ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

Bayesian Analysis and Model Selection for Contingency Tables using Power Priors

By

Katerina Mantzouni

A THESIS

Submitted to the Department of Statistics

of the Athens University of Economics and Business

in partial fulfilment of the requirements for

the degree of

Doctor of Philosophy in Statistics

March 2022
Athens, Greece

ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΔΙΔΑΚΤΟΡΙΚΟ ΠΡΟΓΡΑΜΜΑ

Μπεϋζιανή Ανάλυση και Επιλογή του κατάλληλου Μοντέλου σε Πίνακες Συνάφειας χρησιμοποιώντας Εκ-των-προτέρων Κατανομές Δύναμης

Κατερίνα Μαντζούνη

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής

του Οικονομικού Πανεπιστημίου Αθηνών

ως μέρος των απαιτήσεων για την απόκτηση

Διπλώματος Διδακτορικών Σπουδών στη Στατιστική

Αθήνα
Μάρτιος 2022

Contents

List of Tables

List of Figures

This thesis is dedicated to
my wonderful parents who have raised me to be the person I am today
to my fiancé Dimitris who believes in the richness of learning
and to my beloved grandparents.

Acknowledgments

Throughout the writing of this dissertation, I have received a great deal of support and assistance. First and foremost, I am extremely grateful to my supervisor, Prof. Ioannis Ntzoufras for his invaluable advice, continuous support, and patience during my PhD study. His immense knowledge and plentiful experience have encouraged me during the tdaily strains of my academic research. His unwavering enthusiasm for Bayesian Statistics kept me constantly engaged with my research, and his personal generosity enriched the joy of my time spent as a PhD candidate at the Athens University of Economics and Business (AUEB). I could not imagine a more ideal advisor and mentor for my Ph.D. study without whose help this dissertation would not have been possible.

I owe a deep sense of gratitude to Prof. Dimitris Karlis for his treasured support and for his keen interest at every stage of my research. I would also like to thank Prof. Maria Kateri, who introduce me to the Categorical Data world. My visits to RWTH University in Aachen, Germany gave me the opportunity to learn from one of the leading scientists on association models, which are the main subject of my research. She offered me valuable suggestions and useful comments throughout my research.

I would also like to thank Prof. Claudia Tarantola for her mentorship. I have spent a year at the University of Pavia in Italy, working with her on extending my research on graphical models. She was always there for me to support, encourage, guide, hospitalize and make me love Italy even more. Thus, Prof. Taratola gave me the opportunity to meet Prof. Alan Agresti, a leading figure in the development of

categorical analysis. It was my honor to discuss with him and be advised on my PhD research.

My gratitude extends to the Faculty of Statistics for the funding opportunity to undertake my studies at the Department of Statistics in Athens University of Economics and Business. They gave me the opportunity to participate in three programs: Erasmus+, IKYDA and the Funding program for the research publications of the AUEB Faculty members, run by the Research Center of the Athens University of Economics and Business. My appreciation also extends to my laboratory colleagues.

Last but not the least, I would like to thank my family: my parents, my sister and my nephews for supporting me spiritually throughout my life and particularly during writing my thesis. Especially, I would like to thank my parents, whose love and guidance are with me in whatever I pursue. My deepest gratitude to my loving and supportive fiancée, Dimitris. His encouragement when times got rough are much appreciated and duly noted.

Abstract

Katerina Mantzouni

Bayesian Analysis and Model Selection for Contingency Tables using Power Priors

March, 2022

In this dissertation, a comprehensive Bayesian model comparison approach is proposed for association models in contingency tables. The proposed methodology deals with the suitable specification of the prior distributions, as well as the allied computational issues regarding the estimation of the Bayesian evidence, which is the core component of the posterior model probabilities in Bayesian model comparison, selection and averaging. More specifically, the choice of the prior distribution in Bayesian model comparison and testing is problematic due to the well-known sensitivity of the posterior model odds and the associated Barlett-Lindley paradox (Bartlett, 1957, Lindley, 1957). This fact has led to the development of objective Bayes techniques which refers to the use of reasonably low information priors when no actual prior information is available. Within this framework, the utilization

of the power prior approach is proposed. In order to implement the method, a set of imaginary data from the most parsimonious model is produced satisfying by this way the locality criterion of Bayesian model comparison theory, Bayarri et al. (2012). Then, the prior distribution can be obtained by the product of the likelihood of the model under consideration evaluated at some historical data raised to a power and then multiplied by a pre-prior distribution, Ibrahim and Chen (2000). Here we extend and adapt this method by using instead of historical data imaginary data supporting the null model or hypothesis and we consider a relatively flat pre-prior.

Evaluation of the models under consideration and the related Bayesian tests are obtained by using MCMC based marginal likelihood estimators. We introduce and examine two versions of the importance sampling estimator of Perrakis et al. (2014): the independent and the one-block estimator. The results are compared with two versions of the Laplace approximation: the original one and an MCMC based approximation Lewis and Raftery (1997a). Results have shown that the one-block importance estimator works fast and efficiently even in sparse contingency tables when competitors may fail. We illustrate and compare the proposed methodology in two real data sets (one sparse and one with full cell frequencies) and by using an extended simulation study. In the simulation study, we further examine the model selection consistency of the proposed power-prior based methodology.

Finally, a comprehensive Bayesian analysis expansion is illustrated for graphical models of conditional independence for three way contingency tables using the power prior setup based on the approach of Tarantola and Ntzoufras (2012). More specific, extending the proposed methodology from marginal to conditional independence, involving suitable choices of prior parameters, estimation, model determination, as well as the allied computational issues. Each conditional independence model corresponds to a particular factorization of the cell probabilities and a conjugate analysis based on a Dirichlet prior performed. Unit information interpretation priors are used as a yardstick in order to identify and interpret the effect of any other prior distribution used. The posterior distributions of the graphical models parameters, are obtained using simple Markov chain Monte Carlo (MCMC) schemes.

This dissertation offers an innovative analytical and methodological approach in Bayesian model selection and comparison in categorical data. The first contribution of this thesis is the prior construction using imaginary data and the power prior approach in order to obtain an objective Bayes model comparison approach. The second contribution is the implementation and adaptation of two versions of Perrakis Monte Carlo estimator for obtaining the marginal likelihood in contingency tables. The proposed Monte Carlo estimators are simple to implement and efficient in all examples and simulations illustrated in this thesis. Both estimators can be further used for other practical problems and contexts. A further but secondary contribution is a fact that we identify the problem of zero counts in sparse contingency tables and their effect on the estimation of the marginal likelihood. Here, we propose a way to alleviate this problem, this is an initial study and needs further treatment. Finally, we develop and study a similar prior approach based on the power prior and imaginary data for graphical models in three way contingency tables using conjugate analysis. Future research may include the Bayesian comparison and combination of the two different groups of models (the association and graphical models) consider here and their implementation (using the proposed methods of this thesis) in contingency tables of higher dimension.

The structure of this thesis is the following: Chapter 1 introduces the main ideas of categorical data and Bayes. Chapter 2 illustrates the main Bayesian methodology for model comparison for association models. Chapter 3 provides the implementation in real data and extensive simulation study. Chapter 4 illustrates an Bayesian analysis for graphical models of conditional independence using conjugate analysis. Chapter 5 is a sort discussion and conclusions of this thesis.

ΠΕΡΙΛΗΨΗ

Κατερίνα Μαντζούνη

Μπεϋζιανή Ανάλυση και Επιλογή του κατάλληλου Μοντέλου σε Πίνακες Συνάφειας χρησιμοποιώντας Εκ-των-προτέρων Κατανομές Δύναμης

Μάρτιος, 2022

Κεντρικός πυλώνας της παρούσας διδακτορικής διατριβής είναι η ανάπτυξη προτεινόμενης μεθοδολογίας για τη Μπεϋζιανή ανάλυση κατηγορικών μεταβλητών σε πίνακες συνάφειας με σκοπό την επιλογή του καταλληλότερου μοντέλου. Η Μπεϋζιανή προσέγγιση εφαρμόστηκε τόσο σε μοντέλα συνάφειας (association models), όσο και σε γραφικά μοντέλα (graphical models) σε πίνακες συνάφειας διπλής και τριπλής εισόδου, αντίστοιχα. Η προτεινόμενη μεθοδολογία περιλαμβάνει τον καθορισμό κατάλληλων εκ-των-προτέρων κατανομών, καθώς επίσης και υπολογιστικές τεχνικές για την εκτίμηση Μπεϋζιανών περιθώριων πιθανοφανειών, οι οποίες είναι απαραίτητες

για τον υπολογισμό των εκ-των-υστέρων κατανομών στην Μπεϋζιανή σύγκριση και επιλογή του καταλληλότερου μοντέλου. Πιο συγκεκριμένα, η επιλογή κατάλληλης εκ-των-προτέρων κατανομής στη Μπεϋζιανή σύγκριση μοντέλων και των σχετικών ελέγχων είναι πολλές φορές προβληματική λόγω του γνωστού προβλήματος ευαισθησίας των εκ των υστέρων πιθανοτήτων και του παραδόξου των Barlett-Lindley, (Bartlett, 1957, Lindley, 1957). Το γεγονός αυτό οδήγησε στην ανάπτυξη αντικειμενικών Μπεϋζιανών τεχνικών, οι οποίες προτείνουν τη χρήση μη πληροφοριακών εκ-των-προτέρων κατανονών, όταν δεν υπάρχει κάμια εκ-των-προτέρων πληροφορία για τα δεδομενα. Σε αυτο το πλαίσιο προτείνονται οι εκ-των-προτέρων κατανομές δύναμης. Για την εφαρμογή της προτεινόμενης μεθοδολογίας σε πίνακες συνάφειας, που στόχο έχει την επιλογή του καταλληλότερου μοντέλου συνάφειας, κατασκευάστηκαν δύο σενάρια εκ-των-προτέρων κατανομών με τη χρήση πλασματικών δεδομένων, τα οποία βασίστηκαν στις εκ-των-προτέρων κατανομές δύναμης (Power priors).

Για την αξιολόγηση των υπό εξέταση μοντέλων και Μπεϋζιανών ελέγχων και για τον υπολογισμό της περιθώριας κατανομής χρησιμοποιήθηκαν Monte Carlo εκτιμητές που βασίζονται σε αποτελέσματα MCMC τεχνικών. Εισάγουμε και εξετάζουμε δύο προτεινόμενους εκτιμητές, οι οποίοι βασίζονται στον εκτιμητή που προτείνεται από τον Perrakis et al. (2014). Πραγματοποιήθηκε σύκριση των αποτελεσμάτων με δύο εκδοχές της προσέγγισης κατά Laplace, την απλή και μία όπου συγκεκριμένες ποσότητες του εκτιούνται μέσω MCMC. Τα αποτελέσματα έδειξαν ότι ο εκτίμητής που βασίζεται στην μεθοδολογία του Περράκη, λειτουργεί γρήγορα και αποτελεσματικά ακόμα και σε αραιούς (sparse) πίνακες συνάφειας, όπου οι περισσότεροι εκτιμητές αποτυγχάνουν. Όλες οι τεχνικές εφαρμόστηκαν και ελέγχθηκαν σε πραγματικά δεδομένα αλλά και σε αναλυτικές μελέτες προσομοίωσης. Για να ελεγχθεί η εγκυρότητα της προτεινόμενης μεθοδολογίας χρησιμοποιήθηκαν κριτήρια αντικειμενικών μεθόδων Bayes, όπως συνέπεια επιλογής μοντέλων, συνέπεια πληροφορίας και το κριτήριο της αντιστοίχισης προβλεπτικών κατανομών.

Τέλος, παρουσιάζεται η επέκταση της μεθοδολογίας στη χρήση μεθόδων Μπεϋζιανής ανάλυσης γραφικών μοντέλων σε πίνακες συνάφειας τριπλής εισόδου χρησιμοποιώντας εκ-των-προτέρων κατανομές δύναμης με έμφαση στην προσέγγιση των Tarantola and Ntzoufras (2012). Πιο συγκεκριμένα, η μέθοδος επεκτάθηκε από περιθώρια ανεξαρτησία σε ανεξαρτησία υπό συνθήκη για τρείς κατηγορικές μεταβλητές, και περιλμβάνει κατάλληλες επιλογές εκ-των-προτέρων κατανομών, εκτίμηση και προσδιορισμός του μοντέλου κσι συναφή υπολογιστικά ζητήματα. Σε κάθε μοντέλο ανεξαρτησίας υπό συνθήκη αντιστοιχείται μια συγκεκριμένη παραγοντοποίηση των πιθανοτήτων των κελιών και εφαρμόζεται συζυγής ανάλυση, βασιζόμενη σε Dirichlet εκ-των-προτέρων κατανομές. Εκ-των-προτέρων κατανομές μοναδιαίας ερμηνευτικής πληροφορίας χρησιμοποιούνται σαν μέτρο σύγκρισης με στόχο να ελεγχθεί και να ερμηνευθεί η επίδραση οποιονδήποτε εκ-των-προτέρων κατανομών. Για να υπολογιστούν οι εκ-των-υστέρων κατανομές των παραμέτρων των γραφικών μοντέλων χρησιμοποιήθηκαν MCMC μέθοδοι. Η προτεινόμενη μεθοδολογία εφαρμόστηκε σε πραγματικά αλλά και σε προσομοιωμένα δεδομένα.

Αυτή η διατριβή προσφέρει μια καινοτόμο αναλυτική και μεθοδολογική προσέγγιση για την Μπεϋζιανή επιλογή και σύγκριση μοντέλων σε κατηγορικά δοδομένα. Η πρώτη συβολή της παρούσας διδακτορικής διατριβής είναι η κατασκευή εκ-των-προτέρων κατανομών, για τις οποίες χρησιμοποιήθηκαν πλασματικά δεδομένα και η προσέγγιση της εκ-των-προτέρων κατανομής δύναμης, με σκοπό την εξασφάλιση μιας αντικειμενικής Μπεϋζιανής (Objective Bayes) μεθοδολογίας σύγκρισης μοντέλων. Η δεύτερη συνεισφορά είναι η εφαρμογή και προσαρμογή δύο εκδοχών του Monte Carlo εκτιμητή του Περράκη για την εκτίμηση και τον υπολογισμό της περιθώριας κατανομής σε πίνακες συνάφειας. Οι προτεινόμενοι Monte Carlo εκτιμητές είναι απλοί στην εφαρμογή τους και αποτελεσματικοί σε όλα τα παραδείγματα που εφαρμόστηκαν όπως και στην αναλυτική προσομοίωση που πραγματοποιήθηκε στα πλαίσια αυτού του διδακτορικού. Και οι δύο εκτιμητές μπορούν να χρησιμοποιηθούν και σε άλλα πρακτικά προβλήματα και γενικότερα πλαίσια. Μια επιπλέον, αλλά δευτερεύουσα συνεισφορά, είναι το γεγονός ότι αναγνωρίσαμε και ταυτοποιήσαμε το πρόβλημα των μηδενικών κελιών σε αραιούς πίνακες συνάφειας και την επίδρασή που έχουν στην έκτήμηση της περιθώριας κατανομής. Εδώ προτείνουμε τρόπους για την εξομάλυνση του προβλήματος αυτού, είναι μια αρχική μελέτη και χρειάζεται περαιτέρω και εις βάθος αντιμετώπιση. Τέλος, αναπτύξαμε και μελετήσαμε μια παρόμοια προσέγγιση εκ-των-προτέρων κατανομής βασιζόμενοι στη χρήση εκ-των-προτέρων κατανομών δύναμης και πλασματικών δεδομένων σε γραφικά μοντέλα για πίνακες συνάφειας τριπλής εισόδου χρησιμοποιώνας συζυγή ανάλυση. Σε μελλοντική έρευνα θα μπορούσε να υπάρχει η σύγκριση και ο συνδιασμός των δύο διαφορετικών γκρουπ μοντέλων (μοντέλα συνάφειας και γραφικών μοντέλων) που μελετήθηκαν εδώ καθώς και η εφαρμογή της προτεινόμενης μεθοδολογίας αυτού του διδακτορικού σε πίνακες μεγαλύτερων διαστάσεων.

Η δομή της παρούσας διδακτορικής διατριβής είναι η ακόλουθη: Κεφάλαιο 1 εισάγει τις βασικές έννοιες των κατηγορικών δεδομένων και του Μπέυζ. Το Κεφάλαιο 2 παρουσιάζει την κύρια Μπεϋζιανή μεθοδολογία σύγκρισης μοντέλων εφαρμοσμένη σε μοντέλα συνάφειας. Το Κεφάλαιο 3 παρέχει τα αποτελέσματα της εφαρμογής της προτεινόμενης μεθοδολογίας σε πραγματικά δεδομένα καθώς και τα αποτελέσματα της αναλυτικής μελέτης προσομοίωσης. Το Κεφάλαιο 4 παρουσιάζει μια Μπεϋζιανή ανάλυση σε γραφικά μοντέλα υπό συνθήκης ανεξαρτησίας χρησιμοποιώντας συζυγή ανάλυση. Το Κεφάλαιο 5 αποτελείται από μια σύντομη συζήτηση και τα συμπεράσματα της παρούσας διδακτορικής διατριβής.

Sponsorships and Grants

This research was funded by the Research Center of the Athens University of Economics and Business (Funding program for the research publications of the AUEB Faculty members), by the exchange program IKYDA 2014 of Greek State Scholarships Foundation (IKY) and the German Academic Exchange Service (DAAD) and by the exchange programme Erasmus+ 2017-2018 .

Chapter 1

Bayesian Inference and Hypothesis Testing for Contingency Tables

> Destiny is variable, not fixed; it is forever changing depending upon your free will to make choices for what you want your life to be
>
> *Steven Redhead,*
> *The Solution*

1.1 Introduction and Historical Review to the Analysis of Categorical Data

A contingency table is essentially a way to display cross-classification of two or more categorical variables. Contingency tables are simple frequency tabulations, which present the frequencies for each combination of the levels (or categories) of all variables under consideration.

The literature on categorical data analysis was introduced at the early years of the twentieth century with the emblematic work of Karl Pearson and George Udny Yule. However, Stigler (2002) denoted that the first appearance of 2×2

and the fourfold table dated back to Aristotle era. In the 19th century the work of Quetelet (1849) on measuring association, the hypergeometric analysis for the 2×2 table by Bienaymé (see Heyde and Seneta, 1977), and the introduction of expected values by Galton (1892) formed by the well-known (nowadays) equation:

$$Expected\ Count\ (i,j) = \frac{(Row\ Marginal\ Total\ i) \times (Column\ Marginal\ Total\ j)}{Grand\ Total}. \tag{1.1}$$

Pearson (1900) introduced the chi-squared statistic χ^2; the motivation behind this research was his curiosity to test if the outcome of the roulette wheel in Monte Carlo was random. George Udny Yule first introduced the cross-product ratio (or odds ratio) as a formal statistical tool. He also introduced other related measures of association. In 1904, Pearson introduced the term contingency as a measure of the total deviation of the classification from independence. Yule (1903) showed the potential discrepancy between marginal and conditional associations in contingency tables. This was later studied by Simpson (1951) and he introduced the well-known Simpson's paradox. Fisher (1922) corrected the degrees of freedom originally falsely introduced by Pearson for tests of independence in $I \times J$ tables χ^2 has $df = (I-1)(J-1)$. This finding came in conflict with the work of Pearson (1900, 1904), who claimed that the degrees of freedom were $df = IJ - 1$ for two-way tables. Fisher realized that the degrees of freedom must be adjusted when the cell counts have linear constraints. Later, in 1934, Fisher introduced the Fisher's exact test for 2×2 contingency tables in the fifth edition of Statistical Methods for Research Workers.

The definition for homogeneous association (no interaction) in contingency tables originated in an article by the British statistician Maurice Bartlett (1935) about $2 \times 2 \times 2$ tables. Bartlett showed how to find ML estimates of cell probabilities satisfying the property of equality of odds ratios between two variables at each level of the third. Wilks (1935) introduced the likelihood ratio test as an alternative to

Pearson's chi-square statistic.

Bartlett (1937) used the expression $log\left(\frac{y}{1-y}\right)$ as a response in regression and ANOVA in order to transform observations y that are continuous proportions. Fisher and Yates. (1938) suggested it as a possible transformation of a binomial parameter for analyzing binary data. In 1940, Fisher developed canonical correlation methods for contingency tables. He showed how to assign scores to rows and columns of a contingency table in order to maximize the correlation. Deming and Stephan (1940) introduced the iterative proportional fitting (ITF) method for estimating the cell values in a contingency table subject under constraints coming from known marginal totals, e.g., from a population data set, minimising a least squares criterion called raking.

Berkson (1944) introduced the term logit for this transformation and popularize the logistic regression. Cornfield (1951) used the odds ratio to approximate relative risks in case-control studies. Neyman (1949) introduce a family of best asympotically normal (BAN) estimators, by minimizing chi-squared-type measures comparing observed proportions to proportions predicted by the model. Cochran (1940, 1943, 1950) modeled Poisson and binomial responses with variance-stabilizing transformations, introduced the term of overdispersion and generalized McNemar's test for comparing proportions in several matched samples, respectively. Cochran (1954) proposed the sample size for the chi-square approximation, the directing inference toward narrow alternatives, the partition of χ^2 statistic into components, the Cochran's test of conditional independence in 2×2 tables, the utilisation of ordered categories in $I \times 2$ contingency table and a trend statistic for testing independence by partitioning the Pearson statistic for that hypothesis using a linear probability model. Roy and Mitra (1956) discussed types of independence for three-way tables and their large-sample tests. They derived asymptotic chi-square tests for these different situations, using the union-intersection principle that Roy had developed in his earlier work on multivariate analysis. Additionally, they showed that the "equivalent" hypotheses/designs have the same maximum likelihood estimates and chi-square goodness-of-fit tests. Mantel and Haenszel (1959) proposed a non-model-based test of the null hypothesis for conditional independence in $2 \times 2 \times K$ tables using the response (column) marginal totals as fixed.

Birch's never-submitted Ph.D. thesis at the University of Glasgow was dealing with a variety of important topics in loglinear models. He attained ML estimates of cell probabilities in three-way tables for Poisson and multinomial sampling and extended theoretical results of Cramér and Rao on large-sample distributions for contingency table models (Birch, 1963, 1964a,b, 1965). Caussinus (1965) introduced the quasi-symmetry model for square tables. Bishop (1967, 1969) used Birch's results to derive connections between log-linear models and logit models. She also proposed using a version of the iterative proportional fitting method developed by Deming and Stephan (1940) to perform computations for the MLE, as a practical way to implement the ideas of Birch to higher dimensional tables. She simplify the IPF calculations by multiplicative adjustments to the estimates for marginal tables— an idea related to models with direct multiplicative estimates such as conditional independence.

The book of Goodman and Kruskal (1979) is the most classical reference on association measures, they focused on the $I \times J$ and developed new measures for nominal and ordinal variables. Goodman (1970, 1971) presented methods for analyzing n-way tables using log-linear models and likelihood ratio statistics. In particular, he considered the class of hierarchical log-linear models in which the cell mean vector is expressible in closed form as a rational function of the sufficient statistics. For such models we can compute the MLE directly without resorting to any iterative numerical procedure. Goodman emphasized how these models are interpretable in terms of probability concepts such as independence, conditional independence and equiprobability.

Darroch et al. (1980) introduced the graph theory and Markov properties for modelling interactions of log-linear models for contingency tables. More details about the leading figures in the development of categorical data analysis can be found in Agresti (2002).

1.2 Bayesian development of categorical data analysis

In Bayesian perspective, methods for categorical data analysis in contingency table form have as a starting point the original work of Bayes (1763) and Laplace (1774), where they use a uniform prior in order to estimate the binomial parameter. Good (1965, 1953, 1956) proposed a uniform prior distribution over several categories in estimating the population proportions, log-normal and gamma priors in estimating association factors in contingency tables and methods for estimating multinomial probabilities in contingency tables, using a Dirichlet prior distribution. Good also was innovative in his early use of hierarchical and empirical Bayesian approaches. His interest in this area apparently evolved out of his service as the main statistical assistant in 1941 to Alan Turing on intelligence issues during World War II. Lindley (1964) introduced the Bayesian inference about odds ratios, where used conjugate beta and Dirichlet priors. Althman (1969, 1971) presented Bayesian analogs of small-sample frequentist tests for 2 × 2 tables using conjugate priors. Leonard used a normal prior for logits, which allows greater flexibility. Leonard (1975) and Laird (1978) introduced the non-conjugate priors in Bayesian analysis of log-linear models, using a univariate normal prior to the parameters of the saturated model. Bernardo (1979) attempted to derive non-subjective posterior distributions that satisfy certain natural criteria such as invariance, consistency and admissibility. The intention is that even for small sample sizes the information provided by the data should dominate the prior information. Knuiman and Speed

(1988) proposed a multivariate normal prior for all parameters in log-linear model and extented the approach to multi-way contingency tables. They computed the posterior mode and used the curvature of the log posterior at the mode to measure precision. King and Brooks (2001) also specified a multivariate normal prior on the loglinear parameters, which induces a multivariate log-normal prior on the expected cell counts.

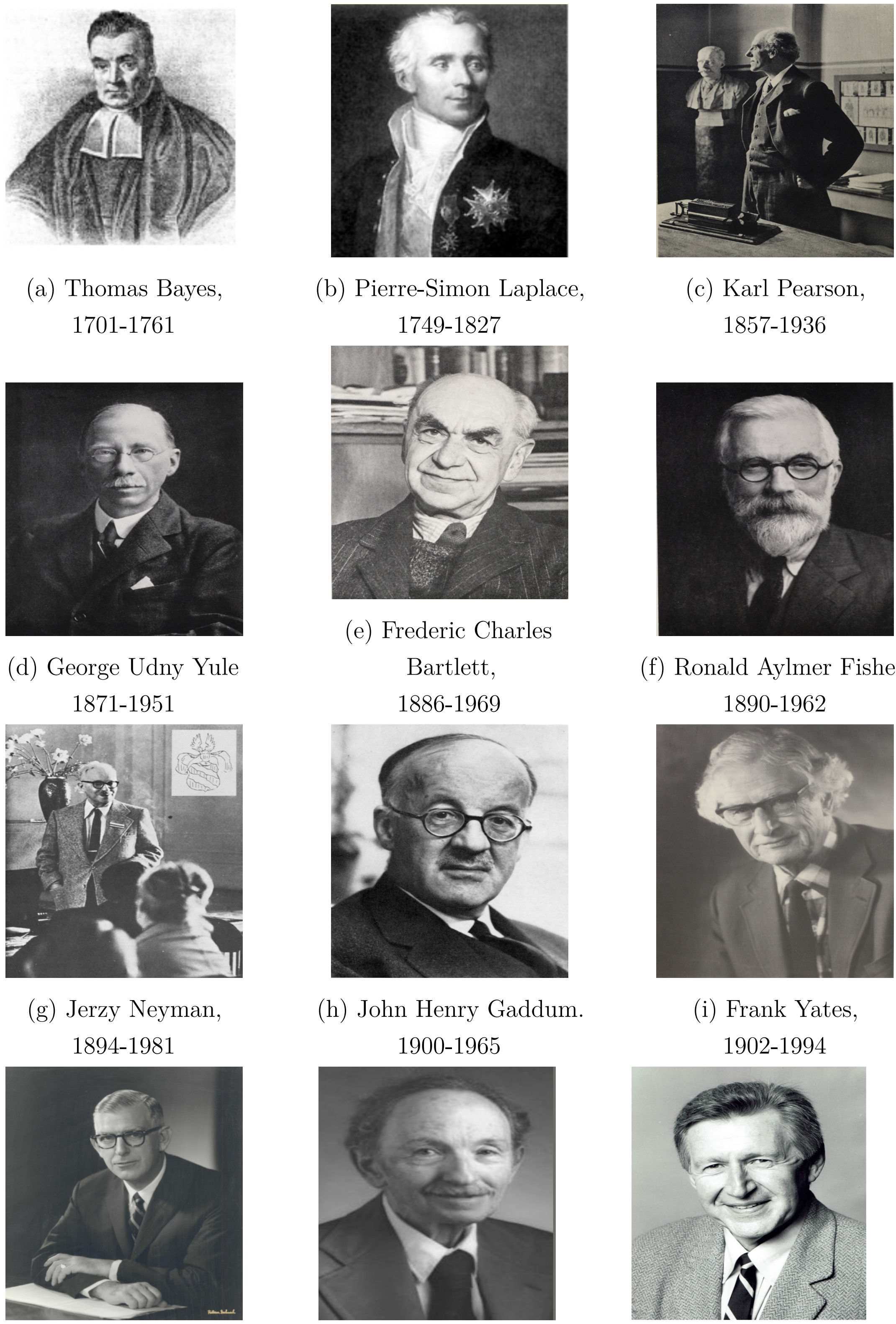
Loglinear model selection using Bayes factors was introduced by Spiegelhalter and Smith (1982a). They also provided an approximate expression for the Bayes factor for a multinomial loglinear model with an improper prior (uniform for the log probabilities) and showed how it is related to the standard chi-squared goodness-of-fit statistic. Shortly after Raftery (1986) noted that this approximation is indeterminate if any cell is empty but is valid when using a Jeffreys prior. He also noted that, for large samples, that the true log of this approximate Bayes factor when multiplicate by -2 is approximately equivalent to the Schwarz's BIC model selection criterion. Albert and Chib (1993) studied probit regression modeling, with extensions to ordered multinomial responses. Madigan and Raftery (1994) proposed a strategy for loglinear model selection with Bayes factors that employs model averaging. Raftery (1996b) used the Laplace approximation to integration in order to obtain the approximate Bayes factors for generalized linear models. Albert (1996) suggested partitioning the loglinear model parameters into subsets and testing whether specific subsets are nonzero. Using normal priors for the parameters, he examined the behavior of the Bayes factor under both normal and Cauchy priors, finding that the Cauchy was more robust to misspecified prior beliefs. Ntzoufras et al. (2000b) developed a MCMC algorithm for loglinear model selection.

Greenland (2001) proposed the approximation of the prior and the likelihood distribution by multivariate normal distribution in logistic and Poisson models with large samples. In the case of sparse data, such approximations may be inadequate.

For sparse data, he recommended exact conjugate analysis. Giving conjugate priors for the coefficient vector in logistic and Poisson models, he introduced a computationally feasible method of augmenting the data with binomial “pseudodata” having an appropriate prior mean and variance. Greenland also discussed the advantages conjugate priors have over non-informative priors in epidemiological studies, showing that flat priors on regression coefficients often imply ridiculous assumptions about the effects of the clinical variables. Congdon (2005) provides a comprehensive introduction to Bayesian methods of categorical data, emphasizing the use of statistical computing and applied data analysis. Forster and Webb (2007) proposed an Bayesian approach to calculate the predictive probabilities for those cells in a contingency table which have small sample frequencies and provides posterior predictive probabilities of identification risk.

Consonni and Pistone (2007) proposed Bayesian analysis of contingency tables with structural zeros based on algebraic statistics. Agresti and Hitchcock (2005b) provide a complementary historical overview over Bayesian inference for categorical data analysis. In Figure 1.1 below some of the leading figures in the development of categorical data analysis are shown.

Figure 1.1: Leading figures in the development of categorical data analysis.



(a) Thomas Bayes, 1701-1761

(b) Pierre-Simon Laplace, 1749-1827

(c) Karl Pearson, 1857-1936

(d) George Udny Yule 1871-1951

(e) Frederic Charles Bartlett, 1886-1969

(f) Ronald Aylmer Fisher, 1890-1962

(g) Jerzy Neyman, 1894-1981

(h) John Henry Gaddum. 1900-1965

(i) Frank Yates, 1902-1994

(j) William G. Cochran, 1909–1980

(k) Irving John Good, 1916–2009

(l) Leo Goodman, 1928–2020

1.3 Basic Principles on Bayesian Hypothesis Testing and Model Comparison

According to the title of the thesis we are going to deal with contingency tables which is a cross-classification of table of one factor versus the other resulting in frequency tabulation. In this thesis we focus on contingency tables which is a tabular representation of categorical data. A contingency table examines the cross-correlation or the cross-classification of two categorical variables X and Y with $I \geq 2$ and $J \geq 2$ levels, respectively, that are cross-classified in a $I \times J$ contingency table. The frequency counts of the contingency table is denoted by n_{ij} for cell (i, j), $i = 1, \ldots, I$, $j = 1, \ldots, J$ and notation-wise the index i stands for the row and j for the column category. The n_{i+} and n_{+j} are the marginal frequency of the ith row and jth column, respectively. In tale form this stated as follows:

n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	n_{1+}
n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}	n_{2+}
$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdots$	$\cdot$	
n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	n_{i+}
$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdots$	$\cdot$	
n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	n_{I+}
n_{+1}	n_{+2}	$\cdots$	n_{+j}	$\cdots$	n_{+J}	n

with $n = \sum_{ij} n_{ij}$ is the total number of observation of the data set, the sample size. With small n_{ij} is the observed frequencies but we are going to treat them as a random variable as well, so whenever the cell frequencires are random variables then we are going to denote them with capital N_{ij}.

In the classical hypothesis testing, let a model M_1 with parameters Θ_1 and density function $f(\boldsymbol{n}|\Theta_1)$ and M_0 with parameters Θ_0 density function $f(\boldsymbol{n}/\Theta_1)$

based on an unknown parameter θ, with $\theta \in \Theta$ we would like to know if $\theta \in \Theta_0$ or $\theta \in \Theta_1$, with $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$.

$$H_0 : \ \theta \in \Theta_0 \text{ is the } \textit{null hypothesis}$$

and to

$$H_1 : \ \theta \in \Theta_1 \text{ is the } \textit{alternative hypothesis.}$$

The probability of rejecting H_0 when it is actually true is called *Type I Error* while accepting H_0 when it is false is called *Type II Error* . The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. When this probability is small enough, in classical approach the null hypothesisis rejected. Therefore, the probability is the rate of committing a false alarm, Type error I, when selecting this specific value as the border of the rejection area and the goal of the decision threshold (usually set at 0.05) is to limit false alarms under control and to this specific value.

In classical statistic hypothesis testing can be framed as a special case of model comparison, but only for nested models. On the other hand in the context of Bayesian inference hypothesis testing is more general and natural. Bayesian hypothesis testing is based on constructing a probability model M, its likelihood $f(\boldsymbol{n}|\boldsymbol{\vartheta}_M, M)$ and the corresponding prior distribution $f(\boldsymbol{\vartheta}_M \mid M)$, where $\boldsymbol{\vartheta}_M$ is a parameter vector, $\boldsymbol{n}$ is the data vector and $f(\boldsymbol{\vartheta}_M \mid \boldsymbol{n},\ M)$. The posterior density is the usual expression where paramerter given the data is proportional to the likelihood times the prior we have specified before $f(\boldsymbol{\vartheta}_M \mid \boldsymbol{n},\ M) \propto f(\boldsymbol{n}|\boldsymbol{\vartheta}_M, M) f(\boldsymbol{\vartheta}_M \mid M)$. Although, the Bayesian inference is primarily based on the posterior distribution $f(\boldsymbol{\vartheta}_M \mid \boldsymbol{n},\ M)$, in Bayesian hypothesis testing we want to quantify the model uncertainty by estimating the posterior model probability $f(M \mid \boldsymbol{n})$. Let us consider two competitive models M_0 and M_1. If $f(M)$ is the prior probability of model M, then using the Bayes theorem, the posterior odds

PO_{01} of model M_0 versus M_1 are given by

$$PO_{01} = \frac{f(M_0|\boldsymbol{n})}{f(M_1|\boldsymbol{n})} = \frac{f(\boldsymbol{n}|M_0)}{f(\boldsymbol{n}|M_1)} \times \frac{f(M_0)}{f(M_1)} = B_{01} \times \frac{f(M_0)}{f(M_1)}. \tag{1.2}$$

The posterior odds are now the main quantity of interest and we are going to use them for model selection or model valuation. Posterior odds are the ratio of posterior model probabilities of the two models and can be written as the product of the Bayes factor B_{01}, which is the ratio of the marginal likelihoods of the two competitive models, multiplied by the prior model odds $\frac{f(M_0)}{f(M_1)}$ of model M_0 against model M_1.

Bayes factor M_0 versus M_1 is defined as the ratio of the marginal likelihoods of the two competitive models The quantity $f(\boldsymbol{n}|M)$ involved in the Bayes factor is defined as the marginal likelihood of model $\frac{f(\boldsymbol{n}|M_0)}{f(\boldsymbol{n}|M_1)}M$ introduced by Jeffreys (1961) and is defined as

$$f(\boldsymbol{n}|M) = \int f(\boldsymbol{n}|\boldsymbol{\vartheta}_M,\ M)f(\boldsymbol{\vartheta}_M|M)d\boldsymbol{\vartheta}_M. \tag{1.3}$$

The Bayes factor B_{01} of model M_1 against M_0, evaluates the evidence against the null hypothesis and the prior model probabilities are equal or expressing ignorant or indifference between the two models under comparison or under consideration and is familiar framework similar to the classical significant tests. Large values of B_{01}, usually greater than 12 (this is one of the most common interpretations first proposed by Jeffreys (1961) and slightly modified by Lee and Wagenmakers (2014)), indicate strong posterior support of model M_0 against model M_1. Alternatively, when we consider a set of competing models $\mathcal{M} = \left\{M_1,\ M_2,\ \dots\ ,\ M_{|\mathcal{M}|}\right\}$, then we focus our interest on the posterior probability model $M_m \in \mathcal{M}$ and $m = 1, \dots, \mathcal{M}$, given by

$$f(M|\boldsymbol{n}) = \frac{f(\boldsymbol{n}|M)f(M)}{\sum_{M_m \in \mathcal{M}} f(\boldsymbol{n}|M_m)f(M_m)} = \left(\sum_{M_m \in \mathcal{M}} PO_{M_m,M}\right)^{-1} \tag{1.4}$$

where $|\mathcal{M}|$ is the number of models under consideration.

The integrals involved in the computation of the posterior model probabilities are mostly analytically intractable. So sometimes asymptotic approximation or alternative computational methods must be applied. One of the most popular techniques is the Markov chain Monte Carlo (MCMC) or the reversible jump MCMC, Green (1995), denoted by RJMCMC, which helps us to account for model uncertainty.

Kass and Wasserman (1995) introduce the utilization of Schwarz (1978) criterion called BIC as an approximation of Bayes factor when unit information priors (UIP) are used. The definition of a unit information prior (UIP) is a prior that contains an information content equivalent to a sample of size one.

Given the log-likelihood function $f(\boldsymbol{n}|\boldsymbol{\vartheta}_M, M)$, the Fisher information matrix $\boldsymbol{I}(\theta)$ is a symmetric $(d_M \times d_M)$ matrix given by $\boldsymbol{I}(\theta) = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} f(\boldsymbol{n}|\boldsymbol{\vartheta}_M, M)$, with $1 \leq i,\ j \leq d_M$, where d_M is the dimension of the parameter vector $\boldsymbol{\vartheta_M}$. The observed Fisher information matrix is simply $\boldsymbol{I}(\theta^{ML})$, the information matrix evaluated at the maximum likelihood estimates (MLE). Given a dataset of size n, the observed Fisher information matrix under model M divided by n can be interpreted as an estimate of the average amount of information in one data point. When $\boldsymbol{\vartheta_M} \in \Theta_{d_M}$ one way to form the UIP is

$$\boldsymbol{\vartheta}_M|M \sim N_{d_M}\Big(\boldsymbol{\mu}_{\vartheta_M},\ n[\mathcal{J}_M^n(\boldsymbol{\mu}_{\vartheta_M})]\Big)^{-1} \tag{1.5}$$

where $\mathcal{J}_M^n(.)$ is the negative of the Hessian matrix (the matrix of second derivatives of the likelihood function with respect to the parameters) of the log-likelihood. Under this prior the logarithm of the Bayes factor is asymptomatically equal to the Schwarz criterion (BIC) and by this way UIP provides an extra tool to Bayesian model selection procedure. The prior mean $\boldsymbol{\mu}_{\vartheta_M}$ can be replaced by the MLE when flat prior is used. A simpler alternative was proposed by Ntzoufras (2009) by considering independent prior distributions with mean set to the corresponding posterior means under a flat prior, while the variance is set equal to the posterior variance under a flat prior inflated by a factor of n in order to account for approximation of one data point. This approach will be used by Consonni et al. (2018) as a method for constructing objective prior distribution, called Unit information principle, a useful tool to objective Bayes analysis that will be discussed analytically in Chapter 2. The posterior model probabilities under this approach will be used as a yardstick in this thesis illustrated examples for comparison with other prior setups. The combination of the unit information principle with the utilization of power prior approach specifying the prior mean $\boldsymbol{\mu}_{\vartheta_M}$ by imaginary data is a sensible choice of prior for model comparison.

