Bayesian Tests for Conditional Independence in Contingency Tables: A simulation Study on Bayes factors for conditional independence tests in 2x2xK contingency tables

Heeman, D.F.

Bayesian Tests for Conditional Independence in Contingency Tables

A simulation study on Bayes factors for conditional independence tests in $2 \times 2 \times K$ contingency tables.

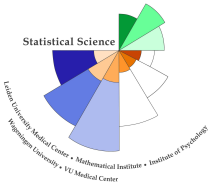
D.F. Heemann (s2086565)

Thesis supervisor: Prof. Dr. P.D. Grünwald
External thesis supervisor: Dr. A. Ly

MASTER THESIS

November 2021

Specialization: Data Science



STATISTICAL SCIENCE
FOR THE LIFE AND BEHAVIOURAL SCIENCES

Abstract

In the analysis of $2 \times 2 \times K$ contingency tables, a common hypothesis is the conditional independence of rows and columns controlling for a third variable. While frequentist versions to test this hypothesis (e.g., the Cochran-Mantel-Haenszel) exist, it was the goal of this thesis to evaluate a Bayes factor alternative for the conditional independence test in contingency tables. Framing the test as a Bayesian model comparison using generalized linear models, multiple g-prior variants were evaluated and compared to each other through a simulation study. The simulation results indicate that priors like the hyper-g/n, intrinsic or the robust prior generally show desirable patterns for medium to large effect sizes, but are prone to lead to wrong conclusions for small underlying effect sizes unless the sample size is large. The R code for the simulation study can be found on GitHub: https://github.com/DHeemann/Bayesian-conditional-independence-simulation-

Acknowledgements

I would like to express my gratitude to my supervisor Alexander Ly for his guidance and help throughout this project. I have learned a lot from his expertise and his insights. Our discussions were both helpful and inspiring and his ideas and feedback had a significant impact on my thesis. I would also like to thank my supervisor Professor Peter Grünwald for making it possible to work on this fascinating and important topic. A big thanks also to the thesis committee, my teachers and advisors/coordinators of the Statistical Science master's program for enabling this project. I especially want to say thank you to Alex, Rolf and Olga, and all my other friends and family who have helped, inspired and supported me.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Overview and problem statement

Observational data and study-based data (e.g., clinical or social studies) often come in the form of contingency tables, which is a data format showing the frequency for (combinations of) one or more variables of interest (Everitt, 2019). An example of a contingency table based on observational data is the publicly available Titanic data set, which among other variables shows the age category, sex and survival (yes or no) of the passengers on the ocean liner Titanic in 1912 (Kaggle.com). In case a contingency table, such as the example above, contains three variables or more, a common null hypothesis is that two variables are conditionally independent given a third variable.

In classical statistics, there is a wide range of existing approaches to test the null hypothesis of conditional independence in three-way contingency tables: For example, the Cochran-Mantel-Haenszel (CMH) test was designed to specifically test the independence of rows i and columns j controlling for layers (or also called strata) k in a $2 \times 2 \times K$ design (Cochran, 1954). In the case of the CMH test, we would expect that the conditional odds ratio of rows and columns equals 1 for each categorical value of the third variable. While tests of conditional independence in three-way contingency tables using p-values have been developed and reviewed widely, Bayesian versions have only been discussed for two-dimensional contingency tables (Jamil et al., 2017). Bayesian conditional independence tests in multidimensional contingency tables have been unavailable for researchers so far. However, it has been shown that the CMH test can be represented as a log-linear model (von Eye and Indurkhya, 2000), which allows for more flexible generalizations of the test. Equivalently, logistic regression models also allow for tests that assess conditional independence in contingency tables (Agresti, 2003).

As both the logistic regression model and the log-linear model are sub-classes of generalized linear models (GLMs), and since Bayes factors for variable selection in GLMs have been widely developed (Li and Clyde, 2018), it is the goal of this thesis to bridge the gap between the classical tests of conditional independence in $2 \times 2 \times K$ contingency tables and the usage of Bayes factors through generalized linear representations. More specifically, the Bayes factor implementation of the linear models used in this thesis is based on the utilization of (mixtures of) g-priors, which have been originally developed for Gaussian models (Zellner, 1986) and then further extended to generalized linear models. Through a simulation study conducted using R (R Core Team, 2019), we studied the suitability of multiple g-prior variants for GLM, as summarised by Li and Clyde (2018), applied to the problem of testing conditional independence in contingency tables. The focus of the simulation study for this thesis was on $2 \times 2 \times 2$ contingency tables. For the calculation of Bayes factors, the BAS package (Bayesian Variable Selection and Model Averaging using Bayesian Adaptive Sampling) developed by Merlise Clyde was used, as it allows for Bayesian nested model comparisons of GLMs using the relevant g-priors (Clyde, 2015).

This thesis is structured as follows: At first, a general overview of the (frequentist) analysis of conditional independence in three-dimensional contingency tables will be given. After that, we will explain the Bayes factor and will give a description of the log-linear and logistic model framework for the test of conditional independence. As part of the model framework, the set-to-zero and the sum-to-zero representation for the log-linear test will be discussed. Furthermore, we will give an overview of the different g-priors that will be explored. As part of the simulation study, the data generation process will be explained. Next, the Bayes factor behavior of different g-priors for the test of conditional independence will be assessed, followed by some practical guidelines for those aiming to implement the Bayes factor test. In addition, a computational modification[1] for the hyper-g/n prior to avoid potential computational issues due to large sample sizes will be proposed.

1.2 Three-dimensional contingency tables

Contingency tables, which consist of count data of multiple categorical variables are widely used across different disciplines and are an efficient way of summarizing how many times a

[1] R script on GitHub: github.com/DHeemann/Bayesian-conditional-independence-simulation-

phenomenon has occurred for each combination of the available categories (Everitt, 2019). For example, epidemiological studies may investigate if a treatment can increase or decrease the prevalence of certain diseases. In social sciences, contingency tables are a popular data format to compare a stimulus to a control condition.

Often, contingency tables contain more than two variables. The Titanic data set mentioned in section 1.1. is a classic case of a contingency table with a dependent variable and more than one factor, allowing researchers to make more nuanced analyses. In this thesis, the focus will be on three-dimensional contingency tables, i.e., count data where a third variable is available, which will be denoted as layers k in this thesis. Although the calculation of expected cell counts in three-dimensional tables follow the same rules as for the two-dimensional case, the number of possible ways to model the cell probabilities is much larger. The different hypotheses differ in the assumptions for the association between rows, columns and layers. In the following section, an overview of possible associations between rows, columns and layers will be given with a focus on the conditional independence between two variables.

1.3 Conditional independence (in $2 \times 2 \times K$ tables)

1.3.1 Conceptual overview

One common test for three-dimensional contingency tables is the test of conditional independence (Agresti, 2003). Conceptually, this means that one can test whether two categorical variables are independent of each other after controlling for the effect of a third variable. This is relevant whenever a third variable influences the strength of association of the two other variables. An example where this would be relevant is when a researcher investigates whether smoking/no-smoking has an influence on whether individuals are categorized as healthy/non-healthy after controlling for the general diet choices of the subjects. The test of conditional independence would then assess if smoking has an effect on the health of an individual beyond the effect of the individual diet choices. Another relevant case where the conditional independence test plays a role is "Simpson's paradox" (Simpson, 1951), which describes a situation where an association between two variables disappears after controlling for a third variable. In such situations, ignoring the effect of the third variable may easily lead to wrong conclusions about the association between two variables of interest (Albers, 2015). Under the model of rows and columns being independent given a layer k, we have the cell probabilities $p_{ijk} = p_{i.k} \times p_{.jk}/p_{..k}$, for which the Maximum Likelihood Estimates (MLE) are

$$\begin{aligned} p_{i.k} &= \frac{\sum_{j=1}^{J} n_{ijk}}{n} \\ p_{.jk} &= \frac{\sum_{i=1}^{I} n_{ijk}}{n} \\ p_{..k} &= \frac{\sum_{i=1}^{I} \sum_{j=1}^{J} n_{ijk}}{n} \end{aligned} \tag{1}$$

where n refers to the total cell count, and n_{ijk} refers to the cell count in row i, column j and layer k (Christensen, 1997). Obtaining the expected cell counts $E_{ijk} = p_{ijk} \times n_{..k}$, it is possible to calculate the Chi-Squared test statistic, based on which the null hypothesis of conditional independence can be rejected or not (Agresti, 2003; McHugh, 2013). For three-way tables, there are seven relevant hypotheses that one can consider (Christensen, 1997):

1. H_0: Rows, columns and layers are independent of each other.
2. H_1: Rows are independent of columns and layers.
3. H_2: Columns are independent of rows and layers.
4. H_3: Layers are independent of rows and columns.

5. H_4: Rows and columns are independent given layers.
6. H_5: Rows and layers are independent given columns.
7. H_6: Layers and columns are independent given rows.
8. H_7: (Conditional) odds ratios of rows and columns is the same (but not zero) for every layer k.

When discussing the GLMs below, we identify for each hypothesis H_i a corresponding generalized linear model $\mathcal{M}_i$. The model comparison of interest is between H_4 and H_7.

1.3.2 The Cochran–Mantel–Haenszel test

The Cochran–Mantel–Haenszel (CMH) test is a special case of testing the conditional independence that can be used for a $2 \times 2 \times K$ design, in which the independence of rows i and columns j conditioned on layer k is assessed (Cochran, 1954). Although the number of rows and columns (i.e., the number of categories of the predictor and outcome variables) is fixed to two, the number of layers (i.e., the number of categories of the third covariate) has no upper bound. Generally, the formula for computing the test statistic of the CMH test is

$$t = \frac{[\sum_{k=1}^{K}(O_{11k} - E_{11k})]^2}{\sum_{k=1}^{K}\left(\frac{\sum_{j=1}^{2} n_{1jk} \sum_{j=1}^{2} n_{2jk} \sum_{i=1}^{2} n_{i1k} \sum_{i=1}^{2} n_{i2k}}{n_k^2(n_k-1)}\right)} \tag{2}$$

where the numerator shows the squared sum of differences between the expected cell count (E_{11k}) and the observed cell count (O_{11k}) of the first cell across layers k. The expected cell count E_{11k} assumes that rows and columns are independent given layers. The denominator shows the variance of the frequency values. Approximate p-values can be computed for this statistic, as it is asymptotically chi-squared distributed with one degree of freedom. It is common to add a continuity correction (to account for the fact that we are approximating a continuous probability distribution with discrete values) by subtracting 0.5 from the absolute difference between the observed and expected cell values in the numerator of the equation. Alternatively to expressing the null hypothesis in terms of the difference of the observed and expected cell counts under independence, one can also express the null hypothesis in terms of an odds ratio: If rows i and columns j are independent for every layer k, then the expected cell count for the MLE can be expressed as

$$\lambda_{ijk} = p_{i.k} \times p_{.jk} \times \frac{1}{p_{..k}} \times n_{..k} \tag{3}$$

(Christensen, 1997) where λ_{ijk} refers to the expected cell count in row i, column j and layer k. When computing the ratio of the expected cell counts, the model of conditional row-column independence results in

$$\frac{\lambda_{1jk}}{\lambda_{2jk}} = \frac{p_{1.k}}{p_{2.k}} \tag{4}$$

for both column 1 and column 2 (noting that the right-hand side in the equation above is independent of j). Consequently, assuming conditional independence of rows and columns for each layer k results in the same ratio of rows for every column j. By stating that

$$\frac{\lambda_{11k}}{\lambda_{21k}} = \frac{\lambda_{12k}}{\lambda_{22k}} \tag{5}$$

we equivalently state that the odds ratio of being assigned to row i across different columns j is equal to 1 for each layer k.

While the CMH test has been widely used by researchers, its reliance on p-values makes the test vulnerable to the known shortcomings of frequentist hypothesis testing. As an alternative to frequentist statistics, the next chapter will focus on how to test conditional independence in contingency tables with Bayesian statistics. At first, we will give a general overview of Bayes factors. Secondly, we will discuss two common GLM structures through which the likelihood needed for the calculation of Bayes factors can be modelled. Lastly, a quick overview of different prior variants (those that were used for the simulation study) will be given.

2 Bayes factors for conditional independence

2.1 Motivation for the usage of Bayes factors

In the social and medical sciences, there are two major paradigms on how statistical inference can be done, one being referred to as frequentist statistics and the other as Bayesian statistics. Since the former has been the standard in research practices for many decades, many methods specifically designed to test hypotheses have been thoroughly developed in the framework of frequentist statistics in the form of Null Hypothesis Significance Testing (NHST), whereas many common tests used in social and medical sciences do not yet have extensively studied Bayesian alternatives to offer. Although NHST has its merits and in some situations should be the preferred way to make statistical inference, the benefits for researchers to seek out alternatives to NHST have been made widely apparent (Wagenmakers et al., 2018). For example, Bayesian inference allows researchers to quantify evidence for the null hypothesis (Rouder, 2014), allows researchers to incorporate prior knowledge/data about the likelihood of hypotheses (Gronau et al., 2019) and in some cases allows for continuous monitoring of test outcomes during the data collection process (Hendriksen et al., 2021).

2.1.1 Definition of Bayes factors

In the Bayesian framework, a common way of testing hypotheses is through the usage of Bayes factors that compare the (posterior) probability of two opposing statistical models (Wagenmakers et al., 2018). These two models can reflect two different hypotheses, e.g., $\mathcal{M}_1$ representing the null model and $\mathcal{M}_2$ expressing the alternative hypothesis. The Bayes factor of $\mathcal{M}_1$ compared to $\mathcal{M}_2$ is then defined as the ratio of the odds of the posterior model probabilities over the odds of the prior model probabilities:

$$\mathrm{BF}_{12} = \frac{P(\mathcal{M}_1 \mid y)/P(\mathcal{M}_2 \mid y)}{P(\mathcal{M}_1)/P(\mathcal{M}_2)} \tag{6}$$

with the posterior probability of $\mathcal{M}_i$ is defined as

$$P(\mathcal{M}_i \mid Y) = \frac{P(Y \mid \mathcal{M}_i)P(\mathcal{M}_i)}{P(Y)} \tag{7}$$

where $P(Y)$ refers to the normalizing constant that ensures that the posterior is between 0 and 1. However, for the calculation of Bayes factors, this term cancels. $P(\mathcal{M}_i)$ refers to the prior probability of $\mathcal{M}_i$ and reflects the prior certainty of model i being true. While this term can reflect a researcher's expectation about the likelihood of the two models, it is often set uniformly to 0.5 for both $\mathcal{M}_1$ and $\mathcal{M}_2$ such that it cancels as well (Ly et al., 2016). The most important part is therefore $P(Y|\mathcal{M}_i)$, which refers to the marginal likelihood of the data at hand given model i. The marginal likelihood for $\mathcal{M}_i$ is defined as

$$P(Y \mid \mathcal{M}_i) = \int \mathcal{L}(Y \mid \theta, \mathcal{M}_i)\pi_i(\theta)\mathrm{d}\theta \tag{8}$$

where $\mathcal{L}(Y|\theta, \mathcal{M}_i)$ refers to the likelihood of the data for the parameter values given $\mathcal{M}_i$. $\pi_i(\theta)$ is the model-specific prior chosen for θ. The product of the likelihood and the priors of the model parameters can be seen as a weighting of the likelihood function by the specified prior probability of the parameter values. Through the integral over θ, we marginalize the choice of θ out of the equation, which allows us to do model comparisons that do not rely on fixed values for the model parameters. The full form of the Bayes factor is thus

$$\mathrm{BF}_{12} = \frac{\int \mathcal{L}(Y \mid \theta, \mathcal{M}_1)\pi_1(\theta)\mathrm{d}\theta}{\int \mathcal{L}(Y \mid \theta, \mathcal{M}_2)\pi_2(\theta)\mathrm{d}\theta} \tag{9}$$

and quantifies the strength of evidence of one hypothesis with respect to another hypothesis (Ly et al., 2016). One common sub-class of hypothesis testing is variable (or model) selection, i.e., where one model includes one or more variables of interest and the competing model does not include these parameter(s). In such nested model comparisons, there will be a set of common parameters and a set of non-common parameters. Notionally, the θ vector

is thus split up into α, representing the parameter common to both models and β which is a test-relevant parameter. By defining a model for each hypothesis, one can then compute Bayes factors to decide whether a variable should be included in the model or not. As can be seen in the equations above, one important component to construct Bayes factors is the likelihood of the data for each of the competing models, which, for instance, can be modelled via generalized linear models. In the next section, we will give an overview of how the conditional independence test for contingency tables can be represented using generalized linear models.

