visual Numerics, 2003, pp. 1307–1310) in Fortran 90.
4. Continue steps 2 and 3 until convergence is reached (at step P) and set $\tilde{B}^m$ and $\tilde{\Sigma}^m$ equal to $\tilde{B}_P^m$ and $\tilde{\Sigma}_P^m$, respectively. We base the convergence criterion on the values of the parameter estimates. Namely, we stop iterating when the absolute values of the elements of $\tilde{B}_p^m - \tilde{B}_{p-1}^m$ and $\tilde{\Sigma}_p^m - \tilde{\Sigma}_{p-1}^m$ are less than $1e-10$.

13.4 The penalty part

In this section, we elaborate on the calculation of the penalty term. First, we briefly describe level probabilities in case Σ is known up to a positive constant. In that case, $\hat{\Sigma}$ in Equation (13.5) is replaced by Σ. Moreover, we give an interpretation of the penalty term. Then, we discuss the consequences of estimating Σ from the data by $\hat{\Sigma}$.

A level probability $w_l(tk, W, \mathcal{C}_m)$ is the probability that there are l levels among the tk order-restricted maximum likelihood estimators, which are in accordance with $\mathcal{C}_m$, given that the parameters β are generated from a normal distribution with a mean vector of zeros and covariance matrix W (see also Anraku (1999); Silvapulle and Sen (2005, pp. 77–83); Robertson et al. (1988, pp. 69)). Stated otherwise, a level probability is the probability that the parameter space in accordance with the active constraints in $\mathcal{C}_m$ is of dimension l.

According to Kuiper et al. (unpublished), all closed convex cones ($r_1 = r_2 = 0$) and relocated ones ($r = [r_1', r_2']' \neq 0$) can be written in the form $H_m : R_1\beta^* \geq 0, R_2\beta^* = 0$, with $\beta^* = \beta$ when $r_1 = r_2 = 0$ and $\beta^* = \beta - q$ and $[R_1', R_2']'q = r$ when $r \neq 0$. Note that q only exist when $[R_1', R_2']'$ is of full rank (after discarding redundant restrictions). Let $\mathcal{C}_m = \{\beta \in R^{tk} : R_1\beta^* \geq 0, R_2\beta^* = 0\}$.

Below, we first assume that Σ is known up to the positive constant σ^2: $\Sigma = \sigma^2 S$ with S a known matrix. After that, we discuss the consequences of Σ being estimated from the data. The calculation of the level probabilities can be done via simulation (Silvapulle & Sen, 2005, pp. 78–81). The simulation consists of 5 steps:

1. Generate z (of length tk) from $\mathcal{N}_{tk}(\beta^0 = 0, W)$, with $W = \sigma^2 S \otimes [X'X]^{-1}$, where S is a known matrix. Silvapulle and Sen (2005, pp. 86) and Robertson et al. (1988, p. 69) prove that the calculation of the level probabilities does not depend on the mean value β^0 for closed convex cones. Furthermore, Robertson et al. (1988, p. 69) demonstrate for closed convex cones that the calculation of the level probabilities are invariant for positive constants like σ^2 and N. However, there is one exception, which is discussed below.

2. Compute $\tilde{z}_m$ via $\tilde{z}_m = \arg\ \min_{\beta^* \in \{\beta^* \in R^{tk}: R_1\beta^* \geq 0, R_2\beta^* = 0\}} (z - \beta^*)' W^{-1} (z - \beta^*)$, such that the parameters are in accordance with $R_1\beta^* \geq 0, R_2\beta^* = 0$, the hypothesis of interest.
   To implement this in software, one requires a quadratic programming algorithm. For example, one can use the IMSL subroutine QPROG (Visual Numerics, 2003, pp. 1307–1310) in Fortran 90.
3. Determine the number of levels in $\tilde{z}_m$ and denote this by L_m. Let restriction a be denoted by $R_{1a}\beta^* \geq 0$ for $a = 1, \ldots, c_{m1}$, $A = \{a : R_{1a}\tilde{z}_m = 0\}$, that is, the set of restriction indices for which the restriction is binding, and $\phi = \{\beta : R_{1a}\beta^* = 0\ \forall\ a \in A, R_2\beta^* = 0\}$. Then, L_m is the dimension of ϕ.
4. Repeat the previous steps T (e.g., $T = 100,000$) times. To examine the stability of the penalty term, one could calculate it a second time with another seed value. If the two penalties are dissimilar, one should increase the value of T.
5. Estimate the level probability $w_l(tk, W, \mathcal{C}_m)$ by the proportion of times L_m is equal to l $(l = 1, \ldots, tk)$ in the T simulations.

As discussed in the first simulation step, the level probabilities are invariant for the mean value β^0 and the variance term σ^2. This holds almost always true for closed convex cones $H_m : R_1\beta \geq 0, R_2\beta = 0$ and relocated ones $H_m : R_1\beta \geq r_1, R_2\beta \geq r_2$ where $[r_1', r_2']' \neq 0$ and $[R_1', R_2']'$ is of full rank after discarding redundant restrictions. There is one exception, namely restrictions of the type $\beta_l \geq r_{11}$ (including $r_{11} = 0$) for $l = 1, \ldots, tk$. When the hypothesis of interest contains this type of restriction, one must use $\beta^0 = 0$. This results in level probabilities that are invariant for the value of σ^2.

Notably, the level probabilities for $H_m : \beta_l \geq r_{11}$ are the same as for $H_m : \beta_l \geq 0$, that is, here is no difference in complexity for these two hypothesis. When sampling z from $\mathcal{N}_1(0, W)$ with W a scalar, half of the time $H_m : z \geq 0$ is valid and $\tilde{z}_m$ has one level; the other time $H_m : z \geq 0$ will be invalid and $\tilde{z}_m$ has zero levels. As a consequent, the expected dimension of β_l for $H_m : \beta_l \geq r_{11}$ is a half.

The penalty term

$$PT_m = 1 + \sum_{l=1}^{tk} w_l(tk, W, \mathcal{C}_m)\ l$$

can be seen as the expected dimension of the parameters. That is, the expected dimension of β values plus 1 because of the unknown variance term σ^2 in $\Sigma = \sigma^2 S$ with S a known matrix.

Until now, we have assumed in the calculation of the level probabilities that Σ is known up to the constant σ^2. Often Σ is unknown, in that case one should estimate it to determine the level probabilities. However, when $t = 1$, no estimation of $\Sigma = \sigma^2$ is required, since the level probabilities are invariant of positive constants like σ^2 (see Step 1). In contrast, Σ needs to be estimated for $t > 1$. One can estimate Σ by $\hat{\Sigma}$, see Equation (13.6); as is done in the software.

If Σ is estimated from the data, the dimension of Σ, which is the number of unknown distinct elements of Σ, is $(t + 1)t/2$ instead of 1. Since the restrictions are always on the β parameters and never on the elements of Σ, the number of unknown

distinct elements is equal for all hypotheses of interest (H_m). So, although the penalty should then (perhaps) be corrected, the correction is equal for all H_m for $m \in \mathcal{M}$.

In the next section, we will demonstrate evaluating hypotheses with the GORIC for different types of models.

