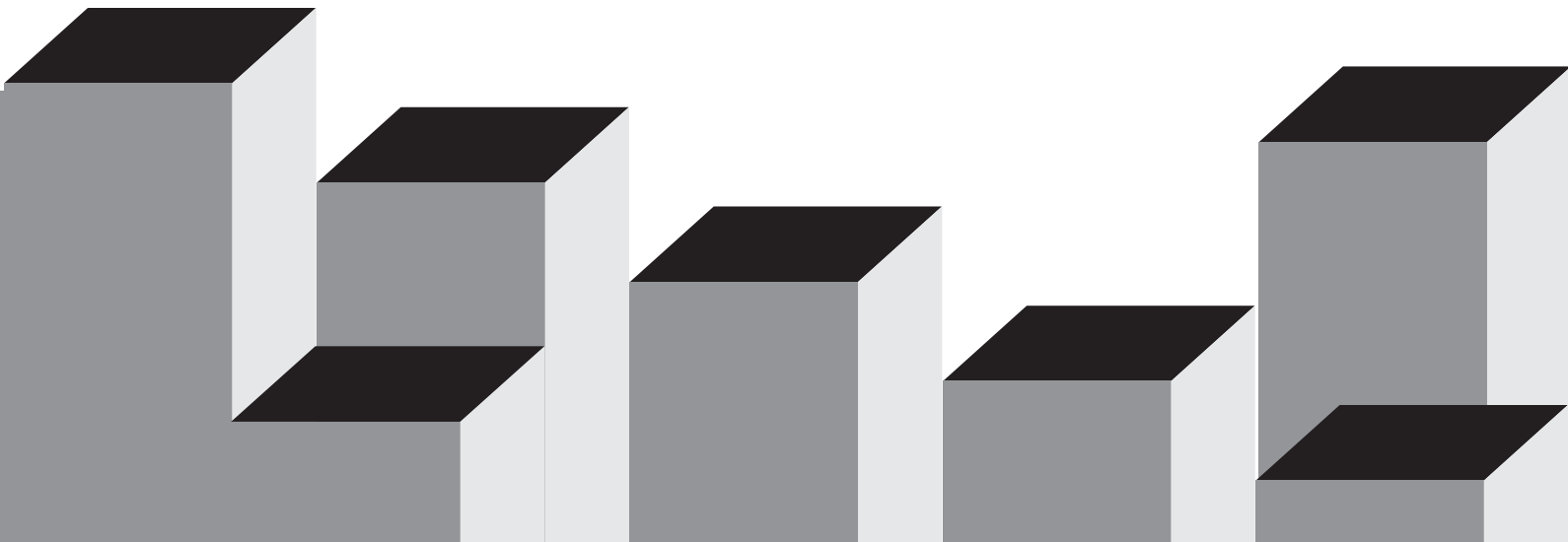
Bayesian Inequality Constrained Models for Categorical Data

Olav Laudy



Bayesian Inequality Constrained Models for Categorical Data

ISBN-10: 90-393-4299-7
ISBN-13: 978-90-393-4299-2

Bayesian Inequality Constrained Models for Categorical Data

Ongelijkheidsrestricties in Bayesiaanse Categorische Data Modellen

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. W. H. Gispen,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

donderdag 27 oktober 2006 des middags te 2.30 uur

door

Olav Laudy

geboren op 27 oktober 1975, te Alkmaar

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

Preface

Five years ago, I intended to do some extra work in the area of statistics. Somehow, I ended up in front of the door of Herbert Hoijtink. I asked him whether he had heard of some statistics called 'Bayesian' statistics, as I heard this term few times before, and somehow it interested me. "Bayesian statistics, that is my hobby", answered Herbert, and although I was not yet graduated, that was the beginning of this book. Herbert, thank you for all the things you taught me, for all your very practical knowledge, your big fun, your patience and your absolute amazing devotion. In Dutch, there is the perfect expression for you: Herbert, je bent een prachtkerel. Irene, in this 5 years we visited conferences in up to four continents, and it has always been a joy discussing matters with you. I remember discussions about statistics in many weird places and times: in the snow during wintersports, somewhere high in the Andes in Chile, climbing on volcano in Tenerife, in the greyhound on our way to Duluth. I tremendously enjoyed the little holidays after the conferences in the exotic places we visited. Some words for my roomie Ardo: though you have been working in Cambridge for the last half year now, I did not forget about you. We shared a room for 6 years, and you were a wonderful interesting roommate. I still miss our flavored discussions. Roland, my big friend, we had the greatest fun together, both in pulling each others legs, as well as those of others. Rolling on the floor laughing, shouting "1-0 for me, buddy" after another silly joke, that is how I will remember those good times. Peter, alias 'big Pete', your open-door policy has proven to be a great success. I have always found it truly valuable to just slip in your room and have a statistical small-talk for 10 minutes. I also like to thank my dear friend Mike, for the many times he raised his mathematical eyebrow over my statistical folly. Frank, Bernet, Pascal, Maureen, Gerty, Henny and many others have helped to make my PhD period to truly great time. They sometimes say that your life as a student is the best life you will ever know. For me this will probably be my life as a PhD-student. Finally, I want to thank Femke, my dearest girlfriend, for just being the best girlfriend I can imagine.

Olav

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the social sciences, formulation of theories is one of the main activities. In general, on the basis of observations, reasoning and earlier research, a researcher has certain expectations or hypotheses about reality. Collecting data serves to verify or falsify these hypotheses. There are two aspects of hypotheses that are important: first, they should describe the data accurately, second, they should provide a parsimonious description of reality.

For example, suppose persons attended a therapy to help them deal with spider phobia. The result of the therapy is either unsuccessful (a person has still spider phobia) or successful (the persons is not hindered in daily life by the phobia). A researcher is interested how success can be predicted by the age of patients. Suppose, the patients are classified into three age groups, young, middle-aged and old. The success probability in a group is given by the number of persons with a successful outcome divided by the total number of persons in that group. This is displayed in Table 1.1.

Table 1.1: Spider Phobia Therapy

	Age		
	Young	Middle-aged	Old
Success probability	π_1	π_2	π_3

A researcher has certain several (conflicting) expectations about these data. As a baseline theory, it is stated that successfulness does not depend on age, or otherwise stated, the probability of success is equal in each age group. This expectation is translated into hypothesis H_0 in Table 1.2, where can be seen that all probabilities are constrained to be equal. The second expectation is that the younger persons are the more successful in the therapy, as younger persons tend to be more flexible. This can be translated into the hypothesis where π_1 is larger than π_2, and π_2 is larger than π_3, which is displayed in Table 1.2 as hypothesis H_1. The third expectation is that older persons are more successful in the therapy, as they are easier taught to rationalize fears. The corresponding hypothesis H_2 is displayed in Table 1.2. A mix between both expectations is that middle-aged persons are most successful in therapy as they still tend to be flexible, but are also able to rationalize their fears. The corresponding hypothesis H_3 is displayed in Table 1.2. The

last hypothesis H_4 specifies no structure on the probabilities. Without any structure (or not being a parsimonious description of reality), the latter hypothesis is not attractive in a scientific sense, however, it may turn out to be the best hypothesis if all other hypotheses fail. Note that all the hypotheses involve careful considerations of the researcher. Taking all possible combinations of constraints in the model set would not lead to better insight in the matter, as the best model for the data may still need interpretation after the analyses are executed. This leads to a more exploratory perspective: 'what can we find in the data?', rather than the confirmatory approach that we advocate: 'which of the ideas we have, is most supported by the data?'

Table 1.2: Theories about Spider Phobia Therapy

	Age				
Model	Young		Middle-aged		Old
H_0:	π_1	$=$	π_2	$=$	π_3
H_1:	π_1	$>$	π_2	$>$	π_3
H_2:	π_1	$<$	π_2	$<$	π_3
H_3:	π_1	$<$	π_2	$>$	π_3
H_4:	π_1		π_2		π_3

Suppose, this experiment is conducted, and has rendered actual data. One is interested in 1) which of the hypotheses accurately and parsimoniously describes the data? 2) which of the hypotheses describes the data best? With the current statistical techniques, the above question cannot always be answered. Using the classical p-value, a researcher is able to calculate the probability that the observed data, or more extreme, stems from the population where hypothesis H_0 is true as compared to hypothesis H_4. If this probability is small (in general smaller than 0.05), the null hypothesis (H_0) is rejected. The p-value lacks the feature that allows to select the best hypothesis from a set of hypotheses, but scientifically this is a relevant question: which of the hypotheses does provide the best combination of an accurate and parsimonious description of the observed data?

A different way to draw conclusions is to use information criteria. Every hypothesis is given a number indicating the support of the data penalized by its complexity, and the hypothesis with the smallest (or largest number) is chosen as the best hypothesis. However, information criteria do not have a interpretable scale, that is, a difference of say 2 point is not known to be small or large. Most information criteria consist of a part measuring how accurate the hypothesis describes the data and a part that penalize for model complexity. The penalty term mostly uses a function of the number of parameters of a hypothesis, however, in inequality constrained hypotheses, this number is not known. In this thesis, the posterior model probability is suggested. This is a Bayesian concept, and has several attractive features. First, it shows the support for a model given the data and other models on a 0-1 scale. Furthermore, it is both possible to investigate whether a hypothesis accurately describes the data and to select the best hypothesis from a set of hypotheses. Moreover, it is easy to implement for inequality constrained hypotheses. An overview is given of the chapters:

Chapter 2

The second chapter deals with confirmatory inequality constrained latent class analysis. Latent class analysis is a technique that can be used to divide a sample of respondents into homogeneous subgroups. These subgroups are called latent classes, because before the analysis it is unknown which respondent belongs to which class. Usually, latent class analysis is used in an exploratory sense, i.e., given a number of classes, the data are partitioned in the best possible subgroups. Then a researcher has to carefully evaluate the parameter estimates in each latent class to give an interpretation. There are two important issues here. First, an exploratory approach may lead to over-interpretation. Given any set of parameters, a reasonable story can be made up, but this may not reflect the state of affairs in the population. Second, researchers are not unprejudiced before the analysis is conducted. Their knowledge consists of a mix of results from previous research and their own thoughts about the matter. In a confirmatory analysis these two issues are dealt with. Before the analyses, a researcher needs to formalize his theories about the data. These formalizations are in terms of (inequality constraints of) the parameters of the model to be estimated. Note that it is encouraged to provide competing theories. The estimation procedure is such that the theories are accounted for in the models. After the analyses, information criteria will indicate which of the models fits the data best, and how the models compare to each other. In the second chapter, theories are translated into inequality constrained latent class models. Posterior predictive checks using the pseudo likelihood ratio test as a discrepancy measure are used to investigate whether the data can be replicated by the model, and the marginal likelihood, estimated by the harmonic mean estimator is used to select between models. Both measures have their problems, and in later chapters attention will be devoted to them. The method in the second chapter is illustrated with two examples. The first example concerns the dimensions of anti-social behavior and the second example concerns the cognitive development of five year old children.

Chapter 3

In the third chapter, the confirmatory latent class analysis is discussed again, but now in more computational depth using an example of the analysis of the Piagetian balance scale task. It is shown how to obtain an empirical representation of the posterior distribution using an inequality constrained Gibbs sampler. Using this representation, parameter estimates and confidence intervals are obtained. It is shown how the sample from the posterior distribution can be used to calculated an estimate of the marginal likelihood and posterior predictive checks using the pseudo-likelihood ratio test as a discrepancy measure. It is well known that the likelihood ratio test does not work well in latent class analyses, since the data are sparse, i.e. there are many more possible data patterns than there are observed data patterns. The pseudo likelihood ratio test deals with this problem by calculating the likelihood ratio test for all sets of two variables. It can be expected that the two-way tables are not sparse, hence the pseudo likelihood ratio test will give an accurate result of the difference between the expected and observed probabilities. Since the distribution of the pseudo likelihood ratio test is not known, a reference distribution is constructed by sampling replicated data from the posterior predictive distribution.

This method is called posterior predictive checks, and will be studied in later chapters in more depth. In the example, a mixed approach is shown: first, the available theories are translated into inequality constrained latent classes analyses, and after obtaining the model that is most supported by the data, more latent classes are added that are unrestricted. This gives the opportunity to see if apart from the hypothesized structures, there are more structures in the data that are not (yet) part of the theory.

Chapter 4

The fourth chapter is concerned with inequality constrained models for contingency tables. Models for contingency tables where interaction terms are allowed are another way to model categorical data. Posterior predictive checks using the likelihood ratio test as a discrepancy measure are used to investigate whether the models were able to replicate the data, and the marginal likelihood is used to investigate which model was most supported by the data. In a saturated contingency table, the inequality constraints can be placed on either the cell probabilities or on functions of cell probabilities. The chapter is illustrated with two examples. In the first example, inequality constraints are placed on odds, or the ratio of two probabilities, and in the second example restrictions are placed on the odds ratios. The marginal likelihood of two or more models can be transformed into posterior model probabilities, indicating the support of each model given the data and the model set. This has an advantage over the use of the classical p-value. Loosely speaking, a p-value is a measure of how much evidence there is against the null hypothesis, whereas the posterior model probability shows the support of both the null and alternative hypothesis. Moreover, the use of posterior model probabilities is not limited to two hypotheses (null and alternative). However, using this approach comes with a price: a prior has to be specified. It is well known that posterior model probability depends on the choice of the prior. However, it is illustrated with a small simulation study that for inequality constrained models, the posterior model probability is not very sensitive to the choice of the prior.

Chapter 5

The fifth chapter is concerned with the performance of several fit measures or information criteria in the context of inequality constrained models for contingency tables. In the previous chapters, the marginal likelihood and posterior predictive checks are used, but their frequency properties for inequality constrained models was not explored yet, hence this study. Three situations are explored: first, a correctly (in the sense of a well-fitting) inequality constrained model is compared to the unconstrained model. Second an incorrectly inequality constrained model is compared to an unconstrained model, and third, a correctly constrained model is compared to an incorrectly constrained model. This has been done for three populations, where in the first simulation the sample size is varied, in the second simulation the effect magnitude and in the third simulation, models with different a-priori likelihoods are compared. Fit measures and information criteria can divided into prior predictive measures and posterior predictive measures. The results showed that the marginal likelihood performed well, and most stable across situations. The previously used posterior predictive checks appeared not to work very well. A general

interesting result was that prior predictive measures (such as the marginal likelihood) are more sensitive for model complexity in inequality constrained models.

Chapter 6

In the sixth chapter, the knowledge gathered in the previous chapters is used to demonstrate the use for the social sciences. In this chapter, data from recent published psychological literature is re-analyzed using constrained hypotheses (both inequality and equality constrained). The previous chapter showed that the marginal likelihood had good frequency properties in choosing among inequality constrained models, and hence it will be used in this chapter as the sole criterion. The calculation of the marginal likelihood was omitted by a simple and direct calculation of Bayes factors and subsequent posterior model probability. This method originally designed for nested inequality constrained models was adapted such that it could also handle equality constrained models. Moreover the derivation of the computational method shows how the Bayes factor consists of a term that captures the fit of the model and a term that captures the penalty for model complexity.

Furthermore, in this chapter detailed simulations are performed to investigate the behavior of the posterior model probability for various data formats. The first simulation is concerned with the magnitude of the posterior model probability as a function of sample size and effect size. A second simulation investigates the counterpart of type 1 error for the posterior model probability. The last simulation investigates the prior sensitivity. It is shown that the Bayes factor for inequality constrained models is not sensitive to the choice of the prior, while equality constrained model are sensitive to the choice of the prior, however, in various situations this sensitivity is more severe than in others. It is concluded that a common prior value of one leads to interpretable results. Also, a common prior value of one has the attractive intuitive interpretation that a-priori each cell probability is equally likely. The chapter is concluded with three more advanced analyses that show the use and benefit of inequality constrained models and the posterior model probability for the social sciences.

Conclusion

This thesis deals with the translation of theories into statistical models, and subsequently testing, and selection of the best model. Several advances are made. First, by translating different hypotheses into various models, the use for social scientists is demonstrated. Second, computational procedures are developed or enhanced and third, through simulation, insight is gathered in the actual performance of several information criteria.

Chapter 2

Applications of Confirmatory Latent Class Analysis in Developmental Psychology*

Abstract

In the field of developmental psychology, researchers may have several competing theories with respect to their research subject. In this paper an approach will be proposed that can be used to select the best of these theories. It will be shown that a theory can be translated in a constrained latent class model (CLCA) using inequality constraints. This can be done for several (possibly competing) theories. Subsequently, fit-measures can be used to determine which model (and thus which theory) is supported most by the data. The approach will be introduced using data with respect to self-reported child and adult antisocial behavior. It will be further illustrated using data obtained using the figural intersection task.

*This chapter has been published as Laudy, O. and Hoijtink, H. (2005) Applications of Confirmatory Latent Class Analysis in Developmental Psychology. *European Journal of Developmental Psycholgy* **2(1)**: 1–15

2.1 Introduction

Exploratory Latent Class Analysis (ELCA)(Clogg, 1981; Goodman, 1974; Haberman, 1988; Vermunt, 1996) is a statistical technique that can be used to divide a sample of respondents into homogeneous subgroups also called latent classes. In developmental psychology, for example, data with respect to some cognitive ability of children, like responses to items on Piaget's balance scale task, Piaget's water level task or the Figural Intersection Task can be subjected to a ELCA. This can result in groups of children using different strategies (e.g. Boom, Hoijtink, Kunnen, 2001; Jansen and van der Maas, 1997; Pascual-Leone and Baillargeon, 1994) and the transitions between the groups (Hoben and Hettsmansperger, 2001). For a methodological overview and applications, see von Eye and Clogg (1994).

A key question in ELCA is: into how many homogeneous subgroups the sample should be divided? Usually fit measures (Everitt, 1988; Lin and Dayton, 1997) are used to determine which number of classes is optimal. Furthermore, the resulting classes have to be interpreted. To illustrate this, consider an ELCA of the responses of 2001 women to several items with respect to self-reported child and adult antisocial behaviors. In order to assess maternal antisocial disorder, the mothers were asked about five childhood antisocial behaviors and four adult behaviors. In Table 2.1 the items (responses to either "yes" or "no") and the response percentages can be found. The women were mothers of five months old infants from a population based longitudinal study of the development of children of the province Quebec, Canada (Zoccolillo, 2000; Jette, Desrosiers, Tremblay and Thibault, 2000).

In the top panel of Table 2.2, three fit measures for exploratory analyses with two, three, and four classes are displayed. Hoijtink (1998, 2001) developed these fit measures to be able to select the best of a number of *exploratory* and *confirmatory* (see below) latent class models. The first measure is -2 log of the marginal likelihood (Kass and Raftery, 1995). This measure can be seen as the Bayesian counterpart to information criteria like AIC, CAIC and BIC (Kass and Raftery, 1995). It is a relative (with respect to the corresponding numbers for other models) fit measure: the smaller the number the better the model. If a priori each model is considered to be equally good, -2 log of the marginal likelihood can be transformed to posterior probabilities (Kass and Raftery, 1995), which are easier to interpret. The third measure is an absolute (how good is the model at hand) fit measure. It is a modification of the likelihood ratio goodness of fit test (see for example, Formann, 1985) that is traditionally used in latent class analysis. For this test, p-values smaller than .05 indicate that the model at hand is not able to adequately reproduce the frequencies with which each response pattern in the data matrix is observed.

As can be seen in Table 2.2, the exploratory model with three classes appears to be the best model. Note that this does not exclude the possibility that in the population the women are grouped in two or four classes. In exploratory latent class analysis the number of classes can only approximately be determined using fit measures (see, for example, Everitt, 1988; Lin and Dayton, 1997; Hoijtink, 1998, 2001). In Table 2.3, estimates of the parameters can be found of the three-class exploratory model. These estimates are obtained using Bayesian computational statistics; the interested reader is

Table 2.1: Items Indicative of Child and Adult Antisocial Behavior

j	Item Phrasing	Item type	%Yes
1	Before the end of high school, did you more than once wipe things from stores or from other children, or steal from your parents or from anyone else?	Child	18
2	Before the end of high school, did you more than once get into fights that you had started?	Child	3
3	Before the end of high school, were you ever involved with Social Services (Department of Youth Protection), in trouble with the police or arrested BECAUSE OF YOUR OWN MISBEHAVIOR?	Child	4
4	Before the end of your high school, did you ever skip school at least twice in one year?	Child	48
5	Before the end of your high school, did you ever run away from home overnight?	Child	9
6	Since leaving of finishing school, have you been FIRED from your job (do not take account layoffs resulting from lack of work)?	Adult	9
7	Since leaving of finishing school, have you ever been arrested for anything OTHER than traffic violations?	Adult	1
8	Since leaving of finishing school, did you ever hit or throw things at your spouse (or partner that you were living with)?	Adult	10
9	Since leaving of finishing school, have you ever been in trouble at work, with the police or with your family, or had a car accident BECAUSE OF DRUGS OR ALCOHOL?	Adult	9

referred to Hoijtink and Molenaar (1997) and Hoijtink (1998). Each class gets a weight, that is, the proportion persons belonging to that class, and each class has its own class specific probabilities, that is, the probability of answering "yes" to each of the items. Note that j will be used to indicate item numbers, ω_1 denotes the proportion of persons in class 1, and π_{2j} denotes the probability of responding "yes" to item j in class 2.

The class specific probabilities can be used to interpret the nature of the class. Looking at Table 2.3, class one will probably be labelled 'school-skippers' because the probability of skipping school at least twice in one year (see Table 2.1 for the item labels) is rather large (37%). However, class one might also be labelled 'normal children' because skipping school is not really unusual behavior, even for 'normal children'. Class two will probably be labelled as persons who have a high probability of some form of antisocial behavior, either in childhood or adulthood. The interpretation of class three is less straightforward. Different researchers may attach different labels to class three. This post-hoc labelling of the latent classes is one of the drawbacks of exploratory latent class analysis. Even if a researcher has a number of theories with respect to the nature of the classes in the population of women, an exploratory analysis may not help to decide which theory is the best, because the number of classes chosen and the labelling attached may not straightforwardly indicate which theory is the best.

A researcher using exploratory analysis behaves as if his research field has not yet been explored very thoroughly, and theories do not exist. This however, is not always

Table 2.2: Evaluation of the Models for the Antisocial Behavior Data

Model	-2log Marginal Likelihood	Post. Prob.	p-values
Exploratory Two Classes	9602	.00	.12
Exploratory Three Classes	9564	.91	.23
Exploratory Four Classes	9570	.09	.22
One-dimensional	9545	.08	.33
Two-dimensional	9540	.92	.38
Three-class	9553	.00	.29

Table 2.3: Class Specific Probabilities for the Three Class Exploratory Model

Item Type	j	Class 1	Class 2	Class 3
Child	1	.07	.74	.54
	2	.01	.60	.09
	3	.01	.80	.12
	4	.37	.80	.81
	5	.02	.74	.30
Adult	6	.07	.47	.16
	7	.01	.46	.04
	8	.05	.37	.26
	9	.01	.42	.03
		ω_1=.77	ω_2=.01	ω_3=.23

an accurate reflection of the true state of affairs. ELCA has been done in areas that have been thoroughly explored, and where theories are well developed (Boom, Hoijtink and Kunnen, 2001; Hoben and Hettsmansperger, 2001; Jansen and van der Maas, 1997; Pascual-Leone and Baillargeon, 1994). The plausibility of real scientific progress is larger if the current state of affairs (existing knowledge and theories) is properly accounted for in the statistical models used for the analysis.

In this paper a specific form of confirmatory latent class analysis (CLCA) will be proposed (Hoijtink and Molenaar, 1997; Hoijtink, 1998; Hoijtink, 2001). It will be shown how theories can be translated into a CLCA using inequality constraints among the parameters of the model. Subsequently, the best model, that is, the best theory, can be selected.

In Section 2.2, the approach proposed will be illustrated continuing the example concerning parental antisocial behavior (Zoccolillo, Pickles, Quinton and Rutter, 1992; Zoccolillo, Price, Ji and Hwu, 1999). In Section 2.3, some of the hypotheses and theories described in Pascual-Leone and Baillargeon (1994) concerning the development of mental attention will investigated using CLCA. The paper will be concluded with a discussion.

2.2 The Dimensions of Antisocial Behavior

In this section, the data with respect to antisocial behavior will be analyzed using confirmatory latent class analysis (CLCA) (Hoijtink and Molenaar, 1997; Hoijtink, 1998; Hoijtink, 2001). This approach consists of four steps. In the first step, a researcher has to

make an inventory of existing theories with respect to the research domain and research questions of interest. In the second step, each theory has to be formalized. As will be illustrated below, this will be done using inequality constraints among the parameters of latent class models. In the third step, the fit measures introduced in the previous section will be used to determine which constrained latent class model (and thus which theory) is supported most by the data. In a possible fourth step, the parameter estimates for the best model are inspected. Note that for each model the inequality constraints are active both when fit measures are computed, and, when parameters are estimated.

For these data, three theories exist. The first model reflects the theory that antisocial behavior is a invariant characteristic of human personality, affecting some persons more than others. The corresponding inequality constraints for this theory restrict the class specific probabilities such that these increase with the number of the latent class, independent of whether it concerns child or adult items. This is illustrated in Table 2.4, where it can be seen that with increasing class number, both the child and adult items have higher class specific probabilities. Note that these constraints imply an one-dimensional structure for the data, low class numbers indicate a low degree of antisocial behavior, high class numbers indicate a high degree.

Table 2.4: Inequality Constraints for the One-Dimensional Theory

Item type:	j	Class 1		Class 2		Class 3		Class 4
Child	1	π_{11}	$<$	π_{21}	$<$	π_{31}	$<$	π_{41}
	..	π_{1j}	$<$	π_{2j}	$<$	π_{3j}	$<$	π_{4j}
	5	π_{15}	$<$	π_{25}	$<$	π_{35}	$<$	π_{45}
Adult	6	π_{16}	$<$	π_{26}	$<$	π_{36}	$<$	π_{46}
	..	π_{1j}	$<$	π_{2j}	$<$	π_{3j}	$<$	π_{4j}
	9	π_{19}	$<$	π_{29}	$<$	π_{39}	$<$	π_{49}

The second model reflects the theory that child and adult antisocial behavior are separate concepts. Being antisocial in childhood does not imply being antisocial in adulthood and visa versa. This is illustrated in Table 2.5. The inequality constraints are implicit, a minus sign indicates that the corresponding class specific probability is smaller than the probabilities associated with a plus sign. Persons in the first and second class are constrained to have lower probabilities of responding positively to the items indicative of childhood antisocial behavior than persons in the third and fourth class. Persons in the first and third class are constrained to have lower probabilities of responding positively to the items indicative of adulthood antisocial behavior than persons in the second and fourth class. This leads to a two-dimensional structure, i.e. the four latent classes separate in not being antisocial in both child- and adult-hood, being antisocial in one of these, and being antisocial in both child- and adult-hood. Note that in Table 2.5 the constraints apply to each row, e.g. the class specific probabilities for item one in class one and two are restricted to be smaller than the class specific probabilities for item one in class three and four.

Longitudinal and epidemiologic studies have found there are broadly three groups of women: those without significant antisocial behavior in childhood or adulthood; those with significant antisocial behavior in childhood but not adulthood; and those with

Table 2.5: Inequality Constraints for the Two-Dimensional Theory

Item type:	j	Class 1	Class 2	Class 3	Class 4
Child	1	-	-	+	+
	..	-	-	+	+
	5	-	-	+	+
Adult	6	-	+	-	+
	..	-	+	-	+
	9	-	+	-	+

Constrains are implicit for each row of the table: - < +

significant antisocial behavior in childhood and adulthood (Zoccolillo, Price, Ji, Hwu, 1999). Women with significant antisocial behavior in adulthood but not childhood are rare. The third model reflects this theory that many antisocial children do not become antisocial adults, but that most antisocial adults were antisocial children. The inequality constraints for this theory are equal to the two-dimensional structure without class two.

In the bottom panel of Table 2.2, it can be seen that the two-dimensional model receives the most support from the data. The p-value of the likelihood ratio test is the highest for the two-dimensional model. This indicates (.38 is larger than .05) that this model adequately reproduces the frequencies with which the response vectors are observed in the data matrix. The two-dimensional model also has the smallest -2 log marginal likelihood. Assuming that a priori each model is equally likely, this implies that after observing the data, the two-dimensional model has a probability of 92% of being the best of the three models under investigation.

The confirmatory approach does not have the drawbacks of the exploratory approach. The number of classes and the labelling of these classes have been specified *before* the analysis, using existing theories and knowledge. Subsequently, the data are used only to decide which of these specifications is the best. This leaves no room for over-interpretation, or interpretations of results (see the previous section) that differ among researchers. This does not imply that the confirmatory approach is always preferred over the exploratory approach. However, if the existing theories and knowledge can be formalized in a number of competing models, it is an excellent tool to select the best model. This includes the situation where an exploratory analysis on a first data set is used to generate theories, and a confirmatory analysis of a second data set is used to select the best of these theories. An elaborate example of the latter will be given in the next section.

In Table 2.6, the estimates of the class specific probabilities and class weights are displayed for the two-dimensional model. Note that these estimates are obtained accounting for the inequality constraints imposed on the class specific probabilities. Class two accounts for the persons that have low probabilities of child antisocial behavior, and high probabilities of adult antisocial behavior. Since this class accounts for 17.4% of all persons, it is clear why the three class model, where this class was left out, does not have a good fit. Class four contains 1% of the persons. This is realistic, because this class contains the persons that are quite antisocial, both in childhood and adulthood.

Table 2.6: Parameter estimates for the Two-Dimensional Model

		Class 1	Class 2	Class 3	Class 4
		Low Child	Low Child	High Child	High Child
Item Type	j	Low Adult	High Adult	Low Adult	High Adult
Child	1	.05	.41	.64	.76
	2	.01	.05	.21	.61
	3	.00	.06	.36	.82
	4	.35	.65	.87	.87
	5	.01	.15	.55	.75
Adult	6	.06	.23	.10	.55
	7	.00	.09	.02	.52
	8	.03	.32	.13	.46
	9	.00	.12	.01	.48
		ω_1=.71	ω_2=.17	ω_3=.10	ω_4=.01

2.3 Cognitive Development of Five Year Old Children

2.3.1 Introduction

In this section models concerning the relation between mental attentional resources and performance will be tested using the Figural Intersection Task (FIT) (Pascual-Leone and Baillargeon, 1994). According to these authors, mental attention, a test's mental demand and performance have very specific relationships, which can be formalized into five Theoretical Structural Predictions (TSP). In their paper they attempt to verify these five predictions with the help of latent class analyses. Although their predictions are presented and described using inequality constraints, the methodology and software necessary to translate their theories into CLCA of the type discussed in this paper was not available at that time. They had to resort to various other types of LCA. In this section part of their theories will be translated into CLCA using inequality constraints. The parts of their theory used in this paper are summarized below. For the complete theory, interested readers are referred to Pascual-Leone and Baillargeon (1994). Subsequently, new data will be used to determine which theory is the best.

The figural intersection task was administered to 106 children and 96 of them completed all test items. These subjects took part in the pilot survey of the Québec Longitudinal Survey of Child Development (QLSCD). The total population for the QLSCD's pilot survey were babies of French- or English speaking mothers residing in 7 administrative regions of the province of Québec, Canada. At wave one, babies were between 58 and 61 months of age.

2.3.2 The Theory

The idea that mental attentional 'energy' or capacity (nowadays often thought of as 'working memory') is essential in cognitive development might go back to Binet (Pascual-Leone and Baillargeon, 1994). In this research the mental capacity is referred to as M-capacity, which is the number of schemes that a subject can simultaneously boost into activation. The M-capacity is a limited resource. A related property of a test item is

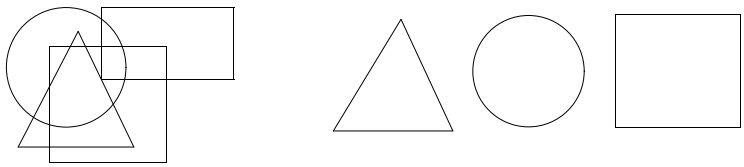


Figure 2.1: An Item from the Figural Intersection Task

M-demand, the minimum M-capacity that a respondent needs to solve an item using a stipulated solution strategy.

Every item of the FIT presents on the right-hand side of the page a number of geometric shapes separated from each other (see Figure 2.1 for an example). Shapes on the right-hand side are the task's relevant figures. Every item has two subtasks. The first is to place a dot inside each figure found on the right-hand side of the page to ensure proper exploration of all relevant figures. The second is to place on the left-hand side a single dot that is inside all relevant figures at the same time, i.e. to find the total intersection. The placement of the final dot determines whether the response is judged to be correct or incorrect.

For five year old children, the appropriate number of figures in each item is two or three. On the left hand there is a figural compound with all relevant (i.e. right-hand side) figures. In other items of the task this figural compound may contain one irrelevant figure, not found on the right-hand side. When present, the irrelevant figure is a distracter that must be ignored.

The version of the FIT used here consists of 8 items, two items per item type: two relevant figures; two relevant figures and one irrelevant figure; three relevant figures; and, three relevant figure and one irrelevant figure.

Children could solve an item by one by one trying all combinations of intersections to find the common intersection. This strategy does not lead a correct estimate of a child's M-capacity, because the different figures or 'schemes' are not boosted at once into the memory. To discourage this partial sectioning strategy children have to practice the FIT, using items that contain figures that are irrelevant, i.e. that do not have a common intersection with the other figures. Thus, in the training the child learns that the partial sectioning strategy is not a good one, because it will not lead to the good solution.

The theoretical structural predictions and an alternative model are displayed in Tables 2.7 and 2.8. The point of departure is a latent class model, where it is assumed that there are four M-demand levels for the items (the most easy are the items with two figures, the most difficult items have three figures and one irrelevant figure). With four different M-demand levels, there can be five M-capacity classes: the first class where children fail any item, the second where the children pass the first item type and fail the others, up to the fifth class where the children pass all items. From this TSP 1 and 2 can be derived: if an item has a higher M-demand than the childs M-capacity, the probability of a correct response is smaller than if the item has a smaller M-demand than the childs M-capacity; and, if a child has a higher M-capacity than the items M-demand, the probability of a correct response is larger than if the child has a smaller M-capacity than the items

M-demand. This implies a boundary as indicated by the $\vee$ and $<$ in Table 2.7. The TSP 3 in the original theory implies that probabilities below the boundary are equal, and, that probabilities above the boundary are equal. This reflects the idea of Pascual-Leone and Baillargeon that the only relevant aspect of M-capacity is whether or not it is bigger than the M-demand. We relaxed this aspect of their theory such that the probabilities (denoted by H(igh)) are restricted to be larger larger than the probabilities (denoted by L(ow)). These three prediction are joined in one set and will be called model TSP123.

Pacual-Leone and Baillargeon also present TSP 4 and 4*. These predictions reflect the theory that if the M-capacity is larger than the M-demand, it is easier to master items with a low M-demand, while having a low M-capacity (for example, responding correctly to items 1 and 2 while in class 2) than to master items with a higher M-demand, while having a higher M-capacity (for example, responding correctly to items 5 and 6 while in class 4). If the M-capacity is smaller than the M-demand, this relation is reversed: it is easier to master items with a low M-demand, while having a low M-capacity (for example, responding correctly to items 1 and 2 while in class 1) than to master items with a higher M-demand, while having a higher M-capacity (for example, responding correctly to items 5 and 6 while in class 3). TSP 4 and 4* reflect the M-capacity / M-demand difference effect (Pascual-Leone and Baillargeon, 1994): whenever one varies simultaneously M-capacity and M-demand the greater the M-capacity/M-demand ratio the greater the probability of success. This can be formalized restricting the upper triangle of Table 2.7, such that the probability of correctly responding to an item gets smaller in the down diagonal direction. This model will be called TSP1234. In the lower triangle this relation is reversed: the probability of correctly responding to an item increases in down diagonal direction. This model will be called TSP1234*. The total set of the predictions will be called model TSP 12344*.

Table 2.7: Inequality Constraints for TSP123 and TSP12344*

j	Class 1		Class 2		Class 3		Class 4		Class 5
1,2	L	$<$	H		H		H		H
		$\nwarrow$	$\vee$	$\searrow$		$\searrow$		$\searrow$	
3,4	L		L	$<$	H		H		H
		$\nwarrow$		$\nwarrow$	$\vee$	$\searrow$		$\searrow$	
5,6	L		L		L	$<$	H		H
		$\nwarrow$		$\nwarrow$		$\nwarrow$	$\vee$	$\searrow$	
7,8	L		L		L		L	$<$	H

TSP 1: $\vee$, TSP 2: $<$, TSP 3: L,H , TSP 4: $\searrow$, TSP 4*: $\nwarrow$

Hoijtink and Molenaar (1997) present the latent class equivalent of a non-parametric item response model that might be appropriate for the data at hand. The five class version of this model is presented in Table 2.8. As can be seen, the main difference with respect to model TSP123 is the absence of a diagonal boundary. This, so called, double monotonous model assumes an ordering of both the items and the classes, i.e. M-capacity increases with class number, M-demand increases with item complexity. As can be seen in Table 2.8, this model implies inequality constraints in two directions. In the horizontal directions the probability of a correct response is restricted to increase with

class number (that is M-capacity), in the vertical direction the probability of a correct response is restricted to decrease with item complexity.

This double monotone model contains two simpler models. One with a row ordering i.e. M-demand increases with item complexity, and one with a column ordering, i.e. M-capacity increases with class number. Both the double monotone and the two simpler models will be included in the set of models that will be investigated.

Table 2.8: Inequality Constraints for the Double Monotone Model

j	Class 1		Class 2		Class 3		Class 4		Class 5
1,2	π_{1j}	$<$	π_{2j}	$<$	π_{3j}	$<$	π_{4j}	$<$	π_{5j}
	$\vee$		$\vee$		$\vee$		$\vee$		$\vee$
3,4	π_{1j}	$<$	π_{2j}	$<$	π_{3j}	$<$	π_{4j}	$<$	π_{5j}
	$\vee$		$\vee$		$\vee$		$\vee$		$\vee$
5,6	π_{1j}	$<$	π_{2j}	$<$	π_{3j}	$<$	π_{4j}	$<$	π_{5j}
	$\vee$		$\vee$		$\vee$		$\vee$		$\vee$
7,8	π_{1j}	$<$	π_{2j}	$<$	π_{3j}	$<$	π_{4j}	$<$	π_{5j}

2.3.3 Results

Table 2.9 shows that TSP1234 is the best model. The p-value of the likelihood ratio test indicates that this model adequately predicts the frequencies with which the each response vector is observed. Furthermore, assuming that a priori each model is equally likely, the posterior probability is 71%, indicating that in comparison with the other models, this model is the best model. Note that TSP1234 contains TSP123 which has a posterior probability of 27%. Stated otherwise, the support for TSP123 is huge (27% plus 71%) and the support for TSP1234 substantial (71%). Estimates of the class-specific probabilities and class weights for the TSP1234 model can be found in Table 10. Between brackets the 95% central credibility interval (the Bayesian counterpart of a confidence interval) can be found. These give the precision with which each of the numbers is estimates. As can be seen, the credibility intervals (also computed accounting for the constraints) have an average range of about .30 (distance between lower and upper bound). This is caused by the relatively small sample of children used for the analysis. Nevertheless, the approach proposed is able to select the best of the seven theories (see Table 9) under investigation.