2.2 Generalized linear models

Instead of using the Cochran–Mantel–Haenszel test, multiple alternatives have been proposed that make use of generalized linear models, which allows for more flexible generalizations and adaptations (Christensen, 1997; von Eye and Indurkhya, 2000). Classically, the conditional independence tests based on generalized linear models have been discussed with a focus on frequentist inference (e.g., p-values). However, one can also use the GLM framework to model the likelihood of the data for constructing Bayes factors (Li and Clyde, 2018). Generalized linear models (GLMs) are derived from the exponential family with the density function

$$f(y;\theta,\phi) = \exp\Big(\frac{y\theta - b(\theta)}{\phi}\Big) + c(y,\theta) \tag{10}$$

where ϕ is a scaling parameter and

$$\theta = g(\mu) = \alpha + \beta\mathbf{X}$$

with g being the (canonical) link function and μ the expectation of the response variable y conditional on $\mathbf{X}$ (McCullagh and Nelder, 2019). In the next two sections, we identify two GLM modelling approaches that can be used to test conditional independence in contingency tables.

2.2.1 Conditional independence test as log-linear model comparison

As has been shown (von Eye and Indurkhya, 2000), the test of conditional independence in contingency tables can also be represented by a nested model comparison of log-linear models which are also known as Poisson regression models. For Poisson regression, the values of the response variable y are assumed to be non-negative integer values and have the probability mass function

$$f(y \mid \lambda) = \frac{\lambda^y e^{-\lambda}}{y!} \tag{11}$$

where λ is a rate parameter and is both the expected value and the variance of y. The link function that is used in Poisson regression is the logarithm, leading to

$$\log(\mathrm{E}(Y \mid \mathbf{X})) = \alpha + \beta\mathbf{X} \tag{12}$$

where $\mathrm{E}(Y|\mathbf{X})$ is the expected value of Y given $\mathbf{X}$ and α and β are the model parameters. The authors represent the model of conditional independence of rows and columns by modelling the expected frequencies of each cell in $I \times J \times K$ contingency tables by assuming that each cell follows a Poisson distribution with the parameter λ_{ijk}. More specifically, the expected cell values are thus modelled via a linear function on the logarithm of λ_{ijk}. The resulting model equation consists of main effects (for rows, columns and layers) and two-way interactions. The conditional independence test refers to testing if a two-way interaction between rows and columns exists. More specifically, this is done through the general log-linear equation

$$\log(\vec{\lambda}) = \mathbf{X}\vec{U}$$

where $\log(\vec{\lambda})$ is a $(IJK \times 1)$ vector of expected cell counts, $\mathbf{X}$ is a $IJK \times 7$ design matrix, and $\vec{U}$ is a 7×1 vector of parameters. The authors use the sum-to-zero constraint for the

design matrix, which in the case of a $2 \times 2 \times 2$ contingency table leads to the following model design (that can easily be extended for $K = 3, 4, 5 \dots$)

$$log \begin{bmatrix} \lambda_{111} \\ \lambda_{211} \\ \lambda_{121} \\ \lambda_{221} \\ \lambda_{112} \\ \lambda_{212} \\ \lambda_{122} \\ \lambda_{222} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & -1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & 1 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \\ u_{13} \\ u_{23} \\ u_{12} \end{bmatrix}$$

Here, u_0 denotes the intercept, u_1 is the row effect, u_2 is the column effect, u_3 is the layer effect, u_{13} is the interaction of rows and layers, u_{23} is the interaction of columns and layers and finally, u_{12} is the interaction of rows and columns. Since we are interested in the association between rows and columns for the test of conditional independence, u_{12} is the test-relevant parameter. Through the inclusion of u_3, u_{13} and u_{23}, the effect of layer k on both rows and columns is accounted for, making u_{12} the interaction effect of rows and columns after controlling for the effect of layers. Thus, the null hypothesis is that $u_{12} = 0$ and the alternative is that $u_{12} \neq 0$. One can now do a nested model comparison by comparing the model including all u parameters mentioned in the matrix equation above with a model that excludes the u_{12} term. In classical statistics this can be done through, for instance, the likelihood ratio test to assess if the null hypothesis holds. For the analysis of $2 \times 2 \times K$ contingency analysis, we only need to model the expected values for two rows and columns and K layers respectively (and their interactions). This means that using the sum-to-zero constraint, only one parameter needs to be estimated for each of the effects that do not depend on layers and $K - 1$ parameters for the ones that include layers. Consequently, the number of parameters is $3K+1$ for the model including the row-column interaction term and $3K$ for the restricted model. As with the original CMH test, the log-linear model comparison test statistic will have an asymptotic Chi-Squared distribution with one $(3K+1-3K)$ degree of freedom.

Note that the model with the u_{12} included corresponds to H_7 in section 1.3.1 whereas the model without u_{12} refers to H_4. These two models will be referred to as $\mathcal{M}_7$ and $\mathcal{M}_4$, respectively. This means that the nested model comparison for $\mathcal{M}_7$ against $\mathcal{M}_4$ compares a model where the odds ratios for each layer are the same for every layer k ($\mathcal{M}_7$), as controlled by the u_{12} parameter, to a model where the odds ratio per layer k is equal to one ($\mathcal{M}_4$). The resulting test statistic is equivalent to computing the MLE estimates $\hat{m}_{ijk}^{(7)}$ for $\mathcal{M}_7$ and $\hat{m}_{ijk}^{(4)}$ for the restricted model and plugging these values into the formula

$$t = \sum_{i=1}^{2} \sum_{j=1}^{2} \sum_{k=1}^{K} \frac{(\hat{m}_{ijk}^{(7)} - \hat{m}_{ijk}^{(4)})^2}{\hat{m}_{ijk}^{(4)}} \tag{13}$$

for any $2 \times 2 \times K$ design (Christensen, 1997). Alternatively to the sum-to-zero restriction which is used by Van Eye (2000), it is also possible to use the set-to-zero restriction (Yandell, 2017), which will not change the number of parameters, but changes the interpretation of the parameter values. For the set-to-zero restriction, it is common to simply set the first (or last) level of each of the factors to 0. In the case of the $2 \times 2 \times 2$ design, this would mean that the row, column and layer effects (and interaction effects) are set to zero for one of the two rows, columns and layers, respectively. For example, setting the row effect for the first row to 0 would mean that this parameter value would show the difference between the first and the second row. For the set-to-zero restriction (using $K = 2$), the design matrix is as follows

$$log \begin{bmatrix} \lambda_{111} \\ \lambda_{211} \\ \lambda_{121} \\ \lambda_{221} \\ \lambda_{112} \\ \lambda_{212} \\ \lambda_{122} \\ \lambda_{222} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \\ u_{13} \\ u_{23} \\ u_{12} \end{bmatrix}$$

The model equation (for the $2 \times 2 \times 2$ design) can be written as

$$\log(\lambda_{ijk}) = u_0 + x_i u_1 + x_j u_2 + x_k u_3 + x_{ik} u_{13} + x_{jk} u_{23} + x_{ij} u_{12} \tag{14}$$

where $x_i, x_j, ..., x_{ij} = 1$ if none of its subscript values are 1 and equal to 0 otherwise. In the set-to-zero representation, $u_0 = log(\lambda_{111})$, thus the expected value of the first cell in the first layer (λ_{111}) is e^{u_0}. Similarly, the value of λ_{211} is $e^{u_0+u_1}$, et cetera. It is important to note that the effect of the parameters on the expected cell count is multiplicative and that each expected cell count is a function containing the intercept u_0 (e.g., $\frac{\lambda_{211}}{\lambda_{111}} = e^{u_0}$). Consequently, adding a constant $\log(C)$ to the intercept is equivalent to multiplying each of the λ_{ijk} with the constant C, thus keeping the ratios of the frequency values unchanged. To see the link between the log-linear representation and the CMH test, the odds ratio of rows and columns for every layer k is the following for $u_{12} = 0$:

$$\begin{aligned} \frac{(\lambda_{11k}/\lambda_{21k})}{(\lambda_{12k}/\lambda_{22k})} &= \frac{e^{(u_0+x_k u_3)-(u_0+u_1+x_k u_3+x_{ik} u_{13})}}{e^{(u_0+u_2+x_k u_3+x_{jk} u_{23})-(u_0+u_1+u_2+x_k u_3+x_{ik} u_{13}+x_{jk} u_{23})}} \\ &= \frac{e^{-u_1-x_{ik} u_{13}}}{e^{-u_1-x_{ik} u_{13}}} \\ &= 1. \end{aligned} \tag{15}$$

If, however, $u_{12} \neq 0$, then

$$\begin{aligned} \frac{(\lambda_{11k}/\lambda_{21k})}{(\lambda_{12k}/\lambda_{22k})} &= \frac{e^{(u_0+x_k u_3)-(u_0+u_1+x_k u_3+x_{ik} u_{13})}}{e^{(u_0+u_2+x_k u_3+x_{jk} u_{23})-(u_0+u_1+u_2+x_k u_3+x_{ik} u_{13}+x_{jk} u_{23}+u_{12})}} \\ &= \frac{e^{-u_1-x_{ik} u_{13}}}{e^{-u_1-x_{ik} u_{13}-u_{12}}} \\ &= e^{u_{12}}. \end{aligned} \tag{16}$$

Thus, u_{12} can be interpreted as the k-conditional log odds ratio between rows and columns given layer k. Testing the hypothesis of odds ratios all being equal to 1 translates to the hypothesis that $u_{12} = 0$. Furthermore, there is an interchangeability of rows and columns: Switching rows and columns in the odds ratio shown on the left-hand side of formula 8, we have

$$\begin{aligned} \frac{\lambda_{11k}/\lambda_{12k}}{\lambda_{21k}/\lambda_{22k}} &= \frac{e^{(u_0+x_k u_3+x_{ik} u_{13})-(u_0+u_2+x_k u_3+x_{jk} u_{23})}}{e^{(u_0+u_1+x_k u_3+x_{ik} u_{13})-(u_0+u_1+u_2+x_k u_3+x_{ik} u_{13}+u_{12})}} \\ &= \frac{e^{-u_2-x_{jk} u_{23}}}{e^{-u_2-x_{jk} u_{23}-u_{12}}} \\ &= e^{u_{12}}. \end{aligned} \tag{17}$$

Consequently, we have

$$\frac{\lambda_{11k}/\lambda_{21k}}{\lambda_{12k}/\lambda_{22k}} = \frac{\lambda_{11k}/\lambda_{12k}}{\lambda_{21k}/\lambda_{22k}} \tag{18}$$

for any layer k. For both ratios, the row effect and row-layer interaction effect cancels in the same manner the column effect and column-layer effect cancels, respectively.

For the frequentist case, the simulation results show that the log-linear representation indeed delivers a close approximation of the CMH test: Calculating the difference between the p-values of the log-linear test and the CMH test for different sample sizes, we observe that as the sample size increases, the difference between p-values of the two test approaches

diminishes logarithmically (as can be seen in Figure 1). One explanation for this is that the CMH test approximates a Chi-Squared distribution under the null for large samples. Furthermore, Agresti (2018) notes that "similarity of results for the likelihood-ratio, Wald, and CMH (score) tests usually happens when the sample size is large." For $n = 3000$, the average absolute difference of the p-values was less than 0.0005.

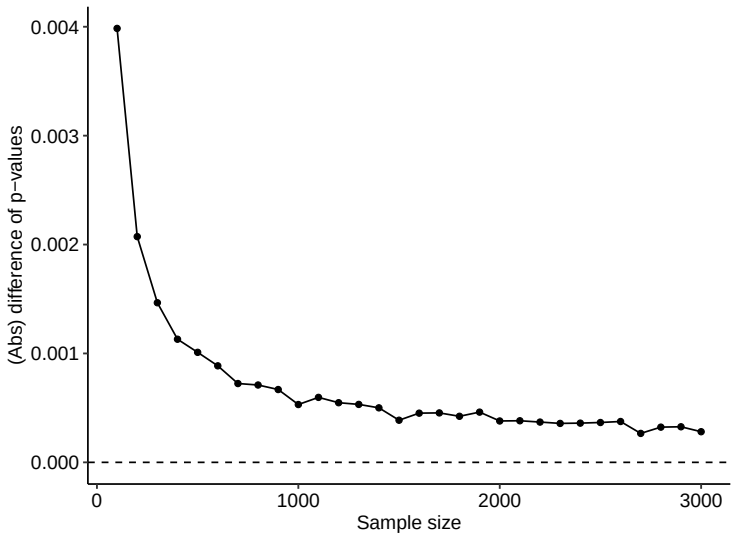


Figure 1: Mean absolute difference (for simulated data) between the p-values for the CMH test (without correction) and its log-linear representation as a function of the sample size n for $u_{12} = 0$

2.2.2 Conditional independence test as logistic model comparison

Alternatively to modelling the individual frequency values in contingency tables, it is possible to model the odds directly (Christensen, 1997). Doing so, one can test the conditional independence through a model comparison of two logistic regression models using the Logit link function (Agresti, 2003). For a $2 \times 2 \times 2$ contingency table, this can be achieved by defining the model

$$\log\left(\frac{p_{1jk}}{1 - p_{1jk}}\right) = \beta_0 + x_j\beta_{column} + x_k\beta_{layer} \tag{19}$$

and a restricted model

$$\log\left(\frac{p_{1.k}}{1 - p_{1.k}}\right) = \beta_0 + x_k\beta_{layer} \tag{20}$$

where $x_j = 1$ when $j = 1$ and $x_j = 0$ when $j = 2$ and (in the case of two layers) $x_k = 1$ when $k = 1$ and $x_k = 0$ when $k = 2$. Similar to the Poisson case, one can now do a model comparison to test the hypothesis of conditional independence. It should be noted that a logistic regression model usually requires a dichotomous dependent variable. However, one can exploit the binary property of rows and columns in $2 \times 2 \times K$ contingency tables, such that either row assignment (or column assignment) is treated as the outcome variable, while the assignment to columns (or rows) and layers are the predictor variables. In order to utilize the information of the frequency values, one would simply replicate each observation m_{ijk} times. As an example, consider the following fictional contingency table:

Layer	Row	Column	
		1	2
1	1	3	2
	2	1	3
2	1	2	4
	2	3	1

Table 1: An example of a $2 \times 2 \times 2$ contingency table

Reformatting this contingency table, we obtain an expanded table, in which the assignment to row 1 is marked as 0 and the assignment to the second row is marked as 1. While the meaning of the effect size parameter is conceptually the same, the log-linear model has the added advantage that it is flexible in terms of the number of rows and columns in the data set, while the logistic regression model allows for only two possible response values (i.e., only two rows or columns). The advantage of the expanded data representation for the logistic regression approach is that the link between the sample size and the total cell count can be more easily established compared to the log-linear regression approach (which by default assumes $n = 2 \times 2 \times K$ instead of the cell counts themselves).