13.5 The GORIC illustrated

13.5.1 Analysis of variance (ANOVA)

In this section, we will illustrate the GORIC supported by real data for which the descriptive statistics are available in Lievens and Sanchez (2007). They investigated the effect of training on the quality of ratings made by consultants. One variable of interest is the signal detection accuracy index, which "refers to the extent to which individuals were accurate in discerning essential from nonessential competencies for a given job" and is measured by "standardized proportion of hits - standardized proportion of false alarms" (Lievens & Sanchez, 2007, p. 817). Three groups of consultants are distinguished: 1) expert, 2) training, and 3) control. There are 21 raters in the expert group, 25 in the training group, and 26 in the control group. Hence, the ANOVA model can be written as Equation (13.1) with $t = 1$, $k' = 3$, and $N = \sum_{j=1}^{k} n_j = 21 + 25 + 26 = 72$, where d_1, d_2, and d_3 denote group membership variables. Since $t = 1$, we will drop the first subscript in the index for ease of notation and use β_j instead of $\beta_{1,j}$. Note that for $t = 1$ no iteration is required between $\tilde{B}^m$ and $\tilde{\Sigma}^m$ (see Section 13.3), and that Σ does not need to be estimated to calculate the level probabilities (see Section 13.4).

The authors expected that accuracy of competency ratings would be higher among experts and trained raters than among raters in the control group (i.e., $\beta_1 \geq \beta_3$ and $\beta_2 \geq \beta_3$) and furthermore, that it would be highest among raters who already had competency modeling experience (i.e., $\beta_1 \geq \beta_2$). These expectations can be represented by the hypothesis $H_1 : \beta_1 \geq \beta_2 \geq \beta_3$. Another theory could be that the accuracy of the training group is at least twice as high as the one in the control group and that that of the export group is higher than that of the training group. This leads to $H_2 : \beta_1 \geq \beta_2 \geq 2\ \beta_3$. Since both can be bad/weak hypotheses, it is informative to evaluate the unconstrained hypothesis (H_u) as well, in which there are no restrictions on the parameters. Namely, its inclusion ensures that no weak hypothesis is selected, since H_u will be preferred if the other two hypotheses are weak / do not fit the data. The set of hypotheses, therefore, consists of

$$
\begin{aligned}
H_1 &: \ \beta_1 \geq \beta_2 \geq \beta_3, \\
H_2 &: \ \beta_1 \geq \beta_2 \geq 2\ \beta_3, \\
H_u &: \ \beta_1,\ \beta_2,\ \beta_3.
\end{aligned}
$$

Table 13.1 displays the order-restricted means $\tilde{\beta}_j^m$ (Equation (13.7)), the log likelihood values $\log f(Y|X, \tilde{B}^m, \tilde{\Sigma}^m)$, the penalty terms PT_m, and the GORIC values (Equation (13.4)), for the three hypotheses of interest. Since the sample means are in accordance with the restrictions in all the three hypotheses, the order-restricted means

of these hypotheses are equal to the sample means. Therefore, the three hypotheses have the same log likelihood and the distinction between the three is based on the penalty, that is, the complexity of the hypotheses. Since H_1 is less complex than H_2 and H_u (i.e, $PT_1 < PT_2$ and $PT_1 < PT_u$), H_1 is the preferred hypothesis. As a result, the first theory is preferred over the second and it is not a weak theory.

m	$\tilde{\beta}_1^m$	$\tilde{\beta}_2^m$	$\tilde{\beta}_3^m$	$\log f(Y\|X, \tilde{B}^m, \tilde{\Sigma}^m)$	PT_m	$GORIC_m$
1	0.79	0.64	0.29	-24.85	2.84	**55.38**
2	0.79	0.64	0.29	-24.85	2.90	55.50
u	0.79	0.64	0.29	-24.85	4.00	57.70

Note. Bolding indicates the lowest value.

Table 13.1: GORIC of the three specified hypotheses.

13.5.2 Multivariate analysis of variance (MANOVA)

In this section, we will illustrate the GORIC supported by real data which are available on page 10 of Silvapulle and Sen (2005) and in a report prepared by Litton Bionetics Inc in 1984. These data were used in an experiment to find out whether vinylidene fluoride gives rise to liver damage. Since increased levels of serum enzyme are inherent in liver damage, the focus is on whether enzyme levels are affected by vinylidene fluoride.

Hence, the variable of interest is the serum enzyme level. Three types of enzymes are inspected, namely SDH, SGOT, and SGPT. To study whether vinylidene fluoride has an influence on the three serum enzymes, four dosages of this substance are examined. In each of these four treatment groups, ten male Fischer-344 rats received the substance. The ANOVA model can be written as Equation (13.1) with $t = 3$, $k' = 4$, and $N = 10$. Hence, $(y_{1i}, y_{2i}, y_{3i})'$ denotes the observations on the three enzymes for rat i, d_1 to d_4 are the group membership variables, and $\beta_{h,j}$ denote the mean response for dose j and dependent variable h.

If vinylidene fluoride induces liver damage, we expect that each serum level increases with the dosage of the substance, see H_1 below. Another theory could be that there is no effect of dosage, see H_0 below. Since both can be bad/weak hypotheses, it is informative to evaluate the unconstrained hypothesis (H_u) in which there are no restrictions on the parameters. The set of hypotheses, therefore, comprises

$$
\begin{aligned}
H_0 &: \ \beta_{h,1} = \beta_{h,2} = \beta_{h,3} = \beta_{h,4} \text{ for all } h = 1, 2, 3, \\
H_1 &: \ \beta_{h,1} \geq \beta_{h,2} \geq \beta_{h,3} \geq \beta_{h,4} \text{ for all } h = 1, 2, 3, \\
H_u &: \ \beta_{h,1}, \ \beta_{h,2}, \ \beta_{h,3}, \ \beta_{h,4} \text{ for all } h = 1, 2, 3.
\end{aligned}
$$

Note that there are twelve parameters in total.

Since the covariance matrix Σ is unknown, it is estimated from the data by the maximum likelihood estimator of Σ:

$$\hat{\Sigma} = \begin{bmatrix} 10.79750 & -0.85750 & -0.07000 \\ -0.85750 & 226.75750 & 21.00500 \\ -0.07000 & 21.00500 & 24.67500 \end{bmatrix}.$$

The formula of $\hat{\Sigma}$ is displayed in Equation (13.6). This estimate, $\hat{\Sigma}$, is used in determining the level probabilities (see Section 13.4).

Table 13.2 displays the order-restricted means $\tilde{\beta}^m_{h,j}$ in Equation (13.7). Furthermore, Table 13.3 presents the log likelihood values ($\log f(Y|X, \tilde{B}^m, \tilde{\Sigma}^m)$), the penalty terms ($PT_m$), and the GORIC values in Equation (13.4), for the three hypotheses of interest. The penalty values for both H_0 and H_1 are low(er), whereas the fit of H_u is high(er). The support in the data for H_u is that much higher that it renders the lowest GORIC value. Therefore, it is concluded that H_u is the preferred hypothesis. Notably, although H_1 is preferred over H_0, H_1 is a weak theory, since it is not preferred over the unconstrained hypothesis H_u.