TSP1234 is obtained adding TSP4 to TSP123. Adding TSP4* to either the TSP123 or the TSP1234 model yields a substantially worse fitting model. In fact, neither the TSP1234* nor TSP12344* model adequately predicts the observed frequencies of each response pattern (p-values smaller than .05). Hence, these results suggest that TSP1, TSP2, TSP3 and TSP4 adequately describe children's performance on the FIT, but TSP4* does not. TSP1234 implies a misleading effect caused by the context when one simultaneously varies M-capacity and M-demand, and when M-capacity is larger than M-demand: the greater the number of partial intersections present in the compound - which increases exponentially with the number of figures - the smaller the probability of a correct answer. As can be seen in Table 2.10 TSP4* is not visible. Looking at the

probabilities in the lower left hand triangle, it is clear that these do not agree with the constraints implied by TSP4*.

Furthermore, these results confirm the all-or-none character of children's performance on the FIT. That is, no matter the distance between a child's M-capacity and a test item's M-demand, as long as the former is greater than the latter, a child will tend to succeed; otherwise, he or she will tend to fail and again the likelihood of doing so is unaffected by the distance between M-capacity and M-demand. As can be seen in Table 2.9, the double monotone, row ordering and column ordering models (which are more subtle than the all-or-none implications of TSP123) can predict the observed frequencies of the response patterns, however, compared to TSP123 and TSP1234 these models have very small posterior probabilities. This is supported by the numbers in Table 2.10, where it is immediately clear that both a column wise and row wise ordering of the class specific probabilities are rather unlikely.

Finally, as can be seen in table 2.10, the majority (23% + 29% + 25% + 3%) of 58-61 month old children can simultaneously coordinate at least two schemes in a goal directed activity. Only 18% of the children does not have this capacity.

Table 2.9: Evaluation of Different Models for the FIT Data

Model	-2log Marginal Likelihood	Post. Prob.	p-value
TSP 123	689	.27	.21
TSP 1234	687	.71	.25
TSP 1234*	699	.00	.03
TSP 12344*	699	.00	.01
Double Monotone	702	.00	.20
Row Ordering	699	.00	.22
Column Ordering	694	.02	.24

Table 2.10: Parameter Estimates of the TSP1234, between () Central Credibility Intervals

j	Class 1	Class 2	Class 3	Class 4	Class 5
1	.23 (.05-.44)	.96 (.87-.99)	.96 (.90-.99)	.89 (.73-.99)	.56 (.17-.97)
2	.06 (.00-.21)	.95 (.85-.99)	.97 (.92-.99)	.94 (.80-.99)	.47 (.06-.94)
3	.05 (.00-.18)	.34 (.04-.62)	.88 (.74-.97)	.90 (.77-.97)	.60 (.25-.90)
4	.04 (.00-.17)	.16 (.01-.43)	.86 (.71-.96)	.88 (.73-.97)	.44 (.09-.85)
5	.06 (.00-.21)	.14 (.01-.33)	.19 (.00-.47)	.74 (.52-.90)	.61 (.24-.89)
6	.10 (.00-.29)	.36 (.13-.59)	.45 (.11-.70)	.78 (.62-.91)	.72 (.44-.90)
7	.10 (.00-.26)	.08 (.00-.26)	.14 (.01-.39)	.38 (.15-.66)	.57 (.34-.80)
8	.04 (.00-.18)	.06 (.00-.19)	.05 (.00-.18)	.18 (.03-.43)	.39 (.13-.73)
ω_i=	.18 (.11-.27)	.23 (.10-.38)	.29 (.09-.49)	.25 (.06-.44)	.03 (.00-.11)

2.4 Conclusion

This paper proposed confirmatory latent class analysis as an alternative for exploratory latent class analysis. Confirmatory latent class analysis is especially appropriate if competing theories can be derived from an existing body of theories, research and

knowledge. As exemplified using two data sets and corresponding theories from the domain of developmental psychology, inequality constraints can be used to superimpose theories on a set of competing latent class models. Subsequently, fit measures can be used to select the best model, and thus, the best theory. Finally, estimates of the model parameters, obtained properly accounting for the constraints, can be used to determine the proportion of persons in each class and the value of the class specific probabilities.

Currently user friendly software containing an implementation of the proposed approach is being developed. Readers interested in this software can write an e-mail to the first author. The e-mail should include a description of the research project at hand, the data and the theories involved.

References

Boom, J., Hoijtink, H. and Kunnen, S. (2001). Rules in the Balance: Classes, Strategies or Rules for the Balance Scale Task. *Cognitive Development,* **16**: 717–735.

Clogg, C.C. (1981). New developments in latent structure analysis. In D. J. Jackson and E. F. Borgatta (Eds), *Factor Analysis and measurement in sociological research* (215–246). Beverly Hills, CA: Sage.

Everit, B.S. (1988). A Monte Carlo investigation of the likelihood ratio test for number of classes in latent class analysis. *Multivariate Behavioral Research,* **23**: 531–538.

Eye, von A. and Clogg, C.C. (1994). *Latent Variable Analysis. Applications for Developmental Research* London: Sage.

Formann, A.K. (1985). Constrained latent class models: theory and applications. *British Journal of Mathematical Statistical Psychology,* **38**: 38–111.

Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach. *American Journal of Sociology,* **79**: 1179–1295.

Haberman, S. J. (1988). *Analysis of quantitative data, vol 2. New developments* New York: Academic Press

Hoben, T. and Hettmansperger, T. P. (2001). Modelling change in cognitive understanding with finite mixtures. *Applied Statistics,* **4**: 435–448.

Hoijtink, H. and Molenaar, I. W. (1997). A multidimensional item response model: constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika,* **62**: 171–180.

Hoijtink, H. (1998). Constrained latent class analysis using the Gibbs sampler and posterior predictive p-values: applications to educational testing. *Statistica Sinica,* **8**: 691–711.

Hoijtink, H. (2001). Confirmatory latent class analysis: model selection using Bayes factors and (pseudo) likelihood ratio statistics. *Multivariate Behavioral Research,* **16**: 717–735.

Jansen, B. R. J. and van der Maas, H. L. J. (1997). Statistical test of rule assessment methodology by latent class analysis. *Developmental Review,* **17**: 321–357.

Jette M., Desrosiers H., Tremblay R.E., Thibault J. (2000). *Longitudinal Study of Child Development in Quebec (LDEQ),* **1(2)**, Qubec, Institut de la statistique du Qubec.

Lin, T. H. and Dayton, C. M. (1997). Model selection information criteria for non nested latent class models. *Journal of Educational and Behavioral Statistics,* **22**: 249–264.

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

Pascual-Leone, J. and Baillargeon, R. (1994). Developmental Measurement of Metal Attention. *International Journal of Behavioral Development,* **17**: 161–200.

Vermunt, J.K. (1996). *Log-linear Event History Analysis. A General Approach with Missing Data, Latent Variables, and Unobserved Heterogeneity* Tilburg: Tilburg University Press.

Zoccolillo, M. (2000). Parent's health and social adjustment, Part II-social adjustment. *Longitudinal Study of Child Development in Qubec (LDEQ),***1(2)**

Zoccolillo, M., Pickles, A., Quinton, D. and Rutter, M. (1992). The outcome of conduct disorder: implications for defining conduct disorder and adult personality disorder. *Psychological Medicine* **22**: 971–986.

Zoccolillo, M., Price, R.K., Ji,. T. and Hwu, H.G. (1999). *Antisocial Personality Disorder: Comparisons of Prevalence, Symptoms, and Correlates in Four Countries. Historical and Geographic Effects on Psychopathology* P. Cohen, C. Slomkowski and Robins L.L.L. (Eds) Erlbaum, New Jersey

Chapter 3

Bayesian Computational Methods for Inequality Constrained Latent Class Analysis *

Abstract

Exploratory latent class analysis is used to group responses of persons to items into classes such that persons with similar responses are assigned to the same class. Before the analysis, researchers may have competing theories with respect to the nature and the number of classes. After the analysis, the latent classes have to be interpreted. Using an exploratory approach, selecting which theory is best given the data, is a difficult task. In this paper an approach is proposed that can be used to select the best of these theories. It will be shown that a theory can be translated in a constrained latent class model (CLCA) using inequality constraints. This can be done for several (possibly competing) theories. Subsequently, fit-measures can be used to determine which model (and thus which theory) is supported most by the data. The approach will be illustrated using data from the Piagetian balance scale task.

*This chapter has been published as Laudy, O. and Hoijtink, H. (2005) *Bayesian Computational Methods for Inequality Constrained Latent Class Analysis* In: New Development in Categorical Data Analysis for the Social and Behavioral Sciences (eds: Van der Ark, A, Croom, M.A. and Sijtsma, K) Erlbaum, Londen

3.1 Introduction

Exploratory Latent Class Analysis (ELCA)(Clogg, 1981; Goodman, 1974; Haberman, 1988; Vermunt, 1997) is used to group responses x_{ij} of persons $i = 1, ..., N$ to items $j = 1, ..., J$ into classes $q = 1, ..., Q$ such that persons with similar responses are assigned to the same class. In this paper we will restrict ourselves to dichotomous data $x_{ij}\epsilon\{0, 1\}$. Each class q is characterized by J class specific probabilities π_{qj} indicating the probability of the response '1' on item j in class q and a weight ω_q indicating the unconditional probability that a person's latent class membership τ equals q. Let $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$, $\boldsymbol{\theta} = [\boldsymbol{\omega}, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_Q]$, $\boldsymbol{\pi}_q = [\pi_{q1}, ..., \pi_{qJ}]$, $\boldsymbol{x}_i = [x_{i1}, ..., x_{iJ}]$ and $\boldsymbol{\omega} = [\omega_1, ..., \omega_Q]$. The density of the data given the parameters of ELCA is then given by

$$\begin{aligned} P(\boldsymbol{X}|\boldsymbol{\theta}) &= \prod_{i=1}^{N} P(\boldsymbol{x}_i \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} \Big[\sum_{q=1}^{Q} P(\boldsymbol{x}_i, \tau = q \mid \boldsymbol{\theta}) \Big] \\ &= \prod_{i=1}^{N} \left[\sum_{q=1}^{Q} \omega_q \prod_{j=1}^{J} \pi_{qj}^{x_{ij}} (1 - \pi_{qj})^{(1-x_{ij})} \right], \end{aligned} \tag{3.1}$$

A key question in ELCA is into how many homogeneous subgroups the sample should be divided? Usually fit measures (Everitt, 1988; Lin and Dayton, 1997) are used to determine which number of classes is optimal. Another question concerns the interpretation of the resulting classes. Sometimes classes can be interpreted independent of other classes. As will be illustrated in the next section, one class may account for persons with highly developed emotional skills, while an other class accounts for persons with highly developed social skills. It can also be that classes can be ordered with respect to one or more underlying dimensions (Croon, 1990). An example of the latter is an ELCA resulting in three latent classes that can be used to order persons with respect to different levels of social skills (a one-dimensional ordering). It even might be the case that the persons can be ordered with respect to two dimensions, for example, the combinations of levels of social skills and the levels of emotional skills.

A researcher using exploratory analysis behaves as if his research field has not yet been explored very thoroughly, and theories are not yet fully developed. After the execution of an exploratory analysis, a researcher has to determine whether the outcome is in accordance with an existing theory, or that a new theory is emerging. This approach has two drawbacks. First of all, it may not at all be clear which theory corresponds best to the outcomes. This may lead to over-interpretation and guessing. Secondly, scientific progress may be larger if the current state of affairs (existing knowledge and theories) are properly accounted for in the statistical models used for the analyses.

ELCA has been done in areas that have been thoroughly explored, and where theories are well developed, for example, Boom, Hoijtink and Kunnen (2001) and Jansen and Van der Maas (1997). They use ELCA to analyze data with respect to the Piagetian Balance Scale Task. In section 3.5.1 new data with respect to this task will be analyzed using Confirmatory Latent Class Analysis (CLCA). There it will also be shown how CLCA can be used to refine (the best of) the existing theories, that is, how a new theory can be generated using the old theory as the point of departure.

In this chapter, a specific form of CLCA will be proposed (Hoijtink and Molenaar, 1997; Hoijtink, 1998; Hoijtink, 2001). The approach allows a theory to be translated into a CLCA using inequality constraints among the parameters of the model. This can be done for several competing theories. Two fit measures will be presented that can be used to select the model that receives the most support from the data.

3.2 Translation of Theories into CLCA

Several models can be constructed using constraints of the following types for $q \neq q'$ and/or $j \neq j'$:

$$\pi_{qj} > \pi_{q'j'}, \tag{3.2}$$

$$\pi_{qj} < \pi_{q'j'}. \tag{3.3}$$

To start with a simple example, suppose that persons have to respond to ten items. The first five items can be answered using skills related to social qualities (e.g. do you think you have a good understanding of other people ?), the others using skills related to emotional qualities(e.g. do you easily succeed in managing yourself?). The answers to these questions are coded 1 (well-developed) and 0 (undeveloped). Thus, the response vector of each respondent has ten scores with realization 1 or 0. Suppose, several theories exist for these data. A researcher thinks skills related social and emotional are not distinct, leading to the conclusion that there are only two groups of persons: persons who have a higher (social/emotional) intelligence, and persons who have a lower intelligence. This *common intelligence theory* can be translated into a latent class model with two latent classes. The class specific probabilities for the first class are all high, the class specific probabilities for the second class are all low, thus meaning that the persons who have both well developed social and emotional skills are allocated in class one, and the less intelligent persons who have both less developed social and emotional skills are allocated in class two. In terms of restrictions (see Table 3.1): for the common intelligence theory, the class specific probabilities of all items in the first class are restricted to be larger than those of all items in the second class. Note that j will be used to indicate item numbers, ω_1 denotes the proportion of persons in class 1, and π_{2j} denotes the probability of responding '1' to item j in class 2.

Table 3.1: Inequality Constraints for the Common Intelligence Theory

Item type	Items	Restrictions		
Social	1-5	π_{1j}	$>$	π_{2j}
Emotional	6-10	π_{1j}	$>$	π_{2j}

Another researcher might not agree with the common intelligence theory and states that there are indeed two groups of persons, but one group has higher social related skills, while the other group has higher emotional related skills. From this *specific (social/emotional) intelligence theory* it can be inferred that in one class the probabilities of responding 'developed' - that is, the response indicates that the person has well-

developed skills - to the social items are higher than for the emotional items, while for the other class the probabilities of responding 'developed' to the emotional items are higher than for the social items. This theory can be translated into a CLCA as indicated in Table 3.2: the first five items in the first class have probabilities that are restricted to be larger than the first five items in the second class. The first five items in the first class are also restricted to be larger than the last five items in the first class. The last five items in the second class have probabilities that are restricted to be larger than the last five items in the first class. The last five items in the second class are restricted to be larger than the first five items in the second class.

Table 3.2: Inequality Constraints for the Specific Intelligence Theory

Item type	Items	Restrictions		
Social	1-5	π_{1j}	$>$	π_{2j}
		$\vee$		$\wedge$
Emotional	6-10	π_{1j}	$<$	π_{2j}

An alterative display of the inequality constraints for the 'specific intelligence' theory is given in Table 3.3. Here the inequality constraints are implicit, for example, a minus sign indicates a class specific probability is restricted to be smaller than all the class specific probabilities corresponding to a plus sign. This type of display will be used in section 3.5.1, where the display with inequality signs is too complicated or impossible. The inequality constrains are implicit: - $<$ +. Note that a minus sign is not restricted with respect to any other minus sign, and and a plus sign is not restricted with respect to any other plus sign.

Table 3.3: Alternative Display for the Specific Intelligence Theory

Item type	Items	Restrictions	
Social	1-5	+	-
Emotional	6-10	-	+

3.3 Estimates for the CLCA

In this section it will be explained how estimates of the parameters are obtained. The general algorithm is described by Gelfand, Smith and Lee (1992) and the direct application to CLCA can be found in Hoijtink (1998). The basic principle is to use the posterior distribution to obtain a sample of the model parameters. This sample can be seen as a discrete representation of the posterior distribution. With this sample, further calculations are easy, for example, the average of the sampled values is the expected a-posteriori (EAP) estimate of a parameter, and the 2.5-th and 97.5-th percentile of the sampled values constitute a 95% central credibility interval. Since it is not trivial to obtain a sample from a multivariate posterior distribution, the Gibbs sampler is applied. This algorithm renders a sample from the joint posterior of the parameters by repeatedly

sampling from conditional distributions, that is, the distribution of the parameter at hand, given all the other parameters.

3.3.1 Posterior distribution

The density of the data given the parameters of the model is given by Equation 3.1. For each model $k = 1, ..., K$, where K denotes the number of models under consideration, the set of inequality constraints will be denoted by H_k. The latter will be included in the posterior distribution via the prior distribution. In this chapter, all the priors are chosen to be uniform for all combinations of parameter values allowed by H_k. Note that since information about the models is included in the prior distributions via inequality constraints, in that respect the priors are informative. The conjugate prior for a (constrained) class specific probability is a (truncated) Beta(1,1) distribution. The conjugate prior for the class weights is a Dirichlet distribution parameterized such that a priori all combinations of weight values summing to one are equally likely, that is, Dirichlet$(\alpha_1, ..., \alpha_Q)$, with $\alpha_q = 1$. The resulting posterior $P(\boldsymbol{\theta} \mid \boldsymbol{X}, H_k)$ is proportional to the product of the density of the data $P(\boldsymbol{X} \mid \boldsymbol{\theta})$ and the (truncated) proportional prior $P(\boldsymbol{\theta} \mid H_k)$, that is,

$$P(\boldsymbol{\theta} \mid \boldsymbol{X}, H_k) \propto P(\boldsymbol{X} \mid \boldsymbol{\theta}) \times P(\boldsymbol{\theta} \mid H_k), \tag{3.4}$$

where $P(\boldsymbol{\theta} \mid H_k)$ has the value 1 if $\boldsymbol{\theta}$ is in accordance with the constraints imposed by H_k, and 0 otherwise.

3.3.2 Gibbs Sampler

The Gibbs sampler is an iterative procedure. In iteration $r = 0$ initial values have to be provided for the class weights and the class specific probabilities. Any set of values that is in agreement with the constraints imposed upon the parameters can be used. Each iteration $r = 1, \ldots, R$ consists of three steps:

Step 1: For $i = 1, \ldots, N$, sample class membership $\tau_{i,r} \in \{1, ..., Q\}$ from its posterior distribution given the current values (that is, the values sampled in iteration $r - 1$) of the class weights, the class specific probabilities and the data. This conditional posterior is a Multinomial distribution with probabilities

$$P(\tau_{i,r} = q \mid \boldsymbol{x}_i, \boldsymbol{\theta}_{r-1}) = \frac{P(\boldsymbol{x}_i, \tau_{i,r} = q \mid \boldsymbol{\theta}_{r-1})}{P(\boldsymbol{x}_i \mid \boldsymbol{\theta}_{r-1})} \tag{3.5}$$

for $q = 1, \ldots, Q$. Note that both the numerator and the denominator in the right-hand side of Equation 3.5 defined in Equation 1.

Step 2: For $q = 1, \ldots, Q$ and $j = 1, \ldots, J$, sample π_{qj} from its posterior distribution given the current values of τ_i for $i = 1, \ldots, N$, and the data and the constraints. This conditional posterior is a (truncated) Beta distribution with parameters $s_{qj,r} + 1$ and $N_{q,r} - s_{qj,r} + 1$, where $N_{q,r}$ denotes the number of persons allocated to class q in iteration r, and $s_{qj,r}$ denotes the number of persons allocated to class q in iteration r that respond 1 to item j. Note that the Beta distribution is truncated because the sampled value for π_{qj} is only acceptable if it is in accordance with the inequality constraints involving π_{qj}.

The naive way to do so is: sample from the correct (non-truncated) Beta distribution until a deviate is sampled that satisfies the constraints. However, this is quite inefficient when only a small range of the distribution is admissible. Inverse probability sampling solves this problem. Let π_{qj} be the parameter that has to be sampled from the truncated Beta distribution. The lowerbound a is given by the largest class specific probability that, according to the constraints imposed by the model at hand, must be smaller than π_{qj}. The upperbound b is the smallest class specific probability that, according to the constraints imposed by the model at hand, must be greater than π_{qj}. The sampling is achieved using a uniform (0,1) deviate U and the computation of

$$\pi_{qj} = \Phi^{-1}_{\pi_{qj}}[\Phi_{\pi_{qj}}(a) + U(\Phi_{\pi_{qj}}(b) - \Phi_{\pi_{qj}}(a))] \tag{3.6}$$

where $\Phi_{\pi_{qj}}(a)$ is the proportion of the conditional posterior distribution (a truncated Beta distribution) of π_{qj} below a and $\Phi_{\pi_{qj}}(b)$ is the proportion of conditional posterior distribution below b. $\Phi^{-1}_{\pi_{qj}}[.]$ denotes the inverse cumulative density evaluated at the argument. This procedure always renders a deviate from the conditional distribution at hand within the bounds a and b (Gelfand et al., 1992).

Step 3: Sample the class weights from their posterior distribution given the current values of τ_i for $i = 1, \ldots, N$. This posterior is a Dirichlet distribution with parameters $N_{1,r} + 1, \ldots, N_{Q,r} + 1$.

For all analyses executed in the chapter, the Gibbs sampler was run for 110,000 iterations. After a burn-in period of 10,000 iterations the values sampled in the second and third step of each 100-th iteration were saved (these iterations will be denoted using the superscript $m = 1, ..., M$). The result is $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^m, \ldots, \boldsymbol{\theta}^{1,000}$. This sample can be used to obtain estimates of the model parameters and the corresponding credibility intervals, taking into account the prior constraints. The expected a posteriori (EAP) estimate of a parameter is simply the average of the 1,000 values of that parameter sampled from the posterior distribution. A 95% central credibility for this parameter is given by the 2.5-th and 97.5-th percentile of the distribution of these 1,000 sampled values. In the next section it will be shown that it is easy to compute and evaluate fit measures using the sample from the posterior distribution.

3.4 Model Selection

After the translation of a number of competing theories into constrained latent class models, the support the data provide for each latent class model has to be determined. Three fit measures that can be evaluated using Bayesian computational methods (the marginal likelihood , posterior model probabilities and the pseudo likelihood ratio test) have been proposed in the literature (Kass and Raftery, 1995, Hoijtink, 2001). For a discussion of the performance of these measures in the context of inequality constrained models, the interested reader is referred to Hoijtink (1998; 2001). These fit measures will be discussed in the next sections.

3.4.1 Marginal Likelihood and Posterior Model Probabilities

Kass and Raftery (1995) present a comprehensive review of the marginal likelihood and posterior probability of a model. The basic idea behind the marginal likelihood factors is the same as the basic idea behind more familiar information criteria like AIC, CAIC and BIC. It can, for example, be shown (see Kass and Raftery, 1995), that the Bayesian Information Criterion (Schwarz, 1978) is an approximation of minus twice the logarithm of the marginal likelihood. Although not explicit in its formulation, the marginal likelihood, like the information criteria, contains a trade off between the likelihood of the parameters given the data and the number of parameters in the model.

In the remainder of this chapter, minus twice the logarithm of the marginal likelihood will be used:

$$-2 \log P(\boldsymbol{X}|H_k) = -2 \log \int_{\boldsymbol{\theta}_k} P(\boldsymbol{X} \mid \boldsymbol{\theta}_k) P(\boldsymbol{\theta}_k \mid H_k) \, d\boldsymbol{\theta}_k, \tag{3.7}$$

which brings comparisons of different models on the same scale as the familiar deviance statistics (Kass and Raftery, 1995). Loosely formulated, minus twice the logarithm of the marginal likelihood can be interpreted as the distance between the model at hand and the true model: the smaller its value, the smaller the distance.

There are many ways to compute Equation 3.7. In this paper, the approach developed by Raftery (1995) will be used. They suggest to sample 99% of the parameter vectors (in our case 990) from the posterior distribution parameter vectors, and to imagine that 1% of the parameter vectors (in our case 10) is sampled from an imaginary distribution where for each θ $P(\boldsymbol{X} \mid \theta)$ is equal to the marginal likelihood. An approximation of $-2 \log P(\boldsymbol{X} \mid H_k)$ is denoted as $-2 \log \hat{P}$ and obtained via a simple iterative algorithm based on the implicit equation:

$$-2 \log \hat{P} = -2 \log \left[\frac{10\hat{P} + \sum_{m=1}^{990} \frac{P(\mathbf{X}|\theta_k^m)}{.01 + P(\mathbf{X}|\theta_k^m)/\hat{P}}}{\hat{P} + \sum_{m=1}^{990} \frac{1}{.01 + P(\mathbf{X}|\theta_k^m)/\hat{P}}} \right] \tag{3.8}$$

If the prior probabilities of the K models under investigation are equal, that is, $P(H_k) = 1/K$ for $k = 1, \ldots, K$, the posterior probability of each model can be computed as:

$$P(H_k|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|H_k)}{\sum_{k=1}^{K} P(\boldsymbol{X}|H_k)}, \tag{3.9}$$

for $k = 1, \ldots, K$. The posterior model probability $P(H_k|\boldsymbol{X})$ denotes the support for model k in the total set of K models given by the data. In this chapter, both the marginal likelihood and the posterior probability of a model will be reported.

3.4.2 Pseudo Likelihood Ratio Test

Hoijtink (2001) shows that the likelihood ratio test (Everitt, 1988; Lin and Dayton, 1997) is not performing very well if the goal is to select the best of a number of inequality constrained models. The performance is much better if the pseudo likelihood ratio

statistic is used. This statistic is denoted by $D_k(\boldsymbol{X}, \boldsymbol{\theta}_k)$ and compares for each pair of items, the expected number of each possible pair of responses (that is, 00, 10, 01 and 11, respectively) to the corresponding observed number. Let $\boldsymbol{n}_{gh}^{vw}$ denote for items g and h the observed frequencies of the response pattern $X_g = v, X_h = w$ where $v, w \in \{0, 1\}$. Furthermore, let $\boldsymbol{m}_{gh|k}^{vw}$ denote the expected frequencies of these response patterns given $\boldsymbol{\theta}_k$. The pseudo likelihood ratio statistic is then defined as

$$\mathrm{D}_k(\boldsymbol{X}, \boldsymbol{\theta}_k) = -2 \sum_{g=1}^{J-1} \sum_{h=g+1}^{J} \sum_{v=0}^{1} \sum_{w=0}^{1} \boldsymbol{n}_{gh}^{vw} \log \frac{\boldsymbol{m}_{gh|k}^{vw}}{\boldsymbol{n}_{gh}^{vw}} \tag{3.10}$$

The expected frequencies conditional on $\boldsymbol{\theta}_k$ are computed using

$$\boldsymbol{m}_{gh|k}^{vw} = N \sum_{q=1}^{Q} \omega_q \left[\pi_{qg}^{v}(1 - \pi_{qg})^{1-v}\right] \left[\pi_{qh}^{w}(1 - \pi_{qh})^{1-w}\right] \tag{3.11}$$

The larger $D(\boldsymbol{X}, \boldsymbol{\theta}_k)$, the larger the discrepancy between the data $\boldsymbol{X}$ and model k. Since $\boldsymbol{\theta}_k$ is unknown, Equation (3.10) cannot be computed. The classical solution is to substitute the unknown quantity with the maximum likelihood estimate of $\boldsymbol{\theta}_k$. The Bayesian solution uses the posterior distribution of $\boldsymbol{\theta}_k$ (Rubin, 1984; Meng, 1994; Gelman, Meng and Stern, 1996). The posterior distribution summarizes the available information with respect to $\boldsymbol{\theta}_k$. The posterior can accurately be represented using a sample $\boldsymbol{\theta}_k^1, \ldots, \boldsymbol{\theta}_k^m, \ldots, \boldsymbol{\theta}_k^{1000}$ from this posterior. Each of the 1000 vectors $\boldsymbol{\theta}_k^m$ can be used to generate a replicated data matrix $\boldsymbol{X}_k^m$ that is in accordance with model k. The procedure is simple: for N persons class membership is sampled from a multinomial distribution with probabilities $\boldsymbol{\omega}_k^m$. Subsequently, the class specific probabilities $\pi_{q1,k}^m, \ldots, \pi_{qJ,k}^m$ of the class to which a person is assigned are compared with a vector of pseudo random numbers sampled from a U(0,1) distribution. If a class specific probability is larger than the corresponding random number, a person gives the response 1, otherwise the response 0 is given. This procedure is repeated for $m = 1, \ldots, 1000$.

For each $\boldsymbol{\theta}_k$ two discrepancies can be computed: $D(\boldsymbol{X}, \boldsymbol{\theta}_k^m)$, which is a discrepancy between the observed data and the model; and, $D(\boldsymbol{X}_k^m, \boldsymbol{\theta}_k^m)$, which is a discrepancy between replicated data and the model. If $D(\boldsymbol{X}_k^m, \boldsymbol{\theta}_k^m) \geq D(\boldsymbol{X}, \boldsymbol{\theta}_k^m)$, the discrepancy between the observed data and the model is equal to or smaller than the discrepancy between the replicated data and the model. The posterior predictive p-value is the proportion of 1000 comparisons for which $D(\boldsymbol{X}_k^m, \boldsymbol{\theta}^m) \geq D(\boldsymbol{X}, \boldsymbol{\theta}^m)$. The posterior predictive p-value is formally defined as

$$p_k = Pr[D(\boldsymbol{X}_k, \boldsymbol{\theta}_k) \geq D(\boldsymbol{X}, \boldsymbol{\theta}_k) \mid \boldsymbol{X}, H_k], \tag{3.12}$$

that is, the probability that the discrepancy between model k and a data-matrix $\boldsymbol{X}_k$ generated using model k is equal to or larger than the discrepancy between model k and the observed data matrix $\boldsymbol{X}$. The p_k is an absolute fit measure, that can be used to test the pseudo likelihood ratio statistic, that is, which can be used to determine whether model k accurately describes the data or not. Stated otherwise, analogous to the interpretation of classical p-values, values smaller than .05 indicate a lack of fit of the model, and values larger than .05 indicate that the model at hand was able to accurately

reproduce the observed frequencies.

3.5 Strategies to Solve the Piagetian Balance Scale Task

The balance scale task was recognized in the early eighties as a way of eliciting different rule-governed response patterns for proportionality reasoning (Siegler, 1981). A picture of a simplified balance scale is shown to children. While the beam is fixed, a number of identical weights are placed on each side at certain distances from the fulcrum. For each of a number of balances (the items) the children have to predict which side will tip, if any. The weights on the balance differ with respect to their number and distance to the center. The formal rule to obtain the correct answer is that the balance is in equilibrium when the product of the number of weights and the distance from the center is equal at both sides of the balance.

Applications of exploratory latent class analysis in the context of the balance task can be found in Boom, Hoijtink et al. (2001) and Jansen and Van der Maas (1997). New balance scale data will be used to determine which of the existing theories that explain children's responses to the items of the balance scale task is the best. Since the result is not conclusive, the best of these theories will be used as the point of departure for further theoretical developments.

Nearly 900 randomly selected Dutch children from 4 to 16 years old participated, with a mean age of 10.35 (standard deviation 2.82). The children were tested individually at home by students and did not receive feedback until the task had been finished. The assessment was part of a training procedure for psychology students. The students were prepared for this specific assessment in small groups and had to follow a strict assessment protocol.

3.5.1 Theories and Hypotheses about the Data

Siegler (1981) distinguished six types of problems. In *balance* problems, weight and distance are equal on both sides. In *weight* problems, the distance is the same on both sides but the number of weights is different. These first two problem types were not used in this study, since they do not differentiate between the postulated rules and were expected to be answered correctly by all children. In *distance* problems, the weight is the same on both sides but the distance is different. In conflict problems, more weight is on one side and greater distance on the other, such that the side with more weight falls (*conflict-weight* problem), the side with the greater distance falls (*conflict-distance* problem), or the balance remains horizontal (*conflict-balance* problem).

Siegler (1981) described four strategies or rules. Each of these strategies can be characterized by a specific pattern of scores on the different item types.

rule 1 Children will only consider the number of weights on each arm. Therefore it can be expected that they have a higher probability of correctly responding to the weight and the conflict-weight items than to the other item types.

rule 2 Children get a grasp of distance: when the number of weights is equal on both sides, they judge the influence of distance correctly, otherwise they ignore distance

and only consider the number of weights. For this strategy, it can be predicted that children have a higher probability of correctly responding to the weight, distance, and conflict-weight items than the other item types.

rule 3 Children will evaluate both the distance and the number of weights correctly, but if one side has more weights and the other side more distance they will be confused and guess. The probability of success will be at chance level (they make a random prediction) for all conflict type of problems.

rule 4 Children will apply the correct (torque) rule. The probability of responding correctly is high for all item types.

As can be seen in Table 3.4, the test used in this paper contains 19 items of the following types: five distance, four conflict-weight (originally 5 but one item had a printing error in the test booklet for half of the sample), five conflict-distance, and five conflict-balance.

In Table 3.4 a translation of Siegler's model into CLCA is elaborated upon. Note that the inequality constraints are implicitly shown: - indicates a low probability of correctly responding to the item, + a high probability of correctly responding to an item, and ± indicates a random prediction. All the probabilities associated with the - signs have to be smaller than the probabilities associated with the ± signs, and all the probabilities associated with the + signs have to be larger than probabilities associated with ±.

Table 3.4: Inequality Constraints for Siegler's Model

		Restrictions			
Item type	Items	Rule 1	Rule 2	Rule 3	Torque
Distance	1-5	-	+	+	+
Conflict Weight	6-9	+	+	±	+
Conflict Distance	10-14	-	-	±	+
Conflict Balance	15-19	-	-	±	+

Wilkenings and Anderson (1982) argue the existence of another strategy. The *addition rule* suggests that the number of weights and the number of distance intervals on the left are summed and compared to the sum of weights and distances on the right: the side with the greater sum is expected to tip. For the existing item types, we designed the items such that the addition rule could be detected because two conflict-weight items and two conflict-balance items evoke an incorrect response whereas the remaining conflict-weight and conflict-balance and all conflict-distance items evoke a correct response when this rule is applied to this set of items. Children applying the addition rule will have a low probability of correctly responding to the items that evoke an incorrect response, and a high probability of correctly responding to the remaining conflict-items. Normandeau, Larivee, Roulin, and Longeot (1989) argue that rule 3 of Siegler is not homogeneous. Their paper supports the existence of the addition rule and they introduce yet another rule: *qualitative proportion rule*. Children using this rule understand that more weights at a small distance from the fulcrum compensates for fewer weights at a far distance, resulting in a prediction of balance for all conflict problems. Thus, the qualitative

proportion rule predicts that all conflict-weight and conflict-distance items have low probabilities of being answered correctly, and all the conflict-balance items have a high (or higher) probability of a correct response. The five resulting latent classes are displayed in Table 3.5. Note that this table is comparable to Table 3.4, but extended to differentiate between the addition and non-addition items. Moreover, there can be seen that rule 3 has been split up into a latent class accounting for the addition rule and a latent class accounting for the qualitative proportion rule. In the current item set all conflict distance items were solvable using the addition rule.

Table 3.5: Inequality Constraints for Normandeau's Model

		Restrictions				
Item Type	Items	Rule 1	Rule 2	Add	QP	Torque
Distance	1-5	-	+	+	+	+
Conflict Weight	6,9	+	+	-	-	+
Conflict Weight Add	7,8	+	+	+	-	+
Conflict Distance	–	-	-	-	-	+
Conflict Distance Add	10-14	-	-	+	-	+
Conflict Balance	16,19	-	-	-	+	+
Conflict Balance Add	15,17,18	-	-	+	+	+

3.5.2 Results

The model selection measures have been computed for Siegler's and Normandeau's model. Note that the pseudo likelihood p-value indicates the absolute fit of the model. A p-value smaller than 0.05 indicates a lack of fit of the model, whereas the p-value is larger than 0.05, the model accurately reproduces the observed frequencies. The marginal likelihood can be interpreted as the distance between the model at hand and the true model: the smaller the value, the smaller the distance. The value of the marginal likelihood is on the same scale as the familiar deviance statistics. Two or more models can be compared using the value of the marginal likelihood. Since the marginal likelihood implicitly uses a parameter penalty, the model with the smallest marginal likelihood value has to be preferred. The marginal likelihoods of all models analyzed can also be transformed into the posterior model probability. This number indicates the support for each model in the total set of models given the data.

In Table 3.6, it can be seen from both the marginal likelihood and the posterior model probability that Normandeau's model is superior. However, as indicated by the p-value of the pseudo likelihood ratio test (smaller than .05), it is questionable whether Normandeau's model adequately reproduces the frequencies with which the response vectors are observed. This lack of fit could be due to existing strategies that are not predicted by the theory.

Figure 3.1 presents the class specific probabilities of Normandeau's model. On the horizontal axis the items are displayed, on the vertical axis the class specific probabilities. Classes one and two clearly represent rule 1 (only considering weight leads to a high probability of answering conflict-weight items (6-9) correct) and rule 2 (high probability of answering conflict-weight (6-9) and distance items (1-5) correct). These rule are

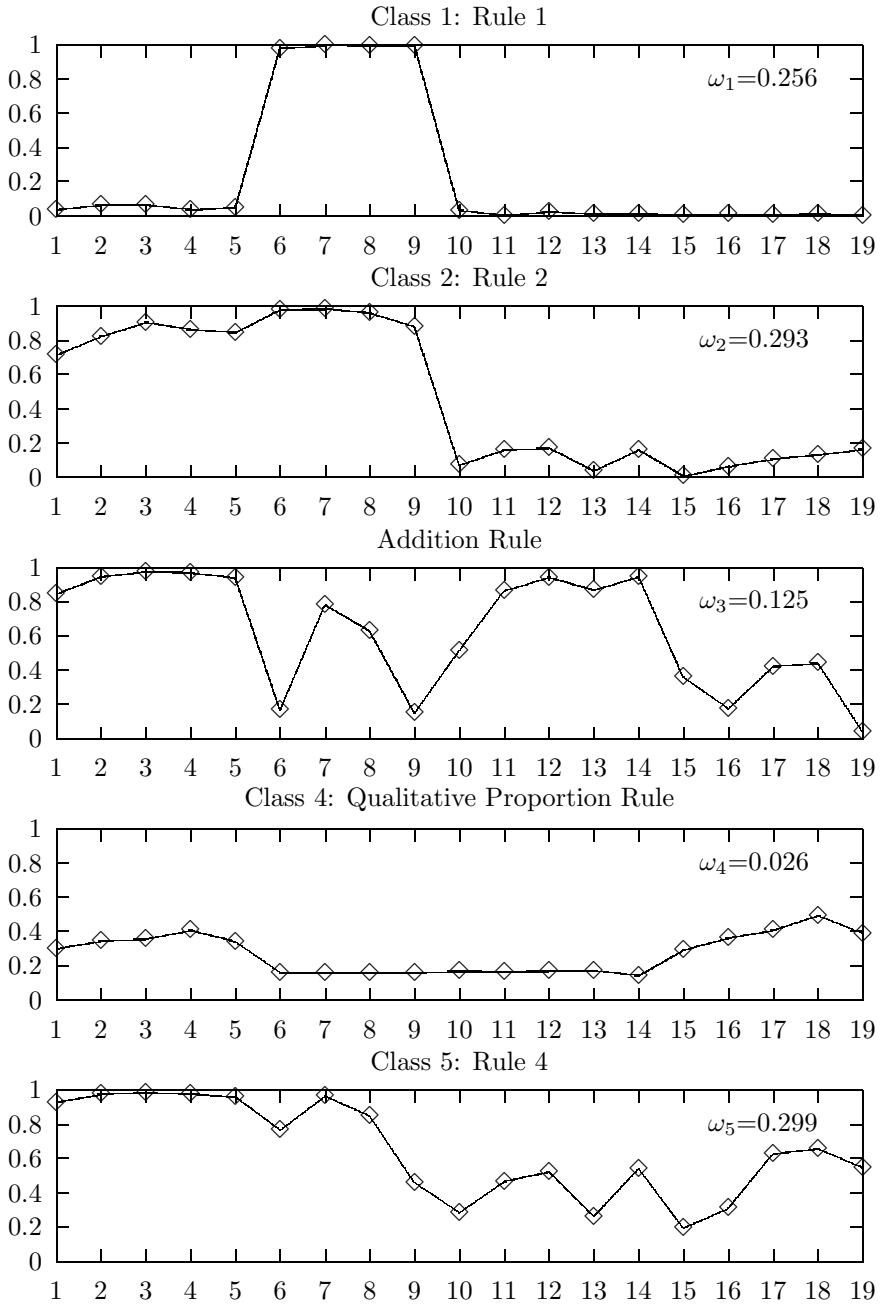


Figure 3.1: Class Specific Probabilities of the Normandeau Theory

Table 3.6: Evaluation of the Models for the Balance Scale Data

Model	Pseudo Likelihood	-2log Marginal Likelihood	Posterior Model Probability
Siegler	0.003	13421	0
Normandeau	0.019	13206	1.0

dominant, since a substantial part of the sample belongs to these classes. The third class represents the addition rule, although the probabilities for items 15, 17 and 18 are lower than expected for this rule.