Var1	Var2	Var3
0	A	A
0	A	A
0	A	A
1	A	A
0	B	A
0	B	A
1	B	A
1	B	A
1	B	A
0	A	B
0	A	B
1	A	B
1	A	B
1	A	B
1	A	B
0	B	B
0	B	B
0	B	B
1	B	B

Table 2: Expanded contingency table for logistic regression

To see the similarity of the logistic and log-linear model, we will focus on the similarity between the parameter u_{12} and β_{column}. In the model with β_{column} included, p_{1jk} represents the probability of being assigned to row 1 for column j and layer k and $\log\left(\frac{p_{1jk}}{1-p_{1jk}}\right)$ are the log odds of being assigned to row 1 for column j and layer k. Therefore, $\left(\frac{p_{1jk}}{1-p_{1jk}}\right) = e^{(\beta_0+\beta_{layer})} \times e^{\beta_{column}}$. Under the null hypothesis, the log odds given layer k are the same across columns and

$$\frac{\left(\frac{p_{1.k}}{1-p_{1.k}}\right)}{\left(\frac{p_{1.k}}{1-p_{1.k}}\right)} = \frac{e^{\beta_0+x_k\beta_{layer}}}{e^{\beta_0+x_k\beta_{layer}}} = 1. \tag{21}$$

Under the alternative hypothesis, we have

$$\frac{\left(\frac{p_{11k}}{1-p_{11k}}\right)}{\left(\frac{p_{12k}}{1-p_{12k}}\right)} = \frac{e^{\beta_0+x_k\beta_{layer}+\beta_{column}}}{e^{\beta_0+x_k\beta_{layer}}} = e^{\beta_{column}}. \tag{22}$$

As with the u_{12} parameter of the log-linear model, the β_{column} parameter represents the log odds ratio for rows and columns conditional on layer k. In fact, our simulation results show that not only are the MLE estimates of β_{column} and u_{12} identical but also the LR-Chi-Square test statistic (and the p-value accordingly) are the same. In line with this, it should be noted that the Score test of the *conditional* logistic regression model is identical to the Cochran–Mantel–Haenszel test (Day and Byar, 1979). One major difference between the Poisson and the logistic approach is that the number of parameters needed for the restricted model and full model are now K (the number of layers) and $K+1$, respectively (as opposed to $3 \times K$ and $3 \times K + 1$). The reason why the number of parameters is reduced in the logistic regression model is that we model the odds directly instead of modelling the individual frequency values.

Using the "glm()" function of the "stats" R-package (R Core Team, 2019), the simulation results show that the nested model comparison with the log-linear model (as defined by Van Eye, 2000) is practically identical to the logistic regression model comparison in terms of the resulting test statistic and p-values. In fact, the simulation results show that the mean absolute difference between the p-values of the Poisson and logistic regression model comparisons is less than 0.00005. This is true for all specifications of the effect parameter u_{12}. Due to the interchangeable results of the p-values for the two model approaches, the remainder of the frequentist comparison will focus on the log-linear version.

2.3 Variants of g-priors

Although the θ parameters are integrated out of equation 9 in section 2.1.1, the Bayes factor nevertheless heavily depends on the choice of priors for the non-common parameters which will therefore have to be chosen with care. For example, the non-common β parameters should be proper (meaning that they must integrate to 1), since non-common parameters appear in only one of the models and the arbitrary constants created through inproper priors do not cancel in the Bayes factor equation (Kass and Raftery, 1995).

For generalized linear models, there are several proposed priors based on which Bayes factors can be constructed that have desirable properties for model selection. One such approach to construct Bayes factors for a GLM is through (mixtures of) g-priors (Li and Clyde, 2018): g-priors have been originally developed for Gaussian models (Zellner, 1986) and can be used as objective priors for regression parameters (for instance in Bayesian variable selection problems). For the Gaussian case, using g-priors on the β coefficients leads to the following prior distribution for the β parameters:

$$\beta \mid \sigma^2 \sim N(\beta_0, g\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) \tag{23}$$

where β_0 is usually set to 0 and g serves the purpose of scaling the covariance matrix of the MLE (and consequently can be used as shrinkage for parameter estimation). While it is possible to define a fixed value for g for both normal linear models and generalized linear models, this is generally advised against (Li and Clyde, 2018). Alternatively, it is also

possible to define a prior distribution for g itself, which is commonly referred to as mixtures of g-priors.

Below, an overview of different variants of these mixtures of g-priors for generalized linear models will be given. The priors used in this thesis refer to a subset of priors mentioned in chapter 3.2 by Li and Clyde (2018), which gives an overview of so-called CHIC ("Confluent Hypergeometric Information Criterion") priors. In addition to the priors mentioned below, we also studied two g-prior variants where g is a fixed value (to benchmark the CHIC prior results). More specifically, this was done by setting $g = 100$ and by setting $g = n$ (Kass and Wasserman, 1995), which can be referred to as unit information prior.

1. Beta-prime prior

   Setting a Beta-prime prior for g (Maruyama and George, 2011), we have the following distribution for $u = 1/(1+g)$ (u not to be confused with the u parameters in the log-linear model equation):

   $$u \sim \text{Beta}(\frac{1}{4}, \frac{n - p_m - 1.5}{2}) \tag{24}$$

2. Hyper-g/n prior

   When specifying $a = 1, b = 2, r = 1.5, s = 0, v = 1, k = 1/n$ for the CHIC g-prior, one recreates the hyper-g/n prior (Liang et al., 2008):

   $$p(g) = \frac{a_h - 2}{2n}\left(\frac{1}{1 + g/n}\right)^{a_h/2} \tag{25}$$

   where $2 < a_h \leq 4$.

3. Robust prior

   In the simulation study we will use the default recommended parameters for the robust prior (Bayarri et al., 2012), leading to a Truncated Gamma prior. The Truncated Gamma prior has the following distribution for u:

   $$u \sim \text{TG}_{(0, \frac{p_m+1}{n+1})}\left(\frac{1}{2}, 0\right) \tag{26}$$

   In the case of the Truncated Gamma prior, there is an upper bound which is increasing as p_m increases and decreasing as n increases (Li and Clyde, 2018).

4. Intrinsic Prior

   Using the intrinsic prior (Berger and Pericchi, 1996), the prior for g in this case is the following:

   $$\begin{aligned} g &= \frac{n}{p_m + 1} \times \frac{1}{\omega} \\ \omega &\sim \text{Beta}(\frac{1}{2}, \frac{1}{2}) \end{aligned} \tag{27}$$

   In this case, there is a lower bound of g set at $n/(p_m + 1)$ (Li and Clyde, 2018).

3 Simulation study

3.1 Overview

The main goal of this simulation study was to investigate the suitability of the proposed Bayes factors of the BAS package to test the conditional independence in contingency tables. More specifically, Bayes factors were computed (constructed from certain CHIC priors) for simulated data sets for which the underlying effect sizes will be specified. The Bayesian version of the test will be compared to the frequentist version. In addition, the suitability of different g-priors using the log-linear model comparison will be put in contrast to the logistic model comparison. The next section shows which criteria were used as guidelines to assess the priors in question.

3.2 Criteria for Bayesian tests of conditional independence

1. Predictive matching and model selection consistency

   Increasing the sample size (total cell count) of the generated data sets should, on average, yield an exponentially increasing Bayes factor for data generated from the alternative with a true underlying association between rows and columns. Likewise, for data generated from the null hypothesis, we would expect that the Bayes factor decreases on average as the sample size increases. In both cases (i.e., the alternative or null hypothesis is true), the Bayes factor should converge towards 1 as the sample size decreases (Bayarri et al., 2012).

2. Test specificity and sensitivity

   In order to assess the practical relevance of the test, the test should (within comparably acceptable error rate ranges) be able to detect (1) the presence of an effect for data generated with a known underlying association, and (2) the absence of any association for data generated from a model of underlying independence. In general, any Bayes factor BF_{12} comparing $\mathcal{M}_1$ with $\mathcal{M}_2$ greater than 1 is (at least slightly) favoring $\mathcal{M}_1$ (the alternative hypothesis) while $\text{BF}_{12} < 1$ indicates that the null hypothesis is favoured. For this reason, $\text{BF}_{12} = 1$ was used as the threshold to investigate the general effects of respective g-priors. In addition, in order to better interpret the results of the Bayes factor, the following guidelines were used as a classification of the strength of evidence provided by the Bayes factors (van Doorn et al., 2021).

Bayes factor BF_{12}	Interpretation
> 100	Extreme evidence for $\mathcal{M}_1$
100 - 30	Very strong evidence for $\mathcal{M}_1$
30 - 10	Strong evidence for $\mathcal{M}_1$
10 - 3	Moderate evidence for $\mathcal{M}_1$
3 - 1	Anecdotal evidence for $\mathcal{M}_1$
1	No Evidence
1/3 - 1	Anecdotal evidence for $\mathcal{M}_2$
1/10 - 1/3	Moderate evidence for $\mathcal{M}_2$
1/30 - 1/10	Strong evidence for $\mathcal{M}_2$
1/100 - 1/30	Very strong evidence for $\mathcal{M}_2$
$< 1/100$	Extreme evidence for $\mathcal{M}_2$

Table 3: Guidelines for the interpretation of Bayes factors comparing $\mathcal{M}_1$ to $\mathcal{M}_2$

3. Influence of nuisance parameters

   As we are testing only one effect parameter, it is assessed to which degree the respective Bayes factor tests can isolate the test-relevant parameter without being affected by changes in the nuisance parameters. To study this, we defined three values for u_1 (-1, 0 and 1) and repeated the simulation process as outlined in Figure 2. While u_1 is fixed across all the samples, the other nuisance parameters are still sampled from a uniform distribution between -1 and 1. Importantly, this is done repeatedly for multiple sets of sampled nuisance parameters to assure that the effect of the remaining nuisance parameters (those that are not fixed) will cancel when studying the average pattern.

4. Model similarity

   Lastly, it is assessed if there is a correspondence between the conditional independence test using the Poisson regression and the logistic regression formulation as is the case in the frequentist framework.

3.3 The BAS package

As a basis for the simulation study of this thesis, the BAS package for R written by Clyde (2015) will be used. The BAS package enables computationally fast implementations of Bayesian Model Selection using Zellner's g-priors and mixtures of g-priors for (generalized)

linear models. Although the package entails several approaches to sampling potential models in a Bayesian variable selection problem, this functionality of the package will not be exploited in this thesis, as the conditional independence test only requires a comparison of two models for which the included parameters are fixed in advance. Using the BAS package, it is possible to compute Bayes factors for both Poisson and Binomial regression models. The relevant function of the package used is 'bas.glm()'.

3.4 Data generating process

Since the CMH test is designed for $2 \times 2 \times K$ contingency tables, $4 \times K$ data points have to be generated for each model comparison and each data point must be an integer value between $[0, \infty)$. In this simulation study, comparisons based on $2 \times 2 \times 2$ tables will be made but can be extended to K-layered designs. In order to sample positive integer values, each count m_{ijk} is assumed to follow the Poisson distribution

$$m_{ijk} \sim Pois(\lambda_{ijk}) \tag{28}$$

where λ_{ijk} represents the expected cell count for the i^{th} row, j^{th} column and k^{th} layer. Each λ_{ijk} term is based on u-parameters through

$$\begin{bmatrix} \lambda_{111} \\ \lambda_{211} \\ \lambda_{121} \\ \lambda_{221} \\ \lambda_{112} \\ \lambda_{212} \\ \lambda_{122} \\ \lambda_{222} \end{bmatrix} = \exp \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \\ u_{13} \\ u_{23} \\ u_{12} \end{bmatrix} \tag{29}$$

The simulation process should give full control over the sample size n (which is the total sum over all m_{ijk}) and the effect size (which is controlled by the u_{12} parameter), whereas the effect of both the sample size and the test-relevant effect size should not be affected by the nuisance parameters u_1, u_2, u_3, u_{13} and u_{23}. In order to achieve these requirements, the following steps describe the simulation process:

1. A value for the test-relevant u_{12} parameter needs to be specified (see 3.4.1).
2. A value n_e for the required sample size needs to be defined. This value will correspond to the total sum of cells that the simulated contingency table is required to have.
3. Next, values for the nuisance u-parameters need to be sampled. These are sampled from a uniform distribution between -1 and 1.
4. λ values are computed through the above-mentioned matrix multiplication of the u parameters and the model matrix, leading to one initial λ value per cell in the contingency table. After having computed the initial λ values, these will be updated through the following:

$$\lambda^*_{ijk} = \frac{\lambda_{ijk}}{\sum_{ijk} \lambda_{ijk}} n_e \tag{30}$$

where n_e represents the required overall sample size value as defined in step 2. This has the same effect as multiplying the intercept with a constant and will not affect the expected ratios of the λ values. The reason why this is done after sampling the u-parameter values is that the sampled u-parameters would otherwise change the overall expected sample size ($\sum \lambda_{ijk}$). Furthermore, this rescaling serves the purpose of eliminating any relationship between the initially sampled u parameters and the sample size (as otherwise higher values for the u parameters would result in a higher cell count).

5. Using the updated λ-values, $2 \times 2 \times K$ values are sampled from a Poisson distribution repeatedly until the desired sample size is reached, ignoring all samples for which the sum of all sampled m_{ijk} values is not the desired sample size.

6. Furthermore, for each specified sample size, U sets of nuisance parameters were sampled. For each set of nuisance parameters, N samples were drawn. By sampling multiple sets of the nuisance parameters, the simulation process cancels out any potential effect of the nuisance parameters.

   For each sample size that was specified, the median over all Bayes factors based on the data sets with the same sample size was calculated in order to prevent extremely large or small Bayes factors to affect the overall pattern. For each g-prior choice, this was repeated for n = [8,20, 30, ..., 90] to investigate the effect in the small sample size range and then n = [100, 200, 300, ..., 3000] to see the effect of larger sample sizes. For each sample size specification, 10000 contingency tables were simulated.

Figure 2 below shows how the simulation process per sample size n and u_{12} value works. The R scripts to create the data can be found on Github. After having created all data sets using the described sampling procedure, Bayes factors comparing the Poisson/logistic regression models $\mathcal{M}_7$ and $\mathcal{M}_4$ were computed for each of the g-prior special cases discussed in section 2.3. As a benchmark, three p-values were computed in parallel for each of the data sets using (1) the deviance statistic of the Poisson GLM representation, (2) the deviance statistic of the logistic GLM representation and (3) the original CMH test.