	SDH				SGOT				SGPT			
m	$\tilde{\beta}^m_{1,1}$	$\tilde{\beta}^m_{1,2}$	$\tilde{\beta}^m_{1,3}$	$\tilde{\beta}^m_{1,4}$	$\tilde{\beta}^m_{2,1}$	$\tilde{\beta}^m_{2,2}$	$\tilde{\beta}^m_{2,3}$	$\tilde{\beta}^m_{2,4}$	$\tilde{\beta}^m_{3,1}$	$\tilde{\beta}^m_{3,2}$	$\tilde{\beta}^m_{3,3}$	$\tilde{\beta}^m_{3,4}$
0	24.13	24.13	24.13	24.13	105.38	105.38	105.38	105.38	59.70	59.70	59.70	59.70
1	24.13	24.13	24.13	24.13	105.37	105.37	105.37	105.37	63.00	63.00	60.64	52.16
u	22.70	22.80	23.70	27.30	99.30	108.40	100.90	112.90	61.90	63.80	60.20	52.90

Table 13.2: The order-restricted means ($\tilde{\beta}^m_{h,j}$) for dependent variable h, predictor j, and Hypothesis H_m.

m	$\log f(Y\|X, \tilde{B}^m, \tilde{\Sigma}^m)$	PT_m	$GORIC_m$
0	-406.54	4.00	821.09
1	-396.85	7.48	808.66
u	-388.80	13.00	**803.61**

Note. Bolding indicates the lowest value.

Table 13.3: The GORIC values of the three specified hypotheses.

13.A *GORIC.exe* user manual

This user manual will describe and illustrate the options available in *GORIC.exe* (published along with this chapter and also available at `http://staff.fss.uu.nl/RMKuiper`). It also includes a directory with the input and output files of the ANOVA and MANOVA example given in this chapter. This program is made in Fortran

90 using the Intel Visual Fortran Compiler 10.0 for Windows. This compiler uses *IMSL* 5.0.

GORIC.exe is free, however, when results obtained with this program are published, please refer to (the article based on) this chapter, Kuiper et al. (2011), and Kuiper et al. (unpublished).

13.A.1 *GORIC.exe*

In the software, the notation differs a bit from the one in Equation (13.1). First, all the d and x variables are combined, resulting in a $N \times k$ matrix X, like in Equation (13.3). Note that a variable of group membership is obtained by filling in ones and zeros at the appropriate places in a predictor/vector. Furthermore, the order of the predictors is not of importance, that is, the group membership variables do not need to come first. In addition, when there are no group variables, one should include an intercept by adding a vector ones in X. Second, the parameters are taken together as well, leading to a vector of tk parameters β with indices 1 to tk. Notably, when $k = 0$, they will be denoted by θ, a vector of t variable / group means. The order of the parameters corresponds to the order of the k predictors and the order of the t dependent variables. Namely, the first k parameters belong to the first dependent variable, $\cdots$, and the last k parameters belong to the last one. Stated differently, $(\beta_1, \ldots, \beta_k, \ldots, \beta_{(t-1)k+1}, \ldots, \beta_{tk})$ corresponds to $\beta = (\beta_{1,1}, \ldots, \beta_{1,k}, \ldots, \beta_{t,1}, \ldots, \beta_{t,k})$. Bear in mind that $\beta_1, \beta_{k+1}, \ldots, \beta_{(t-1)k+1}$ reflect the intercepts when the first column of X consists of ones.

As discussed in Step 4 in Section 13.3, we stop iterating when the absolute values of the elements of $\tilde{B}^m_p - \tilde{B}^m_{p-1}$ and $\tilde{\Sigma}^m_p - \tilde{\Sigma}^m_{p-1}$ are less than $C = 1e-10$. But, to increase computing time, C is lowered to $C = 1e-9$ after $50,000$ iterations and to $C = 1e-8$ after $100,000$ iterations. When still no convergence is achieved after $200,000$ iterations, the program uses the current estimates $\tilde{B}^m_P$ and $\tilde{\Sigma}^m_P$ and displays these estimates together with $\tilde{B}^m_{P-1}$ and $\tilde{\Sigma}^m_{P-1}$ in the dos box and the output file. The consequence of lowering C is that the procedure might not result in good approximations of $\tilde{B}^m$ and $\tilde{\Sigma}^m$. However, slow convergence only occurs when the hypothesis of interest does not fit the data.

13.A.2 Modification input files

No matter what analysis should be performed, two text files have to be modified (such that they apply to your data), namely *Input.txt* and *Data.txt*.

It should be noted that:

- The names of the text files are fixed and cannot be changed. These files have to be text files (also known as ASCII files).
- The format of these files should not be changed, that is, do not add empty lines and do not delete lines containing labels.
- The data in *Data.txt* should be complete, that is, missing data are not allowed.

Data.txt

The file *Data.txt* looks as follows (in the MANOVA example):

```
18 101 65 1 0 0 0
...
27  88 56 1 0 0 0
25 113 65 0 1 0 0
...
27  98 65 0 1 0 0
22  88 54 0 0 1 0
...
21 107 61 0 0 1 0
31 104 57 0 0 0 1
...
29  99 48 0 0 0 1
```

In the data file, a $N \times (t+k)$ matrix must be given. The t dependent variables must be given first, followed by the k predictors. In this example, the predictors only consist of group membership variables, denoted by d in Equation (13.1). In case there are no group membership variables, a vector of ones should be included, which represents the intercept. This can be done by specifying it in the input (see below) or by adding a column of ones to your data file.

It should be stressed that a dot (".") should be used as decimal separator. When a comma (",") is used, only the number proceeding it is read (e.g., "1,9" is read as "1"). Furthermore, text or extra hard returns/enters should not be added to *Data.txt*.

Input.txt

The file *Input.txt* looks as follows (in the MANOVA example):

```
t  k  intercept  N   Stand x  Stand y
3  4  0          40  0        0

Seed  T
123   100000

M
3

Number of Equality (c_2) and Order (c_1) Restrictions for Each Model
 (resulting in M lines with 2 numbers)
9 0
0 9
0 0

R for Model 1
```

```
1 -1 0 0 0 0 0 0 0 0 0  0
...
0  0 0 0 0 0 0 0 0 0 1 -1
R for Model 2
1 -1 0 0 0 0 0 0 0 0 0  0
...
0  0 0 0 0 0 0 0 0 0 1 -1
R for Model 3
r for Model 1
0
...
0
r for Model 2
0
...
0
r for Model 3
```

t, k, and N: t is the number of dependent variable, k the number of predictors, and N the number of observations, see Section 13.2.1; for k see also the item below.

intercept: This should be a 1 if you want the software to incorporate the intercept and a 0 when you do not.

When you want the software to include a vector of ones to the set of predictors, the software will change k into $k + 1$. Consequently, the restrictions should be given for $t(k + 1)$ parameters as opposed to tk. Note that the first parameter (for every dependent variable) will represent the intercept.