1.4 Test of Independence

In contingency tables we usually focus on the underlying association between the two classification variables and consequently, on testing for their independence. From the Bayesian perspective, there are several ways depending on the model (likelihood and prior) assumed. Under conjugacy, we may use the Multinomial-Dirichlet or the Poisson-Gamma models. In log-linear models, there are only two options for modelling the dependence structure in two-way contingency tables. The parsimonious, but restrictive and hardly fulfilled in practice, model of independence and the saturated one.

In Poisson log-linear models, $N_{ij} \sim Poisson(\lambda_{ij})$, with $i = 1, \ldots, I$ and $j = 1, \ldots, J$ the independence model formed by

$$M_0 : \log(\lambda_{ij}) = \lambda_0 + \lambda_i^X + \lambda_j^Y \tag{1.6}$$

which is sum of three parameters, λ_0 is an overall measure of average log-counts and λ_i^X and λ_j^Y are the marginal effect terms for factors X (rows) and Y (columns), respectively. In saturated model, which is formed by

$$M_1 : \log(\lambda_{ij}) = \lambda_0 + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \tag{1.7}$$

has one additional parameter λ_{ij}^{XY} the interaction term. Association between factors is expressed via this interaction term and reflects the deviation from the independence assumption. The basic comparison in contingency tables is:

$$H_0 : \lambda_{ij}^{XY} = 0 \quad \text{versus} \quad H_1 : \lambda_{ij}^{XY} \neq 0, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J$$

If the interaction term is equal to zero then we obtain the independence model.

In Poisson log-linear models, the Bayes factor is not analytically available and can be calculated by using Markov Chain Monte Carlo (MCMC) methods. Gunel and Dickey (1974) considered independence in two-way contingency tables under the Poisson, multinomial, independent multinomial, and hypergeometric sampling models. They showed that the Bayes factor for independence itself factorizes, highlighting the evidence residing in the marginal totals. In classical approach, the maximum likelihood estimates are the same for the four different sample schemes. Similar is the result for the Bayesian framework, where the Bayes factor will be the same for the different sample schemes provided that the specified priors are compatible across different parameter spaces.

Bayesian analysis follows three steps: constructing a probability model, computing the posterior distribution and model evaluation. In this chapter, we are focusing on the third step. It should be stated that a good Bayesian analysis should include mechanisms to check the adequacy of the fit of the model to the data. For example, if a model is poor this can lead to misleading inference.

1.5 Sensitivity of the Bayes Factor

Posterior model probabilities and the Bayes factor are highly sensitive to the prior specification of the model parameters. This result was reported after the publication of Lindley (1957), who reported a surprising behavior of the Bayes factor. When the sample size n increases, then the Bayes factor also increases and tends to infinity, fully supporting the simpler hypothesis in contrast to the standard frequentist hypothesis test, which supports the more complex hypothesis.

$$\boxed{n \to \infty \quad \Rightarrow \quad B_{01} \to \infty.} \tag{1.8}$$

This behaviour is known as the Lindley's Paradox. Subsequently, motivated by the work of Lindley, Bartlett (1957) extended this paradox by observing that the prior variance of the additional parameters in nested models comparisons (when $M_0 \subseteq M_1$), also massively affects the Bayes factor B_{01} as it tends to infinity. This behaviour is known as Jeffreys or Lindley's or Barlett's paradox. For the model comparison of Section 1.9 the Lindley-Bartlett paradox; this can be expressed as:

$$\boxed{Var(\vartheta_{ij}) \to \infty \quad \Rightarrow \quad B_{01} \to \infty} \tag{1.9}$$

for any dataset n. This behavior is the main quantity of interest of this thesis and the reason why we start our research in Bayesian hypothesis testing and deal with the problem of sensitivity of Bayes factor. The problem is that whenever the sample size or the prior variance increase also Bayes factor increases, which means that no matter the data we have, we support the null hypothesis or the simpler model. As a consequence the analysis is informative, without our tension to be informative. The aim of this thesis is to build an objective Bayesian model selection procedure in order to eliminate this effect and use it under the prior ignorance.

Therefore, the specification of the prior variance for the parameters under testing is of primary importance. When we do not have a clear indication about this value, we can perform sensitivity analysis of the Bayes factor with respect to the value of the prior variance for the parameter under testing. This procedure is similar to the regularization plots used for the selection of the shrinkage parameter in Lasso-type methods; see for example Lykou and Ntzoufras (2013). In this thesis, we use power priors to control the effect of the Lindey's paradox. We consider imaginary data and weight them to account for one data point as a reasonable, low-information, choice in Chapter 3.

1.6 Other Bayesian Model Selection Criteria

A variety of information criteria is available in the literature of model comparison and adequacy. The most popular information criteria are AIC,(AIC; Akaike, 1973) and the BIC, (BIC; Schwarz, 1978). More recently the DIC was introduced by Spiegelhalter et al. (2002), as an extension of AIC for Bayesian hierarchical models. All these criteria are penalized likelihood measures giving different weight to complexity and goodness of fit. Lower values indicate better fitted models.

The BIC value for a model M is defined as

$$BIC(M) = D(\widehat{\boldsymbol{\vartheta}}_M, M) + d_M \log n$$

where $D(\boldsymbol{\vartheta}_M, M)$ is the deviance measure of model M evaluated at $\boldsymbol{\vartheta}_M$, $\widehat{\boldsymbol{\vartheta}}_M$ are the maximum likelihood estimates under model M and d_M is the dimension (number of parameters) of the model M. For large n and for specific prior families

$$-2 \log B_{01} \approx BIC(M_0) - BIC(M_1) = \Delta BIC_{01},$$

see, for details Kass and Wasserman (1995). From this form we can obtain approximate posterior model probabilities, assuming the uniform prior distribution for all candidate models.

The Akaike's information criterion is defined as

$$AIC(M) = D(\widehat{\boldsymbol{\vartheta}}_M, M) + 2d_M$$

and is a penalized deviance measure with penalty equal to two for each estimated parameter. The penalty is $\log n$ and 2 for AIC and BIC, respectively. Hence, AIC supports less parsimonious models than BIC for $\log n > 2$.

Spiegelhalter et al. (2002) proposed the deviance information criterion(DIC) which is considered as the Bayesian analogue of AIC. DIC can be directly estimated via the MCMC output and can be applied in a variety of models including the hierarchical, random effects and latent variable models.

1.7 Objective Bayes principles

The need to work without introducing subjective inputs into the Bayesian analysis led to the growth of Objective Bayes techniques, which constitutes the philosopher's stone for the Bayesian community in the last decades. The focus is to search for priors with a minimal impact on the corresponding posterior analysis. Consonni et al. (2018) explain what constitutes an Objective Bayes analysis and the principles for an objective model comparison. Bayarri et al. (2012) introduced the desiderata or criteria of an objective prior distribution for Bayesian model choice. The Objective Bayes criteria are:

- $\boldsymbol{C_1}$: The prior of each model parameter to be proper, so that Bayes factors do not contain different arbitrary normalizing constants across distinct models.
- $\boldsymbol{C_2}$: The model selection consistency criterion. If data have been generated by model M, then the posterior probability of M should converge to one as the sample size diverges.

- $\boldsymbol{C_3}$: The information consistency criterion. If there exists a sequence of datasets with the same sample size n such that the likelihood ratio between M_0 and M_1 goes to infinity, then the corresponding sequence of Bayes factors should also go to infinity.

- $\boldsymbol{C_4}$: The intrinsic consistency criterion. While features of the model and sample size (and possibly even data) frequently affect model selection priors, such features should disappear for large n.

- $\boldsymbol{C_5}$: The predictive matching deal with the minimal sample size. Informally, with a minimal sample size, one should not be able to discriminate between two models, so that the Bayes factor should be close to one, for all samples of minimal size. In particular, exact predictive matching occurs if the Bayes factor equals one.

- $\boldsymbol{C_6}$: Measurement invariance broadly states that answers should not be affected by changes of measurement units. A special type of invariance arises when the families of sampling distributions of models under consideration are such that the model structures are invariant to group transformations.

- $\boldsymbol{C_7}$: The group invariance criterion states that if models M_0 and M_1 are invariant under a group of transformations G, then the conditional priors should be chosen in such a way that the conditional marginal distribution is also invariant under G. This means that if models exhibit an invariance structure, this should be preserved after marginalization.

Three other principles for an objective model comparison are:

- **compatibility of the prior across the models**, priors should be related across models, although in principle they need not be, each being conditional on a given model. We can identify a benchmark model (often the null model), which is nested into every other model under consideration (encompassing from below), so that compatibility is realized between each model and the benchmark, and thus indirectly between any pair of models.

- **validation of Bayesian approaches** using plausible default proper priors. An acceptable Bayesian procedure should correspond, at least asymptotically, to a prior which makes sense in the context where it is applied

- **methods with good frequentist properties**, use prior distributions that lead to good frequentist performances.

Some popular methods for constructing objective prior distributions are: Unit information principle, training samples, imaginary observations, fixed imaginary data, power priors, power expected posterior priors and empirical Bayes approaches. Some of them will be disused in the next Chapter, for more information about the Objective Bayes desiderata and principles see Bayarri et al. (2012) and Consonni et al. (2018).

1.8 Conjugate Multinomial-Dirichlet Setup

Under this setup and under the model of dependence M_1 (saturated), we assume a Multinomial distribution for the random variable N, $N \sim Mult\,(n, \boldsymbol{\pi})$. Hence, the likelihood for model M_1 is given by

$$f(\boldsymbol{n} \mid \boldsymbol{\pi}, M_1) = \frac{n!}{\prod_{i=1}^{I}\prod_{j=1}^{J} n_{ij}} \prod_{i=1}^{I}\prod_{j=1}^{J} \pi_{ij}^{n_{ij}} \quad \text{with} \quad \sum_{i=1}^{I}\sum_{j=1}^{J} \pi_{ij} = 1. \tag{1.10}$$

The parameter of interest is the probability table $\boldsymbol{\pi}$, with entries π_{ij}, $\pi_{ij} = P(X = i, Y = j)$ for the i^{th} row and j^{th} column of the table. Hence, the parameter matrix is denoted by $\boldsymbol{\pi} = \{\pi_{ij};\ i = 1, \dots I,\ j = 1, \dots, J\} = \boldsymbol{\vartheta}_1$.

The conjugate prior for the multinomial parameter $\boldsymbol{\pi}$ is the Dirichlet distribution, a multivariate generalization of the Beta distribution. Therefore, we a-priori assume that

$$\boldsymbol{\pi} \sim Dirichlet(\boldsymbol{k}) \text{ with } \boldsymbol{k} = \{k_{ij} > 0;\ i = 1, \dots I, j = 1, \dots, J\}$$

with density function given by

$$f(\boldsymbol{\pi}) = \frac{\Gamma(K)}{\prod_{i=1}^{I} \prod_{j=1}^{J} \Gamma(k_{ij})} \prod_{i=1}^{I} \prod_{j=1}^{J} \pi_{ij}^{k_{ij}-1},$$

where $\Gamma(K)$ is the Gamma function with $K = \sum_{i=1}^{I} \sum_{j=1}^{J} k_{ij}$.

The posterior distribution is $Dirichlet(\boldsymbol{n} + \boldsymbol{k})$. Moreover, the marginal likelihood under M_1 is now given by

$$f(\boldsymbol{n}|M_1) = c \times \frac{\Gamma(K)}{\Gamma(n+K)} \prod_{i=1}^{I} \prod_{j=1}^{J} \frac{\Gamma(n_{ij} + k_{ij})}{\Gamma(k_{ij})}, \tag{1.11}$$

where $c = \dfrac{n!}{\prod_{i=1}^{r} \prod_{j=1}^{c} n_{ij}!}$.

Under the independence model M_0 we have that $\pi_{ij} = \pi_{i+}\pi_{+j}$ for all $i = 1, \dots, I$ and $j = 1, \dots, J$. So now the parameters of interest are the marginal probabilities vectors $\boldsymbol{\pi}^R = \Big(\pi_{i+};\ i = 1, \dots, I\Big)$ and $\boldsymbol{\pi}^C = \Big(\pi_{+j};\ j = 1, \dots, J\Big)$, that is $\boldsymbol{\vartheta}_0 = (\boldsymbol{\pi}^R, \boldsymbol{\pi}^C)$. By substituting $\pi_{ij} = \pi_{i+}\pi_{+j}$ in (1.10) we obtain

$$f(\boldsymbol{n} \mid \boldsymbol{\pi}^R, \boldsymbol{\pi}^C) = c \times \prod_{i=1}^{I} \pi_{i+}^{n_{i+}} \prod_{j=1}^{J} \pi_{+j}^{n_{+j}}.$$

Assuming independent Dirichlet priors for $\boldsymbol{\pi}^R$ and $\boldsymbol{\pi}^C$ of the following form

$$\boldsymbol{\pi}^R \sim Dirichlet\Big(k_1^R, \cdots, k_I^R\Big) \text{ and } \boldsymbol{\pi}^C \sim Dirichlet\Big(k_1^C, \cdots, k_J^C\Big),$$

where $\boldsymbol{k^R} = \{k_i^R;\ i = 1, \dots I\}$ and $\boldsymbol{k^C} = \{k_j^C; , j = 1, \dots, J\}$ are the parameter of the Dirichlet of each row and column, respectively. We obtain the marginal likelihood under M_0 model given by

$$f(n|M_0) = c \times \frac{\Gamma(K^R)}{\Gamma(N+K^R)} \frac{\Gamma(K^C)}{\Gamma(N+K^C)} \prod_{i=1}^{I} \frac{\Gamma\Big(n_{i+}+k_i^R\Big)}{\Gamma\Big(k_i^R\Big)} \times \prod_{j=1}^{J} \frac{\Gamma\Big(n_{+j}+k_j^C\Big)}{\Gamma\Big(k_j^C\Big)} \tag{1.12}$$

From 1.11 and 1.12, the Bayes factor is given by

$$B_{10} = \frac{\mathcal{B}(n+\boldsymbol{k})}{\mathcal{B}(\boldsymbol{k})} \times \frac{\mathcal{B}(\boldsymbol{k}^R)}{\mathcal{B}(n^R+\boldsymbol{k}^R)} \times \frac{\mathcal{B}(\boldsymbol{k}^C)}{\mathcal{B}(n^C+\boldsymbol{k}^C)} \tag{1.13}$$

where $n^R = (n_{1+}, \ldots, n_{I+})$, $n^C = (n_{+1}, \ldots, n_{+J})$ and $\mathcal{B}$ is the normalizing constant of the Dirichlet distribution (also known as multivariate beta function) given by

$$\mathcal{B}(\boldsymbol{\alpha}) = \frac{\prod\limits_{i=1}^{|\boldsymbol{\alpha}|} \Gamma(\alpha_i)}{\Gamma\left(\sum\limits_{i=1}^{|\boldsymbol{\alpha}|} \alpha_i\right)} \;;$$

for any vector $\boldsymbol{\alpha} = \Big(\alpha_i;\ i = 1, \ldots, |\boldsymbol{\alpha}|\Big)$, with $|\boldsymbol{\alpha}|$ denoting the dimension of $\boldsymbol{\alpha}$. The conjugate analysis of this section is summarized in Table 1.1.

Table 1.1: Summary of the Bayesian testing of the independence in the conjugate Multinomial-Dirichlet setup

H_0: there is no association between the two variables	H_1: there is association between the two variables
$M_0 : n\vert\boldsymbol{\pi}^R, \boldsymbol{\pi}^C \sim Multinomial(n, \boldsymbol{\pi})$	$M_1 : n\vert\boldsymbol{\pi} \sim Multinomial(n, \boldsymbol{\pi})$
$\boldsymbol{\pi} = \boldsymbol{\pi}^R \left[\boldsymbol{\pi}^C\right]^T$	$\boldsymbol{\pi} = (\pi_{ij})$
$\boldsymbol{\pi}^R = (\pi_{i+}) \sim Dirichlet\Big(k_1^R, \cdots, k_I^R\Big)$	$\boldsymbol{\pi} \sim Dirichlet(\boldsymbol{k})$
$\boldsymbol{\pi}^C = (\pi_{+j}) \sim Dirichlet\Big(k_1^C, \cdots, k_J^C\Big)$	$\boldsymbol{k} = (k_{ij})$
$i = 1, \cdots, I$ and $j = 1, \cdots, J$	$i = 1, \cdots, I$ and $j = 1, \cdots, J$

1.9 Poisson Log-linear Models

Poisson log-linear model belong to the family of generalized linear model (GLM) and their response is the cell frequencies of a contingency table. Associations between factors are expressed via interaction terms of the model. In the two-way contingency table, the model takes the form

$$\begin{aligned} &N_{ij} \sim Poisson(\lambda_{ij}), \quad i = 1, \dots, I, \quad j = 1, \dots, J \\ &\log(\lambda_{ij}) = \lambda_0 + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \end{aligned} \tag{1.14}$$

with identifiability constraints

$$\sum_{i=1}^{I} \lambda_i^X = \sum_{j=1}^{J} \lambda_j^Y = \sum_{i=1}^{I} \lambda_{ij}^{XY} = \sum_{j=1}^{J} \lambda_{ij}^{XY} = 0.$$

Parameter λ_0 is an overall measure of average log-counts, while λ_i^X and λ_j^Y are marginal effect terms for factors X (rows) and Y (columns), respectively. The interaction term λ_{ij}^{XY} represents the association between X and Y and reflects the deviations from independence. In the frequentist framework, this model has zero degrees of freedom and thus fits the data perfectly (all residuals are equal to zero). Given the sample size n, the kernel of the likelihood is the same for all three sampling schemes: multinomial, product multinomial and Poisson.

If λ_{ij}^{XY} is equal to zero for all i and j, then we obtain the independence model with linear predictor

$$\log(\lambda_{ij}) = \lambda_0 + \lambda_i^X + \lambda_j^Y. \tag{1.15}$$

Here we illustrate the comparison between the model of independence M_0, given by the log-linear predictor 1.15 versus the saturated model M_1 with log-linear predictor 1.14. By this comparison we test for

$$H_0 : \lambda_{ij}^{XY} = 0 \quad \text{versus} \quad H_1 : \lambda_{ij}^{XY} \neq 0, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J \tag{1.16}$$

which is equivalent to testing for independence.

For all log-linear parameters we consider normal prior distributions with zero means. The prior variances of λ_0, λ_i^X (for $i = 2, \ldots, I$) and λ_j^Y (for $j = 2, \ldots, J$) are set equal to large values in order to express prior ignorance. Note that λ_1^X and λ_1^Y and λ_{i1}^X, λ_{1j}^Y are given as function of the rest of parameters since we use sum-to-zero constraints (STZ). Nevertheless, the prior variance for the interaction parameters λ_{ij}^{XY} needs to be specified with caution due to the sensitivity of the Bayes factors to this value for the parameters under testing. Therefore, the specification of the prior variance of λ_{ij}^{XY} is important for avoiding the activation of the Lindley-Barlett paradox (Bartlett, 1957, Lindley, 1957); see also Section 1.5.

For this comparison, in the illustration of Section 2.4.4 we have used two different prior setups: (a) a Unit Information Empirical prior (UIE) proposed by Ntzoufras (2009, chap. 9) and (b) a prior similar to the one proposed by Dellaportas and Forster (1999) (DF). Regarding the first, we set the prior variance of the interaction parameters equal to their posterior variances of the full model multiplied by the sample size, in order to minimize the information introduced by the double use of the data. Thus, the contribution of the prior information is approximately equal to one data point; see for details Verdinelli and Wasserman (1995). This is an empirical Bayes approach and it cannot be considered as an orthodox Bayesian method since we use information from the data to specify the prior. Nevertheless, it can be used as a rough approximation of the actual objective Bayesian procedure since the effect of the prior information on the posterior is minimized; see for details Consonni et al. (2018). For the second prior setup we follow the approach proposed by Dellaportas and Forster (1999) for the case of $I = J = 2$, hence we set the prior variance equal to two.

We calculate the posterior model probabilities of the two models under consideration by using the Gibbs variable selection sampler of Dellaportas et al. (2002); also see Ntzoufras et al. (2000a) for an implementation in contingency tables using the Stochastic Search Variable Sampler (SSVS). In order to implement any Gibbs variable selection, we introduce the latent binary indicator γ which takes the value of one when $\lambda_{ij}^{XY} \neq 0$ and zero otherwise. Then the log-linear predictors 1.14 and 1.15 are jointly summarized by

$$\log(\lambda_{ij}) = \lambda_0 + \lambda_i^X + \lambda_j^Y + \gamma\lambda_{ij}^{XY}.$$

The full Bayesian procedure is completed by considering a prior $\gamma \sim Binomial(\pi_\gamma)$ with $\pi_\gamma = 0.5$ in order to express indifference between the two models with priors $f(\lambda_i^X, \lambda_j^Y | \gamma = 0)$ and $f(\lambda_i^X, \lambda_j^Y, \lambda_{ij}^{XY} | \gamma = 0)$ The posterior probabilities $P(\gamma = 1|n)$ and $P(\gamma = 0|n)$ will indicate whether the saturated ($\gamma = 1$) or the independence model ($\gamma = 0$), is more suitable for the data in hand. Gibbs variable selection can be implemented in a straightforward manner using standard Bayesian statistical packages, such as WinBUGS (see Ntzoufras, 2009, Chapter 11).

1.10 Illustrative Examples

In this section, we implement the Bayesian tests of independence using data from several real examples. We illustrate two examples, in the first example we use three 2 × 2 contingency tables and testing the independence. The second example provides a conjugate analysis in $I \times J$ contingency tables and tests the sensitivity of Bayes factor.

1.10.1 Examples of 2×2 Contingency Tables

First, we implement Bayesian tests via log-linear models in three different 2 × 2 contingency tables: The first table is a cross-classification of 100 individuals randomly sampled from a large population as part of a study of sex differences by handedness (dataset a; Source: *http://en.wikipedia.org*). The second contingency table (dataset b) refers to the number of ant and vertebrate dispersed plant species by seed and vegetative regeneration (Source: French and Westoby, 1996) the third one consists of 200 randomly selected cancer patients examined for the cell phone usage by the brain cancer diagnosis (dataset c; Source: *http://wiki.stat.ucla.edu.*).

Table 1.2: Datasets (a)–(c) used for testing independence via log-linear models

	Right Handed	Left Handed
Males	43	9
Females	44	4

Dataset (a)

Sex by Handness

	Ant	Vertebrate
Seed only	25	6
Vegetative	36	21

Dataset (b)

Dispersal by Regeneration

		Brain cancer	
		yes	no
Cell	yes	18	80
Phone	no	7	95

Dataset (c)

Brain Cancer by Cell Phone Usage

Table 1.3 presents the estimated Odds Ratio (OR), p-values for testing independence based on the Likelihood Ratio Test (LRT) with $\alpha = 0.05$ and the information criteria: DIC and BIC. For the two first examples H_0 is not rejected based on LRT in contrast to the third contingency table. In comparison with the information criteria, for the second contingency table both DIC and BIC indicate the saturated model as the best model. Hence, the three selected datasets cover a variety of different levels of dependence which make them appropriate to illustrate model uncertainty via the Bayesian paradigm.

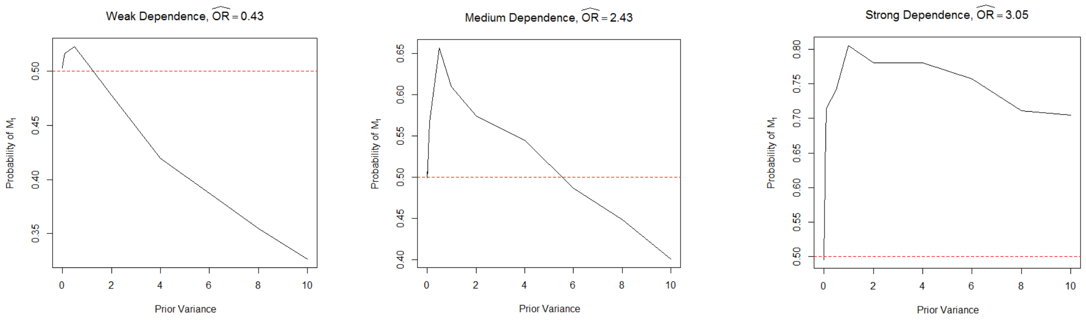
Table 1.3: Bayesian information criteria for datasets (a)–(c)

				DIC		**BIC**	
Dataset	**Dependence**	$\widehat{OR}$	**LRT**	M_0	M_1	M_0	M_1
(c)	Weak	0.43	Not rejected, p-value=0.30	26.38	26.73	24.53	29.85
(d)	Medium	2.43	Not rejected, p-value=0.14	28.07	26.91	30.54	30.34
(e)	Strong	3.05	Rejected, p-value=0.02	33.36	28.96	36.68	32.46

LRT: Decision concerning the independence (H_0) based on the likelihood ratio test with $\alpha = 0.05$

M_0: independence model; M_1: saturated model

Using Gibbs variable selection, we have estimated the posterior model probability $P(M_1|n) = P(\gamma = 1|n)$ for a range of prior variance values for the interaction term λ_{22}^{XY}, which are depicted in Figure 1.2. From these figures two points are evident: (i) the surface between the line of the posterior model probability and 0.5 increases as we move from dataset (a) to dataset (c) and therefore with $\widehat{OR}$, and (ii) the posterior model probability of M_1 decreases as the prior variance increases after a threshold value. Concerning point (i), the posterior probability for dataset (a) reaches its maximum value (which is just above 0.5) for prior variance lower than two. In dataset (b), the maximum of the posterior model probability is about 0.65, while for (c) is around 0.8. The main characteristic of the latter is that the posterior model probability of M_1 remains above 0.5 for all values of the prior variance presented here, indicating the strong association between brain cancer and cell phone usage. Point (ii) clearly depicts a realization of the Lindley-Bartlett paradox. For datasets with weak association (as in dataset a), the posterior probability of M_1 shrinks fast towards zero as the prior variance of the interaction term increases, while for strong associations (as in dataset c) the paradox is delayed and only appears for large values of the prior variance.



(a) Dataset (b) Dataset (c) Dataset

Figure 1.2: Posterior model probability of M_1 versus the prior variance of the interaction for datasets (a)–(c) of Table 1.2.

1.10.2 Examples of $I \times J$ Contigeny Tables

We proceed by illustrating the conjugate analysis in two datasets: (d) a 5×3 cross-classification of the change in the clinical condition by degree of infiltration (i.e. the skin damage) from leprosy at the start of the experiment (Source: Agresti, 2013) and (e) a 4×2 contingency table of snoring by heart disease (Source: Norton and Dunn, 1985).

Table 1.4: Datasets (d) and (e) used for testing independence via conjugate analysis

	Degree of Infiltration	
Clinical Change	High	Low
Worse	1	11
Stationary	13	53
Slight improvement	16	42
Moderate improvement	15	27
Marked improvement	7	11

(d) Leprosy dataset (Agresti, 2013)

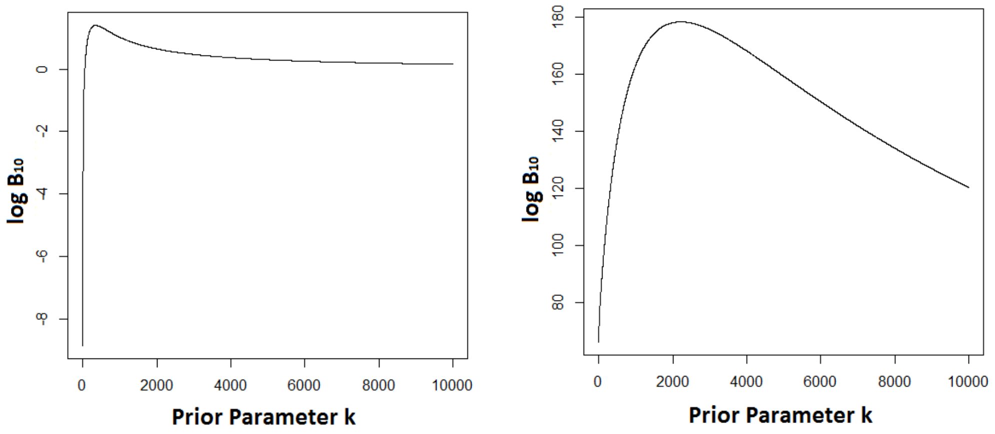
	Heart Disease	
Snoring	Yes	No
Never	24	1355
Occasionally	35	603
Nearly every night	21	192
Every night	30	224

(e) Snoring dataset

(Norton and Dunn, 1985)

For the first dataset, the independence assumption is supported by the Likelihood Ratio Test (LRT $p - value = 0.122$), AIC ($\Delta AIC_{01} = -0.72$) and BIC ($\Delta BIC_{10} = -13.84$) measures. For the second dataset, all corresponding measures support the dependence between the two factors under consideration

(LRT $p - value \ll 0.001$, $\Delta AIC_{01} = 59.9$ and $\Delta BIC_{01} = 42.45$).



(d) Leprosy dataset

(e) Snoring dataset

Figure 1.3: Sensitivity of Bayes factor to the Dirichlet prior parameter $\boldsymbol{k}$ for datasets (d) and (e) of Table 1.4.

In this Section, we implement the Bayesian independence tests based on the conjugate analysis of Section 1.8. We consider prior values $k_{ij} = k_i^R = k_j^C = \boldsymbol{k}$ and we perform sensitivity analysis over different values of prior parameter vector $\boldsymbol{k}$. From Figure 1.3 we observe that in dataset (d), the independence model is supported for a wide range of values, while for large values of $\boldsymbol{k}$ (accumulated prior information) the log-Bayes factor is stabilized in a value just above zero. In contrast, in the second example (where the dependence is strongly supported by other methods) the Bayes factor offers conflicting evidence against M_0 for all values of $\boldsymbol{k}$. It should be noted that in both cases, the Bayes factor decreases for values of $\boldsymbol{k}$ close to zero. This is due to Lindley-Barlett paradox since in this case the prior variance of the probability parameters (induced by the imposed Dirichlet prior) is increasing.

1.11 Chapter Structure

This thesis is focused on models comparison for contingency tables. Initially, we present Bayesian hypothesis tests for the independence between two categorical variables and we implement conjugate analysis based on the Multinomial-Dirichlet setup. Then we compute the Bayes factor and assess the sensitivity of the results to the prior distribution. Next, we focus on log-linear models. We evaluate the saturated versus the model of independence with the utilization of the Bayes factor, compared with other several model selection criteria. We illustrate all methods using real datasets.

In Chapter 2 we present the Bayesian independence tests based on the conjugate Multinomial-Dirichlet models (Section 1.8) and the Poisson log-linear models (Section 1.9). We also describe the well-known sensitivity of the Bayes factor on the prior parameters (Section 1.5), other Bayesian measures of model comparison (Section 1.6) and we conclude with illustrative examples, in order to implement and compare these approaches.

In Chapter 3 we propose a comprehensive Bayesian model comparison approach for association models in contingency tables. The proposed methodology deals with the suitable specification of the prior distributions, as well as the allied computational issues regarding the estimation of the Bayesian so-called evidence, which is the core component of the posterior model probabilities in Bayesian model comparison, selection and averaging. Specifically the choice of the prior distribution in Bayesian model comparison and testing is problematic due to the well-known sensitivity of the posterior model odds and the Barlett-Lindley paradox (Bartlett, 1957, Lindley, 1957). This fact had led to the development of objective Bayes techniques, which refer to the use of reasonably low information priors when no actual prior information is available. Within this framework, we propose the utilization of the power prior approach. In order to implement the method, we devise a set of imaginary data from the most parsimonious model. Then, the prior distribution can be obtained by the product of the likelihood of the model under consideration evaluated at the imaginary data and raised to a power; then it is multiplied by (a relatively flat) pre-prior distribution.

Evaluation of the models under consideration and the related Bayesian tests are obtained by using MCMC-based estimators of the (Bayesian) marginal likelihood of the data. We introduce and examine two versions of the importance sampling estimator of Perrakis et al. (2014). The results are compared with two versions of the Laplace approximation: the original one and an MCMC based approximation. We illustrate and compare the proposed methodology using two real data sets (one sparse and one with full cell frequencies) and an extended simulation study. In the simulation study, we further examine the model selection consistency of the proposed power-prior based methodology.

1.12 Concluding remarks

In this chapter, we have presented, reviewed and implemented Bayesian independence hypothesis tests in two-way contingency tables. We have also explored the sensitivity of Bayes factor and model selection on prior variances of model parameters. We have presented Bayesian model comparisons based both on conjugate and log-linear analysis. In the next chapters, we will focus on developing compatible priors for model comparison in two-way tables. The use of power-priors will be used for this reason. Moreover, we will work for the construction of efficient algorithms for the computation of the marginal likelihood both in small and large scale problems. The mathematical properties of the Bayes factor for the comparison of association models will also be explored.

Chapter 2

Bayesian Methodology for Association Models

> You are the salt of the earth. But remember that salt is useful when in association, but useless in isolation.
>
> *Israelmore Ayivor*

2.1 Background Information

In contingency tables, fitting a Poisson log-linear model is the standard way to analyse the association and the interaction patterns between categorical factors. Although log-linear models are useful to describe associations of conditional independence in multi-way contingency tables, in the case of two categorical variables involved only two models can be considered: (a) the parsimonious but restrictive model of *independence* and (b) the *saturated* model. The first implies independence between the two variables while the latter does not impose any structure on the underlying association. Therefore, for the saturated model, the number of parameters is equal to the number of cells of the contingency table. In addition to standard log-linear models, the association models, introduced by

Goodman (1979) in their present form; earlier related results by other authors [see 1 df test of Tukey (1949), Nelder and Wedderburn (1972), Simon (1974), Haberman (1974), fill the gap between these two cases by imposing a specific structure on the local associations of the contingency table. This can be achieved by writing the interaction term as multiplicative function of scores of the row and the column levels. This way considers dependence between the variables under consideration and, on the same time, reduces the number of the interaction parameters compared to the saturated model by imposing specific structure. Therefore association models can be considered as in termediate models of dependence. Moreover, they may be used to analyse the associations between ordinal classification variables since they attribute (and estimate) scores for each level of the categorical variables.

The general model of association, where both column and row scores are parameters to be estimated, is called Row-Column association model (RC). The RC model does not require ordinality for any of the classification variables and it is invariant to re-ordering of columns or rows. Moreover, special cases of this model can be obtained if we consider fixed/predefined scores for row and/or column scores. To be more precise, if the scores are fixed for the categories of all row and column variables, the model is called linear-by-linear association model (LL) and has one additional parameter to the independence model. The most characteristic LL model is the Uniform (U), where the scores are equidistant for the successive categories. This model requires both classification variables to be ordinal, as this model is sensitive in re-ordering of the rows or columns. When the scores of the column variables are fixed but the scores of the row variable are parameters under estimation, the model is called Row (R) effect association model. The R model is invariant to re-ordering of the rows of the table and the corresponding classification variable need not be necessarily ordinal. Equivalently, the Column-effect association model (C) considers column scores as parameters that need to be estimated while the row scores are fixed. For a comprehensive and detailed description of the association models, see in Kateri (2014, Chapter 6).

Concerning the Bayesian analysis of association models, the simple U model has been considered by Agresti and Chuang (1989). They imposed a Dirichlet prior distribution on the probability table for the component means in the U model. Alternatively to the conjugate prior-type analysis, they proposed the Bayesian log-linear analysis by considering independent uniform (improper) priors for the main effect parameters and normal priors for the interaction parameters. The first attempt for fitting the RC association model within the Bayesian framework was by Chuang (1982). He used independent uniform (improper) priors on the main effect parameters and normal priors on the parametric row and column scores and proceeded with empirical variance estimation. Evans et al. (1993) based their analysis on the Bayesian estimation of the saturated log-linear model with normal priors on all parameters and then obtained posterior estimates for the RC model by considering the Euclidean projection from the posterior of the saturated log-linear model. Albert (1997) provided an interesting Bayesian approach for testing the fit of simple models such as independence and uniform association models.

Kateri et al. (2005a) developed the Bayesian inference for the more general $RC(M)$ association model. Iliopoulos et al. (2007) introduced an approach for identifying possible score equalities for association models with order-constrained parametric scores. This approach was based on calculating the posterior probabilities of possible order violations for successive categories in the unrestricted model. Tarantola et al. (2008) adopted a methodology from product partition models to make inferences about clustering of scores in the row effect model. Iliopoulos et al. (2009) focused on the estimation of posterior model probabilities of the RC order-constrained model, in a full Bayesian way, by allowing for ties in the prior distribution level. They constructed a *trans*-dimensional MCMC algorithm (reversible jump MCMC; Green, 1995) for assessing the equality of successive row and column scores. For two-group comparison of an ordinal scale, Kateri and

Agresti (2013) discussed stochastic orderings, based on generalized odds ratios for ordinal responses for $2 \times I$ contingency tables, from the Bayesian point of view.

Demirhan (2013) proposed a prior setup for the Bayesian estimation of association models with and without order restrictions on scores. He used a previously introduced multivariate prior in the unrestricted case and an order statistics approach in the order-restricted case. Specifically, he reformulated the approach of Chen and Dunson (2003) to decompose the covariance matrix to reflect the degree of belief in prior knowledge for scores and model parameters. His approach is composed of three steps: a) decompose the covariance matrix of scores using the Cholesky decomposition, b) write the Pearson correlation coefficient in terms of the decomposed elements and c) induce a uniform prior for each correlation coefficient. When there is no order restriction on the scores, a generalized multivariate log-gamma prior is chosen for the scores and independent log-gamma priors are chosen for the main effect parameters of the association model. In case of order restrictions on scores, he used the joint probability density function (pdf) of order statistics and assumed the independence of row to column scores. Finally, he adapted the approach of Chib and Jeliazkov (2001) for the calculation of the Bayes factor for model comparison. In their approach, for the calculation of marginal likelihood using multi-block bridge sampling estimators, the output of the Metropolis–Hastings algorithm is used directly,and the required Bayes factors are easily obtained at the end of each run of the algorithm.

More recently, Oh (2014) suggested a Bayesian model selection procedure, that simultaneously compares all possible combinations of equalities of successive score parameters in the order restricted RC association model. This method introduces normal latent variables into the model and uses an approximation to the likelihood function so the full conditional posterior distributions of elements of the parameter are given as truncated normal distributions. The basic idea of his approach is based on data augmentation by introducing appropriate normal latent variables which, in combination with an approximation of the cumulative distribution function of the Poisson, results in obtaining all the full conditional posterior densities such as the truncated normal distributions. Given their convenient form, the generation of the posterior values and the estimation of the Bayes factor can be conducted in an automated way. The Gibbs sampling algorithm of Gelfand and Smith (1990) is employed to generate posterior samples from the full model in which all the scores are strictly ordered. Finally, the formulation of the method of Oh (1999) allows the calculation of the Bayes factors of the models under comparison by using the Savage-Dickey density ratio which requires only one set of posterior samples from the full model.

2.2 Contribution and Aim of this Chapter

Although, Bayesian bibliography has treated several aspects of association models (including model comparison issues), it lacks simple, practical Bayes approaches, which incorporate recent findings in objective Bayes model comparisons. Hence, the aim of this chapter is dual: (a) to propose a default, objective method for Bayesian comparison of association models in contingency tables when no prior information is available, and (b) to introduce and study reliable and efficient MCMC based estimators of the marginal likelihood required in order to compare models or to implement Bayesian model averaging.