Class four represents the qualitative proportion rule. As can be seen only a small proportion of the children belong to this class. Furthermore, although the class specific probabilities are in accordance with the constraints, especially the probabilities for the first and the last five items should have been higher to obtain a convincing representation of a qualitative proportion rule. It can be a 'true' strategy, but maybe it should be specified in more detail than simply by "all conflict items except conflict balance items are answered incorrectly". It could be, for example, that children do have an intuitive idea how distance and weight work together, but only when there is a large difference between the products of weight and distance on both sides.

Knowing that there are very few children that can actually solve the balance scale task, a class size of 29% for rule 4 seems extremely large. Furthermore, the class specific probabilities for this correct strategy are all predicted to be high, but as can be seen in the figure, this is not the case. Stated otherwise, class five does not yet give a convincing representation of rule 4.

The results for Normandeau's model are not conclusive. The p-value of the pseudo likelihood ratio test indicates that the data are not adequately reproduced. Furthermore, for classes four and five the class specific probabilities do not give a clear representation of the presumed underlying rule.

It could have been an option to represent the torque rule by a class for which $\pi_{qj} > .90$ for $j = 1, \ldots, J$. However, this value of 0.90 seems rather arbitrary. From the theory can be predicted that the probability of a correct response in class five has to be higher than the probabilities associated with a random prediction. Note that the method of testing models via the incorporation of inequality constraints on the model parameters explicitly shows that one chooses the best theory from a set of reasonable theories. This means that not all possible models are included, nor guarantees this procedure that the 'true' model is in the set. In the next section, it will be shown how this best theory can be used as the point of departure for theory refinement.

3.5.3 Theory Refinement

In this section, Normandeau's model will be extended with one and two unconstrained classes, respectively. This constitutes an example of scientific exploration using the current state of knowledge as the point of departure. Note that the results of this exploration are indefinite. To confirm the exploratory results, these findings have to be translated into inequality constrains and they have to be analyzed using new data.

Table 3.7: Refined Evaluation of the Models for the Balance Scale Data

Model	Pseudo Likelihood	-2log Marginal Likelihood	Post
Siegler	0.003	13421	-
Normandeau	0.019	13206	0.0
Normandeau + 1 class	0.054	12909	0.0
Normandeau + 2 classes	0.082	12837	1.0

As can be seen in Table 3.7, the Normandeau model with two unconstrained classes receives the most support from the data. Note, that the p-value of the pseudo likelihood ratio test now indicates that the data are adequately reproduced by this model. Furthermore, comparing the posterior probabilities of the three Normandeau models, it is clear that the variant with two extra unconstrained classes is superior.

As can be seen in Figure 3.2, the interpretation of the first four classes is similar to the interpretation given for the Normadeau model without extra classes. Note, however, that the probabilities for items 17 and 18 in class three have increased, that is, class three gives a better representation of the addition rule. The same holds for the first and last five items in class four, which now give a better representation of the Qualitative Proportion rule. Class five now represents rule 4 and contains only a small proportion of the children (as is expected).

Class six accounts for a fairly large proportion of persons. This class is difficult to associate with a strategy or rule. Our best guess is that it is class of children who are somewhere between rule 2 and the addition rule.

The second unrestricted class (class seven) is a global pattern, only grouping children that have in common that they do not consider the answer 'balance' an option (the last five items are almost never correctly answered). This class was also mentioned by Jansen and van der Maas (1997).

In the current version of the balance scale task, the items are chosen on the basis of being of a certain type (e.g. conflict balance item). The magnitude of the physical quantities is not varied systematically. We suggest that in further research one chooses the items of the same type more carefully, such that the role of the physical quantities can be asserted. For example, choose addition items within the conflict distance item such that the items vary from a big difference between addition torque to a small difference in a controlled way.

3.6 Conclusion

This chapter illustrated that theories can be included in latent class models using inequality constraints among the class specific probabilities. An example from the domain of developmental psychology was used to illustrate the resulting confirmatory latent class analysis. If, in a certain research domain, one or more theories exist, confirmatory latent class analysis has advantages over exploratory latent class analysis. First of all it provides a straightforward way to select the best of a number of competing theories. Secondly, as illustrated using the balance scale data, it allows theory refinement using the current

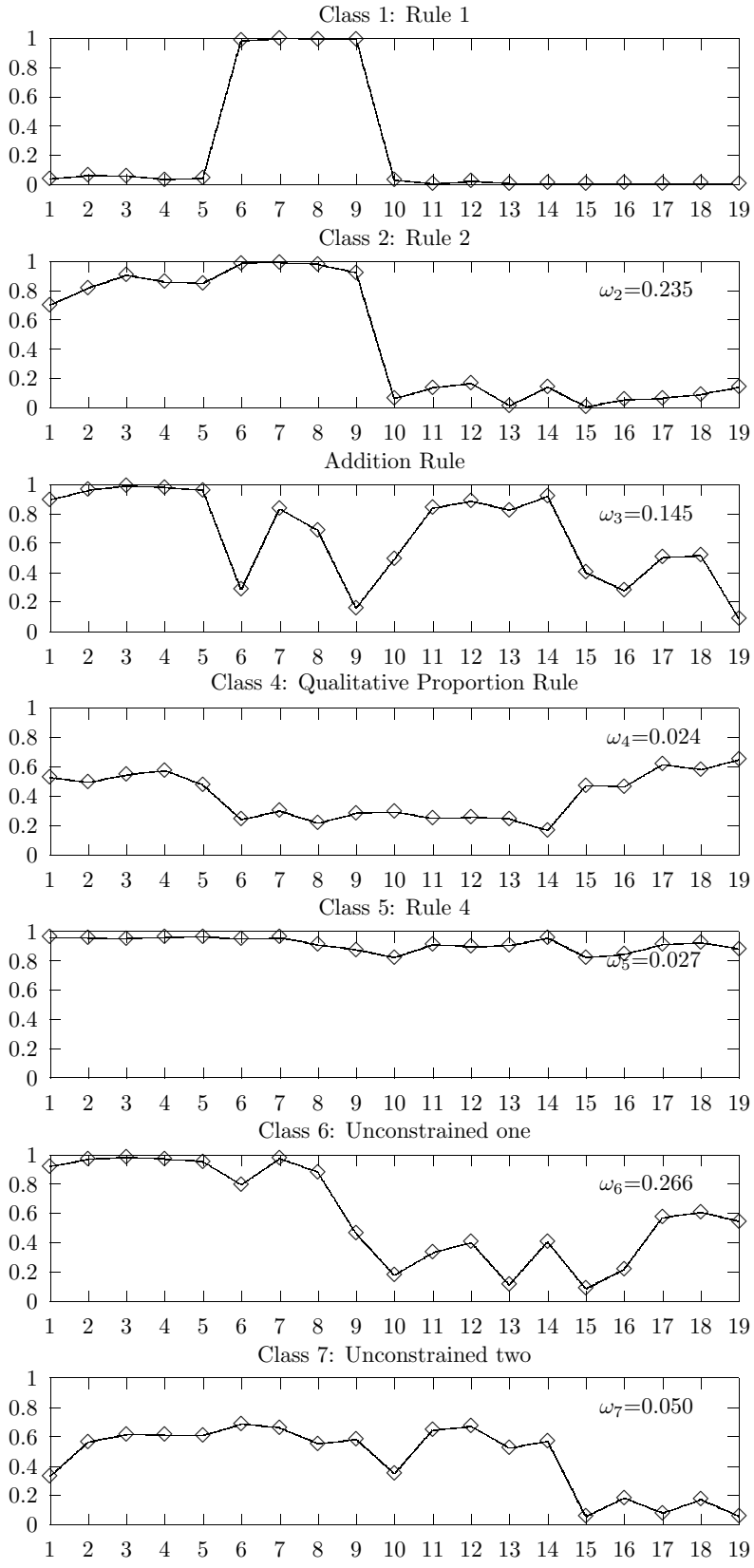


Figure 3.2: Class Specific Probabilities of the Normandeau Theory + Two Classes

state of knowledge as the point of departure.

Siegler and Chen (2002) mention some disadvantages of the LCA. One is the arbitrary alpha level of 5% and the unclear interpretation of it. This is acknowledged in Bayesian statistics for quite some time, and a solution has been found in the form of posterior model probability (Sellke, Bayarri and Berger, 2001). We use this solution, because instead of a probability of incorrectly rejecting the null hypothesis, the posterior model probability gives the probability of the data given the model among other models.

Currently user friendly software containing an implementation of the proposed approach is being developed. Readers interested in this software can write an e-mail to the first author. The e-mail should include a description of the research at hand, the data and the theories involved.

References

Boom, J., Hoijtink, H. and Kunnen, S. (2001). Rules in the balance: classes, strategies or rules for the balance scale task. *Cognitive Development,* **16**: 717–735.

Clogg, C.C. (1981). New developments in latent structure analysis. In D.J. Jackson and E. F. Borgatta (Eds), *Factor analysis and measurement in sociological research*, 215–246. Beverly Hills, CA: Sage.

Croon, M.A. (1990). Latent class analysis with ordered latent classes. *British Journal of Mathematical and Statistical Psychology,* **43**: 171–192.

Everit, B.S. (1988). A Monte Carlo investigation of the likelihood ratio test for number of classes in latent class analysis. *Multivariate Behavioral Research,* **23**: 531–538.

Gelfand, A.E., Smith, A.F.M. and Lee, T. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association,* **87**: 523–532.

Gelman, A., Meng, X. and Stern, H.S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica,* **6**: 733–807.

Goodman, L.A. (1974) The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach. *American Journal of Sociology,* **79**: 1179–1295.

Haberman, S.J. (1988). *Analysis of quantitative data, vol 2. New developments* New York: Academic Press

Hoijtink, H. and Molenaar, I. W. (1997). A multidimensional item response model: constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika,* **62**: 171–180.

Hoijtink, H. (1998). Constrained latent class analysis using the Gibbs sampler and posterior predictive p-values: applications to educational testing. *Statistica Sinica,* **8**: 691–711.

Hoijtink, H. (2001). Confirmatory latent class analysis: model selection using Bayes factors and (pseudo) likelihood ratio statistics. *Multivariate Behavioral Research,* **16**: 717–735.

Jansen, B.R.J. and Van der Maas, H.L.J. (1997). Statistical test of rule assessment methodology by latent class analysis. *Developmental Review,* **17**: 321–357.

Lin, T.H. and Dayton, C. M. (1997). Model selection information criteria for non nested latent class models. *Journal of Educational and Behavioral Statistics,* **22**: 249–264.

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

Meng, X. (1994). Posterior predictive p-values. *The Annals of Statistics,* **22**: 1142–1160.

Newton M.A. and Raftery A.E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society, Series B,* **56**: 3–48.

Normandeau, S., Larivee, S., Roulin, J. and Longeot, F. (1989). The balance-scale dilemma: Either the subject or the experimenter muddles through. *Journal of Genetic Psychology* **150**: 237–250.

Rubin, D.R. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics,* **12**: 1151–1172.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics,* **6**: 461–464.

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

Siegler, R.S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development* **46(2,189)**.

Siegler, R.S. and Chen, Z. (2002). Development of rules and strategies: Balancing the old and the new. *Journal of Experimental Child Psychology,* **81**: 446–457.

Vermunt, J.K. (1997). *Log-linear models for event histories. Advanced Quantitative Techniques in the Social Sciences Series, vol. 8*, Thousand Oakes CA: Sage

Wilkenings, F. and Anderson, N.H. (1982). Comparison of two rule-assessment methodologies for studying cognitive development and knowledge structure. *Psychological Bulletin,* **92**: 215–237.

Chapter 4

Bayesian Methods for the Analysis of Inequality Constrained Contingency Tables*

Abstract

We present a Bayesian methodology for analyzing inequality constrained models for contingency tables. The problem of interest lies in obtaining estimates of functions of cell probabilities subject to inequality constraints, testing hypotheses and selection of the best model. Constraints on conditional cell probabilities and on local, global, continuation and cumulative odds ratios are discussed. We use a Gibbs sampler to obtain a discrete representation of the posterior distribution of the inequality constrained parameters. Using this discrete representation, the credibility regions of functions of cell probabilities can be constructed. Posterior model probabilities are used for model selection and hypotheses are tested using posterior predictive checks. The Bayesian methodology proposed is illustrated in two examples.

*This chapter has been accepted for publication as Laudy, O. and Hoijtink, H. (2006) Bayesian Computational Methods for the Analysis of Inequality Constrained Contingency Tables *Statistical Methods in Medical Research* to appear in 2006

4.1 Introduction

In recent years, there has been growing interest in statistical models incorporating inequality constraints on model parameters. This is because the omnibus hypotheses can be replaced by more specific inequality constrained hypotheses Agresti and Coull (1998). In the extensive review by Agresti and Coull (2002), literature is discussed on order-restricted statistical models for contingency tables. A great deal of the literature on ordinal data deals with order constraints. Order constraints are considered a subset of inequality constraints, which is the topic of this paper. The approaches for ordinal data can be divided into two mainstreams. Firstly, ordered responses are traditionally parametrically modelled by using various logits, such as the adjacent-categories logits, the continuation-ratio logits and cumulative logits (Agresti, 1990; Goodman, 1985; McCullagh, 1980). Secondly, in the non-parametric approach, various types of orderings such as stochastic ordering, hazard ratio ordering and likelihood ratio ordering are used to compare probability distributions (El Barmi and Kochar, 1994; Dykstr, Lee and Yan, 1996; Charles, Lee and Yan, 2002). In larger contingency tables, these orderings give rise to various types of odds ratios, for example, cumulative odds ratios, local odds ratios, global odds ratios and continuation odds ratios (Agresti and Coull, 1998, 2002; Douglas and Fienberg, 1990).

In this paper, first we propose a simple estimation procedure for inequality constrained contingency tables that can handle inequality constraints with respect to functions of cell probabilities. The estimation procedure also yields credibility intervals. Examples include constraints on conditional probabilities and odds ratios. We also set up a framework that allows for hypothesis testing using posterior predictive p-values. The null hypothesis is the inequality constrained model, which is tested against the saturated model. We then demonstrate how to select the best model from a set of non-nested inequality constrained models using posterior model probabilities. In the next three sections, these topics are discussed in greater depth.

The first topic of this paper pertains to estimation. In classical statistics, the estimation procedures for inequality constraints usually rely on the theory of convex cones (Barlow, Bartholomew, Bremner and Brunk, 1972). In simple applications, they lead to the pool adjacent violators algorithm (Robertson, Wright and Dykstra, 1988; Bhattacharya, 1995; Tebbs and Swallow, 2003). In more complex situations, quadratic programming algorithms are used to obtain parameter estimates (Agresti and Coull, 1998). Bartolucci and Forcina (2002) use a constrained Fisher scoring algorithm. References to Bayesian estimation procedures for order constrained inference can be found in (Evans, Gilula, Guttman and Swartz, 1997; Hoijtink and Molenaar, 1997; Laudy and Hoijtink, 2004). This paper applies theory developed by Gelman , Carlin, Stern and Rubin (1995), who demonstrate the ease of inequality constrained Bayesian estimation procedures. An advantage of Bayesian methods is the discrete representation of the posterior. Using this representation, credibility regions of functions of inequality constrained cell probabilities are readily available.

The second topic is hypothesis testing. In our search, we have discovered that most of the focus is on a null hypothesis that restricts a parameter to zero against a constrained alternative (Agresti and Coull, 1998; Bartolucci and Scaccia, 2004). If this test renders

a small p-value, the null hypothesis is rejected, i.e., the data are not likely to have arisen under the parameter fixed at zero, and therefore the constrained alternative is taken as the model. In this paper, we focus on testing an inequality constrained model against a wider unconstrained alternative. A small p-value then indicates that the data are not likely to have arisen under the inequality constrained model. The Bayesian approach we advocate makes use of the posterior predictive checks introduced by Rubin (1984). Replicated data are generated from the posterior predictive distribution and compared to the observed data using a test statistic. Large differences between observed and replicated data indicate a model misfit.

The third topic is model selection. As far as we know, in classical statistics, hardly any model selection procedures are available for nested or non-nested models in the context of inequality constraints (Anraku, 1999). The problem is that most model selection criteria such as Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) have a penalty for the number of parameters, and in inequality constrained models, this number is not clear. In this paper, we use the traditional Bayesian model selection criterion: the marginal likelihood and subsequent posterior model probabilities (Kass and Raftery, 1995).

This paper is structured as follows. In Section 4.2, we illustrate the possibilities of the Bayesian methodology for applied researchers. The estimation procedure is discussed in Section 4.3. The notation for the contingency tables is introduced and we subsequently explain how to sample from the constrained posterior. We also demonstrate how to obtain parameter estimates and credibility intervals for functions of cell probabilities. Hypothesis testing is discussed in Section 4.4. We demonstrate that posterior predictive p-values can be an alternative to the classical p-value. In Section 4.5, we discuss model selection using the marginal likelihood and subsequent posterior model probabilities. Two examples are given to illustrate the method. The first example shows how inequality constraints can take the ordinal nature of variables into account. In the second example various types of odds ratios are constrained.

4.2 Motivating Examples

4.2.1 Example One: Oral Cancer

Consider a case control study by Rothman and Keller (1972). The data are displayed in Table 4.1 and consists of 483 patients with oral cancer and 456 controls. The cases and controls are classified according to their smoking behavior and alcohol consumption. The research question is whether alcohol consumption and smoking *increase* the probability of oral cancer.

Suppose, there are three models for these data. The first model states that with increasing alcohol consumption, the odds of oral cancer also increase, which can be translated into the model in Table 4.2 in the sub-table 'Theory one'. The table shows that the odds of oral cancer increases as alcohol consumption increases, for all categories of smokers.

The second model states that with an increasing number of cigarettes per day, the odds of oral cancer also increase, which can be translated into the model in Table 4.2 in

Table 4.1: Cross-Classification of Oral Cancer, Alcohol Consumption and Smoking Behavior, for Cases/Controls

Cigarettes	Alcohol (oz/day)			
(equiv./day)	0	0.1-0.3	0.4-1.5	1.6+
0	10/38	7/27	4/12	5/8
1-19	11/26	16/35	18/16	21/29
20-39	13/36	50/60	60/49	125/52
40+	9/8	16/19	27/14	91/27

Table 4.2: Models for the Oral Cancer Data

		Alcohol						
	Smoking	0		0.1-0.3		0.4-1.5		1.6+
Theory 1	0	π_{111}/π_{112}	$<$	π_{121}/π_{122}	$<$	π_{131}/π_{132}	$<$	$\pi_{141}\pi_{142}$
	1-19	π_{211}/π_{212}	$<$	π_{221}/π_{222}	$<$	π_{231}/π_{232}	$<$	$\pi_{241}\pi_{242}$
	20-39	π_{311}/π_{312}	$<$	π_{321}/π_{322}	$<$	π_{331}/π_{332}	$<$	$\pi_{341}\pi_{342}$
	40+	π_{411}/π_{412}	$<$	π_{421}/π_{422}	$<$	π_{431}/π_{432}	$<$	$\pi_{441}\pi_{442}$
Theory 2	0	π_{111}/π_{112}		π_{121}/π_{122}		π_{131}/π_{132}		$\pi_{141}\pi_{142}$
		$\wedge$		$\wedge$		$\wedge$		$\wedge$
	1-19	π_{211}/π_{212}		π_{221}/π_{222}		π_{231}/π_{232}		$\pi_{241}\pi_{242}$
		$\wedge$		$\wedge$		$\wedge$		$\wedge$
	20-39	π_{311}/π_{312}		π_{321}/π_{322}		π_{331}/π_{332}		$\pi_{341}\pi_{342}$
		$\wedge$		$\wedge$		$\wedge$		$\wedge$
	40+	π_{411}/π_{412}		π_{421}/π_{422}		π_{431}/π_{432}		$\pi_{441}\pi_{442}$
Theory 3	0	π_{111}/π_{112}	$<$	π_{121}/π_{122}	$<$	π_{131}/π_{132}	$<$	$\pi_{141}\pi_{142}$
		$\wedge$		$\wedge$		$\wedge$		$\wedge$
	1-19	π_{211}/π_{212}	$<$	π_{221}/π_{222}	$<$	π_{231}/π_{232}	$<$	$\pi_{241}\pi_{242}$
		$\wedge$		$\wedge$		$\wedge$		$\wedge$
	20-39	π_{311}/π_{312}	$<$	π_{321}/π_{322}	$<$	π_{331}/π_{332}	$<$	$\pi_{341}\pi_{342}$
		$\wedge$		$\wedge$		$\wedge$		$\wedge$
	40+	π_{411}/π_{412}	$<$	π_{421}/π_{422}	$<$	π_{431}/π_{432}	$<$	$\pi_{441}\pi_{442}$

the sub-table 'Theory two'. This table shows that with increasing number of cigarettes, the odds of oral cancer also increase.

A third theory states that smoking and alcohol consumption both increase the odds of oral cancer. This theory can be translated into the model in Table 4.2 in the sub-table 'Theory three'. The table shows that this is a combination of Theory one and two.

Note that there is many debate over the shape of the dose-response curves in the area of human heath-risk assessment (Conolly and Lutz, 2004). Common models use monotone increasing or decreasing response functions. However, for reaction of complex biological systems to a toxicant, non-monotone dose-response relations may be observed, showing a decrease at low dose followed by an increase at high dose (called U-shape or J-shape response curves). These models may also be included in the model set. A J-shaped response curve requires knowledge about the location of the lowest probability and can subsequently be modelled by requiring decreasing probabilities up to the lowest

probability, and increasing probabilities from the lowest probability. Models that deal with a specific form of the response curve may also be included. An example of the latter is a response curve that increases faster than linear. This can be modelled by inequality constraints that require increasing differences between the successive probabilities

In Section 4.6.1, the results of the oral cancer example are discussed. To investigate the absolute fit of the three models, hypotheses are tested. Agresti and Coull (1998) developed an exact test, for testing the independence model against the ordered model. We show a hypothesis test that tests the ordered model against the data. If the p-value of this test is larger than 0.05, the model at hand can accurately reproduce the observed frequencies.

A second question is which of the three models fits the data best. Model selection procedures are uncommon in the context of inequality constraints, since fit measures like BIC, AIC and Consistent Akaike Information Criterion (CAIC) can not be used, because the number of parameters is not known. In Section 4.5, the marginal likelihood and the subsequent posterior model probability are used to determine the relative fit of a model in a set of models at hand.

4.2.2 Example Two: Subarachnoid Hemorrhage

Table 4.3 is taken from Agresti and Coull (1998). The data describe the outcome of patients with a trauma due to subarachnoid hemorrhage. The patients' health is described in five categories, ranging from 'death' to 'good recovery'. Patients are in four treatment groups, one group is given a placebo and three groups are given increasing doses of a certain medicine. The question is whether increasing the dose of the medicine improves the health of the patient.

Table 4.3: Responses from a Clinical Trial Comparing Treatments on Extent of Trauma due to Subarachnoid Hemorrhage

	Outcome				
Treatment group	Death	Vegetative state	Major disability	Minor disability	Good recovery
Placebo	59	25	46	48	32
Low dose	48	21	44	47	30
Medium dose	44	14	54	64	31
High dose	43	4	49	58	41

Agresti and Coull (1998) discuss four definitions of odds ratios. The odds ratio is a measure of the association between two categorical variables. Let T(reatment) and O(utcome) be two ordinal variables. Let $i = 1, ..., I$ index the levels of variable T and $j = 1, ..., J$ the levels of variable O. Denote π_{ij} as the probability that variable $T = i$ and $O = j$, and let $\pi_{i|j} = Pr(T = i|O = j)$

When there is a positive association between variables T and O, the (I-1)(J-1) odds ratios are larger than one.

- The local odds ratio constraint (LO) is:

$$\frac{\pi_{j+1|i}/\pi_{j|i}}{\pi_{j+1|i+1}/\pi_{j|i+1}} > 1,$$

for $i = 1, ..., I-1$, $j = 1, ..., J-1$. This constraint is also referred to as total positivity of order two (Douglas and Fienberg, 1990; Douglas, Fienberg, Lee, Sampson and Whitaker, 1990). By generalizing the local odds ratio, various odds ratios can be defined, each partitioning the contingency table in a different way, leading to more flexible definitions of association.

Agresti and Coull also propose the following definitions of constrained odds ratios :

- The cumulative odds ratio constraint (CU)

$$\frac{\pi_{O\leq j|i}/\pi_{O>j|i}}{\pi_{O\leq j|i+1}/\pi_{O>j|i+1}} > 1$$

- The continuation odds ratio constraint (CO)

$$\frac{\pi_{j|i}/\pi_{O>j|i}}{\pi_{j|T>i}/\pi_{O>j|T>i}} > 1$$

- The global odds ratio constraint (GO)

$$\frac{\pi_{O\leq j|T\leq i}/\pi_{O>j|T\leq i}}{\pi_{O\leq j|T>i}/\pi_{O>j|T>i}} > 1$$

Note that the continuation odds ratio is defined in a way that variable O is the dependent variable. These constrained definitions of odds ratios differ with respect to the strength of the association. The odds ratios are ordered below according to their strength of association. If the LO holds, CO, CU, GO hold; if, for example, CO holds, only GO holds.

$$\text{LO} \Rightarrow \text{CO} \Rightarrow \text{GO}$$
$$\text{LO} \Rightarrow \text{CU} \Rightarrow \text{GO}$$

Suppose a researcher wants to test for a positive association between two variables, but does not know what type of odds ratios describe the association best. Agresti and Coull (1998) and Bartolucci and Scaccia (2004) provide exact tests for these models, testing the null hypothesis that no association exists against the alternative hypothesis that a specific type of odds ratio is larger than one. However, it cannot be decided on the basis of this p-value which of the models fits the data best. In Section 4.5, an approach will be shown to select the best model from a set of inequality constrained models. Furthermore, credibility intervals around the constrained odds ratios will be provided. This example will be analyzed in Section 4.6.2.

4.3 Model Estimation

4.3.1 Posterior Distribution

We use the standard notation for contingency tables with three variables A, B, and C with respective indexes $i = 1, ..., I$, $j = 1, ..., J$, and $k = 1, ..., K$. Generalizations to tables with more than three dimensions are straightforward. The total number of cells is given by $T = I \times J \times K$. Let f_{ijk} denote the observed frequency for cell (i, j, k). Let π_{ijk} denote the cell probabilities in the contingency table. Let $\boldsymbol{f} = \{f_{ijk} | i = 1, ..., I, j = 1, ..., J, k = 1, ..., K\}$ and $\boldsymbol{\pi} = \{\pi_{ijk} | i = 1, ..., I, j = 1, ..., J, k = 1, ..., K\}$. We assume that $\boldsymbol{f}$ has a multinomial distribution, $\boldsymbol{f}|\boldsymbol{\pi} \sim M(\boldsymbol{\pi}, N)$.

Let $\boldsymbol{\alpha}$ denote the parameters of the conjugate prior distribution, $Pr(\boldsymbol{\pi}|\boldsymbol{\alpha}) \sim Dirichlet(\boldsymbol{\alpha}) \propto \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} \pi_{ijk}^{\alpha_{ijk}-1}$. Before the analysis, a choice has to be made with respect to $\boldsymbol{\alpha}$. A common choice is a constant for each element of the vector $\boldsymbol{\alpha}$. In a saturated model, if an improper prior with hyperparameter $\boldsymbol{\alpha} = \mathbf{0}$ is used, the posterior mean is $\pi_{ijk} = f_{ijk}/N$. However, a hyperparameter of zero will never be used, since the calculation of the marginal likelihood requires a proper prior. Another common choice is $\boldsymbol{\alpha} = \mathbf{1}$. This leads to the estimate $\pi_{ijk} = f_{ijk}/N$ for the posterior mode. In the examples, a common value of one is used for the prior.

Denote inequality constraint z as $r_z(\boldsymbol{\pi})$ for $z = 1, ..., Z$. The joint constraints are $R(\boldsymbol{\pi}) = (r_1(\boldsymbol{\pi}), ..., r_Z(\boldsymbol{\pi}))$. The inequality constraints are accounted for in the prior distribution as follows

$$Pr(\boldsymbol{\pi}|R(\boldsymbol{\pi}), \boldsymbol{\alpha}) = \frac{Pr(\boldsymbol{\pi}|\boldsymbol{\alpha}) I_{\boldsymbol{\pi} \epsilon R(\boldsymbol{\pi})}}{\int Pr(\boldsymbol{\pi}|\boldsymbol{\alpha}) I_{\boldsymbol{\pi} \epsilon R(\boldsymbol{\pi})} d\boldsymbol{\pi}.}$$

Denote $I_{\boldsymbol{\pi} \in R(\boldsymbol{\pi})}$ as the indicator function over the inequality constraints, where $I_{\boldsymbol{\pi} \in R(\boldsymbol{\pi})} = 1$ if $\boldsymbol{\pi}$ is in accordance with $R(\boldsymbol{\pi})$, otherwise it is zero. The posterior for the contingency table is

$$P(\boldsymbol{\pi}|\boldsymbol{f}) \propto \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} \pi_{ijk}^{f_{ijk}+\alpha_{ijk}-1} I_{\boldsymbol{\pi} \in R(\boldsymbol{\pi})},$$

restricted such that $\sum_{ijk} \pi_{ijk} = 1$.

Since the cell probabilities are restricted to sum to one, the cell probabilities cannot be sampled successively. Narayanan (1990) shows that a Dirichlet distribution can be parameterized into a gamma distribution in a way that the sampling procedure is simplified. Let $\gamma_{ijk} \sim Gamma(f_{ijk} + \alpha_{ijk}, 1)$, then the vector $(\pi_{111}, ..., \pi_{222})$ where $\pi_{ijk} = \gamma_{ijk}/\gamma_{+++}$ is distributed as $Dirichlet(f_{111} + \alpha_{111}, ..., f_{222} + \alpha_{222})$. Under this parameterization the posterior becomes:

$$P(\boldsymbol{\pi}|\boldsymbol{f}) \propto \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} \left(\frac{\gamma_{ijk}}{\gamma_{+++}} \right)^{f_{ijk}+\alpha_{ijk}-1} I_{\boldsymbol{\gamma} \in R(\boldsymbol{\gamma})}, \tag{4.1}$$

where $\gamma_{+++} = \sum_{ijk} \gamma_{ijk}$. Note that the inequality constraints on $\boldsymbol{\pi}$ are now also parameterized into inequality constraints on $\boldsymbol{\gamma}$, since $R(\boldsymbol{\pi}) = R(\frac{\boldsymbol{\gamma}}{\gamma_{+++}}) = R(\boldsymbol{\gamma})$. For example, suppose we have the following ordering $\pi_{111} > \pi_{112} > \pi_{113}$ which equals $\frac{\gamma_{111}}{\gamma_{+++}} > \frac{\gamma_{112}}{\gamma_{+++}} > \frac{\gamma_{113}}{\gamma_{+++}}$. This reduces to $\gamma_{111} > \gamma_{112} > \gamma_{113}$.

It is now explained how the model parameters are sampled from the constrained

posterior distribution.

A sample is taken from $\gamma_{ijk} \sim Gamma(f_{ijk} + \alpha_{ijk}, 1)$ for $i = 1, ..., I, j = 1, ..., J, k = 1, ..., K$. Without constraints, the parameters are independent, and an i.i.d. sample from the successive gamma distributions provides a draw from the posterior distribution. If inequality constraints are incorporated in the model, the successive draws of the gamma distribution are dependent, and we will resort to the Gibbs sampler (Gelman , Carlin, Stern and Rubin, 1995; Gelfand, Smith and Lee, 1992).

The Gibbs sampler is an iterative procedure. Suppose we want to take a sample from the joint posterior distribution of γ under an inequality constraint $R(\boldsymbol{\gamma})$. Taking the full conditionals required for the Gibbs sampler reduces the multivariate constraints to univariate constraints. In the Gibbs sampler, the gamma distribution of a parameter thus has a lower and an upper bound, conditional upon the constraints and values of all the other parameters. These values are denoted by $bounds(\gamma_{ijk}^{(t)}) = (l, u)$, where l denotes the maximum lower bound and u the minimum upper bound over all Z constraints in iteration t. To keep the notation simple, we omit indexes for l and u.

In iteration $t = 0$, initial values have to be provided for γ_{ijk}. Any set of values that is in agreement with the constraints imposed upon the parameters can be used. Each iteration $t = 1, \ldots, T$ consists of the following steps:

(i) Cycle Step 1 $\forall\ i, j, k$

(i_a) Calculate $bounds(\gamma_{ijk}^{(t)}) = (l, u)$ given the current values of the parameters, and

(i_b) Sample $\gamma_{ijk}^{(t+1)} \sim Gamma(f_{ijk} + \alpha_{ijk}, 1|l, u)$

(ii) Compute $\pi_{ijk} = \gamma_{ijk}/\gamma_{+++}\ \forall\ i, j, k$ and deliver $(\pi_{111}, ..., \pi_{ijk}, ..., \pi_{IJK})$ as a draw of the correct truncated posterior.

citetGelfand Smith Lee 1992 show that in iteration t as $t \to \infty$ under mild conditions, the Gibbs sampler provides parameters that come from the correct constrained joint posterior distribution.

The naive way to sample from a truncated gamma distribution is to sample from the non-truncated gamma distribution until a deviate is sampled that satisfies the constraints. However, this is quite inefficient if only a small range of the distribution is admissible. Inverse probability sampling solves this problem. Let γ_{ijk} be the parameter to be sampled from the truncated gamma distribution. The lower bound l is given by the largest γ that, according to the constraints imposed by the model at hand, must be smaller than γ_{ijk}. The upper bound u is the smallest γ that, according to the constraints imposed by the model at hand, must be greater than γ_{ijk}. The sampling is achieved using a uniform (0,1) deviate U and the computation of

$$\gamma_{ijk} = \Phi_{\gamma_{ijk}}^{-1}[\Phi_{\gamma_{ijk}}(l) + U(\Phi_{\gamma_{ijk}}(u) - \Phi_{\gamma_{ijk}}(l))] \tag{4.2}$$

where $\Phi_{\gamma_{ijk}}(l)$ is the proportion of the conditional posterior distribution of γ_{ijk} below l and $\Phi_{\gamma_{ijk}}(u)$ is the proportion of conditional posterior distribution below u. $\Phi_{\gamma_{ijk}}^{-1}[.]$ denotes the inverse cumulative density evaluated at the argument. This procedure always renders a deviate from the conditional distribution at hand within the bounds l and u (Gelfand, Smith and Lee, 1992).

With regard to the sampling procedure, in our experience 9 000 iterations and a burn-in of 1 000 iterations generally lead to stable estimates. The burn-in period will be longer, depending on how flexibly the sampled parameters can move through the parameter space, due to the constraints. The mixing of the MCMC sampler is visually checked by plotting $\boldsymbol{\pi}^{(t)}$ against t, for $t = 1, \ldots, T$ iterations (Cowles and Carlin, 1996).

For all the restrictions used in this paper, the estimation procedure is basically the same. The authors developed a software package with an interface that allows the users to specify a model and input the inequality constraints as text. To evaluate the inequality constraints in each sample of the posterior, a public domain function parser is used (Fortran 90 Funktionenparser, 2005). This parser reads the inequality constraints as text elements, and translates it into equations. Afterwards, the equations are solved numerically, using a simple root finder. Once the bounds are obtained, the further estimation procedure is explained above.

4.3.2 Parameter Estimates, Posterior Standard Deviations and Credibility Intervals

After removing the burn-in, the sample of T iterations from the constrained posterior can be summarized to obtain parameter estimates, posterior standard deviations and credibility intervals. By taking the averages over the T values, the Expected APosteriori estimates (EAP) are obtained. The posterior standard deviations are obtained by taking the standard deviations over the T values. The 90% central credibility intervals can be calculated by taking the 5th and 95th percentiles of the posterior sample. In the context of inequality constraints, the posterior distribution may be skewed. The credibility interval then correctly provides an asymmetric interval.

Furthermore, the summary measures can also be calculated for functions of the parameters. Let $g(\boldsymbol{\pi})$ describe a function of interest, then the T iterations of the posterior of the cell probabilities can be transformed according to $g(\boldsymbol{\pi})$. Summary measures can be calculated for this newly created vector. For example, suppose a contingency table with $I = J = K = 2$ is estimated using the constraint that the odds ratio of the collapsed table $(\pi_{11+}\pi_{22+})/(\pi_{21+}\pi_{12+}) > 1$. Let $g(\boldsymbol{\pi}) = (\pi_{11+}\pi_{22+})/(\pi_{21+}\pi_{12+})$. The vector $g^{(t)}(\boldsymbol{\pi})$ for $t = 1, ..., T$, contains the posterior distribution of these odds ratio. The EAP, the posterior standard deviation, and a c% central credibility interval can subsequently be calculated.

4.4 Model Fit

After obtaining the parameter estimates, the question is whether or not the inequality constrained model at hand accurately describes the data. A likelihood ratio test is used to investigate the absolute fit, testing the inequality constrained model (null-hypothesis) against the data.

As a test-statistic, the likelihood ratio test will be used. The likelihood ratio test for the constrained model against the unconstrained alternative is

$$LR = -2 \sum_i \sum_j \sum_k f_{ijk} \cdot \log \left(\frac{\pi_{ijk}}{f_{ijk}/N} \right)$$

The reference distribution of the test statistic under an inequality constrained model is a mixture of distributions, with the mixing probabilities generally unknown for more complex constraints. Robertson, Wright and Dykstra (1988) show how to obtain a p-value in a variety of models in the context of inequality constraints. However, a general approach evaluating arbitrary constraints is not available.