When your data (represented by the $N \times k$ matrix X) includes a vector of ones, the number of predictors (k) should include the intercept (see Section 13.2.1). In that case, "intercept" should be set to 0, otherwise the program will fail to continue.

Stand x and Stand y: If you set "Stand x" to 1, the predictors (X) will be standardized. The analogue hods true for "Stand y".

The parameters regarding the same dependent variable are only comparable when the x variables are standardized (see Section 13.2.2). Additionally, the parameters belonging to different dependent variables can solely be examined if both the dependent variables and the corresponding x variables (if any) are standardized.

Seed and T: The seed value is represented by "Seed" and the number of iterations required for computing the penalty part of the GORIC by T. These are discussed in Simulation step 4 in Section 13.4.

M, c_2, and c_1: M denotes the number of models/hypotheses, and $c\_2 = c_2$ and $c\_1 = c_1$ the number of equality and order restrictions, respectively, see Section 13.2.2.

R and r: R is the restriction matrix and equals $[R_2', R_1']'$ and r the right hand side and equals $[r_2', r_1']'$ (see Sections 13.2.2).

The models are of the form $H_m : R_2\beta = r_2, R_1\beta \geq r_1$. It should be stressed that the order of the restrictions are of importance: the c_2 equality restrictions must be given first and the c_1 order restrictions second.

One must give a restriction matrix ($R = [R_2', R_1']'$) and a right hand side ($r =$

$[r_2', r_1']'$) for each model. Hence, you need to fill in M restriction matrices with each a heading and then M right hand side vectors with each a heading. Note that there is only a heading when there are no restrictions, that is, in case of the unconstrained model. Bear in mind that the ordering of the columns in the restriction matrix depend on the ordering of the parameters. In the software, the first k parameters belong to the first dependent variable ($h = 1$), $\cdots$, and the last k to the last dependent variable ($h = t$). Hence, in the example, β_1 corresponds to $\beta_{1,1}$, β_2 to $\beta_{1,2}$, $\cdots$, β_5 to $\beta_{2,1}$, $\cdots$, and β_{12} to $\beta_{3,4}$.

As in *Data.txt*, text or extra hard returns/enters should not be added to *Input.txt*, except for headings for additional models.

13.A.3 Error messages

In the program *GORIC.exe*, error messages are incorporated to detect wrongly stated input. However, it is possible to make a mistake that we have not foreseen. In that case, check the input and compare it to the data. If you cannot solve the problem, send the input and data file to R.M.Kuiper@uu.nl.

The requirement that $R = [R_2', R_1']'$ should be of full rank when $r = [r_2', r_1']' \neq 0$ (see Kuiper et al. (2011) and Section 13.4) is investigated in the software. However, note that R is not examined on redundant restrictions. Therefore, the software does not detect hypotheses that are no relocated closed convex cones. A warning appears when R is not of full rank when $r \neq 0$ and the user is asked to investigate whether the additional restrictions are redundant. By pressing the enter button, the program proceeds. It should be stressed that the program stops without a warning in case of conflicting restrictions (e.g., $H_m : \beta_l \leq -r_{11}, \beta_l \geq r_{11}$ for $r_{11} > 0$). Moreover, the GORIC is calculated in presence of non-redundant restrictions, like range restrictions (e.g., $H_m : \beta_l \geq -r_{11}, \beta_l \leq r_{11}$ for $r_{11} > 0$), which is not a (relocated) closed convex cone. In that case, the GORIC should be interpret with care for two reasons. First, the GORIC is not (yet) defined for these types of restrictions. Second, the level probabilities are now no longer invariant for β^0 and σ^2. In the software, we use $\beta^0 = 0$. As a consequence, $H_m : \beta_l = 0$ is examined in determining the penalty.

13.A.4 Save and close

When you have modified *Input.txt* and *Data.txt* (such that it applies to your data), you should save and close it.

13.A.5 Run *GORIC.exe*

When *GORIC.exe* is completed, the output file *Output.txt* will be created in the folder you are working in.

Output.txt

The output is given in *Output.txt* and will look as follows (in case of the MANOVA example):

```
This program is free. However, when results obtained with this program
  are published, please refer to:

Rebecca M. Kuiper, Herbert Hoijtink, and Mervyn J. Silvapulle (2011).
An Akaike-type Information Criterion for Model Selection under
 Inequality Constraints.
Biometrika, 98 (2), 495-501.

Rebecca M. Kuiper, Herbert Hoijtink, and Mervyn J. Silvapulle
 (unpublished).
Generalization of the Order-Restricted Information Criterion for
 Multivariate Normal Linear Models.

Rebecca M. Kuiper and Herbert Hoijtink (unpublished).
A Fortran 90 Program for the Generalization of the Order-Restricted
 Information Criterion.

N.B. The latter is included in this software and the second is
 available upon request (R.M.Kuiper@uu.nl).

-- Summary of observed data --

- Number of observations (N) -

N =  40

- Sigma estimated from the data -

      h,      estimated Sigma
      1      10.79750      -0.85750      -0.07000
      2      -0.85750     226.75750      21.00500
      3      -0.07000      21.00500      24.67500

- Order-restricted betas -

Note that the first   4 parameters belong to the first dependent
 variable, ..., and the last   4 to the last dependent variable.

Group number: 1     2     3     4     5     6     7     8     9     10    11    12
Sample betas: 22.70 22.80 23.70 27.30  99.30 108.40 100.90 112.90 61.90 63.80 60.20 52.90

Hypothesis 1  24.13 24.13 24.13 24.13 105.38 105.38 105.38 105.38 59.70 59.70 59.70 59.70
Hypothesis 2  24.13 24.13 24.13 24.13 105.37 105.37 105.37 105.37 63.00 63.00 60.64 52.16
Hypothesis 3  22.70 22.80 23.70 27.30  99.30 108.40 100.90 112.90 61.90 63.80 60.20 52.90

-- GORIC --

The value of the Generalized Order-Restricted Information Criterion
 (GORIC) = -2 * log likelihood + 2 * penalty:
```

```
for Hypothesis   1, GORIC = -2 *   -406.54 + 2 *    4.00 =   821.09
for Hypothesis   2, GORIC = -2 *   -396.85 + 2 *    7.48 =   808.66
for Hypothesis   3, GORIC = -2 *   -388.80 + 2 *   13.00 =   803.61

According to the Generalized Order-Restricted Information Criterion,
 out of  the set of hypotheses the preferred one is number   3,
which is the unconstrained model, that is, the model without
 restrictions on the parameters.
```

Number of observations (N): See Section 13.2.1.

Sigma estimated from the data: In the software, Σ is estimated by $\hat{\Sigma}$ (Equation (13.6)), the maximum likelihood estimator of Σ. Bear in mind that Σ is only estimated when $t > 1$.

For more details see Section 13.4.