Regarding (a), we propose a comprehensive Bayesian model comparison approach for association models in contingency tables, which is based on imaginary data and the use of power prior approach. In Bayesian model comparison and testing the choice of the prior distribution is of paramount importance due to the well-known sensitivity of the posterior model odds and the Barlett-Lindley paradox (Bartlett, 1957, Lindley, 1957). Using prior distributions with large prior variance leads to the selection of the simplest models regardless of the data we may have. This is also the case regarding Bayes factors which cannot be calculated due to the undetermined normalizing constants involved even when improper priors have been used. For this reason, within the context of association models for contingency tables, we need to develop objective Bayes approaches (O'Bayes for short) and default priors for model comparisons. Following Consonni et al. (2018), we select one of the main tools for constructing O'Bayes model comparisons which is the power prior approach combined with the use of imaginary data. In order to apply this method, we produce a set of imaginary data from the most parsimonious model. Thus, we satisfy the "group invariance" (or locality) criterion, which is one of the main criteria for building sensible priors for O'Bayes model comparisons (Bayarri et al., 2012) and can be interpreted as centering our prior to the alternative/simpler models. Then, the prior distribution can be obtained by the product of the likelihood of the model under consideration evaluated at the imaginary data and raised to a power and then multiplied by (a relatively flat) pre-prior distribution. Under this perspective, we also quest for model selection consistency, which is one of the fundamental criteria of Bayarri et al. (2012), and whether it is satisfied via an extensive simulation study.

Concerning (b), i.e. the MCMC estimators of the marginal likelihoods, we introduce and examine two versions of the importance sampling estimator of Perrakis et al. (2014), namely the independent and the one-block estimator. The results are compared with the original Laplace approximation and the MCMC based one. Results have shown that the one-block importance estimator works fast and efficiently even in sparse contingency tables when competitors may fail.

In the next sections, we will proceed as follows. In the Section 2.3, we introduce the general model formulation of association models and the candidate models under consideration for comparison or selection along with some basic theory about Bayes model comparisons, Bayes factors and posterior model probabilities and their sensitivity on prior probabilities. In the Section 2.6, we present the prior specification based on the power priors and their advantages. The computational estimators and methods of the marginal likelihood are described in the Section 2.4. The Section 2.6.1 provides a step-by-step description of the proposed methodology. The Chapter concludes with a short discussion and the closing remarks in the Section 2.7.

2.3 Model Formulation

Association models are describing the structure of the association between the two categorical variables assigning scores to the classification variables, which can be either fixed and prespecified or unknown (to be estimated) parameters. We denote by $\{\mu_1, .., \mu_I\}$ the row scores and by $\{\nu_1, .., \nu_I\}$ the column scores. In the association models (cf. Goodman, 1985), the interaction term is now written as $\lambda_{ij}^{XY} = \phi\mu_i\nu_j$ and therefore the linear predictor (1.14) is substituted by

$$\log(\lambda_{ij}) = \lambda_0 + \lambda_i^X + \lambda_j^Y + \phi\mu_i\nu_j \tag{2.1}$$

for $i = 1, \ldots, I$ and $j = 1, \ldots, J$, where μ_i and ν_j are the i-th row and j-th column scores. For identifiability purposes, the sum-to-zero constraints are imposed on row and column main effects

$$\sum_{i=1}^{I} \lambda_i^X = \sum_{j=1}^{J} \lambda_j^Y = 0, \tag{2.2}$$

while the row and column scores are usually standardized

$$\sum_{i=1}^{I} \mu_i = \sum_{j=1}^{J} \nu_j = 0 \quad \text{and} \quad \sum_{i=1}^{I} \mu_i^2 = \sum_{j=1}^{J} \nu_j^2 = 1.$$

Under this constraints ϕ is given by

$$\phi = \sum_{i,j} \mu_i \nu_j \log \lambda_{ij}.$$

The parameter ϕ is a global measure of association under certain parametrization and measures the correlation between row and column scores, called intrinsic association parameter. Whenever a set of scores is unknown, phi is redundant. For a comprehensive and detailed description of association models, along with their frequentists implementation in R, *see* Kateri (2014).

There are three types of association models, depending on the type of the row and column parameter score:

- The linear-by-linear association model (LL) with fixed row and column scores. The most characteristic LL model is the Uniform association model (U), in case the scores are equidistant for the successive category.

- The Row association model (R) with unknown row and fixed column scores and Column effect association model (C) with unknown column and fixed row scores.

- The Row-Colum association model (RC) with both row and column scores to be parameters under estimation.

Iliopoulos et al. (2009) adopted alternative parametrizations for the RC model $\mu_1 = \nu_1 = 0$ and $\mu_I = \nu_J = 1$, which fix the scores of the first and last level of each classification variable (under ordinality constraints this corresponds to the minimum and maximum score). By this parametrization, these score parameters are fixed and prespecified.

In this thesis, we generally follow the parametrization of Iliopoulos et al. (2009) with the exception of parameter ϕ, where we set it equal to one for all models where the row or column scores are stochastic, to-be-estimated, parameters (i.e. for models R, C and RC). As a consequence, we leave the last row score in the R

model and the last column score in C and RC models to be free i.e. unconstrained parameters under estimation. With this parametrization we avoid problems of auto-correlation which appear in the implemented MCMC algorithms due to the multiplicative structure of the interaction term. Moreover, this approach is in compliance with the earlier work of Iliopoulos et al. (2007).

2.3.1 Models Under Consideration

Model selection techniques are used to identify which model fits best the data and provides accurate estimates of the quantities of interest. In this paper, we consider the six candidate models for describing the association between two categorical variables. The models under consideration are the following:

- Independence model (I): It is the simplest model and is specified by Equation 1.15. The parameter vector is given by $\boldsymbol{\vartheta}_I = \left(\lambda_0,\ \lambda^{\mathbb{X}},\ \lambda^{\mathbb{Y}}\right)$, where $\lambda^{\mathbb{X}} = \left(\lambda_1^X,\ \ldots,\ \lambda_I^X\right)$ and $\lambda^{\mathbb{Y}} = \left(\lambda_1^Y,\ \ldots,\ \lambda_J^Y\right)$. For the main effects, we impose sum-to-zero (STZ) constraints (see Eq. 2.2) and this is the case also for the remaining models.

- Uniform (U): It is specified by the log-expected frequencies of Equation 2.1. The parameters vector is now given by $\boldsymbol{\vartheta}_U = \left(\lambda_0,\ \lambda^{\mathbb{X}},\ \lambda^{\mathbb{Y}},\ \phi\right)$. The row and column scores are equidistant for the successive categories, $\boldsymbol{\mu} = (\mu_i = i,\ i = 1, \ldots, I)$ and $\boldsymbol{\nu} = (\nu_j = j,\ j = 1, \ldots, J)$. This model assumes that both X and Y are ordinal, while the additional parameter ϕ is equal to the log-odds ratio of the 2×2 contingency sub-table obtained by successive categories of both ordinal factors under consideration.

- Row effect association model (R): This model is specified by Equation 2.1 with parameters $\boldsymbol{\vartheta}_R = \left(\lambda_0,\ \lambda^{\mathbb{X}},\ \lambda^{\mathbb{Y}},\ \mu\right)$. We generally adopt the parametrization of Iliopoulos et al. (2007) with $\phi = 1$ and $\mu_1 = 0$ for the row scores and as a consequence the model R will have $I - 2$ additional parameters than model

U, corresponding to the row score. Ordinality is required only for columns, $\boldsymbol{\nu} = \{\nu_j = j,\ j = 1, \dots, J\}$.

- Column effect association (C): It is defined as analogous to the R association model by using expression 2.1 and interchanging the role of row and column scores which are now fixed and unknown (to-be-estimated) parameters, respectively.

- Row-Column association model (RC): This is the more general association model and it is also characterized by Equation 2.1 but with both row and column scores to be parameters under estimation, i.e. $\boldsymbol{\vartheta}_{RC} = \Big(\lambda_0,\ \lambda^{\mathbb{X}},\ \lambda^{\mathbb{Y}},\ \mu,\ \nu\Big)$. We adopt the parametrization of Iliopoulos et al. (2007) with $\phi = 1$, $\mu_1 = \nu_1 = 0$ and $\nu_J = 1$ for the row and column scores.

- Saturated model (S): This model is the most complex model which assumes no structure in terms of association between X and Y. It has as many parameters as the number of cells, that is IJ. Model S is given by Equation 1.14. In the set of parameters of the independence model, we further consider the interaction parameters λ_{ij}^{XY} resulting in a model parameter vector given by $\boldsymbol{\vartheta}_S = \Big(\lambda_0,\ \lambda^{\mathbb{X}},\ \lambda^{\mathbb{Y}},\ \lambda^{\mathbb{XY}}\Big)$. Additionally to the sum-to-zero constraints for the main effects, we impose STZ contraints also for the interaction parameters, i.e. $\lambda_{1j}^{XY} = -\sum\limits_{i=2}^{J} \lambda_{ij}^{XY}$ for all $j = 1, \dots, J$ and $\lambda_{i1}^{XY} = -\sum\limits_{j=2}^{I} \lambda_{ij}^{XY}$ for all $i = 1, \dots, I$.

All the models under consideration are members of the log-linear model family except of the RC model which is log-multiplicative, since the predictor is a multiplicative function in the row and column parameters μ_i and ν_j. When one set of parameter scores is fixed, the RC model simplifies to the R or C model, for which the log-expected value is a linear function of its parameters. We denote with $M \in \mathcal{M} = \{I, U, R, C, RC, S\}$ the set with all candidate models for model selection. The problem of model selection within the framework of association models is directly connected to the analysis of the association of categorical factors in a contingency table and it is strongly based on the interconnection between the models since $I \subset U \subset R$ *or* $C \subset RC \subset S$.

Estimation of parameters of the RC model, which is not linear in its parameters, can be arduous since the likelihood may not be concave and therefore multiple local maxima may exist. For this reason, MCMC methods can be used to explore the whole posterior space. Goodman (1979) suggested an iterative model-fitting algorithm. An empirical Bayes analysis of the RC model has been considered by Chuang (1982) and Evans et al. (1993) as it mentioned in Section 2.1. In their approach, they set a prior distribution for parameters of the saturated log-linear model and then they estimated the posterior distribution of the parameters of the RC model through minimization of the Euclidean squared distance between the interaction terms of the two models. Kateri et al. (2005b) provided Bayesian inference for the general association model.

2.4 Marginal Likelihood Computation

Historically, a barrier for establishment of the Bayesian approach as a standard way to analyse data has been the difficulty in the computation of the posterior distribution when the prior is not conjugate. This problem has been solved in the early 90s with the development of MCMC methods for sampling from the posterior distribution of interest. Nevertheless, the computation of the marginal likelihood remains cumbersome; this provides a motivation to researchers to develop alternative estimation methods. Bayesian model comparison via the Bayes factor, posterior model probabilities and odds (Kass and Raftery, 1995), requires the computation of the Bayesian marginal likelihood given by (**??**), where, in this context, M is one of the six candidate models, that take part in model comparison, $M \in \mathcal{M} = \{I, U, R, C, RC, S\}$.

Many methods have been introduced to compute the marginal likelihood, but simplicity is not the strongest point of most of these methods. Another disadvantage is that the proposed methods are not universal in the sense that they work under certain assumptions/requirements which are easily checked and therefore, there are cases where they fail to provide reliable estimates of the marginal likelihood. Though methods exist to directly compute the Bayes factor or the posterior odds (see e.g. Dickey (1971) and Verdinelli and Wasserman (1995) for the (generalized) Savage-Dickey density ratio), computing directly the marginal likelihood is conceptually the simplest approach. Only in very special cases, most notably for the exponential likelihood with conjugate priors, the marginal likelihood can be calculated analytically as the integrating constant of the posterior kernel. Usually, the marginal likelihoods are estimated by using MCMC estimators or asymptotic methods which are relatively easy to implement in comparison to other alternatives such as the trans-dimensional MCMC approaches (e.g. **?**). Nevertheless, they are rather computationally costly since they demand to run multiple MCMC algorithms, i.e. one for every model under consideration. This fact makes their implementation rather difficult especially when the model space under consideration is large. Moreover, all these methods require a considerable amount of fine tuning and adaptation to each problem at hand and may fail when the posteriors are complicated with many modes or extreme asymmetries.

Historically, the required calculation of the integration to estimate the marginal likelihood has been achieved by taking advantage of the conjugacy, by assuming an approximate posterior normality or by using numerical quadrature (i.e. the Laplace method or Monte Carlo integration); see Kass and Raftery, 1995. Recently, it has become possible to estimate a wider range of models, using posterior simulation methods such as the Monte Carlo sampling methods to avoid the analytical computation of the marginal likelihood (Neal, 2000, Perrakis et al., 2014). Lartillot and Philippe (2006a) introduced a technique called thermodynamic integration to approximate the marginal likelihood. More recently, a similar method called stepping-stone-sampling (Fan et al., 2011, Xie et al., 2011), has been proposed (see also Baele et al. (2013), Baele and Lemey (2013), Friel et al. (2014), for a summary and comparison of these methods).

The previous ways of calculating integrated likelihoods cannot be often used for models estimated via MCMC or their posterior simulation methods. A standard method to approximate the marginal likelihood is the Laplace approximation (Tierney and Kadane, 1986) and its MCMC based version (Lewis and Raftery, 1997b). This approach is based on assuming a multivariate normal approximation of the posterior and works reasonably well for moderately large datasets and symmetric, well behaved posterior distributions. A popular alternative is to use MCMC based estimators where there is a wide variety of methods with the most notable the harmonic-mean and the prior/posterior mixture importance sampling estimators (Newton and Raftery, 1994) as well as the more recent stabilized version of it (Raftery et al., 2007), the bridge-sampling methods (Meng and Wong, 1996), the candidate's estimators for Gibbs sampling (Chib, 1995) and Metropolis–Hastings sampling (Chib and Jeliazkov, 2001), and the power posterior method (Friel and Pettitt, 2008, Lartillot and Philippe, 2006b). For a detailed review and comparison of methods we refer to Friel and Wyse (2012) and Ardia et al. (2012). In this thesis two methods for estimating the marginal likelihood, based on the Laplace approximation method are presented. The one is the Laplace approximation approach and the other is the MCMC based version of it (namely Laplace-Metropolis estimator). In the following section we introduce two versions of the importance sampling marginal estimator of Perrakis et al. (2014): the independent and the one-block importance sampling estimators. Laplace approximation and Laplace-Metropolis estimator will be used as the gold-standards for our comparisons presented in the illustrative examples.

2.4.1 Laplace Approximation

One of the most popular methods for the computation of the marginal likelihood is the Laplace approximation based on the normal distribution. The Laplace method has been used from the late 80s for approximating Bayes factors in GLMs. It first appeared in Raftery (1988) and later Leonard et al. (1989) facilitated Laplace approximations for evaluating the posterior of summary measures of interest in contingency tables. The Laplace approximation for an integral of the form $\int e^{h(u)}du$ is found using a Taylor series expansion of a real-valued function of a P-dimensional vector u. Rosenkranz (1992) reported that the Laplace method produces much more accurate estimates of the marginal likelihood than posterior simulation for a variety of models. Raftery (1996a) employed the method to obtain approximate Bayes factors for GLMs. Under this approach, the marginal likelihood for contingency tables can be approximated by

$$\log f\left(\boldsymbol{n}|M\right) \approx \frac{d_M}{2}\log(2\pi) + \frac{1}{2}\log \mid H^* \mid + \log f\left(\boldsymbol{\vartheta}_M^* \mid M\right) + \log f\left(n \mid \boldsymbol{\vartheta}_M^*, M\right),$$

where$\boldsymbol{\vartheta}_M^*$is the posterior mode of model M, and H^* is minus the inverse Hessian evaluated at the posterior mode. This approximation works efficiently when the posterior distribution is relatively symmetric and unimodal, *see* Tierney and Kadane (1986).

2.4.2 Laplace-Metropolis Estimator

Lewis and Raftery (1997a) describe a way to use the posterior simulation output to estimate integrated likelihoods. The Laplace method is often not applicable because it requires the evaluation of derivatives that are not easily available. This is particularly true for complex models for which posterior simulation, especially MCMC, is often used. The idea of the Laplace-Metropolis estimator is to overcome the limitations of the Laplace method by using posterior simulation to estimate the quantities needed. To avoid analytic calculation of H^* and $\boldsymbol{\vartheta}^*_M$ in Laplace-Metropolis estimator, we estimate them by the posterior mean $\bar{\boldsymbol{\vartheta}}_M$ and the posterior variance-covariance matrix of $\boldsymbol{\vartheta}$, respectively, obtained by the MCMC output. Hence, under this approach, the marginal likelihood is given by

$$\log f\left(\mathbf{n}|M\right) \approx \hat{f}(\mathbf{n}|M),$$
$$\hat{f}(\mathbf{n}|M) = \frac{d_M}{2}\log(2\pi) + \frac{1}{2}\log \mid \boldsymbol{H}^* \mid + \log f\left(\bar{\boldsymbol{\vartheta}}_M \mid M\right) + \log f\left(\mathbf{n} \mid \bar{\boldsymbol{\vartheta}}_M, M\right).$$

Although, the choice of the posterior mean for $\boldsymbol{\vartheta}^*_M$ is the most popular choice, it may be problematic when the actual posterior departs from symmetry and being unimodal. Hence, there are several alternative ways of estimating $\boldsymbol{\vartheta}^*_M$ an MCMC sample:

- Estimate $\boldsymbol{\vartheta}^*_M$ as that $\boldsymbol{\vartheta}_M$ in the sample at which $h(\boldsymbol{\vartheta}_M) = f(\boldsymbol{n}|\boldsymbol{\vartheta}_M, M)f(\boldsymbol{\vartheta}_M|M)$ achieves its maximum.

- Estimate the components of $\boldsymbol{\vartheta}^*_M$ by finding the componentwise posterior means.

- Estimate the components of $\boldsymbol{\vartheta}^*_M$ by finding the componentwise posterior medians.

- Estimate $\boldsymbol{\vartheta}^*_M$ by finding the multivariate median.

The first of these methods is the simplest conceptually and usually the most accurate. However, it involves calculating the likelihood many times and therefore it might be computationally expensive. In such cases, the multivariate median might be used instead, as this does not require too many computing resources. Moreover, the MCMC estimated posterior median is more robust than the corresponding posterior mean, which is influenced by outliers. Furthermore, the median is a more accurate proxy of the model than the mean for a wide range of distributions (Johnson and Kotz, 1985).

The other quantity required for the calculation of the Laplace-Metropolis estimator is H^*. This is asymptotically equal to the posterior variance matrix, and we could estimate it by the sample covariance matrix of the posterior simulation output, $\boldsymbol{S}_M = \dfrac{1}{T-1}\sum_{t=1}^{T}(\boldsymbol{\vartheta}_M^{(t)} - \bar{\boldsymbol{\vartheta}}_M)(\boldsymbol{\vartheta}_M^{(t)} - \bar{\boldsymbol{\vartheta}}_M)^T$. However, because MCMC trajectories take occasional distant excursions, it is recommended to use a robust estimator of the posterior variance matrix. One such estimator is the weighted variance matrix estimate with weights based on the minimum volume ellipsoid estimate of Rousseeuw and van Zomeren (1990).

2.4.3 Common and Different Characteristics of Laplace approaches

The Laplace approximation is obtained by using any optimization method, while the Laplace-Metropolis estimator is based on the output of an MCMC algorithm. The first method is faster and free of any Monte Carlo error, but it is not always applicable due to computational problems and not being unimodal. The second approach is more computationally demanding, but the implementation is direct when the MCMC sample is available. Both estimators are approximations of the Bayesian marginal likelihood and therefore they require a large enough sample size in order to obtain accurate results. Therefore, they will not be accurate for small or sparse datasets. Both approaches have specific regularity conditions. These approximations are more efficient when the posterior distributions are symmetric, otherwise log, logit or Box-Cox transformations may be applied in order to convert the Laplace approximation on transformations of parameters of interest that are a-posteriori normally distributed.

2.4.4 Illustrated Example

In this section, we use the data of a survey conducted at the University of Ioannina (Greece) in 1995 which studies the association between the use of cannabis and the alcohol consumption among students (Marselos et al., 1997). The frequency of alcohol consumption is measured in a four-level scale while the use of cannabis through a three-level factor (never tried, once or twice, more frequently). These two ordinal factors are cross-classified in a 4×3 table presented in Table 2.1.

Table 2.1: Cannabis dataset (Marselos, 1997).

	Cannabis Use			
Alcochol consumption	Never	Once or twice	More often	Total
At most once/month	204	6	1	211
Twice/month	211	13	5	229
Twice/week	357	44	38	439
More often	92	34	49	175
Total	864	97	93	1054

Using this example, we illustrate the Bayesian model comparison of association models for two-way contingency tables by approximating all marginal likelihoods using the Laplace-Metropolis estimator. Two normal prior distributions with zero means were used: (a) Unit Information Empirical prior (UIE) where the prior variance of each interaction parameter is set equal to the corresponding posterior variance multiplied by the sample size, and (b) a prior similar to the one proposed by Dellaportas and Forster (1999) with prior variance equal to two (DF).

Table 2.2: Bayesian measures of model evaluation for all models $M \in \mathcal{M}$ in the Cannabis dataset.

$\boldsymbol{M_j}$	**Model**	**DIC**	**BIC**	**Dif. BIC**	$\boldsymbol{-2logBF_{j2}}$		**Posterior Probabilities**		
					$\boldsymbol{UIE^{\star}}$	$\boldsymbol{DF^{\dagger}}$	**BIC**	$\boldsymbol{UIE^{\star}}$	$\boldsymbol{DF^{\dagger}}$
1	Independence (I)	228.9	258.7	144.1	139.4	145.5	<0.01	<0.01	<0.01
2	Uniform (U)	81.1	114.8	0.0	0.0	0.0	0.956	0.973	0.801
3	Row (R)	82.8	128.1	13.4	17.9	9.1	<0.01	<0.01	<0.01
4	Column (C)	81.5	120.9	6.2	7.2	2.9	0.043	0.027	0.191
5	Row-Column (RC)	84.9	134.4	19.6	21.8	20.1	<0.01	<0.01	<0.01
6	Saturated (S)	88.2	147.7	32.9	40.4	20.5	<0.01	<0.01	<0.01

**UIE: Unit Information Empirical prior*

†*DF: Dellaportas and Foster(1999) prior*

Table 2.2 presents several measures (DIC, BIC, logBF and posterior probabilities under the two priors discussed previously) for the six models under consideration (I, U, R, C, RC, S). All model comparison measures indicate that the Uniform model is the best, which has equidistant scores for the rows and columns, $\boldsymbol{\mu} = \{1,\ 2,\ 3\}$ and $\boldsymbol{\nu} = \{1,\ 2,\ 3,\ 4\}$. From Table 2.2 we observe that the DIC and BIC values for this model are clearly lower than those of other models. Similar is the picture from the posterior model probabilities where the U model is the Maximum a-posteriori model (MAP) with probabilities 0.956, 0.973 and 0.801 for BIC, UIE prior and DF prior, respectively. Results under the two different prior setups differ due to the effect of the Lindley-Bartlett paradox. The posterior model probabilities under the empirical prior are closer to the corresponding BIC based probabilities due to the approximate unit information interpretation of UIE. The column effect association model appears to be the second best but with considerably lower probabilities (0.043, 0.027 and 0.191 for BIC, UIE prior and DF prior, respectively). These results are in agreement with the frequentists analysis (see Kateri, 2014, Chapter 6).

2.5 Importance sampling marginal likelihood estimators

Perrakis et al. (2014) proposed an importance sampling estimator of the marginal likelihood in which the product of the marginal posterior distributions is used as an importance sampling function. This approach is generally applicable to multivariate parameter vector $\boldsymbol{\vartheta}$ that can be split in multiple (multivariate) blocks of parameters $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_{d_\vartheta})$ (multi-block parameter vector settings). It does not require additional Markov Chain Monte Carlo (MCMC) sampling and does not depend on the type of MCMC scheme used to sample from the posterior. It can be applied in a wide range of models including regression models, finite normal mixtures and longitudinal Poisson models. It leads to accurate marginal likelihood estimates provided that the importance sampling distribution used captures the main characteristics of the posterior such as asymmetries and correlations.

In this thesis, we propose two estimators based on the original estimator Perrakis et al. (2014): the independent and the one-block Perrakis estimators. In the first choice $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_{d_\vartheta})$ becomes univariate scalar while in the second $\boldsymbol{\vartheta}$ is considered as multivariate block where a multivariate approximation of the posterior is used as proposal. A critical feature in differentiating the independent Perrakis estimator (IP) from the one-block (OBP) version is the distribution of the importance sampling function. In the first case (IP), the importance sampling function g is simply a product of independent normal distributions, as opposed to one-block variant where g is a multivariate normal distribution. The two variants of the estimator are used here in the context of association models in Bayesian contingency tables analysis.

The marginal likelihood estimator of Perrakis et al. (2014) is based on the use of a proper importance sampling density g which is introduced by the marginal likelihood given by (**??**). Hence, the marginal likelihood can be expressed as an expectation with respect to g (instead of the prior) given by

$$f(\boldsymbol{n}|M) = \int \frac{f(\boldsymbol{n}|\boldsymbol{\vartheta}_M, M) f(\boldsymbol{\vartheta}_M|M)}{g(\boldsymbol{\vartheta}_M|M)} g(\boldsymbol{\vartheta}_M|M) d\boldsymbol{\vartheta}_M = E_g \left[\frac{f(\boldsymbol{n}|\boldsymbol{\vartheta}_M, M) f(\boldsymbol{\vartheta}_M|M)}{g(\boldsymbol{\vartheta}_M|M)} \right]. \tag{2.3}$$

This quantity can be easily estimated by

$$\widehat{f}(\boldsymbol{n}|M) = \frac{1}{T} \sum_{n=1}^{N} \frac{f(\boldsymbol{n}|\boldsymbol{\vartheta}_M^{(t)}, M) f(\boldsymbol{\vartheta}_M^{(t)}|M)}{g(\boldsymbol{\vartheta}_M^{(t)}|M)} \tag{2.4}$$

where T is the number of randomly generated values of $\boldsymbol{\vartheta}_M$ from $g(\boldsymbol{\vartheta}_M|M)$ which will be called importance sampling size. Each generated value is denoted by $\boldsymbol{\vartheta}_M^{(t)}$, for $t = 1, 2, \ldots, T$. An ideal importance sampling density should provide easiness in sampling from it and it should be close to the posterior distribution.

A key issue of the above method is to select appropriately the importance function $g(\boldsymbol{\vartheta}_M|M)$ which should be easy-to-generate from the distribution. Such distributions can be built by estimating posterior summaries from the MCMC output and by selecting known distributions which match the marginal posterior distributions of interest. Another critical point, is the selection of the dimensional complexity of g. For models of high dimension, the selection of a multivariate g that is a good proxy of the target posterior may be difficult to be constructed. So we proceed by identifying blocks of parameters $\boldsymbol{\vartheta}_M = (\boldsymbol{\vartheta}_{b_1,M}, \boldsymbol{\vartheta}_{b_2,M}, \ldots, \boldsymbol{\vartheta}_{b_B,M})$ with $\text{Cor}(\boldsymbol{\vartheta}_{b_l,M}, \boldsymbol{\vartheta}_{b_\ell,M}) \approx 0$ for every $l \neq \ell \in \{1, 2, \ldots, B\}$; where b_ℓ specify the different blocks of parameters and B the number of selected blocks. Under this approach we re-write the density of the importance function as

$$g(\boldsymbol{\vartheta}_M|M) = \prod_{\ell=1}^{B} g_\ell(\boldsymbol{\vartheta}_{b_\ell,M}|M)$$

and we generate each parameter block from

$$\boldsymbol{\vartheta}_{b_\ell,M} \sim g_\ell(\boldsymbol{\vartheta}_{b_\ell,M}|M) \text{ for } \ell = 1, \ldots, B.$$

Next, we describe in detail the two special cases of the general importance sampling estimator introduced here: the independent and the one-block Perrakis estimators.

2.5.1 Independent Perrakis Estimator

In this case, we assume that each parameter is a single block, that is, $B = d_M$ with $d_M = |\boldsymbol{\vartheta}_M|$ denoting the number of parameters of model M. Although posterior independence is not frequently met in practice, the product marginal posterior can serve as a good approximation to the joint posterior even if $\boldsymbol{\vartheta}_M$ is not completely independent a-posteriori. The parameter blocks could be close to orthogonal regardless of whether the elements within $\boldsymbol{\vartheta}_M$ are strongly correlated. Furthermore, appropriate reparameterizations can be used in order to form parameter blocks which are orthogonal or close to orthogonality (see Gilks and Roberts, 1996). By using this simplified approach, the simulation of the importance sample will be accelerated since all distributions that we will sample from are univariate. Regarding association models, the underlying independence assumption of the parameters is not realistic; but we intuitively assume that all posterior correlations between parameters are not high enough to cause severe problems to the estimation of the marginal likelihood. The importance sampling function is now written as $g(\boldsymbol{\vartheta}_M|M) = \prod_{j=1}^{d_M} g_j(\vartheta_{j,M}|M)$ and each single parameter $\vartheta_{j,M}$ is generated from $g_j(\vartheta_{j,M}|M)$.

For the association models, we can obtain an estimate of the marginal likelihood using the independent Perrakis estimator by implementing the Algorithm 1. Initially, we run the MCMC algorithm. From the MCMC output we estimate the posterior mean $\tilde{\vartheta}_{j,M}$ and the posterior variance $S^2_{\vartheta_{j,M}}$ of each parameter $\vartheta_{j,M}$. We generate the importance sampling values by

$$\vartheta^{(t)}_{j,M} \sim N(\tilde{\vartheta}_{j,M}, S^2_{\vartheta_{j,M}}) \text{ for } j = 1, \ldots, d_M$$

for $t = 1, 2, \ldots, T$. Finally, we estimate the marginal likelihood using 2.4, which now becomes

$$\widehat{f}_I(\boldsymbol{n}|M) = \frac{1}{T}\sum_{t=1}^{T}\frac{f(\boldsymbol{n}|\boldsymbol{\vartheta}_M^{(t)}, M)f(\boldsymbol{\vartheta}_M^{(t)}|M)}{\prod_{j=1}^{d_M}\frac{1}{S_{\vartheta_{j,M}}}\phi\left(\frac{\vartheta_{j,M}^{(t)}-\tilde{\vartheta}_{j,M}}{S_{\vartheta_{j,M}}}\right)} \tag{2.5}$$

where $\phi(z)$ is the density function of the standardized normal distribution. See Algorithm 1 for a summary.

Algorithm 1 Independence Monte Carlo estimator algorithm

```
Input:
    MCMC output: $\boldsymbol{\vartheta_M^{(t)}} = \left(\vartheta_{j,M}^{(t)},\ j = 1, \ldots, d_M\right)$, $t = 1, \ldots, T_0$,
    $T_0 = T_{MCMC}$: the size of MCMC output
    T: the importance sample size
for $j = 1$ to $d_M$ do
    $\tilde{\vartheta}_{j,M} = \frac{1}{T_0}\sum_{t=1}^{T_0}\vartheta_{j,M}^{(t)}$                ▷ Estimate the posterior mean
    $S_{\vartheta_{j,M}}^2 = \frac{1}{T_0 - 1}\sum_{t=1}^{T_0}(\vartheta_{j,M}^{(t)} - \tilde{\vartheta}_{j,M}^{(t)})^2$    ▷ Estimate posterior variance
end for
for $t = 1$ to $T$ do
    for $j = 1$ to $d_M$ do
        $\vartheta_{j,M}^{(t)} \sim N(\tilde{\vartheta}_{j,M}, S_{\vartheta_{j,M}}^2)$           ▷ Generate importance sampling values
    end for
end for
$\widehat{f}_I(\boldsymbol{n}|M) = \frac{1}{T}\sum_{t=1}^{T}\frac{f(\boldsymbol{n}|\boldsymbol{\vartheta}_M^{(t)}, M)f(\boldsymbol{\vartheta}_M^{(t)}|M)}{\prod_{j=1}^{d_M}\frac{1}{S_{\vartheta_{j,M}}}\phi\left(\frac{\vartheta_{j,M}^{(t)}-\tilde{\vartheta}_{j,M}}{S_{\vartheta_{j,M}}}\right)}$    ▷ Calculate the marginal likelihood
Output:
    The marginal likelihood estimate: $\widehat{f}_I(\boldsymbol{n}|M)$
```

2.5.2 One-Block Perrakis Estimator

In this case, we consider all parameters as a single block (i.e. $B = 1$) and generate all importance sampling values from a single multivariate distribution. Although this will be generally inefficient for high dimensional or complex models, in our case we expect that such an estimator will behave reasonably well since the dimension of association models for two-way tables is usually limited; also empirical evidence has shown that the posterior distributions in such models are relatively symmetric and therefore they can be approximated reasonably well from multivariate normal distributions. The advantage of this approach, in comparison to the independence approach, is that the importance sampling will be definitely closer in terms of shape to the posterior distribution since it will incorporate the correlations between the parameters.

Under this approach, the procedure can be described by the steps of Algorithm 2. First we run the MCMC algorithm. From the MCMC output, we estimate the posterior mean $\widetilde{\boldsymbol{\vartheta}}_M$ and posterior variance-covariance matrix $\boldsymbol{S}_{\boldsymbol{\vartheta}_M}$. We generate the importance sampling values from a multivariate normal distribution $\boldsymbol{\vartheta}_M^{(t)} \sim N_{d_M}(\widetilde{\boldsymbol{\vartheta}}_M, \boldsymbol{S}_{\boldsymbol{\vartheta}_M})$, for $t = 1, \ldots, T$. Finally, we estimate the marginal likelihood using 2.4, which now becomes

$$\widehat{f}_{B_1}(\boldsymbol{n}|M) = \frac{1}{T}\sum_{t=1}^{T} \frac{f(\boldsymbol{n}|\boldsymbol{\vartheta}_M^{(t)}, M) f(\boldsymbol{\vartheta}_M^{(t)}|M)}{f_{N_{d_M}}(\boldsymbol{\vartheta}_M^{(t)}; \widetilde{\boldsymbol{\vartheta}}_M, \boldsymbol{S}_{\boldsymbol{\vartheta}_M})}. \tag{2.6}$$

In the above procedure, $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is used to denote the p-dimensional normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ and $f_{N_p}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote the corresponding density function. In the following of this thesis, we will refer to this approach as the one-block Perrakis estimator. See Algorithm 2 for a summary.

Algorithm 2 One-block Monte Carlo estimator algorithm

```
Input:
    MCMC output: $\boldsymbol{\vartheta}_{\boldsymbol{M}}^{(\boldsymbol{t})} = \left(\vartheta_{j,M}^{(t)},\ j = 1, \ldots, d_M\right),\ t = 1, \ldots, T_0,$
    $T_0 = T_{MCMC}$: the size of MCMC output
    T: the importance sample size
for $j = 1$ to $d_M$ do
    $\tilde{\vartheta}_{j,M} = \frac{1}{T_0} \sum_{t=1}^{T_0} \vartheta_{j,M}^{(t)}$        ▷ Estimate posterior mean
end for
Set: $\tilde{\boldsymbol{\vartheta}}_M = \left(\tilde{\vartheta}_{1,M}, \tilde{\vartheta}_{2,M}, \ldots, \tilde{\vartheta}_{d_M,M}\right)$        ▷ Posterior mean vector
$\boldsymbol{S}_{\boldsymbol{\vartheta}_M} = \frac{1}{T_0 - 1} \sum_{t=1}^{T_0} \left(\boldsymbol{\vartheta}_{\boldsymbol{M}}^{(\boldsymbol{t})} - \tilde{\boldsymbol{\vartheta}}_{\boldsymbol{M}}^{(\boldsymbol{t})}\right)\left(\boldsymbol{\vartheta}_{\boldsymbol{M}}^{(\boldsymbol{t})} - \tilde{\boldsymbol{\vartheta}}_{\boldsymbol{M}}^{(\boldsymbol{t})}\right)^\top$        ▷ Estimate posterior variance-covariance matrix
for $t = 1$ to $T$ do
    $\boldsymbol{\vartheta}_M^{(t)} \sim N_{d_M}(\tilde{\boldsymbol{\vartheta}}_M, \boldsymbol{S}_{\boldsymbol{\vartheta}_M})$        ▷ Generate importance sampling values
end for
$\widehat{f}_{B_1}(\boldsymbol{n}|M) = \frac{1}{T} \sum_{t=1}^{T} \frac{f(\boldsymbol{n}|\boldsymbol{\vartheta}_M^{(t)}, M) f(\boldsymbol{\vartheta}_M^{(t)}|M)}{f_{N_{d_M}}(\boldsymbol{\vartheta}_M^{(t)}; \tilde{\boldsymbol{\vartheta}}_M, \boldsymbol{S}_{\boldsymbol{\vartheta}_M})}$        ▷ Estimate marginal likelihood
Output:
    The marginal likelihood estimate: $\widehat{f}_{B_1}(\boldsymbol{n}|M)$
```

2.6 Prior Specification via Power Priors and Imaginary Data

Posterior model probabilities and the Bayes factor are highly sensitive to the prior specification of the model parameters. This behavior is known by the name of Lindley's and Barlett paradox, respectively. Hence, the Bayes factor is quite sensitive to the choices of hyperparameters of vague proper priors, and thus one cannot simply specify vague proper priors in model selection contexts to avoid informative prior elicitation. When no prior information is available, a non-informative prior such as a uniform prior or a Jeffreys prior can be used (see Kass and Wasserman, 1996) as possible of choices. When non informative or improper prior are used, the prior surface is flat. Moreover, such priors do not make use of real prior information, which is available for a specific application. Thus, informative priors, such as power-priors (Ibrahim and Chen, 2000), can be of great value in such situations and, more generally, in applied research settings where the researcher has access to previous studies measuring the same response and covariates. Real prior information such as historical data or data from previous similar studies are often available in applied research. Also, in experiments conducted over time, data from previous time periods can be used as prior information by quantifying it with a suitable prior distribution on the mode parameters.

Power prior distributions are based on the idea of raising the likelihood function of historical data to a power w, where $0 \leq w \leq 1$. The power prior approach of Ibrahim and Chen (2000) and Chen et al. (2000b) is adopted to advocate sensible values for the prior parameters of model used to fit contingency table data. The initial idea of the power prior can be traced back to Diaconis and Ylvisaker (1979) and Morris (1983), where they studied conjugate priors for exponential families. However, these two authors considered only the case of the power w as a fixed constant.

In this work, the idea of imaginary data is adopted in order to alleviate the effect of the Lindley-Bartlett paradox and, therefore, specify reasonable priors which will lead to a sensible Bayesian model comparison. We consider imaginary data coming from the simplest model of independence and weight them in order to obtain a contribution equivalent to one additional data point to the posterior. Thus, we indirectly avoid the use of improper priors by the use of imaginary data. The initial idea of the imaginary data can be traced in Good (1950). Spiegelhalter and Smith (1982b), based on the original ideas of Good, considered a thought experiment with an appropriate dataset that was used to specify the normalizing constants involved in the Bayes factor when using improper priors. When no information is available, a common procedure is to produce or generate imaginary data with some specific properties. Usually, we consider data coming from the simplest model (especially in nested model comparisons) in order to center our priors to the null hypothesis. By this way to satisfy the "locality" principle; see for details Consonni et al. (2018). Moreover, by adopting the unit information principle, we minimize the amount of information on our posterior by the imaginary data to a piece of information equivalent to one additional data point. The use of common "imaginary" data in building the prior distributions as (re-weighted) posteriors, ensures that they will be "compatible" across models in the sense that the same prior information is infused in all models. The term of compatibility refers to the fact that priors should be somehow related across models (see Consonni and Veronese, 2008, Dawid and Lauritzen, 2011). Finally, power-priors of unit information can help us ensure that the prior will be consistent in terms of model selection in the sense that the posterior probability of the true model will tend to one when the sample size grows. This is a desirable property which will be examined extensively here using a simulation study.