The posterior predictive p-value p_{pp} (Rubin, 1984; Meng, 1994) was introduced as a Bayesian alternative for the classical p-value. The idea is to replicate data under the null model and compare it with the observed data. Systematic differences between the observed and the replicated data indicate a misfit of the model. Replicated data are generated from the posterior predictive distribution,

$$P(\boldsymbol{f^{rep}}|\boldsymbol{f}) = \int P(\boldsymbol{f}^{rep}|\boldsymbol{\pi}) P(\boldsymbol{\pi}|\boldsymbol{f}) d\boldsymbol{\pi}$$

In practice, from each of the T sampled parameters vectors from the posterior, one replicated dataset is generated. For each replicate t, the likelihood ratio test is then calculated.

$$LR_t^{rep} = -2 \sum_i \sum_j \sum_k f_{ijk}^t \cdot log \left(\frac{\pi_{ijk}^t}{f_{ijk}^t/N} \right)$$

For each sampled parameter vector from the posterior, the likelihood ratio test is also calculated using the original data, and compared with the likelihood ratio test of the replicated data. The calculation of the posterior predictive p-value reduces to counting the number of more extreme values for the LR^{rep}:

$$p_{pp} = \frac{1}{T} \sum_{t=1}^{T} [LR_t^{rep} \geq LR_t]$$

Analogous to the interpretation of classical p-values, values smaller than .05 indicate a lack of model fit, and values larger than 0.05 indicate that the model at hand was able to reproduce the observed frequencies.

The advantage of this method is that a p-value can be obtained in virtually any situation without much effort. A disadvantage of this method is that the frequency properties may not be optimal (Bayarri and Berger, 2000).

4.5 Model Selection

If two or more plausible models are fitted to the data and all the models indicate a reasonable absolute fit, the question is which model fits the data best. Classical model selection procedures use information criteria like AIC, CAIC and BIC. The information criteria consist of two components, one part indicating the fit in terms of the likelihood and a penalty for the number of parameters. In the context of inequality constraints, however, the number of parameters is not exactly known so that

the traditional information criteria can not be used.

In Bayesian statistics, the marginal likelihood was proposed as a measure of fit by Jeffreys (1935). As is the case with information criteria, the marginal likelihood contains a trade-off between the likelihood of the parameters given the data and the number of parameters in the model, but the number of parameters is implicitly accounted for (Smith and Spiegelhalter, 1980). This makes the marginal likelihood appropriate to select between inequality constrained models. The marginal likelihood for model k is

$$P(\boldsymbol{f}|M_k) = \int P(\boldsymbol{f}|\boldsymbol{\pi}, M_k) Pr(\boldsymbol{\pi}|M_k) d\boldsymbol{\pi} \tag{4.3}$$

As is clear from (4.3), the likelihood is integrated with respect to the prior. A general result is that the marginal likelihood is sensitive to the prior, which is an undesirable property if one wants to compare the fit of models. However, for the models discussed in this paper, the marginal likelihood is not very sensitive to the prior. This is illustrated by small simulation study in Section 4.6.2.

For many models, there is no marginal likelihood in closed form. Approximations such as the Laplace approximation assume multivariate normality around the posterior mode, which is inappropriate in the context of inequality constraints. To do a Monte Carlo integration of (4.3), it might seem feasible to take a sample from the prior and evaluate the likelihood for the sampled parameter values. The marginal likelihood would then be given by the average of all the likelihoods. This, however, results in an unstable estimate, since sampled values from the prior from a region with high likelihood are unlikely, though these values contribute significantly to the marginal likelihood (Kass and Raftery, 1995).

A solution to the instability is to use importance sampling, in such a way that values with high likelihood are more frequently sampled. Generate a sample $\boldsymbol{\pi}^t$ for $t = 1, ..., T$ from importance sampling density $P^*(\boldsymbol{\pi})$. The Monte Carlo estimate of the marginal likelihood is then

$$\hat{P}(\boldsymbol{f}|M_k) = \frac{\sum_{t=1}^{T} w_t P(\boldsymbol{f}|\boldsymbol{\pi}^t, M_k)}{\sum_t w_t}, \text{ with } w_t = \frac{P(\boldsymbol{\pi}^t)}{P^*(\boldsymbol{\pi}^t)}$$

Newton and Raftery (1994) evaluate the case where the importance sampling function is the posterior. This leads to an expression of the marginal likelihood that has infinite variance, and is therefore less desirable. In this paper we use an other importance function suggested by Newton and Raftery (1994), i.e., a mixture of the posterior and the prior with mixing probabilities respectively $\delta = 99\%$ and $(1 - \delta) = 1\%$ or $p^*(\boldsymbol{\pi}) = \delta p(\boldsymbol{\pi}|\boldsymbol{f}, M_k) + (1 - \delta) p(\boldsymbol{\pi}, M_k)$.

To compare more than two model-fits, the posterior model probability is used. If the prior probability of the models is denoted by $P(M_k)$, the posterior model probability for model k is:

$$P(M_k|\boldsymbol{f}) = \frac{P(\boldsymbol{f}|M_k)P(M_k)}{\sum_{k=1}^{K} P(\boldsymbol{f}|M_k)P(M_k)}$$

Note that in this paper, all the models have equal prior probability. With respect to the interpretation of the posterior model probability, if two models are compared, a posterior

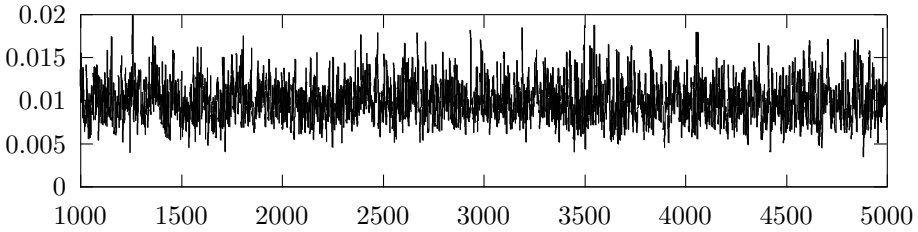


Figure 4.1: 5000 Iterations of π_{111}

probability of 0.5 for each of the models indicates that both models fit the data equally well, while a 0.9 probability for the one model against a 0.1 probability for the other model clearly shows a better fit for the first.

4.6 Examples

4.6.1 Example One

In this section, we continue with the Oral Cancer data example of Section 4.2. The research question is whether alcohol consumption and smoking increase the probability of oral cancer. The models to be fitted are in Table 4.2. An analysis was performed with 10 000 iterations. The first 1 000 iterations are considered the burn-in, and are excluded from the further calculations. In Figure 4.1, the mixing can visually be inspected for π_{111} for the first 4 000 samples. So far as visual inspection allows conclusions, the sample is converged to the stationary distribution. The same pattern was observed for all other analyses in this and the next example.

Table 4.4 shows the posterior predictive p-values for the constrained model against the data and the posterior model probabilities. With respect to the posterior predictive p-value, it can be concluded that none of the inequality constrained models are rejected. The posterior model probability indicates that the model that includes an effect for smoking as well as alcohol consumption is the best model.

In Table 4.5, the observed and estimated odds from the Theory three are displayed. The observed data show increasing odds in row and column direction alike. Since the alcohol+smoking model is the best model, the violations of the ordering in the observed data can be assumed to be due to sample fluctuations. The estimated odds show a monotone increase in the row and column directions alike, and they show that there are 3.45 as many heavy smokers and drinkers who develop oral cancer as heavy smokers and drinkers who do not develop oral cancer.

4.6.2 Example Two

In this section, the second example of Section 4.2 is continued. The research question is which of the association models fits the data best. Five models are fit to the data: M_1: unconstrained, M_2: LO >1, M_3: CU >1, M_4: CO >1 and M_5: GO >1. For the different definitions of the odds ratios, a table can be made of the observed odds ratios. This table shows that there are five local odds ratios smaller than one, two cumulative

Table 4.4: Posterior Predictive p-values and Posterior Model Probabilities for the Oral Cancer Data

Model	p_{pp}	$P(M_k\|\boldsymbol{f})$
M_0:Unconstrained	-	.02
M_1:Alcohol	.55	.16
M_2:Smoking	.61	.30
M_3:Alcohol+Smoking	.62	.52

Table 4.5: Observed/Estimated Odds for the Oral Cancer Example

	Alcohol			
Smoking	0	0.1-0.3	0.4-1.5	1.6+
0	0.26/0.20	0.26/0.28	0.33/0.41	0.63/0.66
1-19	0.42/0.33	0.46/0.50	1.13/0.81	0.72/1.02
20-39	0.36/0.44	0.83/0.83	1.22/1.27	2.40/2.34
40+	1.13/0.77	0.84/1.04	1.93/1.89	3.37/3.45

odds ratios smaller than one, three continuation odds ratios smaller than one and that all the global odds ratios are larger than one.

For the models at hand, the posterior model probabilities are shown in Table 4.6. The cumulative odds ratio model is clearly the best model, with 0.52 of the probability mass. This might look surprising, since the global odds ratio model had no observed odds ratios smaller than one. However, posterior probabilities have an implicit penalty for parameters, and the cumulative odds ratio model has fewer parameters than the global odds ratio model, since the cumulative odds ratio model implies the global odds ratio model.

As is noted in Section 4.3.2, the posterior distribution of the odds ratios can be calculated from the posterior distribution of the cell probabilities. From this posterior, the EAP can be found in Table 4.7, together with the 95 credibility interval. The table shows that all the odds ratios are larger than one, and the credibility intervals reflect the skewness of the posterior distribution.

Table 4.6: Posterior Model Probabilities for the hypotheses of the Subarachnoid Hemorrhage Data

Model	$P(M_k\|\boldsymbol{f})$
M_1: unconstrained	0.14
M_2: LO >1	.00
M_3: CU >1	.52
M_4: CO >1	.15
M_5: GO >1	.19

Table 4.7: Estimated Cumulative Odds Ratios, with 95% Central Credibility Interval

	1-2	2-3	3-4	4-5
1-2	1.26 (1.01-1.71)	1.30 (1.03-1.75)	1.25 (1.02-1.65)	1.21 (1.01-1.74)
2-3	1.24 (1.01-1.64)	1.45 (1.09-1.92)	1.30 (1.03-1.71)	1.15 (1.00-1.54)
3-4	1.17 (1.01-1.57)	1.45 (1.11-1.94)	1.29 (1.02-1.80)	1.40 (1.02-2.01)

4.6.3 Prior Sensitivity

Berger and Pericchi (1996) and Berger and Pericchi (2001) conclude that for some models the value of the posterior model probabilities heavily depend on the values of the prior. In this section is demonstrated that this is not likely for the models used in this paper. To investigate the prior sensitivity, all the models in Example two are analyzed repeatedly using different values for the prior. The prior was varied from $\boldsymbol{\alpha} = 1$ to $\boldsymbol{\alpha} = 0.2$. For each value of the prior, the models were analyzed 500 times. Table 4.8 shows the average values for the posterior model probabilities. It is firstly noted that the changes are small. If those changes allow an interpretation: if the value of the prior decreases, the posterior model probability of the best model (M_3) shows the largest increase, while the posterior model probability of the unconstrained model M_1 shows the largest decrease. This can be understood by realizing that a higher value of the prior smoothes the table towards independence. Since model M_3 fits the data, decreasing the value the prior leads to a better fit. Among the inequality constrained models, there is no model that describes independence. When the table is smoothes towards independence (a higher value for the prior), the unconstrained model, which includes independence, shows a better fit.

Table 4.8: Posterior Model Probabilities for Different Values of the Prior

	$\boldsymbol{\alpha}$				
Model	1	.8	.6	.4	.2
M_1: unconstrained	.14	.14	.13	.11	.09
M_2: LO >1	.00	.00	.00	.00	.00
M_3: CU >1	.52	.52	.53	.55	.57
M_4: CO >1	.15	.15	.15	.16	.16
M_5: GO >1	.19	.19	.19	.18	.18

4.7 Concluding Remarks

The Bayesian approach can provide a relatively straightforward solution to inequality constrained models for contingency table. First, we discussed how inequality constrained parameter estimates and central credibility intervals have been obtained. In contrast to the Bayesian method we proposed, in classical statistics, the estimation procedure and obtaining credibility intervals can be difficult or impossible. Furthermore, we have also demonstrated how posterior predictive p-values can be an alternative to classical

p-values, and how the best of a set of competing models can be selected using posterior model probabilities.

The posterior predictive p-value is currently a topic of discussion, since p-values are not uniformly distributed in some models (Bayarri and Berger, 2000). However, the posterior predictive p-values were used to test an inequality constrained hypothesis against the data. This approach is hardly available in the literature, and therefore the relatively simple approach to the posterior predictive p-value may be of use.

Frequency evaluations of the posterior predictive p-values and the posterior model probability are beyond the scope of this paper, and will be investigated in a future paper. Readers who are interested in the software can e-mail Olav Laudy at o.laudy@fss.uu.nl.

References

Agresti A. and Coull B.A. (1998) Order-restricted inference for monotone trend alternatives in contingency tables. *Computational Statististics and Data Analysis* **28**: 139–155.

Agresti A. *Categorical Data Analysis* Wiley: New York, NY, 1990.

Agresti A. and Coull B.A. (2002) The analysis of contingency tables under inequality constraints. *Journal of Statistical Planning and Inference* **107**(1-2): 45–73.

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

Bartolucci F. and Forcina A. (2002) Extended RC association models allowing for order restrictions and marginal modelling *Journal of the American Statistical Association* **97**: 1192–1199.

Barlow R.E. and Bartholomew D.J., Bremner J.M., Brunk H.D. *Statistical Inference Under Order Restrictions* Wiley: New York, NY, 1972.

Bartolucci F. and Scaccia L. (2004) Testing for positive association in contingency tables with fixed margins *Computational Statistics and Data Analysis* **47**: 195–210.

Bayarri M.J. and Berger J. (2000) P-values for composite null models [with discussion]. *Journal of the American Statistical Association* **95**: 1127–1142.

El Barmi H. and Kochar S. (1994) Likelihood ratio tests for bivariate symmetry against ordered alternatives in a square contingency tables. *Statistics and Probability Letters* **22**: 167–173.

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

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

Bhattacharya B. (1995) Restricted tests for and against the increasing failure rate ordering on multinomial parameters. *Statistics and Probability Letters* **25**: 309–316.

Charles Lee C. and Yan X. (2002) Chi-squared tests for and against uniform stochastic ordering on multinomial parameters. *Journal of Statistical Planning and Inference* **107**(1-2): 267–280.

Conolly R.B. and Lutz W.K. (2004) Nonmonotonic dose-response relationships: mechanic basis, kinetic modeling and implications for risk assessment *Toxicological Sciences* **77**: 151–157.

Cowles M.K. and Carlin B.P. (1996) Markov chain monte carlo convergence diagnostics: A comparative review *Journal of the American Statistical Association* **91**(434): 883–904.

Douglas R. and Fienberg S.E. (1990) An overview of dependency models for cross-classified categorical data involving ordinal variables. In *Topics in Statistical Dependence* Block, H.W., Sampson, A.R., Savits, T.S. (eds). Institute of Mathematical Statistics Lecture Notes: Hayward, CA. 167–188.

Douglas R., Fienberg S.E., Lee M.L.T., Sampson A.R. and Whitaker L.R. (1990) Positive dependence concepts for ordinal contingency tables. *Topics in Statistical Dependence* Block H.W., Sampson A.R., Savits T.S. (eds). Institute of Mathematical Statistics Lecture Notes: Hayward, CA. 189–202.

Dykstra R.L., Lee C.C. and Yan X. (1996) Multinomial estimation procedures for two stochastically ordered distributions. *Statistics and Probability Letters* **30**: 353–361.

Evans M., Gilula Z., Guttman I. and Swartz T. (1997) Bayesian analysis of stochastically ordered distributions of categorical variables. *Journal of the American Statistical Association* **92**: 208–214.

Fortran 90 Funktionenparser.
http://www.its.uni-karlsruhe.de/ schmehl/functionparser.html [31 Januari 2005].

Gelfand A.E. ,Smith A.F.M. and Lee T.M. (1992) Bayesian analysis of constrained parameter and trucated data problems using Gibbs sampling. *Journal of the American Statistical Association* **87**: 523–532.

Gelman A., Carlin J.B., Stern H.S. and Rubin D.B. *Bayesian Data Analysis* Chapman and Hall: London, UK., 1995;

Goodman L.A. (1985) The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Annals of Statistics* **13**: 10–69.

Jeffreys H. (1995) Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society* **31**: 203–222.

Hoijtink H. and Molenaar W. (1997) A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika* **62**: 171–189.

Laudy O. and Hoijtink H. Bayesian computational methods for inequality constrained latent class analysis. In: *New Developments in Categorical Data Analysis for the Social and Behavioral Sciences* Van der Ark, A., Croon, M., Sijtsma,K. (eds.) Erlbaum: Mahwah, NJ., 2005;

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

McCullagh P. (1980) Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society Series B* **42**: 109–142.

Meng X.L. (1994) Posterior predictive p-values. *Annals of Statistics* **22**: 1142–1160.

Narayanan A. (1990) Computer generation of Dirichlet random vectors. *Journal of Statistical Computation and Simulation* **36**: 19–30.

Newton M.A. and Raftery A.E. (1994) Approximate Bayesian inference by the weighted likelihood bootstrap (with Discussion). *Journal of the Royal Statistical Society Series B* **56**: 3–48.

Robertson T., Wright F.T. and Dykstra R.L. *Order Restricted Statistical Inference* Wiley: New York, NY, 1988; 229–230.

Rothman K. and Keller A. (1972) The effect of joint exposure to alcohol and tobacco on risk of cancer of the mouth and pharynx *Journal of Chronic Diseases Volume* **25**(12): 711–716.

Rubin D.B. (1984) Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician *Annals of Statistics* **12**(4): 1151–1172.

Smith A.F.M. and Spiegelhalter D.J. (1980) Bayes Factors and Choice Criteria for Linear Models. *Journal of the Royal Statistical Society Series B* **42**: 213–220.

Tebbs J. and Swallow W. (2003) More powerful likelihood ratio tests for isotonic binomial proportions. *Biometrical Journal* **45**: 618–630.

Chapter 5

Evaluation of Bayesian Model Selection Criteria in the Context of Inequality Constrained Contingency Tables*

Abstract

In this paper we present an evaluation of Bayesian model selection criteria in the context of inequality constrained contingency tables. Model selection is performed with either posterior predictive inference (L-criterion, DIC, posterior predictive checks) or prior predictive inference (posterior model probabilities, prior predictive checks and the prior predictive L-criterion). Three simulations are performed. First, the effect of the sample size on the frequency properties of the model selection criteria is investigated. Second, the magnitude of the effect is varied, and third, two models with different a-priori likelihood are compared. The results show that prior predictive model selection criteria are more sensitive to the a-priori likelihood of the models. Overall, the posterior model probability and the prior L-criterion seems to perform best: the results are stable across simulations, and although not always the most powerful for each simulation, their performance is satisfactory.

*This chapter has been submitted for publication as Laudy, O. and Hoijtink, H. Evaluation of Bayesian Model Selection Criteria in the Context of Inequality Constrained Contingency Tables *Bayesian Analysis*

5.1 Introduction

In recent years, new insights led many statisticians to believe that model selection is unavoidable (Raftery, 1995a,b; Berger and Sellke, 1987; Sellke, Bayarri and Berger, 2001; Cohen, 1994; Kass, 1993). Laudy and Hoijtink (2006) show an approach to model selection in the context of inequality constrained contingency tables. In this context, the traditional information criteria can not be used. Model selection criteria like the Bayesian Information Criterion (BIC, Schwarz, 1978; Raftery, 1986), and Akaike's Information Criterion (AIC, Akaike, 1974) use a function of the number of parameters as a penalty for model complexity. For inequality constrained models, however, the number of parameters is not known. An exception is Anraku (1999) who shows an approach to model selection of means under simple ordering using classical information criteria. Laudy and Hoijtink (2006) use the marginal likelihood as a model selection criterion. The marginal likelihood uses an implicit penalty for model complexity (Newton and Raftery, 1994; Raftery, 1995a,b). However, its frequency properties are not yet fully explored and compared to other Bayesian alternatives in the context of inequality constrained models. This comparison is the subject of this paper.

The following measures are investigated: posterior predictive checks (Rubin, 1984; Meng, 1994; Gelman, Meng and Stern, 1996), prior predictive checks (Box, 1980), the L-criterion (Laud and Ibrahim, 1995; Gelfand and Gosh, 1998), the DIC (Spiegelhalter et. al., 2002; Van der Linde, 2005), posterior model probabilities (Jeffreys, 1961; Kass and Raftery, 1995) and the prior predictive L-criterion, which is the prior predictive counterpart of L-criterion that is presented in the literature.

These model selection criteria can be categorized into prior predictive model selection criteria and posterior predictive model selection criteria. Posterior predictive model selection criteria evaluate the observed data with respect to the posterior distribution of the model parameters, while prior predictive model selection criteria evaluate the observed data with respect to the prior distribution of the model parameters.

In this paper, the model selection criteria are evaluated using a simulation study. In the simulations, data are repeatedly sampled from a population. Two or more competing models are specified. Subsequently, for each model the model selection criteria are calculated and it is evaluated how often each criterion picks the correct model.

Three simulations are performed. The first simulation investigates the effect of sample size. In the second simulation, sample size is held constant and the magnitude of the effect is varied. In the former two simulations, attention is devoted to three aspects: comparing two inequality constrained models, comparing an unconstrained model to an inequality constrained model that correctly describes the population, and comparing an unconstrained model to an inequality constrained model that incorrectly describes the population. In the third simulation, consists of two parts. First, it is investigated how the model selection criteria perform when the model set consists of two models that correctly describe the population. Second, it is investigated how the model selection criteria perform when the model set consists of two models that incorrectly describe the population.

The article is built up as follows. First, attention is devoted to the prior, likelihood and posterior of inequality constrained contingency tables. In Section 5.3 the differences

between prior and posterior predictive inference are discussed. In Section 5.4, the posterior predictive model selection criteria are discussed. In Section 5.5, the prior predictive model selection criteria are discussed. In Section 5.6, three simulation studies are executed. The paper is concluded with a discussion in Section 5.7

5.2 The Posterior of Constrained Contingency Tables

We use the standard notation for contingency tables with three variables A, B, and C with respective indexes $i = 1, ..., I$, $j = 1, ..., J$, and $k = 1, ..., K$. Generalizations to tables with more than 3 dimensions are straightforward. The total number of cells is given by $I \times J \times K$. Let f_{ijk} denote the observed frequency for cell (i, j, k). Let π_{ijk} denote the cell probabilities in the contingency table. Let $\boldsymbol{f} = \{f_{ijk} | i = 1, ..., I, j = 1, ..., J, k = 1, ..., K\}$ and $\boldsymbol{\pi} = \{\pi_{ijk} | i = 1, ..., I, j = 1, ..., J, k = 1, ..., K\}$. We assume that $\boldsymbol{f}$ has a multinomial distribution, $P(\boldsymbol{f}|\boldsymbol{\pi}) = M(\boldsymbol{\pi}, N)$.

Let $\boldsymbol{\alpha} = \{\alpha_{ijk} | i = 1, ..., I, j = 1, ..., J, k = 1, ..., K\}$ denote the parameters of the conjugate prior distribution, $Pr(\boldsymbol{\pi}|\boldsymbol{\alpha}) = Dirichlet(\boldsymbol{\alpha})$. Before the analysis $\boldsymbol{\alpha}$ has to be specified. In the simulations, $\boldsymbol{\alpha}$ is set to $\mathbf{1}$, which makes the posterior proportional to the likelihood (Shafer, 1997).

Denote inequality constraint z as $r_z(\boldsymbol{\pi})$ for $z = 1, ..., Z$. The joint set of constraints is given by $R(\boldsymbol{\pi}) = (r_1(\boldsymbol{\pi}), ..., r_Z(\boldsymbol{\pi}))$. Examples of $R(\boldsymbol{\pi})$ are orderings like $\pi_{111} > \pi_{112}$, $\pi_{112} > \pi_{113}$ or odd ratio restrictions like $\pi_{111}\pi_{122}/\pi_{112}\pi_{121} > 1$. The inequality constraints are accounted for in the constrained prior by normalizing the prior of the unconstrained model, that is,

$$Pr(\boldsymbol{\pi}|R(\boldsymbol{\pi}), \boldsymbol{\alpha}) = \frac{Pr(\boldsymbol{\pi}|\boldsymbol{\alpha}) I_{\boldsymbol{\pi} \epsilon R(\boldsymbol{\pi})}}{\int Pr(\boldsymbol{\pi}|\boldsymbol{\alpha}) I_{\boldsymbol{\pi} \epsilon R(\boldsymbol{\pi})} d\boldsymbol{\pi}},$$

where $I_{\boldsymbol{\pi} \in R(\boldsymbol{\pi})}$ denotes an indicator function, with $I_{\boldsymbol{\pi} \in R(\boldsymbol{\pi})} = 1$ if $\boldsymbol{\pi}$ is in accordance with $R(\boldsymbol{\pi})$, and 0 otherwise. The posterior for the contingency table is given by

$$P(\boldsymbol{\pi}|\boldsymbol{f}) \propto \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} \pi_{ijk}^{f_{ijk} + \alpha_{ijk} - 1} I_{\boldsymbol{\pi} \in R(\boldsymbol{\pi})}, \tag{5.1}$$

restricted such that $\sum_{ijk} \pi_{ijk} = 1$.

5.3 Prior Predictive versus Posterior Predictive Inference

In prior predictive inference, observed data are evaluated with respect to the prior distribution of the model. Large differences indicate that the observed data are unlikely to have occurred as a realization from the parameters of the prior distribution of the model. In posterior predictive inference, observed data are compared to the posterior distribution. Large differences indicate that the observed data are unlikely to have occurred as a realization from the parameters of the posterior distribution of the model. Philosophically, the prior predictive inference embraces the idea that both new parameters and data can occur. In contrast, posterior predictive inferences is based on the assumption that parameter values do not change when new data are collected

(Gelman, Meng and Stern, 1996).

Prior predictive inference is criticized on two aspects. First, the outcome is sensitive to the prior distribution, even if the sample size is large. Second, it is not defined for improper prior distributions. A critique against posterior predictive inference is that it makes double use of the data (Bayarri and Berger, 2000). The data are used to obtain the posterior distribution of the model, and subsequently, the data are evaluated with respect to this posterior distribution

It seems that the status of the prior is important here. In the view of a practical Bayesian, a prior has to be specified to profit from the ease of Bayesian computational methods. Non-informative and vague priors are used to obtain similar results to likelihood inference. Since the prior is rather arbitrary, as a matter of course, one only learns from the observed data. Hence, posterior predictive inference might be most appropriate. A more classical Bayesian viewpoint is that the prior is constructed using of knowledge from past research. In this type of research, one is interested in evaluating knowledge from past research in the light of newly collected data, hence, it might be more appropriate to use prior predictive inference.

As an illustrative example, consider a model for two binomial probabilities, π_1 and π_2. Suppose, data are collected and the sample probabilities show that $p_1 > p_2$. The upper figures show the situation for prior predictive inference, and the lower figures show the situation for posterior predictive inference. The solid contour lines in the figures show the likelihood of the data. In the upper figures, the shaded areas indicate the restricted parameter space. The dashed contour lines in the lower figures show the posterior distribution.

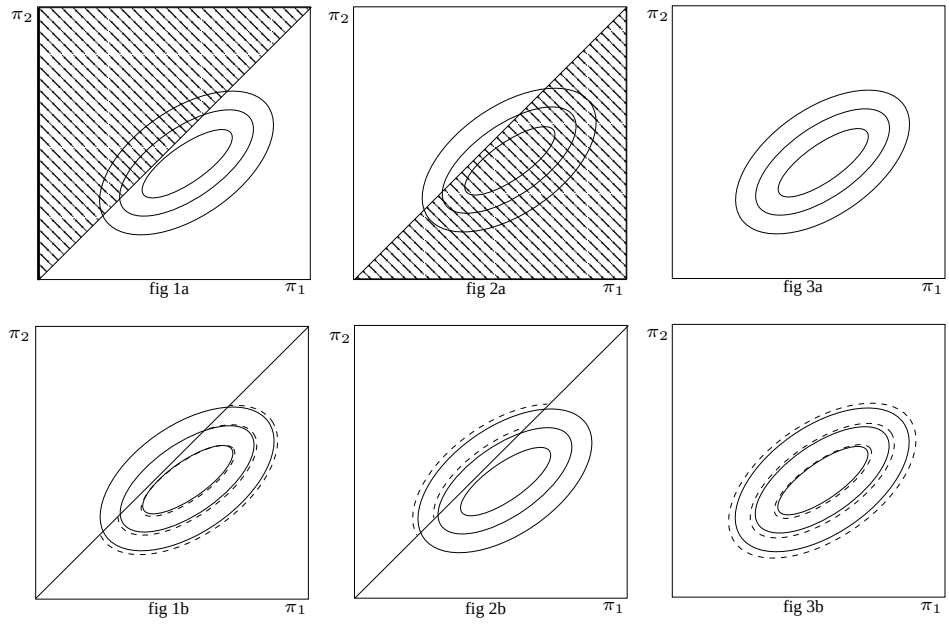


Figure 5.1: Differences between Prior Predictive Inference and Posterior Predictive Inference

Three models exist for these data. In Figure 5.1, model 1 is displayed in figures 1a and 1b, where π_1 is restricted to be larger than π_2, which is in accordance with the sampled

probabilities. In Figures 2a and 2b, model 2 is displayed, where the reverse holds: π_2 is restricted to be larger than π_1. Finally, model 3, the unrestricted model is displayed in Figures 3a and 3b.

The differences between prior and posterior predictive inference are discussed in terms of sampling. For prior predictive inference, a sample of parameters is drawn from the prior of the corresponding model. In the figures, this matches a sample from a non-informative prior, a uniform distribution, of the non-shaded areas in the upper figures. The observed data are evaluated with respect to the parameter samples from the prior. This can be done either a) evaluating the likelihood with respect to the sample of parameters from the prior, or b) comparing the observed data to replicated data that is generated using the sampled parameters from the prior. Without going into details (an elaboration will follow the next section), both methods result in a number indicating the fit. For the posterior predictive inference, a sample is drawn from the posterior (dashed lines, due to the non-informative prior similar to the likelihood) of the corresponding model and the observed data are evaluated with respect to the samples from the posterior using either a) or b).

Three situations are examined. First, suppose that model 1 and model 2 are to be compared. A larger part of the likelihood of model 1 is unrestricted than in model 2, hence procedure a) and b) will result in a better fit for model 1. Note that the result depends on the prior, irrespective of the sample size of the observed data. This implies that the prior must be chosen on well considered grounds. For the posterior predictive example, a larger part of the likelihood of model 1 is unrestricted than of model 2, leading to a better fit for model 1 than for model 2.

Second, model 1 and model 3 are compared. For the prior predictive example, in large parts of the prior for model 3, the likelihood is very low (method a), or replicated data generated from the sampled parameters from the prior show large differences with the observed data (method b), hence, model 1 will be selected. In the posterior predictive example, for a sample of parameters from the posterior of model 1, there are more parameters with a high likelihood compared to a sample of parameters from the posterior of model 3, hence model 1 will be the best model. In this situation, the double use of the data becomes clear: first the data are used to construct the posterior distribution, second, the data are evaluated with respect to the posterior.

Third, model 2 and model 3 are compared. For the prior predictive example, both model 2 and 3 will yield many samples of parameters from the prior with with a low likelihood, however, model 3 will also yield parameters from the prior with a high likelihood, hence model 3 will be the best model. For the posterior predictive example, samples from the posterior parameters of model 3 will have high likelihood values compared to samples from the posterior parameters of model 2, hence model 3 will fit best. In the simulations, these three situations are investigated.

5.4 Posterior Predictive Inference

5.4.1 Deviance Information Criterion

Model selection criteria like Bayesian Information Criterion (Schwarz, 1978) or Akaike's Information Criterion (Akaike, 1974) use a function of the number of parameters in the model as a penalty for model complexity. In model with inequality constraints, however, the number of parameters is not known, and therefore, the common model selection criteria are not available in the context of inequality constrained models. Spiegelhalter et. al. (2002) suggests an information criterion that is based on the posterior distribution of the deviance statistic.

$$D(\boldsymbol{\pi}|\boldsymbol{f}) = -2 \log P(\boldsymbol{f}|\boldsymbol{\pi}) + 2 \log h(\boldsymbol{f}),$$

where $P(\boldsymbol{f}|\boldsymbol{\pi})$ is the likelihood of data $\boldsymbol{f}$ given parameters $\boldsymbol{\pi}$. The part $h(\boldsymbol{f})$ is a function of the data alone, and thus is irrelevant for model selection.

The Deviance Information Criterion (DIC) has a part that captures the fit of the model, and a part that penalizes the complexity of the model. The fit of a model is defined as the expected a-posteriori deviance, $\bar{D} = I\!E[D(\boldsymbol{\pi}|\boldsymbol{f})]$. The complexity is captured by the number of effective parameters P_D. Spiegelhalter et. al. (2002) shows that a reasonable definition for P_D is the expected deviance minus the deviance evaluated at the expected a posteriori values (EAP) of the parameters.

$$p_D = I\!E[D(\boldsymbol{\pi}|\boldsymbol{f})] - D[I\!E(\boldsymbol{\pi}|\boldsymbol{f})] = \bar{D} - D(\bar{\boldsymbol{\pi}})$$

The Deviance information criterion is thus defined as

$$DIC = \bar{D} + p_D = 2\bar{D} - D(\bar{\boldsymbol{\pi}}) \tag{5.2}$$

Smaller values of the DIC indicate a better fit. In an MCMC run (Gelfand and Smith, 1990; Gelfand, Smith and Lee, 1992), the DIC is readily available. For each sampled $\boldsymbol{\pi}_t$, for $t = 1, ..., T$ from the posterior, the deviance $D(\boldsymbol{\pi}_t)$ can be calculated. The estimate of the DIC is given by twice the mean of the sampled deviances minus the deviance evaluated at the EAP of $\boldsymbol{\pi}$.

5.4.2 L-criterion

Laud and Ibrahim (1995) propose a model selection criterion that uses the posterior predictive space. The posterior predictive L-criterion also emerges using decision theoretic arguments (Gelfand and Gosh, 1998). Denote the posterior predictive distribution as

$$P(\boldsymbol{f}^{rep}|\boldsymbol{f}^{obs}) = \int P(\boldsymbol{f}^{rep}|\boldsymbol{\pi}) P(\boldsymbol{\pi}|\boldsymbol{f}^{obs}) d\boldsymbol{\pi}, \tag{5.3}$$

where $\boldsymbol{f}^{rep}$ denotes replicated data sampled from the posterior predictive distribution. First, a discrepancy function $d(\boldsymbol{f}^{rep}, \boldsymbol{f}^{obs})$ has to be chosen. Suppose there are L models,

then, for each model M_l, for $l = 1, ..., L$ at hand, one computes

$$I\!E[d(\boldsymbol{f}^{rep}, \boldsymbol{f}^{obs})|\boldsymbol{f}^{obs}, M_l],$$

and chooses the model that has the smallest expected discrepancy. For categorical data, Laud and Ibrahim (1995) suggest to take the deviance criterion as a discrepancy measure. Furthermore, to avoid extreme values, they add .5 to each frequency. The resulting deviance criterion is:

$$d(\boldsymbol{f}^{rep}, \boldsymbol{f}^{obs}) = 2 \sum_{cells} ((\boldsymbol{f}^{obs} + 0.5) \log \frac{\boldsymbol{f}^{obs} + 0.5}{\boldsymbol{f}^{rep} + 0.5}), \tag{5.4}$$

where $\sum_{cells}$ sums over all cells of the contingency table. Laud and Ibrahim (1995) show that the L-criterion consists of a fit measure and a penalty. In an MCMC run, the L-criterion is readily available. For each sampled $\boldsymbol{\pi}_t$, for $t = 1, ..., T$ from the posterior, new data $\boldsymbol{f}_t^{rep}$ are generated from a multinomial distribution $M(\boldsymbol{\pi}_t, N)$. The calculation reduces to taking the average of $d(\boldsymbol{f}^{obs}, \boldsymbol{f}_t^{rep})$ for $t = 1, ...T$.

5.4.3 Posterior Predictive Checks

Posterior predictive checks were introduced by Rubin (1984) and elaborated by Meng (1994) and Gelman, Meng and Stern (1996). The general idea is to replicate data from the posterior predictive distribution and compared it to the observed data using a test statistic or discrepancy measure. Large differences indicate a misfit of the model.

In this paper, the likelihood ratio is used as discrepancy measure. The likelihood ratio for the constrained model against the unconstrained alternative is given by

$$d(\boldsymbol{\pi}, \boldsymbol{f}^{obs}) = 2 \sum_{cells} (\boldsymbol{f}^{obs} \log \frac{\boldsymbol{f}^{obs}}{\boldsymbol{\pi} \cdot N}) \tag{5.5}$$

Replicated data are generated from the posterior predictive distribution, see (5.3). In practice, for each $\boldsymbol{\pi}_t$ from the posterior, one replicated data set is generated from $M(\boldsymbol{\pi}_t, N)$. Subsequently for each $\boldsymbol{\pi}_t$, two likelihood ratios are calculated, $d(\boldsymbol{\pi}_t, \boldsymbol{f}^{obs})$ and $d(\boldsymbol{\pi}_t, \boldsymbol{f}_t^{rep})$. The most common use of the posterior predictive checks is to plot $d(\boldsymbol{\pi}_t, \boldsymbol{f}^{obs})$ against $d(\boldsymbol{\pi}_t, \boldsymbol{f}_t^{rep})$ for $t = 1, ..., T$. The posterior predictive p-value is given by the proportion of $d(\boldsymbol{\pi}_t, \boldsymbol{f}_t^{rep}) > d(\boldsymbol{\pi}_t, \boldsymbol{f}^{obs})$. The posterior predictive p-value shows whether the model fits the data, and it is well known that this value can not be used to select the best of a set of models (Gelman, Carlin, Stern and Rubin, 1995). All simulations done in this paper confirmed this, and the posterior predictive p-value will not be reported.

To be able to select the best of a set of models, we suggest, in analogy to the L-criterion, to calculate the expectation of the difference between the likelihood ratio of the original data and the likelihood ratio of the replicated data:

$$I\!E\left[d(\boldsymbol{\pi}_t, \boldsymbol{f}_t^{rep}) - d(\boldsymbol{\pi}_t, \boldsymbol{f}^{obs})\right],$$

and choose that model that has the smallest expected discrepancy. In an MCMC run,

the posterior predictive check is readily available. For each sampled $\boldsymbol{\pi}_t$, for $t = 1, ..., T$ from the posterior, new data are generated from a multinomial distribution $M(\boldsymbol{\pi}_t, N)$. The calculation reduces to taking the average of $d(\boldsymbol{\pi}_t, \boldsymbol{f}_t^{rep}) - d(\boldsymbol{\pi}_t, \boldsymbol{f}^{obs})$ for $t = 1, ..., T$.

5.5 Prior Predictive Inference

5.5.1 Prior Predictive Checks

Prior predictive checks were first suggested by Box (1980). The approach is similar to the posterior predictive checks, but instead of generating replicated data from the posterior predictive distribution, replicated data are generated from the prior predictive distribution.

$$P(\boldsymbol{f}^{rep}) = \int P(\boldsymbol{f}^{rep}|\boldsymbol{\pi}) Pr(\boldsymbol{\pi}|\boldsymbol{\alpha}) d\boldsymbol{\pi} \tag{5.6}$$

As a discrepancy measure, the likelihood ratio from (5.5) is used. Replicated data are generated from the prior predictive distribution, see Equation 5.6. The subsequent calculations are the same as in Section 5.4.3.

5.5.2 Prior Predictive L-criterion

In Section 5.4.3, the resemblance of the posterior predictive checks and the L-criterion was pointed out. In Section 5.5.1, it was shown that samples can be taken from the prior predictive distribution instead of the posterior distribution. A simple modification of the L-criterion uses a sample from the prior predictive distribution instead of the posterior predictive distribution.