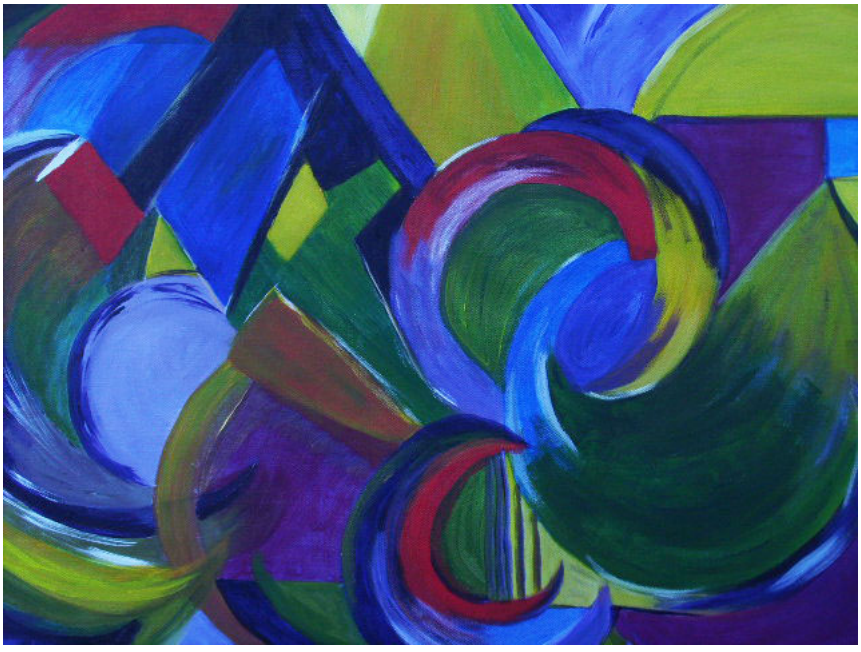
Order-restricted betas: The order-restricted βs can be found in Equation (13.7) see also Sec-tion 13.A.1. Note that the subscripts are 1 to 12 in the software, where $\tilde{\beta}_1^m$ corresponds to $\tilde{\beta}_{1,1}^m$, $\tilde{\beta}_2^m$ to $\tilde{\beta}_{1,2}^m$, $\cdots$, $\tilde{\beta}_5^m$ to $\tilde{\beta}_{2,1}^m$, $\cdots$, and $\tilde{\beta}_{12}^m$ to $\tilde{\beta}_{3,4}^m$ in Equation (13.1).

GORIC: The expression of the GORIC is displayed in Equation (13.4).

The model/hypothesis with the lowest GORIC value is the preferred one: Hypothesis "number 3", that is, $H_u : \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}, \beta_{11}, \beta_{12}$.

Part VI

References, Dutch Summary, Acknowledgments, and About the Author



Vergeten van het weten by Marga Klungel

References

Akaike, H. (1973). Information theory and an etension of the maximum likelihood principle. In B. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control*, *19*, 716–723.

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

Arbuckle, J. L. (2007). *Amos 16.0 [computer software].* Chicago: SPSS. Available from `http://www.smallwaters.com/amos`

Batenburg, R. S., Raub, W., & Snijders, C. (2003). Contacts and contracts: Temporal embeddedness and the contractual behavior of firms. *Research in the Sociology of Organizations*, *20*, 135–88.

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

Berger, J. O., & Mortera, J. (1999). Default Bayes factors for nonnested hypothesis testing. *Journal of the American Statistical Association*, *94*, 542–554.

Berger, J. O., & Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, *91 (433)*, 109–122.

Berger, J. O., & Pericchi, L. R. (2004). Training samples in objective Bayesian model selection. *The Annals of Statistics*, *32 (3)*, 841–869.

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

Berzonsky, W. A., Kleven, S. L., & Leach, G. D. (2003). The effects of parthenogenesis on wheat embryo formation and haploid production with and without maize pollination. *Euphytica*, *133*, 285–290.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: Effect of inequality of variance in the one-way classification. *Annals Mathematical Statistics*, *25*, 290–302.

Box, G. E. P., & Youle, P. V. (1955). The exploration and exploitation of response surfaces: an example of the link between the fitted surface and the basic mechanism of the system. *Biometrics*, *11*, 287–323.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (second ed.). New York:

Springer-Verlag.
Buskens, V., & Raub, W. (2002). Embedded trust: Control and learning. *Advances in Group Processes*, *19*, 167–202.
Buskens, V., & Raub, W. (2010). *Rational choice research on social dilemmas: Embeddedness effects on trust.* ISCORE paper 200, Utrecht University.
Buskens, V., Raub, W., & van der Veer, J. (2010). Trust in triads: An experimental study. *Social Networks*, *32*, 301–312.
Buskens, V., & Weesie, J. (2000). An experiment on the effects of embeddedness in trust situations: Buying a used car. *Rationality and Society*, *12*, 227–53.
Cavanaugh, J. E., & Shumway, R. H. (1998). An Akaike information criterion for model selection in the presence of incomplete data. *Journal of Statistical Planning and Inference*, *67*, 45–65.
Chen, M., & Sungduk, K. (2008). The Bayes factor versus other model selection criteria for the selection of constrained models. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 155–180). New York: Springer.
Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, *90*, 1313–1321.
Claeskens, G., & Consentino, F. (2008). Variable selection with incomplete covariate data. *Biometrics*, *64*, 1062–1096.
Claeskens, G., & Hjort, N. (2008). Minimising average risk in regression models. *Econometric Theory*, *24*, 493–527.
Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.
Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (second ed.). New York: Sage Foundation.
Cribbie, R. A., & Keselman, H. J. (2000). Annual meeting of the american educational research association, 2000, New Orleans. In *A power comparison of pairwise multiple comparison procedures: A model testing approach versus stepwise procedures.*
Dayton, C. M. (1998). Information criteria for paired-comparison problem. *American Statistician*, *52*, 144–151.
Dayton, C. M. (2001). Subset: Best subsets using information criteria. *Journal of Statistical Software*, *6*, 1–10.
Dayton, C. M. (2003). Information criteria for pairwise comparisons. *Psychological Methods*, *8*, 61–71.
Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, *39*, 1–38.
Dickey, J. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *Annals of Mathematical Statistics*, *42*, 204–223.
Ferguson, T. S. (1996). *A course in large sample theory.* Chapman & Hall.
Gelfand, A. E., Smith, A. F. M., & Lee, T. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, *87*, 523–532.
Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (second ed.). London: Chapman and Hall.

Gourieroux, C., & Monfort, A. (1995). *Statistics and econometric models* (Vol. 2). Cambridge: Cambridge University Press.

Gupta, V. K., Turban, D. B., & Bhawe, N. M. (2008). The effect of gender stereotype activation on entrepreneurial intentions. *Journal of Applied Psychology*, *93*, 1053–1061.

Hartley, H. O. (1950). The maximum F-ratio as a short-cut test for heterogeneity of variance. *Biometrika*, *37*, 308–312.

Hasel, L. E., & Kassin, S. M. (2009). On the presumption of evidentiary independence: Can confessions corrupt eyewitness identifications? *Psychological Science*, *20*, 122–126.

Hens, N., Aerts, M., & Molenberghs, G. (2006). Model selection for incomplete and design-based samples. *Statistics in Medicine*, *25*, 2502–2520.

Hoijtink, H., Huntjens, R., Reijntjes, A., Kuiper, R., & Boelen, P. A. (2008). An evaluation of Bayesian inequality constrained analysis of variance. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 85–108). New York: Springer.