The power prior, in the general frameworkwas setup for any under this setup statistical model M with response Y, covariates $X_1, \ldots, X_p$ and parameters $\boldsymbol{\vartheta} = (\boldsymbol{\beta}, \boldsymbol{\psi})$. Chen et al. (2000a) denoted by $\boldsymbol{y}^*$ a historical data response vector of size $n_0 \times 1$ and by $\boldsymbol{X}^*$ the corresponding data/design matrix of dimension $n_0 \times p$ that corresponds to covariates $X_1, \ldots, X_p$. Also, let $\boldsymbol{x}_i^{*T} = (x_{i1}^*, x_{i2}^*, \ldots, x_{ik}^*)$ to be the i^{th} row of $\boldsymbol{X}^*$ with $x_{i1}^* = 1$ (for all $i = 1, \ldots, n$) being the elements of the first column which correspond to the intercept. Moreover, $\eta_i^* = \boldsymbol{x}_i^{*T} \boldsymbol{\beta}$ denotes the linear predictor based on the historical data, and $D^* = (\boldsymbol{y}^*,\ \boldsymbol{X}^*)$ denotes the historical data. Then a prior can be obtained as a posterior based on the historical data D^* and a flat (pre)prior can be obtained by setting $\pi(\boldsymbol{\vartheta}_M | M) = f(\boldsymbol{\vartheta}_M | D^*) \propto f(\boldsymbol{y}^* | \boldsymbol{\vartheta}_M, \boldsymbol{X}^*)$.A problem with this induced prior is that each historical data point will account for a value of information same to that of the actual observed data. Moreover, if the size of historical data is large, then the posterior including both actual and historical data will be governed by the “past ”and not by the actually observed data. The problem is even more acute, when “imaginary”data are used instead of historical data in the case of prior ignorance with the aim to build a set of sensible, data driven Bayesian procedure with a prior compatible across models. For this reason, Chen et al. (2000a) introduced a “power ”parameter which calibrates or tunes the amount of information infused in our posterior distribution. Under this approach the “power”prior is written as

$$\pi(\boldsymbol{\vartheta}|w) \propto f(\boldsymbol{y}^*|\boldsymbol{\vartheta}, \boldsymbol{X}^*)^w, \tag{2.7}$$

where $0 \leq w \leq 1$ is a scalar prior precision parameter that weights the historical data relative to the likelihood of the current study. The prior specification is complete by specifying a prior distribution or specific value for w. Imaginary data are assigned higher weight than actual observations when $w > 1$. This is generally not desirable, especially when no prior information is available. If we take $w = 1$ and n_0 equal to the observed sample size then the prior and data will account for 50% of the information used in the posterior. The choice of $w = 1/n_0$ means that the prior data $\boldsymbol{y}^*$ will account for information of one data point. Hence, we add information equivalent to adding one observation and, by construction, this will support simpler models that are closer to the uniform-cells assumption.

Returning back to the notation used for contingency tables, for a set of imaginary data $\boldsymbol{n}^*$, the power prior for a model M can be defined as

$$\pi(\boldsymbol{\vartheta}_M|w, M) \propto f(\boldsymbol{n}^*|\boldsymbol{\vartheta}_M, \boldsymbol{X})^w f_0(\boldsymbol{\vartheta}_M), \tag{2.8}$$

since $\boldsymbol{X}^* = \boldsymbol{X}$ this is a design matrix, fixed by experiment. In the above formulation, we have also added the pre-prior $f_0(\boldsymbol{\vartheta}_M)$ for generality but this can be eliminated by considering a uniform, improper prior or almost eliminated by considering a proper prior with extremely large variance. To simplify the choice of the imaginary data, we consider data from the simplest possible model which is the model of uniform cell frequencies (i.e. all cells have the same frequencies). Furthermore, in order to consider the unit information approach we set $w = 1/n^*$ where $n^* = \sum_{i=1}^{I}\sum_{j=1}^{J} n^*_{ij}$ is the total sample size of the imaginary contingency table. Thus, we specify

$$n^*_{ij} = \xi \text{ and } w = \tfrac{1}{\xi IJ}.$$

For the common value ξ of the imaginary cell frequencies, two are the "natural" choices:

Prior Choice 1: $\xi = 1$ which is the minimum value that can be used to fit a relevant model without identifiability problems.

Prior Choice 2: $\xi = \overline{n} = N/(IJ)$ which ensures that the constant of all models will be centered to the "correct" value in terms of effect size. Some may argue that this prior is (even minimally) empirical due to the use of N, but one remark against it is that we introduce information to our prior which is based only on the characteristics of the experiment (i.e. sample size and number of levels of each categorical factor) and these are fixed before observing the data. Moreover, no data driven information about the association between variables is included in our prior building procedure.

Under this approach, the marginal likelihood is now given by

$$\begin{aligned} f(\boldsymbol{n}|M) &= \int f(\boldsymbol{n}|\boldsymbol{\vartheta}_M, M)\pi(\boldsymbol{\vartheta}_M|w, M)d\boldsymbol{\vartheta}_M \\ &= \frac{\int f(\boldsymbol{n}|\boldsymbol{\vartheta}_M, M) f(\boldsymbol{n}^*|\boldsymbol{\vartheta}_M, \boldsymbol{X})^w f_0(\boldsymbol{\vartheta}_M)d\boldsymbol{\vartheta}_M}{\int f(\boldsymbol{n}^*|\boldsymbol{\vartheta}_M, \boldsymbol{X})^w df_0(\boldsymbol{\vartheta}_M)\boldsymbol{\vartheta}_M}. \end{aligned} \tag{2.9}$$

Hence, this approach requires to calculate two distinct Bayesian marginal likelihoods: one for the imaginary data and one about both the imaginary and the actual data. Note that the resulting (overall) marginal likelihood can be calculated even in the case we consider an improper pre-prior distribution. The reason is that the use of the imaginary data make the actual power prior proper provided that the imaginary data provide identifiable information for every parameter. Moreover, the unknown normalizing constant of the improper pre-prior appears in both the numerator and the denominator of (2.9) and, thus, they cancel out without creating any problem to the computation of the marginal likelihood

On the other hand, the requirement of double computations makes the above procedure unattractive. For this reason, we use a normal approximation of the posterior distribution using imaginary data and then re-weight them appropriately. The normal approximation typically works satisfactorily for GLMs; see for example in **?**. Following the arguments used in the Zellner's g-prior and its connection with the power prior (Fouskakis and Ntzoufras, 2016, **?**), we may consider an approximation as the following setup

$$\pi(\boldsymbol{\vartheta}_M|w, M) \sim N_{d_{\boldsymbol{\vartheta}_M}}(\widetilde{\boldsymbol{\vartheta}}^*_M, w^{-1}\widetilde{\boldsymbol{\Sigma}}^*_{\boldsymbol{\vartheta}_M}) \tag{2.10}$$

where $\widetilde{\boldsymbol{\vartheta}}^*_M$ and $\widetilde{\boldsymbol{\Sigma}}^*_{\boldsymbol{\vartheta}_M}$ are the posterior means and variance-covariance matrix of $\boldsymbol{\vartheta}_M$ estimated using data $\boldsymbol{n}^*$. Estimates of these values can be obtained by an MCMC of each model M using the imaginary data $\boldsymbol{n}^*$. or from the corresponding MLE estimates. In order to further simplify the approach, we consider $\widetilde{\boldsymbol{\Sigma}}^*_{\boldsymbol{\vartheta}_M}$ to be diagonal. Thus one need only to estimate the posterior variances by using the imaginary data. The effect of this simplification is minor since the actual prior information induced by our proposed prior is minimal and equal to one data point.

2.6.1 Implementation of the Proposed Methodology in Association Models

The proposed methodology for association models in contingency tables can be described by the following five steps.

- **Step 1:** For $w = 1/(IJ)$, set imaginary data $n^*_{ij} = \xi$ and with $\xi = 1$ for prior choice 1 and $\xi = N/(IJ)$ (or the closest integer) for the second prior.

- **Step 2:** Set an improper uniform prior or a normal pre-prior with very large variance.

- **Step 3:** Compute posterior means and variances of $\boldsymbol{\vartheta_M}$ for each model M (either using MCMC or any other method) using the imaginary data $\boldsymbol{n}^*$.

- **Step 4:** For the actual data, use a normal prior of the type (2.10) for the model parameters with mean and variance obtained by Step 3.

- **Step 5:** Compute the marginal densities for all M using a marginal likelihood estimator (see Section 2.4 for proposed approaches).

Algorithm 3 Bayesian model evaluation using power prior for contingency tables

```
Notation:
    Parameter vector of model $M$: $\boldsymbol{\vartheta_M}$
Input:
    Data: $\boldsymbol{n} = \Big(n_{ij},\ i = 1, \ldots, I,\ j = 1, \ldots, J\Big)$
    Imaginary data: $\boldsymbol{n^*} = \Big(n^*_{ij},\ i = 1, \ldots, I,\ j = 1, \ldots, J\Big)$
    MCMC output of parameter: $\boldsymbol{\vartheta^{(t)}_M} = \Big(\vartheta^{(t)}_{j,M},\ j = 1, \ldots, d_M\Big),\ t = 1, \ldots, T_0,$
    Size of MCMC output: $T_0 = T_{MCMC}$
    Importance sample size: $T$
    Weight of imaginary data: $w$
    Value of each $n^*_{ij}$: $\xi$ (i.e. $n^*_{ij} = \xi$)
for $M$ in $\mathcal{M} = \{I, U, R, C, RC, S\}$ do
    Specify prior type: $\boldsymbol{Prior\ 1:}\ w = \frac{1}{IJ} \leftarrow n^*_{ij} = \xi = 1$   $\boldsymbol{or}$

                        $\boldsymbol{Prior\ 2:}\ w = \frac{1}{IJ} \leftarrow n^*_{ij} = \xi = \frac{n}{(IJ)}$
    Specify pre-prior: $\boldsymbol{i.}\ \vartheta_{j,M} \propto 1\ (improper)$   $\boldsymbol{or}$

                        $\boldsymbol{ii.}\ \vartheta_{j,M} \sim N(0, \sigma^2_{\vartheta_{j,M}}),\ \ \sigma^2_{\vartheta_{j,M}} \leftarrow$ large
    for $j = 1$ to $d_M$ do
        $\tilde{\vartheta}^*_{j,M} = \frac{1}{T_0} \sum_{t=1}^{T_0} {\vartheta^*_{j,M}}^{(t)}$          ▷ Estimate the posterior mean for imaginary data
        $S^2_{\vartheta^*_{j,M}} = \frac{1}{T_0 - 1} \sum_{t=1}^{T_0} ({\vartheta^*_{j,M}}^{(t)} - \tilde{\vartheta}^*_{j,M})^2$     ▷ Estimate posterior variance for imaginary data
    end for
    Set prior: $\vartheta_{j,M} \sim N(\tilde{\vartheta}^*_{j,M},\ nS^2_{\vartheta^*_{j,M}}),\ j = 1, \ldots, d_M$
    Run MCMC with data $\boldsymbol{n}$ and the above prior
    Select Algorithm 1 or 2
    Estimate the marginal likelihood of model $M$ using MCMC output: $\vartheta^{(t)}_M,\ t = 1, \ldots, T_0$
end for
Calculate posterior model probabilities
Identify the MAP model
Output:
    Marginal likelihood estimates $\widehat{f}(\boldsymbol{n}|M)$ for all $M \in \mathcal{M}$,
    Posterior Model Probabilities $f(M|\boldsymbol{n})$, for all $M \in \mathcal{M}$,
    MAP model
```

2.7 Discussion

In this chapter we introduced the main idea of the proposed methodology. Two prior setups are proposed in order to advocate sensible choices of the prior using imaginary data and the power prior approach. Additionally, two version of the Monte Carlo estimators are proposed, independence and one-block Monte Carlo estimators of the marginal likelihood. They are straightforward to use when the MCMC output is available, efficient and no computational demanding. In the next chapter we provide the implementation of the proposed methodology in two real datasets along with an extensive simulation study.

Chapter 3

Implementation of Bayesian Model Comparison Methodology

All laws are simulations of reality.

John C. Lilly

3.1 Real Data Examples

3.1.1 Example 1: Association of Dreams Disturbance Subscales

The classical dataset of Maxwell (1961), in which the severity of dreams disturbance of 223 boys is cross classified with their age, has been used to illustrate the proposed methodology. Maxwell discusses an analysis of a 5×4 contingency table giving the number of boys with four different ratings for disturbed dreams in five different age groups, see Table 3.1. The higher the rating, the more the boy suffers from disturbed dreams.

We set all cells of imaginary data equal to one and impose a normal non-informative pre-prior with large variance. The posterior mean and posterior

Table 3.1: Cross-classification of 223 boys by severity of disturbances of dreams and age.

Age group	Disturbance (from low to high)				
	1	2	3	4	Total
5-7	7	4	3	7	21
8-9	10	15	11	13	49
10-11	23	9	11	7	50
12-13	28	9	12	10	59
14-15	32	5	4	3	44
Total	100	42	41	40	223

variance for all parameters in each model using the imaginary data are estimated via MCMC. These values are used to build an approximation of the power prior. The results of the new MCMC output are listed in Table 3.2 along with the estimated log-marginal likelihood, the Laplace approximation and the Laplace-Metropolis estimator approach. The results of the two approaches, supporting the same model, are very close.

Table 3.2: Estimated logarithm of marginal likelihood for all the competitive models with the two versions of Laplace method.

Mj	**Model**	**log-marginal**	
		Laplace	**Laplace-Metropolis**
1	Independence (I)	−91.399	−91.296
2	Uniform(U)	−90.167	−90.596
3	Row (R)	−103.771	−103.159
4	Column (C)	−97.652	−97.096
5	Row-Column (RC)	−107.365	−107.446
6	Saturated (S)	−131.665	−131.253

After implementing the proposed methodology using $\mu_{min} = \nu_{min} = 0$, $\nu_{max} = 1$ and $\phi = 1$ parametrization for the RC model, only two models were found with posterior model probabilities higher than 1%; see Table 3.3. All model comparison measures indicate that the Uniform model is the best. According to Table 3.3 the highest probability model is the Uniform association model with fix row and column scores $77,4\%$ and $66,8\%$ for the Laplace Metropolis and the Laplace Approximation respectively. The independence model is supported as the second best but with considerably lower probabilities (0.331 and 0.226 for Laplace approximation and Laplace-Metropolis approach, respectively).

Table 3.3: Estimated logarithm of Bayes factor and posterior model probabilities for all the competitive models with the two version of Laplace method.

Mj	**Model**	**Log-BF_{j2}**		**Posterior Probabilities**	
		Laplace	***Laplace-Metropolis***	***Laplace***	***Laplace-Metropolis***
1	Independence (I)	−0.7	−1.2	0.331	0.226
2	Uniform (U)	0.0	0.0	0.668	0.774
3	Row (R)	−12.6	−13.6	<0.01	<0.01
4	Column (C)	−6.5	−7.5	<0.01	<0.01
5	Row-Column (RC)	−16.8	−17.2	<0.01	<0.01
6	Saturated (S)	−40.7	−41.5	<0.01	<0.01

3.1.2 Example 2: Association of Schizotypal Personality Subscales

In this illustration, we re-analyze the dataset of Table 3.4, which presents the cross-classification of 202 students of the survey according to “social anxiety” and “odd behaviour”. These variables refer to two of the nine specific characteristics of a “schizotypal personality” as they are defined in the DSM-III-R diagnostic and statistical manual of mental disorders, edited by the American Psychiatric Association (1987). Social anxiety refers to excessive stress, nervousness, or feeling extremely uncomfortable when being with other people which does not disappear with familiarity. Odd behaviour is related to eccentric appearance, unusual habits, and peculiar actions that may not be acceptable in society. This dataset was originally considered by Iliopoulos et al. (2009) and extracted from a larger university survey conducted in Greece by Iliopoulou (2004). The main aim of this survey was to assess the association between schizotypal traits and impulsive and compulsive buying behaviour of university students. This dataset is a 5 × 5 contingency table with many cells with zero frequencies. This fact makes our analysis more complicated because we have an extra problem to concern and this is the sparsity. Many methods break down when the table is sparse or when the sample size of the table is small.

Table 3.4: Schizotypy data: Cross-classification of 202 students by social anxiety and odd-behavior sub-scales

Social anxiety score	Odd behavior score						
	0	1	2	3	4	5-7	Total
0	11	5	1	0	1	0	18
1	13	8	8	2	2	3	36
2	8	9	4	1	4	0	26
3	6	7	5	4	4	1	27
4	6	9	5	3	2	4	29
5	3	13	5	4	1	5	31
6-8	0	11	5	10	3	6	35
Total	47	62	33	24	17	19	202

Under the first prior scenario, we set all the imaginary data equal to one, i.e. $n_i^* = 1$. We impose normal non-informative pre-priors with large variances. These imaginary data are then used to build an approximation of the power prior

Table 3.5: Estimated log-marginal likelihood for all the competing models.

			Log-marginal						
Prior	M_j	Model	Laplace	Laplace		Independence		One-Block	
				Metropolis		Perrakis		Perrakis	
Prior 1	1	Independence (I)	-151.63	-152.04	(0.024)	-151.64	(0.015)	-151.59	(0.002)
	2	Uniform (U)	-148.01	-148.64	(0.025)	-147.07	(0.166)	-147.95	(0.004)
	3	Row (R)	-173.61	-175.05	(0.045)	-174.08	(1.242)	-173.51	(0.012)
	4	Column (C)	-158.72	-160.16	(0.049)	-159.85	(0.776)	-158.62	(0.007)
	5	Row-Column (RC)	-182.86	-182.02	(0.058)	-184.02	(2.496)	-182.66	(0.048)
	6	Saturated (S)	-234.43*	-289.80	(0.122)	-362.79	(11.312)	-234.85	(5.944)
Prior 2	1	Independence (I)	-160.88	-161.38	(0.026)	-160.61	(0.015)	-160.83	(0.002)
	2	Uniform (U)	-157.53	-158.21	(0.023)	-158.01	(0.123)	-157.49	(0.004)
	3	Row (R)	-186.19	-187.65	(0.038)	-191.32	(1.934)	-186.08	(0.012)
	4	Column (C)	-172.05	-173.52	(0.042)	-176.57	(0.448)	-171.96	(0.015)
	5	Row-Column (RC)	-202.24	-203.01	(0.077)	-205.55	(3.377)	-202.21	(0.028)
	6	Saturated (S)	-302.73*	-287.48	(0.102)	-348.57	(5.625)	-308.40	(4.366)

* *Laplace was obtained after removing cells with zero frequencies.*
Importance sampling size $T = 15,000$ *and MCMC iterations* $T_{mcmc} = 11,000$ *and additional* $1,000$ *burn-in.*

as described in Section 2.6. Then, the marginal likelihood estimates are estimated as described in Section 2.4. All estimates (using Laplace, Laplace-Metropolis, independent Perrakis, and one-block Perrakis estimators) are presented in Table 3.5. The Laplace and Laplace-Metropolis estimators can serve as the gold-standard for models that fall within the GLM framework (i.e. for all models except the RC model) given that the sample size is large. On the contrary, results using these two approaches are only indicative for small samples or for the RC model.

From Table 3.5, we observe that all marginal likelihood estimates are very close except those of the Saturated model. The reason for this instability is the existence of zero cell frequencies in the observed contingency table. For this model, the Laplace approximation (after removing zero frequencies) is in agreement with the one-block Perrakis estimator. In order to test the accuracy and compare the efficiency of the estimation of the marginal likelihood of the three estimators, Laplace Metropolis, Independent Perrakis and One-Block Perrakis, we compute the

Monte Carlo error. MC error measures the variability of each estimate due to the number of the importance sample size.The Laplace-Metropolis and the independent version of Perrakis seem to considerably fail (especially the latter) for the saturated model with extremely high Monte Carlo error (≈ 11.3). The one-block Perrakis estimator is adequately accurate but the Monte Carlo error is still high (≈ 5.9 units in terms of log-marginal likelihood). Generally, the Monte Carlo error of the independence sampler is almost twice as high as the corresponding error of the one-block sampler of the saturated model, while for the rest of the models this MCE ratio between the two methods ranged from 7.5 to 110 (the ratio of MCE of Column association model $\frac{0.776}{0.007} = 110$ between independence and one-block estimators).

Similar are the results of the second scenario where all the imaginary data are set equal to the cell mean of the contingency table (here $4.8 \approx 5$). This approach can be thought of as minimally empirical since the power parameter is set equal to $1/n$ so that our prior approximately accounts in the posterior for one data point in total; see the lower part of Table 3.6. Note that, for this prior choice, the MCMC is significantly improved (5.6 and 4.4 for the independent and one-block estimators respectively) mainly because the prior imaginary data were adjusted to reflect the overall mean and this has greater influence on the intercept of the model and the a-priori expected number of cells assumed and introduced in our analysis.

Table 3.6 and Figure 3.1 present the estimated log-Bayes factors of each model against the MAP which, here, is the uniform. Again all estimators are similar and the one-block estimator is very accurate in all cases except for the saturated model. Note that for this estimator, all log-Bayes factors are calculated with very low Monte Carlo error even for relatively small samples of 1000 or 3000 generated values. In Figure 3.1, cyan and red shaded areas represent the plus-minus two times Monte Carlo error for the one-block and independent estimators. It is clear that the accuracy of the one-block estimator is much higher than that of the

Figure 3.1: Estimated log-Bayes factors of each model with the uniform association model (MAP in this analysis). Shaded areas represent ±2× Monte Carlo errors.

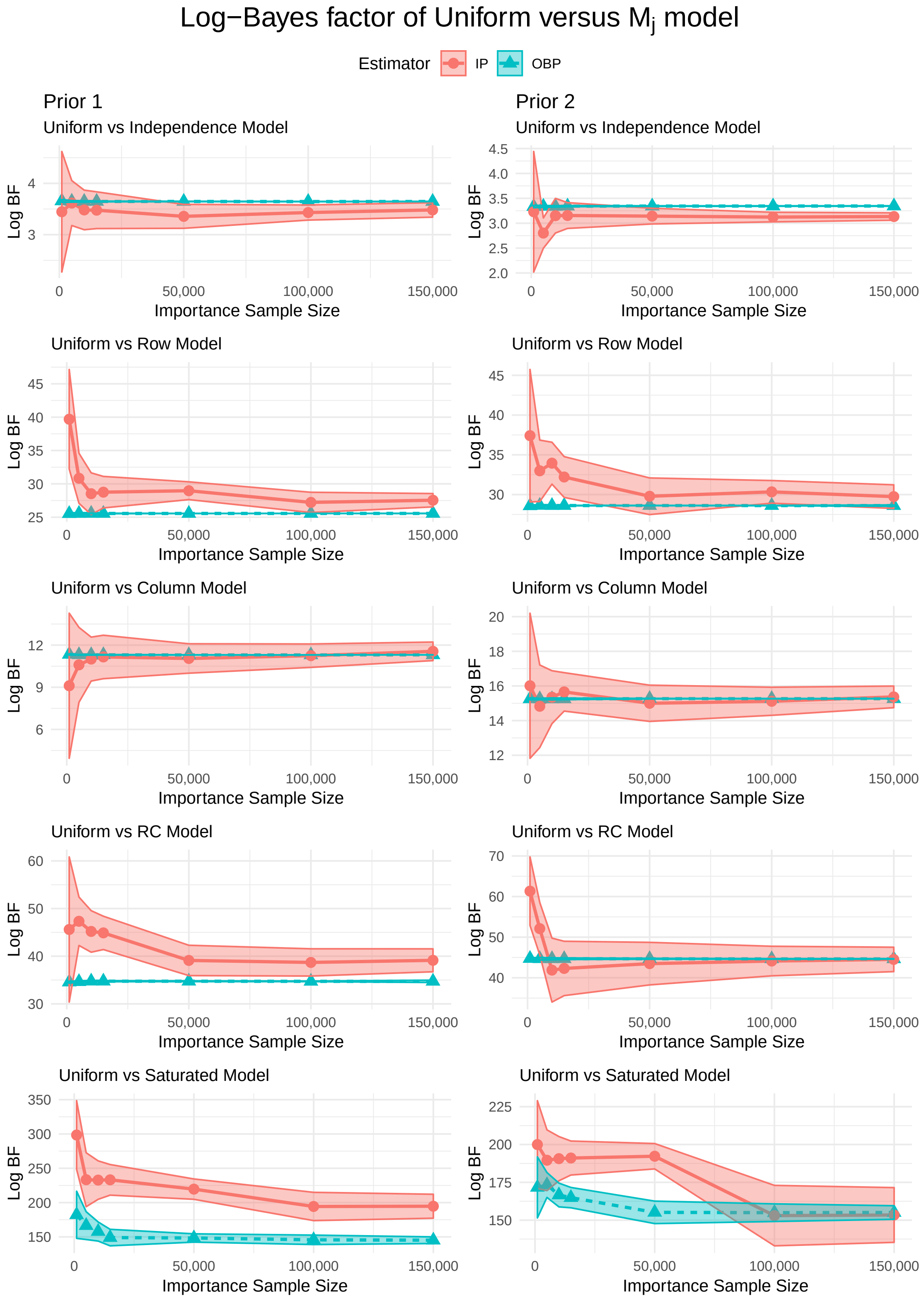


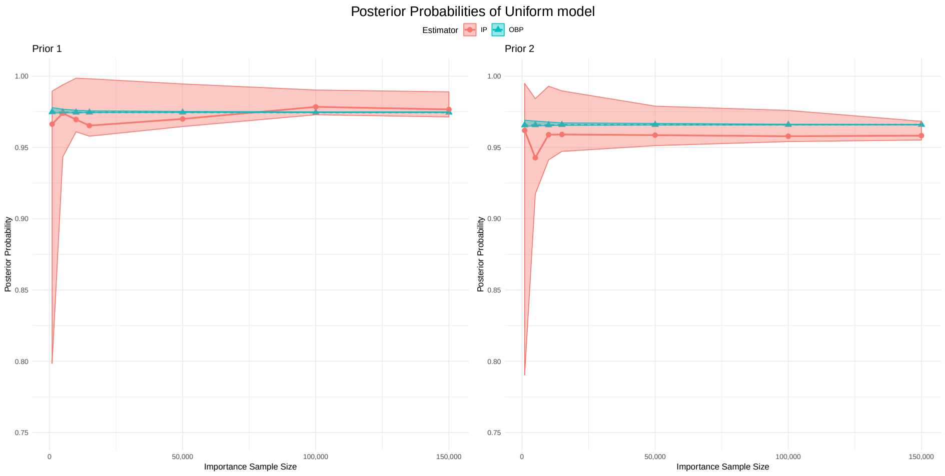
Table 3.6: Estimated log-Bayes factor of all the models compared with the maximum a-posteriori model (Uniform); Results are in ascending order of log-Bayes factors.

			Log-Bayes Factor of Uniform model vs. M_j						
Prior	M_j	Model	Laplace	Laplace Metropolis		Independence Perrakis		One-Block Perrakis	
Prior 1	1	Uniform (U)	0.00	0.00	—	0.00	—	0.00	—
	2	Independence (I)	3.63	3.39	(0.034)	4.57	(0.179)	3.64	(0.003)
	3	Column (C)	10.71	11.51	(0.054)	12.78	(0.773)	14.89	(0.009)
	4	Row (R)	25.61	26.40	(0.048)	27.01	(1.182)	25.56	(0.019)
	5	Row-Column (RC)	34.85	33.37	(0.069)	36.95	(1.745)	34.71	(0.053)
	6	Saturated (S)	86.43^*	141.15	(0.123)	215.72	(11.151)	86.90	(6.011)
Prior 2	1	Uniform (U)	0.00	0.00	—	0.00	—	0.00	—
	2	Independence (I)	3.34	3.17	(0.034)	2.59	(0.130)	3.34	(0.005)
	3	Column (C)	14.52	15.31	(0.055)	18.56	(0.555)	14.47	(0.021)
	4	Row (R)	28.66	29.44	(0.042)	33.32	(1.283)	28.59	(0.015)
	5	Row-Column (RC)	44.71	44.81	(0.081)	47.54	(3.346)	44.71	(0.096)
	6	Saturated (S)	150.97^*	129.27	(0.101)	190.57	(5.616)	150.91	(3.361)

* *Laplace was obtained after removing cells with zero frequencies.*
Importance sampling size $T = 15,000$ *and MCMC iterations* $T_{mcmc} = 11,000$ *and additional* $1,000$ *burn-in.*

independent estimator for all models since the cyan shaded area is much smaller than the corresponding red ones for all models except for the saturated. For the saturated model, the one-block estimator demonstrates again smaller intervals but the improvement in the corresponding intervals of the independent estimator is considerably higher.

Figure 3.2: Estimated posterior probability of the uniform association model (MAP in this analysis) for various importance sampling sizes. Shaded areas represent ±2× Monte Carlo error (restricted to the zero-one interval); $T_{MCMC} = 11000$ iterations and additional 1000 iterations as burn-in.



The same conclusions can be drawn if we focus on the posterior probability of the maximum a-posteriori (MAP) model which is the uniform model in our analysis; see Table 3.2. Even if we observe an instability in the estimation of the marginal likelihood of the saturated model, all approaches indicate that the Uniform model is the maximum a-posteriori model (MAP) with 97% posterior probability (and 99% for the importance Perrakis estimator). The second best model, is the independence model with posterior probability around 3% for all methods except for the IP estimator (≈ 1%). All the other models have negligible posterior model probabilities, lower than 0.1%. The second prior scenario with larger differences in the estimated of the log-marginal likelihoods, also, complies to our expectations. The one-block estimator provides an accurate estimate even when the importance sample size does not exceed 1000. The Monte Carlo error is equal to 0.033 while the importance sampling estimator required considerably higher size (about 15,000 generated observations) to reach similar levels of precision. The variability of the estimated posterior probability was very high for sizes of 5000 observations or lower, with MC error ranging from 1.5% to 8.4% for 1000 importance sampling generated values.

Concerning the required number of importance sampling iterations T, we have specified it by using the percentage change of the estimated log-marginal between subsequent Monte Carlo simulations of different size. To calculate this, we have considered $T \in \{1K, 5K, 10K, 15K, 50K, 100K, 150K\}$ iterations; where K stands for thousands. The value of $\alpha = 1\%$ accuracy has been selected as threshold for the percentage differences. If all differences are lower than 1%, then we assume that the required number of iterations needed for the estimators is at most equal to 1000. And indeed, this is the case for most of the simulations we have considered here; see columns 6 and 9 of Table 3.7. For some limited cases (models Row, RC and saturated for the independent sampler and the saturated model only for the one-block estimator), we needed to increase the importance sampling size and/or the precision level α in order to get a stabilized estimator in the sense that the reported precision is achieved for all samples sizes higher than the reported sample size. For these cases, in Table 3.8 we report the required importance sampling size for different levels of precision $\alpha \in \{1\%, 2\%, 3\%, 4\%, 5\%, 6\%, 12\%, 15\%\}$. For

Table 3.7: Required importance sample size and precision α as percentage change of the estimated log-marginal likelihood.

		Prior 1			Prior 2		
Model	Estimator	Max(%)>α	Iterations (>α)	α	Max(%)>α	Iterations (>α)	α
Independence	Ind	0.00031	1000	0.01	0.00029	1000	0.01
	OB	0.00001	1000	0.01	0.00003	1000	0.01
Uniform	Ind	0.00095	1000	0.01	0.00302	1000	0.01
	OB	0.00005	1000	0.01	0.00004	1000	0.01
Row	Ind	0.01217	5000	0.02	0.01264	5000	0.02
	OB	0.00017	1000	0.01	0.00022	1000	0.01
Column	Ind	0.00868	1000	0.01	0.00409	1000	0.01
	OB	0.00013	1000	0.01	0.00006	1000	0.01
RC	Ind	0.02929	1000	0.03	0.00602	10000	0.01
	OB	0.00062	1000	0.01	0.00034	1000	0.01
Saturated	Ind	0.06891	5000	0.07	0.11207	1000	0.12
	OB	0.00899	15.000	0.01	0.03022	1000	0.04

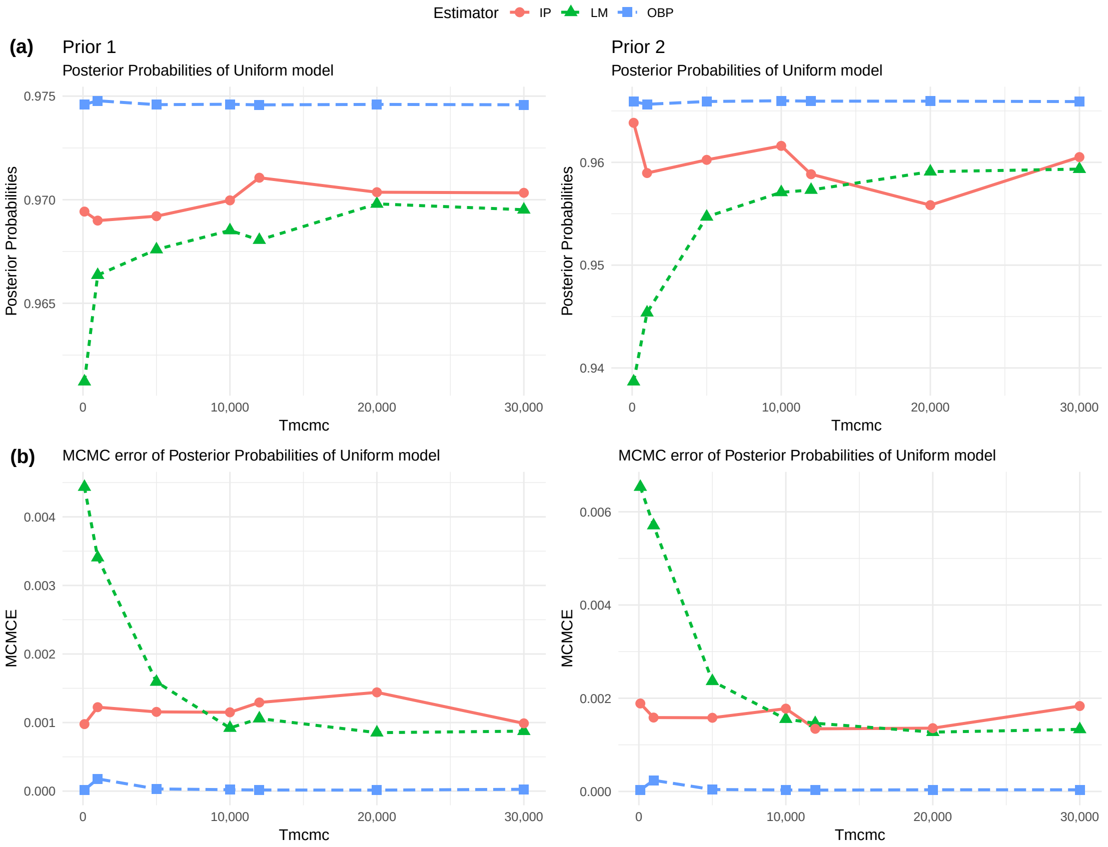
example, for the Row model the precision of $\alpha = 2\%$ is achieved for 5000 generated samples (after that the Monte Carlo error is smaller than 2%) for both prior setups. For the RC model, a precision of 3% is achieved with 1000 generated samples for prior 1 while for prior 2 we can achieve precision of 1% but with the increased size of 10000 generated values. Finally, the saturated model has been proven the most problematic in terms of estimation due to the zeros appearing in their its structure. For the one-block estimator we can achieve precision of 1% and 4% with samples of 15000 and 1000 for priors 1 and 2, respectively. The independent estimator does not seem cost-effective since the precision of 1% is achieved only when 100 thousand generated values are obtained; see Table 3.8. Hence, a reasonable choice to settle with an increased precision of 7% and 12% obtained using 5000 and 1000 generated samples as reported in Table 3.7.

Table 3.8: Required importance sample size for different levels of precision α as percentage difference of the estimated log-marginal likelihood; only reported for models that required higher than 1000 importance sample siz to achieve precision of 1% as indicated in Table 3.7.

Prior	Model	Method	$\alpha = 1\%$	$\alpha = 2\%$	$\alpha = 3\%$	$\alpha = 4\%$	$\alpha = 5\%$	$\alpha = 6\%$	$\alpha = 12\%$	$\alpha = 15\%$
Prior 1	Row	Ind	10K (1.1%)	5K (1.2%)	5K (1.2%)	5K (1.2%)	1K (4.8%)			
	RC	Ind	50K (0.3%)	50K (0.3%)	1K (2.9%)					
	Saturated	Ind	100K (0.1%)	100K (0.1%)	100K (0.1%)	100K (0.1%)	100K (0.1%)	100K (0.1%)	5K (6.9%)	1K (14.6%)
		OB	15K (0.9%)	15K (0.9%)	5K (2.8%)	5K (2.8%)	1K (4.7%)			
Prior 2	Row	Ind	50K (0.3%)	5K (1.3%)	1K (2.1%)					
	RC	Ind	10K (0.6%)	10K (0.6%)	10K (0.6%)	10K (0.6%)	10K (0.6%)	1K (5.1%)		
	Saturated	Ind	100K (0.1%)	100K (0.1%)	100K (0.1%)	100K (0.1%)	100K (0.1%)	100K (0.1%)	1K (11.2%)	
		OB	50K (0.1%)	50K (0.1%)	50K (0.1%)	1K (3.1%)				

Regarding the number of MCMC iterations, Figure 3.3 presents the evolution of the posterior model probability (and its MCMC error) of the Uniform model (MAP in this example) over different MCMC sample sizes (with importance sample size kept to 12000 to ensure minimum variability due to the Monte Carlo variability of the estimators). Results indicate that the one-block estimator converges very fast (even for 1000 MCMC iterations) while for both the independent and the Laplace Metropolis estimators a much higher MCMC sample size is required (about 10-20000 iterations).

Figure 3.3: Posterior probabilities of the MAP (Uniform) model (a) and the corresponding MCMC errors (b) vs. the number of MCMC iterations (T_{MCMC}); Importance sample size equal to 15000.



Finally, the comparison of the computational time between the two estimators didn’t reveal adequate distinctions (processor characteristics: Intel(R) Core(TM), i7-7700 CPU, 3.60 GHz, RAM 16,0 GB). The independent Perrakis estimator approach with 12.000 MCMC iterations and 15.000 importance sampling observations needed 2.27 *min* to compute the marginal likelihood and posterior model probabilities of all models in our example. The one-block estimator needed 2.82 *min* for the same computations. The computational time difference between the two estimators is at most 2.81 *secs* for the saturated model which is the maximum of the differences across all models under comparison. The mean difference between the two estimators (across all six models under consideration) was found to be about

2.4 *secs*. However, if we consider that the one-block estimator converges much faster in terms of importance sampling size we conclude that the attractiveness of the independent estimator is debatable.

3.1.3 Comparison One-Block Importance sampling with Bridge sampling marginal likelihood estimator.

For this example we compare the One-Block Perrakis estimator with the most recently popular marginal likelihood estimator the bridge sampling. Meng and Wong (1996) introduced an efficient Monte Carlo estimator on the basis of a simulation sampling and is given by

$$\hat{f}_{br}(n|M) = \frac{\frac{1}{T_1}\sum_{t=1}^{T_1} h(\vartheta_M^{\star(t)}) f(\vartheta_M^{\star(t)}|M) f(\boldsymbol{n}|\vartheta_M^{\star(t)}, M)}{\frac{1}{T_2}\sum_{t=1}^{T_2} h(\vartheta_M^{(t)}) g(\vartheta_M^{(t)})}$$

where $g(\vartheta_M)$ is the proposal distribution, $h(\vartheta_M)$ is an arbitrary bridge function, $\vartheta_M^{\star(1)}, \ldots, \vartheta_M^{\star(t)}, \ldots \vartheta_M^{\star(T_1)}$ is sample from the proposal distribution and $\vartheta_M^{(1)}, \ldots, \vartheta_M^{(t)}, \ldots \vartheta_M^{0(T_2)}$ is sample from the posterior distribution usually taken from MCMC algorithm. The efficiency of this estimator depends on the selection of the proposal distribution g and the bridge function h. The proposal distribution must be close to the target posterior distribution and the function h plays the role of the bridge that links the two distributions. The results, Table A.7, show that the two estimators, One-Block and the bridge sampling, are quite comparable and efficient, with one-block being considerably simpler to implement. Bridge sampling estimator required the specification of both the proposal distribution and the bridge function, which make the approach unattractive for inexperienced users. Of course these are initial results and should be assessed thoroughly the behaviour of the two estimators.