5.5.3 Marginal Likelihood

In Bayesian statistics, the marginal likelihood has been proposed as a measure of fit by Jeffreys (1961). It has gained much popularity since Newton and Raftery (1994) showed approaches to estimate the marginal likelihood. As with information criteria, the marginal likelihood contains a trade off between the likelihood of the parameters given the data and the number of parameters in the model, but the number of parameters is implicitly accounted for (Newton and Raftery, 1994; Raftery, 1995a,b). This makes the marginal likelihood appropriate to select between inequality constrained models. The marginal likelihood for model l is given by

$$P(\boldsymbol{f}|M_l) = \int P(\boldsymbol{f}|\boldsymbol{\pi}, M_l) Pr(\boldsymbol{\pi}|\boldsymbol{\alpha}, M_l) d\boldsymbol{\pi}, \tag{5.7}$$

the larger the marginal likelihood, the better the model fits. As can be seen in (5.7), the likelihood is integrated with respect to the prior. A general result is that the marginal likelihood is sensitive to the prior, which is an undesirable property if one wants to compare models with respect to their fit. For the models discussed in this paper, we found that the marginal likelihood is not sensitive to the prior, see Laudy and Hoijtink (2006) and Klugkist, Kato, and Hoijtink (2005) for more details.

The Bayes factors is given by the ratio of two marginal likelihoods:

$$BF_{M_1,M_0} = \frac{P(\boldsymbol{f}|M_1)}{P(\boldsymbol{f}|M_0)} \tag{5.8}$$

Klugkist, Kato, and Hoijtink (2005) show that the calculation of Bayes factors is greatly simplified in inequality constrained models. Denote M_0 as the unconstrained model, and M_1 as its constrained alternative, then the Bayes factors is given by:

$$BF_{M_1,M_0} = \frac{c_1}{d_1}, \tag{5.9}$$

where c_1 is the proportions of samples from the unconstrained posterior in agreement with the constrained model, and d_1 is the proportions of samples from the unconstrained prior in agreement with the constrained model. If there are L models, the posterior model probability for model l is given by

$$PM_l = \frac{BF_{M_l,M_0}}{1 + \sum_{w=1}^{L} BF_{M_w,M_0}} \tag{5.10}$$

Note that in this paper all models have equal prior probability, since we are interested in which model is most supported by the data. The resulting posterior model probability gives an easily interpretable scale that shows the support for each model given the data and the model set.

5.6 Simulations

In the following subsections several simulations are performed. The general procedure is as follows: first a population is specified. From each population 500 data sets are generated. Two or more competing models are specified. For each of the models, a sample of both the prior and the posterior distribution is obtained using a Gibbs sampler (Gelfand, Smith and Lee, 1992; Laudy and Hoijtink, 2006), with sample size 11000. The first 1000 samples are considered the burn-in and are removed from further calculations. The model selection criteria are calculated using the sample from either the prior or posterior. Finally, it is evaluated how often a model selection criterion has chosen each of the models. Table 5.1 displays the abbreviations of the model selection criteria that are used in the subsequent tables.

Table 5.1: Abbreviations of the Model Selection Criteria

PM	Posterior model probability
PC_{prior}	Prior predictive checks
$L\text{–}crit_{prior}$	Prior L-criterion
DIC	Deviance Information Criterion
PC_{post}	Posterior predictive checks
$L\text{–}crit_{post}$	Posterior L-criterion

5.6.1 Sample Size

The first simulation investigates the effect of the sample size. The population consists of a three-way contingency table, with variables A,B and C, indexed by i,j and k and $I = J = 2$ and $K = 3$. Odds ratio θ_k is defined as $(\pi_{11k}\pi_{22k})/(\pi_{12k}\pi_{21k})$. Table 5.2 specifies the cell probabilities and odds ratios in the population. From the population in Table 5.2, data sets are generated with six sample sizes, respectively, 40 60, 80, 240, 500 and 1000. For each sample size, 500 data sets are generated.

Table 5.2: Population for the Sample Size Simulation

C		1		2		3	
B		1	2	1	2	1	2
A	1	.05	.05	.15	.10	.10	.05
	2	.05	.05	.10	.15	.05	.10
Odds ratios		1		2.25		4	

The lower half of Table 5.2 shows that the odds ratio in sub table k increases as k increases. Model M_1 displayed in Table 5.3 requires the odds ratios to increase, which correctly describes the population. Model M_2 requires the odds ratios to decrease and model M_3 is the unrestricted model.

Table 5.3: Models for the Population in Simulation 1

Model	$k = 1$		$k = 2$		$k = 3$
M_1	θ_1	$<$	θ_2	$<$	θ_3
M_2	θ_1	$>$	θ_2	$>$	θ_3
M_3	θ_1		θ_2		θ_3

Three simulations are performed. In the first simulation, the correctly constrained model M_1 is compared to the incorrectly constrained model M_2. In the second simulation, the correctly constrained model M_1 is compared to the unconstrained model M_3. In the third simulation, the incorrectly constrained model M_2 is compared to the unrestricted model M_3.

Correctly Constrained Model M_1 versus Incorrectly Constrained Model M_2

The results in Table 5.4 display for each model selection criterion how often model M_1 was chosen over model M_2. With a sample size of $N = 40$, the posterior model probability chose in 80.4% of the samples the correct model. For all the model selection criteria holds that with increasing sample size, the correct model is chosen more often. With a sample size of $N = 1000$, the correct model is always chosen by all model selection criteria. The PC_{prior} appears to be the most powerful in the sense that with small sample sizes, it chooses the correct model more often than the other model selection criteria. A summary of the results can also be found in Table 5.18.

Table 5.4: Results for the Correctly Constrained Model M_1 versus Incorrectly Constrained Model M_2

Sample size	PM	PC_{prior}	$L\text{–}crit_{prior}$	DIC	PC_{post}	$L\text{–}crit_{post}$
40	80.4	83.4	75.4	75.1	78.2	75.2
60	86.0	92.3	77.4	83.4	83.5	80.0
80	86.8	92.0	80.2	86.4	86.0	84.6
240	96.2	98.6	93.0	96.8	96.2	95.2
500	100	100	96.2	100	100	100
1000	100	100	100	100	100	100

The table displays the percentage that model M_1 is chosen over model M_2

Correctly Constrained Model M_1 versus the Unconstrained Model M_3

Table 5.5 displays the results of a comparison of correctly constrained model M_1 and incorrectly constrained model M_3. The prior predictive check appears to be the most powerful. Both the posterior predictive check and the posterior L-criterion do not perform well when the sample size is small. Even with a sample size of $N = 1000$, the posterior L-criterion only chooses in 91% of the samples for the correct model, while all the other model selection criteria always choose for the correct model.

Table 5.5: Results for Correctly Constrained Model M_1 versus Unconstrained Model M_3

Sample size	PM	PC_{prior}	$L\text{–}crit_{prior}$	DIC	PC_{post}	$L\text{–}crit_{post}$
40	74.1	92.2	72.2	75.0	59.4	59.2
60	80.0	94.8	72.8	83.2	63.7	59.6
80	82.4	98.2	78.0	88.4	76.2	69.8
240	93.6	100	88.4	98.2	90.0	80.4
500	96.2	100	95.1	100	96.4	86.2
1000	100	100	100	100	100	91.2

The table displays the percentage that model M_1 is chosen over model M_3

Incorrectly Constrained Model M_2 versus the Unconstrained Model M_3

The incorrectly constrained model M_2 displayed in Table 5.3 is compared to the unconstrained model M_3. The results in Table 5.6 displays that the prior predictive checks performs worst when the sample size is small. Furthermore, even with a sample size of $N = 1000$, the DIC chooses in 10% of the data matrices sampled from the population for the incorrectly constrained model.

Conclusions from the Sample Size Simulation

In the first simulation that compared a correctly constrained model to a incorrectly constrained model, the performance of the model selection criteria were similar, and satisfactory. The prior predictive checks appeared to be slightly more powerful than the other criteria. In the second simulation that compared a correctly constrained model to the unconstrained model, the prior predictive check also performed best, while the

Table 5.6: Results for the Incorrectly Constrained Model M_2 versus Unconstrained Model M_3

Population	PM	PC_{prior}	$L\text{–}crit_{prior}$	DIC	PC_{post}	$L\text{–}crit_{post}$
40	19.2	34.2	23.8	31.2	15.0	16.4
60	14.4	21.4	21.4	28.4	12.2	12.8
80	13.2	19.0	15.0	27.4	11.2	10.9
240	5.8	12.2	6.0	25.0	7.8	5.0
500	0	3.2	3.2	14.2	1.2	0
1000	0	1.0	0	10.8	0	0

The table displays the percentage that model M_2 is chosen over model M_3

posterior predictive checks and the posterior L-criterion performed worst. In the third simulation that compared an incorrectly constrained model to an unconstrained model, the prior predictive checks and the DIC performed worst, while the posterior predictive checks and the posterior L-criterion perform best.

The simultaneous interpretation of simulation two and three may indicate that the prior predictive checks may have a bias against unconstrained models, while both the posterior predictive checks and the posterior L-criterion may show a bias towards unconstrained models.

The posterior model probability may not outperform the other model selection criteria, but it appeared to be most stable across simulations.

5.6.2 Effect Magnitude

In this simulation, the magnitude of the effect is varied, that is, with N held constant, four populations are specified with increasing difference between the frequency counts of the ordered categories. Three simulations are performed. First, the best of two inequality constrained models has to be selected. The first model accurately describes the data, the second model imposes wrong ordering restrictions. The second simulation compares the correctly constrained model to the unrestricted model. The third simulation compares an incorrectly constrained model against an unrestricted model.

The population is a cross-classification of variables A and B, indexed by $i = 1, ..., 4$ and $j = 1, 2$. To interpret the magnitude of the effect, the population is displayed in terms of an ideal sample, that is, the population probabilities multiplied by the sample size. Table 5.7 displays an ideal sample from 4 populations. The magnitude of the effect increases from P_1 to P_4.

Table 5.7: Ideal Samples for the Effect Magnitude Simulation

Population		P_1		P_2		P_3		P_4	
	B	1	2	1	2	1	2	1	2
A	1	20	20	20	20	20	20	20	20
	2	20	20	19	21	18	22	17	23
	3	20	20	18	22	16	24	14	26
	4	20	20	17	23	14	26	11	29

In Table 5.8, the inequality constrained models are displayed. Model M_1 is the correctly constrained model for P_2, P_3 and P_4. The probabilities in the first column of probabilities are restricted to decrease, and the probabilities in the second column are restricted to increase. The incorrectly constrained model for P_2, P_3 and P_4 is model M_2. The restrictions require a reversed ordering in comparison to the first model. Model M_3 is the unrestricted model.

Table 5.8: Models for the Effect Magnitude Simulation

Model		M_1		M_2		M_3	
B		1	2	1	2	1	2
A	1	π_{11}	π_{12}	π_{11}	π_{12}	π_{11}	π_{12}
		$\vee$	$\wedge$	$\wedge$	$\vee$		
	2	π_{21}	π_{22}	π_{21}	π_{22}	π_{21}	π_{22}
		$\vee$	$\wedge$	$\wedge$	$\vee$		
	3	π_{31}	π_{32}	π_{31}	π_{32}	π_{31}	π_{32}
		$\vee$	$\wedge$	$\wedge$	$\vee$		
	4	π_{41}	π_{42}	π_{41}	π_{42}	π_{41}	π_{42}

Correctly Constrained Model M_1 versus incorrectly Constrained Model M_2

The models M_1 and M_2 displayed in Table 5.8 are compared. Model M_1 accurately describes the data for population P_2,P_3 and P_4, while model M_2 reverse the ordering of the probabilities, and is not true for population P_2, P_3 and P_4.

Table 5.9: Results for the Correctly Constrained Model M_1 versus Incorrect Constrained Model M_2

Population	PM	PC_{prior}	$L\text{–}crit_{prior}$	DIC	PC_{post}	$L\text{–}crit_{post}$
P_1	48.0	48.8	49.2	48.6	49.0	49.4
P_2	79.8	78.0	78.0	76.6	78.6	79.4
P_3	89.2	88.2	88.0	86.9	87.7	88.4
P_4	98.8	98.8	98.7	98.8	98.9	98.0

The table displays the percentage that model M_1 is chosen over model M_2

The results are displayed in Table 5.9. For population one, none of the model selection criteria prefers the model M_1 over model M_2. This occurs because the restriction are precisely opposite, and the population is neutral w.r.t. the restrictions. It the magnitude of the effect increases all model selection criteria in equal percentages choose the correct model over the incorrect model.

Correctly Constrained Model M_1 versus Unconstrained Model M_3

The correctly constrained model M_1 displayed in Table 5.8 is compared to the unconstrained model M_3. The results in the are displayed in the upper half of Table 5.10. Note that the populations are displayed in reverse order to enhance the interpretation of the simulation two and three simultaneously. Both the posterior predictive check

and the posterior L-criterion perform worse than the other model selection criteria. For population P_4 in only respectively 56% and 52% of the samples, the correct model is chosen. In population P_1, the posterior model probability chooses in 23% of the samples for the constrained model. For the other model selection criteria, this is much higher.

Table 5.10: Results for the Correctly Constrained Model M_1 versus Unconstrained Model M_3 (Upper Table) and the Incorrectly Constrained Model M_2 versus Unconstrained Model M_3 (Lower Table)

Population	PM	PC_{prior}	$L\text{–}crit_{prior}$	DIC	PC_{post}	$L\text{–}crit_{post}$
P_4	98.2	100	98.2	94.0	57.4	51.6
P_3	69.8	90.0	89.6	89.8	37.2	29.8
P_2	45.0	79.2	75.2	82.7	25.2	20.6
P_1	23.2	52.8	58.0	68.2	13.0	9.4
P_1	20.0	48.4	54.2	65.2	12.4	8.0
P_2	8.4	23.4	24.8	40.4	5.2	5.9
P_3	3.2	12.6	14.6	27.3	2.2	1.2
P_4	0	1.0	2.2	6.8	0	0

The upper table displays the percentage that model M_1 is chosen over model M_3
The lower table displays the percentage that model M_2 is chosen over model M_3

Incorrectly Constrained Model M_2 versus Unconstrained Model M_3

The incorrectly constrained model M_2 displayed in Table 5.8 is compared to the unconstrained model M_3. The results in the are displayed in the lower half of Table 5.10. In population P_4, the DIC performs worst. In population P_1, the DIC also chooses the inequality constrained model very often.

Conclusions from the Effect Magnitude Simulation

The second and third simulation can be interpreted at once. In the upper half of Table 5.10, in the first row, the results for the population with the strongest effect is displayed. In each subsequent row, the magnitude of the effect decreases, and reverses in the lower half of the table. The posterior model probability performs best, since with large effect, the correct model is mostly chosen, and when the magnitude of the effect decreases, it shows the steepest descent. The performance of both the posterior predictive checks and the posterior L-criterion are poor, since with large effect, they choose the unconstrained model too frequently, which may indicate a bias against constrained models. The performance of the prior predictive checks and the prior L-criterion is also poor: in population P_1, with all probabilities in the contingency table being equal, the constrained model imposing the ordering is chosen in half of the samples, which may indicate a bias towards constrained models. The performance of the DIC is unsatisfactory for two aspects. First, in population P_1 it chooses in around 66% of the data sets for the constrained model, and with large effect, the incorrect model is chosen in 7% of the samples.

5.6.3 Models with Different A-priori Likelihood

The third simulation concerns models for the cross-classification of variables A and B. Let $i, j = 1, ..., 3$ index the contingency table. Parameter θ_{ij} denotes the local odds ratio and is defined as $\theta_{ij} = \pi_{ij}\pi_{i+1,j+1}/(\pi_{i+1,j}\pi_{i,j+1})$. The focus in this simulation is to investigate how the model selection criteria perform in the presence of two models that correctly describe the population. It is of interest how the model complexity is taken into account by the model selection criteria. In the second simulation, the population is the same, but the inequality constraints are reversed, so that none of the inequality constrained models correctly describe the population.

Table 5.11 displays the population, and 500 data sets are generated from this population with sample size 225. The odds ratios in the population are displayed in Table 5.12; the odds ratios on the diagonal are larger than 1 and the diagonal odds ratios are ordered from large to small.

Table 5.11: Population for the A-priori Likelihood Simulation

B		1	2	3
A	1	.128	.066	.111
	2	.066	.155	.102
	3	.111	.102	.155

Table 5.12: Odds Ratios θ_{ij} for the Population of the A-priori Likelihood Simulation

B		1-2	2-3
A	l-2	4.51	.39
	2-3	.39	2.3

Two Correct Models

The models displayed in Table 5.13 correctly describe the population. Model M_1 states that the the odds ratios in the diagonal are larger than 1. Model M_2 states that the diagonal odds ratios are in decreasing order. Model M_3 specifies the unconstrained model. In contrast to the previous simulations, the best out of three models is chosen, thus note that in Table 5.14 the results add up too 100% in each column.

Table 5.13: Two Correctly Constrained Models

Model
M_1: $\theta_{ii} > 1$ for $i = 1, 2$
M_2: $\theta_{ii} > \theta_{i+1i+1}$ for $i = 1$
M_3: unconstrained

The results show a major difference between prior predictive and posterior predictive model selection criteria. The prior predictive model selection criteria choose model M_1 as

the best model, while the posterior predictive model selection criteria do not show clear preference of model M_1 over model M_2. It is true that both model M_1 and M_2 correctly describe the population, however, model M_1 restricts the parameter space more than model M_2, stated otherwise, is a smaller model. This can be seen as follows: a sample is taken from the unconstrained prior, and the percentage in agreement with the restrictions is calculated. Table 5.15 displays that model M_1 is a-priori more unlikely than model M_2. The prior predictive model selection criteria correctly incorporate this into the results, favoring model M_1 over model M_2.

Table 5.14: Results for the Correctly Constrained Models

Population	PM	PC_{prior}	$L\text{–}crit_{prior}$	DIC	PC_{post}	$L\text{–}crit_{post}$
M_1	95.6	99.2	99.6	49.4	57.5	38.8
M_2	4.4	.8	.4	50.6	42.5	56.7
M_3	0	0	0	0	.2	4.4

The table displays the percentage that model M_g is chosen as the best model, $g = 1, ..., 3$

Table 5.15: A-priori Likelihood of the Models

	M_1	M_2	M_3
A-priori Likelihood	0.289	0.500	1

Two Incorrectly Constrained Models

In this simulation, the model set consists of three models, in which the inequality constrained models incorrectly describe the population. To make the results comparable to the simulation in the previous subsection, the inequality constraints are reversed, and the population remains the same.

Table 5.16: Two Incorrectly Constrained Models

Model
M_1: $\theta_{ii} < 1$ for $i = 1, 2$
M_2: $\theta_{ii} < \theta_{i+1i+1}$ for $i = 1$
M_3: unconstrained

Model M_1 in Table 5.16 requires the odds ratios to be smaller than 1, and model M_2 requires the odds ratios to be increasing on the diagonal. Table 5.17 displays the results. All the model selection criteria choose the unconstrained model M_3 most frequently as the best model. The prior and posterior predictive L-criterion and the posterior model probability choose the unconstrained model most often. The DIC seems to be the worst performing model selection criterion, since in 36% of the samples, the incorrect model M_2 is chosen.

Table 5.17: Results the Two Incorrectly Constrained Models

Population	PM	PC_{prior}	$L\text{–}crit_{prior}$	DIC	PC_{post}	$L\text{–}crit_{post}$
M_1	0	0	0	0	0	0
M_2	11	21.0	8.6	36.7	17.8	9.6
M_3	89	79.0	91.4	63.3	82.2	90.4

The table displays the percentage that model M_g is chosen as the best model, $g = 1, ..., 3$

Conclusions for the A-priori Likelihood Simulation

The results using the first model set show that the prior predictive model selection criteria use different penalties for model complexity and the results suggest that prior predictive model selection criteria are more sensitive to the a-priori likelihood of the models. In the second model set, there was no clear distinction between prior and posterior predictive model selection criteria. In the second model set, the DIC performed worst, while both the prior and posterior L-criteria performed best.

5.7 Discussion

Three simulations were performed to investigate the difference between various model selection criteria. In Table 5.18, an overview is displayed of the results. A '++' is used to indicate which criterion performed best in a simulation. If two criteria perform very well, both are granted a '++'. A '+' is used to indicate that the performance was satisfactory, a '0' is used to indicate that the performance was neither good nor poor, and a '-' is used to indicate poor performance. The total is calculated by counting a '++' as 2, '+' as 1, '0' as a 0, and '-' as -1. look to 5.6.1.

Table 5.18: Overview of the all Simulation Results

Section	5.6.1	5.6.1&5.6.1	5.6.2	5.6.2&5.6.2	5.6.3	5.6.3	Total
PM	+	++	+	++	+	+	8
PC_{prior}	++	0	+	+	++	+	7
$L\text{–}crit_{prior}$	+	+	+	+	++	++	8
DIC	+	-	+	-	-	-	-2
PC_{post}	+	0	+	-	-	-	-1
$L\text{–}crit_{post}$	+	0	+	-	-	++	2

++: best +: satisfactory 0: indecisive -: poor

The posterior predictive check and the posterior L-criterion showed similar performance. In the sample size simulation, with small sample size in the comparison of the correctly constrained model to the unconstrained model, the criteria were not able to choose the correctly constrained model with high frequency. When the incorrectly constrained model was compared to the unconstrained model, they performed very well, even with small sample size. Similar results were obtained in the simulation where magnitude of the effect was varied. These results may indicate a bias against constrained models.

The reverse may be true for the both the prior predictive checks and the prior L-criterion. In the sample size simulation, even with small sample size, the correctly constrained model was chosen with high frequency. However, when the incorrectly constrained model was compared to the unconstrained model, the performance of both the prior predictive checks and the prior L-criterion was worse than the other model selection criteria. Similar results were obtained for the effect magnitude simulation. Hence, these results may indicate a bias against unconstrained models.

Those results may also be explained in the light of how the criteria are composed of a fit and a penalty for model complexity. Prior predictive checks and the prior L-criterion give large weight to model complexity: even if the fit of a constrained model is not very well, it is reward for being a small model. Posterior predictive check and the posterior L-criterion give large importance to the fit: if a very small constrained model does not show a very good fit, the unconstrained model is preferred. These results are in agreement to the simulation of the models with different a-priori likelihood, where the results also showed that the prior predictive model selection criteria are more sensitive to the a-priori likelihood of the models.

Overall, the prior predictive model selection criteria perform better in the context of inequality constrained contingency tables. The posterior model probability and the prior L-criterion seems to perform best: the results are stable over simulations, and although not the most powerful in all situations, their performance is satisfactory.

References

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Automat. Control AC* **19**: 716–723

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

Barlow, R.E., Bartholomew, D.J., Bremner, J.M. and Brunk, H.D. *Statistical Inference Under Order Restrictions* Wiley, New York, 1972

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

Berger J.O. and Sellke T. (1987) Testing a point null hypothesis: The irreconcilability of p-values and evidence (with discussion). *Journal of the American Statistical Association* **82**: 112–139

Box G.E.P. (1980) Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistics Society, Series A* **143(4)**: 383–430

Cohen, J. (1994) The earth is round ($p < .05$). *American Psychologist* **49**: 997–1003

Gelfand A.E. and Ghosh, S.K. (1998) Model choice: A minimum posterior predictive loss approach. *Biometrika* **85(1)**: 1–11

Gelfand A.E. and Smith A.F.M. (1990) Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* 1990; **85**: 398–409

Gelfand A.E., Smith A.F.M. and Lee T.M. (1992) Bayesian analysis of constrained parameter and trucated data problems using Gibbs sampling. *Journal of the American Statistical Association* 1992; **87**: 523–532

Gelman, A. Carlin B.P., Stern H.S. and Rubin D.B. *Bayesian Data Analysis*. Chapman and Hall: New York, 1995.

Gelman, A., Meng, X.L. and Stern, H.S. (1996) Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica* **6**: 733–807

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

Kass, R.E. (1993) Bayes factors in practice. *The Statistician* **42**: 551–560

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

Klugkist, I., Kato, B. and Hoijtink, H. (2005) Bayesian model selection using encompassing priors *Statistica Neerlandica* **59(1)**: 57–69

Laud, P.W. and Ibrahim, J.G. (1995) Predictive model selection. *Journal of the Royal Statistics Society, Series B* **57**: 247–262

Laudy, O. and Hoijtink, H. (2006) Bayesian methods for the analysis of inequality constrained contingency tables. *Statistical methods in medical research* to appear in 2006.

Van der Linde, A. (2005) DIC in variable selection. *Statistica Neerlandica* **59(1)**: 45-56

Meng, X.L. (1994) Posterior predictive p-values. *The Annals of Statistics* **22**: 1142–1160

Newton M.A. and Raftery A.E. (1994) Approximate Bayesian inference by the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B* **3**: 3–48

Raftery, A.E. (1986) Choosing models for cross-classifications. *American Sociological Review* **51(1)**: 145-146

Raftery, A.E. (1995a) Bayesian Model Selection in Social Research. *Sociological Methodology* **25**: 111–164

Raftery, A.E. (1995b) Rejoinder: Model Selection Is Unavoidable in Social Research. *Sociological Methodology* **25**: 185–196

Rubin, D. B. (1984) Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *Annals of Statistics* **12**: 1151–1172

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics* **6**: 461–464

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

Shafer, J. *Analysis of Incomplete Multivariate Data* Chapman and Hall: New York, 1997

Spiegelhalter, D.J., Best, N.G., Carlin B.P. and Van der Linde A. (2002) Bayesian Measures of Model Complexity and Fit (with Discussion). *Journal of the Royal Statistical Society, Series B* **64(4)**: 583–639

Chapter 6

Bayesian Evaluation of Equality and Inequality Constrained Hypotheses for Contingency Tables*

Abstract

In this paper an approach is introduced to select the best of a set of equality and inequality constrained hypotheses for contingency tables. Currently available methods require extensive knowledge about various parameterizations of log-linear models and parameter estimates may be difficult to interpret. Moreover, inequality constrained hypotheses testing is virtually absent in many software packages. In the approach discussed in this paper, models for contingency tables are presented in terms of cell probabilities rather than log-linear parameters. This allows researchers to test equality and inequality constrained hypotheses in a format that is directly related to the data, and in a way that researchers are used to think of hypotheses. User friendly software is provided that allows researchers to apply the method in this paper to their own data. Contingency tables can be modelled using a wide range of models including ordinary log-linear models, ordinal logit models and association models using ordinal odds ratios. We provide examples of the above mentioned models using recent published literature from psychological journals.

*This chapter has been submitted for publication as Laudy, O., Klugkist, I, and Hoijtink, H. Bayesian Evaluation of Equality and Inequality Constrained Hypotheses for Contingency Tables *Psychological Methods*

6.1 Introduction

Advances in Bayesian methodology in recent years have resulted in a great expansion of applications of Bayesian statistics in a wide variety of fields. Bayesian papers now make up a substantial percentage of the papers published in the top statistical journals. However, to understand and apply Bayesian statistical models, a fair amount of technical knowledge is required. In this paper, we discuss Bayesian model selection in the context of contingency tables at a non-technical level. The hypotheses are presented in terms of cell probabilities rather than log-linear parameters. This allows researchers to test equality and inequality constrained hypotheses in a format that is directly related to the data, and in a way that researchers are used to think about expected outcomes or theories about their data.

Consider the following example, where the relation between internal assets and being sent from class is investigated (Nash and Bowen, 2002). Data are collected using the School Success Profile (SSP) by Bowen, Richman and Brewster (1998). The SSP is a self-administered instrument designed for students in grades 6 to 12. The internal assets index is a dichotomized composite of 10 items that assesses the adolescent's perception of his or her strength and resources (health, exercises, or involvement in sports). A dichotomous measure was created using a students response to a single SSP item asking whether, during the previous 30 days, the student had been sent from class due to his or her behavior. The data are presented in Table 6.1.

Table 6.1: Cross-classification of Internal Assets and Being Sent from Class

		Sent from class	
		yes	no
Internal assets	low	220	1060
	high	96	609

The (classical) null hypothesis testing approach defines the possible relation between internal assets and being sent from class in terms of the odds ratio (denoted by θ), formally defined as

$$\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}},$$

where π_{ij} denote the cell probabilities in the contingency table, $i = 1, ..., I$ denote the categories of the row variable, $j = 1, ..., J$ denote the categories of the column variable. In this example $I = J = 2$.

A standard (classical) test evaluates the fit of the data under the null hypothesis $H_0 : \theta = 1$, representing the independence model, that is: 'no association between the two variables'. The resulting p-value is the probability that the observed data, or a more extreme outcome, would have occurred under the null-hypothesis. Loosely speaking, a p-value is a measure of how much evidence we have against the null hypothesis. In the preceding test we used the alternative hypothesis stating "not H_0", that is, the unconstrained model, without restrictions on θ, formally represented by $H_1 : \theta \neq 1$. It is however also possible to test against a more informative alternative. In this example, it is for instance rather straightforward to test against the alternative $H_1 : \theta > 1$, representing

the expectation that there will be a positive association between the two variables.

Directional testing within the classical framework becomes less straightforward when larger contingency tables and different informative alternative hypotheses are considered. For an extensive overview of classical methods for inequality constrained hypotheses see Barlow, Bartholomew, Bremner and Brunk (1972); Robertson, Wright and Dykstra (1988). Consider for example a $2 \times 2 \times 2$ cross tabulation of internal assets (low, high), being sent from class and the additional variable gender. The cell probabilities are denoted by π_{ijk}, where $i = 1, 2$ refers to respectively females and males, $j = 1, 2$ refers to low and high internal assets, and $k = 1, 2$ to being sent form class or not. The hypotheses below are formulated in terms of odds ratios, with θ_1 denoting the relation between internal assets and being sent form class for females, and θ_2 the association for males. Several hypotheses can be formulated. For example: the relation between internal assets and being sent from class is one for females and males ($H_1 : \theta_1 = \theta_2 = 1$), or for females and males the relation is equal ($H_2 : \theta_1 = \theta_2$). Interesting hypotheses using inequality constraints are for example: for both males and females the relation between internal assets and being sent from class is positive ($H_3 : \theta_1 > 1, \theta_2 > 1$), or, this relation is stronger for females than for males ($H_4 : \theta_1 > \theta_2$), or, a combination of H_3 and H_4: both relations are positive and the relation is stronger for females than for males ($H_5 : \theta_1 > \theta_2 > 1$). This example with hypotheses $H_1 - H_5$ (and more) is further discussed in Section 6.3.2. Comparing several alternative hypotheses using p-values is not straightforward. Although each of the alternatives can be tested against the null ($H_0 : \theta_1 = \theta_2 = 1$), this does not provide information about the *relative fit* of the alternatives. On the basis of multiple p-values, it can not be decided which alternative hypothesis is the best. In this paper, a Bayesian hypothesis selection approach is presented that can incorporate equality and inequality constraints on cell probabilities (as well as on odds ratios) in the context of (complex) contingency tables. In this approach, posterior model probabilities for each of the hypotheses under consideration are obtained (see Section 2). There are several advantages of posterior probabilities over the p-value. Firstly, the use of posterior probabilities is not limited to two hypotheses (null and alternative) but can be used to select the best of a set of hypotheses. Secondly, the interpretation of a posterior probability is straightforward: it is the probability assigned to the hypothesis given the data and the set of hypotheses under consideration, taking the fit and complexity of the hypothesis into account. Thirdly, the method is flexible in the sense that models and constraints can be formulated in terms of cell probabilities or functions of cell probabilities, like odds ratios. Finally, the method is general in the sense that it can also easily be applied to complex (for instance three-way or four-way) contingency tables, with simple and complex sets of inequality constraints.

However, the ease of the posterior model probability comes with a price: a prior distribution has to be specified for the model parameters. This will be elaborated in the next section. For equality constrained hypotheses, the results of the model selection depend on the prior specification, however, as will be shown, inequality constrained hypotheses are not sensitive to the prior specifications. In Section 6.5, we will discuss the prior sensitivity of the model selection in the context of equality and inequality constrained contingency tables. The examples consist of the re-analysis of recently published data to show how the posterior model probability can be used to select the

best of a set of competing hypotheses.

6.2 Bayesian Model Selection

6.2.1 Prior and Posterior Distributions

In Bayesian statistics the parameters of a model follow a random distribution. The distribution assumed for the parameters before observing any data is called the prior distribution. Before the analysis, this prior distribution has to be specified. After data have been observed, the prior distribution is 'updated' by multiplying the prior with the likelihood of the data. This leads to the posterior distribution of the parameters. Thus, the prior distribution specifies the knowledge that is available before the data are observed. The posterior distribution specifies what is known after observing the data. For a general comprehensive introduction in Bayesian statistics, we refer to Gelman, Carlin, Stern and Rubin (1995). For contingency tables the three ingredients: the likelihood, the prior and the posterior are introduced.

We use the standard notation for contingency tables with three variables A, B, and C with respective indexes $i = 1, ..., I$, $j = 1, ..., J$, and $k = 1, ..., K$. Generalizations to tables with more than three dimensions are straightforward. Let f_{ijk} denote the observed frequency for cell (i, j, k). Let π_{ijk} denote the cell probabilities in the contingency table. Let $\boldsymbol{f} = \{f_{ijk} | i = 1, ..., I, j = 1, ..., J, k = 1, ..., K\}$ and $\boldsymbol{\pi} = \{\pi_{ijk} | i = 1, ..., I, j = 1, ..., J, k = 1, ..., K\}$. We assume that $\boldsymbol{f}$ has a multinomial distribution, $\boldsymbol{f}|\boldsymbol{\pi} \sim M(\boldsymbol{\pi}, N) \propto \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} \pi_{ijk}^{f_{ijk}}$.

Let $\boldsymbol{\alpha} = \{\alpha_{ijk} | i = 1, ..., I, j = 1, ..., J, k = 1, ..., K\}$ denote the parameters of the conjugate prior distribution, $Pr(\boldsymbol{\pi}|\boldsymbol{\alpha}) \sim Dirichlet(\boldsymbol{\alpha}) \propto \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{K} \pi_{ijk}^{\alpha_{ijk}-1}$.

The parameters of the prior distribution $\boldsymbol{\alpha}$ can be viewed as the number of persons that are in the contingency table before observing the data. Before the analysis, a choice has to be made with respect to $\boldsymbol{\alpha}$. A common choice is a constant for each element of the vector $\boldsymbol{\alpha}$. The choice of this constant will be discussed in Section 6.5, as it will turn out that the posterior model probability is sensitive to the choice of $\boldsymbol{\alpha}$. The encompassing prior approach is used, that is, one prior is specified for the unconstrained hypothesis, and the priors for the constrained hypotheses follow from the procedure explained below.

The inequality constrained hypothesis is translated into inequality constraints on the cell probabilities. Denote inequality constraint z as $r_z(\boldsymbol{\pi})$ for $z = 1, ..., Z$. The joint constraints are $R(\boldsymbol{\pi}) = (r_1(\boldsymbol{\pi}), ..., r_Z(\boldsymbol{\pi}))$. The inequality constraints are accounted for in the prior distribution as follows

$$Pr(\boldsymbol{\pi}|R(\boldsymbol{\pi}), \boldsymbol{\alpha}) = \frac{Pr(\boldsymbol{\pi}|\boldsymbol{\alpha}) I_{\boldsymbol{\pi} \epsilon R(\boldsymbol{\pi})}}{\int Pr(\boldsymbol{\pi}|\boldsymbol{\alpha}) I_{\boldsymbol{\pi} \epsilon R(\boldsymbol{\pi})} d\boldsymbol{\pi}.}$$

For example, consider the restriction $\pi_1 > \pi_2$ and let $\boldsymbol{\alpha}$ equal 1. The inequality constrained prior equals zero for all combinations of π_1 and π_2 where $\pi_1 < \pi_2$ and a constant for all combinations of π_1 and π_2 where $\pi_1 > \pi_2$. This constant is determined such that the prior integrates to one, e.g., the restriction $\pi_1 > \pi_2$ constrains half the parameter space, hence the constant equals 2 for all combinations of π_1 and π_2 where $\pi_1 > \pi_2$

Denote $I_{\boldsymbol{\pi}\in R(\boldsymbol{\pi})}$ as the indicator function over the inequality constraints, where $I_{\boldsymbol{\pi}\in R(\boldsymbol{\pi})} = 1$ if $\boldsymbol{\pi}$ is in accordance with $R(\boldsymbol{\pi})$, otherwise it is zero. The posterior for the contingency table is

$$P(\boldsymbol{\pi}|\boldsymbol{f}) \propto \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{k=1}^{K} \pi_{ijk}^{f_{ijk}+\alpha_{ijk}-1} I_{\boldsymbol{\pi}\in R(\boldsymbol{\pi})},$$

restricted such that $\sum_{ijk} \pi_{ijk} = 1$.

The focus of this paper is on selection of the best of a set of hypotheses, hence, parameter estimates and confidence intervals will not be presented, however, they are easily obtained from a sample from the posterior distribution. For details see (Laudy and Hoijtink, to appear)

For hypotheses selection a sample from both the prior distribution and the posterior distribution is needed. This will be elaborated upon in the next section.

6.2.2 Bayes Factors and Posterior Model Probabilities

The interesting question for a researcher with competing theories or hypotheses is: What are the probabilities of each of these hypotheses given the observed data? These probabilities are the posterior model probabilities. Three ingredients are required to compute posterior model probabilities: a finite set of models or hypotheses, a prior model probability for each hypothesis in the set and marginal likelihoods or Bayes factors. The first require careful considerations of the researcher: which theories or expectations should be included? Should the unconstrained model be part of the set of hypotheses? With respect to the second ingredient, throughout this paper (the conventional) equal prior model probabilities are used. So, for a set of T models the prior probability for model t ($t = 1, \ldots, T$) equals $1/T$. The last ingredient, the marginal likelihood can be interpreted as the likelihood that the data are observed given that the hypothesis at hand is true. The fit of two hypotheses or models can be compared by examining the ratio of the marginal likelihoods, the Bayes factor (Kass, 1993; Kass and Raftery, 1995). A great deal of literature shows that the computation of marginal likelihoods can be burdensome (Chib, 1995; Gelfand and Smith, 1990; Verdinelli and Wasserman, 1995). However, the hypotheses considered in this paper are all constrained versions of the unconstrained model. Stated otherwise, the (in)equality constrained hypotheses are nested in the unconstrained model. Klugkist, Kato, and Hoijtink (2005) show that the calculation of a Bayes factor for nested hypotheses (that is, the Bayes factor for any constrained model with respect to the unconstrained model) is greatly simplified and does not require the computation of marginal likelihoods. Estimation of Bayes factors for nested contingency tables is discussed separately for inequality constrained hypotheses (Section 6.2.2) and equality constrained hypotheses (Appedix 6.8).

Inequality Constrained Hypotheses

Klugkist, Kato, and Hoijtink (2005) show that the calculations of Bayes factors is greatly simplified in the context of inequality constrained hypotheses. Denote the unconstrained hypothesis as $H_0 : \boldsymbol{\pi}$ and the t-th constrained hypothesis as $H_t : r_t(\boldsymbol{\pi})$, where r_t is

a function that imposes restrictions on $\boldsymbol{\pi}$. Note that in the sequel, the unconstrained hypothesis is always denoted by H_0. The Bayes factor BF_{t0} can be written as:

$$BF_{t0} = c_t/d_t, \tag{6.1}$$

where c_t denotes the proportion of the unconstrained posterior that is in accordance with the constrained model, and d_t denotes the proportion of the unconstrained prior that is in accordance to the constrained prior. A sample estimate of c_t is obtained by sampling from the unconstrained posterior distribution and calculating the proportion of parameter vectors that is in agreement with hypothesis H_t. A sample estimate of d_t is obtained by sampling from the unconstrained prior distribution and calculating the proportion of samples that is in agreement with the hypothesis H_t. Any software package that allows sampling from a distribution can be used to obtain a sample from both the unconstrained prior and posterior distribution.