Horton, N. J., & Lipsitz, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statistician*, *55*, 244–254.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications.* Mahwah: Erlbaum.

Hughes, A. W., & King, M. L. (2003). Model selection using AIC in the presence of one-sided information. *Journal of Statistical Planning and Inference*, *115*, 397–411.

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*, 297–307.

Jamshidian, M. (2004). Strategies for analysis of incomplete data. In M. Hardy & A. Breiman (Eds.), *Handbook of data analysis* (pp. 113 – 130). London: Sage.

Jamshidian, M., & Bentler, P. M. (1999). ML estimation of mean and covariance structures and missing data using complete data routines. *Journal of Educational and Behavioral Statistics*, *24*, 21–41.

Jeffreys, H. (1961). *Theory of probability.* New York: Oxford University Press.

Johnson, V. E. (2005). Bayes factors based on test statistics. *Journal of the Royal Statistical Society: Series B*, *67*, 689–701.

Johnston, J., & DiNardo, J. (1997). *Econometric methods.* New York: McGraw-Hill.

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

Klugkist, I. (2008). Encompassing prior based model selection for inequality constrained analysis of variance. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 53–83). New York: Springer.

Klugkist, I., & Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics and Data Analysis*, *51*, 6367–6379.

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*, 57–69.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*, 477–493.

Kramer, C. Y. (1956). Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics*, *12*, 307–310.

Kramer, C. Y. (1957). Extension of multiple range tests to group correlated adjusted means. *Biometrics*, *13*, 13–18.

Kuiper, R. M., & Hoijtink, H. (2010). Comparisons of means using exploratory and confirmatory approaches. *Psychological Methods*, *15*, 69–86.

Kuiper, R. M., Hoijtink, H., & Silvapulle, M. J. (2011). An Akaike-type information criterion for model selection under inequality constraints. *Biometrika*, *98 (2)*, 495–501.

Kuiper, R. M., Hoijtink, H., & Silvapulle, M. J. (unpublished). Generalization of the order-restricted information criterion for multivariate normal linear models.

Kuiper, R. M., Klugkist, I., & Hoijtink, H. (2010). A Fortran 90 program for confirmatory analysis of variance. *Journal of Statistical Software*, *34(8)*, 1–31.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 79–86.

Lievens, F., & Sanchez, J. I. (2007). Can training improve the quality of inferences made by raters in competence modeling? A quasi-experiment. *Journal of Applied Psychology*, *92*, 812–819.

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

Lipsey, M. W., & Wilson, D. B. (2000). *Practical meta-analysis.* London: Sage Publications.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Liu, M., Wei, L., & Zhang, J. (2006). Review of guidelines and literature for handling missing data in longitudinal clinical trials with case study. *Pharmaceutical Statistics*, *5*, 7–18.

Lucas, J. W. (2003). Status processes and the institutionalization of women as leaders. *American Sociological Review*, *68*, 464–480.

Martin, S. A., Toothaker, L. E., & Nixon, S. J. (1989). A Monte Carlo comparison of multiple comparison procedures under optimal and nonoptimal conditions. In *annual meeting of the southwestern psychological association, Houston.*

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9 (2)*, 147–163.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models.* London: Chapman and Hall.

McCulloch, C. E., & Searle, S. R. (2005). *Generalized, linear, and mixed models.* Hoboken: Wiley.

McQuarrie, A. D. R., & Tsai, C. L. (1998). *Regression and time series model selection.* Singapore: World Scientific Publications.

Miller, R. G., Jr. (1986). *Beyond anova: Basics of applied statistics.* New York: Wiley.

Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: Objective model selection using constrained

posterior priors. *Journal of Statistical Planning and Inference*, *140*, 887–906.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus: Statistical analysis with latent variables: User's guide.* Los Angeles: Muthén & Muthén. Available from `http://www.statmodel.com`

Neath, A. A., & Cavanaugh, J. E. (2006). A Bayesian approach to the multiple comparisons problem. *Journal of Data Science*, *4*, 131–146.

Nomakuchi, K. (2002). A monotonicity of moments concerned with order restricted statistical inference. *Annals of the Institute of Statistical Mathematics*, *54*, 621–625.

Palmer, E. J., & Gough, K. (2007). Childhood experiences of parenting and causal attributions for criminal behavior among young offenders and non-offenders. *Journal of Applied Social Psychology*, *37*, 790–806.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Development Core Team. (2009). nlme: Linear and nonlinear mixed effects models. R package version 3.1-92. Available from `http://CRAN.R-project.org/package=nlme`

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

Raub, W., & Buskens, V. (2008). Theory and empirical research in analytical sociology: The case of cooperation in problematic social situations. *Analyse und Kritik*, *30*, 689–722.

Rencher, A. C. (1995). *Methods of multivariate analysis.* New York: Wiley.

Rice, J. (1995). *Mathematical statistics and data analysis* (second ed.). Duxbury Press.

Robertson, T., Wright, F. T., & Dykstra, R. L. (1988). *Order restricted statistical inference.* Chichester: Wiley.

Rooks, G., Raub, W., Selten, R., & Tazelaar, F. (2000). Cooperation between buyer and supplier: Effects of social embeddedness on negotiation effort. *Acta Sociologica*, *43*, 123–37.

Rossell, D., Baladandayuthapani, V., & Johnson, V. E. (2008). Bayes factors based on test statistics under order restrictions. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 111–129). New York: Springer.

Royall, R. (1997). *Statistical evidence: A likelihood paradigm.* London: Chapman and Hall.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* London: Chapman and Hall.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of art. *Psychological Methods*, *7*, 147–177.

Scheffer, J. (2002). Dealing with missing data. *Research Letters in the Information and Mathematical Sciences*, *3*, 153–160.

Schumacker, R. E., & Akers, A. (2001). *Understanding statistical concepts using S-plus.* Mahwah: Erlbaum.

Shaffer, J. P. (1979). Comparison of means: An F test followed by a modified multiple range procedure. *Journal of Educational Statistics*, *4*, 14–23.

Silvapulle, M. J. (1996). On an F-type statistic for testing one-sided hypotheses and computation of chi-bar-squared weights. *Statistics and Probability Letters*, *28*,

137–141.

Silvapulle, M. J., & Sen, P. K. (2005). *Constrained statistical inference: Inequality, order, and shape restrictions*. New York: Wiley.

Smith, A. F. M., & Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, *55*, 3–24.

SPSS Inc. (2006). SPSS for Windows, release 15 [Computer software manual]. Chicago. Available from `http://www.spss.com/`

Stevens, J. (1999). *Intermediate statistics: A modern approach*. Mahwah: Erlbaum.

Sugiura, N. (1978). Further analysts of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics - Theory and Methods*, *7(1)*, 13–26.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Boston: Allyn and Bacon.

Toothaker, L. E. (1993). *Multiple comparison procedures*. Newbury Park: Sage.

Tukey, J. W. (1953). *The problem of multiple comparisons*. Unpublished manuscript, Princeton University.

Visual Numerics. (2003). IMSL Fortran library user's guide math/library Volume 2 of 2: Mathematical functions in Fortran [Computer software manual].