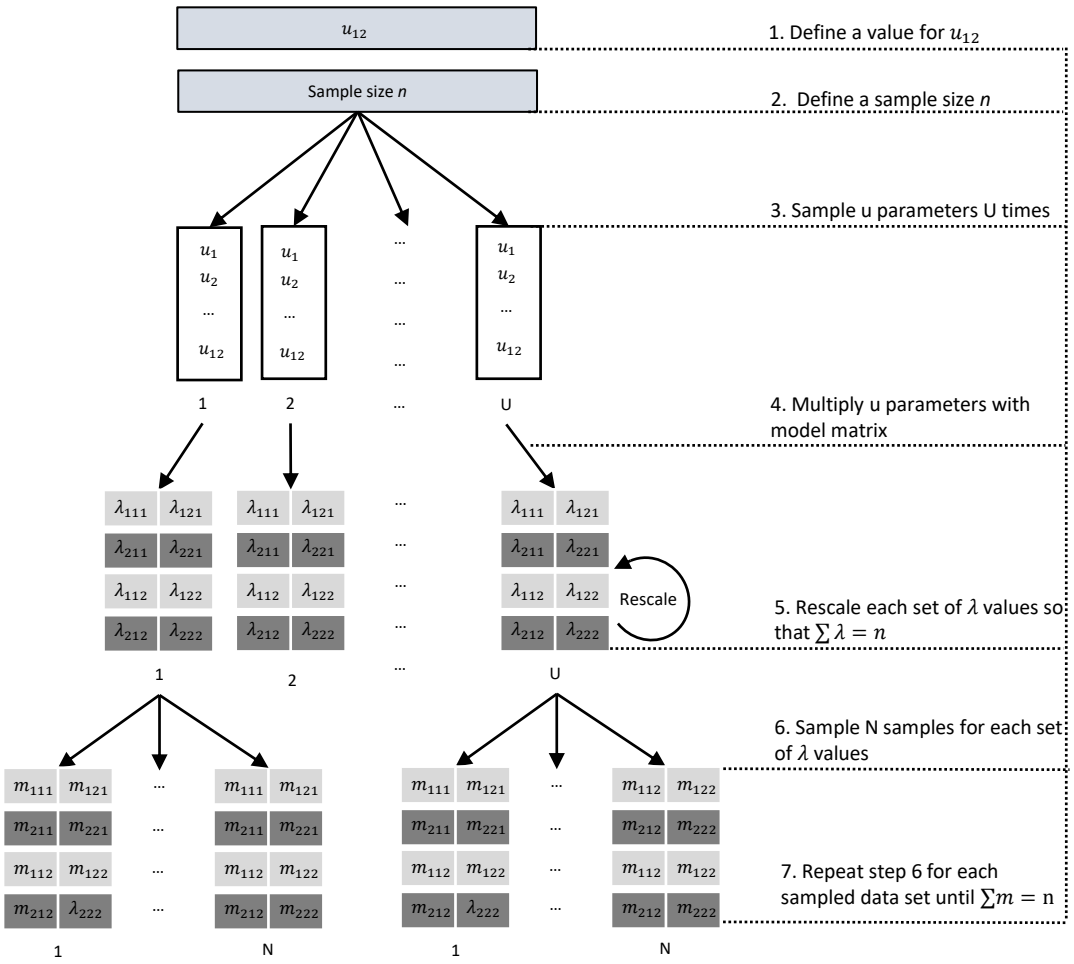


Figure 2: Simulation process per specified sample size n and u_{12} parameter.

3.4.1 Choice of the u_{12} parameter

Since the Bayes Factor behaves differently depending on whether the chosen effect size is large or small, it is necessary to define what can be considered a small or large effect size. As the u_{12} parameter can be interpreted as an odds ratio, threshold numbers to categorize odds ratios as small, medium or large need to be defined. The (conditional) odds ratio is standardized for any 2×2 design as well for the $2 \times 2 \times K$ design.

There have been different rules of thumbs to define these thresholds. For example, it has been proposed that for 2×2 designs, odds ratios of 1.68, 3.47, and 6.71 correspond to small, medium and large effects respectively, which are based on transformations of the odds ratio to Cohen's d = 0.2 (small), 0.5 (medium), and 0.8 (large) (Chen et al., 2010). Similarly, Hedges, Higgins and Rothstein (2009), also link Cohen's d classification of small medium and large effect sizes to odds ratios, ultimately leading to $u_{12} = 0.363$, $u_{12} = 0.907$ and $u_{12} = 1.451$ for a small, medium and large effect size, respectively. To verify that the Bayes factor is not affected by the sign of the u_{12} values, the simulations were also repeated with u_{12} = [-0.363, -0.907, -1.451], leading to 7 different u_{12} values (including $u_{12} = 0$).

3.4.2 Sample size considerations

At first, it is important to clarify how the sample size is defined with each of the testing approaches implemented in the simulation study. For tests that are specifically designed for contingency tables, such as the Chi-squared test or the Cochran–Mantel–Haenszel test, the sample size refers to the total count of cell values across all rows, columns and layers. However, in the case of Poisson regression models, the sample size would generally speaking refer to the number of observations in our data set. In the case of contingency tables, this implies that the sample size would always be fixed to $2 \times 2 \times K$, which means that increasing the cell count would not increase the sample size as normally understood for Poisson regression models.

For the logistic regression representation of the conditional independence test, the outcome variable is the assignment to the first or second row (or columns, as this can be done interchangeably). Consequently, we can expand our contingency table in such a way that each cell count represents one observation and has 1 or 0 as the outcome variable and the associated column and layer assignment as predictor variables. In this case, the definition of the sample size for both the original CMH test and the GLM representation is the same. For the remainder of this thesis, we will also use the term sample size when referring to the total cell count in the GLM representation. As the concept of the sample size plays an important role in the use of some of the g-priors for generalized linear models, it will be tested how the Bayesian logistic representation (i.e., models where the sample size in the calculations increases as the total cell count increases) compares to the log-linear representation for those g-priors in which the sample size is explicitly relevant. For example, for the hyper-g/n prior, the parameter k is set to $1/n$ (Li and Clyde, 2018). In the Poisson representation, k will therefore always be set to 1/8, regardless of the observed cell count.

In addition, a minimal sample size needs to be defined for the simulated data sets. Since any cells that contain a zero-value are potentially structurally zero, i.e., cells that by definition cannot be larger than 0, these structurally zero values "make the rejection of independence a forgone and uninteresting conclusion" (Wickens, 2014). To account for this potential limitation, it was decided that each cell of the $2 \times 2 \times K$ contingency table should be larger than 0, resulting in a minimal overall sum of the cell counts of $4 \times K$. Furthermore, this also accounts for the fact that the original CMH test can only be computed if $\sum_k^K m_{.jk}$ and $\sum_k^K m_{i.k}$ for all i and j are larger than 0 (i.e., the sum of each of the row and column sums across all layer k) are larger than 0 - as otherwise the denominator becomes zero in the calculation of the test statistic. In this case, some programs add a constant (e.g., 0.5) in case of zero value cells for the CMH test. However, for the purpose of this simulation, all samples which had any zero values in them were discarded.

4 Simulation findings

In the following sections, results are shown for Bayes factor variants as outlined in 2.3. For all prior variants, the same set of simulated contingency tables were used. Before presenting the results of each of the priors in detail, we will at first show the effect of the nuisance parameters in the conditional independence test for contingency tables.

4.1 Impact of the nuisance parameters

To see if the nuisance parameters play any role for the g-prior based Bayes factors for the conditional independence test in contingency tables, a new set of contingency tables was simulated. These data sets were created using a similar procedure as explained in section 3.4 with the exception that one of the nuisance parameters (e.g u_1) was fixed to a specified value between -1 and 1 while the other nuisance parameters were still sampled from a uniform distribution between 0 and 1. Importantly, multiple samples were simulated to assure that the effect of the remaining nuisance parameters cancels. This was then repeated for a range of values for that fixed nuisance parameter. Figure 3 shows how different values for u_1 affect the median Bayes factor across multiple sample size values when $u_{12} = 0.4$. As the figure suggests, the median Bayes factor appears to be systematically lower when $u_1 = 1$ compared to $u_1 = 0$ or $u_1 = -1$, indicating that the nuisance parameters do have an effect on the resulting Bayes factors. For the remaining studied g-prior variants, there are similar effects of the nuisance parameters. The respective figures can be found in the appendix. As a potential explanation, the effect of the nuisance parameters could be due to the sampling procedure for the simulated data sets. For this reason, the following changes were implemented:

1. Retaining all sampled data sets

   In the original simulation procedure, all samples that do not match the required sample size are discarded. As this could create a bias, the sampling process was repeated also while retaining all samples regardless of their actual cell count. Implementing this change, the effect of the nuisance parameters is still present.

2. Removing the re-scaling step

   Removing the step of re-scaling of the λ parameters also did not remove the effect of the nuisance parameters. The effect was also not removed by adjusting the initial value for the intercept u_0.

In addition, testing the effect of the nuisance parameters was repeated using the same approach as explained above for samples generated with $u_{12} = 0$. The results of this simulation are shown in Figure 4. In contrast to the results for $u_{12} = 0.4$, the results for $u_{12} = 0$ indicate that for data generated under the null, the effect of the nuisance parameters disappears. This can be seen by the lack of any systematic difference between the different values for u_1. While the above-mentioned tweaks indicate that the nuisance parameters influence the test of conditional independence in contingency tables, the lack of that effect for $u_{12} = 0$ leaves the possibility that the sampling procedure plays a role in the observed bias.

In the next section, simulation results focusing on the change in the Bayes factor as a function of the sample size and the effect size will be presented. The general structure for the figures of the different prior variants is the same: At first the results for all four effect size choices will be shown for $n \leq 100$ followed by results for just $u_{12} = 0$ and $u_{12} = 0.363$ when $n \geq 100$. Furthermore, the majority of figures relevant for the model selection criteria include a categorization of the Bayes factors in order to allow for easier visual interpretation of the plots. It is also important to note that the simulation results for $n \geq 100$ were done in sample size steps of 100. This means that the reported n thresholds (e.g., when a Bayes factor surpasses BF = 1) for larger n is also reported similarly.

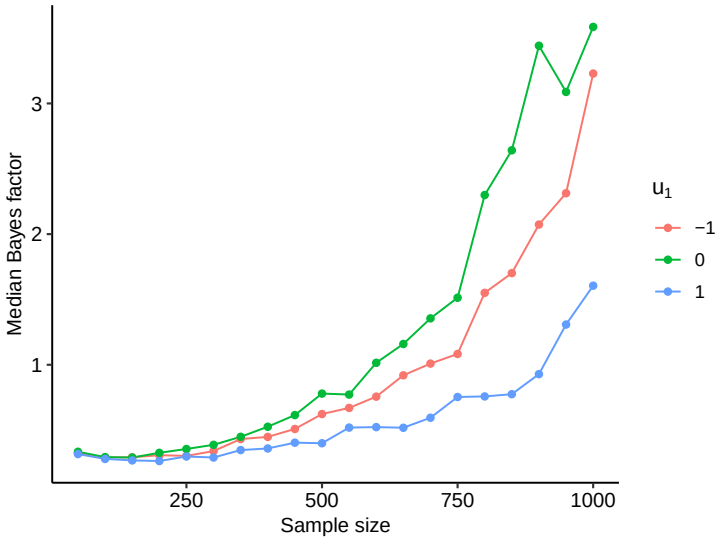


Figure 3: Median Bayes factor for different fixed u_1 values using the logistic regression model with the hyper-g/n prior and $u_{12} = 0.4$.

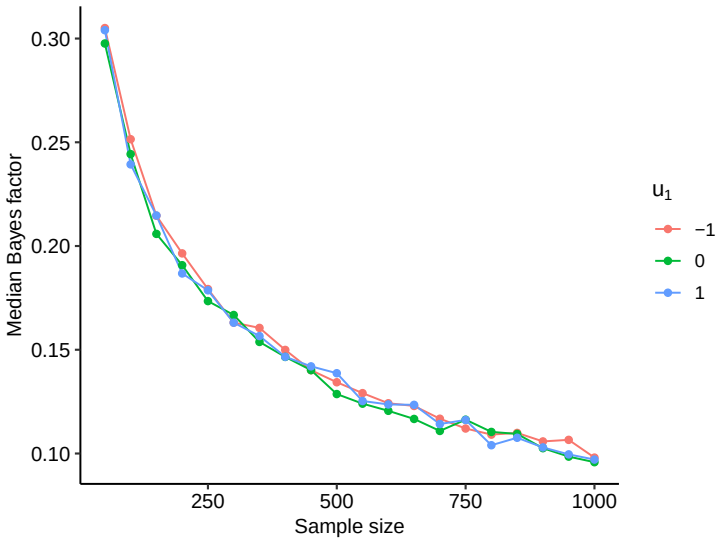


Figure 4: Median Bayes factor for different fixed u_1 values using the logistic regression model with the hyper-g/n prior and $u_{12} = 0$.

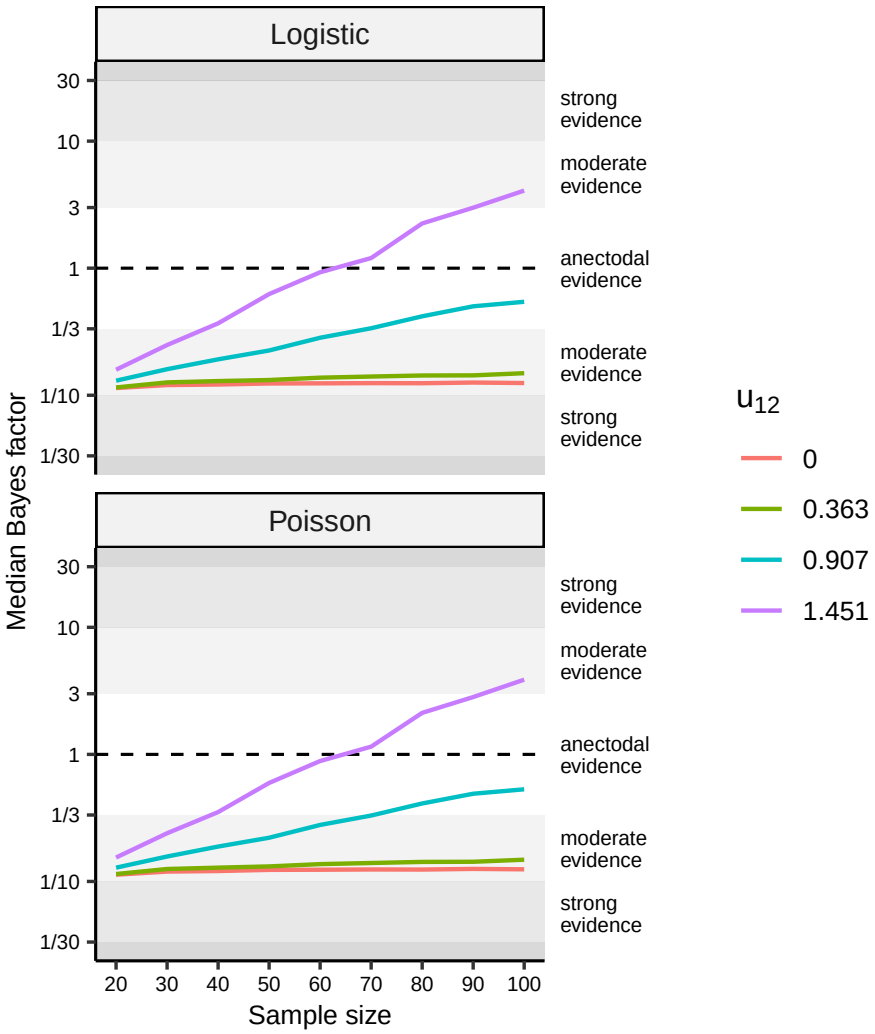


Figure 5: Median Bayes factor using a fixed g-prior ($g = 100$) for the Poisson and logistic regression model comparison for small n

4.2 Simulation results for different prior choices

4.2.1 Fixed g ($g = 100$)

While the focus of the simulation study is on the variants of g-priors which depend on the sample size n, results for the simplest form of the g-prior (a fixed value of g) are shown as a benchmark for the remaining approaches. Using an arbitrary fixed value for g ($g = 100$), Figure 6 shows the median Bayes factor calculated from the individual Bayes factors for each combination of u_{12} and n. Importantly, the y-axis is shown on a log-scale to differentiate the behavior of the Bayes factor for smaller sample sizes. Since any Bayes factor greater than 1 (at least anecdotally) favours the alternative hypothesis, whereas a Bayes factor smaller 1 (at least anecdotally) supports the null hypothesis, the dashed line at BF = 1 highlights if the Bayes factors are favoring the null or the alternative hypothesis.