3.2 Simulation Study

Following Galindo-Garre et al. (2004) and Iliopoulos et al. (2009) simulation study format, we have also conducted a Monte Carlo study in order to assess the accuracy and the efficiency of the proposed methodology. To be more specific, with this extensive simulation study we assess whether the proposed methodology is model selection consistent (i.e. selects the true model structure). Then, we compare the efficiency of the two proposed estimators and we also test for the sensitivity of the posterior results by using different values of the association parameter ϕ of the uniform model. We also examine the effect of the total sample size of the contingency table on the selection of the true model.

First, we consider six different sample sizes, $n \in \{20, 50, 100, 1.000, 10.000, 100.000\}$ for six different scenarios. For each scenario we generate 100 contingency tables of dimension 5×5 from the following models: independence, uniform, row, column, RC and the saturated, that is $\mathcal{M} = \{I, U, R, C, RC, S\}$. Therefore, for every scenario (true model structure) we generate a total of 600 contingency tables (100 tables for every sample size under consideration). Details about the exact sampling schemes and the parameter values we have considered are available at Appendix A.4 along with the table of the expected cell frequencies θ_{ij} used to generate each contingency table. For all simulations we have used $T_{MCMC} = 11000$ with additional 1000 burn-in period and importance sampling size $T = 15000$.

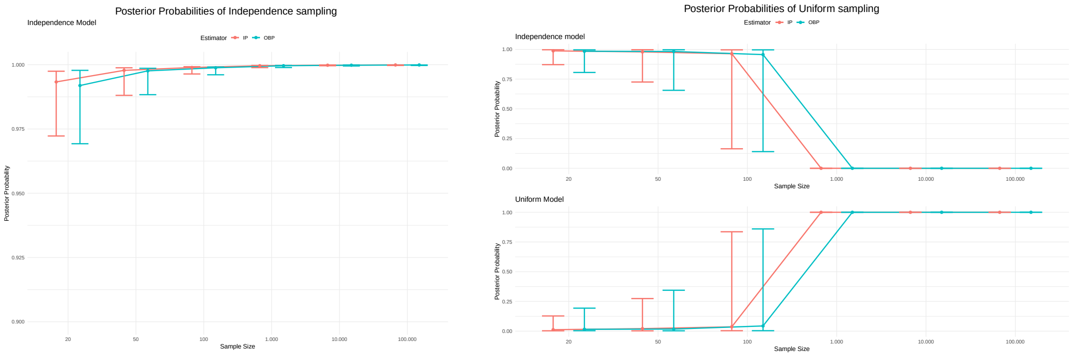
Note that, for small sample sizes, the generated contingency tables are usually very sparse with zero or small frequencies appearing often in specific cells, which results in the generation of contingency tables with high variability concerning their structure. This sparsity and uncertainty concerning the structure of the generated contingency table, makes it difficult to detect the true model for any statistical procedure. For this reason, in practice we need an increased sample size in order to be able to identify the true generating model for most cases.

3.2.1 Assessing model selection consistency

Here we present the evolution of the estimated posterior model probabilities for each of the selected sample sizes $n \in \{20, 50, 100, 1.000, 10.000, 100.000\}$. The results of this analysis are graphically displayed in figures presenting how the posterior model probabilities vary with different sample sizes. Focus is given to the true and the highest a-posteriori models of each simulation scenario. Results for the prior setup 1 are presented in detail while a small summary of the corresponding results using the second prior setup is provided in Section 3.2.4. The aim here is to confirm that the proposed methodology leads to a consistent model selection procedure, i.e. the proposed method identifies the true model structure with increasing posterior model probability as n increases, eventually leading to the selection of the true model with probability one for sufficiently large sample size. Indirectly, we also examine which is the required sample size to identify the true model structure under the assumed simulation scenarios. Results are presented using a variety of different graphical representations of the posterior model probabilities obtained by 100 generated contingency tables for each scenario. The type of plot has been chosen with regard to the visibility and the scale of the results. In Appendix A.5 we further provide boxplots, error bars and line plots for all scenarios. Note that the confidence intervals in the error bars are constructed by using the 5% and 95% percentiles of the posterior model probabilities over the 100 generated samples/contingency tables while the centered value refers to the median of the posterior model probabilities obtained across the generated samples/contingency tables.

For simulation Scenario 1, we have generated values from the independence model $M = I$ with linear predictor (1.15) and parameters λ_0, $\lambda^{\mathbb{X}}$ and $\lambda^{\mathbb{Y}}$ given at the first column of the table in Appendix A.11, with sum-to-zero constraints for the main effects. Figure 3.6a presents the results of the simulation study for this scenario by using error bars of 90% confidence intervals calculated over the 100 generated samples. For this scenario, the true model is identified by the proposed methods even for small samples with average posterior probability of the independence (true) model higher than 97.5% for $n = 20$. In over 95% of the samples the posterior of the true model is a-posteriori supported with probability higher than 92.5%. The support of the independence model is increased when the sample size increases, not only in terms of posterior model probabilities, but also in terms of sampling variability.

Figure 3.6: Error bars of the posterior model probabilities for 100 generated samples (Scenarios 1 and 2); The lines are connected at the median of the posterior probability of each model



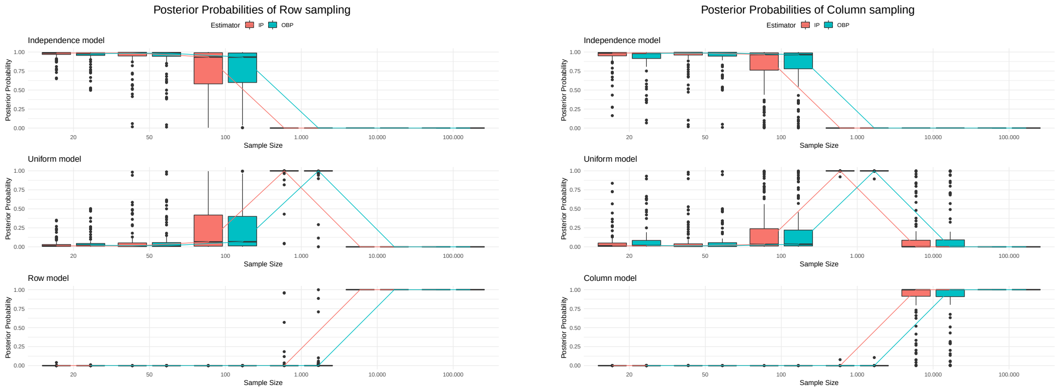
(a) Scenario 1: Data generated from the Independence model

(b) Scenario 2: Data generated from the Uniform model

For simulation Scenario 2, we generate values from the uniform model $M = U$ with linear predictor (2.1) and parameters λ_0, $\lambda^{\mathbb{X}}$ and $\lambda^{\mathbb{Y}}$ given at the second column of Table A.11, with sum-to-zero constraints for the main effects. The additional association parameter of the Uniform model was set equal to $\phi = -0.22$. For the score parameters μ and ν, we have considered fixed scores taking values from one to five. Figure 3.6b presents the results of the simulation study for this scenario by using error bars of 90% confidence intervals over the 100 generated samples for the two highest a-posteriori models, the independence and the uniform (true) model. The true model (uniform) is identified by the proposed methods for sample size equal to $n = 1000$ with average posterior probability of the uniform (true) model higher than 99%. Small sample sizes $n = 20$ and $n = 50$ indicate the independence model as the MAP with average posterior probability 92.1% and 88.8%, respectively. The length of the bars is increased as the sample size increases up to $n = 100$, where we observe the maximum uncertainty between the two competing models. Afterwards our proposed method indicates the correct model revealing that the criterion of the model selection consistency is satisfied in this scenario.

For simulation Scenario 3, we generate values from the row model $M = R$ with linear predictor (2.1) and parameters λ_0, $\lambda^{\mathbb{X}}$ and $\lambda^{\mathbb{Y}}$ given at the third column of Table A.11, with sum-to-zero constraints for the main effects. Additionally, we consider $\phi = 1$, row score parameters given by $\mu = (0, 0.66, 0.61, 0.16, 0.08)$ and fixed column scores $\nu = (1, 2, 3, 4, 5)$. From Figure 3.7a, we observe that the posterior probability of the true model (i.e. R here), tends to one with a slower rate requiring a higher number of observations to identify the true structure. For very small datasets, the model of Independence is a-posteriori supported. For samples of size equal to 1000, the Uniform model is (falsely) indicated as the MAP model. The Row model is identified as the MAP, when increasing the sample size over 1000 observations. The same happens in Scenario 4 where the Column model was used to generate 100 samples; see for details in Figure 3.7b.

Figure 3.7: Boxplots of the posterior model probabilities for 100 generated samples (Scenarios 3 and 4)

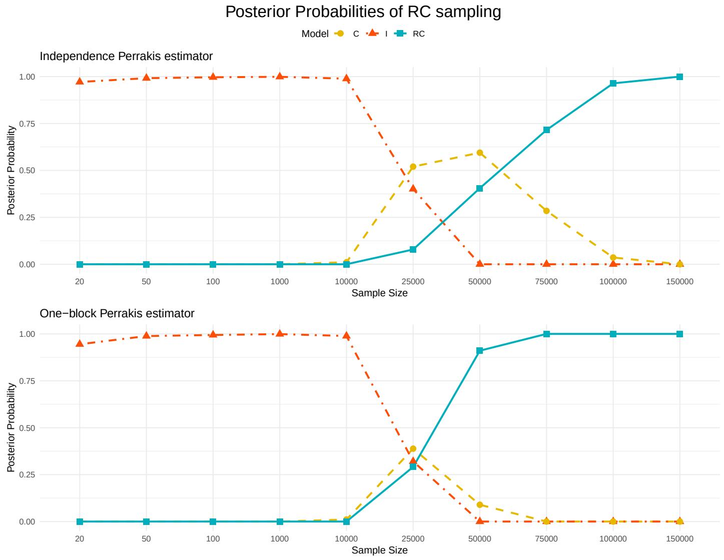


(a) Scenario 3: Data generated from the Row model

(b) Scenario 4: Data generated from the Column model

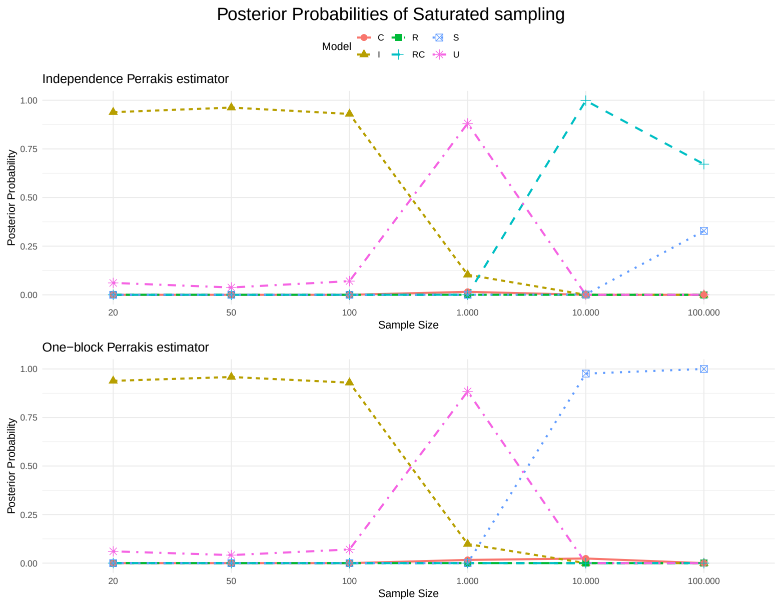
For simulation Scenario 5, the RC model is used as the true generating structure with linear predictor (2.1). We have used the same main effect parameters as in the simulation Scenario 1, $\phi = 1$, row score parameters as in the simulation Scenario 3 and column score parameters as in the simulation Scenario 4. For this case, we have additionally generated samples of 25000, 50000 and 75000 observations in order to smoothly depict the change of the posterior support from the Independence to the Column and finally to the true (RC) model. Figure 3.8 shows that for samples of size up to 10.000, the Independence model is indicated as the MAP model. Under this scenario and for $n = 25000$, we observe (for the first time in this simulation study) notable differences between the two estimators. It seems that the one-block Perrakis is better behaved than the independence estimator spotting the true RC model for small sample sizes with higher posterior model probabilities.

Figure 3.8: Arithmetic means of the posterior model probabilities for 100 simulated datasets-Scenario 5 with the RC being the true generating model



For simulation Scenario 6, the Saturated model is used as the true generating structure with linear predictor (1.14). In this scenario we had to deal with many difficulties. We have the same behavior as the previous Scenario 5.

Figure 3.9: Arithmetic means of the posterior model probabilities for 100 simulated datasets-Scenario 6 with data generated from the saturated model



The procedure identify the true model with slower rate requiring a higher number of observations. This fact makes sense since we usually go to the simpler model where we have a good approximation of the dataset. the procedure failed to identify the true model when we use the independent Perrakis estimator. This failure was our start point to investigate this behavior. We noticed that the procedure fails under the presence of zeros.

3.2.2 Marginal Likelihood computation under the presence of zeros

Contingency tables having small cell counts are said to be sparse and usually contain cells with $n_{ij} = 0$. When data are sparse or the contingency table has many zero cell counts, these can have undesirable features. A count of zero is a permissible outcome for a Poisson or multinomial variate. It contributes to the likelihood function and model fitting. Early applications of Bayesian methods to contingency tables involved smoothing cell counts to improve estimation of cell probabilities with small samples. One possible but naive solution to deal with zeros in contingency tables is to just ignore them and throw them away. In log-linear models we must specify the value of $log(\lambda_{ij})$ for every cell. When we consider the saturated model, the estimated probability of a cell with zero frequency is also estimated to be zero and as a consequence also λ_{ij} is estimated to be equal to zero. The result: $log(\lambda_{ij})$ is undefined.

Many methods break down with the presence of zero counts in statistical inference. The saturated model can be massively affected in model comparison than the rest of the models. Although the problem mainly appears in the saturated model, this leads to the breakdown of the whole model comparison since the saturated serves as reference to evaluate the goodness of fit.

Initial we have started to experiment with the importance sample size, since we knew from the previous simulation study that the independence Perrakis estimator had increased Monte Carlo error in comparison with the one-block estimator. We have tried to see if this was the reason of the failure increasing a lot the number of iterations of importance sample size. We managed to identify the true model in 300K importance sample size!!! So one problem here was the Monte Carlo variability of the method but this was not the major problem eventually because we still want a method that estimate the marginal likelihood distribution with less importance sampling demands. In this section we propose some techniques to avoid the problem of estimating the marginal likelihood of the model.

First technique is to set the zeros into NA. This implies that the corresponding probability was not observed which can be invalid interpenetration given that if we increase the sample size n, this will be eventually observed. With this trick the cell will not carry any information, so the analysis will not be affected by this cell. We use sum-to-zero constraints. The results of this approach are summarized in the Table 3.9. Table 3.10 shows that the minimum Monte Carlo error is 6.2 for the saturated model when the importance sample size is equal to

150.000. The procedure indicates the saturated model as thee best fitted model from all the competitive models with posterior probability equal to 1 and the log-marginal likelihood to be equal to -321.6

Second technique that seems working efficiently is to avoid the zero cell counts, considering as structural zero. With this way the zero cell count will be had zero contribution to the analysis. The Table 3.10 shows that the minimum Monte Carlo error is 11.9 for the saturated model when the importance sample size is equal to 500.000. The procedure indicates the saturated model as the best fitted model from all the competitive models with posterior probability equal to 1 and the log-marginal equal to -292

Third technique under sum-to-zero constraints for the main effects and corner constraints for the interaction parameter, we set a normal prior $N(0, 100)$ to the parameter of the interaction term of the cell with zeros. As we can see from Figure 3.10 in order to detect the true model importance sample size equal to 500.00 is required.

The results of the three techniques are summarized in Tables 3.9 and 3.10

Table 3.9: Estimated the posterior probabilities and the log-marginal for all competitive models, when the minimum Monte Carlo error occurred at t=500.000 importance sample size.

M_j	Model	Posterior Probabibilies	Log-marginal		
			1st	2nd	3rd
1	Independence (I)	0	-1105.743	-1105.743	-1106.86
2	Uniform (U)	0	-1107.305	-1107.305	-1109.29
3	Row (R)	0	-421.3887	-421.389	-421.96
4	Column (C)	0	-1079.270	-1079.271	-1080.34
5	Row-Column (RC)	0	-431.7123	-431.409	-655.64
6	Saturated (S)	1	-321.6371	-292.002	-399.22

Table 3.10: Estimated the Monte Carlo error for all competitive models occurred at t=500.000 importance sample size.

Mj	Model	Monte Carlo error		
		1st	2nd	3rd
1	Independence (I)	0.0004	0.0002	0.0003
2	Uniform (U)	0.0003	0.0002	0.0002
3	Row (R)	0.0009	0.0006	0.0007
4	Column (C)	0.0015	0.0008	0.0014
5	Row-Column (RC)	0.1998	0.1687	8.4529
6	Saturated (S)	6.8107	11.927	35.9917

Figure 3.10: Estimated Monte Carlo Error for the three techniques via the importance sample size

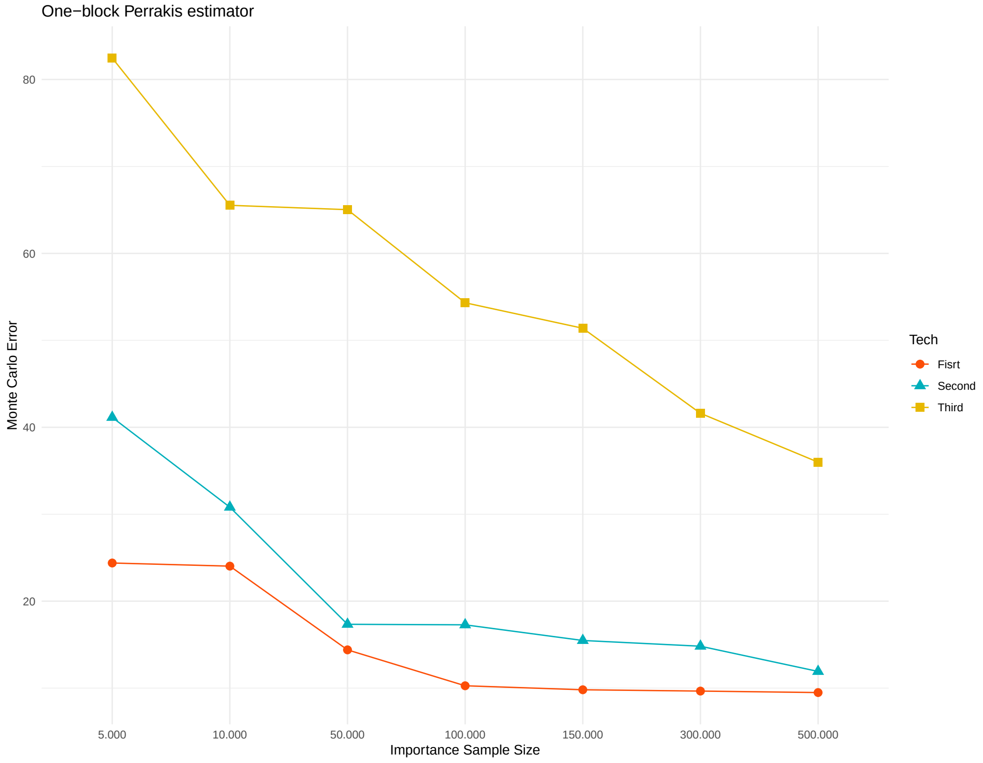


Table 3.9 performs the estimated log-marginal for the six candidates model

using the three techniques, when the minimum Monte Carlo error occurred. From the results we can see that when the complexity of the model is low the three techniques start to deviate, with the third technique diverge from the other two. So, we consider that the third technique fails to give a good estimation of the log-marginal. The estimated Monte Carlo error of log-marginal likelihood confirms this assumption with the third technique performs the higher Monte Carlo error, see Table 3.10. Technique 1 and 2 have similar Monte Carlo error for all candidate models except the saturated where the first technique has lower MC error.

3.2.3 Concluding Comments based using Prior 1

The simulation study shows that the proposed method performs well in detecting the true model. More specifically, the results showed that when the total sample size is small, the probability to detect the true model is quite small. Only when the true structure is independence, we could detect it by assuming small samples with $n \geq 20$.

3.2.4 Summary of results for Prior 2

The simulation study results for prior 2 are quite similar with those of prior 1. Although the results referring to the prior 2 are not provided here in details, one can refer to Appendix A.5.2 for associated graphical illustrations. Here we only present in line plots the evaluation of the posterior model probabilities over different samples size n for prior 2. From this Figure, prior 2 appears to be stricter than prior 1 for more complex models and it requires large sampler sizes in order to detect the true model. Nevertheless this method also appears to be consistent in terms of the selected model.

3.2.5 Comparison of Marginal Likelihood Estimators

The gold standard for our comparisons, which are presented in our illustrative examples, are the two versions of Laplace estimators. They both provide a reasonable solution for the marginal likelihood. Laplace approximation is an optimization method, while the Laplace-Metrolopis estimator is based on the output of an MCMC algorithm. The first method is faster and without any Monte Carlo error, but is not always easy to apply due to computational problems. This is the case in Example 1, where we have sparse data and the Laplace estimation fails to accurately compute the marginal likelihood. The second approach is slower because it requires to run an MCMC algorithm, but when the output is provided the implementation is direct. For the Laplace-Metropolis we require only the MCMC estimated posterior mean vector and the posterior covariance matrix. Both approaches provide approximations of the marginal likelihood and the required adequately large sample size. They do not work efficiently in small or sparse datasets. Both approaches have some regularities conditions (see Kass and Wasserman (1995)). These approximations work efficiently when the posterior distributions are symmetric, otherwise transformations are applied (log, logit or Box-Cox).

The independent and one-block Perrakis estimators require as input only the posterior marginal samples, which are used to build the importance sampling function. The approach of Perrakis et al. (2014) is non-iterative and does not require adaptations in MCMC sampling. The method can fail when the product of marginal posteriors is a poor approximation to the joint posterior. The proposed approach is applied efficiently to association models, and leads to accurate marginal likelihood estimates. As illustrated, in model comparison, the one-block estimator is more accurate in capturing the true model in both scenarios of the priors. In general, the overall performance of the estimator depends on; i) the efficiency of approximating the posterior through independent univariate or multivariate marginals and ii) the accuracy in estimating marginal posterior densities.

3.2.6 Parameter ϕ sensitivity

The Uniform association model has only one parameter additional to the parameters of the independence model. It can describe a positive or a negative trend association between the variables. The multiplicative interaction term $\phi\mu_i\nu_j$ represents the deviation of log counts from the independence pattern. The greater it departs from independence, the higher are the frequencies in specific corners of the table. The direction and strength of association depends on ϕ. When $\phi > 0$ the categorical variable X increases and then the categorical variable Y will also increase. When $\phi < 0$, Y tends to decrease as X increases. Small values of ϕ indicate weak association between the two variables, so it would be relatively more difficult to detect the true association model. The simulation results showed that when ϕ is equal to zero, then our proposed model comparison procedure correctly suggests that the Independence model is the best fit model. As ϕ increases, the proposed methodology supports a-posteriori more and more the true model.

To test the effect of parameter ϕ, we consider simulated scenarios with $\phi = k|\phi_{S_2}|$ where $\phi_{S_2} = -0.214$ is the corresponding parameter ϕ in the initial simulation Scenario 2 (see Table A.11 at the Appendix for details) and k takes values from -20 to 20 with increment equal to 2. For each value of ϕ we have generated 100 samples. In Figure 3.12a, $Y{-}axis$ represents the mean of the posterior probabilities of the ten simulated samples for each value of ϕ. Small values of $|\phi|$ indicate weak association between the two variables. In such cases, it is relatively more difficult to detect the true association model.

The simulation results indicate that when ϕ is equal to zero, then the procedure correctly identifies the Independence model as the best fit model. From Figure 3.12b and for $k > 0$ we observe that our method correctly identifies the uniform model with increasing posterior probability as ϕ increases. For $k < 0$, the posterior probability of the uniform model increases as $|k|$ increases with an exception at values around of $k = -10$.

To further examine this irregularity, in reference to the above simulation scenario (denoted as S1–Φ), we have considered two additional simulation scenarios for different choices of fixed effects parameters: $\lambda^{\mathbb{X}} = \lambda^{\mathbb{Y}} = (0.3, 0.2, 0.1, 0, -0.6)$ (S2–Φ) and $\lambda^{\mathbb{X}} = \lambda^{\mathbb{Y}} = (0, 0, 0, 0, 0)$ (S3–Φ). In S2–Φ, for $k = 0$ the joint probabilities of the upper left corner are slightly higher than the corresponding frequencies of the lower right corner (ranging from 5.5% to 6.7% vs 1.1% to 3.7%, respectively). Increasing the value of k causes a movement of the cell relative toward the lower right corner even for small values of k. Reducing the value of k (going to negative values), creates the opposite movement. For these three examples, we have observed that all relative cell frequencies are gathered in the lower right corner of the table for smaller values of ϕ and k compared to the corresponding concentration in the upper-left corner of the table which results when $k < 0$ but for much larger values of $|\phi|$ and $|k|$. To be more specific, all values are gathered to the lower right corner (with 100% relative frequency) for $k = 12$ while for $k = 6$ the relative frequency of n_{55} is equal to 83% for S1–Φ and over 99% for the other two scenarios (S2–Φ and S3–Φ). Only for $k > -14$ we achieve high concentration in the upper-right corner of the contingency table with relative frequency of $n_{11} > 90\%$. Thus, in the negative values of ϕ (and k) there is a large range of values where the dominance of the uniform model is a-posteriori unclear since the corresponding posterior probability is close to 50%.

3.3 Discussion and Concluding remarks

In this chapter, we deal with the problem of Bayesian model selection for association models used in two-way contingency tables. We have focused on the comparison of association models and the Independence and Saturated model,

$\mathcal{M} = \{I, U, R, C, RC, S\}$, using power priors. To achieve this, we test four estimators to compute the marginal likelihood: Laplace, Laplace-Metropolis, independent Perrakis and one-block Perrakis estimators. We illustrate a comprehensive comparison of the four estimators in which the One-Block Perrakis outperformed the others in every example. The utilization of power priors provides the researcher with a reasonable prior avoiding the effect of Lindley's paradox. Moreover, we achieve compatibility of priors across models under consideration due to the use of common imaginary data across models. With the term of compatibility we denote that priors should be related across models, each being conditional on the given model. Our approach can easily handle sparse tables and the best a posteriori models are provided automatically by the Bayes factor. The imaginary data were generated from the simplest model in order to support more parsimonious models. This way, we achieve a sensible default choice of prior, in the Bayesian context analysis, with the minimum informative cost. The use of product marginal as an importance faction makes the method very simple to use and computationally efficient. Also, the simulation study and the real data analysis suggest that the proposed methodology performs well.

An obvious extension of our proposed method is to embody to our algorithm different scenarios of priors between the two cases we present here. Other interesting issues for future research may include a two-block Perrakis estimator of the marginal likelihood, where the two block will be the correlated and uncorrelated parameters.

Both Laplace approximation and Laplace-Metropolis estimator provide a reasonable solution for low information estimation of marginal likelihood. The utilization of power priors provides good argument for a reasonable prior and avoids the effect of Lindley's paradox. Moreover, we achieve compatibility of priors across models under consideration due to the use of common imaginary data across models. In the future, we will focus on the implementation of other alternative estimation methods for the computation of the marginal likelihood, for example by using the Monte Carlo estimate proposed by Perrakis et al. (2014) or the Chib (1995) Chib's (1995) marginal likelihood estimator.

Figure 3.4: Estimated MCMC Error of the log–marginal vs the Number of MCMC iterations (T_{MCMC}).

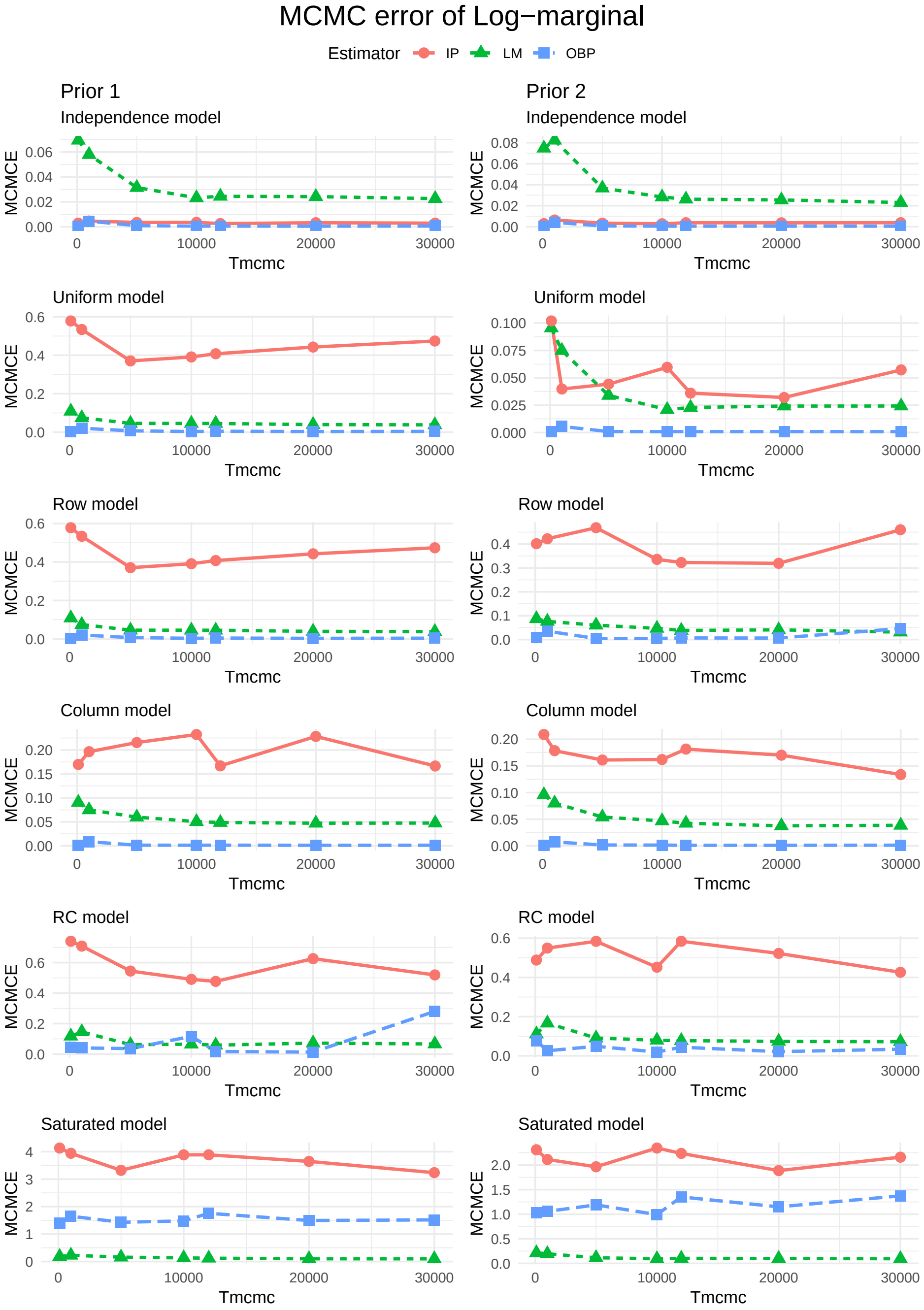


Figure 3.5: Estimated MCMC Error of the log–Bayes factor vs the Number of MCMC iterations (T_{MCMC}).

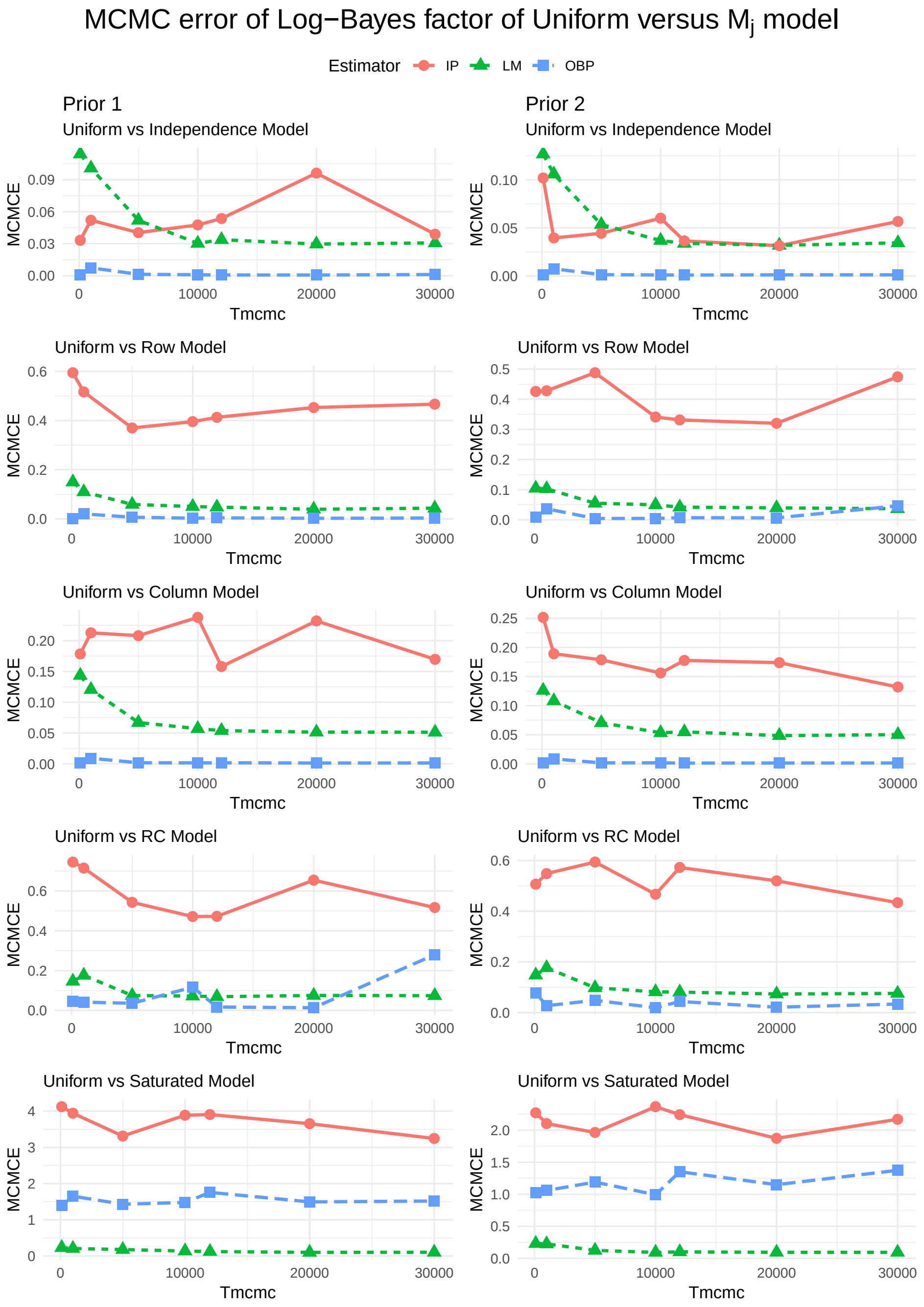


Figure 3.11: Line plots for the posterior model probabilities over 100 simulated datasets for Scenarios 1–6 for Prior 2

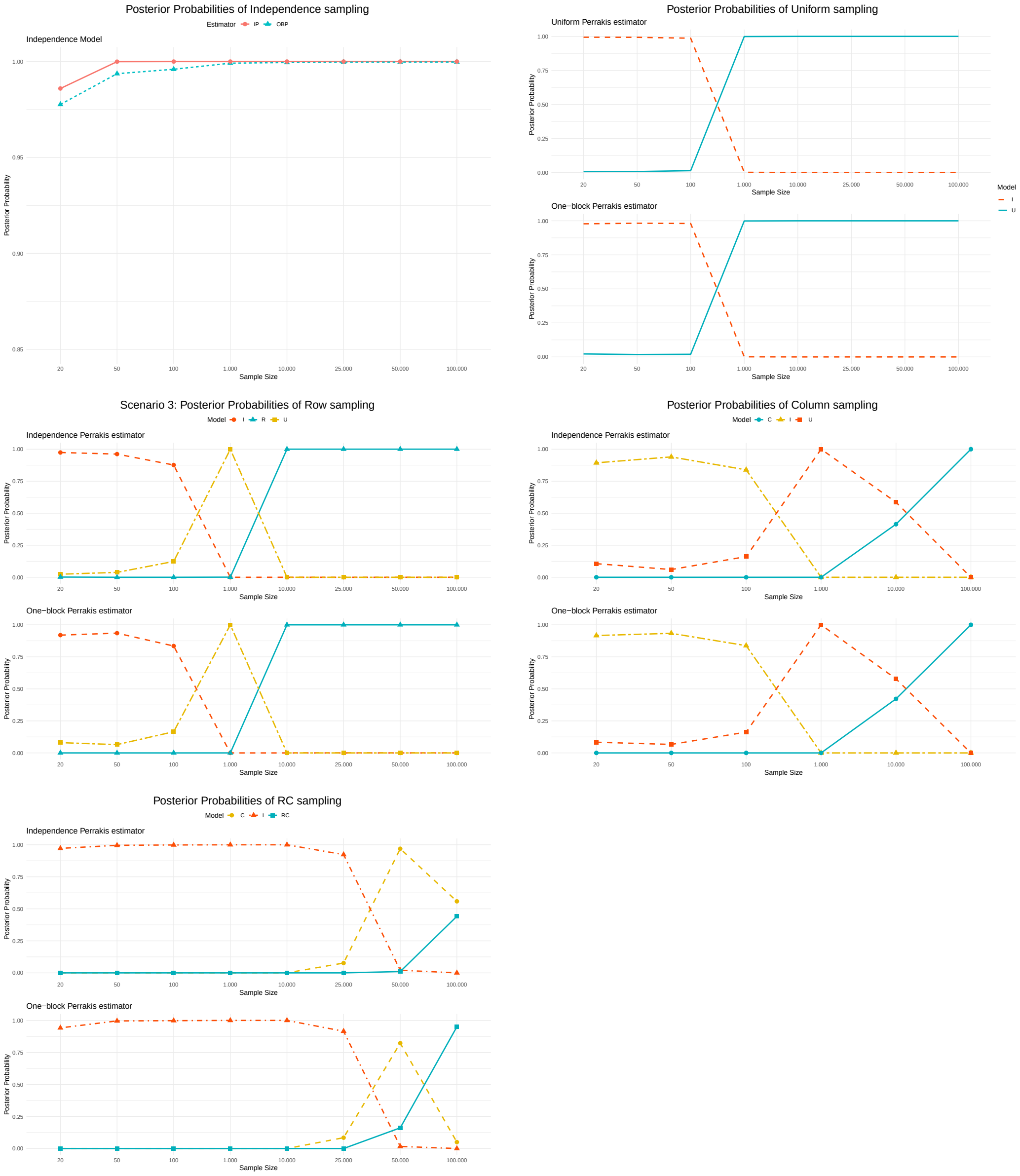
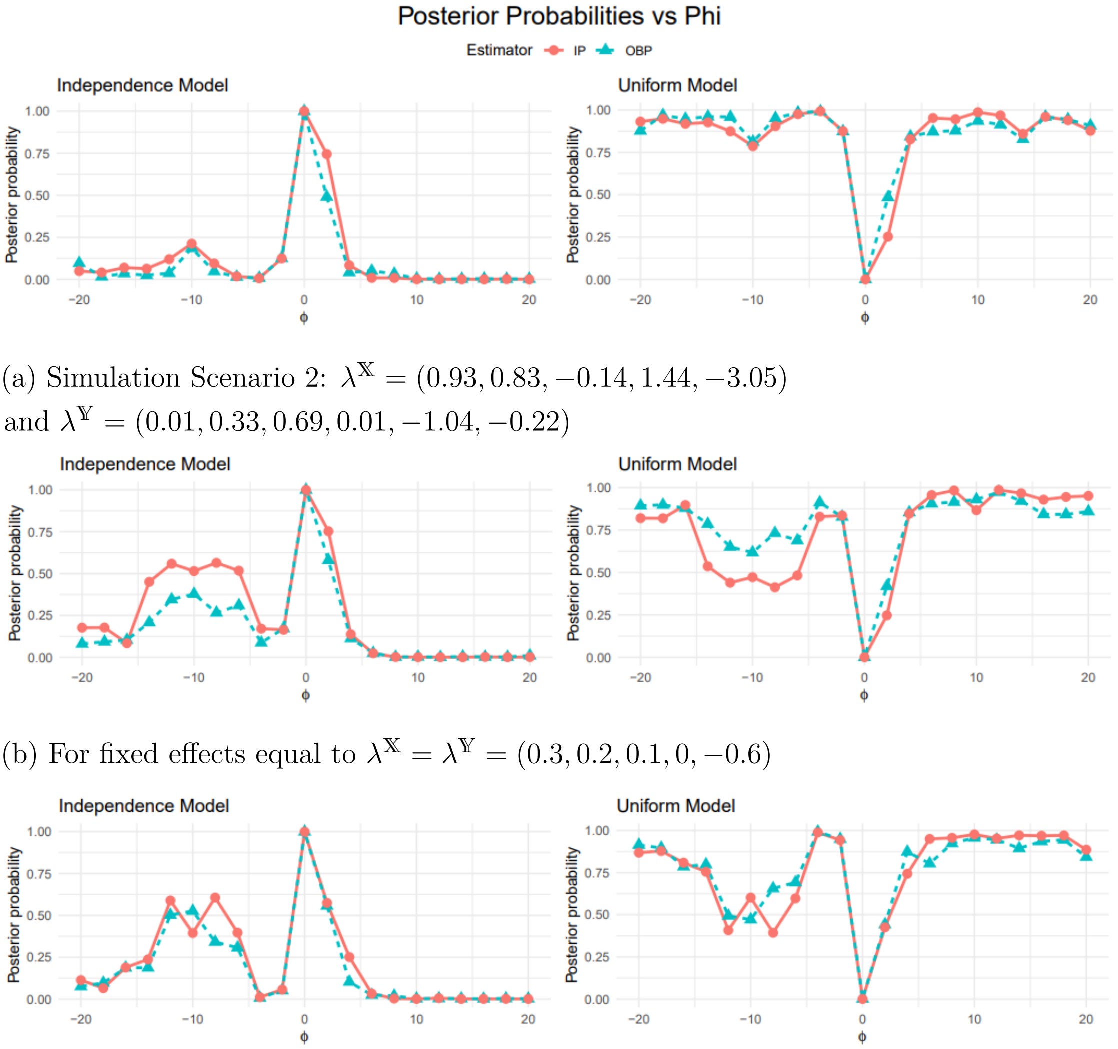


Figure 3.12: Mean of the posterior probability by the parameter ϕ for the three scenarios of main effect $\lambda^{\mathbb{X}}$ and $\lambda^{\mathbb{Y}}$



(a) Simulation Scenario 2: $\lambda^{\mathbb{X}} = (0.93, 0.83, -0.14, 1.44, -3.05)$ and $\lambda^{\mathbb{Y}} = (0.01, 0.33, 0.69, 0.01, -1.04, -0.22)$

(b) For fixed effects equal to $\lambda^{\mathbb{X}} = \lambda^{\mathbb{Y}} = (0.3, 0.2, 0.1, 0, -0.6)$

(c) For fixed effects equal to $\lambda^{\mathbb{X}} = \lambda^{\mathbb{Y}} = (0, 0, 0, 0, 0)$

Chapter 4

Power Priors for the Bayesian Comparison of Graphical Models in Three Way Tables

> Numerical quantities focus on
> expected values, graphical
> summaries on unexpected values
>
> *John Tukey*

4.1 Introduction

In this chapter a Bayesian model comparison procedure of Graphical models in three way contingency tables using power priors is analysed. So far we have seen a Bayesian model selection procedure of association models for two-way contingency tables. Now we proceed in three way contingency tables where we have different types independence, conditional independence, marginal independence, between the variables. For this reason we choose to use graphical models in order to express the different types of independence and build a Bayesian model selection procedure using also here the power prior approach and the utilization of imaginary data in order to build an objective default prior. We illustrate a comprehensive Bayesian analysis of graphical models of conditional independence, involving suitable choices of prior parameters, estimation, model determination, as well as the related computational issues for three-way contingency tables. Each conditional independence model corresponds to a particular factorization of the cell probabilities and a conjugate analysis based on a Dirichlet prior. We present an imaginary data approach for prior specification and we compare alternative prior set-ups. We adopt the idea of Ibrahim and Chen (2000) and Chen et al. (2000b) of power prior approach in order to advocate sensible values for the Dirichlet prior parameters. Unit information interpretation priors are used as a yardstick to identify and interpret the effect of any other distribution used. The posterior distributions of the graphical models parameters, are obtained using simple Markov chain Monte Carlo (MCMC) schemes. A real data application will be analytically presented.