Welsch, R. E. (1977). Stepwise multiple comparison procedures. *Journal of the American Statistical Association*, *72*, 566–575.

Wesel, F., Hoijtink, H., & Klugkist, I. (in press). Choosing priors for constrained analysis of variance: Methods based on training data. *Scandinavian Journal of Statistics*.

Wiener, R. L., Holtje, M., Winter, R. J., Cantone, J. A., & Gross, K. (2007). Consumer credit card use: The roles of creditor disclosure and anticipated emotion. *Journal of Experimental Psychology: Applied*, *13*, 32–46.

Woods, H., Steiner, H., & Starke, H. (1932). Effects of composition of portland cement on heat evolved during hardening. *Industrial and Engineering Chemistry*, *24*, 1207–1214.

Zelano, P. R., Zelano, N. A., & Kolb, S. (1972). Walking in the newborn. *Science*, *176*, 314–315.

Dutch summary: Samenvatting

Menig onderzoeker heeft op voorhand een verwachting over de onderlinge relatie tussen model parameters, bijvoorbeeld groepsgemiddelden. Zo verwachten Lievens and Sanchez (2007) voor hun drie groepen dat de groepsgemiddelden afnemen over de groepen, zie pagina 11. Dit kan worden weergegeven door de hypothese H_1 : $\theta_1 > \theta_2 > \theta_3$. Restricties met een groter-dan-teken ($>$) en/of een kleiner-dan-teken ($<$) worden ongelijkheidsrestricties genoemd. Ondanks dat er vaak hypothesen zijn te formuleren met betrekking tot de relatie tussen de parameters, wordt veelal de klassieke nulhypothese (in het voorbeeld: $\theta_1 = \theta_2 = \theta_3$) getoetst. In nulhypothese toetsen kan alleen de nulhypothese verworpen worden. Ongeacht het resultaat, geeft dit nog steeds geen inzicht in de verwachting van de onderzoeker. Namelijk, wanneer de nulhypothese niet wordt verworpen, weet de onderzoeker helemaal niets over de relatie tussen de parameters; indien het wel wordt verworpen, weet de onderzoeker alleen maar dat niet alle parameters gelijk zijn. De onderzoeker zal dus nog meer toetsen (namelijk, paarsgewijze toetsen) moeten uitvoeren om er achter te komen welke parameters niet gelijk zijn. Let wel, meerdere keren toetsen verhoogt de Type I fout, dat is, de kans dat een nulhypothese verworpen wordt terwijl deze correct is. Daarnaast geeft paarsgewijs toetsen paren/groepen van parameters die wel of niet significant van elkaar verschillen. Dit geeft vaak nog steeds weinig inzicht in de hypothese waarin je geïnteresseerd bent. Daarom zou een onderzoeker zijn verwachtingen direct moeten evalueren. In dit proefschrift zijn meerdere methoden die hypothesen evalueren met elkaar vergeleken en zijn sommigen van hen uitgebreid zodat ze toepasbaar zijn op meer algemene restricties en/of meerdere typen modellen. Verder wordt er ingegaan op het omgaan met missende waarden in de data en op het combineren van resultaten uit meerdere onderzoeken.

In Part I zijn verscheidene methoden met elkaar vergeleken die toegepast kunnen worden op het variantie-analyse (ANOVA) model. Hierbij is gekeken naar hypothese toetsen en model selectie technieken. Een ander onderscheid dat gemaakt is is exploratieve en confirmatieve methoden. Met de eerste worden alle mogelijke gelijkheidsrestricties bekeken, dat is, restricties zonder een groter-dan-teken (“$>$”) en een kleiner-dan-teken (“$<$”); oftewel restricties waarin een groepsgemiddelde wel (“$=$”) of niet (“,” of “$\neq$”) gelijk aan een ander gemiddelde is. Met de tweede worden specifieke hypothesen geëvalueerd die op voorhand door de onderzoeker zijn gespecificeerd. Hoofdstuk 2 en 3 laten zien dat confirmatieve model selectie de voorkeur verdient boven de andere methoden. De twee confirmatieve model selectie technieken zijn het ongelijkheidsgerestricteerde informatie criterium (*order-restricted information criterion*; ORIC) en Bayesiaanse model selectie (BMS). Let wel, indien

geen zinnige hypothesen kunnen worden gespecificeerd, zal er exploratief te werk moeten worden gegaan. Verder zijn in Hoofdstuk 3 de eigenschappen van confirmatieve technieken onderzocht wanneer niet aan de homogeniteitsassumptie voldaan wordt, dat is, wanneer de groepsvarianties niet gelijk zijn. Indien de groepsgroottes gelijk zijn, is er geen (noemenswaardig) effect van de schending op het functioneren van de confirmatieve methoden. In de andere gevallen is er wel een effect. De richting en de grootte van het effect hangen af van de grootte van de verschillen in gemiddelden (effectgrootte) en van de verhouding tussen de groepsvariantie en groepsgroottes. Wanneer de groepen met de kleinste varianties behoren tot de groepen met de grootste groepsgroottes, is het effect van de schending op het functioneren van de confirmatieve methoden het grootst.

In Part II is het ORIC van Anraku (1999) uitgebreid. Het ORIC kan namelijk alleen worden toegepast in ANOVA modellen op simpele ongelijkheidsrestricties (*simple order restrictions*): $\theta_1 \leq \ldots \leq \theta_k$ waarbij "$\leq$" vervangen mag worden door "$=$". In Hoofdstuk 4 is het ORIC zo aangepast dat het kan worden toegepast in ANOVA modellen op hypothesen met meer algemene restricties: $R_1\theta \leq 0, R_2\theta = 0$ met R_1 en R_2 restrictiematrices. Met deze uitbreiding, genoemd het GORIC (*generalized* ORIC), kunnen alle lineaire combinaties van gemiddelden onderzocht worden. Een simulatie studie laat zien dat het GORIC een goede methode is om de beste uit een verzameling van hypothesen te selecteren. In Hoofdstuk 5 is het GORIC uitgebreid zodat het (meer) algemene restricties in multivariate lineaire modellen kan evalueren. Verder is er een GORIC afgeleid (genaamd de GORICC) die toegepast kan worden indien er een kleine steekproef is. Deze werkt met name goed in regressie modellen. In ANOVA modellen kan wanneer er weinig data zijn ook het GORIC zelf gebruikt worden.

Tot Part III zijn methoden besproken die toepasbaar zijn op compleet geobserveerde data sets, maar vaak zijn niet alle data punten geobserveerd en zijn er dus missende waarden. Er bestaan methodes en software programma's om met missende waarden om te gaan bij het schatten van parameters. Er is echter weinig bekend over het omgaan met missende waarden in model selectie op basis van informatie criteria. Hoofdstuk 8 beschrijft hoe een onderzoeker dit kan doen. Het meest belangrijke is het model dat wordt aangenomen als onderliggend data model, welke wordt gebruikt om de missende waarden te schatten. Men moet het meest ruime model nemen, dat is, het model met alle mogelijke verklarende variabelen / voorspellers, zodat de schattingen van de missende waarden zuiver zijn. Indien de schattingen worden gebaseerd op de te onderzoeken hypothese, zijn de schattingen van de missende waarden vaak niet zuiver en geven ze meer steun aan die hypothese dan wanneer het meeste ruime model wordt gebruikt. Wanneer er veel missende waarden zijn, kan dit leiden tot het verkiezen van de verkeerde hypothese. Let daarom goed op hoe er in bestaande software programma's wordt omgegaan met missende waarden en met name welk onderliggend data model wordt gebruikt om de missende waarden te schatten.