As seen in Figure 5 and 6, the results of the logistic and the Poisson tests are almost the same. This means that the criterion of model similarity is achieved for the fixed-g case. For both models, independently of the effect size, smaller sample sizes will lead to median Bayes factors smaller than 1, thereby favoring the null hypothesis. For $n = 8$ (where all cell counts are 1) the Bayes factor is equal to 0.1, which indicates that the Bayes factor does not converge towards 1 as the sample size decreases. For data based on $u_{12} = 0$, the median Bayes factor slightly decreases before converging towards a value of approximately 0.125 as n increases. For all simulated data generated with $u_{12} > 0$, there

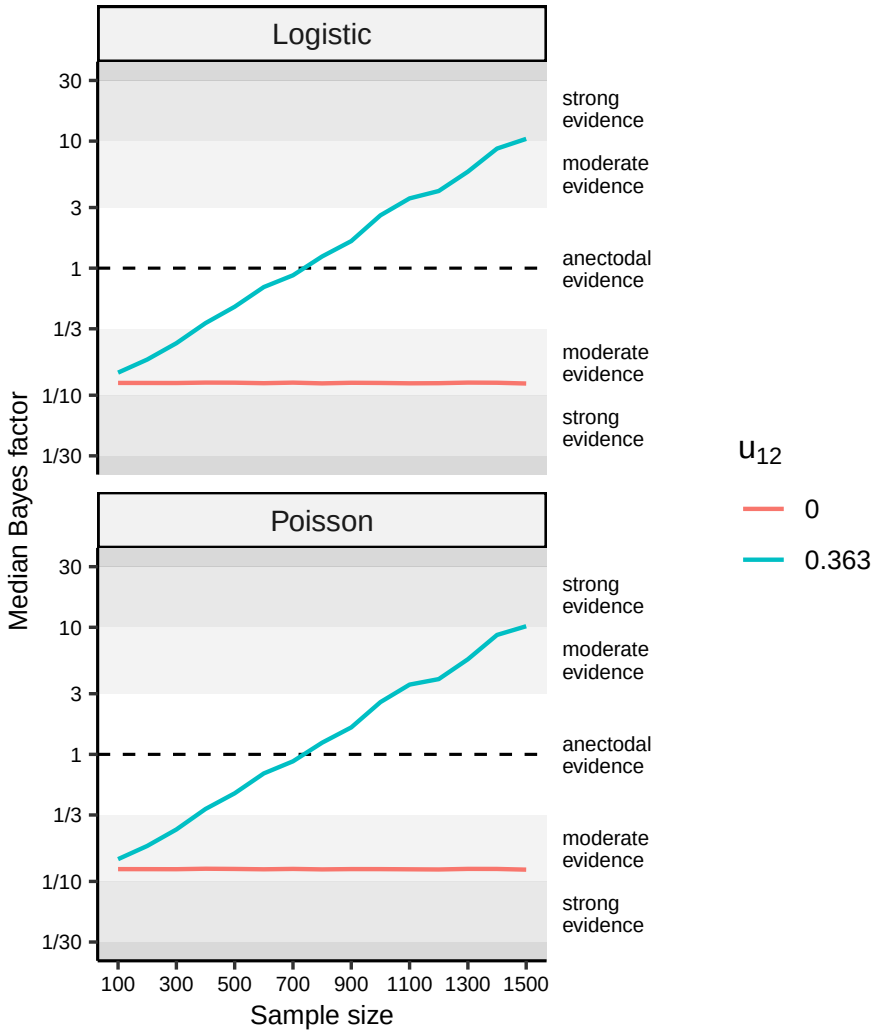


Figure 6: Median Bayes factor using a fixed g-prior ($g = 100$) for the Poisson and logistic regression model comparison for large n

is a positive exponential relationship between the Bayes factors and the sample size n with the strength of the relationship depending on the value of the effect size. For a small effect size ($u_{12} = 0.363$), the median Bayes factor is smaller than 1 for $n \leq 700$, i.e., the average error rate (using BF = 1 as a threshold) was larger than 50 % for data sets where the total count was less than 800.

4.2.2 Unit Information prior

Figure 7 and Figure 8 show the results for the unit information prior ($g = n$) for $n \leq 100$ and $n \geq 100$, respectively. When calculating the Bayes factors for the simulated data, the sample size n (and as such also the value for g) was chosen to be $n = \sum m_{ijk}$ for both the Poisson and the logistic regression tests. Interestingly, the results show again almost identical results for both the log-linear and the logistic testing approaches, which again meets the model similarity criterion mentioned in chapter 3.2. For $n = 8$, the Bayes factor is 1/3 for both versions of the test. While this is closer to 1 compared to the prior variant above, there is a high risk of wrongly finding moderate evidence for the null hypothesis as n decreases towards $n = 8$. With respect to the model selection consistency and sensitivity, the median Bayes factor stays lower than 1 for $n < 50$ for all effect sizes - at which point the median Bayes factor surpasses 1 for the large effect data sets. For small effect size data, the median Bayes factor decreases similarly as the Bayes factor for $u_{12} = 0$ and only surpasses BF = 1 at $n \geq 1100$. In fact, up until $n = 700$, the median Bayes factor for small effect data even shows moderate evidence for the null hypothesis. In case the null hypothesis is true, there is a strong decrease in the median Bayes factor as n increases: Already for $n \geq 200$, the median Bayes factor for $u_{12} = 0$ results in strong evidence for the null hypothesis, while for $n \geq 1500$, the median Bayes factor indicates extreme evidence in favour of the null hypothesis.

4.2.3 Hyper-g/n prior

In this section, the results for the hyper-g/n prior will be discussed. Similar to Figure 6, Figure 9 shows the median Bayes factor using the hyper-g/n prior for the Poisson and the logistic regression model. Both the Poisson and the logistic approach yield median Bayes factors lower than 1 for smaller sample sizes regardless of the effect size. For medium and large effect sizes, the median Bayes factor increases exponentially as the sample size increases. However, for small effect sizes (for both of the models), the median Bayes factor first decreases as n increases. For small effect data, the median Bayes factor only starts increasing with larger sample sizes as can be seen in Figure 10. For $u_{12} = 0$, the median Bayes factor decreases immediately and continuously as n increases and shows a logarithmic pattern.

Comparing the log-linear and the logistic model approaches, it can be noticed that the logistic regression test overall results in lower Bayes factors compared to the Poisson regression model. It takes considerably larger sample sizes until the median Bayes factor starts reliably favoring the alternative hypothesis using the logistic regression model. For simulated data sets of $u_{12} = 0.363$ (small effect size), the threshold where the median Bayes factor is larger than 1 is at $n = 900$ for the logistic regression model and $n = 600$ for the Poisson regression model, respectively. As such, the two modelling approaches are not as similar as the two previous prior variants where no prior is placed on g.

As mentioned in chapter 2.2, implementing the Poisson regression model comparison with the package BAS implies, by default, that the sample size is fixed at $2 \times 2 \times K$. On the other hand, we have used an expanded format of the contingency table for the logistic version, such that every cell count corresponds to one observation ($n = \sum m_{ijk}$). This has the effect that the k parameter (not to be confused with the notation for layers) used in the hyper-g/n prior, defined as $1/n$ (Li and Clyde, 2018), has a different value depending on the definition of the sample size. For this reason, Figure 11 shows another version of the hyper-g/n Poisson regression test, for which k is set to $1/\sum m_{ijk}$ instead.

Figure 11 illustrates that the effect of setting $k = 1/\sum M_{ijk}$ instead of $k = 1/8$ does not change the overall pattern of the Bayes factor for different effect sizes and sample sizes. Therefore, another simulation was conducted to isolate the effect of the parameter k. This was done by generating one fixed contingency table and applying the hyper- g/n prior using

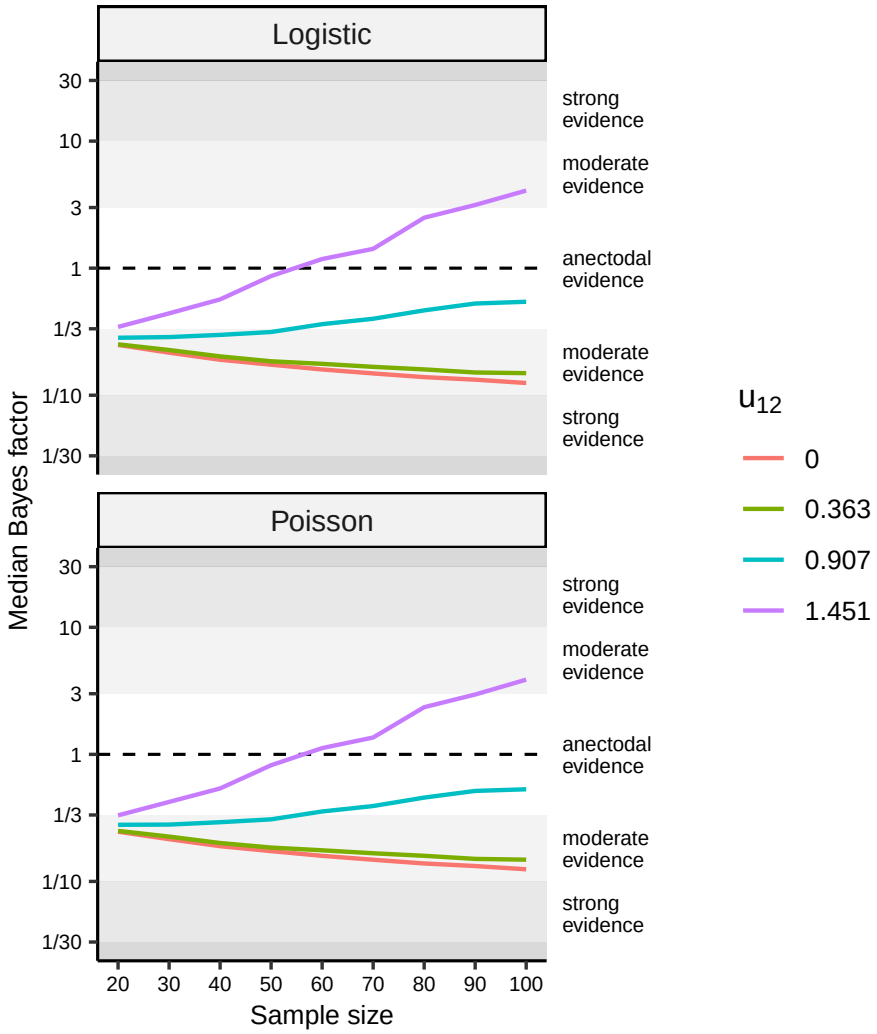


Figure 7: Median Bayes factor using the unit information prior ($g = n = \sum m_{ijk}$) for $n <= 100$

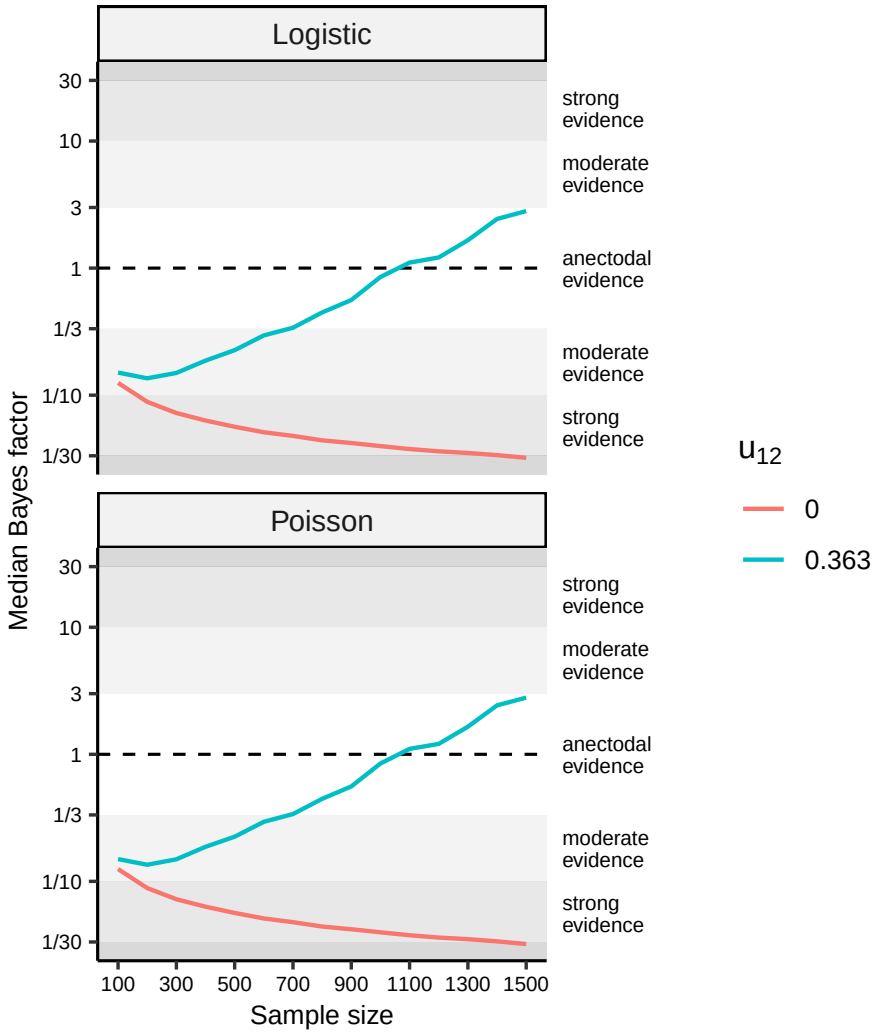


Figure 8: Median Bayes factor using the unit information ($g = n = \sum m_{ijk}$) for $n >= 100$