4.2 Graphical Models of Conditional Independence

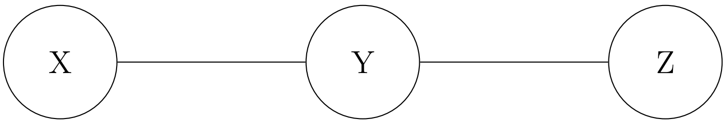
Models that have interpretations in terms of conditional independence are known as graphical models. In this chapter we have more than two categorical variables, a K-way contingency table cross-classifying variables X_v, $v \in \{1, ..., K\}$. A graphical representation for associations indicates the pairs of conditionally independent variables. The use of graphical models to describe association between categorical variables dates back to the work of Darroch et al. (1980), where graphical log-linear models were introduced. Graphical models has been an efficient methodology for categorical data analysis and in this paper we focus on graphical models of conditional independence. Conditional independence is important when modelling highly complex systems.

A undirected graph $G = (\mathcal{V}, E)$ is characterized by a vertex set $\mathcal{V}$, where the cardinality of the set is K and every vertex corresponds to a classification variable of the contingency table, and an edge set E. The pair of nodes $\{i, j\} \in \mathcal{V}$ are adjacent if $(i, j) \in E$, and a subset $C \subset \mathcal{V}$ is said to be complete if all its elements are adjacent to each other. A complete subgraph that is maximal (i.e. not contained within another complete subgraph) is called a clique. An ordering of the cliques of an undirected graph, $(C_1, \ldots, C_{n_c})$ is said to be perfect if the vertices of each clique C_i also contained in any previous clique $C_1, C_2, \ldots C_{i-1}$ are all members of one previous clique; that is, for $i = 2, 3, \ldots, n_c$, $H_i = C_i \cap \cup_{j=1}^{i-1} C_j \subseteq C_j$ for some $h \in \{1, 2, \cdots, i-1\}$. The sets H_i, for $i = 1, \ldots, n_{c-1}$, are called separators. We write $S_1, \ldots, S_{n_s}$ the non-empty separators (some might appear multiple times). If an undirected graph admits a perfect ordering it is said to be decomposable. The widespread use of decomposable models is due to the resulting factorization of densities.

We associate to each vertex a random variable. For sets X, Y and $Z \subset \mathcal{V}$, X and Z are conditionally independent given Y, whenever X and Z are separated by Y. All paths in the graph connecting X and Z pass through Y. The three discrete random variables X, Y and $Z \in \mathcal{V}$ defined by undirected graphs and the cliques of the graph correspond to the maximal terms in log-linear model.

Dawid and Lauritzen (1993) described the Bayesian framework for decomposable models. Figure 4.1 shows such a model which also embodies the assumption that X and Z are independent given Y.

Figure 4.1: A decomposable undirected graphical model



The cliques of this graph are $\{X, Y\}$ and $\{Y, Z\}$ and there is a single intersection Y. The factorization form of the joint distribution is given by:

$$P(XZY) = P(XZ|Y)P(Y) = P(X|Y)P(Z|Y)P(Y) \tag{4.1}$$

In this section we focus on decomposable models in three-way contingency tales, for which the graph is chordal. We are going to deal with conjugate models where their parameter can be estimated without the utilization of iterative methods.

In three-way contingency table a graphical model of conditional independence can be parametrized in terms of cell-probabilities. For every three-way contingency table eight possible graphical models exist which can be represented by four different types of graphs:

- the independence
- the saturated
- the edge (only one edge)
- the gammma structure graph (a single path of length two)

The different types of graphs are represented in Figure 4.2.

Gamma model is characterized by its corner node Y which is connected with both variables X and Z. A gamma structured model implies that

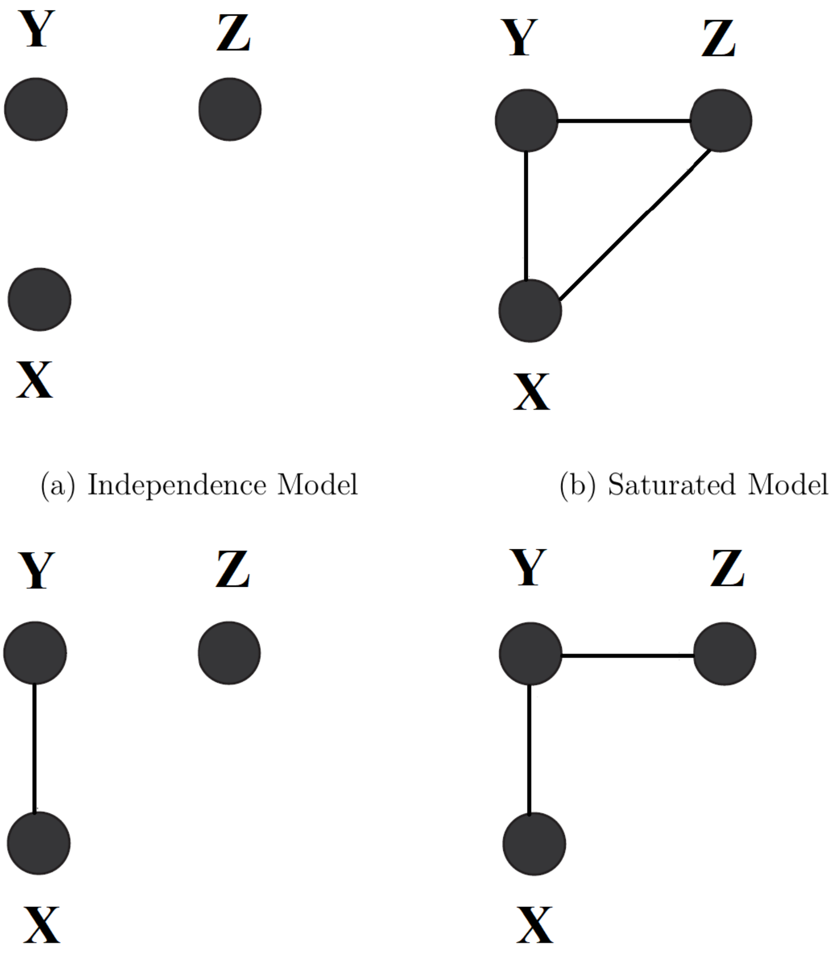
$$X \text{ is separated from } Z \text{ by } Y.$$

Then the likelihood is factorized by this expression

$$\begin{aligned} f(\boldsymbol{n}|\boldsymbol{\pi}^G, G) = & \prod_{j=1}^{|\mathcal{I}_Y|} \left(\prod_{i=1}^{|\mathcal{I}_X|} \Big[\pi_{X|Y}(ij)\Big]^{n_{ij}} \right) \\ & \prod_{j=1}^{|\mathcal{I}_Y|} \left(\prod_{z=1}^{|\mathcal{I}_Z|} \Big[\pi_{Z|Y}(z|j)\Big]^{n_{zj}} \right) \prod_{j=1}^{|\mathcal{I}_Y|} (\pi_Y(j))^{n_j} \end{aligned} \tag{4.2}$$

where $\boldsymbol{\pi}^G = \Big(\boldsymbol{\pi}_{X|Y}, \boldsymbol{\pi}_Y, \boldsymbol{\pi}_{Z|Y}\Big)$.

Figure 4.2: Type of Graphs in Three Way Tables.



(a) Independence Model

(b) Saturated Model

(c) Edge Model

(d) Gamma Model

4.3 Conjugate Priors

Next, we use conjugate priors on the probability parameters and then calculate the corresponding log-linear parameters using Monte Carlo schemes. For the specification of the prior distribution on the probability parameter vector we initially consider a Dirichlet distribution with parameters $\boldsymbol{\alpha} = \Big(\alpha(i),\, i \in \mathcal{I}\ \Big) = \Big(\alpha_{ijz},\, i = 1, \ldots, |\mathcal{I}_X|, j = 1, \ldots, |\mathcal{I}_Y|, z = 1, \ldots, |\mathcal{I}_Z|\ \Big)$ for the vector of the joint probabilities $\boldsymbol{\pi}$ of the full table, expanded to a (I x 1) vector, where $I = I_X * I_Y * I_Z$.

Hence, for the full table $\boldsymbol{\pi} \sim \mathcal{D}i(\boldsymbol{\alpha})$ with prior density given by

$$\begin{aligned} f(\boldsymbol{\pi}) &= \frac{\Gamma(\alpha)}{\prod\limits_{i\in\mathcal{I}} \Gamma\Big(\alpha(i)\Big)} \prod_{i\in\mathcal{I}} \pi(i)^{\alpha(i)-1} \\ &= \frac{\Gamma(\alpha)}{\prod\limits_{i=1}^{|\mathcal{I}_X|} \prod\limits_{j=1}^{|\mathcal{I}_Y|} \prod\limits_{z=1}^{|\mathcal{I}_Z|} \Gamma\Big(\alpha_{ijz}\Big)} \prod_{i=1}^{|\mathcal{I}_X|} \prod_{j=1}^{|\mathcal{I}_Y|} \prod_{z=1}^{|\mathcal{I}_Z|} \pi_{ijz}^{\Big(\alpha_{ijz}\ -1\Big)} = f_{\mathcal{D}i}(\boldsymbol{\pi};\ \boldsymbol{\alpha}) \end{aligned} \tag{4.3}$$

where $f_{\mathcal{D}i}\Big(\boldsymbol{\pi};\ \boldsymbol{\alpha}\Big)$ is the density function of the Dirichlet distribution evaluated at $\boldsymbol{\pi}$ with parameters $\boldsymbol{\alpha}$ and $\alpha = \sum_{i\in\mathcal{I}} \alpha(i)$, which control the precision and plays the role of prior information, volume the prior data like the sample size to the actual data.

When no prior information is available then we usually set all $\alpha(i) = \frac{\alpha}{|\mathcal{I}|}$ resulting to

$$E\Big[\pi(i)\Big] = \frac{1}{|\mathcal{I}|} \ \text{ and } \ V\Big[\pi(i)\Big] = \frac{|\mathcal{I}| - 1}{|\mathcal{I}|^2(\alpha + 1)}\ .$$

Small values of α increase the variance of each cell probability parameter. Usual choices for α are the values $|\mathcal{I}|/2$ (Jeffrey's prior), $|\mathcal{I}|$ and 1 (corresponding to $\alpha(i)$ equal to 1/2, 1 and $1/|\mathcal{I}|$ respectively); for details see Dellaportas and Forster (1999). The choice of this prior parameter value is of prominent importance for the model comparison due to the well-known sensitivity of the posterior model odds and the Bartlett-Lindley paradox (Lindley, 1957, Bartlett, 1957). Here for two reasons this effect is not so adverse, as in usual variable selection for generalized linear models. Firstly, even if we consider the limiting case where $\alpha(i) = \frac{\alpha}{|\mathcal{I}|}$ with $\alpha \to 0$, the variance is finite and equal to $(|\mathcal{I}| - 1)/|\mathcal{I}|^2$. Secondly, the distributions of all models are constructed from a common distribution of the full model/table making the prior distributions 'compatible' across different models (Dawid and Lauritzen, 2000 and Roverato and Consonni, 2004).

A Dirichlet distribution with parameters $\boldsymbol{\alpha} = \Big(\alpha(i),\, i \in \mathcal{I}\Big) = \Big(\alpha_{ijz},\, i = 1, \ldots, |\mathcal{I}_X|, j = 1, \ldots, |\mathcal{I}_Y|, z = 1, \ldots, |\mathcal{I}_Z|\Big)$ for the vector of the joint probabilities of the full table $\boldsymbol{\pi}$.

$$
\begin{aligned}
p\left(\boldsymbol{\pi}^G\right) = \prod_{j=1}^{|\mathcal{I}_Y|} &\left\{ \frac{\Gamma\Big(\alpha_{X|Y}\Big)}{\prod_{i=1}^{|\mathcal{I}_X|} \Gamma\Big(\alpha_{X|Y}(i|j)\Big)} \left(\prod_{i=1}^{|\mathcal{I}_X|} \Big[\pi_{X|Y}(i|j)\Big]^{\alpha_{X|Y}(i|j)-1} \right) \right\} \\
\times \prod_{j=1}^{|\mathcal{I}_Y|} &\left\{ \frac{\Gamma\Big(\alpha_{Z|Y}\Big)}{\prod_{z=1}^{|\mathcal{I}_C|} \Gamma\Big(\alpha_{Z|Y}(z|j)\Big)} \left(\prod_{z=1}^{|\mathcal{I}_Z|} \Big[\pi_{Z|Y}(z|j)\Big]^{\alpha_{Z|Y}(z|j)-1} \right) \right\} \\
\times &\frac{\Gamma\left(\alpha_Y\right)}{\prod_{j=1}^{|\mathcal{I}_Y|} \Gamma\Big(\alpha_Y(j)\Big)} \prod_{j=1}^{|\mathcal{I}_Y|} \pi_j^{\Big(\alpha_Y(j)\ -1\Big)}
\end{aligned}
\tag{4.4}
$$

where $\alpha_{X|Y} = \sum_{i=1}^{|\mathcal{I}_X|} \alpha_{X|Y}(i|j)$ and $\alpha_{Z|Y} = \sum_{z=1}^{|\mathcal{I}_Z|} \alpha_{Z|Y}(z|j)$.

4.4 Power Priors

The debate in graphical models literature concerns the use of conjugate priors based on Dirichlet distributions; see for example in Steck and Jaakkola (2002), Steck (2008) and Ueno (2008). It is clear that the parameters of the Dirichlet prior should be carefully specified and in this thesis we adopt ideas based on the power prior approach of Ibrahim and Chen (2000) and Chen *et al.* (2000). We use their approach to advocate sensible values for the Dirichlet prior parameters on the full table and the corresponding induced values for the rest of the graphs.

Imaginary set of data $\boldsymbol{n}^* = (n^*(i), i \in \mathcal{I})$: the table expanded to a vector of the imaginary data $n^* = \sum_{i \in \mathcal{I}} n^*(i)$: the total sample size of the imaginary data

$f_{\mathcal{D}i}\Big(\boldsymbol{\pi};\ \alpha(i) = \alpha_0, i \in \mathcal{I}\Big)$: a Dirichlet 'pre-prior' with all parameters equal to α_0.

Then unnormalized prior distribution can be obtained by:

$$\begin{aligned} f(\boldsymbol{\pi}) &\propto f(\boldsymbol{n}^*|\boldsymbol{\pi})^w \times f_{\mathcal{D}i}\Big(\boldsymbol{\pi};\ \alpha(i) = \alpha_0, i \in \mathcal{I}\Big) \\ &\propto \prod_{i \in \mathcal{I}} \pi(i)^{w\, n^*(i) + \alpha_0 - 1} \\ &= f_{\mathcal{D}i}\Big(\boldsymbol{\pi};\ \alpha(i) = w\, n^*(i) + \alpha_0, i \in \mathcal{I}\Big) . \end{aligned} \tag{4.5}$$

- $w = 1$: each imaginary observation has the same weight as the observations
- $w < 1$: each imaginary observation less weight than the observations
- $w > 1$: will increase the weight of believe on the prior/imaginary data
- $w = 1$, $n^* = n$ and $\alpha_0 \to 0$: both the prior and data will account for 50% of the information used in the posterior
- $w = 1/n^*$, $\alpha(i) = p^*(i) + \alpha_0$, $p^*(i) = n^*(i)/n^*$: the prior data $\boldsymbol{n}^*$ will account for information of one data point. Information additional to one data point will be introduced by pre-prior parameters α_0.
- $w = 1/n^*$, $\alpha(i) = p^*(i) + \alpha_0$, $p^*(i) = n^*(i)/n^*$, $\alpha_0 = 0$: Unit information prior (UIP), equivalent to a single observation. No pre-prior information is introduced.
- When no information is available, set $n^*(i) = n^*$, $n^* = n^* \times |\mathcal{I}|$ and $w = 1/n^* = \frac{1}{n^* \times |\mathcal{I}|}$ resulting to

  $$\boldsymbol{\pi} \sim \mathcal{D}i\Big(\alpha(i) = 1/|\mathcal{I}|,\ i \in \mathcal{I}\Big) .$$

  This is a UIP with zero pre-prior information and uniform distributed prior cell

4.4.1 Specification of Prior Parameters Using Imaginary Data

Let us consider an imaginary set of data represented by the frequency table $\boldsymbol{n}^* = (n^*(i), i \in \mathcal{I})$ of total sample size $n^* = \sum_{i \in \mathcal{I}} n^*(i)$ and a Dirichlet 'pre-prior' with all parameters equal to α_0. Then the unnormalized prior distribution can be obtained by the product of the likelihood of $\boldsymbol{n}^*$ raised to a power w multiplied by the 'pre-prior' distribution. Hence

$$\begin{aligned} f(\boldsymbol{\pi}) &\propto f(\boldsymbol{n}^*|\boldsymbol{\pi})^w \times f_{\mathcal{D}i}\Big(\boldsymbol{\pi};\ \alpha(i) = \alpha_0, i \in \mathcal{I}\Big) \\ &\propto \prod_{i \in \mathcal{I}} \pi(i)^{w\, n^*(i) + \alpha_0 - 1} \\ &= f_{\mathcal{D}i}\Big(\boldsymbol{\pi};\ \alpha(i) = w\, n^*(i) + \alpha_0, i \in \mathcal{I}\Big)\ . \end{aligned} \tag{4.6}$$

Using the above prior set up, we a priori expect to observe a total number of $w\, n^* + |\mathcal{I}|\alpha_0$ observations. The parameter w is used to specify the steepness of the prior distribution and the weight of belief on each prior observation. For $w = 1$ each imaginary observation has the same weight as the actual observation. Values of $w < 1$ will give less weight to each imaginary observation while $w > 1$ will increase the weight of believe on the prior/imaginary data. Overall, the prior will account for the $(w\, n^* + |\mathcal{I}|\alpha_0)/(w\, n^* + n + |\mathcal{I}|\alpha_0)$ of the total information used in the posterior distribution. Hence for $w = 1$, $n^* = n$ and $\alpha_0 \to 0$ then both the prior and the data will account for 50% of the information used in the posterior.

For $w = 1/n^*$ then $\alpha(i) = p^*(i) + \alpha_0$ with $p^*(i) = n^*(i)/n^*$, the prior data $\boldsymbol{n}^*$ will account for information of one data point while the total weight of the prior will be equal to $(1 + |\mathcal{I}|\alpha_0)/(1 + n + |\mathcal{I}|\alpha_0)$. If we further set $\alpha_0 = 0$, then the prior distribution (4.5) will account for information equivalent to a single observation. This prior set-up will be referred in this paper as the unit information prior (UIP). When no information is available, then we may further consider the choice of equal cell frequencies $n^*(i) = \xi$ for the imaginary data in order to support the simplest possible model under consideration. Under this approach $n^* = \xi \times |\mathcal{I}|$ and $w = 1/n^* = \frac{1}{\xi \times |\mathcal{I}|}$ with a Dirichlet pre-prior with all parameters equal to $\alpha_0 \rightarrow$, resulting to $\boldsymbol{\pi} \sim \mathcal{D}i\Big(\alpha(i) = 1/|\mathcal{I}|,\, i \in \mathcal{I}\Big)$. The latter prior is equivalent to the one advocated by Perks (1947).

4.4.2 Comparison of Prior Set-ups

Since Perks' prior (with $\alpha(i) = 1/|\mathcal{I}|$) has a unit information interpretation, it can be used as a yardstick in order to identify and interpret the effect of any other prior distribution used. Prior distributions with $\alpha(i) < 1/|\mathcal{I}|$, or equivalently $\alpha < 1$, result in larger variance than the one imposed by our proposed unit information prior and hence they a posteriori support models, which are more parsimonious. On the contrary, prior distributions with $\alpha(i) > 1/|\mathcal{I}|$, or $\alpha > 1$, result in lower prior variance and hence they a posteriori support models with more complicated graphical structure. So the variance ratio between a Dirichlet prior with $\alpha(i) = \alpha/|\mathcal{I}|$ and Perks prior is equal to

$$VR = \frac{V\Big(\pi(i)\Big|\ \alpha(i) = \frac{\alpha}{|\mathcal{I}|}\Big)}{V\Big(\pi(i)\Big|\ \alpha(i) = |\mathcal{I}|^{-1}\Big)} = \frac{2}{\alpha+1}\ .$$

In this chapter we considered the comparison of the information from the following prior choices:

(i) the Jeffrey's prior with $\alpha(i) = 1/2$;

(ii) the Unit Expected Cells prior (UEC) with $\alpha(i) = 1$;

(iii) the Unit Information Prior (UIP) which is derived by a power prior with $\alpha(i) = p^*(i)$, $w = 1/n^*$ and $a_0 = 0$; where $p^*(i)$ is the sample proportion of cell i estimated from a set of imaginary data $n^*(i)$;

- Perks' prior (UIP-Perks') with $\alpha(i) = 1/|\mathcal{I}|$ which is equivalent to UIP coming from a table of imaginary data with all cell frequencies equal to one;

- the Unit Information Empirical Bayes Prior (UI-EBP), which is derived by UIP with $p^*(i)$ set equal to the sample proportions $p(i) = n(i)/n$. This will inflate the actual data by a factor of $\frac{n+1}{n}$

It is observed that Jeffreys' prior variance is lower than the corresponding Perks' prior. The reduction is even greater for the Unit Expected Cell prior reaching. Finally, for the Empirical Bayes prior, based on the UIP approach, the prior variance for each $\pi(i)$ is equal to $V[\pi(i)] = \frac{1}{2}p(i)\Big(1 - p(i)\Big)$. Hence it depends on the observed proportion and can vary from zero (if $p(i) = 0$ or 1) to 1/8 if $p(i) = 1/2$. For values in the interval $(0.058, 0.942)$ the variance of the UI-EBP is higher than the corresponding UIP variance reaching its maximum when $p(i) = 1/2$ where it is 4.6 times the corresponding UIP prior variance. For $p(i) = 0.058$ or 0.942 the variances of the UIP and UI-EBP are equal while for the remaining values, UIP variance is higher. Table 4.1 summarizes five of the most common prior setups and shows the comparison of the information under the following prior choices.

Table 4.1: A summary of the five most common prior setups.

$\boldsymbol{Prior}$	$\boldsymbol{Parameter}$	$\boldsymbol{V[\pi(i)]}$	$\boldsymbol{n^*(i) = \xi}$
$\boldsymbol{Jeffreys}$	$\alpha(i) = \frac{1}{2}$	$2\dfrac{\lvert I\rvert - 1}{\lvert I\rvert^2\lvert I\rvert + 2}$	$\xi = \frac{4}{\lvert I\rvert+2}$
$\boldsymbol{UnitExpectedCell(UEC)}$	$\alpha(i) = 1$	$\frac{\lvert I\rvert-1}{\lvert I\rvert^2(\lvert I\rvert+1)}$	$\xi = \frac{2}{\lvert I\rvert+1}$
$\boldsymbol{UnitInformationPrior(UIP)}$	$\alpha(i) = p^*(i)$	$\frac{1}{2}p^*(i)\Big(1 - p^*(i)\Big)$	$\xi = \dfrac{\lvert I\rvert^2}{\lvert I\rvert - 1}p^*(i)\Big(1 - p^*(i)\Big)$
$\boldsymbol{Perk's Prior(UIP - Perks)}$	$\alpha(i) = \frac{1}{\lvert I\rvert}$	$\frac{\lvert I\rvert-1}{2\lvert I\rvert^2}$	$\xi = 1$
$\boldsymbol{UnitInformationEmpirical}$ $\boldsymbol{BayesPrior(UI - EBP)}$	$\alpha(i) = p(i)$	$\frac{1}{2}p(i)\Big(1 - p(i)\Big)$	$\xi = \frac{\lvert I\rvert^2}{\lvert I\rvert-1}V\Big(\pi(i)\Big)$

4.4.3 Marginal Likelihood of Each Graphical Model

The marginal likelihood can be calculated analytically since the above prior set-up is conjugate.

For the saturated model the marginal likelihood is given by

$$f(\boldsymbol{n}|G) = C(\boldsymbol{n}) \times \frac{B(\tilde{\boldsymbol{\alpha}})}{B(\boldsymbol{\alpha})}$$

where $C(\boldsymbol{n})$ is the multinomial constant and $B(\boldsymbol{\alpha})$ is the normalizing constant of the multinomial beta function given by $B(\boldsymbol{\alpha}) = \frac{\prod_{i \in I} \Gamma\Big(\alpha(i)\Big)}{\Gamma\Big(\sum_{i \in I} \alpha(i)\Big)}$. For the independence model the marginal likelihood is given by

$$f(\boldsymbol{n}|G) = C(\boldsymbol{n}) \times \frac{B(\tilde{\boldsymbol{\alpha}}_X)B(\tilde{\boldsymbol{\alpha}}_Y)B(\tilde{\boldsymbol{\alpha}}_Z)}{B(\boldsymbol{\alpha}_X)B(\boldsymbol{\alpha}_Y)B(\boldsymbol{\alpha}_Z)}$$

while for the edge model the marginal likelihood is calculated by

$$f(\boldsymbol{n}|G) = C(\boldsymbol{n}) \times \frac{B(\tilde{\boldsymbol{\alpha}}_Z)B(\tilde{\boldsymbol{\alpha}}_{XY})}{B(\boldsymbol{\alpha}_Z)B(\boldsymbol{\alpha}_{XY})}$$

Finally, for the gamma structure the marginal likelihood is given by

$$f(\boldsymbol{n}|G) = C(\boldsymbol{n}) \times \frac{B(\tilde{\boldsymbol{\alpha}}_{X|Y})B(\tilde{\boldsymbol{\alpha}}_{Z|Y})B(\tilde{\boldsymbol{\alpha}}_Y)}{B(\boldsymbol{\alpha}_{X|Y})B(\boldsymbol{\alpha}_{Z|Y})B(\boldsymbol{\alpha}_Y)}$$

.

4.4.4 Illustrative examples

We consider a data set presented by Healy (1988) regarding a study on the relationship between patient condition (more or less severe), assumption of antitoxin (yes or not) and survival status (survived or not); see Table 4.2. In Table 4.3 we compare posterior model probabilities under the four different prior set-ups.

Table 4.2: Antitoxin data

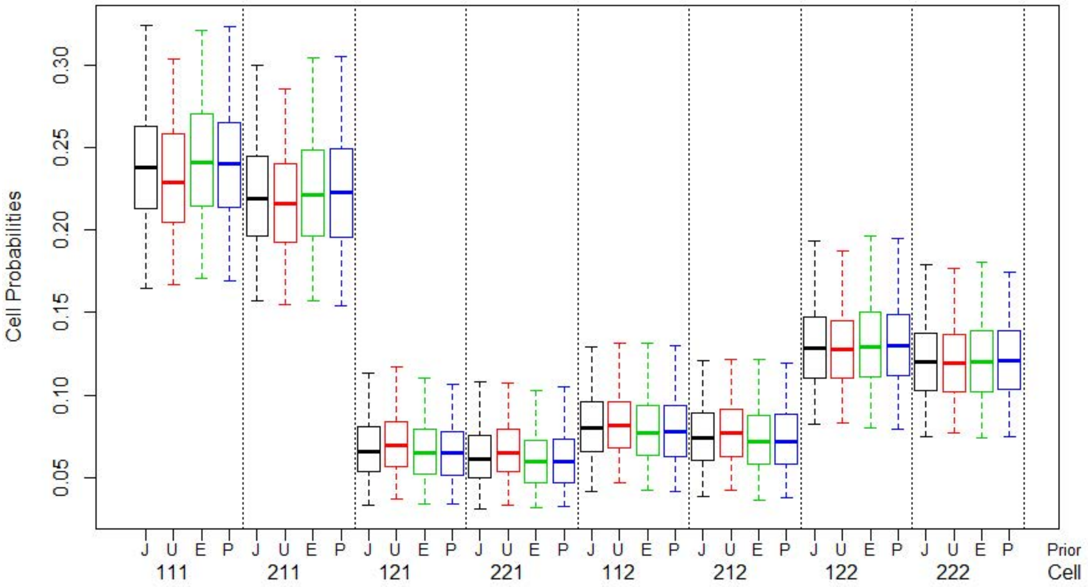
		Survival (Y)	
Condition (Z)	Antitoxin (X)	No	Yes
More Severe	Yes	15	6
	No	22	4
Less Severe	Yes	5	15
	No	7	5

Table 4.3: Posterior model probabilities (%) for the Antitoxin data for Jeffreys', Unit Expected Cell, Empirical Bayes, UIP-Perks' and Dellaportas and Foster prior set-up

Model	***Prior Disribution***				
	Jeffreys'	*UEC*	*Empirical*	*Perks'*	*DF*
X+Y+Z	0.09	0.07	0.62	0.42	0.17
XY+Z	0.41	0.36	0.93	0.75	0.53
XZ+Y	0.06	0.06	0.13	0.10	0.07
YZ+X	15.88	12.24	36.09	32.51	2.83
XY+XZ	0.25	0.31	0.20	0.17	0.21
XY+YZ	69.99	67.69	54.30	58.38	67.17
XZ+YZ	9.78	10.63	7.59	7.39	8.82
XYZ	3.55	8.65	0.14	0.28	1.20

We compare the results obtained with our yardstick prior, the UIP-Perks' prior ($\alpha(i) = 1/|I|$), with those obtained using Jeffrey's ($\alpha(i) = 1/2$), Unit Expected Cell ($\alpha(i) = 1$), and unit information Empirical Bayes ($\alpha(i) = p(i)$) priors. Under all prior assumptions the maximum a posterior model (MAP) is $XY + YZ$, assuming that the Antitoxin is conditional independence from Condition given Survival. Under Empirical Bayes and UIP-Perks priors the posterior distribution

Figure 4.3: Antitoxin data: Boxplots summarizing 2.5%, 97.5% posterior percentiles and quantiles of the joint probabilities $\pi_{XYZ}(i,j,k)$ for the MAP model (YZ+X) for all prior set-ups (J=Jeffreys', U=Unit Expected Cell, E=Empirical Bayes, P=Perks') .



is concentrate on the MAP model. It takes into account 69.99% and 67.69%, respectively of the posterior model probabilities. The posterior model probabilities under Jeffreys and the Unit Expected prior setups are lower, 54.30% and 58.38%, respectively. The edge model $YZ + X$, where Antitoxin is independent from the connected variables Survival and Condition, is the model with the second highest posterior probabilities under UIP-Perks and the Empirical prior with posterior model probabilities equal to 36% and 32.5% and for Jeffreys and UEC 15.9% and 12.2%, respectively. So, we can see here that the four prior setups are grouped into two pairs, the fisrt ywo and the last two. Figure 4.3 presents box-plots summarizing 2.5% and 97.5% posterior percentiles and quantiles of the joint probabilities for the MAP model $(XY + YZ)$ for the four prior setups. From the Figure 4.3 we observe minor deferences between the posterior distributions obtained under the

Unit Expected cell and the posterior distributions under the three other prior setups. Differences are higher for the first two cell probabilities $\pi(1,1,1)$ and $\pi(2,1,1)$. from the graph we can see that within each model the posterior results are robust no matter prior we use.

4.5 Discussion

We use the power approach with imaginary data for prior specification. Unit information interpretation priors are used as a yardstick. We employ this framework to interpret standard prior choices (Jeffreys, Empirical, Perks, Dellaportas-Foster) previously used in graphical models and their effect on model comparison. The approach is general for any dimension.

In the three way case all the considered models are Markov equivalent to a DAG; see Pearl and Wermuth (1994) and Drton and Richardson (2008). An immediate question which arises is whether the the graphical structures implied using the parameterization illustrated in this paper are the same with the ones that we would derive using the parameterization implied by the corresponding DAG representation. For example, in our approach the parameters for the edge model $\{XZ,\ Y\}$ are given by $\pi^G = (\pi_{XZ}, \pi_Y)$ while the parameterization implied by the corresponding DAG structure is either $\pi^G = (\pi_{X|Y}, \pi_Y, \pi_Z)$. The answer is given by Heckerman et al. (2013) Heckerman et al. (1995) where they prove that the posterior distributions and the marginal likelihoods will be the same if the priors are compatible across models since some normalizing constants cancel out. This result can be easily confirmed in the above imple example with model $\{XZ, Y\}$.

An obvious extension of this work is to implement the same approach in tables of higher dimension starting from four way tables. Although most of the models in a four way contingency table can be factorized and analysed in a similar manner, two type of graphs (the 4-chain and the cordless four-cycle graphs) cannot be decomposed in the above way. These models are not Markov equivalent to any directed acyclic graph (DAG). In fact each bi-directed graph (which corresponds to a marginal association model) is equivalent to a DAG, i.e. a conditional association model, with the same set of variables if and only if it does not contain any 4-chain, see Pearl and Wermuth (1994). We believe that also in higher dimensional problems our approach can be applied to bi-directed graphs that admit a DAG representation. For the graph that do not factorize, more sophisticated techniques must be adopted in order to obtain the posterior distribution of interest and the corresponding marginal likelihood needed for the model comparison.

Chapter 5

Discussion and Concluding Remarks

> Our goal is to work against things we have not seen that we will see in the near future
>
> *Nick Saban*

5.1 Conclusions, Discussion and Future Work

In this thesis, an efficient Bayesian model selection procedure is proposed for contingency tables using the power prior approach. When no information is available for the problem in hand the utilization of imaginary data is proposed as a possible way to specify an objective data-driven approach. The power prior has been constructed from imaginary data and its use has been demonstrated in detail for association models. This class of priors is useful for controlling the sensitivity of the Bayes factor due to the prior choice. By this way, the prior specification procedure is automated for all models under consideration. Moreover power priors are quite robust under a variety of model setups. The proposed methodology can be described by the following steps: (a) consider data from the uniform constant model (i.e. the simplest model we may consider for this problem), (b) build the prior and (c) compute the Bayes factor utilizing the two versions of Perrakis et al.

(2014) estimator. The simulation study results show that the Bayes factors of association models are not sensitive to the choice of prior parameter when power priors are used. Two prior setups are introduced in order to advocate sensible prior choices avoiding the Lindley-Bartlett paradox. The first prior scenario is to set all the cell counts equal to one and the second scenario equal to the mean of the cell counts in order to asses whether the size of imaginary data influence the past results. The use of imaginary data combined with the real data using the power prior approach makes the proposed method computationally efficient and straightforward to implement. In summary we build in this thesis an objective default prior using the imaginary data and the power prior approach. Due to the utilization of common imaginary data we achieve compatibility of priors across models under consideration.

In order to compute the Bayes factor, the computation of the marginal likelihood is required. Here we have proposed two versions of Monte Carlo estimators of the marginal likelihood based on the importance estimator of Perrakis et al. (2014): the one-block and the independent Perrakis estimators. In both cases we use the product marginal posteriors as importance sample density. Both of these proposed MCMC marginal likelihood estimators are straightforward to implement even in complex setups as they are obtained from a non-iterative approach from the MCMC output. Moreover they do not require to adapt or adjust the MCMC algorithm. As illustrated the one block estimator is accurate in computing the marginal likelihood under the several prior setups in Section 3 delivering accurate estimates even in sparse data.