Naast model selectie op basis van informatie criteria kan men ook BMS gebruiken om ongelijkheidsrestricties te evalueren. BMS staat bekend om het meenemen van voorkennis en om het kwantificeren van het bewijs voor een hypothese. Hoofdstuk 10 in Part IV laat zien hoe je het bewijs voor een teken van een parameter (positief of negatief) uit meerderde onderzoeken kan combineren, waarbij de variabelen in elk onderzoek hetzelfde concept meten. Dit is een praktische methode aangezien er geen

originele data sets beschikbaar hoeven te zijn, maar alleen twee parameterschattingen per onderzoek. De evaluatie van een voorbeeld gebaseerd op echte studies en van vijf hypothetische situaties laat zien hoe de methode werkt en dat het goed werkt.

Om de beschreven methoden toegankelijk te maken voor de onderzoeker, is er voor iedere methode een software applicatie gemaakt. De applicaties met betrekking tot de confirmatieve ANOVA methoden en het GORIC zijn uitgebreid beschreven in Part V. De software applicaties van alle methoden beschreven in dit proefschrift is te vinden op

`http://staff.fss.uu.nl/RMKuiper`

Acknowledgments

I would like to thank my promotor and supervisor Herbert Hoijtink for structuring my thoughts and texts. Especially your questions helped me a lot.

Furthermore, I would like to thank my fellow PhD students and colleagues at Methods & Statistics. I always enjoyed our walks and talks, playing squash and beach volleyball, sitting in a restaurant or at a bar, joining IOPS-conferences or courses, writing a book-chapter or article, organizing and joining *M&T-uitjes* and drinks, or trying to mimic emotionlessness! In addition, I would like to thank Ben, my friends, and family (in law) for the talks, dinners, and playing board games.

I like to end (just for statistical fun and no, this is not a contradictio in terminis) with a quote of Gofdried Bomans (in Dutch): "Een statisticus waadde vol vertrouwen door een rivier, die gemiddeld één meter diep was. Hij verdronk jammerlijk."

About the Author

Rebecca Margaretha Kuiper was born in 1982 on January 11 in Groningen, the Netherlands. She finished her preparatory university education (Gymnasium; at Dollard College in Winschoten) in 2000. Subsequently, she started her studies in Econometrics and Operation Research at Groningen University. She finished both majors Operations Research and Econometrics. In the latter, she obtained her degree in 2005. During this period, she was a teaching-/research-assistant in several fields (namely at the Faculties of Management Sciences, Marketing & Market Research, and Econometrics & Operations Research); she gave extra lessons in several courses to university students of various faculties and in economics and mathematics to high schools students; and she was an active member of the study union (several positions and committees). Furthermore, she studied for half a year at the Hanken School of Economics in Helsinki. In September 2005, she started the Research Master called Human Behaviour in Social Contexts. In March 2007, she obtained her degree in Psychometrics and Statistics (cum laude). During this period, she worked as a statistical and methodological consultant for students of the Faculty of Behavioral and Social Sciences and she was an active member of the floorbal sports union.

In May 2007, Rebecca started as a Ph.D. student at the Department of Methodology and Statistics of University Utrecht under supervision of Prof. Dr. Herbert Hoijtink. From May 2007 till April 2008 she worked for .8 fte, from April 2008 till October 2010 for 1 fte, from October 2010 till April 2011 for .4 fte, and from April 2011 till September 2011 for .6 fte (i.e., in total for 3.7 fte). During this period, she published several articles and two book chapters, which are listed below. In addition, she provided several software packages (available from `http://staff.fss.uu.nl/RMKuiper`). Furthermore, she presented her work on several workshops and conferences. In October 2010, she started working as a junior statistician (for .6 fte; from April 2011 on, for .4 fte) for TNO (*Dutch*: de Nederlandse Organisatie voor toegepast-natuurwetenschappelijk onderzoek), where she analyzes data and provides statistical and methodological consulting in projects of TNO. In September 2011, she will start (for .6 fte) as a postdoctoral fellow ("post doc") at the Department of Methodology and Statistics of University Utrecht under supervision of associate professor Dr. Ellen Hamaker.

Articles:

Kuiper, R. M. and Hoijtink, H (2010). Comparisons of Means Using Exploratory and Confirmatory Approaches. Psychological Methods, 15, 69-86.

Kuiper, R. M., Klugkist, I., and Hoijtink (2010). A Fortran 90 Program for Confirmatory Analysis of Variance. Journal of Statistical Software, 34(8), 1-31.

Kuiper, R. M., Hoijtink, H, and Silvapulle, M. J. (2011). An Akaike-type Information Criterion for Model Selection Under Inequality Constraints. Biometrika, 98 (2), 495-501.

Kuiper, R. M. and Hoijtink, H (2010). Generalization of the Order-Restricted Information Criterion: Illustrated. Proceedings of the 25th International Workshop on Statistical Modelling, University of Glasgow 5-9th July, 2010, 303-306.

Kuiper, R. M. and Hoijtink, H (2011). How to Handle Missing Data in Regression Models Using Information Criteria. Statistica Neerlandica, 65(4), 489-506.

Kuiper, R. M., Raub, W., Buskens, V., and Hoijtink, H (accepted). Combining Statistical Evidence from Several Studies: Positive Past Effects on Trust. Sociological Methods and Research.

Book chapters:

Hoijtink, H., Huntjens, R., Reijntjes, A., Kuiper, R., and Boelen, P. A. (2008). An Evaluation of Bayesian Inequality Constrained Analysis of Variance. In Hoijtink, H., Klugkist, I., and Boelen, P. A. (Ed.), Bayesian evaluation of informative hypotheses. (pp. 85-108). New York : Springer.

Hamaker, E. L., van Hattum, P., Kuiper, R. M., and Hoijtink, H. (2011). Model Selection Based on Information Criteria in Multilevel Modeling. In Hox, J. and Roberts, K. (Ed.), Handbook of Advanced Multilevel Analysis. (pp. 231-255). New York: Taylor and Francis Group.

Submitted Articles:

Kuiper, R. M., Nederhoff, T., and Klugkist, I., and Hoijtink, H (submitted). Performance and Robustness of Confirmatory Approaches.

Kuiper, R. M., Hoijtink, H, and Silvapulle, M. J. (submitted). Generalization of the Order-Restricted Information Criterion for Multivariate Normal Linear Models.

Kuiper, R. M., and Hoijtink, H (submitted). A Fortran 90 Program for the Generalization of the Order-Restricted Information Criterion.

Kuiper, R. M. (submitted). Model Selection under Inequality Constraints in Small Samples.