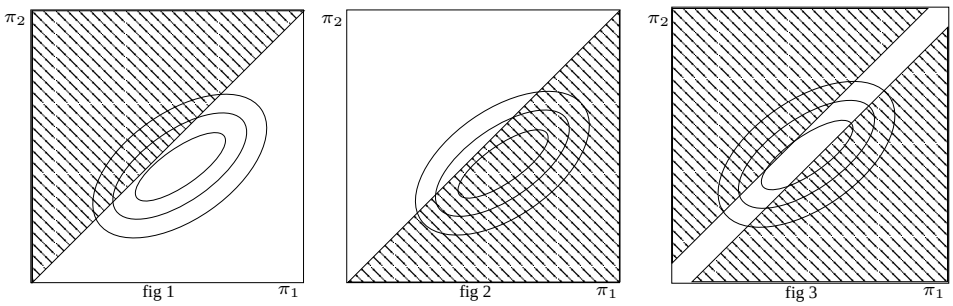


Figure 6.1: Prior and posterior for three models

As an illustration, consider the left panel in Figure 6.1. There are two parameters π_1 and π_2. The unconstrained hypothesis is given by $H_0 : \pi_1, \pi_2$ and the constrained hypothesis is $H_1 : \pi_1 > \pi_2$. The non-shaded area indicates the prior for the constrained hypothesis; the ellipses indicate the posterior distribution of the unconstrained hypothesis. Note that the posterior distribution of the constrained hypothesis is indicated by the ellipses in the non-shaded area. Sampling from the unconstrained prior and calculating the proportion that is in agreement to the constrained prior leads to an estimate of d_1, that equals .5. Note that in this case, it is not necessary to sample, since it can easily be seen that the constrained area is half the area of the unconstrained hypothesis, however for more complex for that is not the case. A sample from the unconstrained posterior that is in agreement with the constrained hypothesis leads to an estimate of c_1, which will be larger than .5, as the posterior concentrates in the area where $H_1 : \pi_1 > \pi_2$ is true. Consequently, the BF_{10} will be larger than one, indicating that the constrained hypothesis is more likely than the unconstrained hypothesis.

Consider the middle penal in Figure 6.1. The constrained hypothesis is given by $H_1 : \pi_1 < \pi_2$. The quantity d_1 is still .5, while the proportion of samples from the unconstrained posterior that is in agreement with the constrained hypothesis will be smaller than .5. Consequently, the BF_{10} will be smaller than one, indicating that the unconstrained hypothesis is more likely than the constrained hypothesis. So far, we state that the quantity c_1 can be seen as the fit of the model. The next illustration will show

how the quantity d_1 acts as the penalty for model complexity.

Consider the right penal in Figure 6.1. The constrained hypothesis is given by $H_1 : |\pi_1 - \pi_2| < .1$. A sample from the unconstrained prior that is in agreement with the constrained hypothesis will be .19 (which is the area of the non-shaded part, and can also be calculated by hand). The proportion of samples from the unconstrained posterior that is in agreement with the constrained hypothesis will be (relatively) large, as the unconstrained posterior concentrates in the non-shaded area. Consequently, the BF_{10} will also be large, indicating that the constrained hypothesis is more likely than the unconstrained hypothesis. We argue that the quantity d_1 acts a penalty for model complexity. The constrained hypothesis $H_1 : |\pi_1 - \pi_2| < .1$ parsimoniously describes the data, i.e., imposes strict restrictions on the parameters. Therefore, if under these restrictions the data fits well, i.e, the quantity c_1 is reasonably large, the BF_{10} will indicate that the constrained hypothesis is more likely than the unconstrained hypothesis.

Summarizing: the proportion of samples in agreement with the constraints after observing the data (d_1) is compared to (or stated differently, penalized by) the proportion of samples in agreement with the constraints a priori (c_1). This approach works for inequality constrained hypotheses, however, not for equality constrained hypotheses. In the appendix, this method is adjusted such that also the Bayes factors for equality constrained hypotheses can be calculated.

Posterior Model Probabilities

A Bayes factor provides the posterior odds of two hypotheses. For a finite set of hypotheses, representing a set of competing theories or expectations, posterior model probabilities for all models in the set can be computed from the Bayes factors.

Consider a set of just the unconstrained model (H_0) and one alternative hypothesis (H_1). The posterior model probability $P_{H_1|H_0,H_1}$, that is, the probability that hypothesis H_1 is true given the set of hypotheses H_0, H_1, is given by:

$$P_{H_1|H_0,H_1} = \frac{BF_{10}}{1 + BF_{10}}.$$

The probability of H_0 is $1 - P_{H_1|H_0,H_1}$. Note that a-priori both hypotheses are equally likely. When $P_{H_1|H_0,H_1} = .9$, there is much evidence that the hypothesis is true. However, when $P_{H_1|H_0,H_1} = .6$ the evidence for hypothesis H_1 is much weaker. For a constrained hypothesis against the unconstrained hypothesis, Table 6.2 shows the strength of evidence (Kass and Raftery, 1995).

Table 6.2: Interpretation of the Posterior Model Probability

.5 - .75	Not worth than a bare mentioning
.75 - .9	Substantial
.9 -.99	Strong
>.99	Decisive

If the set of hypotheses of interest consists of the unconstrained model H_0 and several alternative hypotheses H_t ($t = 1, \ldots, T$), the posterior probabilities are computed

using all Bayes factors of the constrained models with the unconstrained model (BF_{t0}), applying:

$$P_{H_t|H_0,\ldots,H_T} = \frac{BF_{t0}}{1 + BF_{10} + \ldots + BF_{T0}}$$

To obtain the posterior probability of the unconstrained model ($P_{H_0|H_0,\ldots,H_T}$) the numerator in the previous equation is replaced by the value 1.

Note that it is not always interesting to incorporate the unconstrained hypothesis into the set of hypotheses. Incorporating the unconstrained hypothesis H_0 shows whether the constrained hypotheses in the set provide a good description of the data: if the posterior probability of H_0 is large, none of the constrained hypotheses is supported by the data. However, it may be of interest to choose between the best of two restricted hypotheses. In that case the unrestricted model (H_0) is not included in the set of hypotheses. Note that it is still part of the analyses because the unconstrained model is used to compute all the Bayes factors BF_{t0}. The posterior model probabilities of models H_t ($t = 1, \ldots, T$), exclusive the unconstrained model, are computed using

$$P_{H_t|H_1,\ldots,H_T} = \frac{BF_{t0}}{BF_{10} + \ldots + BF_{T0}}$$

6.3 Examples

In this section, the application of the posterior model probability is illustrated by analyzing simple examples.The data come from recently published literature, hence it is shown that even for simple examples, one is able to select between various inequality constrained hypotheses. For each example, several simulation studies are conducted to illustrate the behavior of the posterior model probability. In the analyses, the value of $\boldsymbol{\alpha}$ is set to one, unless otherwise specified. In Section 6.5, it is argued why an $\boldsymbol{\alpha}$ of one leads to good results.

6.3.1 One Odds Ratio

The results of the first example (Nash and Bowen, 2002) of Section 6.1 are discussed. The data are displayed in Table 6.1, and the research question is whether scholars with high internal assets are less frequently sent from class than scholars with low internal assets. Table 6.3 displays the hypotheses in terms of odds ratios. The odds ratio is a measure of association between being sent from class and internal assets, with a value larger than one supporting a positive association.

Table 6.3: Hypotheses for the Data in Table 6.1

H_0:	θ
H_1:	$\theta = 1$
H_2:	$\theta > 1$

The observed odds ratio of 1.31 ($(220 * 609)/(96 * 1060)$) shows that scholars with low internal assets are being sent from class more frequently than scholars with high internal

assets, however, it is to be investigated whether this positive odds ratio is due to sample fluctuation or not.

Table 6.4 shows the posterior model probabilities for each of the hypotheses. Note that the sum of the posterior model probabilities from the models that are taken into account always sums to 1. For example, in the first comparison, $H_1 : \theta = 1$ is compared to the unconstrained hypothesis $H_0 : \theta$, denoted by $P(H_1|H_0, H_1)$. For illustrative purposes, the two posterior model probabilities are displayed sum to one, thus also displaying $P(H_0|H_0, H_1)$. In the remainder of this paper, the latter will not be shown. Table 6.4 shows a posterior model probability of .697 for the the hypothesis $\theta = 1$ against .303 for the unconstrained model. According to Table 6.2, this is not worth more than a bare mentioning. The next hypothesis investigates whether the odds ratio is larger than one. This also results in a posterior model probability not worth more than a bare mentioning. Hypotheses H_1 and H_2 from Table 6.3 can be combined in a single Bayes factors. This is just an other way of representing the results, and this can be done to enhance the interpretation. The results show that there is a slight preference for the hypothesis that the odds ratio equals 1. Thus, there appears no relation between internal assets and being sent from class, however, the evidence is not strong.

Table 6.4: Posterior Model Probabilities for Hypotheses in Table 6.3

Comparison	Post. prob.
$P(H_1\|H_0, H_1)$:	.697
$P(H_0\|H_0, H_1)$:	.338
$P(H_2\|H_0, H_2)$:	.662
$P(H_0\|H_0, H_2)$:	.303
$P(H_1\|H_1, H_2)$:	.540
$P(H_2\|H_1, H_2)$:	.460
$P(H_0\|H_0, H_1, H_2)$:	.190
$P(H_1\|H_0, H_1, H_2)$:	.437
$P(H_2\|H_0, H_1, H_2)$:	.372

6.3.2 Two Odds Ratios

The second research question of Nash and Bowen (2002) is whether the relation between internal assets and being sent form class differs between gender. Table 6.5 displays the data. Note that summing over gender gives Table 6.1. The observed odds ratios are 2.25 and 1.19 for females and males respectively.

In the previous example, the hypothesis most supported by the data, however not with concluding evidence, was that the odds ratio equalled one. For this example, several hypotheses can be investigated and are presented in Table 6.6. In Table 6.7 the posterior model probabilities for various hypotheses are displayed. In the first hypothesis it is investigated whether both odds ratios can be assumed to result from a population where there is no association between internals assets and being sent from class for both males and females. The posterior model probability (.280) indicates that this hypothesis is not much supported by the data. The next hypothesis is whether the odds ratio is equal

Table 6.5: Cross-classification of Internals Assets and Being Sent from Class by Gender

		Sent from class	
		yes	no
Sex	Internal assets		
Female	low	79	629
	high	18	323
Male	low	141	431
	high	78	286

for males and females (H_2). The posterior model probability (.194) shows that this hypothesis is not supported by the data.

A further hypothesis is whether the odds ratios for both males and females are larger than one (H_3). The posterior model probability (.767) shows substantial support for this hypothesis, making it likely that both males and females with low internal assets are more often being sent from class in comparison to males and females with high internal assets. A further hypothesis is whether the odds ratio for females is larger than for males(H_4). The posterior model probability (.665) shows moderate support for this hypothesis. A more specific hypothesis is whether the odds ratios for females is larger than for males, and both odds ratios are larger than one (H_5). The posterior model probability (.872) shows substantive support for this hypothesis.

One may be interested in contrasting two of the above hypotheses. The posterior model probability $P(H_3|H_1, H_3)$ shows substantive support for the hypothesis that both odds ratios are larger than one, compared to both odds ratios being equal to one. The posterior model probability $P(H_5|H_1, H_5)$ shows substantive support for the hypothesis that the odds ratio for females is larger than for males, and both odds ratios are larger than one, compared to the hypothesis that both odds ratios equal one, thus concluding that there is a substantive evidence that students with high internals assets are less being sent from class, and females with high internal assents are less being sent from class than males with high internal assets.

Table 6.6: Hypotheses for the Data in Table 6.5

H_0:	θ_1,θ_2
H_1:	$\theta_1 = \theta_2 = 1$
H_2:	$\theta_1 = \theta_2$
H_3:	$\theta_1 > 1$,$\theta_2 > 1$
H_4:	$\theta_1 > \theta_2$
H_5:	$\theta_1 > \theta_2 > 1$

6.4 Behavior of the Posterior Model Probability

In this section the effect of sample size and effect size on the posterior model probability is explored. For various sample sizes and various odds ratios, the posterior model

Table 6.7: Posterior Model Probabilities for Hypotheses in Table 6.6

Comparison	Post. prob.
$P(H_1\|H_0, H_1)$:	.280
$P(H_2\|H_0, H_2)$:	.194
$P(H_3\|H_0, H_3)$:	.767
$P(H_4\|H_0, H_4)$:	.665
$P(H_5\|H_0, H_5)$:	.872
$P(H_3\|H_1, H_3)$:	.898
$P(H_5\|H_1, H_5)$:	.933

probability is computed using the data format of the examples from Sections 6.3.1 and 6.3.2.

6.4.1 One Odds Ratio

To explore the behavior of the posterior model probability for various sample sizes and odds ratios, new data for the 2 by 2 contingency table like Table 6.1 are constructed. The procedure to create data are set up as follows. For a chosen sample size, say 20, the first data set is constructed to have an exact odds ratio of 1, that is, each cell contains a frequency of 5. To increase the odds ratio in the next data set, the diagonal cells are increased with 1 and the off-diagonal cells are decreased with 1. In the next data set, the diagonal cells are again increased with 1, and the off-diagonal cell are decreased with 1, etc. The resulting sequence of tables is displayed in Table 6.8. Following a similar procedure, a sequence of odds ratios smaller than 1 can be constructed via a stepwise subtraction of 1 from the diagonal cells and addition of 1 to the off-diagonal cells. Note that this sequence is not presented in Table 6.8. This procedure is followed for sample size 20, 40, 100, 200, 400 and 2000. For the larger sample sizes, the (off) diagonal steps are increased as indicated between the brackets 20(2), 40(5), 100(10), 200(25), 400(50).

Table 6.8: Constructed Data N=20 for a 2x2 Contingency Table

		B		B		B			B	
		1	2	1	2	1	2		1	2
A	1	5	5	6	4	7	3	$\cdots$	9	1
	2	5	5	4	6	3	7	$\cdots$	1	9
Odds ratio		1		2.25		5.44		$\cdots$	16	

For each constructed data set, the posterior model probability of hypotheses $H_1 : \theta = 1$ and $H_0 : \theta$ is evaluated. The results are presented visually in Figure 6.2. Each line represents a different sample size. The x-axis displays the odds ratio. The figure shows that when the odds ratio increases, the posterior model probability for the hypothesis $H_1 : \theta = 1$ decreases. For a given odds ratio larger than 1, the posterior model probability for $H_1 : \theta = 1$ for a data set with a large sample size is lower than for a data set with smaller sample size. Note that if the odds ratio exactly equals 1, the posterior model probability is not automatically 1, however, the posterior model probability increases to

1, if the sample size is large enough. This is appropriate behavior, since a small data set could never show great persuasiveness.

In Figure 6.3, the posterior model probability for the hypotheses $H_1 : \theta > 1$ versus $H_0 : \theta$ is displayed. Note that the maximum posterior model probability is 2/3. This can be understood in two ways. First, a hypothesis is associated with a parsimony. If a hypothesis describes the data parsimoniously, and the data are in agreement to the hypothesis, it is rewarded for its parsimony in the form of a high posterior model probability. The hypothesis that $H_1 : \theta > 1$ is not very parsimonious. The posterior model probability incorporated this, and gives a maximum of 2/3. However, when $H_1 : \theta > 1$ is not true, the posterior model probability goes to 0. Second, a maximum of 2/3 can also be understood theoretically: when the observed odds ratio is much larger than 1, a sample from the unconstrained posterior distribution of the model will show that most of the sampled values from the posterior distribution are in agreement to the hypothesis H_1. A sample from the prior distribution will show that half of the samples is in agreement with the hypothesis, resulting in a Bayes factors B_{10} of (1/.5=2) according to Formula 6.1. The posterior model probability then equals 2/(2+1)=2/3. This is appropriate behavior, since the posterior model probability shows the support of the data for each of the hypotheses corrected for the complexity of each of the hypotheses. If θ is much larger than 1, both hypotheses are true.

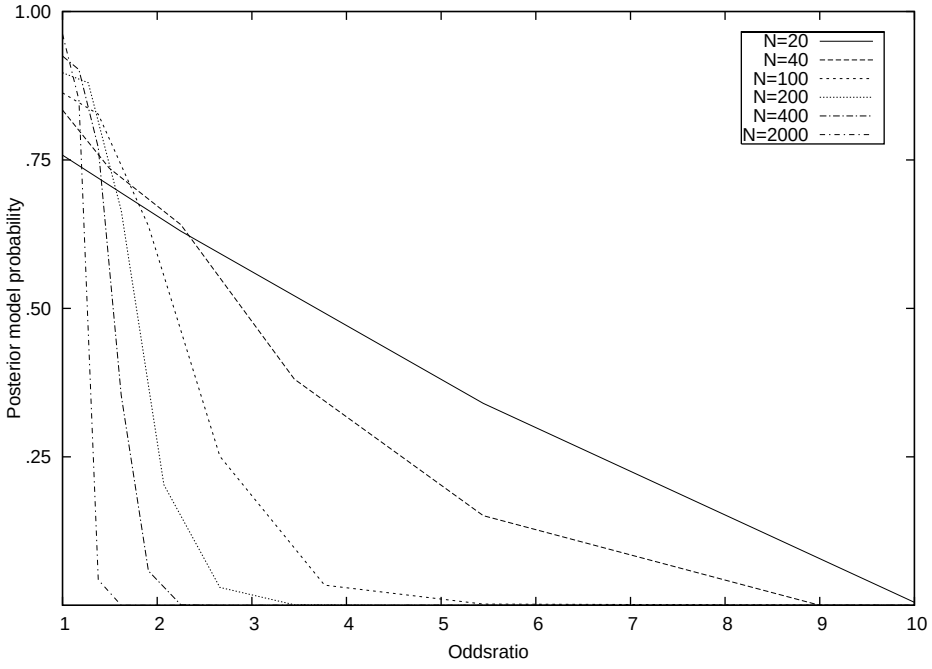


Figure 6.2: $\theta = 1$ versus θ

6.4.2 Two Odds Ratios

The behavior of the posterior model probability is also explored for the format of the data in Table 6.5.

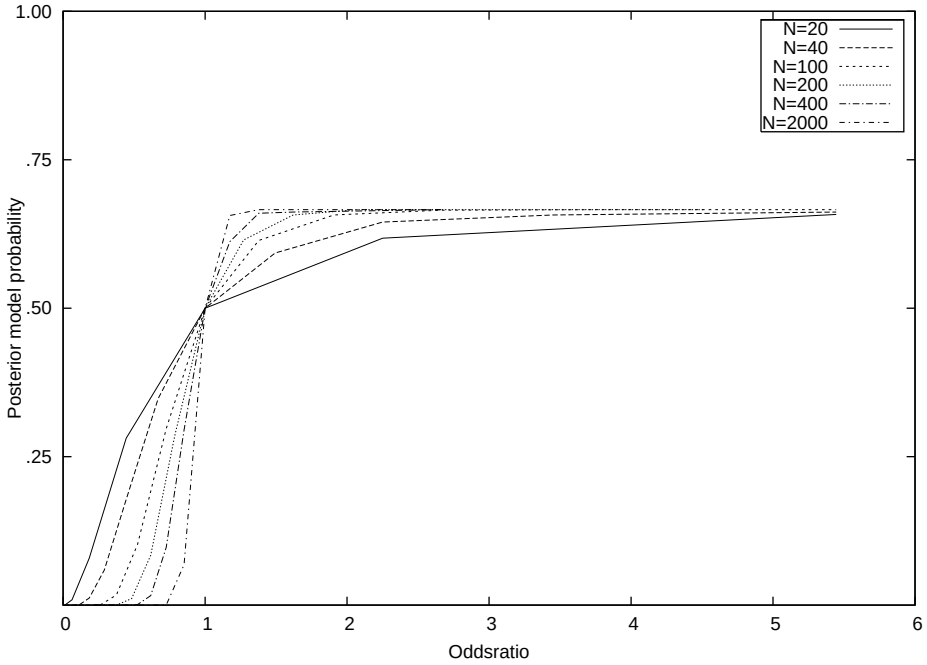


Figure 6.3: $\theta > 1$ versus θ

The contingency table displayed in Table 6.9 consists of two two-way tables. To construct data, for say sample size 40, the first two-way table (C=1) contains in each cell 5 observations, while the subsequent two-way tables (C=1) vary exactly as in Section 6.3.1. The new sequence of data sets obtained is presented in Table 6.9. Note that the sequence of odds ratios smaller than one is constructed in a similar way.

Table 6.9: Constructed Data N=20 for a 2x2x2 Contingency Table

		B		B		B			B	
C	A	1	2	1	2	1	2		1	2
1	1	5	5	5	5	5	5	$\cdots$	5	5
	2	5	5	5	5	5	5	$\cdots$	5	5
2	1	5	5	6	4	7	3	$\cdots$	9	1
	2	5	5	4	6	3	7	$\cdots$	1	9
θ_1		1		1		1		$\cdots$	1	
θ_2		1		2.25		5.44		$\cdots$	16	

The odds ratios θ_1 and θ_2 for C=1 and C=2, respectively are defined as $\pi_{111}\pi_{221}/(\pi_{211}\pi_{121})$ and $\pi_{112}\pi_{222}/(\pi_{212}\pi_{122})$. For each constructed data set, the hypothesis $\theta_1 = \theta_2$ is tested against θ_1, θ_2 and the results are displayed in Figure 6.4. The odds ratio in the second two-way table is displayed on the x-axis. Note that the figures for the sequence of odds ratios smaller than one is symmetric to the one presented, and hence are not displayed. The figure shows that with a sample size of 20 or 40 and the odds ratios both being equal to 1, the hypothesis that $\theta_1 = \theta_2$ always has a posterior model probability smaller than .5. For sample sizes over 100, the data increasingly support the hypothesis that both odds ratios are equal. The posterior model probability shows appropriate behavior for

$\theta_1 = \theta_2$: the larger the sample size, the more easy a difference between odds ratios θ_1 and θ_2 is detected.

In Figure 6.5, the setup is similar to Figure 6.4, and the hypothesis $\theta_1 > \theta_2$ is compared with θ_1, θ_2. Again can be seen that the larger θ_2, the larger the posterior model probability is for $\theta_1 > \theta_2$. Observe that the posterior model probability can not exceed 2/3, as explained in the previous example. Furthermore note that when the odds ratios are both equal to 1, the posterior model probability equals .5 for all sample sizes.

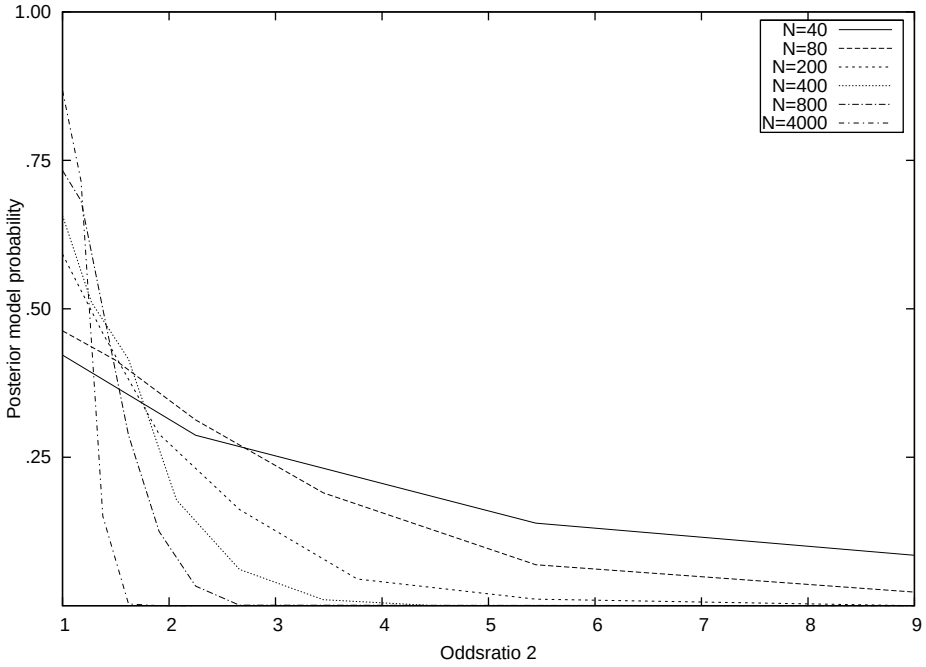


Figure 6.4: $\theta_1 = \theta_2$ versus θ_1, θ_2

6.4.3 Power versus Type 1 Error Rate

In classic hypothesis testing, Type 1 error, or α is fixed on .05. In terms of simulation, if samples are repeatedly drawn from a population where the null hypothesis is true, in 5% of the samples, the null hypothesis is (incorrectly) rejected. Power is closely related to the Type 1 error. Power is the probability that the null hypothesis is rejected when repeated samples are drawn from a population where the null hypothesis is not true. Decreasing the Type 1 error also decreases the power.

The posterior model probability is not associated with a fixed type one error rate. The posterior model probability shows the support for each model given the observed data, that is, if one chooses a model with a posterior model probability that equals .8, the probability of an incorrect decision is .2.

The performance in terms of Type 1 error rate of the posterior model probability can be investigated by sampling from a population where the null-hypothesis is true, and by displaying the relation between a value for the posterior model probability and the proportion of samples that has an equal or smaller value for the posterior model probability. This is different from Figures 6.2 and 6.3, since those display ideal samples,

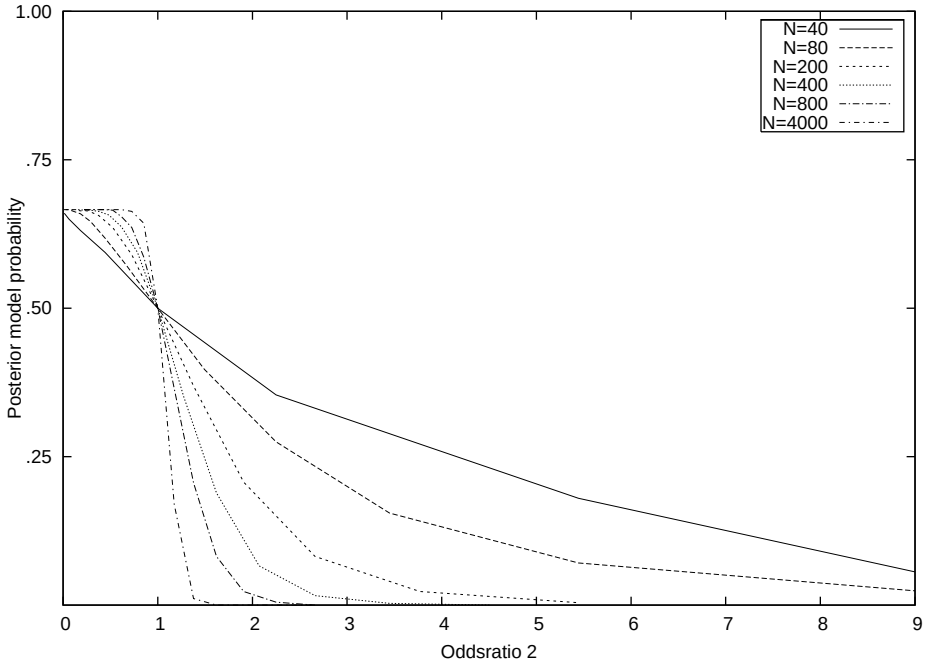


Figure 6.5: $\theta_2 > \theta_1$ versus θ_1, θ_2

i.e., what is the posterior model probability if the odds ratio has a particular value in the sample. When sampling from the population where the null hypothesis is true (i.e. in the population $\theta = 1$), some posterior model probabilities turn out very large or small as a result of the variability of the sampling distribution.

Consider a 2x2 contingency table with a probability of .25 in each cell. The odds ratio for this table in the population equals one. For various sample sizes 2000 data sets are sampled from this population. Suppose the posterior model probability for the hypotheses $H_1 : \theta = 1$ and $H_0 : \theta$ is evaluated. Figure 6.6 displays the cumulative distribution of the posterior model probability for H_1, that is, the relation between a value for the posterior model probability and the proportion of samples that has that a smaller or equal value for the posterior model probability. Note if there are two hypotheses, $H_1 : \theta = 1$ and $H_0 : \theta$, a posterior model probability of .5 shows equal support both hypotheses, making that .5 a cut-off value. The results in Figure 6.6 show for a sample size of $N = 20$ in 10% of the samples, the posterior model probability is smaller than .5. In other words, in 10% of the samples, the posterior model probability indicates that the odds ratio is unequal to one (while in the population, the odds ratio equals one). Note that if the sample size increases, the percentage that shows support for the odds being unequal to one, decreases. With a large sample size of $N = 2000$, none of the samples shows a posterior model probability smaller than .3, and in 1% of the samples, the posterior model probability is smaller than .5.

In Figure 6.7, the same procedure is executed for the hypotheses $\theta > 1$ and θ for N=20. The distribution of $\theta > 1$ is not clearly defined, thus for various values of θ, a sample has been drawn from the corresponding population. First, repeated samples are drawn for the population $\theta = .44$. Clearly, $\theta > 1$ does not hold. For a cut-off value of .5,

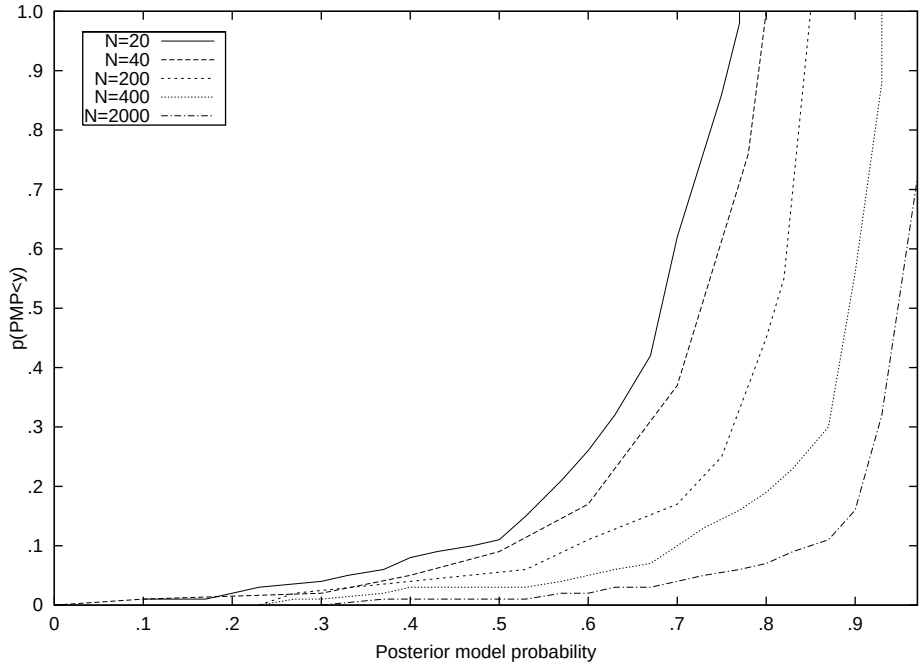


Figure 6.6: Error Rate for $\theta = 1$ versus θ

around 70% of the samples show a posterior model probability less than .5, thus in 30% of the samples, an odds ratio larger than 1 is detected. If the odds ratio in the population equals 1, 50% of the samples from this population show a posterior model probability less than .5. As the odds ratio increases, the percentage of samples that is detected with an odds ratio larger than 1, rapidly increases. If repeated samples are drawn from a population where the odds ratio equals 5.44, in around 97% of the samples, an odds ratio larger than 1 is detected.

For a sample size of $N = 400$, the results are displayed in Figure 6.8. The result look similar to Figure 6.7, however, the odds ratios are smaller: with a larger sample size, an odds ratio slightly larger than one, is easily detected as being larger than one.

The results in this section can be summarized as follows: it has been shown that when the hypothesis (i.e. $\theta = 1$, $\theta_1 = \theta_2$, $\theta > 1$ or $\theta_2 > \theta_1$) is not true, the posterior model probability for that hypothesis approaches to zero as the sample size increases. When the equality hypothesis is true, the sample size determines the value of the posterior model probability; the larger the sample size, the more the posterior model probability approaches 1. When an inequality constrained hypothesis is true, the quantity c_t, the proportion of samples from the unconstrained prior in agreement with hypothesis t determines the maximum value of the posterior model probability; the larger the sample size, the more the posterior model probability approaches that maximum. In the error rate simulations it was shown that for increasing sample sizes the probability of incorrectly rejecting both the equality and the inequality constrained hypothesis against the unconstrained hypothesis decreases. The simulations performed in Section 6.4 show that the properties of the posterior model probability are desirable in the context of selection of hypotheses. The posterior model probability compares the fit

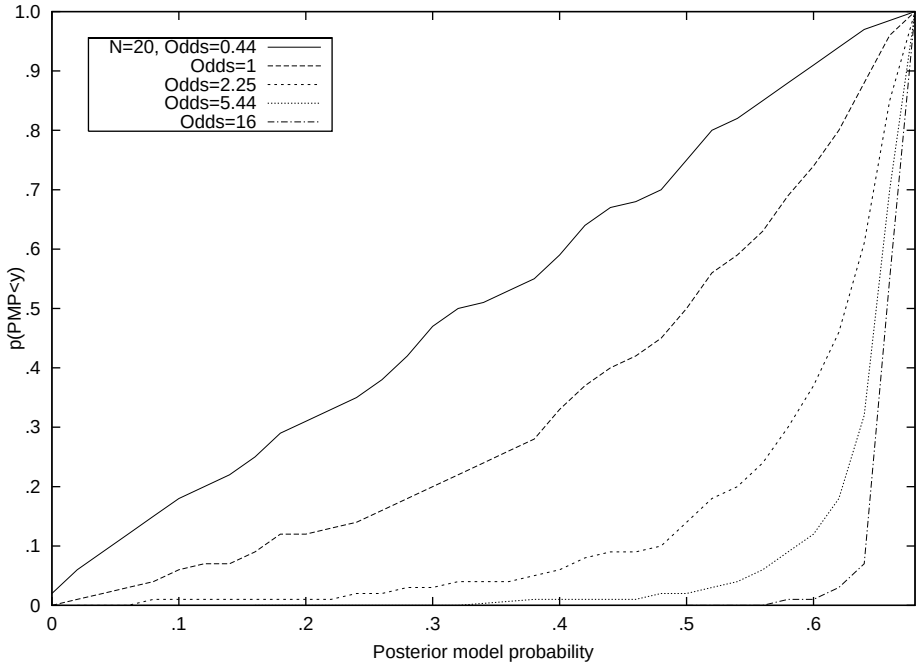


Figure 6.7: Error Rate for $\theta > 1$ versus θ, N=20

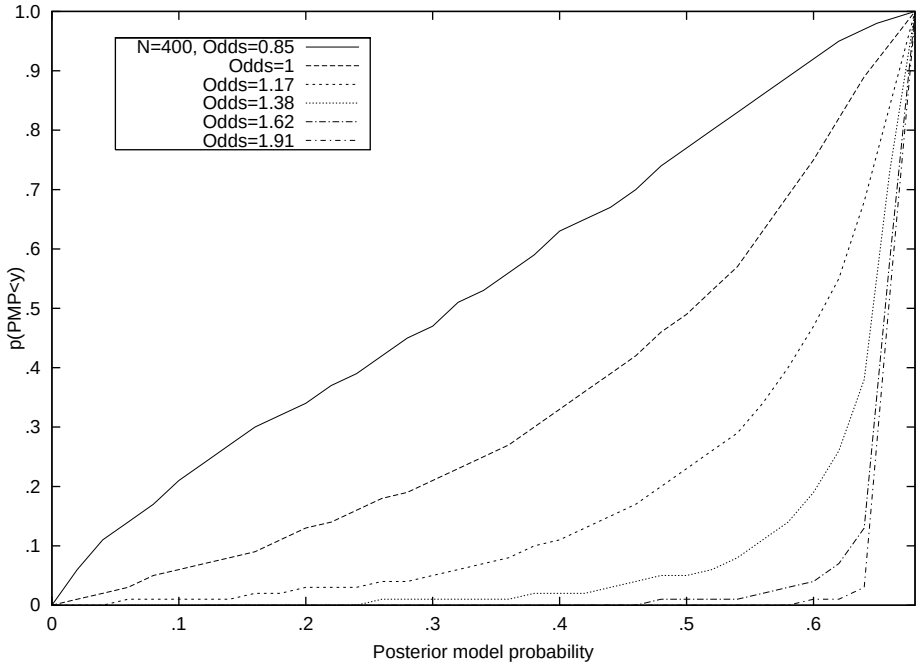


Figure 6.8: Error Rate for $\theta > 1$ versus θ, N=400

of two hypotheses while implicitly taking a penalty for model complexity into account.

To calculate the posterior model probability, a choice has to be made for the value of $\boldsymbol{\alpha}$. In the next section, the influence of the value for $\boldsymbol{\alpha}$ on the posterior model probability is discussed.

6.5 Sensitivity to the Prior

As was mentioned in Section 6.1, for equality constrained hypotheses, the posterior model probability may depend on the prior specification. In all previous analyses, a value for $\boldsymbol{\alpha}$ of 1 is used. This corresponds to assuming that a-priori (i.e. before observing the data) there is one person in each cell of the contingency table. An other way of viewing this is that a-priori cell probabilities uniformly distributed in the interval [0,1], that is the cell probabilities are as likely to be equal as being more extreme. If an $\boldsymbol{\alpha}$ smaller than 1 is chosen, a-priori is assumed that the cell probabilities are more extreme, i.e. before observing the data, one assumes that the cell probabilities are either small or large. In contrast, if an value for $\boldsymbol{\alpha}$ larger than 1 is chosen, cell probabilities tend to be more similar. We will show that for values of $\boldsymbol{\alpha}$ other than one, the properties of the posterior model probability are less desirable.

6.5.1 Sensitivity to the Prior for Odds Ratio Hypotheses

The prior sensitivity study for hypotheses in terms of odds ratios is set up as follows. The data from Table 6.8 is used again. For each data set the posterior model probability of $\theta = 1$ versus θ is calculated used different values of $\boldsymbol{\alpha}$. In Figure 6.9, the results are displayed for a sample size of $N = 20$. The figure shows that the posterior model probability depends on $\boldsymbol{\alpha}$, however, not too much. When a smaller $\boldsymbol{\alpha}$ is chosen, the posterior model probability of model $\theta = 1$ increases. In Figure 6.10 the sensitivity analysis is conducted using a sample size of $N = 400$. The figure shows that the sensitivity to $\boldsymbol{\alpha}$ is not related to sample size. In contrast, Figure 6.11 displays the posterior model probability of $\theta > 1$ versus θ for various values of $\boldsymbol{\alpha}$ and sample sizes 20 and 400. As it appears, prior sensitivity is completely absent when inequality constrained hypotheses are compared to unconstrained hypotheses.