In contingency tables with zero counts, the computation of the marginal likelihood of the saturated model has proven challenging. In this dissertation we have identified and initially dealt with the problem. We have implemented different techniques in order to estimate the corresponding marginal likelihood using the conjugate approach (with slightly different prior), which seems to be the most appropriate approach which avoids large MCMC errors due to zeros. We came to the conclusion as in Greenland (2001) and Giudici (1998) where conjugate prior setups were used in sparse contingency tables. This issue has also discussed in detail in Agresti and Hitchcock (2005a). The prior parameters of the Gamma distribution in the Poisson-Gamma conjugate analysis for the saturated model in the full table specified by the power prior approach. Even this prior approach for the saturated model is not exactly the same as the proposed methodology, it results from the same idea of unit information using the same imaginary data and hence it conveys similar amount of information (equivalent to one data point). Hence, the two approaches should be intuitively close in terms of model selection.

A Bayesian analysis for graphical models of conditional independence for three way contingency tables using power prior has proposed in Section 4. This method is an extension of the approach of Tarantola and Ntzoufras (2012). Each conditional independence model blend with a particular parametrization of the cell probabilities and a conjugate Bayesian analysis based on Dirichlet prior is proposed. The parametrization that imposed proceed from the constraints under the conditional association structure of a graph. The resulting parametrization and the corresponding decomposition of the likelihood simplifies the problem and automatically imposes the conditional independences represented by the considered graph. By this way the posterior model probabilities and the posterior distribution for the parameters can be analytically calculated. The proposed method is straightforward, fast and automatically applicable to any type of datasets, even in sparse contingency tables. Under this perspective we compare the results under different priors which fit in this framework.

Our proposed methodology is easily to handle and detect the true model in automatic way using the Bayes factor. From the simulation study we saw the the criterion of the model selection consistency is satisfied and the proposed methodology achieve to detect the true model. Nevertheless in some simulation scenarios were required large sample sizes in order to detect the true model, so in the future we want to tune the prior parameters in order to improve the properties of model selection consistency criterion.

5.2 Discussion

The need to work without introducing subjective inputs into the Bayesian analysis led to the growth of Objective Bayes techniques. In this thesis we accomplished the goal of finding a prior satisfying two of the basic Objective Bayes principles, consistency and compatibility of the prior across the model, which are proved via the extended simulation study in Section 3.2. We test the proposed methodology under the consistency criterion and the results show a good performance. Additionally methods for constructing objective prior distribution are applied in the proposed Bayesian model comparison procedure. More specifically, the unit information principle is identified by the utilization of unit information prior (UIP) and the combination with the power prior approach, setting the prior mean equal with the imaginary data mean. One of the main approaches used to construct prior distributions for objective Bayes methods is the concept of imaginary observation in combination with the power prior approach.

5.2.1 Further Considerations about the presence of zeros

The estimation of the marginal likelihood in contingency tables under the presence of zeros ***Conjugate Analysis*** is used for the saturated model, getting all parameters from the full table. The density of the likelihood distribution for the saturated model is

$$f(\boldsymbol{n}|\lambda_{ij}) = \prod_{i=1}^{I}\prod_{j=1}^{J}\frac{e^{-\lambda_{ij}}\lambda_{ij}^{n_{ij}}}{n_{ij}!} \tag{5.1}$$

Let us consider a set of imaginary data all equal to one $n^*_{ij} = 1$ and $n^* = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} = IJ$ and a Gamma pre-prior. Then the unnormalized prior distribution can be obtained by the product of the of the likelihood raised to a power $w = \frac{1}{n}$ multiplied by the pre-prior distribution

$$f(\boldsymbol{\lambda}) = f(\lambda_{ij}) = f(\lambda_{ij}|n^*_{ij}) \propto \Big(\prod_{i=1}^{I}\prod_{j=1}^{J} e^{-\lambda_{ij}}\lambda_{ij}^{n_{ij}}\Big)^{\frac{1}{n}} f_0(\lambda_{ij}) \tag{5.2}$$

where $f_0(\lambda_{ij})$ is a pre-prior. Here we consider $f_0 \sim Gamma(a_0,\ b_0)$ and $n^*_{ij} = 1$. For $a_0 = b_0 = 1$ we obtain

$$\lambda_{ij} \sim Gamma(\frac{1}{IJ}, \frac{1}{IJ}) \tag{5.3}$$

For the saturated model the posterior likelihood is given by

$$f(\lambda_{ij}|\boldsymbol{n}) \propto \frac{e^{-\lambda_{ij}}\lambda_{ij}^{n_{ij}} e^{-\frac{\lambda_{ij}}{IJ}} \lambda^{\frac{1}{IJ}-1}}{\Gamma(\frac{1}{IJ})} \tag{5.4}$$

For the saturate model the marginal likelihood is given by

$$\begin{aligned} f(n_{ij}|\lambda_{ij}) &= \Big(\frac{1}{IJ}\Big)^{\frac{1}{IJ}} \int \frac{e^{-\lambda_{ij}}\lambda_{ij}^{n_{ij}} e^{-\frac{\lambda_{ij}}{IJ}} \lambda^{\frac{1}{IJ}-1}}{\Gamma(\frac{1}{IJ})} d\lambda_{ij} \\ f(\boldsymbol{n}) = \prod_{i=1}^{I}\prod_{j=1}^{J} f(n_{ij}) &= \prod_{i=1}^{I}\prod_{j=1}^{J} \frac{\Big(\frac{1}{IJ}\Big)^{\frac{1}{IJ}}}{\Gamma(\frac{1}{IJ})} \frac{\Gamma(n_{ij}+\frac{1}{IJ})}{\Big(1+\frac{1}{IJ}\Big)^{n_{ij}+\frac{1}{IJ}}} \end{aligned} \tag{5.5}$$

and the logarithm of the marginal likelihood is given by

$$\begin{aligned} log(f(\boldsymbol{n})) &= \sum \frac{1}{IJ} log(\frac{1}{IJ}) - \sum_{i=1}^{I}\sum_{j=1}^{J}(n_{ij} + \frac{1}{IJ}) log(1+\frac{1}{IJ}) \\ &+ \sum_{i=1}^{I}\sum_{j=1}^{J} log(\Gamma(n_{ij}+\frac{1}{IJ})) - \sum_{i=1}^{I}\sum_{j=1}^{J} log(\Gamma(\frac{1}{IJ})) \\ log(f(\boldsymbol{n})) &= -log(IJ) - (N+1) log(1+\frac{1}{IJ}) + \sum_{i=1}^{I}\sum_{j=1}^{J} log(\Gamma(n_{ij}+\frac{1}{IJ})) + IJlog(IJ) \end{aligned} \tag{5.6}$$

With this approach one term is explode and the result are not compatible with the proposed methodology. The identification of the prior mean and variance provides one possible solution or the utilization of other prior parameters.

5.3 Future Work

An obvious extension of our work is to the Bayesian model selection procedure between association models and graphical models and test the type of association in three or tables of higher dimension.

In the future we want to expand and use the methodology of the two proposed marginal likelihood estimators, the independent and the one-block Perrakis estimators, in higher dimensional problems, using the product marginal divided by blocks. In three way contingency tables or higher the multi block parameter approach will be a solution for computational demands. With some additional effort one could consider to incorporate with bridge sampling estimation using the product marginal as a approximation density for categorical data in contingency table form.

Another interesting introduction to our proposed methodology could be the utilization of random imaginary data. This class of prior distribution treats imaginary data as stochastic components. Fouskakis et al. (2015) introduced the power-expected-posterior (PEP) prior approach in order to alleviate the amount of information introduced by the size of the training dataset. They combine the idea of power prior method with the unit information prior approach in order to procedure a minimal informative prior, and at the same time to reduce the effect of training sample.

With respect to model selection consistency, detailed empirical evidence via extended simulation study for association model using power priors in contingency tables is provided in this thesis. Our intension for the future is to further investigate whether empirical evidence also suggests that the rest of the criteria of objective Bayesian model comparison are also valid.

Appendix A

Results

A.1 Convergence diagnostics with R package boa

In order to test the converge of the MCMC algorithm we use the R package **boa**, which is a program for carrying out convergence diagnostics and statistical and graphical analysis of Monte Carlo sampling output and can be used as a post-processor for the WinBUGS software. We applied three diagnostic methods for MCMC sampler output: the Geweke, the Heidelberger and Welch and the Raftery and Lewis test.

In Geweke diagnostic Geweke (1992) the convergence is assessed by comparing the sample mean in an early segment of the chain to the mean in a later segment and is valid when the two segments are independent. The statistic has the general form

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\hat{S}_1(0)/n_1 - \hat{S}_2(0)/n_2}}$$

where the variance estimate $\hat{S}(0)$ is calculated as the spectral density at frequency zero to account for serial correlation in the sampler output. If the two segments are from the same stationary distribution, the limiting distribution for this statistic is a standard normal. The p-value is a measure of evidence against the two sequences being from a common stationary distribution.If statistical significance is assess at the 5% level, these results would be deemed non-significant the diagnostic does not provide evidence of non-convergence.

In Heidelberger and Welch (1983) diagnostic an estimate of the number of samples that should be discarded as a burn-in sequence and a formal test for non-convergence are provided. Given an MCMC chain, the null hypothesis of convergence uses the Cramer-von-Mises test statistic

$$\int_0^1 B_n(t)^2 dt$$

where

$$B_n(t) = \frac{T_{nt} - nt\bar{x}}{\sqrt{nS(0)}} \quad \text{with} \quad T_k = \begin{cases} 0, & k = 0 \\ \sum_{j=1}^{k} x_j, & k \geqslant 1 \end{cases}$$

and $S(0)$ is the spectral density evaluated at frequency zero, given an MCMC chain $\{x_j : \ j = 1, \ \ldots, \ T\}$. In calculating the test statistic, the spectral density is estimated from the second half of the original chain. If the null hypothesis is rejected, then the first $0.1n$ of the samples are discarded and the test reapplied to the resulting chain. This processes is repeated until the test is either non-significant or 50% of the samples have been discarded, at which point the chain is declared to be non-stationary. If convergence is not rejected in the final step, a half-width test is performed by computing the mean and associated $(1 - \alpha)100\%$ confidence interval. This test is passed if the half-width of the confidence interval is less than a user-specified level of accuracy ε, otherwise the test is failed.

The third diagnostic we used to test the convergence is the Raftery and Lewis (1992) method, which estimate the number of MCMC samples needed when quantiles are the posterior summaries of interest. A summary of the results of the three diagnostics are provided in Tables A1-A6:

Table A.1: Summary results of convergence diagnostics: ***Independence Model***

Parameters	***Geweke***		***Heidelberger & Welch***		***Raftery & Lewis***
	p-value	*pass*	*Stationarity Test*	*Halfwidth Test*	*MCMC samples*
m	0.041	✗*	passed	passed	3776
a[1]	0.195	✓	passed	passed	3820
a[2]	0.338	✓	passed	passed	3724
a[3]	0.497	✓	passed	passed	3808
a[4]	0.371	✓	passed	passed	3832
a[5]	0.537	✓	passed	passed	3771
b[1]	0.189	✓	passed	passed	3824
b[2]	0.679	✓	passed	passed	3945
b[3]	0.017	✗*	passed	passed	3865
b[4]	0.829	✓	passed	passed	3759
b[5]	0.249	✓	passed	passed	3815

* Failed for $\alpha = 0.05$ but passed for $\alpha = 0.10$

Table A.2: Summary results of convergence diagnostics: ***Uniformm Model***

Parameters	***Geweke***		***Heidelberger & Welch***		***Raftery & Lewis***
	p-value	*pass*	*Stationarity Test*	*Halfwidth Test*	*MCMC samples*
m	0.599	✓	passed	passed	3780
a[1]	0.748	✓	passed	passed	3751
a[2]	0.071	✓	passed	passed	3641
a[3]	0.312	✓	passed	passed	3669
a[4]	0.084	✓	passed	passed	3708
a[5]	0.801	✓	passed	passed	3798
b[1]	0.693	✓	passed	passed	3943
b[2]	0.492	✓	passed	passed	3865
b[3]	0.776	✓	passed	passed	3810
b[4]	0.995	✓	passed	passed	3696
b[5]	0.951	✓	passed	passed	3736
phi	0.652	✓	passed	passed	6311

* Failed for $\alpha = 0.05$ but passed for $\alpha = 0.10$

Table A.3: Summary results of convergence diagnostics: ***Row Model***

Parameters	***Geweke***		***Heidelberger & Welch***		***Raftery & Lewis***
	p-value	*pass*	*Stationarity Test*	*Halfwidth Test*	*MCMC samples*
m	0.934	✓	passed	passed	3704
a[1]	0.971	✓	passed	passed	3724
a[2]	0.185	✓	passed	passed	3774
a[3]	0.042	✗*	passed	passed	3792
a[4]	0.189	✓	passed	passed	3695
a[5]	0.129	✓	passed	passed	3798
b[1]	0.665	✓	passed	passed	3744
b[2]	0.782	✓	passed	passed	3839
b[3]	0.437	✓	passed	passed	3877
b[4]	0.351	✓	passed	passed	3783
b[5]	0.115	✓	passed	passed	3653
mu[2]	0.437	✓	passed	passed	3696
mu[3]	0.291	✓	passed	passed	3808
mu[4]	0.226	✓	passed	passed	3780
mu[5]	0.221	✓	passed	passed	3759

* Failed for $\alpha = 0.05$ but passed for $\alpha = 0.10$

Table A.4: Summary results of convergence diagnostics: ***Column Model***

Parameters	***Geweke***		***Heidelberger & Welch***		***Raftery & Lewis***
	p-value	*pass*	*Stationarity Test*	*Halfwidth Test*	*MCMC samples*
m	0.037	✗*	passed	passed	3807
a[1]	0.078	✓	passed	passed	3696
a[2]	0.071	✓	passed	passed	3780
a[3]	0.256	✓	passed	passed	3724
a[4]	0.059	✓	passed	passed	3696
a[5]	0.129	✓	passed	passed	3642
b[1]	0.153	✓	passed	passed	3808
b[2]	0.997	✓	passed	passed	3848
b[3]	0.955	✓	passed	passed	3763
b[4]	0.172	✓	passed	passed	3779
b[5]	0.499	✓	passed	passed	3669
nu[2]	0.128	✓	passed	passed	3696
nu[3]	0.079	✓	passed	passed	3724
nu[4]	0.076	✓	passed	passed	3736
nu[5]	0.039	✗*	passed	passed	3736

* Failed for $\alpha = 0.05$ but passed for $\alpha = 0.10$

Table A.5: Summary results of convergence diagnostics: ***Row-Column Model***

Parameters	***Geweke***		***Heidelberger & Welch***		***Raftery & Lewis***
	p-value	*pass*	*Stationarity Test*	*Halfwidth Test*	*MCMC samples*
m	0.206	✓	passed	passed	3798
a[1]	0.081	✓	passed	passed	3696
a[2]	0.499	✓	passed	passed	3923
a[3]	0.212	✓	passed	passed	3738
a[4]	0.033	✗*	passed	passed	3600
a[5]	0.063	✓	passed	passed	3792
b[1]	0.785	✓	passed	passed	3848
b[2]	0.025	✗*	passed	passed	3716
b[3]	0.521	✓	passed	passed	3751
b[4]	0.025	✗*	passed	passed	3808
b[5]	0.251	✓	passed	passed	3724
mu[2]	0.083	✓	passed	passed	3780
mu[3]	0.073	✓	passed	passed	3865
mu[4]	0.025	✗*	passed	passed	3724
mu[5]	0.032	✗*	passed	passed	3865
nu[2]	0.172	✓	passed	passed	3798
nu[3]	0.138	✓	passed	passed	3780
nu[4]	0.534	✓	passed	passed	3780

* Failed for $\alpha = 0.05$ but passed for $\alpha = 0.10$

Table A.6: Summary results of convergence diagnostics: ***Saturated Model***

Parameters	***Geweke***		***Heidelberger & Welch***		***Raftery & Lewis***
	p-value	*pass*	*Stationarity Test*	*Halfwidth Test*	*MCMC samples*
m	0.962	✓	passed	passed	3820
a[1]	0.941	✓	passed	passed	3780
a[2]	0.035	✗*	passed	passed	3837
a[3]	0.191	✓	passed	passed	3669
a[4]	0.298	✓	passed	passed	3696
a[5]	0.234	✓	passed	passed	3848
b[1]	0.853	✓	passed	passed	3786
b[2]	0.819	✓	passed	passed	3720
b[3]	0.214	✓	passed	passed	3653
b[4]	0.601	✓	passed	passed	3642
b[5]	0.596	✓	passed	passed	3724
ab[2,2]	0.129	✓	passed	passed	3837
ab[2,3]	0.129	✓	passed	passed	3751
ab[2,4]	0.132	✓	passed	passed	3780
ab[2,5]	0.091	✓	passed	passed	3865
ab[3,2]	0.195	✓	passed	passed	3732
ab[3,3]	0.364	✓	passed	passed	3696
ab[3,4]	0.891	✓	passed	passed	3837
ab[3,5]	0.491	✓	passed	passed	3708
ab[4,2]	0.324	✓	passed	passed	3808
ab[4,3]	0.630	✓	passed	passed	3810
ab[4,4]	0.537	✓	passed	passed	3894
ab[4,5]	0.327	✓	passed	passed	3867
ab[5,2]	0.277	✓	passed	passed	3815
ab[5,3]	0.173	✓	passed	passed	3751
ab[5,4]	0.901	✓*	passed	failed	3724
ab[5,5]	0.715	✓	passed	failed	3724

* Failed for $\alpha = 0.05$ but passed for $\alpha = 0.10$

A.2 Comparison one-block importance sampling and bridge sampling estimators.

Table A.7: Estimated log-marginal likelihood for all the competing models with the two competitive estimators: one-block and bridge sampling.

		Log-marginal			
Mj	Model	One-block		Bridge sampling	
1	Independence (I)	-151.59	(0.002)	-151.59	(0.002)
2	Uniform (U)	-147.95	(0.004)	-147.97	(0.003)
3	Row (R)	-173.51	(0.012)	-173.68	(0.009)
4	Column (C)	-158.62	(0.007)	-159.31	(0.008)
5	Row-Column (RC)	-182.66	(0.048)	-182.96	(0.047)
6	Saturated (S)	-234.85	(5.944)	-234.80	(5.047)

A.3 Results about the proposed techniques under the presence of zeros 3.2.2

A.3.1 The initial results

Table A.8: Estimated log-marginal in the presence of zeros in Section 3.2.2

	t	Independence (I)	Uniform (U)	Row (R)	Column (C)	Row-Column (RC)	Saturated (S)
1	5000	-1105.74	-1107.31	-421.39	-1079.27	-431.95	-947.55
2	10000	-1105.74	-1107.30	-421.39	-1079.28	-430.61	-738.70
3	50000	-1105.74	-1107.31	-421.39	-1079.27	-431.23	-481.06
4	100000	-1105.74	-1107.31	-421.39	-1079.27	-431.39	-481.76
5	150000	-1105.74	-1107.31	-421.39	-1079.27	-431.33	-482.16
6	300000	-1105.74	-1107.30	-421.39	-1079.27	-431.41	-482.85
7	500000	-1105.74	-1107.30	-421.39	-1079.27	-431.28	-483.36

Table A.9: Estimated posterior probabilities in the presence of zeros in Section 3.2.2

	t	Independence (I)	Uniform (U)	Row (R)	Column (C)	Row-Column (RC)	Saturated (S)
1	5000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
2	10000	0.0000	0.9999	0.0000	0.0001	0.0000	0.0000
3	50000	0.0000	0.9999	0.0000	0.0001	0.0000	0.0000
4	100000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
5	150000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
6	300000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
7	500000	0.0000	0.0000	0.9999	0.0000	0.0001	0.0000

Table A.10: Estimated Monte Carlo error in the presence of zeros in Section 3.2.2

	t	Independence (I)	Uniform (U)	Row (R)	Column (C)	Row-Column (RC)	Saturated (S)
1	5000	0.0016	0.0016	0.0046	0.0073	1.1544	97.8587
2	10000	0.0011	0.0014	0.0030	0.0051	0.8409	88.5555
3	50000	0.0006	0.0005	0.0019	0.0024	0.2944	81.3714
4	100000	0.0005	0.0004	0.0010	0.0015	0.2275	66.7635
5	150000	0.0003	0.0003	0.0009	0.0013	0.2074	61.8563
6	300000	0.0002	0.0002	0.0007	0.0008	0.1266	44.0097
7	500000	0.0002	0.0002	0.0004	0.0007	0.1271	37.3232

A.4 Simulation study specifics: Details about the simulated scenarios (Section 3.2)

Table A.11: Coefficients of the true generating models for simulation scenarios 1–6 used in Section 3.2

Coefficients	**Simulation Scenario and True Generating Model**					
	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6
	Independence	Uniform	Row	Column	RC	Saturated
λ_0	4.322	6.161	3.315	3.376	4.322	4.322
λ_1^X	1.436	0.925	0.665	1.831	1.436	1.436
λ_2^X	0.772	0.832	1.148	0.831	0.772	0.772
λ_3^X	-0.715	-0.144	-0.152	-1.052	-0.715	-0.715
λ_4^X	0.415	1.444	1.153	-0.375	0.415	0.415
λ_5^X	-1.907^{*}	-3.053^{*}	-2.814^{*}	-1.234^{*}	-1.908^{*}	-1.908
λ_1^Y	0.571	0.007	0.991	0.0667	0.571	0.571
λ_2^Y	0.276	0.328	0.341	0.512	0.276	0.276
λ_3^Y	0.079	0.694	-0.277	0.782	0.079	0.079
λ_4^Y	-1.131	0.007	-1.966	-0.332	-1.132	-1.132
λ_5^Y	0.206^{*}	-1.036^{*}	0.913^{*}	-1.028^{*}	0.206	0.206
ϕ		-0.215	$1.00^{\dagger}$	$1.00^{\dagger}$	$1.00^{\dagger}$	
μ_1			$0.00^{\dagger}$		$0.00^{\dagger}$	
μ_2			0.655		0.654	
μ_3			0.607		0.607	
μ_4			0.164		0.164	
μ_5			0.081		0.081	
ν_1				$0.00^{\dagger}$	$0.00^{\dagger}$	
ν_2				0.709	0.709	
ν_3				0.479	0.479	
ν_4				0.218	0.218	
ν_5				0.039	$1.00^{\dagger}$	

* *Sum to zero constrains for the main effects.*

† *Row and column score constrains: $\phi = 1$, $\mu_1 = \nu_1 = 0$ and $\nu_J = 1$ for the RC moodel. $\phi = 1$, $\mu_1 = 0$ and $\phi = 1$, $\nu_1 = 0$ for the Row and Column model, respectively.*

Table A.12: Interaction terms coefficients of true saturated model used in the simulated scenario 6 of Section 3.2

	$\boldsymbol{\lambda_{i1}^{XY}}$	$\boldsymbol{\lambda_{i2}^{XY}}$	$\boldsymbol{\lambda_{i3}^{XY}}$	$\boldsymbol{\lambda_{i4}^{XY}}$	$\boldsymbol{\lambda_{i5}^{XY}}$
$\boldsymbol{\lambda_{1j}^{XY}}$	2.060 923 2	−1.068 515 07	−0.721 939 81	−0.329 162 84	−0.058 861 680
$\boldsymbol{\lambda_{2j}^{XY}}$	−0.945 544 7	0.463 777 03	0.313 349 92	0.142 869 45	0.025 548 255
$\boldsymbol{\lambda_{3j}^{XY}}$	−0.878 025 5	0.430 659 79	0.290 974 32	0.132 667 48	0.023 723 913
$\boldsymbol{\lambda_{4j}^{XY}}$	−0.237 353 0	0.116 418 49	0.078 657 90	0.035 863 45	0.006 413 188
$\boldsymbol{\lambda_{5j}^{XY}}$	−0.117 556 2	0.057 659 75	0.038 957 68	0.017 762 45	0.003 176 324

Table A.13: Expected cell values for each of the six simulated scenarios used in Section 3.2

Expected	**Model**					
cell	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6
parameter	Independence	Uniform	Row	Column	RC	Saturated
θ_{11}	13.747 890	6.396 697	7.910 405	6.184 587	13.747 890	107.965 157
θ_{12}	19.794 174	14.654 328	16.428 680	14.692 932	19.794 174	6.799 660
θ_{13}	14.741 084	16.288 631	16.487 926	17.675 335	14.741 084	7.161 351
θ_{14}	12.109 992	18.964 883	17.096 489	19.339 539	12.109 992	8.713 452
θ_{15}	3.606 861	7.695 460	6.076 500	6.107 607	3.606 861	3.400 683
θ_{21}	389.237 136	275.714 818	244.971 505	269.551 751	389.237 136	151.206 145
θ_{22}	560.422 543	509.583 018	485.562 014	508.759 836	891.110 177	891.110 177
θ_{23}	417.356 929	456.960 167	465.086 008	471.545 594	570.944 334	570.944 334
θ_{24}	342.864 144	429.228 585	460.255 924	431.188 293	395.520 943	395.520 943
θ_{25}	102.119 249	140.513 412	156.124 549	130.954 526	196.353 098	104.761 830
θ_{31}	200.418 459	202.687 846	254.824 173	201.616 966	200.418 459	83.294 454
θ_{32}	288.561 938	302.222 827	324.330 114	302.322 621	443.886 786	443.886 786
θ_{33}	214.897 359	218.642 899	199.477 246	215.890 338	287.474 695	287.474 695
θ_{34}	176.540 975	165.687 745	126.758 472	164.983 793	201.586 885	201.586 885
θ_{35}	52.581 269	43.758 683	27.609 995	48.186 281	96.490 912	53.843 617
θ_{41}	45.325 075	61.592 690	63.926 609	62.372 467	45.325 075	35.748 467
θ_{42}	65.258 916	74.092 363	74.895 457	74.303 522	73.316 170	73.316 170
θ_{43}	48.599 510	43.244 042	42.402 232	40.881 201	52.576 610	52.576 610
θ_{44}	39.925 129	26.437 842	24.802 731	26.109 379	41.382 968	41.382 968
θ_{45}	11.891 369	5.633 063	4.972 972	7.333 431	14.012 149	11.967 876
θ_{51}	140.271 440	242.607 949	217.367 308	249.274 229	140.271 440	124.714 011
θ_{52}	201.962 429	235.447 464	234.783 736	235.921 089	213.949 805	213.949 805
θ_{53}	150.405 118	110.864 261	122.546 589	100.007 531	156.380 185	156.380 185
θ_{54}	123.559 760	54.680 945	66.086 384	53.378 996	125.774 092	125.774 092
θ_{55}	36.801 252	9.399 382	12.215 984	14.418 154	39.917 435	36.918 331

Table A.14: Monte Carlo error of the estimation of the log-marginal as a function of the number of Importance sample size for all the competitive models with the Independence and One-Block Perrakis estimators and the two prior scenarios.

			Monte Carlo error of the log-marginal													
			T=1.000		T=5.000		T=10.000		T=15.000		T=50.000		T=100.000		T=150.000	
Prior	M_j	Model	IP	OBP	IP	OBP	IP	OBP	IP	OBP	IP	OBP	IP	OBP	IP	OBP
Prior 1	1	Independence	0.0562	0.0096	0.0316	0.0034	0.0194	0.0030	0.0165	0.0028	0.0102	0.0017	0.0071	0.0010	0.0063	0.0009
	2	Uniform	0.8633	0.0127	0.2223	0.0067	0.1646	0.0050	0.1199	0.0036	0.0744	0.0020	0.0544	0.0011	0.0476	0.0011
	3	Row	3.5787	0.0336	2.6479	0.0181	2.2862	0.0195	1.7931	0.0190	1.0985	0.0070	0.9857	0.0070	0.7768	0.0100
	4	Column	2.5436	0.0230	1.2775	0.0087	0.9960	0.0089	0.7663	0.0081	0.3742	0.0035	0.3434	0.0029	0.2873	0.0022
	5	RC	8.1690	0.1077	3.2277	0.0457	2.5878	0.0564	2.2685	0.0530	2.0819	0.0423	1.5897	0.0291	1.3182	0.1211
	6	Saturated	21.1382	17.1724	20.3462	10.1307	14.5277	6.9642	11.3116	6.0112	8.3744	3.0267	10.5953	3.4265	7.9363	2.4154
Prior 2	1	Independence	0.0781	0.0114	0.0279	0.0041	0.0184	0.0034	0.0155	0.0022	0.0098	0.0014	0.0077	0.0010	0.0062	0.0009
	2	Uniform	0.6227	0.0155	0.1579	0.0053	0.1734	0.0045	0.1277	0.0038	0.0802	0.0025	0.0504	0.0017	0.0374	0.0012
	3	Row	4.3775	0.0423	1.9420	0.0210	1.3146	0.0232	1.3250	0.0156	1.1424	0.0113	0.7105	0.0056	0.7491	0.0074
	4	Column	2.3907	0.0287	1.2036	0.0129	0.7987	0.0110	0.5615	0.0106	0.5106	0.0039	0.4019	0.0033	0.3102	0.0026
	5	RC	4.4522	0.1168	3.2691	0.0463	3.9943	0.0295	3.3771	0.0280	2.6385	0.0380	1.8261	0.0432	1.5091	0.0320
	6	Saturated	14.8354	9.5993	10.0692	4.1040	7.3528	4.9028	5.6252	4.3661	4.2050	3.7218	9.9941	2.2958	9.0480	2.0478

Table A.15: Monte Carlo error of the estimation of the log-Bayes Factor as a function of the number of Importance sample size for all the competitive models with the Independence and One-Block Perrakis estimators for the two scenarios of imaginary data.

			Monte Carlo error of the $log - BF_{UM_j}$													
			T=1.000		T=5.000		T=10.000		T=15.000		T=50.000		T=100.000		T=150.000	
Prior	M_j	Model	IP	OBP	IP	OBP	IP	OBP	IP	OBP	IP	OBP	IP	OBP	IP	OBP
Prior 1	1	Independence	0.5875	0.0138	0.2188	0.0074	0.1929	0.0055	0.1791	0.0033	0.1167	0.0022	0.0736	0.0016	0.0705	0.0015
	2	Row	3.7339	0.0328	1.8760	0.0178	1.5641	0.0202	1.1822	0.0190	0.6672	0.0071	0.7591	0.0071	0.5002	0.0104
	3	Column	2.5826	0.0236	1.3314	0.0097	0.7815	0.0112	0.7731	0.0094	0.5232	0.0043	0.4166	0.0034	0.3300	0.0028
	4	RC	7.6176	0.1077	2.5384	0.0482	2.1801	0.0579	1.7495	0.0531	1.5919	0.0421	1.4295	0.0288	1.2036	0.1213
	5	Saturated	25.0342	17.1713	19.6509	10.1310	14.0071	6.9647	11.1509	6.0112	7.4078	3.0269	10.2971	3.4266	8.7271	2.415
Prior 2	1	Independence	0.6057	0.0120	0.1522	0.0094	0.1739	0.0064	0.1297	0.0053	0.0788	0.0029	0.0482	0.0022	0.0373	0.0024
	2	Row	4.1636	0.0617	1.9307	0.0252	1.3162	0.0129	1.2826	0.0151	1.1517	0.0099	0.7119	0.0111	0.7441	0.0426
	3	Column	2.0950	0.0304	1.1916	0.0153	0.7634	0.0282	0.5545	0.0211	0.5231	0.0084	0.4055	0.0055	0.3122	0.0040
	4	RC	4.2171	0.0983	3.2248	0.1559	3.9369	0.1186	3.3461	0.0965	2.6201	0.0554	1.8198	0.0367	1.5019	0.0297
	6	Saturated	14.5093	10.0915	10.0336	4.1113	7.2991	3.9209	5.6166	3.3607	4.1939	3.7120	9.9873	2.9121	9.0441	2.2440

A.5 Summary plots for the simulation study of Section 3.2

A.5.1 Results using Prior 1

Figure A.1: Boxplots of the posterior model probabilities over 100 simulated datasets for Scenarios 1–6 for Prior 1

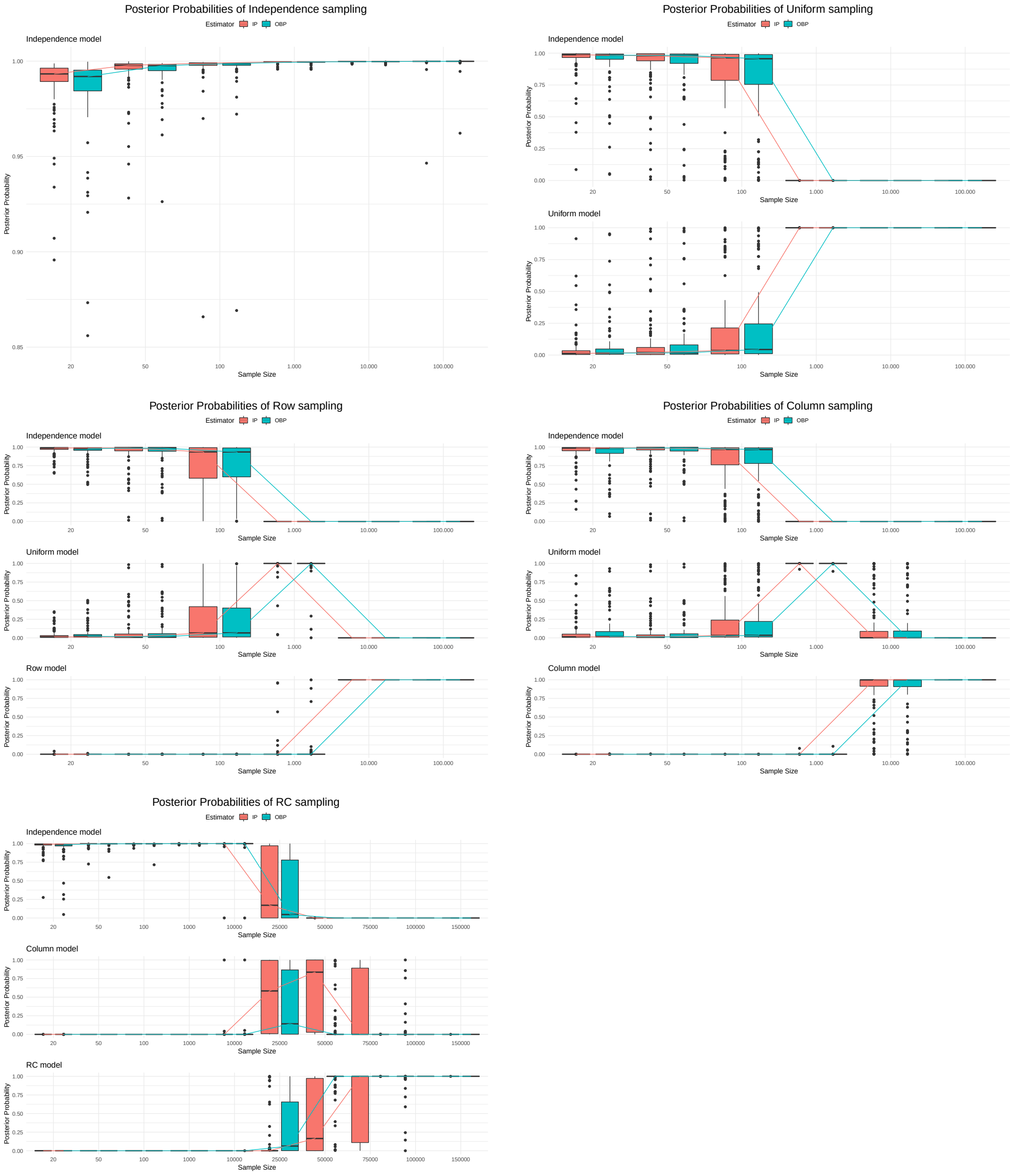


Figure A.2: Error bars of the 90% confidence intervals for the posterior model probabilities over 100 simulated datasets for Scenarios 1–6 for Prior 1

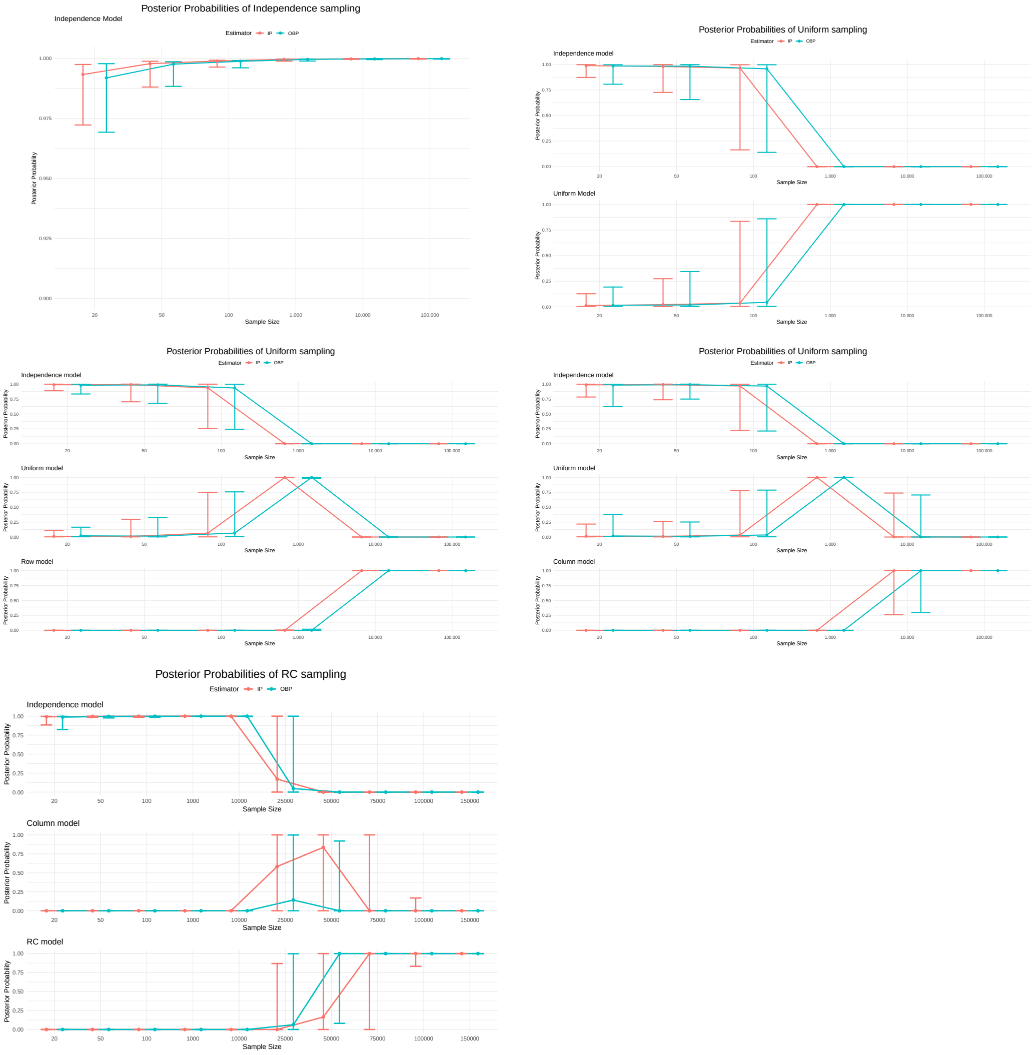
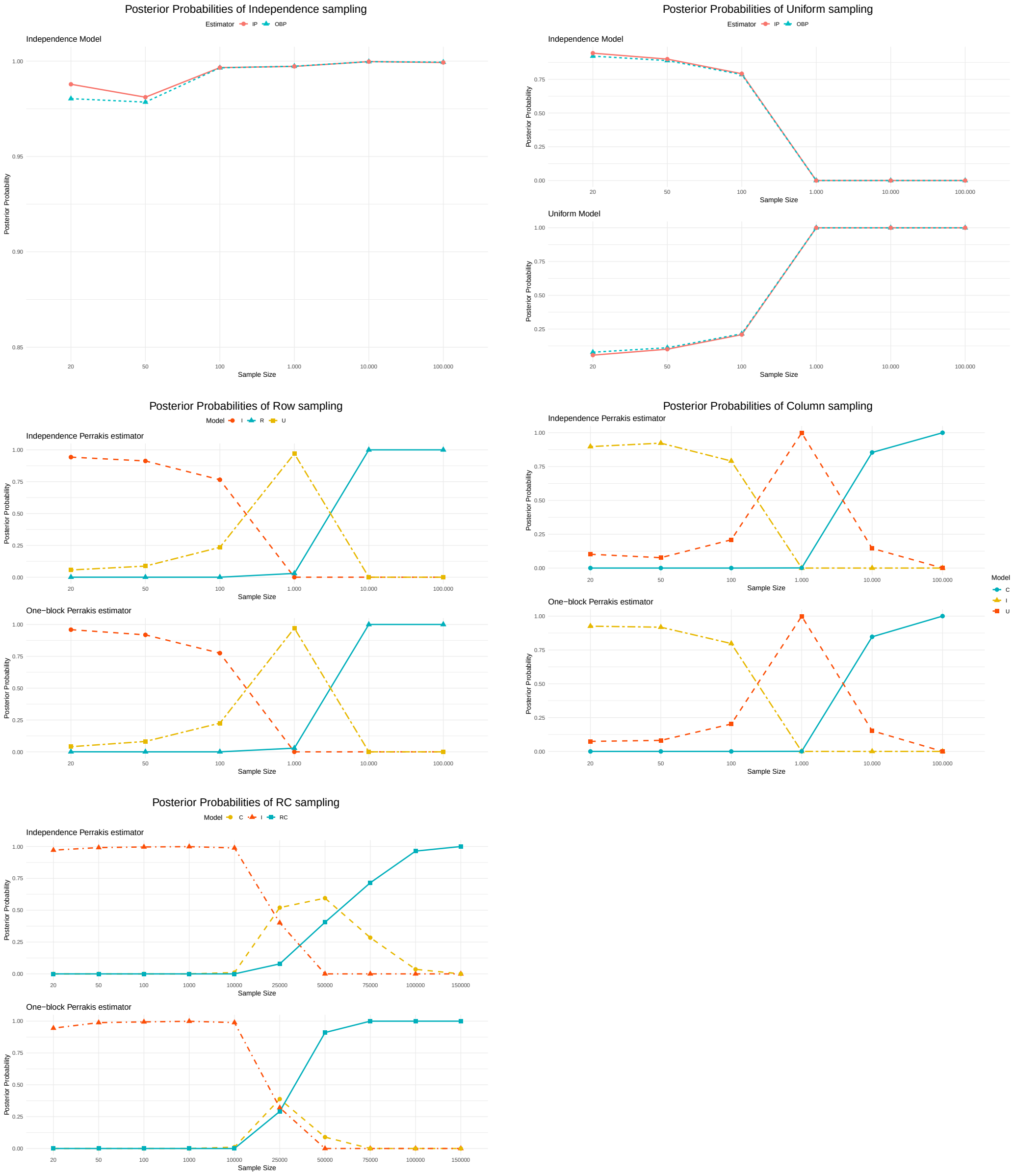