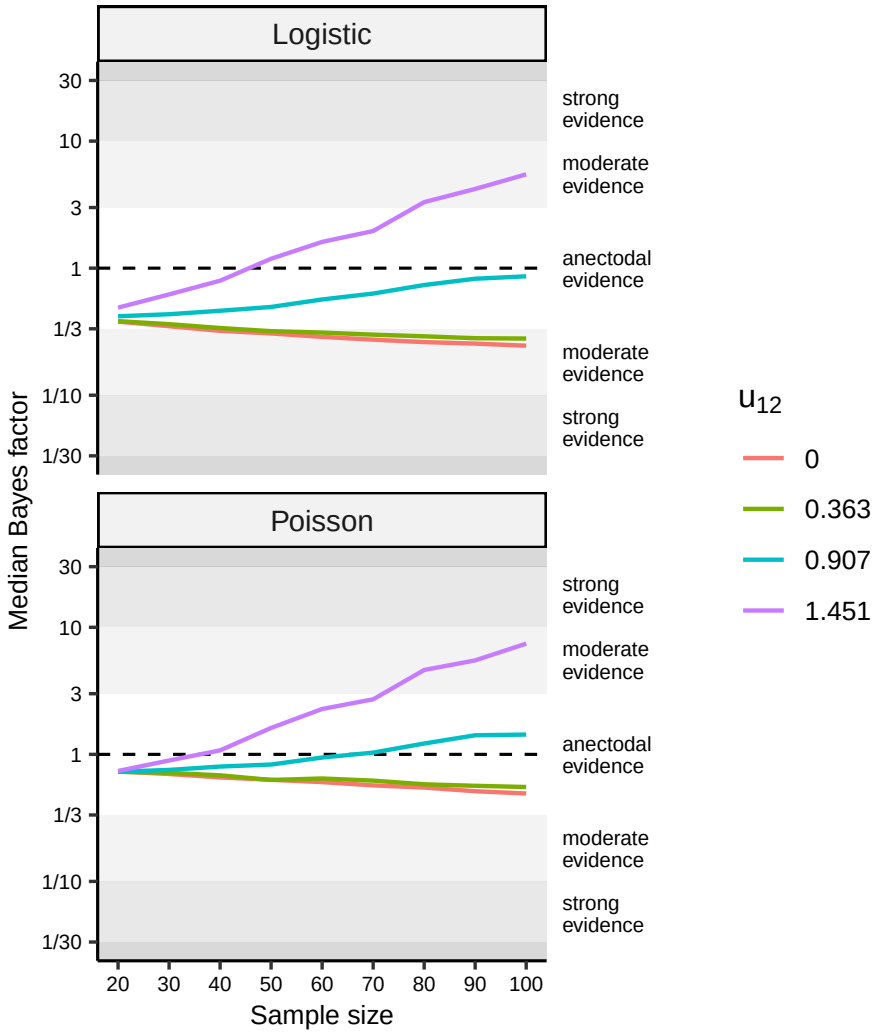


Figure 9: Median Bayes factor using the hyper-g/n prior with the Poisson and logistic regression model for small n

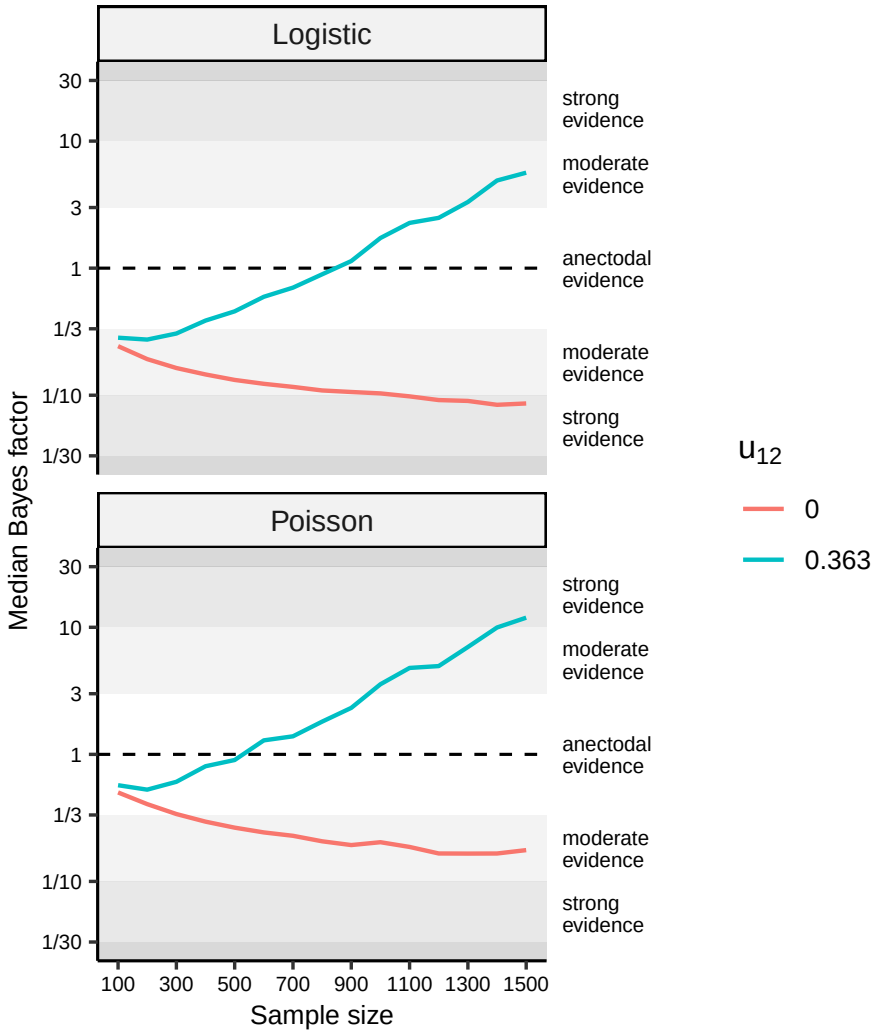


Figure 10: Median Bayes factor using the hyper-g/n prior with the Poisson and logistic regression model for large n

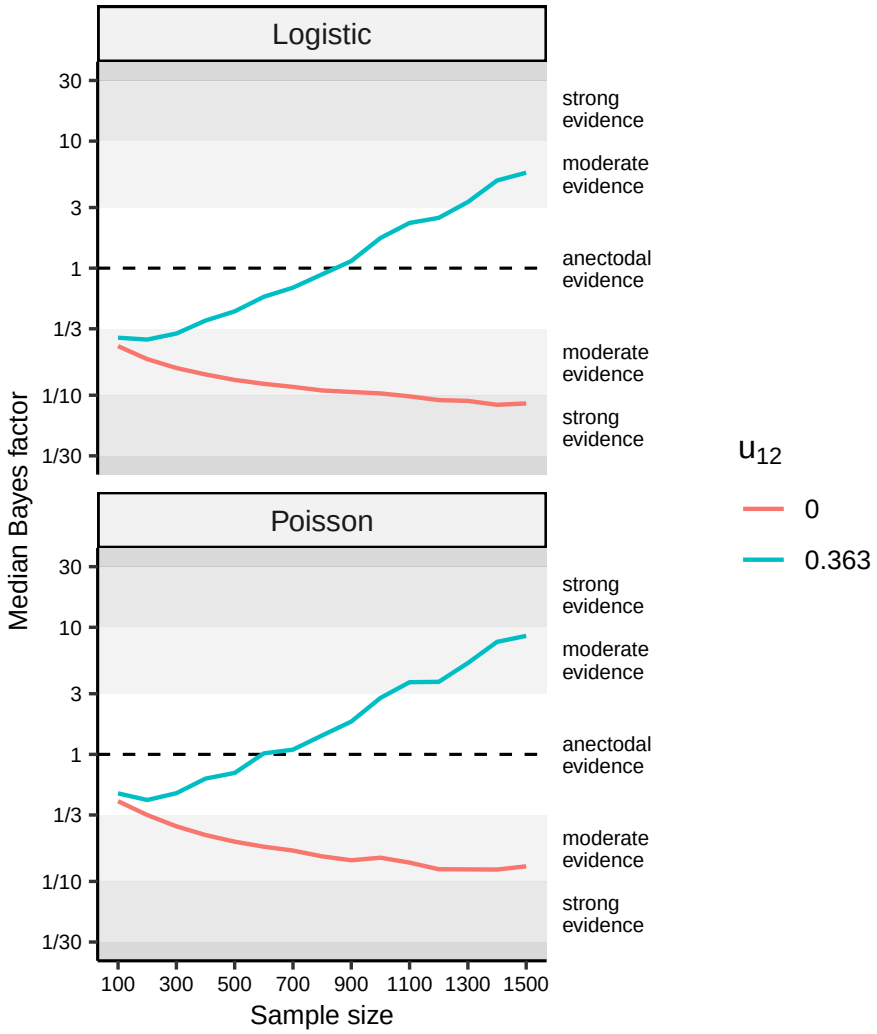


Figure 11: Median Bayes factor using the hyper g/n prior with $k = 1/\sum m_{ijk}$ for the Poisson Bayes factor

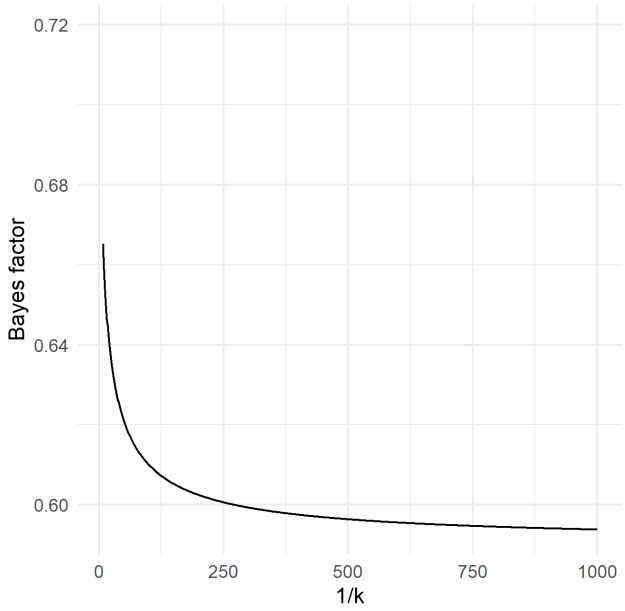


Figure 12: Bayes factor using the hyper-g/n prior as a function of $1/k$ (n) for a fixed data set.

different k-values ranging from 1 to 1/1000. The results for this simulation can be seen in Figure 12.

As Figure 12 indicates, there is a logarithmic relation between n ($1/k$) and the Bayes factor. Even though the impact of changing k is higher for lower values of $1/k$, the overall percentage changes are rather modest. This explains why no overall change in the behavior of the Bayes factor can be noticed in Figure 11. Consequently, the observed differences between the log-linear and the logistic testing approach cannot be explained just by changes in the parameters dependent on the sample size n.

Concerning the test specificity, Table 4 below shows the rate of Bayes factors smaller than 1 per effect size and per sample size for the Poisson test. As Table 4 indicates, the percentage of Bayes factors smaller than 1 based on the hyper-g/n prior for data generated from the null hypothesis increases as the sample size increases. With respect to the model selection error rates, the median Bayes factor incorrectly favours the null hypothesis for $n \leq 500$ for small effect data using the Poisson model comparison and $n \leq 800$ using the logistic test. For the medium effect size, the test achieves a power (when using BF = 1 as threshold) of more than 80 % as $n > 200$ for both the Poisson and the logistic approach. For large effect sizes, this threshold is already achieved at $n \geq 100$ for the Poisson model and at $n \geq 200$ for the logistic model.

As discussed before, due to the restriction of cell counts being larger than 0, the smallest possible $2 \times 2 \times 2$ contingency table has a sample size of 8 with all 1's for each count. The null hypothesis is slightly favoured for this minimally informative data set: The Bayes factor is 0.792 for the Poisson model comparison and 0.447 with the Logit model comparison. As such, specifying $n = 8$ results in only anecdotal evidence for the null hypothesis using either the logistic or the Poisson model. With respect to the requirement that the Bayes factor should converge towards 1 as n decreases, the Poisson model is more favourable compared to the logistic model, whereas both approaches are still more preferable than the previous prior variants discussed above.

4.2.4 Beta-prime prior

For the Beta-prime prior, the comparison of the Poisson and the logistic regression model in the small sample size range ($n \leq 100$) is depicted in Figure 13.

While the hyper-g/n prior shows very similar results in the comparison of the log-linear and the logistic regression models, the Bayes factors for the Beta-prime prior are vastly

	u_{12}			
n	0	0.363	0.907	1.451
50	0.812	0.779	0.57	0.368
70	0.833	0.751	0.49	0.288
100	0.853	0.747	0.406	0.194
200	0.885	0.694	0.208	0.069
500	0.929	0.524	0.044	0.017
1000	0.946	0.299	0.006	0.001
1500	0.956	0.194	0.000	0.000

Table 4: Proportion of Bayes factors < 1 for different sample sizes using the Poisson regression hyper-g/n Bayes factor

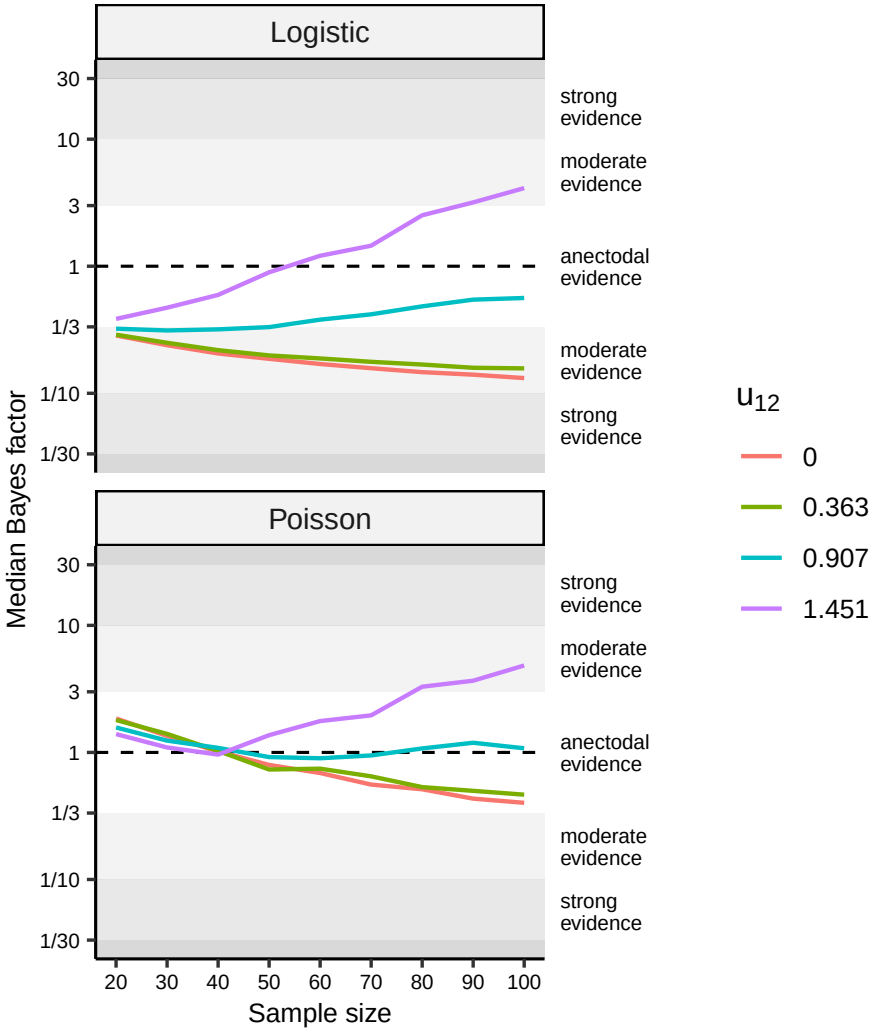


Figure 13: Median Bayes factor using the Beta-prime prior with the Poisson and logistic regression model for small sample sizes.