The prior sensitivity is also investigated for the hypotheses $\theta_1 = \theta_2$ versus θ_1, θ_2 using the data from Table 6.9. In Figure 6.12, a prior sensitivity analysis is displayed for a sample size of $N = 40$. The line associated with an $\boldsymbol{\alpha}$ of one is the same as in Figure 6.4. The figure shows that the posterior model probability depends on $\boldsymbol{\alpha}$. When an $\boldsymbol{\alpha}$ of .5 is chosen, the highest posterior model probability is equal to .1. Also note that an $\boldsymbol{\alpha}$ greater than 1, has smaller effect than less than one. In Figure 6.13 the sensitivity analysis is conducted using a sample size of $N = 400$. This figure shows that using an $\boldsymbol{\alpha}$ of .5, even with a sample size of $N = 400$, if the odds are equal, the posterior model probability still remains around .1. If the odds are equal, with an $\boldsymbol{\alpha}$ of .8, the posterior model probability is slightly larger than .5. This behavior is not desirable. In an experiment with a sample size of $N = 400$, with odds ratios being exactly equal, firm conclusions should be allowed. Furthermore, note that $\boldsymbol{\alpha}$'s over one have more similar posterior model probabilities than $\boldsymbol{\alpha}$'s less than one. We recommend using an $\boldsymbol{\alpha}$ of

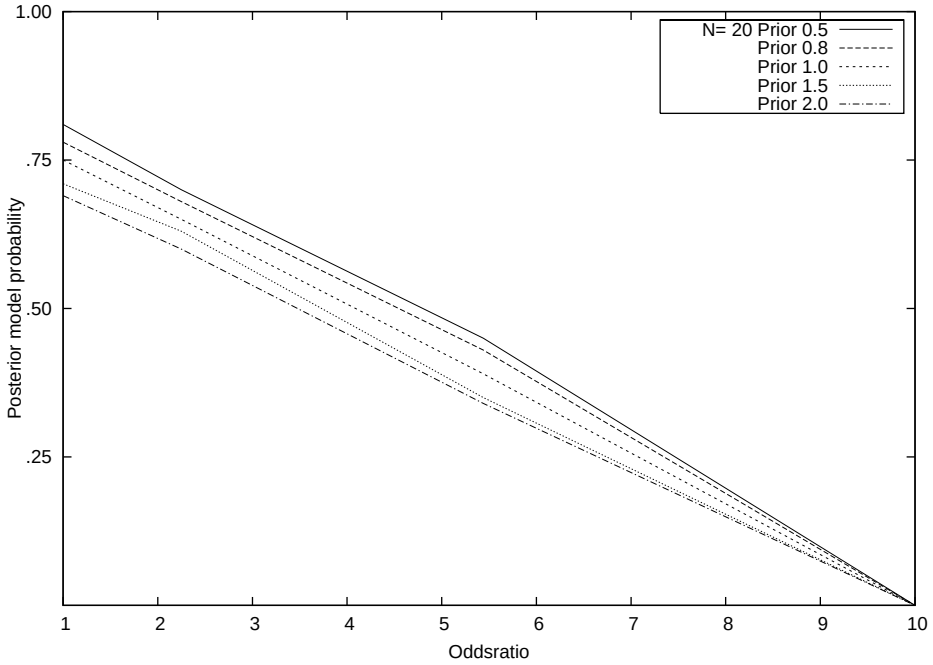


Figure 6.9: $\theta = 1$ versus θ, $N = 20$

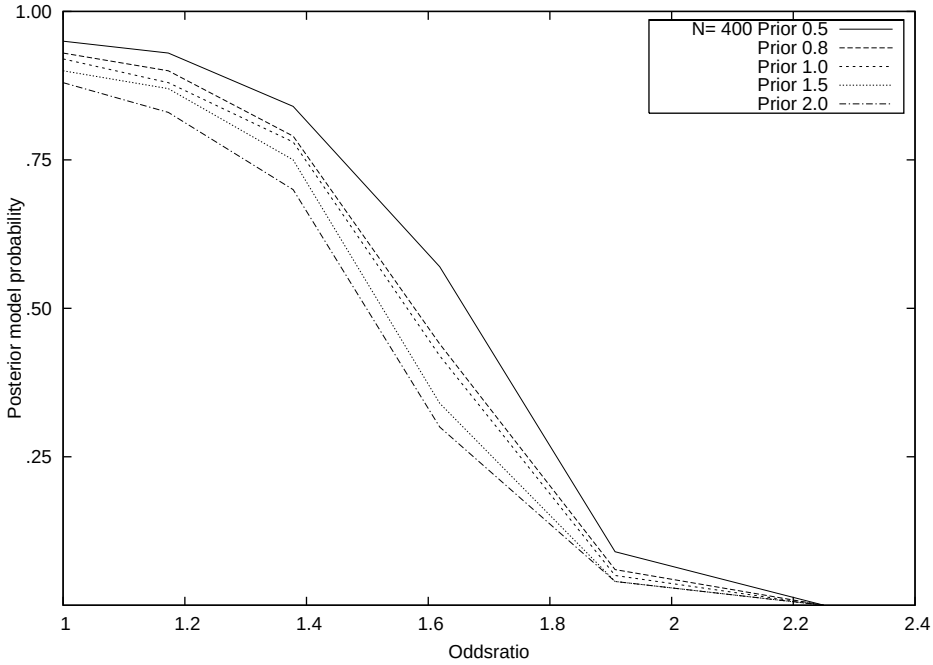


Figure 6.10: $\theta = 1$ versus θ, $N = 400$

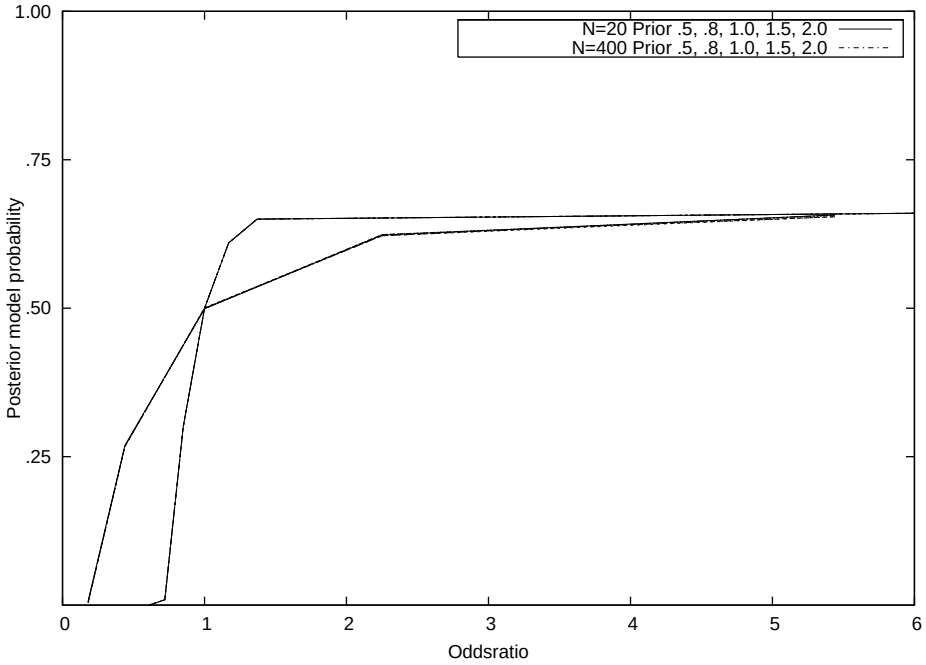


Figure 6.11: $\theta > 1$ versus θ

one. This value has the intuitive interpretation that one does not favor certain values in terms of cell probabilities. Moreover, if hypothesis are stated in terms of odds ratios, the posterior model probability show appropriate behavior, with large enough power.

In contrast to the equality constrained hypotheses, inequality constrained hypotheses do not show sensitivity to the prior. In Figure 6.14, the posterior model probability of $\theta_2 > \theta_1$ versus θ_1, θ_2 is evaluated, using different $\boldsymbol{\alpha}$'s. With sample size $N = 40$, the sensitivity is small, mainly due to the effect of adding prior observations to each cell has. With a sample size of $N = 400$, no sensitivity to the prior can be detected.

Finally we show the results of the two analyses of Nash and Bowen (2002) when different $\boldsymbol{\alpha}$'s are used. Table 6.10 shows the results for $\boldsymbol{\alpha}$ varying from .5 to 2. The results show that the inequality constrained model appears not to be sensitive, while the equality constrained model shows little sensitivity. Thus, the conclusions do not change if a different $\boldsymbol{\alpha}$ is used.

Table 6.10: Posterior Model Probabilities for the Hypotheses in Table 6.3 for Different Values of the Prior

	Post. prob.		
Comparison	$\boldsymbol{\alpha}$=.5	$\boldsymbol{\alpha}$=1.0	$\boldsymbol{\alpha}$=2.0
$P(H_1 \mid H_0, H_1)$:	.74	.70	.60
$P(H_2 \mid H_0, H_2)$:	.67	.67	.67

The results of various $\boldsymbol{\alpha}$'s for the hypotheses presented in Table 6.6 are shown in Table 6.11. The equality constrained models show sensitivity to the prior, while the inequality constrained hypotheses do not show sensitivity to the prior.

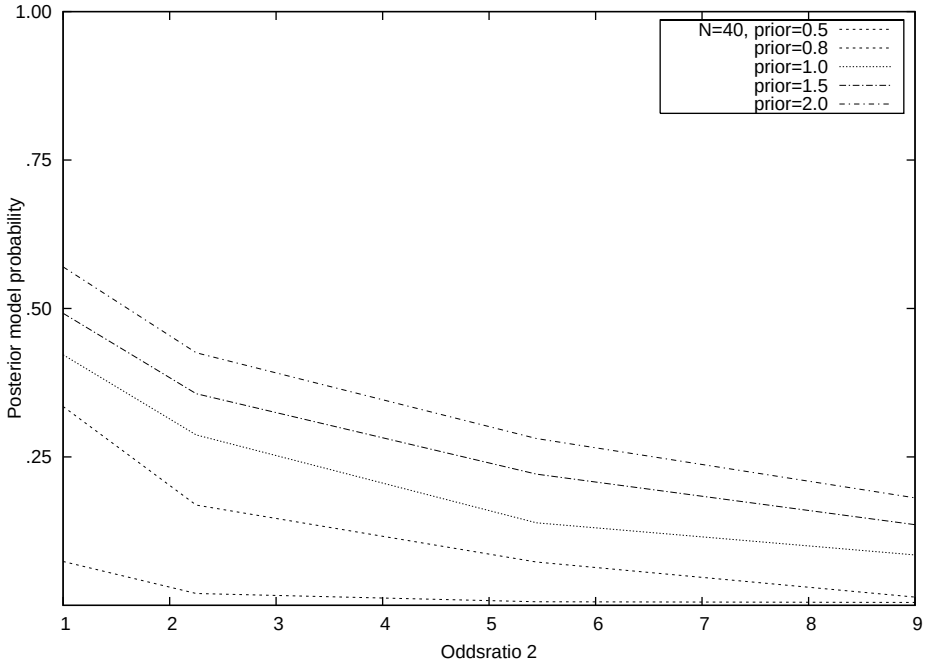


Figure 6.12: $\theta_1 = \theta_2$ versus θ_1, θ_2

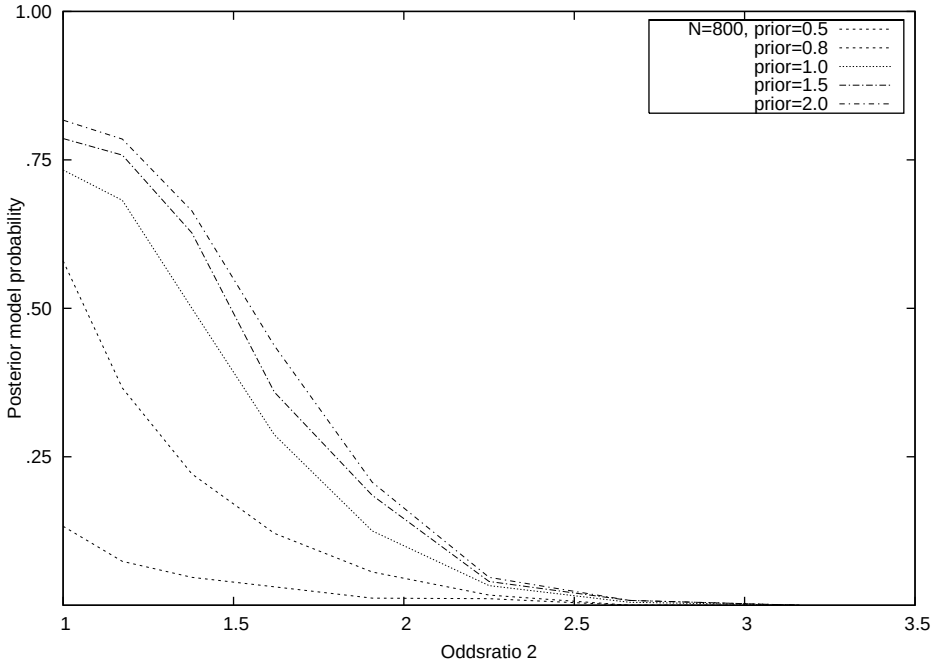


Figure 6.13: $\theta_1 = \theta_2$ versus θ_1, θ_2

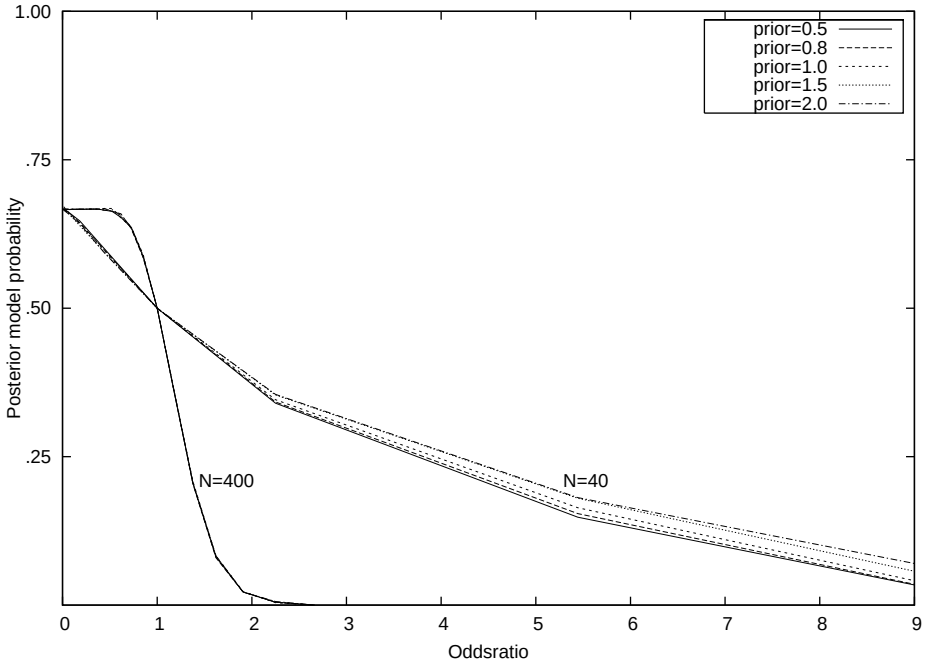


Figure 6.14: $\theta_1 > \theta_2$ versus θ_1, θ_2

Table 6.11: Posterior Model Probabilities for the Hypotheses in Table 6.6 for Different Values of the Prior

	Post. prob.		
Comparison	$\boldsymbol{\alpha}$=.5	$\boldsymbol{\alpha}$=1.0	$\boldsymbol{\alpha}$=2.0
$P(H_1 \mid H_0, H_1)$:	.53	.28	.26
$P(H_2 \mid H_0, H_2)$:	.01	.19	.33
$P(H_3 \mid H_0, H_3)$:	.77	.77	.77
$P(H_4 \mid H_0, H_4)$:	.67	.67	.67
$P(H_5 \mid H_0, H_5)$:	.87	.87	.87

6.5.2 Sensitivity to the Prior for Hypotheses in Terms of Odds

The hypotheses discussed in the previous section have been posed in terms of odds ratios. However another way to analyze contingency tables is in terms of the ratio of two probabilities, or odds. In Section 6.6.1, an example of real data is given with hypotheses in terms of odds.

Consider a 3 x 2 contingency table as displayed in Table 6.12. The hypotheses of interest are displayed in Table 6.13. Hypothesis H_0 is the unconstrained model. Hypothesis H_1 restricts the odds to be equal in the columns, and hypothesis H_2 requires the odds to be ordered. As in the previous simulations, hypothesis H_1 and H_2 are contrasted against the unconstrained model H_0.

The data for the simulation is constructed in the following way: the sample size is held constant with $N = 300$. The constructed data are displayed in Table 6.14. The first contingency table displays data that is not in agreement to hypothesis H_2, and with each table below, the data conforms more and more to hypothesis H_2. The right hand side

Table 6.12: 3 x 2 Contingency Table

	A		
B	1	2	3
1	π_{11}	π_{12}	π_{13}
2	π_{21}	π_{22}	π_{23}

Table 6.13: Hypotheses for the 3 x 2 Contingency Table

H_0:	$\pi_{11}/\pi_{21}, \quad \pi_{11}/\pi_{21}, \quad \pi_{11}/\pi_{21}$
H_1:	$\pi_{11}/\pi_{21} = \pi_{11}/\pi_{21} = \pi_{11}/\pi_{21}$
H_2:	$\pi_{11}/\pi_{21} > \pi_{11}/\pi_{21} > \pi_{11}/\pi_{21}$

of the table displays the odds for each contingency table. Note that with respect to the equality hypothesis H_1, only the table where each entry equals 100 is true. Moreover, the deviation from hypothesis H_1 is symmetric in either direction from the table where each entry equals 100, hence for the equality hypothesis only the results from the table where each entry equals 100 and below are displayed.

Table 6.14: Constructed Data for the 3 x 2 Contingency Table

	Cell counts			Odds		
	A			A		
B	1	2	3	1	2	3
1	80	90	100			
2	120	110	100	.66	.82	1
1	90	95	100			
2	110	105	100	.82	.90	1
1	100	100	100			
2	100	100	100	1	1	1
	⋮	⋮	⋮			
1	180	140	100			
2	20	60	100	9	2.33	1

The results for the equality hypothesis H_1 against the unconstrained hypothesis H_0 are displayed in Figure 6.15. On the x-axis the frequency of call (1,1) is displayed, corresponding to Table 6.14. The figure shows small sensitivity to the prior only for $\boldsymbol{\alpha}$ equals .5. Furthermore, an $\boldsymbol{\alpha}$e of 1 does give reasonable results. The results for the ordering hypothesis H_2 against the unconstrained hypothesis H_0 are displayed in Figure 6.16. Again can be observed that inequality constrained hypotheses are not sensitive to the value of $\boldsymbol{\alpha}$.

This section can be summarized as follows: the simulations showed that the equality constrained hypotheses show sensitivity to $\boldsymbol{\alpha}$. For the hypothesis $\theta = 1$ this sensitivity was not large, and different conclusions will not been drawn if a different $\boldsymbol{\alpha}$ is used. For the hypothesis $\theta_1 = \theta_2$, however, the sensitivity was large, and hence the conclusion may

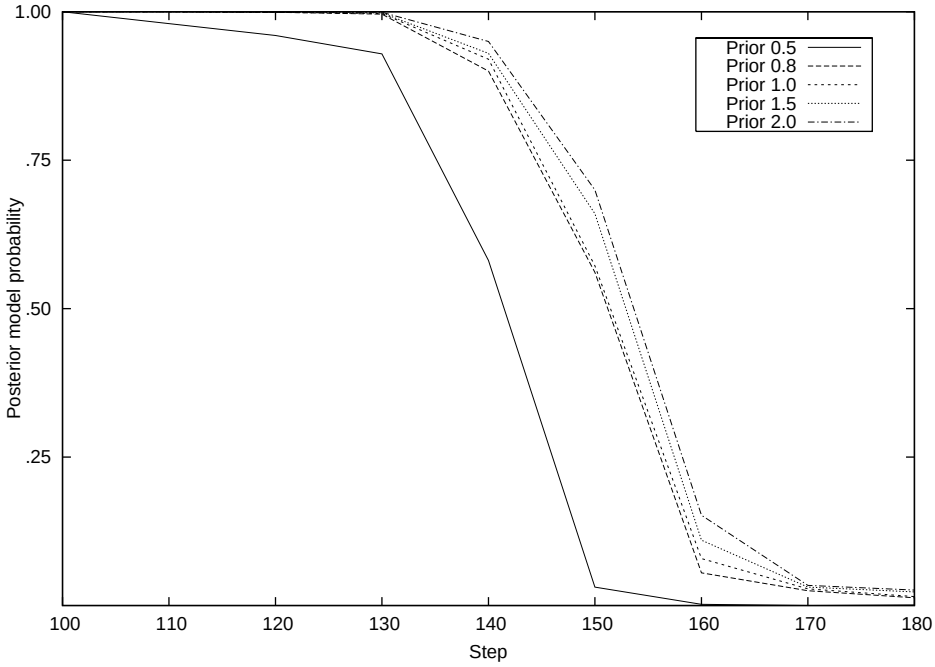


Figure 6.15: $\pi_1 = \pi_2 = \pi_3$ versus π_1, π_2, π_3

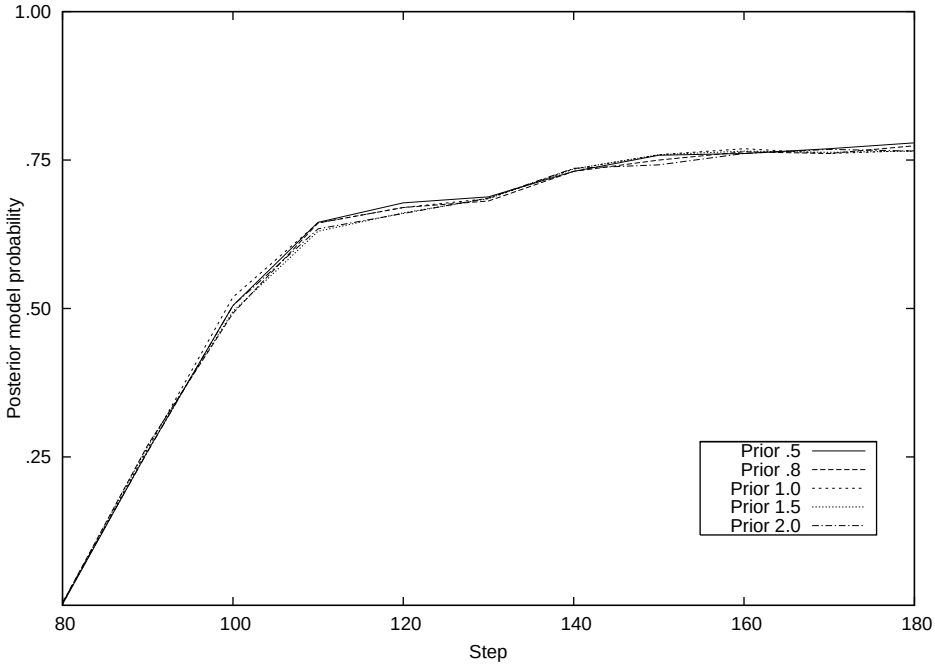


Figure 6.16: $\pi_1 > \pi_2 > \pi_3$ versus π_1, π_2, π_3

depend on the chosen value for $\boldsymbol{\alpha}$. For the hypothesis in terms of odds, the sensitivity was not very large, apart from an $\boldsymbol{\alpha}$ of .5. Inequality constrained hypotheses at the other hand do not show sensitivity to the value of $\boldsymbol{\alpha}$. We advocate an $\boldsymbol{\alpha}$ of 1, as this value shows reasonable results and has an intuitive interpretation of not favoring certain values of cell probabilities over others. In the remainder of this paper, an $\boldsymbol{\alpha}$ of 1 will be used.

6.6 Elaborate Examples

In this section, more elaborate examples are analyzed. These data are taken from recently published research. It is shown how theories are translated into inequality constrained hypotheses in various situations. Note that it is not possible to compare the outcome of the posterior model probability to the outcomes reported in the articles, since the hypotheses are more informative. However, it is shown how the original research question was posed, and how more detailed knowledge can be obtained from using inequality constrained hypotheses.

The first example concerns hypotheses about odds. In the second example, it is shown how interaction terms can be included using inequality constrained hypotheses. In the third example a larger contingency table is analyzed, showing how theoretic consideration about the data give rise to a set of inequality constrained hypotheses. The last example deals with ordinal data and a set of inequality constraints that takes the ordinal nature of the data into account.

6.6.1 Sexual Abuse and Bulimic Behavior

Perkins and Luster (1999) investigate the relationship between sexual abuse and bulimic behavior, namely, purging. A sample of 7,903 female adolescents, ages 12– 17 years, was collected from a large Midwestern state. The Search Institutes Profiles of Student Life: Attitude and Behavior Questionnaire (ABQ), a 152-item inventory developed by the Search Institute (Benson, 1990; Blyth, 1993) was administered by classroom teachers.

Purging is the bulimic behavior measured by one item on the ABQ. It was originally registered using a 5 point scale: "How often do you vomit (throw up) on purpose after eating?" The range of responses were: never, once a month or less, 2-3 times a month, once a week, 2 or more times a week. Based on DSM-IV criteria, responses on the last category were assigned to the problem group. Those in the intermediate categories, once per month to once per week, were excluded from the analysis. The amount of purging these adolescents engaged in fell short of the DSM-IV criterion for bulimia but these adolescents were not viewed as appropriate for the no-problem group. Sexual abuse was measured by one item: "Have you ever been sexually abused?" The range of choices was: never, once, 2-3 times, 4-10 times, more than 10 times. Table 6.15 shows the data. The table classifies purging by sexual abuse. To specify the hypotheses, the rows are indexed with i, $i = 1, ..., 2$, the columns are indexed by j, with $j = 1, ..., 5$.

The objective of the study was to examine the relationship between sexual abuse and purging. Table 6.16 displays the hypotheses for the data. Hypothesis H_0 is the unconstrained hypothesis. The odds are defined in the following way: π_{11}/π_{21} indicates the ratio that a respondent who has never been sexually abused purges more than twice

Table 6.15: Cross-classification of Purging by Sexual Abuse

	Sexual abuse				
	Never	Once	2-3 Times	4-10 Times	>10 Times
Purging >2x/week	103	18	17	8	8
No purging	5259	644	332	108	121

a week over not purging. Hypothesis H_1 states that the odds for all categories of sexual abuse are equal. The second hypothesis states that the odds of purging increase as the frequency of sexual abuse increases. It is of interest to contrast hypotheses H_2 with H_1.

Table 6.16: Hypotheses for the Data in Table 6.15

H_0 :	Unconstrained								
H_1 :	π_{11}/π_{21}	$=$	π_{12}/π_{22}	$=$	π_{13}/π_{23}	$=$	π_{14}/π_{24}	$=$	π_{15}/π_{25}
H_2 :	π_{11}/π_{21}	$<$	π_{12}/π_{22}	$<$	π_{13}/π_{23}	$<$	π_{14}/π_{24}	$<$	π_{15}/π_{25}

The results are displayed in Table 6.17. The posterior model probability of H_1 indicates that equal odds for each category of abuse is not supported by the data. The posterior model probability of the hypotheses H_2 that required increasing odds shows very decisive support for the data. When hypothesis H_2 is contrasted with hypothesis H_1, the posterior model probability (.995) shows that the ordered odds are far more supported than the hypothesis of equal odds.

Table 6.17: Posterior Model Probabilities of the Hypotheses in Table 6.21

Comparison	Post. prob.
$P(H_1 \mid H_0, H_1)$:	.124
$P(H_2 \mid H_0, H_2)$:	.965
$P(H_2 \mid H_1, H_2)$:	.995

6.6.2 Attachment Patterns and Suicidal Ideation

Lessard and Moretti (1998) conducted a study to investigate the relationship between attachment patterns and suicidal ideation. Respondents were assessed on the level of current ideation through a self-report questionnaire. Quality of attachment to care-givers based on semi-structured clinical interviews was assessed using Bartholomew's two-dimensional model of attachment (Bartholomew, 1990). The first dimension consists of Positivity of the Self (the degree of self-worth versus anxiety and dependency on others approval), the second dimension consists of positivity of Other (the degree to which one tends to seek out or avoid closeness in relationships). This two-dimensional model yields 4 attachment patterns, displayed in Figure 6.17

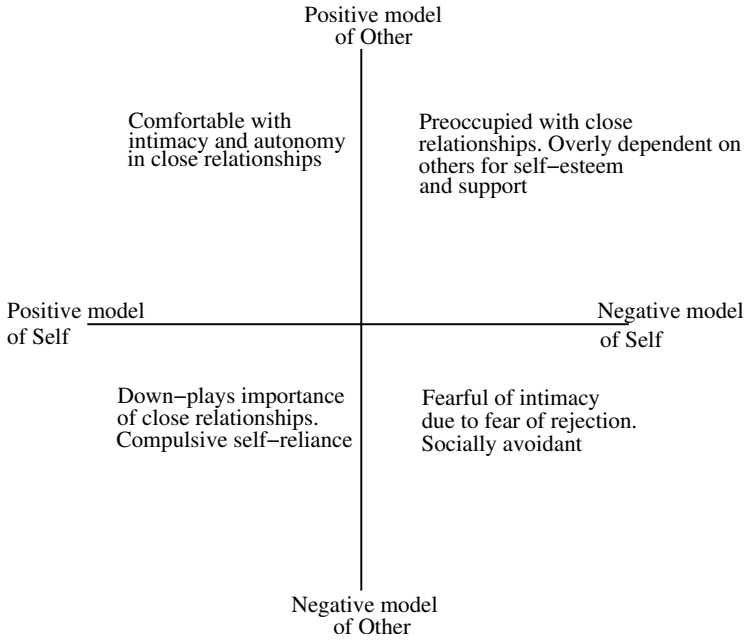


Figure 6.17: 4 Attachment Patterns

Table 6.18 displays the 116 cross-classified respondents. To discuss the hypotheses, the variables are indexed. Self-image is indexed by $i = 1, 2$, Other-image by $j = 1, 2$, and Suicidal ideation is indexed by $k = 1, 2$.

Table 6.18: Cross-classification of Self-image and Other-image by Suicidal Ideation

		Suicidal ideation	
		yes	no
Self image	Other image		
Positive	positive	4	5
	negative	8	14
Negative	positive	19	8
	negative	37	9

Table 6.19 displays the hypotheses for the the data. First, the unconstrained model is displayed. Lessard and Moretti (1998) are interested in the relationship between attachment patterns and suicidal ideation. It is of interest whether a positive of negative Self-image is related to suicidal ideation. To investigate this hypothesis, odds are defined: π_{1+1}/π_{1+2} is the odds that respondents with a positive Self-image show suicidal ideation over non-suicidal ideation. Note that the +-sign indicated summing over the corresponding index: $\sum_{j=1}^{2} \pi_{1j1}$ Hypothesis H_1 displays the hypothesis that respondents with a positive Self-image have smaller odds for suicidal ideation than respondents with a negative Self-image. Next, it is of interest whether positive or negative Other-image is related to suicidal ideation. Hypothesis H_2 displays the hypothesis that positive Other-image is related to a lower odds for suicidal ideation than negative Other-image. Finally, it is of interest whether on each level of Self-image, Other-image has an effect with respect

to suicidal ideation. Hypothesis H_3 states that respondents with both a positive Self and Other-image have the smallest odds on suicidal ideation. The largest odds on suicidal ideation is related to the cell in the contingency table where both Self-image en Other-image are negative. The odds on suicidal ideation where either one of the categories of Self-image or Other-image is negative are restricted to lie in between.

Table 6.19: Hypotheses for the Data in Table 6.18

H_0:	Unconstrained
H_1:	$\pi_{1+1}/\pi_{1+2} < \pi_{2+1}/\pi_{2+2}$
H_2:	$\pi_{+11}/\pi_{+12} < \pi_{+11}/\pi_{+12}$
H_3:	$(\pi_{111}/\pi_{112} < \pi_{121}/\pi_{122} < \pi_{221}/\pi_{222})$, $(\pi_{111}/\pi_{112} < \pi_{211}/\pi_{212} < \pi_{221}/\pi_{222})$

The results are displayed in Table 6.20. It shows that both Self-image and Other-image influence the odds on suicidal ideation. The hypothesis that on each level of Self-image, Other-image has an effect with respect to suicidal ideation has the highest posterior model probability. Since each hypotheses fits better than the unconstrained hypothesis, the best of the three hypotheses is selected by contrasting the three without including the unconstrained hypothesis. It is concluded that respondents having both a positive Self-image and Other-image show less suicidal ideation as compared to respondents with both a negative Self- and Other-image.

Table 6.20: Posterior Model Probabilities of the Hypotheses in Table 6.18

Comparison	Post. prob.
$P(H_1\|H_0, H_1)$:	.667
$P(H_2\|H_0, H_2)$:	.550
$P(H_3\|H_0, H_4)$:	.768
$P(H_1\|H_1, H_2, H_3)$:	.336
$P(H_2\|H_1, H_2, H_3)$:	.277
$P(H_3\|H_1, H_2, H_3)$:	.387

6.6.3 The Effect of Prior Dispositions on Current Disposition for Young Delinquents

This example shows how two competing theories can be translated into specific constraints on sums of cell probabilities. Subsequently, the best of the two hypotheses is selected. Matarazzo, Carrington and Hiscott (2001) investigate the effect of prior dispositions on current disposition for young delinquents. This study uses data from the Canadian Youth Court Survey (YCS) for fiscal 1993-1994. The unit of analysis is the "case" which is operationalized as all the charges pertaining to an offender which were disposed of (sentenced) at the same court hearing. In order to study the effect and sequencing of prior dispositions, the data were limited to the 16,636 cases involving young offenders with at least two previous cases that reached disposition. There are four types of disposition: Other, a residual category of dispositions including a fine not

exceeding $1000, community service order, order for compensation or restitution, etc; Probation; Open custody; and, Secure custody. A small number (94) of youth court cases involving very serious charges and-or lengthy criminal histories were transferred to ordinary (adult) criminal court and are excluded from this study.

"Most recent prior disposition" was defined as the most serious disposition in the case with the most recent date of disposition before the date of the disposition of the current case. "Second most recent prior disposition" was defined as the most serious disposition in the case with the most recent date of disposition before the date of disposition for the most recent prior disposition. The data are displayed in Table 6.21

Table 6.21: Cross-classification of Current, Most Recent and Second Most Recent Disposition

		Current prior disposition			
		Other	Probation	Open	Secure
2-nd recent	Most recent				
Other	Other	373	294	100	106
Probation		427	507	239	191
Open		75	74	74	80
Secure		55	58	48	101
Other	Probation	379	496	242	168
Probation		612	1399	814	499
Open		87	168	167	171
Secure		77	127	90	172
Other	Open	72	109	168	138
Probation		177	356	696	421
Open		103	207	537	473
Secure		42	82	183	311
Other	Secure	62	78	51	174
Probation		122	220	205	442
Open		85	145	269	576
Secure		125	195	269	1074

One idea in criminology is stabilization. The pattern of the relationship between prior dispositions and current dispositions is characterized primarily by stabilization: the current disposition tends to be similar to prior dispositions. It is of interest to determine which hypothesis is best: H_1 : the most recent prior disposition is similar to the current prior disposition and H_2 : the second recent prior disposition is similar to the current prior disposition.

To clarify the research question, two new tables are constructed. Table 6.21 is indexed by i, j, k, where i indexes the most recent prior disposition, j the second recent prior disposition and k the current disposition. The upper half of Table 6.22 displays the data summed over the second recent prior disposition, and normalized such that the row sum equals 1. The hypothesis H_1 states that the current disposition is similar to the most prior disposition. This implies that the diagonal probabilities are larger than the probabilities in the corresponding row and column. This is formalized in Table 6.23 in hypothesis H_1. In the lower half of Table 6.22, the data are summed over the most recent prior disposition and normalized such that the row sum equals 1. The hypothesis H_2 states that the second recent prior disposition is similar to the current disposition. In

terms of probabilities this means that all the diagonal probabilities are restricted to be larger than the probabilities in the corresponding row and column. It is of interest which hypothesis is most supported by the data.

Table 6.22: Turnover Tables for the Data in Table 6.21

	Current prior disposition			
	Other	Probation	Open	Secure
Most recent				
Other	.29	.32	.19	.19
Probation	.18	.34	.27	.21
Open	.11	.18	.32	.40
Custody	.10	.15	.20	.55
Second most recent				
Other	.33	.33	.16	.17
Probation	.20	.39	.23	.18
Open	.10	.19	.39	.33
Custody	.10	.16	.19	.55

Table 6.23: Hypotheses for the Data in Table 6.21

H_0 :	Unconstrained
H_1 :	$\pi_{1+1}/\pi_{1++} > \pi_{1+k}/\pi_{1++}$, $k \neq 1$
	$\pi_{1+1}/\pi_{1++} > \pi_{i+1}/\pi_{i++}$, $i \neq 1$
	$\pi_{2+2}/\pi_{2++} > \pi_{1+k}/\pi_{1++}$, $k \neq 2$
	$\pi_{2+2}/\pi_{2++} > \pi_{i+1}/\pi_{i++}$, $i \neq 2$
	$\pi_{3+3}/\pi_{3++} > \pi_{1+k}/\pi_{1++}$, $k \neq 3$
	$\pi_{3+3}/\pi_{3++} > \pi_{i+1}/\pi_{i++}$, $i \neq 3$
	$\pi_{4+4}/\pi_{4++} > \pi_{1+k}/\pi_{1++}$, $k \neq 4$
	$\pi_{4+4}/\pi_{4++} > \pi_{i+1}/\pi_{i++}$, $i \neq 4$
H_2 :	$\pi_{+11}/\pi_{+1+} > \pi_{+1k}/\pi_{+1+}$, $k \neq 1$
	$\pi_{+11}/\pi_{+1+} > \pi_{+j1}/\pi_{+j+}$, $j \neq 1$
	$\pi_{+11}/\pi_{+1+} > \pi_{+1k}/\pi_{+1+}$, $k \neq 2$
	$\pi_{+11}/\pi_{+1+} > \pi_{+j1}/\pi_{+j+}$, $j \neq 2$
	$\pi_{+11}/\pi_{+1+} > \pi_{+1k}/\pi_{+1+}$, $k \neq 3$
	$\pi_{+11}/\pi_{+1+} > \pi_{+j1}/\pi_{+j+}$, $j \neq 3$
	$\pi_{+11}/\pi_{+1+} > \pi_{+1k}/\pi_{+1+}$, $k \neq 4$
	$\pi_{+11}/\pi_{+1+} > \pi_{+j1}/\pi_{+j+}$, $j \neq 4$

The results are displayed in Table 6.24. The hypothesis that the most recent prior disposition is similar to the current prior disposition is not supported by the data. In contrast, the posterior model probability shows decisive support for the hypothesis that the second most prior disposition is similar to the current prior disposition. Finally, the two hypotheses are contrasted against each other. The posterior model probability clearly shows that hypothesis H_2 is most supported by the data. It is concluded that the stabilization hypothesis is a good hypothesis for the data, and that the second most prior disposition is more similar to the current disposition that the most recent prior disposition.

Table 6.24: Posterior Model Probabilities of the Hypotheses in Table 6.21

Comparison	Post. prob.
$P(H_1\|H_0, H_1)$:	.000
$P(H_2\|H_0, H_2)$:	.993
$P(H_1\|H_1, H_2)$:	.000
$P(H_2\|H_1, H_2)$:	1.000

6.6.4 The Relation between the Number of Sibling and Happiness

In the last example, more elaborate hypotheses are tested. Vermunt (1991) uses a contingency table that classifies the number of siblings by happiness, where both variables are ordinal. Table 6.25 displays the 1517 respondents classified in three categories of increasing happiness and in 5 categories of increasing number of siblings.