Figure A.3: Line plots for the posterior model probabilities over 100 simulated datasets for Scenarios 1–6 for Prior 1



A.5.2 Results using Prior 2

Figure A.4: Boxplots of the posterior model probabilities over 100 simulated datasets for Scenarios 1–6 for Prior 2

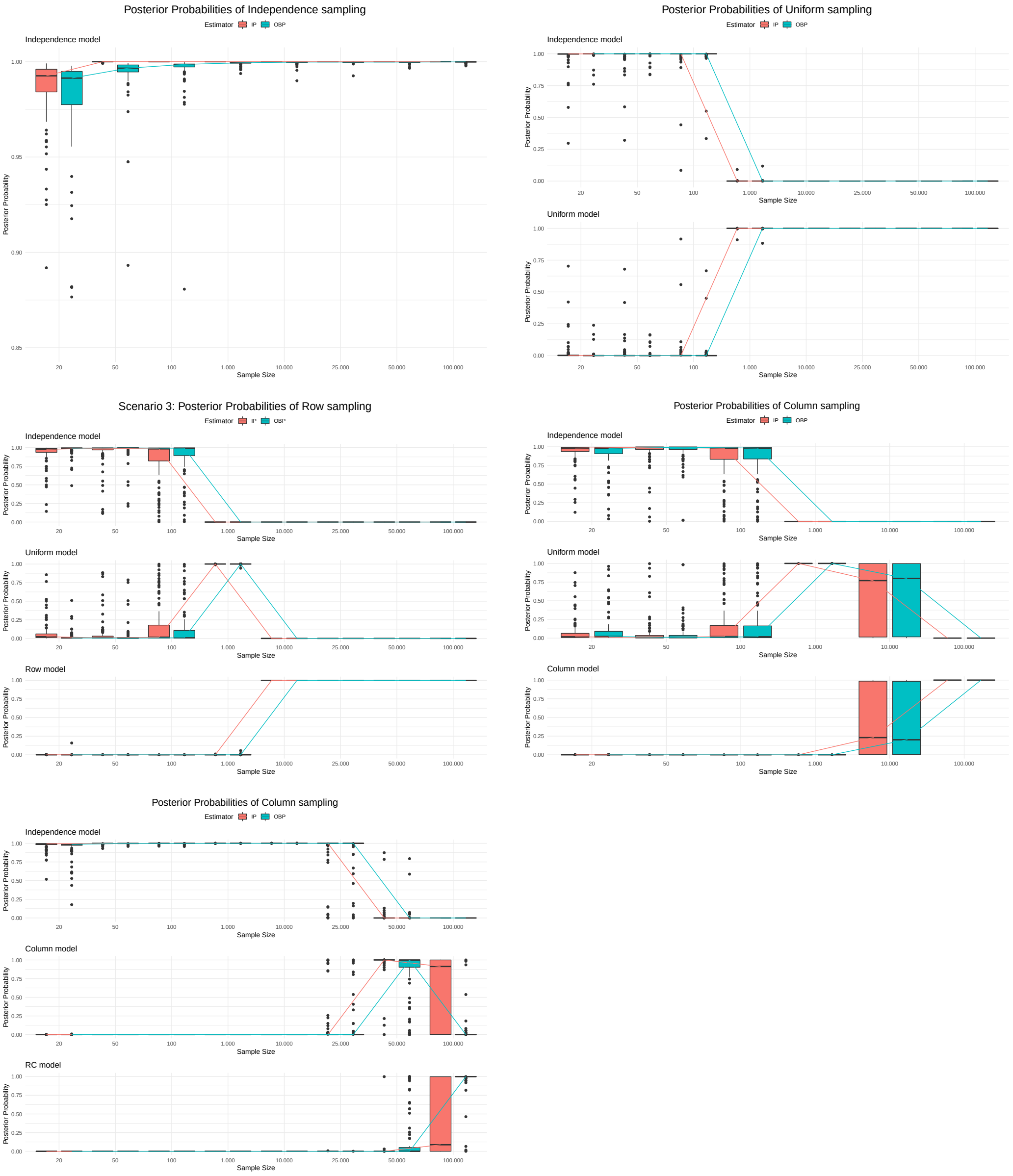


Figure A.5: Error bars of the 90% confidence intervals for the posterior model probabilities over 100 simulated datasets for Scenarios 1–6 for Prior 2

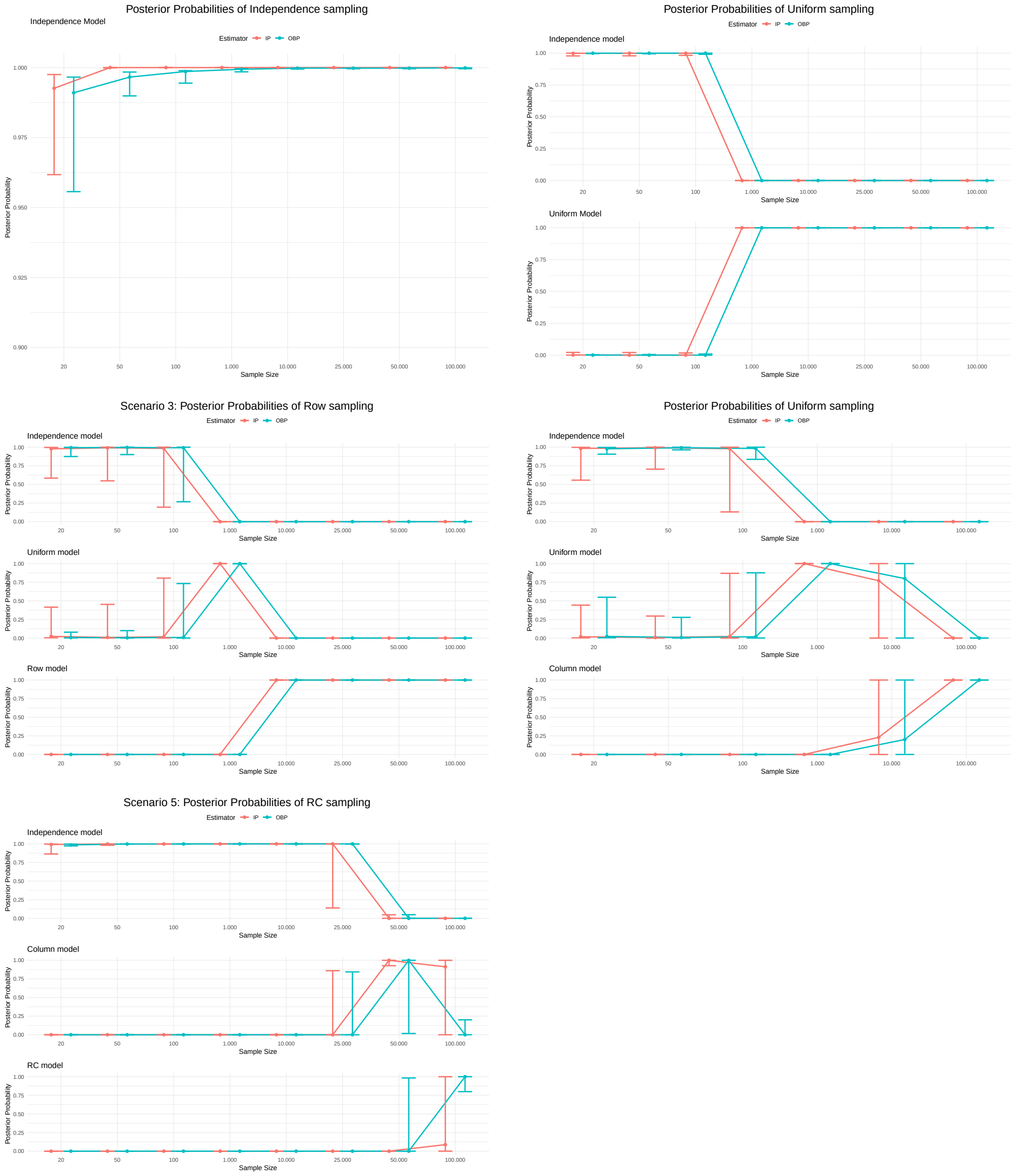
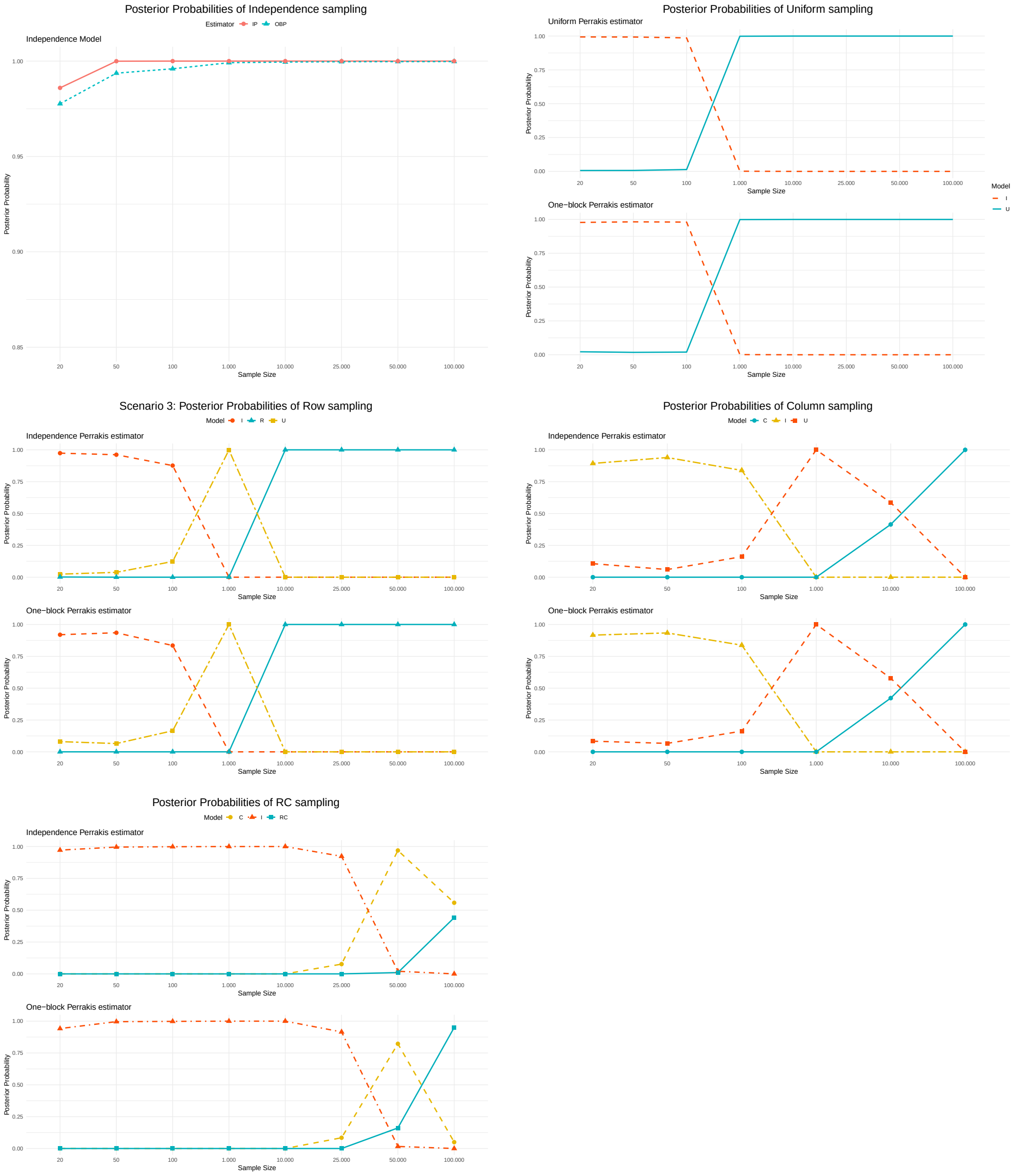


Figure A.6: Line plots for the posterior model probabilities over 100 simulated datasets for Scenarios 1–6 for Prior 2



Appendix B

R code

B.1 Packages and Functions in R

B.1.1 Required R-Packages in this thesis

The following R packages in alphabetical order of were used:

- `ggplot2`: Create elegant data visualisations using the grammar of graphics (H. Wickham, W. Chang, L. Henry *et al.*)
- `ggpubr`: Easy-to-use functions for creating and customizing `ggplot2` (A. Kassambara)
- `gnm`: Generalized nonlinear models (H. Turner and D. Firth).
- `MASS`: Support functions and datasets for Venables and Ripley's MASS (B. Ripley, B. Venables, D.M. Bates *et al.*)
- `mvtnorm`:Multivariate Normal and t Distributions (A. Genz, F. Bretz, T. Miwa *et al.*)
- `R2WinBUGS`: Running WinBUGS and OpenBUGS from R/S-PLUS (A. Gelman, S. Sturtz and U. Ligges).

B.1.2 Functions in R

```
# freq: vector of the frequencies, given by rows
# I: number of rows
# J: number of columns
```

```
row <- gl(I,J,length=I*J)
col <- gl(J,1,length=I*J)
data<-data.frame(freq, row, col)
# Sum-to-zero contrasts
contrasts(data$row)<-contr.sum(I)
contrasts(data$row)<-contr.sum(I)[c(I,1:(I-1)),]
contrasts(data$col)<-contr.sum(J)
contrasts(data$col)<-contr.sum(J)[c(J,1:(J-1)),]
I.model <- glm(freq ~ row+col, family=poisson)
U.model <- glm(freq ~ row+col+mu:nu, family=poisson)
# Variance-Covariance matrix
X<-model.matrix(model)

# Estimated the Poisson parameter
lamda<-NULL
for (i in 1:length(t)){
        lamda[[i]]<-t(apply(prop_theta[[i]], 1 , function(
           w) w%*%t(X)))
        lamda[[i]]<-exp(lamda[[i]])
}

# Generate the importance sample
# t: importance sample size
# l: posterior mean
# S: posterior variance
temp<-mvrnorm(t, l, S)
prop_theta<- NULL
for (i in 1:length(t)){
        prop_theta[[i]]<-temp[1:t[i],]
}


# Linear predictor for the Row-Column Association model
# beta: the parameter vector from the MCMC output
# Nrow: number of rows
# Ncol: number of columns
linear.predictor.rc <- function(beta, Nrow=I, Ncol=J) {
        J<-Ncol
        I<-Nrow
        m  <- beta[1]
        temp <- beta[2:I]
        a   <- c( -sum(temp), temp )
        temp <- beta[(I+1):(I+J-1)]
        b   <- c( -sum(temp), temp )
        mu <-beta[(I+J):(2*I+J-2)]
        mu <- c(0, mu)
        nu <-beta[(2*I+J-1):(2*I+2*J-4)]
        nu <- c(0, nu, 1)
        phi <- 1
        A <- matrix( a, I, J  )
        B <- matrix( b, I, J, byrow=TRUE  )
        MU <- matrix( mu, I, J  )
        NU <- matrix( nu, I, J, byrow=TRUE  )
        result <- as.vector(  t(m + A + B + phi*MU*NU) )
        return(result)
}


# Compute the marginal likeligood using
# the independence Perrakis estimator
# PARAMETERS:
# lamda: poisson parameter
# y: the data
# prop_theta: the genetated importance sampling
# l: posterior mean
# sd: posterior standar deviation
# prior.mean: prior mean
# prior.sd: prior standad deviation
# RETURNS:
# log_lik: the logarithm of the likelihood
# log_prior: the logarithm of the prior
# log_g: the logarithm of the imortance function
Independence_Perrakis<- function(lamda, y, prop_theta, l,
   sd, prior.mean, prior.sd){
        log_lik<-apply(lamda, 1, function(x) sum(dpois(y,x
            , log=TRUE)))
        log_prior<-apply(prop_theta,1, function(k) sum(
            dnorm(k, prior.mean, prior.sd, log=TRUE)))
        log_g<-apply(prop_theta, 1, function(z) sum(dnorm(
            z, l, sd, log=TRUE)))
        w<-log_lik+log_prior-log_g
        maxw<-max(w)
        w <- w-maxw
        perrakis<-log(mean(exp(w)))+maxw
        return( perrakis )
}

# Compute the marginal likeligood using
# the one-block Perrakis estimator
# PARAMETERS:
# lamda
# y: the data
# prop_theta: the genetated importance sampling
# l: posterior mean
# S: posterior variance
# prior.mean: prior mean
# prior.sd: prior standad deviation
# RETURNS:
# log_lik: the logarithm of the likelihood
# log_prior: the logarithm of the prior
# log_g: the logarithm of the imortance function
One_Block_Perrakis<- function(lamda, y, prop_theta, l, S,
   prior.mean, prior.sd){
        log_lik<-apply(lamda, 1, function(x) sum(dpois(y,x
           , log=TRUE)))
        log_prior<-apply(prop_theta,1, function(k) sum(
           dnorm(k, prior.mean, prior.sd, log=TRUE)))
        log_g<-apply(prop_theta, 1, function(z)
        sum(dmvnorm(z, 1, S, log=TRUE)))
        w<-log_lik+log_prior-log_g
        maxw<-max(w)
        w <- w-maxw
        perrakis<-log(mean(exp(w)))+maxw
        return( perrakis )
}
```

Bibliography

Agresti, A. (2002). *Categorical Data Analysis.* 2nd ed. Wiley-Interscience, Hoboken, NJ.

Agresti, A. (2013). *Caterorical Data Analysis.* 3rd edition. Wiley & Sons.

Agresti, A. and Chuang, C. (1989). Model-based bayesian methods for estimating cell proportions in cross-classification tables having ordered categories. *Computational Statistics & Data Analysis*, 7(3):245 – 258.

Agresti, A. and Hitchcock, D. (2005a). Bayesian inference for categorical data analysis. *Statistical Methods and Applications*, 14:297–330.

Agresti, A. and Hitchcock, D. B. (2005b). Bayesian inference for categorical data analysis. *Statistical Methods & Applications*, 14(3):297–330.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory, Budapest: Akademia Kiado*, pages 267–281.

Albert, J. (1997). Teaching bayes rule: A data-oriented approach. *The American Statistician*, 51(3):247–253.

Albert, J. H. (1996). Bayesian selection of log-linear models. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 24(3):327–347.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

Althman, P. M. E. (1969). Exact Bayesian analysis of a 2 × 2 contingency table, and Fisher's "exact" significance test. *Journal of the Royal Statistical Society, Series B, Methodological*, 31:261–569.

Althman, P. M. E. (1971). The analysis of matched proportions. *Biometrika*, 58(3):561–576.

Ardia, D., Baştürk, N., Hoogerheide, L., and van Dijk, H. (2012). A comparative study of monte carlo methods for efficient evaluation of marginal likelihood. *Computational Statistics and Data Analysis*, 56:3398–3414.

Baele, G. and Lemey, P. (2013). Bayesian evolutionary model testing in the phylogenomics era: matching model complexity with computational efficiency. *Bioinformatics*, 29:1970–1979.

Baele, G., Lemey, P., and Vansteelandt, S. (2013). Make the most of your samples: Bayes factor estimators for high-dimensional models of sequence evolution. *BMC bioinformatics*, 14:85.

Bartlett, M. (1957). Comment on D.V. Lindleys statistical paradox. *Biometrika*, 44:533–534.

Bartlett, M. S. (1935). Contingency table interactions. *Supplement to the Journal of the Royal Statistical Society*, 2(2):248–252.

Bartlett, M. S. (1937). Some examples of statistical methods of research in agriculture and applied biology. *Supplement to the Journal of the Royal Statistical Society*, 4(2):137–183.

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). Criteria for bayesian model choice with application to variable selection. *Ann. Statist.*, 40(3):1550–1577.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions*, 53:370–418.

Berkson, J. M. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39(227):357–365.

Bernardo, J. M. (1979). Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):113–147.

Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 25(1):220–233.

Birch, M. W. (1964a). A New Proof of the Pearson-Fisher Theorem. *The Annals of Mathematical Statistics*, 35(2):817 – 824.

Birch, M. W. (1964b). The detection of partial association, i: The 2 × 2 case. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):313–324.

Birch, M. W. (1965). The detection of partial association, ii: The general case. *Journal of the Royal Statistical Society. Series B (Methodological)*, 27(1):111–124.

Bishop, Y. M. M. (1967). *Multi-dimensional Contingency Tables: Cell Estimates.* PhD dissertation, Department of Statistics, Harvard University.

Bishop, Y. M. M. (1969). Full contingency tables, logits, and split contingency tables. *Biometrics*, 25(2):383–399.

Caussinus, H. (1965). Contribution à l'analyse statistique des tableaux de corrélation. *Annales de la Faculté des sciences de Toulouse : Mathématiques*, 4e série, 29:77–183.

Chen, C. and Dunson, D. (2003). Random effects selection in linear mixed models. *Biometrics*, 59:762–769.

Chen, M., Ibrahim, J., and Shao, Q. (2000a). Power prior distributions for generalized linear models. *Journal of Statistical Planning and Inference*, 84(1-2):121–137.

Chen, M., Ibrahim, J. G., and Shao, Q.-M. (2000b). Power prior distributions for generalized linear models. *Journal of Statistical Planning and Inference*, 84:121–137.

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

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the metropolis–hastings output. *Journal of the American Statistical Association*, 96:270–281.

Chuang, C. (1982). Empirical bayes methods for a two-way multiplicative-interaction model. *Communications in Statistics: Theory and Methods*, 11:2977–2989.

Cochran, W. G. (1940). The Analysis of Variance when Experimental Errors Follow the Poisson or Binomial Laws. *The Annals of Mathematical Statistics*, 11(3):335 – 347.

Cochran, W. G. (1943). Analysis of variance for percentages based on unequal numbers. *Journal of the American Statistical Association*, 38(223):287–301.

Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika*, 37(3/4):256–266.

Cochran, W. G. (1954). Some methods for strengthening the common x^2 tests. *Biometrics*, 10(4):417–451.

Congdon, P. (2005). *Bayesian Models for Categorical Data.* Wiley, 1 edition.

Consonni, G., Fouskakis, D., Liseo, B., and Ntzoufras, I. (2018). Prior distributions for objective bayesian analysis. *Bayesian Anal.*, 13(2):627–679.

Consonni, G. and Pistone, G. (2007). Algebraic bayesian analysis of contingency tables with possibly zero-probability cells. *Statistica Sinica*, 17(4):1355–1370.

Consonni, G. and Veronese, P. (2008). Compatibility of prior specifications across linear models. *Statist. Sci.*, 23(3):332–353.

Cornfield, J. (1951). A Method of Estimating Comparative Rates from Clinical Data. Applications to Cancer of the Lung, Breast, and Cervix. *JNCI: Journal of the National Cancer Institute*, 11(6):1269–1275.

Darroch, J. N., Lauritzen, S. L., and Speed, T. P. (1980). Markov Fields and Log-Linear Interaction Models for Contingency Tables. *The Annals of Statistics*, 8(3):522 – 539.

Dawid, A. P. and Lauritzen, S. (2011). Compatible prior distributions. *In George, E. I. (ed.), Bayesian Methods with Applications to Science, Policy and Official Statistics. Proceedings of the 6th World Meeting*, (2):109–118.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models. *The Annals of Statistics*, 21(3):1272 – 1317.

Dellaportas, P. and Forster, J. (1999). Markov chain monte carlo model determination for hierarchical and graphical log-linear models. *Biometrika*, 3:615–633.

Dellaportas, P., Forster, J., and Ntzoufras, I. (2002). On bayesian model and variable selection using mcmc. *Statistics and Computing*, 12:27–36.

Deming, W. E. and Stephan, F. F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *The Annals of Mathematical Statistics*, 11(4):427 – 444.

Demirhan, H. (2013). Bayesian estimation of order-restricted and unrestricted association models. *Journal of Multivariate Analysis*, 121:109–126.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *Ann. Statist.*, 7(2):269–281.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *Ann. Math. Statist.*, 42(1):204–223.

Drton, M. and Richardson, T. (2008). Binary models for marginal independence. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70(2):287–309.

Evans, M., Gilula, Z., Guttman, I., and Swar, T. (1993). "computational issues in the bayesian analysis of categorical data: log-linear and goodman's rc mode". *Statistica Sinica*, 3:391–406.

Fan, Y., Wu, R., Chen, M.-H., Kuo, L., and Lewis, P. O. (2011). Choosing among partition models in bayesian phylogenetics. *Molecular Biology and Evolution*, 28:523–532.

Fisher, R. A. (1922). On the interpretation of x^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94.

Fisher, R. A. and Yates., F. (1938). *Statistical Tables.* Edinburgh: Oliver and Boyd.

Forster, J. J. and Webb, E. L. (2007). Bayesian disclosure risk assessment: predicting small frequencies in contingency tables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 56(5):551–570.

Fouskakis, D. and Ntzoufras, I. (2016). Power-conditional-expected priors: Using g-priors with random imaginary data for variable selection. *Journal of Computational and Graphical Statistics*, 25:647–664.

Fouskakis, D., Ntzoufras, I., and Draper, D. (2015). Power-expected-posterior priors for variable selection in Gaussian linear models. *Bayesian Analysis*, 10:75–107.

French, K. and Westoby, M. (1996). Vertebrate-dispersed species in a fire-prone environment. *Australian Journal of Ecology*, 21:379–385.

Friel, N., Hurn, M., and Wyse, J. (2014). Improving power posterior estimation of statistical evidence. *Statistics and Computing*, 24:709–723.

Friel, N. and Pettitt, A. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society B*, 70:589–607.

Friel, N. and Wyse, J. (2012). Estimating the evidence – a review. *Statistica Neerlandica*, 66:288–308.

Galindo-Garre, F., Vermunt, J. K., and Bergsma, W. P. (2004). Bayesian posterior estimation of logit parameters with small samples. *Sociological Methods & Research*, 33(1):88–117.

Galton, F. (1892). *Finger Prints.* London: Macmillan.

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.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics*, volume 4, pages 169–193. Oxford University Press, New York.

Gilks, W. and Roberts, G. (1996). Strategies for improving mcmc. In Markov Chain Monte Carlo in Practice, 6, eds. W.R. Gilks, S. Richardson and D.J. Spiegelhalte, pages 89–114. Chapman & Hall, London, UK.

Giudici, P. (1998). Smooth sparse contingency tables: a graphical bayesian approach. *Metron*, 56.

Good, I. (1965). *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* M.I.T. Press research monographs. M.I.T. Press.

Good, I. J. (1950). *Probability and the Weighing of Evidence*, volume 647.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3/4):237–264.

Good, I. J. (1956). On the estimation of small frequencies in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 18(1):113–124.

Goodman, L. (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.

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, 74(367):537–552.

Goodman, L. A. and Kruskal, W. H. (1979). *Measures of Association for Cross Classifications III: Approximate Sampling Theory*, pages 76–130. Springer New York, New York, NY.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732.

Greenland, S. (2001). Putting background information about relative risks into conjugate prior distributions. *Biometrics*, 57(3):663–670.

Gunel, E. and Dickey, J. (1974). Bayes factors for independence in contingency tables. *Biometrika*, 61(3):545–557.

Haberman, S. J. (1974). Log-Linear Models for Frequency Tables Derived by Indirect Observation: Maximum Likelihood Equations. *The Annals of Statistics*, 2(5):911 – 924.

Healy, M. J. R. (1988). *Glim: An Introduction.* Clarendon Press, USA.

Heckerman, D., Geiger, D., and Chickering, D. (2013). Learning bayesian networks: The combination of knowledge and statistical data. *CoRR*, abs/1302.6815.

Heidelberger, P. and Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31(6):1109–1144.

Heyde, C. C. and Seneta, E. (1977). *I. J. Bienayme : statistical theory anticipated / C. C. Heyde, E. Seneta.* Springer-Verlag New York.

Ibrahim, J. G. and Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science*, 15:46–60.

Iliopoulos, G., Kateri, M., and Ntzoufras, I. (2007). Bayesian estimation of unrestricted and order-restricted association models for a two-way contingency table. *Computational Statistics and Data Analysis*, 51:4643–4655.

Iliopoulos, G., Kateri, M., and Ntzoufras, I. (2009). Bayesian model comparison for the order restricted rc association model. *Psychometrika*, 74:561–587.

Iliopoulou, K. (2004). Schizotypy and consumer behavior (in greek). Master's thesis, Department of Business Administration, University of the Aegean, Chios, Greece.

Jeffreys, H. (1961). *Theory of probability.* Oxford: Oxford University Press.

Johnson, N. L. and Kotz, S. (1985). Some distributions arising as a consequence of errors in inspection. *Naval Research Logistics Quarterly*, 32(1):35–43.

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

Kass, R. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91:1343–1370.

Kass, R. E. and Wasserman, L. (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association*, 90:928–934.

Kateri, M. (2014). *Contingency Table Analysis: Methods and Implementation Using R*. Birkhäuser.

Kateri, M. and Agresti, A. (2013). Bayesian inference about odds ratio structure in ordinal contingency tables. *Environmetrics*, 24:281–288.

Kateri, M., Nicolaou, A., and Ntzoufras, I. (2005a). Bayesian inference for the RC(m) association model. *Journal of Computational and Graphical Statistics*, 14:116–138.

Kateri, M., Nicolaou, A., and Ntzoufras, I. (2005b). Bayesian inference for the RC(m) association model. *Journal of Computational and Graphical Statistics*, 14:116–138.

King, R. and Brooks, S. (2001). Prior induction in log-linear models for general contingency table analysis. *The Annals of Statistics*, 29(3):715 – 747.

Knuiman, M. W. and Speed, T. P. (1988). Incorporating prior information into the analysis of contingency tables. *Biometrics*, 44(4):1061–1071.

Laird, N. M. (1978). Empirical Bayes methods for two-way contingency tables. *Biometrika*, 65(3):581–590.

Laplace, P. S. (1774). Mémoire sur la Probabilité des Causes par les évènemens. *" Mémoires de Mathematique et de Physique, Presentés à l'Académie Royale des Sciences, Par Divers Savans & Lus Dans ses Assemblées, Tome Sixièm*, 53:621–656.

Lartillot, N. and Philippe, H. (2006a). Computing Bayes Factors Using Thermodynamic Integration. *Systematic Biology*, 55(2):195–207.

Lartillot, N. and Philippe, H. (2006b). Computing Bayes factors using Thermodynamic Integration. *Systematic Biology*, 55:195–207.

Lee, M. and Wagenmakers, E. (2014). *Bayesian Cognitive Modeling: A Practical Course*. Bayesian Cognitive Modeling: A Practical Course. Cambridge University Press.

Leonard, T. (1972). Bayesian methods for binomial data. *Biometrika*, 59(3):581–589.

Leonard, T. (1975). Bayesian estimation methods for two-way contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 37(1):23–37.

Leonard, T., Hsu, J., and Tsui, K. (1989). Bayesian marginal inference. *Journal of the American Statistical Association*, 84:1051–1057.

Lewis, S. and Raftery, A. (1997a). Estimating bayes factor via posterior simulation with the laplace-metropolis estimator. *Journal of the American Statistical Association*, 92:648–655.

Lewis, S. M. and Raftery, A. E. (1997b). Estimating bayes factors via posterior simulation with the laplace-metropolis estimator. *Journal of the American Statistical Association*, 92(438):648–655.

Lindley, D. (1957). A statistical paradox. *Biometrika*, 44:187–192.

Lindley, D. V. (1964). The bayesian analysis of contingency tables. *The Annals of Mathematical Statistics*, 35(4):1622–1643.

Lykou, A. and Ntzoufras, I. (2013). On bayesian lasso variable selection and the specification of the shrinkage parameter. *Statistics and Computing*, 23:361–390.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546.

Mantel, N. and Haenszel, W. (1959). Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *JNCI: Journal of the National Cancer Institute*, 22(4):719–748.

Marselos, M., Boutsouris, K., Liapi, H., Malamas, M., Kateri, M., and Papaioannou, T. (1997). Epidemiological aspects on the use of cannabis among university students in greece. *European Addiction Research*, 3:184–191.

Maxwell, A. (1961). Analyzing qualitative data. *Methuen, London.*

Meng, X.-L. and Wong, W. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 6:831–860.

Morris, C. N. (1983). Natural exponential families with quadratic variance functions: Statistical theory. *The Annals of Statistics*, 11(2):515–529.

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9:249–265.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

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

Neyman, J. (1949). Contributions to the theory of the x^2 test. in proceedings of the berkeley symposium on mathematical statistical probability. *Berkeley: University of California Press*, pages 239–273.

Norton, P. and Dunn, E. (1985). Snoring as a risk factor for disease: an epidemiological survey. *British Medical Journal*, 291:630–632.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS.* Wiley Series in Computational Statistics. Hoboken, NJ.

Ntzoufras, I., Dellaportas, P., and Forster, J. (2000a). Stochastic search variable selection for log-linear models. *Journal of Statistical Computation and Simulation*, 68:23–38.

Ntzoufras, I., Forster, J. J., and Dellaportas, P. (2000b). Stochastic search variable selection for log-linear models. *Journal of Statistical Computation and Simulation*, 68(1):23–37.

Oh, M.-S. (1999). Estimation of posterior density functions from a posterior sample. *Computational Statistics & Data Analysis*, 29(4):411–427.

Oh, M.-S. (2014). Bayesian test on equality of score parameters in the order restricted rc association model. *Computational Statistics and Data Analysis*, 72:147–157.

Pearl, J. and Wermuth, N. (1994). When can association graphs admit a causal interpretation? pages 205–214.

Pearson, K. F. (1900). X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Pearson, K. F. (1904). Mathematical contributions to the theory of evolution xiii: On the theory of contingency and its relation to association and normal correlation. *Draper's Co. Research Memoirs, Biometric Series (Reprinted in Karl Pearson's Early Papers, ed. E. S. Pearson, Cambridge: Cambridge University Press, 1948)*, (1).

Perks, W. (1947). Some observations on inverse probability including a new indifference rule. *Journal of the Institute of Actuaries*, 73(2):285–334.

Perrakis, K., Ntzoufras, I., and Tsionas, E. G. (2014). On the use of marginal posteriors in marginal likelihood estimation via importance sampling. *Computational Statistics & Data Analysis*, 77:54 – 69.

Quetelet, A. (1849). *Letters addressed to H.R.H. the Grand Duke of Saxe Coburg and Gotha, on the theory of probabilities [microform] : as applied to the moral and political sciences / by A. Quetelet ; translated from the French by Olinthus Gregory Downes.* C. & E. Layton London.

Raftery, A. (1988). Approximate bayes factors for generalized linear models. *Technical Report, Department of Statistics, University of Washington*, 121.

Raftery, A. (1996a). Hypothesis testing and model selection. in W. Gilks, S. Richardson, and D. Spiegelhalter, eds., *Markov Chain Monte Carlo in Practice*, pages 163–188. Chapman & Hall, Suffolk, UK.

Raftery, A. and Lewis, S. (1992). How many iterations in the Gibbs sampler? In Bernardo, J., Berger, J., Dawid, A., and Smith, A., editors, *Bayesian Statistics, Vol. 4*, pages 763–774. Claredon Press, Oxford.

Raftery, A., Newton, M., Satagopan, J., and Krivitsky, P. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity (with discussion). In Bernardo, J., Bayarri, M., and J.O., B., editors, *Bayesian Statistics, Vol. 8*, pages 1–45. Oxford University Press.

Raftery, A. E. (1986). A note on bayes factors for lod-linear contingency table models with vague prior information. *Journal of the royal statistical society series b-methodological*, 48:249–250.

Raftery, A. E. (1996b). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, 83(2):251–266.

Rosenkranz, S. (1992). *The Bayes factor for model evaluation in a hierarchical Poisson model for area counts*. PhD thesis, Department of Biostatistics, University of Washington.

Rousseeuw, P. and van Zomeren, B. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639.

Roy, S. N. and Mitra, S. K. (1956). AN INTRODUCTION TO SOME NON-PARAMETRIC GENERALIZATIONS OF ANALYSIS OF VARIANCE AND MULTIVARIATE ANALYSIS*. *Biometrika*, 43(3-4):361–376.

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

Simon, G. (1974). Alternative analyses for the singly-ordered contingency table. *Journal of the American Statistical Association*, 69(348):971–976.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(2):238–241.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Spiegelhalter, D. J. and Smith, A. F. M. (1982a). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(3):377–387.

Spiegelhalter, D. J. and Smith, A. F. M. (1982b). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(3):377–387.

Stigler, S. (2002). The missing early history of contingency tables. *Annales de la Faculté des sciences de Toulouse : Mathématiques*, Ser. 6, 11(4):563–573.

Tarantola, C., Consonni, G., and Dellaportas, P. (2008). Bayesian clustering for row effects models. *Journal of Statistical Planning and Inference*, 138(7):2223–2235.

Tarantola, C. and Ntzoufras, I. (2012). Bayesian analysis of graphical models of marginal independence for three way contingency tables. Quaderni di Dipartimento 172, Pavia.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

Tukey, J. W. (1949). One degree of freedom for non-additivity. *Biometrics*, 5(3):232–242.

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.

Wilks, S. S. (1935). The Likelihood Test of Independence in Contingency Tables. *The Annals of Mathematical Statistics*, 6(4):190 – 196.

Xie, W., Lewis, P., Fan, Y., Kuo, L., and Chen, M. (2011). Improving marginal likelihood estimation for bayesian phylogenetic model selection. *Systematic Biology*, 60:150–160.

Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika*, 2(2):121–134.