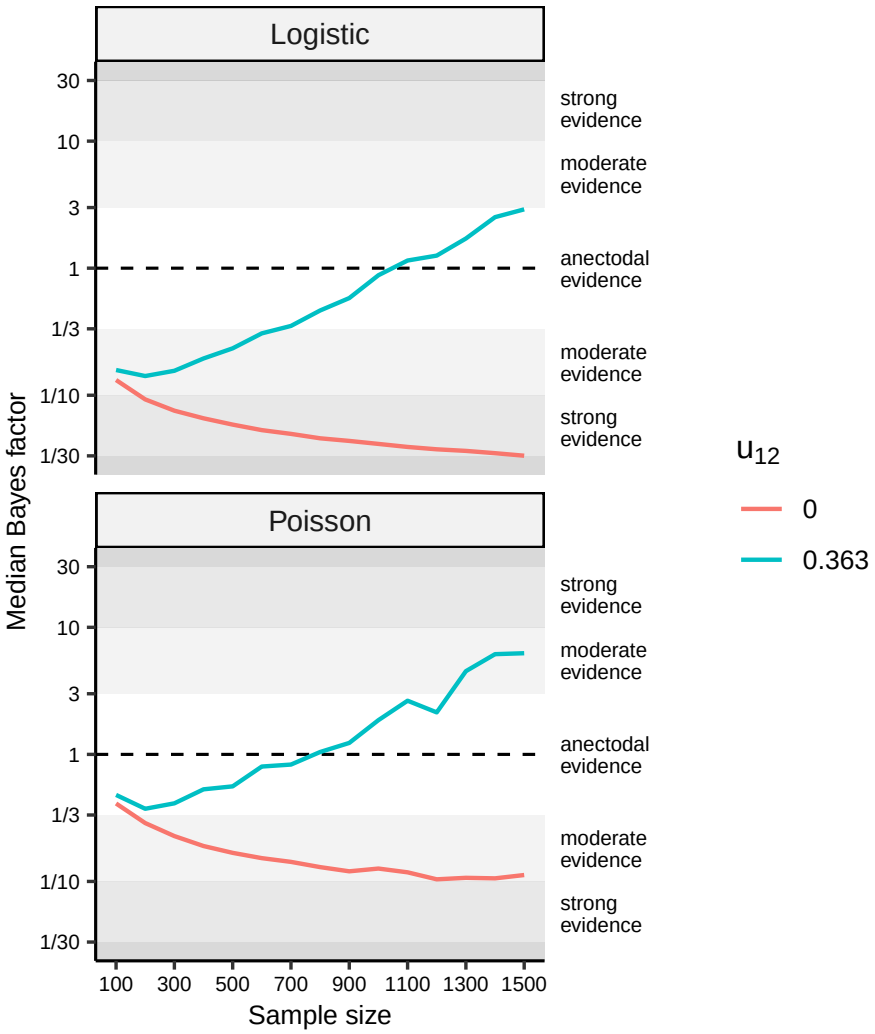


Figure 14: Median Bayes factor using the Beta-prime prior using the Poisson and logistic regression model with large n

different for both modelling approaches: For the Poisson Bayes factors, the median Bayes factors for all effect sizes initially favour the alternative hypothesis, whereas the Logit Bayes factors anecdotally favour the null hypothesis for small sample sizes.

For a contingency table with all cells counts of 1, the Logit Bayes factor is 0.492 while the log-linear Bayes factor is 2.81. Not only does this indicate that the model similarity is not met, but also that the Bayes factor does not converge towards 1 as the sample size decreases. Furthermore, for all specifications of u_{12}, the Poisson Bayes factors initially decrease in size as n increases. Figure 13 indicates that for medium and large effect sizes, the Bayes factor stops decreasing at around $n = 50$. Figure 14 depicts the Bayes factor behavior using the Beta-prime prior for larger sample sizes. For $u_{12} = 0$, the Bayes factor decreases logarithmically, while for small effect sizes, the median Bayes factor again initially decreases (until approximately $n = 300$ for both modelling approaches) and only starts surpassing BF $= 1$ at approximately $n = 800$ for the Poisson model and $n = 1100$ using the logistic model. When the null hypothesis is true, at already n= 50, less than 5 percent of all samples favour the alternative hypothesis (with the threshold BF=1) using the logistic regression approach. For the Poisson regression approach, however, this type I error rate of 5 percent is only reached at $n = 400$.

4.2.5 Robust prior

For the robust prior, the comparison of Bayes factors for the log-linear and the logistic regression tests for small sample sizes are shown in Figure 15. For sample sizes smaller than 40, the median Bayes factor for all effect sizes is slightly smaller than 1 in the case of the logistic test, while the Poisson regression is very close to BF = 1 for all effect sizes. While the difference between the two models is not as severe compared to the Beta-prime prior, the model similarity (and convergence towards BF = 1 with decreasing n) is comparable to the results of the hyper-g/n prior.

For larger n, the general pattern of the robust prior becomes very similar to the results of the hyper-g/n prior, as can be seen in Figure 16. Similar to the previous priors, the median Bayes factor for small effect sizes at first decreases as n increases. With $n \geq 600$, the median Bayes factor based on small effect data starts favouring the alternative hypothesis (using the threshold BF = 1) using the Poisson test. Using the logistic test, this threshold lies at $n = 1000$.

The Bayes factor based on small effect data favors the alternative hypothesis in 79 % of the cases at $n = 1500$ using the Poisson approach. However, using the logistic regression approach, only 67 % of the resulting Bayes factors were favoring the alternative hypothesis at $n = 1500$. Furthermore, the median Bayes factor decreases continuously as n increases for the logistic and the Poisson model comparison when the null hypothesis is true. With $n \geq 1100$, less than 5% of the Bayes factors incorrectly favour the alternative hypothesis for data generated where $u_{12} = 0$ (using the threshold BF = 1). For the logistic model comparison, this threshold is reached with $n \geq 90$.

4.2.6 Intrinsic Prior

Figure 17 shows the median Bayes factors using the intrinsic prior for small sample sizes while Figure 18 illustrates the effect for $n \geq 100$. Similar to the robust prior, Figure 17 demonstrates that for all effect sizes, the median Bayes factor using the Poisson model comparison is slightly closer to 1 compared to the Bayes factors of the logistic regression approach. For the contingency table with a cell count of 1 in each cell, the logistic regression Bayes factor is 0.49 and the log-linear regression Bayes factor is 0.8. Again, the main difference between the Poisson and the logistic regression test is that the logistic test is considerably more conservative compared to the Poisson test. Similar to the robust prior, with $n = 1500$, the Bayes factor based on small effect data favours the alternative hypothesis in 79 % of the cases using the Poisson approach. For the logistic regression approach, the simulation results show that 70% of the Bayes factors favour the alternative hypothesis at $n = 1500$ for small effect data. Using the Poisson model comparison, the model selection error rate (using BF = 1 as threshold) for data generated under the null is smaller than 5 percent when $n \geq 1100$. For the logistic model comparison this threshold lies at $n = 200$.

4.3 Summary of results

To provide an overview of the simulations results, Table 5 summarizes the main findings for each prior variant. Strictly speaking, one could argue that the model similarity is only met for the prior variants with a fixed value of g - especially when using the frequentist results as a benchmark for the model similarity. However, since this criterion is more violated for the Beta-prime prior, whereas the other prior variants generally favour the same hypothesis for both modelling approaches, the hyper-g/n, intrinsic and robust priors were said to meet this criterion as well. Similar reasoning was adopted to define which of the priors result in Bayes factors converging to 1 as n decreases. For this criterion, it should also be noted that the Poisson Bayes factors are more favourable than ones resulting from the logistic test. For the type I error rate mentioned in the summary table, BF > 1 was used as the threshold. As the log-linear model comparison overall yields more favourable results compared to the logistic model comparison, the results for the type I error rate and the minimal n where BF > 1 (at $u_{12} = 0.363$) are shown for the log-linear model only. In chapter 6, a more detailed interpretation of the simulation findings together with recommendations for research practitioners will be provided.

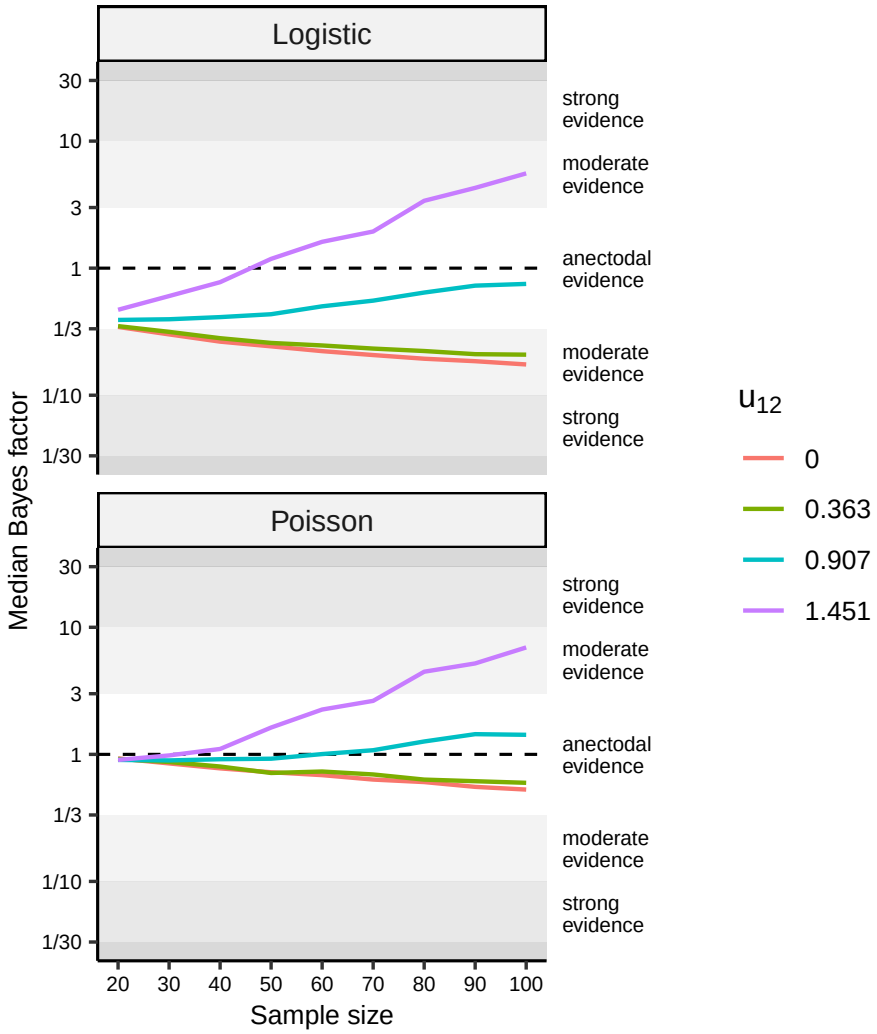


Figure 15: Median Bayes factor using the robust prior (logistic and Poisson model) for small sample sizes

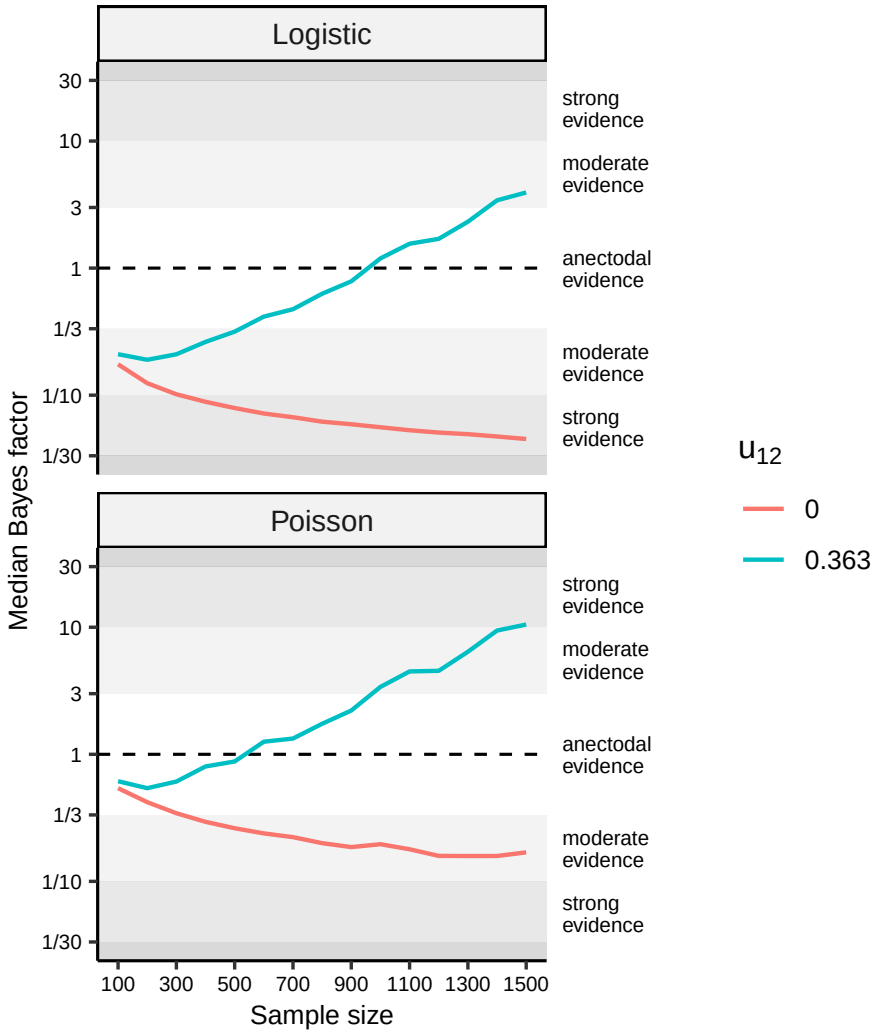


Figure 16: Median Bayes factor using the robust prior ($n \geq 100$)