Table 6.25: Cross-classification of the Number of Sibling by Happiness

	Happiness		
	Not too happy	Pretty Happy	Very happy
Number of Siblings			
0-1	99	155	19
2-3	153	238	43
4-5	115	163	40
6-7	63	133	32
8+	99	118	47

The main research question is whether respondent with more siblings are happier. The first hypothesis in Table 6.26 is the unconstrained model. To incorporate the ordinal nature of happiness, cumulative probabilities π^C are defined. The first column of cumulative probabilities are defined as $\pi^C_{i,1} = \pi_{i1}/\pi_{i+}$, for $i = 1, ..., 5$. The second column is defined as $\pi^C_{i,2} = (\pi_{i1} + \pi_{i2})/\pi_{i+}$, , for $i = 1, ..., 5$. Note that the third column, defined as $\pi^C_{i,3} = (\pi_{i+})/\pi_{i+}$, for $i = 1, ..., 5$ equals 1. Cumulative probabilities provide a way to use the last category of Table 6.25 as a reference category. If the number of siblings and happiness are positively related, the cumulative probabilities are expected to be decreasing in the first two columns. The upper half of Table 6.27 shows that the observed cumulative probabilities are not all in decreasing order. It is to be investigated whether this can be assumed to be structural, or due to sampling fluctuations. Hypothesis H_1 investigates whether for each column the cumulative probabilities can be assumed equal. This will provide a reference for hypothesis H_2, where the cumulative probabilities for each column are decreasing with increasing number of siblings. Note that the above definition of cumulative probabilities incorporates the ordinal nature of happiness, but neglects the ordinal nature of the number of siblings. Incorporating the ordinal nature of the number of siblings, that is, relaxing the ordering, may result in a better fit. This is done by defining double cumulative probabilities $\pi^{CC}_{i,1} = \sum_{k=1}^{i} \pi_{i1} / \sum_{k=1}^{i} \pi_{i+}$, for $i = 1, ..., 5$, thus, in each row all preceding probabilities are summed. For the second

row, the double cumulative probabilities become $\pi_{i,2}^{CC} = \sum_{k=1}^{i}(\pi_{i1}+\pi_{i2}+)/\sum_{k=1}^{i}\pi_{i+}$, for $i = 1, ..., 5$. The observed double cumulative probabilities are displayed in the lower half of Table 6.27, and as can be seen, there is only one violation of a decreased ordering, instead of two for the single cumulative ordering. Hypothesis H_3 states that in each column, all double cumulative probabilities are equal. This provides a reference to hypothesis H_4, where for each column, all double cumulative probabilities are decreasing as the number of siblings increases.

Table 6.26: Hypotheses for the Data in Table 6.25

H_0 :	Unconstrained	
H_1 :	$\pi_{i1}^{C} = \pi_{i+1,1}^{C}$, for $i = 1, ..., 4$ $\pi_{i2}^{C} = \pi_{i+1,2}^{C}$, for $i = 1, ..., 4$	
H_2	$\pi_{i1}^{C} > \pi_{i+1,1}^{C}$, for $i = 1, ..., 4$ $\pi_{i2}^{C} > \pi_{i+1,2}^{C}$, for $i = 1, ..., 4$	
H_3:	$\pi_{i,1}^{CC} = \pi_{i+1,1}^{CC}$, for $i = 1, ..., 4$ $\pi_{i,2}^{CC} = \pi_{i+1,2}^{CC}$, for $i = 1, ..., 4$	
H_4:	$\pi_{i,1}^{CC} > \pi_{i+1,1}^{CC}$, for $i = 1, ..., 4$ $\pi_{i,2}^{CC} > \pi_{i+1,2}^{CC}$, for $i = 1, ..., 4$	
H_5 :	$\theta_{ij} = 1$,for $i = 1, ..., 4$,$j = 1, 2$	(independence)
H_6 :	$\theta_{ij} > 1$, for $i = 1, ..., 4$, $j = 1, 2$	(positive association)
H_7 :	$\theta_{ij} = \beta$,for $i = 1, ..., 4$,$j = 1, 2$	(uniform association)
H_8 :	$\theta_{ij} = \beta > 1$,for $i = 1, ..., 4$,$j = 1, 2$	(positive uniform association)
H_9 :	$\theta_{ij} > \theta_{i,j+1}$,for $i = 1, ..., 4$	(positive column association)
H_{10} :	$\theta_{ij} > \theta_{i+1,j}$,for $i = 1, ..., 4, j = 1, 2$	(positive row association)
H_{11} :	$\theta_{ij} > \theta_{i,j+1}$,for $i = 1, ..., 4$ $\theta_{ij} > \theta_{i+1,j}$,for $i = 1, ..., 4, j = 1, 2$	(row and column association)

Table 6.27: Top panel: Row cumulative, Bottom panel: Row+Column Cumulative Probabilities for the Data in Table 6.25

	Happiness		
	1	1+2	1+2+3
Number of Siblings			
0-1	.36	.93	1
2-3	.35	.90	1
4-5	.36	.87	1
6-7	.28	.86	1
8+	.38	.82	1
Number of Siblings			
0-1	.36	.93	1
2-3	.36	.91	1
4-5	.36	.90	1
6-7	.34	.89	1
8+	.35	.88	1

An other approach to answer the research question is to use odds ratios instead of cumulative probabilities. Using odds ratios may provide a more flexible way to incorporate structure in the contingency table. The local odds ratio θ_{ij} is defined as $\pi_{ij} * \pi_{i+1,j+1}/\pi_{i+1,j} * \pi_{i,j+1}$. Odds ratios throughout the contingency table that are larger

than 1 indicate a positive association between the number of siblings and happiness. The observed odds ratios are displayed in Table 6.28, and it shows that four out of six odds ratios are smaller than 1. It is to be investigated whether this can be assumed to be due to sampling fluctuations. Hypothesis H_5 in Table 6.26 states that all the odds ratios are equal to 1, that is, no association between the number of siblings and happiness. This model is called the independence model. This hypothesis provides a reference for hypothesis H_6, where all odds ratios are larger than 1. If all odds ratios are equal, a parsimonious hypothesis for the data are the uniform association model as displayed in hypothesis H_7. This hypothesis states that the association in the contingency table can be described with a single odds ratio with value β. A naturally follow-up hypothesis is H_8, where the common odds ratio β is larger than 1.

Table 6.28: Odds ratios of the Data in Table 6.25

	Happiness	
	1-2	2-3
Number of Siblings		
1-2	.99	1.47
2-3	.91	1.36
3-4	1.49	.98
4-5	.56	1.66

Finally, various stochastic ordering hypotheses are discussed. If the positive association hypotheses are not supported by the data, one may resort to requiring the odds ratios to be ordered rather than requiring the odds ratios to be larger than 1. Hypothesis H_9 describes the structure where the odds ratios in the first column of Table 6.28 are smaller than in the second column, that is, the happier respondents are, the stronger the association to the number of siblings. A related hypothesis is H_{10}, that states that the more siblings a respondents have, the stronger the association to happiness. Hypothesis H_{11} investigates whether hypotheses H_9 and H_{10} simultaneously are supported by the data.

The results are displayed in Table 6.29. To prevent an information overload, the posterior model probabilities of each model and the unconstrained model are omitted. The posterior model probability for hypothesis H_1 and H_2 shows very little support for the hypothesis of ordered cumulative probabilities. If the hypothesis of ordered double cumulative probabilities is compared to the hypothesis of the equality of the double cumulative probabilities, it shows that the latter is a far better hypothesis than the former. When the best of hypotheses H_5 to H_8 is chosen, the posterior model probabilities show large support for the positive uniform association hypothesis (H_8). To satisfy curiosity, β, the common odds ratio is displayed, indicating a small positive association between the number of sibling and happiness. The best of the stochastic ordering hypotheses is the hypothesis of positive column association (H_9).

Overall, for the different types of hypotheses, the results seemed to indicate that there was no relation between the number of siblings and happiness, as seen from $P(H_1|H_1, H_2)$,$P(H_3|H_3, H_4)$ and the relative high $P(H_5|H_5, H_6, H_7, H_8)$. In general, an independence model parsimoniously describes the data, however, it is not specially

interesting. We showed that by carefully exploring different types of hypotheses, and contrasting various hypotheses against each other, a model can be found that shows more structure that the independence model and is highly supported by the data,

Table 6.29: Posterior Model Probabilities of the Hypotheses in Table 6.25

	Post. prob.	
$P(H_1\|H_1, H_2)$:	.987	
$P(H_3\|H_3, H_4)$:	1.000	
$P(H_5\|H_5, H_6, H_7, H_8)$:	.276	
$P(H_6\|H_5, H_6, H_7, H_8)$:	.000	
$P(H_7\|H_5, H_6, H_7, H_8)$:	.002	
$P(H_8\|H_5, H_6, H_7, H_8)$:	.723	$\beta = 1.075$
$P(H_9\|H_9, H_{10}, H_{11})$:	.630	
$P(H_{10}\|H_9, H_{10}, H_{11})$:	.051	
$P(H_1 1\|H_9, H_{10}, H_{11})$:	.318	

6.7 Discussion

In this paper, we showed an approach to hypotheses selection in a Bayesian context. The advantages may be clear: first, it is both a measure of evidence for the null-hypothesis and for the alternative hypothesis. Second, the interpretation is straightforward: it is the probability that the hypothesis is true, given the data and the set of hypotheses. Third, the use of the posterior model probability is not limited to two hypotheses. Fourth, the posterior model probability can both be used in inequality constrained testing and equality constrained testing. Fifth, the posterior model probability consists of the fit of the hypotheses and an implicit penalty for model complexity.

For inequality constrained hypotheses, there is no prior sensitivity, however, for equality constrained hypotheses the prior sensitivity may still pose a problem as choosing different values for $\boldsymbol{\alpha}$, results may be arbitrary posterior model probability. However, we showed that choosing a prior value of 1, the posterior model probability performs as may be expected, and the common intuition behind a prior value of 1 may provide results with face validity. Software can be downloaded via the web site of the publisher. A simple console program is provided, together with a comprehensive manual and the examples discussed in this paper. Questions regarding the software can be emailed to the first author.

References

Barlow R.E., Bartholomew D.J., Bremner J.M., Brunk H.D. *Statistical Inference Under Order Restrictions* Wiley: New York, NY, 1972

Bartholomew, K. (1990) Avoidance of intimacy: An attachment perspective. *Journal of Social and Personal Relationships,* **7**: 147–178.

Benson, P. L. (1990) *The troubled journey: A portrait of 6th–12th grade youth* Minneapolis, MN: Search Institute.

Blyth, D. A. (1993) *Healthy communities; healthy youth: How communities contribute to positive youth development* Minneapolis, MN: Search Institute.

Bowen, G.L., Richman, J.M., Brewster, A., Bowen, N.K. (1998) Sense of school coherence, perceptions of danger at school, and teacher support among youth at risk of school failure *Child and Adolescent Social Work Journal* **15**: 273–286

Chib, S. (1995) Marginal likelihood from the Gibbs output *Journal of the American Statistical Association* **90(432)**: 1313–1321

Gelfand, A.E. and Smith, A.F.M. (1990) Sampling-based approaches to calculating marginal densities *Journal of the American Statistical Association* **85(410)**: 398–409

Gelman, A. Carlin B.P., Stern H.S. and Rubin D.B. *Bayesian Data Analysis*. Chapman and Hall: New York, 1995.

Kass, R.E. (1993) Bayes factors in practice. *The Statistician* **42**: 551–560

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

Klugkist, I., Kato, B. and Hoijtink, H. (2005) Bayesian model selection using encompassing priors *Statistica Neerlandica* **59(1)**: 57–69

Klugkist, I., Laudy, O. and Hoijtink, H. (2005) Inequality Constrained Analysis Of Variance: A Bayesian Approach. *Psychological Methods* **10(4)**: 477–493.

Laudy, O. and Hoijtink, H. (to appear). Bayesian methods for the analysis of order restricted contingency tables. Statistical Methods in Medical Research.

Lessard J.C. and Moretti M.M. (1998) Suicidal ideation in an adolescent clinical sample: attachment patterns and clinical implications. *Journal of Adolescence* **21(4)**: 383-395

Matarazzo, A., Carrington, P.J. and Hiscott, R.D. (2001) The Effect of Prior Youth Court Dispositions on Current Disposition: An Application of Societal-Reaction Theory. *Journal of Quantitatie Criminology,* **17** (2) pp. 197–228

Nash,J.K. and Bowen, G.L. (2002) Defining and Estimating Risk and Protection: An Illustration from the School Success Profile Child and Adolescent *Social Work Journal* **19(3)**: 247–261

Perkins D.F. and Luster T. (1999) The relationship between sexual abuse and purging: findings from community-wide surveys of female adolescents *Child Abuse Neglect* **23(4)**: 371-382

Robertson T., Wright F.T., Dykstra R.L. *Order Restricted Statistical Inference* Wiley: New York, NY, 1988

Verdinelli, I. and Wasserman, L. (1995) Computing Bayes factors using a generalization of the Savage-Dickey density ratio *Journal of the American Statistical Association* **90(430)**: 614–618

Vermunt, J.K. (1991) A general class of nonparametric models for ordinal categorical data *Sociological methodology,* **29** pp. 187–223

6.8 Appendix: Bayes Factors for Equality Constrained Hypotheses

To illustrate the computation of Bayes factors for an equality constrained hypothesis, $H_1 : \pi_{11} = \pi_{12}$ and the unconstrained $H_0 : \pi_{11}, \pi_{12}$ are considered.

Note that straightforward application of (6.1) is not possible, because the numbers of samples from both the unconstrained prior and unconstrained posterior that are in agreement with the constraint of H_1 (that is, π_{11} and π_{12} are exactly equal) will always be zero. However, the Bayes factor (BF_{10}) can be approximated by a stepwise procedure, starting with a restriction that allows a (not too) small part of the parameter space (e.g. $H_{10} : |\pi_{11} - \pi_{12}| \leq 0.1$). In several subsequent steps the parameter space is more restricted in each step (see $H_{1,1} - H_{1,5}$ in Table 6.30). Note, that the set of new hypotheses is constructed such that $H_{1,5} \subset H_{1,4} \subset H_{1,3} \subset H_{1,2} \subset H_{1,1} \subset H_0$.

Table 6.30: Approximating Hypotheses for the Estimation of BF_{21}

H_0:	π_{11}, π_{12} (unconstrained)
$H_{1,1}$:	$\lvert\pi_{11} - \pi_{12}\rvert \leq 0.1$
$H_{1,2}$:	$\lvert\pi_{11} - \pi_{12}\rvert \leq 0.01$
$H_{1,3}$:	$\lvert\pi_{11} - \pi_{12}\rvert \leq 0.001$
$H_{1,4}$:	$\lvert\pi_{11} - \pi_{12})\rvert \leq 0.0001$
$H_{1,5}$:	$\lvert\pi_{11} - \pi_{12}\rvert \leq 0.00001$
H_1:	$\pi_{11} = \pi_{12}$

The Bayes factor for hypotheses $H_{1,1}$ and H_0 (denoted by $\mathrm{BF}_{(1,1)0}$) is given by (6.1), that is, samples from both the unconstrained prior and posterior provide the proportions of these samples that are in agreement with $H_{1,1}$. The number of samples from both prior and posterior in agreement with $H_{1,1}$ will not be zero (given that enough samples are drawn) and the procedure is rather efficient. However, $\mathrm{BF}_{(1,1)0}$ is not yet a sufficient approximation of BF_{10}.

In the next step, the Bayes factor for hypotheses $H_{1,2}$ and $H_{1,1}$ (denoted by $\mathrm{BF}_{(1,2)(1,1)}$) is obtained by sampling from both the constrained prior and posterior of hypothesis $H_{1,1}$, and calculating the proportions of samples that are in agreement with hypothesis $H_{1,2}$. This procedure is continued till the Bayes factor for hypotheses $H_{1,5}$ and $H_{1,4}$ is calculated.

The Bayes factor of interest, BF_{10} is estimated by $\mathrm{BF}_{(1,5)0}$, which is computed applying the product rule:

$$BF_{(1,5)0} = BF_{(1,5)(1,4)} \times BF_{(1,4)(1,3)} \times \cdots \times BF_{(1,1)0}$$

It can easily be checked if BF_{10} is accurately approximated by $\mathrm{BF}_{(1,5)0}$, that is, if the last interval is small enough: the Bayes factor $\mathrm{BF}_{(1,5)(1,4)}$ will be very close to one.

It is not trivial anymore to sample from the constrained prior and posterior distribution. However, the authors provide a simple software package that allows to calculate Bayes factors and the posterior model probability for arbitrary constraints and contingency tables of arbitrary dimensions.

Samenvatting

Inleiding

Het formuleren van theorieën is één van de hoofdactiviteiten in de sociale wetenschappen. Een onderzoeker heeft hypothesen of theorieën over het uit te voeren onderzoek op basis van eerder onderzoek en filosofische beschouwingen. Het opnieuw verzamelen van data heeft tot doel de hypothesen te verifiëren dan wel te falsificeren. Met betrekking tot de hypothesen zijn twee aspecten van belang: een hypothese moet de geobserveerde data accuraat beschrijven, en een hypothese moet niet meer elementen bevatten dan noodzakelijk, oftewel, deze moet de werkelijkheid spaarzaam beschrijven.

Beschouw ter illustratie een groep personen die een therapie volgt die hen helpt om te gaan met een spinnenfobie. Stel vervolgens dat de uitkomst van de therapie tweeledig is: onsuccesvol (een persoon heeft nog steeds last van zijn/haar spinnenfobie) of succesvol (een persoon wordt in zijn/haar dagelijkse leven niet meer gehinderd door de fobie). Een onderzoeker is geïnteresseerd in hoe het succes van de therapie samenhangt met leeftijd. Stel nu dat alle personen worden ingedeeld in drie leeftijdscategorieën: jong, middelbaar en oud. De kans op een succesvolle therapie wordt per leeftijdscategorie gegeven door het aantal personen met een succesvol resultaat gedeeld door het totaal aantal personen in die leeftijdscategorie. Dit is weergegeven in Tabel 6.31.

Table 6.31: Spinnenfobie Therapie

	Leeftijd		
	Jong	Middelbaar	Oud
Succeskans	π_1	π_2	π_3

Een onderzoeker heeft verschillende (en mogelijk tegenstrijdige) verwachtingen over deze data. De eerste verwachting is dat de succeskans niet afhangt van de leeftijd, of anders gezegd, de kans op succes is gelijk in elke leeftijdscategorie. Deze verwachting kan worden vertaald in de hypothese H_0 in Tabel 6.32, waar men kan zien dat de succeskansen gelijk zijn in elke leeftijdscategorie. Een andere mogelijke verwachting is dat jongere personen meer succesvol zijn in de therapie, bijvoorbeeld omdat zij meer flexibel zijn dan oudere mensen. Deze verwachting kan worden vertaald in de hypothese dat π_1 groter is dan π_2, en π_2 vervolgens groter is dan π_3. De hypothese is weergegeven als H_1 in Tabel 6.32. Een derde mogelijke hypothese is dat oudere personen een grotere succeskans hebben, bijvoorbeeld omdat zij makkelijker hun angst kunnen rationaliseren. De corresponderende hypothese is weergegeven als H_2 in Tabel 6.32. Een mix tussen

beide verwachtingen is dat de personen van middelbare leeftijd het meest succesvol zijn in de therapie, omdat zij nog steeds flexibel zijn maar ook makkelijker kunnen kunnen rationaliseren. Deze verwachting kan worden vertaald in de hypothese dat zowel jongere als oudere personen minder succesvol zijn in de therapie dan personen van middelbare leeftijd. Dit is weergegeven als hypothese H_3 in Tabel 6.32. Als laatste wordt hypothese H_4 gesteld. In deze hypothese wordt geen enkele verwachting gespecificeerd. Zonder enige structuur (een weinig spaarzame beschrijving van de werkelijkheid) is deze hypothese wetenschappelijk niet aantrekkelijk, maar het zou de beste hypothese kunnen zijn in het geval dat alle andere hypothesen niet waar blijken. Merk op dat alle hypothesen zorgvuldig overwogen moeten worden door de onderzoeker. Men zou op het idee kunnen komen alle mogelijke combinaties van restricties mee te nemen als hypothesen. Echter, dit zal niet leiden tot een groter inzicht, omdat na analyse van de data alsnog een interpretatie moet worden gemaakt van de beste hypothese. Dit leidt tot een meer exploratieve benadering: wat kunnen we vinden in de data? In dit proefschrift wordt voor het beantwoorden van deze vraag een confirmatieve benadering nagestreefd: welke theorieën die we vóór het analyseren van de data hadden, worden het meest ondersteund door de data?

Table 6.32: Hypothesen over de Uitkomsten van de Spinnenfobie Therapie

	Leeftijd				
Model	Jong		Middelbaar		Oud
H_0:	π_1	$=$	π_2	$=$	π_3
H_1:	π_1	$>$	π_2	$>$	π_3
H_2:	π_1	$<$	π_2	$<$	π_3
H_3:	π_1	$<$	π_2	$>$	π_3
H_4:	π_1		π_2		π_3

Ervan uitgaande dat dit experiment is uitgevoerd en de onderzoeker de beschikking heeft over de verzamelde data, zijn er twee vragen van belang: 1) welke hypothese beschrijft de data op een voldoende accurate en spaarzame manier? En 2) welke van de hypothesen beschrijft de data het beste? Met de huidige statistische technieken is het niet altijd mogelijk bovenstaande vragen te beantwoorden.

Gebruikmakend van de klassieke p-waarde is een onderzoeker in staat om de kans te berekenen op de geobserveerde data, of extremer gegeven, dat de hypothese H_0 waar is, afgezet tegen hypothese H_4. Als deze p-waarde klein is (doorgaans kleiner dan 0.05), dan wordt de nulhypothese (H_0) verworpen. Deze klassieke p-waarde staat echter niet toe de beste hypothese te selecteren uit een set van hypothesen, alhoewel dit wetenschappelijk wel een relevante vraag is. De kernvraag is immers welke van de door de onderzoeker geformuleerde hypothesen de beste combinatie verschaft van een accurate en spaarzame beschrijving van de geobserveerde data.

Naast de klassieke p-waarde wordt veel gebruik gemaakt van informatie-criteria om data te analyseren. Hierbij wordt iedere hypothese geassocieerd met een getal dat aangeeft hoeveel de hypothese wordt ondersteund door de data, gecorrigeerd voor het niet-spaarzaam zijn van de hypothese. De hypothese met het grootste

(of kleinste) getal wordt gekozen als beste hypothese. Echter, informatie-criteria hebben geen interpreteerbare schaal, dit is, is een verschil van 2 punten groot of klein? Zoals aangegeven, de meeste informatie criteria bestaan uit twee delen: een gedeelte dat aangeeft hoe goed de hypothese de data beschrijft, en een gedeelte dat aangeeft in hoeverre de hypothese een spaarzame beschrijving van de werkelijkheid geeft; de strafmaat. Deze strafmaat is een functie van het aantal parameters in een hypothese. Echter, wanneer gebruik wordt gemaakt van ongelijkheidrestricties, is dit aantal parameters onbekend. In dit proefschrift wordt gebruik gemaakt van deze posterior modelkans. Dit is een Bayesiaans informatie-criterium, en heeft een aantal aantrekkelijke eigenschappen boven de klassieke p-waarde. Ten eerste geeft de posterior modelkans aan hoeveel een hypothese wordt ondersteund door de data in de aanwezigheid van andere modellen op een 0-1 schaal, zodat de verschillen tussen modellen interpreteerbaar zijn. Tevens is het mogelijk te onderzoeken of een hypothese de data accuraat en spaarzaam beschrijft, en is het mogelijk de beste hypothese uit een set van hypothesen te kiezen. Verder is de posterior modelkans makkelijk te berekenen voor ongelijkheids gerestricteerde hypothesen. In voorliggend proefschrift wordt uitgebreid ingegaan op de bovengenoemde Bayesiaanse techniek.

Hoofdstuk 1 en 2

Hoofdstuk een van het proefschrift verschaft een inleiding op de Bayesiaanse technieken waarna in het tweede hoofdstuk wordt ingegaan op confirmatieve, ongelijkheids gerestricteerde latente klassen analyse. Latente klassen analyse is een techniek die kan worden gebruikt om een groep respondenten in te delen in homogene subgroepen. Deze subgroepen heten latente klassen, omdat vóór de analyse onbekend is welke respondent in welke klasse hoort. Meestel wordt latente klassen analyse gebruikt in een exploratieve zin: gegeven het aantal klassen wordt de data opgedeeld in de best mogelijke subgroepen. Na de analyse is het de taak van de onderzoeker elke latente klasse te interpreteren. Hierin schuilen twee gevaren. Ten eerste is er het gevaar van over-interpretatie. Gegeven de uitkomsten van een latente klassen analyse kan er altijd een redelijk verhaal worden gemaakt, echter dit hoeft helemaal niet overeen te stemmen met de werkelijkheid. Ten tweede zijn onderzoekers niet onbevooroordeeld voordat ze de analyse starten. Hun voorkennis bestaat uit een mix van resultaten van eerder onderzoek en hun eigen filosofische beschouwingen van de materie. In een confirmatieve analyse wordt rekening gehouden met beide zaken. Voordat de analyse wordt uitgevoerd moet een onderzoeker de theorieën met betrekking tot de data formaliseren. Deze formaliseringen worden gesteld in termen van (ongelijkheden van) de parameters van het model. Merk op dat een onderzoeker wordt aangemoedigd om tegenstrijdige theorieën te formuleren. De schattingsprocedure is zodanig dat de invloed van de theorieën wordt verdisconteerd in het betrefffende model. Nadat de analyse is uitgevoerd, kunnen informatiecriteria worden gebruikt om te onderzoeken welke theorie de data het beste en meest spaarzaam beschrijft, en hoe de modellen zich onderling vergelijken.

In hoofdstuk twee worden theorieën vertaald in ongelijkheids gerestricteerde latente klassen modellen. Posterior predictive checks, met gebruikmaking van de pseudo-likelihoodratio test, worden gebruikt om te kijken of de geobserveerde data goed kunnen

worden terugvoorspeld door de modellen. Vervolgens wordt de marginale likelihood gebruikt om het beste model te selecteren. Het voorafgaande wordt in hoofdstuk twee geïllustreerd aan de hand van twee voorbeelden. Samenvattend, in plaats van een exploratieve analyse uit te voeren, en na de analyse de parameterschattingen van het model te interpreteren, worden met deze confirmatieve benadering, de theorieën over de data meegenomen in de analyse, waarna een statistisch criterium aangeeft welk van de theorieën het meest ondersteund wordt door de data.

Hoofdstuk 3

In het derde hoofdstuk wordt de confirmatieve latente klassen analyse in meer technisch detail besproken. Ter illustratie worden data geanalyseerd afkomstig van de door Piaget ontwikkelde balanstaak. Er wordt getoond hoe met de Gibbs sampler een empirische representatie wordt verkregen van de posterior verdeling van het ongelijkheids gerestricteerde latente klassen model. Uit deze empirische representatie kunnen de parameters en de betrouwbaarheidsintervallen worden verkregen. Tevens wordt getoond hoe met een steekproef uit de posterior verdeling van het model een schatting kan worden gemaakt van de marginale likelihood en de posterior predictive checks, waarbij de posterior predictive checks gebruik maakt van de pseudo-likelihoodratio test als discrepantie maat.

Het is bekend dat de likelihoodratio test niet optimaal werkt in de context van latente klassen analyse omdat de data 'sparse' is.Dit houdt in dat er veel meer datapatronen mogelijk zijn dan de geobserveerde datapatronen. De pseudo-likelihoodratio test lost dit op door de likelihoodratio test te berekenen voor alle combinaties van twee variabelen in plaats van over het hele datapatroon. Er kan worden verwacht dat alle mogelijke twee-weg tabellen weinig tot geen lege cellen bevatten, waardoor een accurater beeld ontstaat. De verdeling van de pseudo-likelihoodratio test is niet bekend en als zodanig zal de referentieverdeling worden gesampled uit de posterior verdeling van het model. Dit is de methode van de posterior predictive checks. Deze methode wordt in latere hoofdstukken nader bestudeerd. In het voorbeeld wordt een mix van confirmatieve en exploratieve benaderingen getoond. Eerst worden de theorieën vertaald in een gerestricteerd latente klassen model, en nadat het beste model gegeven de data is gekozen, worden ongerestricteerde klassen toegevoegd. Dit verschaft de mogelijkheid te onderzoeken of naast de verwachte patronen in de data er nog meer patronen zijn terug te vinden welke nog niet door de theorie worden voorspeld.

Hoofdstuk 4

Het vierde hoofdstuk betreft modellen voor ongelijkheids gerestricteerde kruistabellen. Modellen voor kruistabellen waar interacties worden toegelaten zijn een andere manier om categorische data te analyseren. Posterior predictive checks, met gebruikmaking van de likelihoodratio test als discrepantie maat, zullen worden gebruikt om te onderzoeken of een model in staat is de geobserveerde data terug te voorspellen, en de marginale likelihood wordt gebruikt om te onderzoeken welk model het meest ondersteund wordt door de data. In modellen voor verzadigde kruistabellen kunnen ongelijkheidsrestricties worden gelegd op de celkansen of op functies van celkansen. Het

hoofdstuk wordt geïllustreerd met twee voorbeelden. In het eerste voorbeeld worden de ongelijkheidsrestricties geplaatst op de odds (de verhouding tussen twee celkansen) en in het tweede voorbeeld worden restricties geplaatst op de odds ratios (de verhouding van twee odds). De marginale likelihood van twee of meer modellen kan worden getransformeerd naar posterior modelkansen, zodanig dat kan worden gezien in welke mate elk van de modellen ondersteund wordt door de data, gegeven alle andere modellen in de set. Dit heeft een voordeel boven het gebruik van de klassieke p-waarden. Los geformuleerd is de klassieke p-waarde een maat die aangeeft hoeveel bewijs er is tegen de nulhypothese, terwijl de posterior modelkans aangeeft wat het bewijs is voor zowel de nul- als de alternatieve hypothese. Bovendien is het gebruik van de posterior modelkans niet beperkt tot het vergelijken van twee modellen. Echter, het gebruik van de posterior modelkans heeft ook een prijs: er moet een prior worden gespecificeerd. Het is bekend dat de posterior modelkans afhangt van de specificaties van de prior, echter voor ongelijkheids gerestricteerde modellen is de prior sensitiviteit niet groot. Dit wordt geïllustreerd aan de hand van een kleine simulatiestudie.

Hoofdstuk 5

Het vijfde hoofdstuk onderzoekt de prestaties van verschillende Bayesiaanse informatiematen in de context van ongelijkheids gerestricteerde modellen voor kruistabellen. In de voorafgaande hoofdstukken is gebruik gemaakt van de marginale likelihood en de posterior predictive checks, echter de frequentie-eigenschappen van deze maten zijn nog niet volledig bekend. Drie situaties worden bekeken: ten eerste wordt een correct (in de zin van een goed passend) ongelijkheids gerestricteerd model vergeleken met een ongerestricteerd model. Ten tweede wordt een incorrect ongelijkheids gerestricteerd model vergeleken met het ongerestricteerde model, en ten derde wordt een correct ongelijkheids gerestricteerd model vergeleken met een incorrect ongelijkheids gerestricteerd model. Voor deze vergelijkingen worden simulatiestudies uitgevoerd voor drie populaties. In de eerste simulatie wordt de steekproefgrootte gevarieerd, in de tweede simulatie wordt de grootte van het effect gevarieerd en in de derde simulatie worden modellen vergeleken met verschillende a-priori likelihoods. De fit-en informatiematen kunnen worden onderverdeeld naar prior predictieve maten en posterior predictieve maten. De resultaten laten zien dat de marginale likelihood over alle situaties heen stabiel en bevredigend presteert. De eerder gebruikte posterior predictive checks blijken niet goed te presteren. Een interessant resultaat is dat de prior predictieve maten (zoals de marginale likelihood) meer sensitief zijn voor de modelcomplexiteit in ongelijkheids gerestricteerde modellen.

Hoofdstuk 6

In het zesde en laatste hoofdstuk wordt de toepasbaarheid van de verzamelde kennis uit de voorgaande hoofdstukken voor de sociale wetenschappen gedemonstreerd. Data uit recent gepubliceerde literatuur wordt opnieuw geanalyseerd met modellen voor kruistabellen waarbij wordt ingegaan op zowel ongelijkheids- als gelijkheidsrestricties. Uit de analyses in het vijfde hoofdstuk bleek dat de marginale likelihood het beste presteerde in vergelijking met andere maten, het berekenen van de marginale likelihood

is in veel gevallen echter erg lastig. In het laatste hoofdstuk wordt getoond dat de posterior modelkans op een simpelere manier kan worden berekend dan via de marginale likelihood. Deze methode is in eerste instantie ontwikkeld voor de berekening van posterior modelkansen voor ongelijkheids gerestricteerde modellen, maar wordt zodanig aangepast dat ook de posterior modelkans voor gelijkheids gerestricteerde modellen kan worden berekend. Door de afleiding van de computationele methode kan bovendien worden getoond dat de posterior modelkans bestaat uit een term die de fit van een model weergeeft en een term die de strafmaat voor modelcomplexiteit weergeeft.

Verder zullen in hoofdstuk zes voor verschillende dataformaten gedetailleerde simulaties worden uitgevoerd. Een eerste simulatie onderzoekt de grootte van de posterior modelkans bij verschillende steekproefgrootten en verschillende effectgrootten. Een tweede simulatie onderzoekt het equivalent van de type 1 fout van de posterior modelkans. De laatste simulatie onderzoekt de sensitiviteit voor de prior. De resultaten laten zien dat ongelijkheids gerestricteerde hypothesen ongevoelig zijn voor de keuze van de prior, terwijl de gelijkheids gerestricteerde hypothesen wel sensitief zijn, echter in sommige situaties meer dan in andere. De conclusie van de prior sensitiviteit is dat een parameter waarde van één voor de prior tot goed interpreteerbare uitkomsten leidt. Bovendien heeft de waarde één voor de parameter van de prior de aantrekkelijke eigenschap dat a-priori elke celkans even waarschijnlijk is. Het hoofdstuk wordt afgesloten met drie meer geavanceerde analyses die de brede toepasbaarheid van gerestricteerde modellen en de posterior modelkansen tonen in de sociale wetenschappen.

Conclusies

Voorliggend proefschrift behandelt de vertaling van theorieën in ongelijkheids gerestricteerde statistische modellen. Vervolgens kan worden onderzocht welk van de modellen het meest wordt ondersteund door de data. Verscheidene resultaten zijn geboekt: Ten eerste, door het vertalen van theorieën in verschillende modellen is getoond hoe deze toepassing nuttig kan zijn voor de sociale wetenschappen. Ten tweede zijn computationele methoden ontwikkeld en verbeterd en ten derde, door simulatie is kennis verkregen van de daadwerkelijke prestatie van verschillende informatie-criteria.

Curriculum Vitae

Olav Laudy was born on October 27, 1975 at 21.05 in Alkmaar, The Netherlands. In 1994, he completed pre-university education (VWO) at Bernardus Alfrink College in Schagen. He studied Psychology at the University of Utrecht from 1997, until graduation in 2001. During the last two years of that period he worked as a student assistant at the Department of Methodology and Statistics at the Faculty of Social Sciences at Utrecht University, where in 2001 he started as a PhD-student for four days a week and as a junior researcher for one day a week.

Scientific articles

Laudy, O. and Hoijtink, H. (2005) *Bayesian Computational Methods for Inequality Constrained Latent Class Analysis* In: New Development in Categorical Data Analysis for the Social and Behavioral Sciences (eds: Van der Ark, A, Croom, M.A. and Sijtsma, K) Erlbaum, Londen, 2005

Laudy, O. and Hoijtink, H. (2005) Applications of Confirmatory Latent Class Analysis in Developmental Psychology. *European Journal of Developmental Psycholgy* **2(1)**: 1–15

Laudy, O. and Hoijtink, H. (2006) Bayesian Computational Methods for the Analysis of Inequality Constrained Contingency Tables *Statistical Methods in Medical Research* to appear in 2006

Klugkist, K., Laudy, O. and Hoijtink, H. (2005) Inequality Constrained Analysis of Variance: A Bayesian Approach (with discussion) *Psychological Methods* **10**: 477–493.

Laudy, O. and Hoijtink, H. (submitted to Bayesian Analysis) Evaluation of Bayesian Model Selection Criteria in the Context of Inequality Constrained Contingency Tables

Laudy, O., Klugkist, I and Hoijtink, H. (submitted to Psychological Methods) Bayesian Selection of Equality and Inequality Hypotheses in Contingency Tables.

Klugkist, K., Laudy, O. and Hoijtink, H. (2005) Bayesian Eggs and Omelettes *Psychological Methods* **10**: 500–503.

Lensvelt-Mulders, G., Van der Heijden P. and Laudy O. (2005) A validation of a Computer Assisted Randomized Response Survey to Estimate the Prevalence of Fraud in Social Security. *Journal of the Royal Statistical Society. Series A* **169(2)**: 305–318

Laudy, O. and Hoijtink, H. (unpublished manuscript) ANOVA with a Categorical Dependent Variable.

Applied research articles

Laudy, O. (2001) Daderprofielen van Veroordeelde Delinquenten. Openbaar Ministerie.

Breeman, L. and Laudy, O. (2002) Genderanalyse Medewerkersmonitor. Afdeling Universitair Strategisch Programma, Universiteit Utrecht.

Menenti, L. and Laudy, O. (2002) Vragenlijst Analyse Medewerkersmonitor. Afdeling Universitair Strategisch Programma, Universiteit Utrecht.

Breeman, L. and Laudy, O. (2003) Answertree Analyses Medewerkers Monitor: Vastgelopen in het Werk & Wantrouwen in de Leiding. Afdeling Universitair Strategisch Programma, Universiteit Utrecht.

Gils, G. van, Heijden, P.G.M. van der, Laudy, O. and Ross, R. (2003). Regelovertreding in de Sociale Zekerheid. De Tweede Meting van het Periodiek Onderzoek naar Regelovetreding in de Sociale Zekerheidsregelingen WAO, WW en Abw. Doetichem: Reed Business Information

Gils, G. van, Heijden, P.G.M. van der, and Laudy, O. (2003). Fraude in Particuliere Inboedelverzekeringen. Utrecht: BOA Utrecht.

Gils, G. van, Heijden, P.G.M. van der, and Laudy, O. (2004). Beleving en Overtreding van Regels van de Huursubsidiewet

Gils, G. van, Heijden, P.G.M. van der, and Laudy, O. (2005). Regelovertreding in de Volksverzekeringen. Een Onderzoek naar Regelovertreding in de Volksverzekeringen AOW, AKW en ANW. Den Haag: Ministerie van Sociale Zaken en Werkgelegenheid.

Heijden, P.G.M. van der, Gils, G. van, and Laudy, O. (2005). Regelovertreding in de WAO, WW en Abw (vergeleken met de jaren 2000 en 2002). Ministerie van Sociale Zaken en Werkgelegenheid.

Breeman, L. and Laudy, O. (2005) Betrouwbaarheidsanalyse 360-graden Feedback. Afdeling Universitair Strategisch Programma, Universiteit Utrecht.

Index