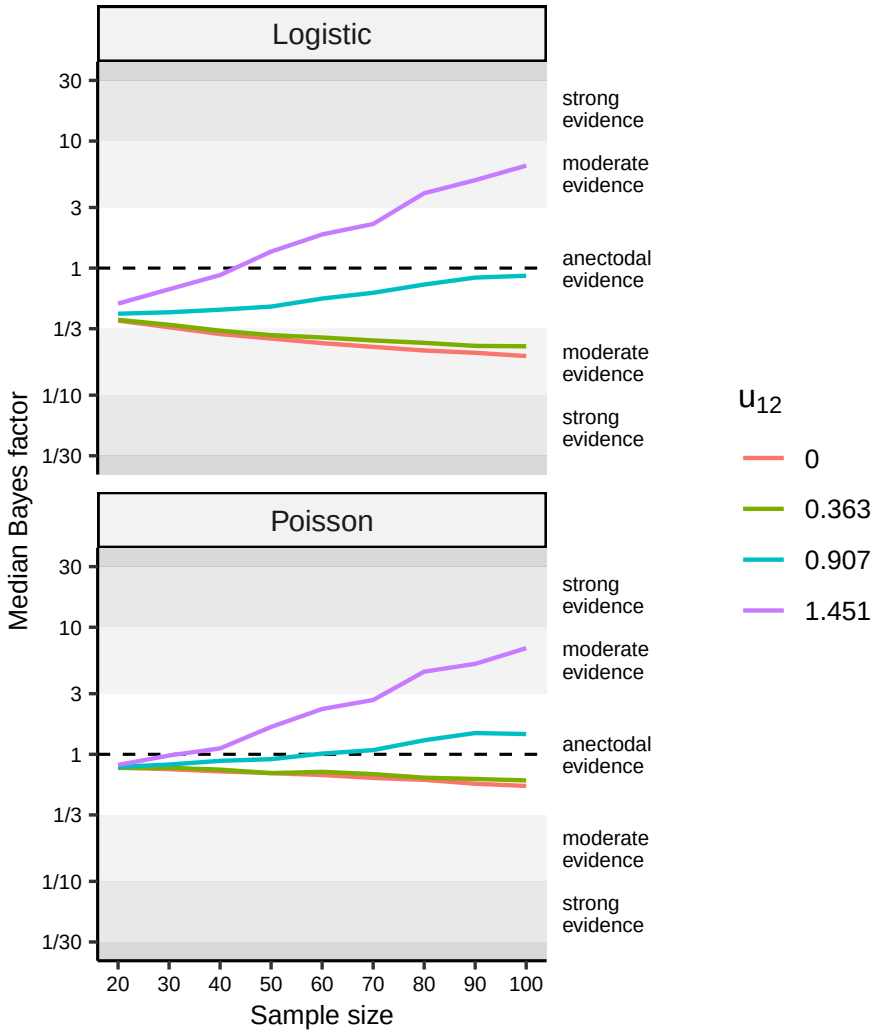


Figure 17: Median Bayes factor using the intrinsic prior (logistic and Poisson model)

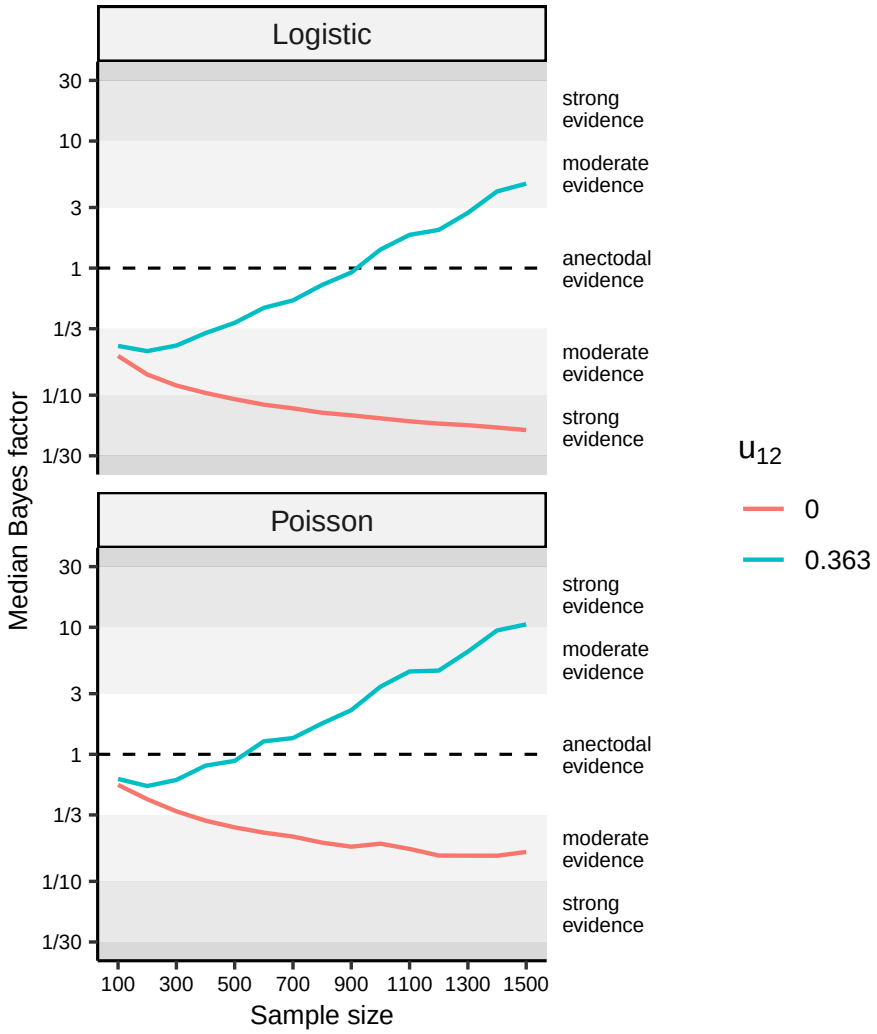


Figure 18: Median Bayes factor using the intrinsic prior (logistic and Poisson model)

	$g = 100$	$g = n$	Beta-prime	Hyper-g/n	Intrinsic	Robust
Impact of nuisance parameters	Yes	Yes	Yes	Yes	Yes	Yes
Convergence to BF $= 1$ as $n \to 0$	No	Yes	No	Yes	Yes	Yes
Convergence to BF $= 0$ as $n \to \infty$ ($u_{12} = 0$)	No	Yes	Yes	Yes	Yes	Yes
Poisson & logistic similarity	Yes	Yes	No	Yes	Yes	Yes
Type I error rate at $n = 1000$ for $u_{12} = 0$	0.031	0.008	0.040	0.054	0.052	0.052
min n where median BF ≥ 1 ($u_{12} = 0.363$)	800	1100	800	600	600	600

Table 5: Summary of simulation results per prior variant

5 Computational modification (hyper-g/n BF)

5.1 Computational limitations of the hyper-g/n prior

Computing the Bayes factors for the simulated data sets using the hyper-g/n prior of the Poisson model comparison approach, it is apparent that the R function of the BAS (Clyde, 2015) package used to calculate the Bayes factor returns NA (Not Available) in some of the cases. Figure 19 shows the proportion of these NA values for the hyper-g/n prior for different effect sizes and sample size values. It is to be noted that similar computational issues can happen to other CHIC priors as well (and the below mentioned solution would be the same), but for the purpose of demonstration, we will focus on the hyper g/n prior in this section.

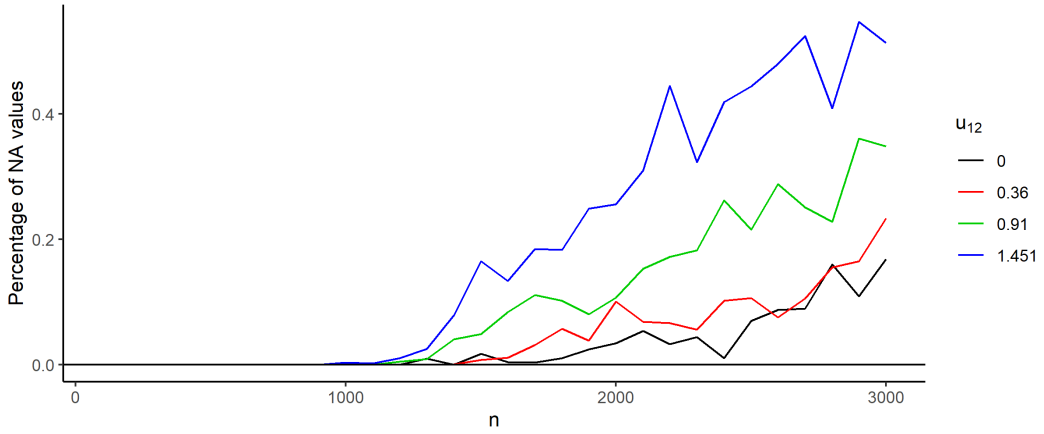


Figure 19: Proportion of data sets where no Bayes factor could be returned using the standard functions of the BAS package

As Figure 19 shows, there are almost no computational issues with smaller sample sizes, whereas for $n > 1000$ an increasing number of simulated data sets do not return any Bayes factor. For a large effect size and sample size of more than 3000, more than 40 percent of the resulting Bayes factors could not be computed. By breaking down the individual components of the hyper-g/n prior, we can see that the computational issues are caused by the Humbert series ϕ_1 (Humbert, 1920) defined as

$$\phi_1(a, b, c, x, y) = \frac{\Gamma(c)}{\Gamma(c-a)} \int_0^1 t^{a-1}(1-t)^{c-a-1}(1-xt)^{-b} e^{yt} \mathrm{d}t. \tag{31}$$

For the hyper-g/n prior, y is defined as $\frac{s+Q_M}{2v}$ with s=0 and v=1 and Q_M being the

Wald statistic of Model M. Consequently, if the Wald statistic of either $\mathcal{M}_4$ or $\mathcal{M}_7$ is large, implementing $e^{\frac{Q_M}{2}}$ in the formula above can quickly become infeasible using the default Double-precision floating-point format (Becker, 2018). In fact, since the maximum absolute value that can be evaluated in R is roughly 2×10^{307} (R Core Team, 2019), any Bayes factor that is based on a Wald Statistics $Q_M > \log(2 \times 10^{307})$ cannot be computed in R using the standard approach outlined above. To allow for a wider range of feasible Bayes factor computations, two alterations were done to the Humbert Series ϕ_1, which do not change the resulting Bayes factor but avoid the computational limitations outlined above.

5.2 Technical alternative

Below, we demonstrate how the Bayes factor function using the hyper-g/n prior is being altered. These steps can also be reviewed and tested through the R function[2] written as part of this thesis that implements the modified version of the function. In line with section 1.3.1, the two competing models in the Bayes factor equation will be referred to as $\mathcal{M}_7$ and $\mathcal{M}_4$, respectively.

5.2.1 Modification 1

The first change takes advantage of the product rule for exponents which allows us to split up the exponent:

$$e^x = \prod_i^I e^{\frac{x}{I}}. \tag{32}$$

In the Humbert Series defined in equation 31 above, it is possible to split up e^{yt} into $e^{\frac{yt}{i}} \times e^{\frac{yt}{i}} \times e^{\frac{yt}{i}} \times .... \times e^{\frac{yt}{i}}$ where i should be chosen such that is is numerically feasible to compute $e^{\frac{yt}{i}}$. For example, it is not feasible to compute e^{800} in R, whereas the individual components $e^{400} \times e^{400}$ (the product of which is e^{800}) can be computed in R. While this split allows us to compute the individual components without affecting the overall result of the Humbert Series, this change alone will of course not provide any computational relief, as the product of the split-up components would still potentially yield numbers too large to be computed. For this reason, a second modification needs to be implemented, which in combination with the first modification can overcome the limitations of large exponents in the hyper-g/n Bayes factor.

5.2.2 Modification 2

The second change is based on the fact that the Humbert series is computed for both competing models in the Bayes factor equation, and only the ratio of them is used in the formula for the Bayes factor. This allows us to multiply the Humbert Series with an arbitrarily small constant, as this constant will cancel out and not affect the Bayes factor. However, it is not sufficient to merely multiply e^{yt} with a small constant because (1) R would evaluate the exponent and the small constant separately, which would still yield NA values as output and (2), the small constant needed might also not be feasible to compute in R (yielding 0). For this reason, each $e^{\frac{yt}{i}}$ component from the first modification was multiplied with a small constant. It is important that this constant is the same for the Humbert Series of both models in the Bayes factor equation to ensure that they cancel each other out and do not affect the overall Bayes factor.

5.2.3 Implementing the two modifications

Putting both changes together, we are left with the following version of the Humbert Series

$$\phi_{1mod}(a,b,c,x,y) = \frac{\Gamma(c)}{\Gamma(c-a)} \int_0^1 t^{a-1}(1-t)^{c-a-1}(1-xt)^{-b} \prod_i^I S \times e^{\frac{y}{I}} \mathrm{d}t \tag{33}$$

where I is the factor indicating how many times we split and divide e^{yt} and S is the constant by which we multiply each of the components in the exponent. To implement both modifications above, appropriate values for S and I need to be chosen. In the modified R function, S and I need to be chosen dynamically according to the value of the Wald Statistic in model 7 and model 4. The reason why S and I need to be chosen dynamically (and not set to any small arbitrary number) is that the function might return 0 in case the Wald Statistics are small. In the new R function, which is implementing the modified version of the Humbert Series, the following steps are implemented:

1. At first we need to define a maximum value (MV) for which we know that e^{MV} can be computed without any issues in R.

[2]GitHub link: github.com/DHeemann/Bayesian-conditional-independence-simulation-

2. Next, we find MY as the maximum of $\frac{s+Q_{\mathcal{M}_7}}{2v}$ and $\frac{s+Q_{\mathcal{M}_4}}{2v}$.

3. I (how many times we split up the exponent) is then found by rounding up the ratio of MY and MV $\lceil \frac{MY}{MV} \rceil$.

4. S is then computed as

$$\frac{1}{e^{max(0, \frac{MY-MV}{I})}}. \tag{34}$$

With this approach, if both $e^{\frac{s+Q_{\mathcal{M}_7}}{2v}}$ and $e^{\frac{s+Q_{\mathcal{M}_4}}{2v}}$ will be smaller than e^{MV}, then S and I will both be equal to 1. Therefore, the original version of the Humbert Series (as used in the BAS package) will be used if the Wald Statistic of either $\mathcal{M}_7$ or $\mathcal{M}_4$ are not too large. Testing the modified R function for data sets where the original hyper-g/n BF implementation from the BAS package is able to compute the Bayes factor, the results using the modified function in R are *practically* identical (with small differences likely due to the "integrate()" function used in R).

5.2.4 Example of the modified hyper-g/n BF function

As an example, Table 6 below shows one of the simulated data set for which the original BAS package would not be able to compute the Bayes factor.

Layer	Row	Column	
		1	2
1	1	67	136
	2	199	414
2	1	172	169
	2	865	778

Table 6: An example of a $2 \times 2 \times 2$ contingency table

Implementing the log-linear representation of the CMH test, the Wald statistic for model 7 and model 4 are $Q_{\mathcal{M}_7} = 1481.2$ and $Q_{\mathcal{M}_4} = 1483$, respectively. In the Humbert Series, this results in $y_7 = 740.6$ and $y_4 = 741.4$ and both $e^{740.6}$ and $e^{741.4}$ cannot be computed on most machines using Standard R (and will thus not return any Bayes factor using the BAS package).

Following the modifications outlined above, with MV (the maximum value) set to 400, I is equal to $\lceil \frac{741.4}{400} \rceil = 2$, i.e., $e^{741.4}$ will be represented as $e^{370.7} \times e^{370.7}$ according to the first modification. Secondly, we calculate $S = \frac{1}{e^{max(0, \frac{741.4-400}{2})}} = 7.03 \times e^{-75}$. Taking both steps together, we replace $e^{741.4t}$ in the Humbert Series with $7.03e^{-75}e^{370.7t} \times 7.03e^{-75}e^{370.7t}$. The final Bayes factor of the example data set in Table 6 equals 0.07.

Figure 20 below shows the Bayes factor results of 100 simulated contingency tables per specified sample size for $u_{12} = 1.451$. The red points indicate those Bayes factors which could not be computed using the BAS package for the simulated data. With the modified approach (setting MV to 300), all of the simulated data sets, which could not be computed using the BAS package, were feasible. Noteworthy, while the chance of infeasible Bayes factor calculations increases with the sample size, the magnitude of the resulting Bayes factors does not seem to have any effect on the feasibility of the calculation.

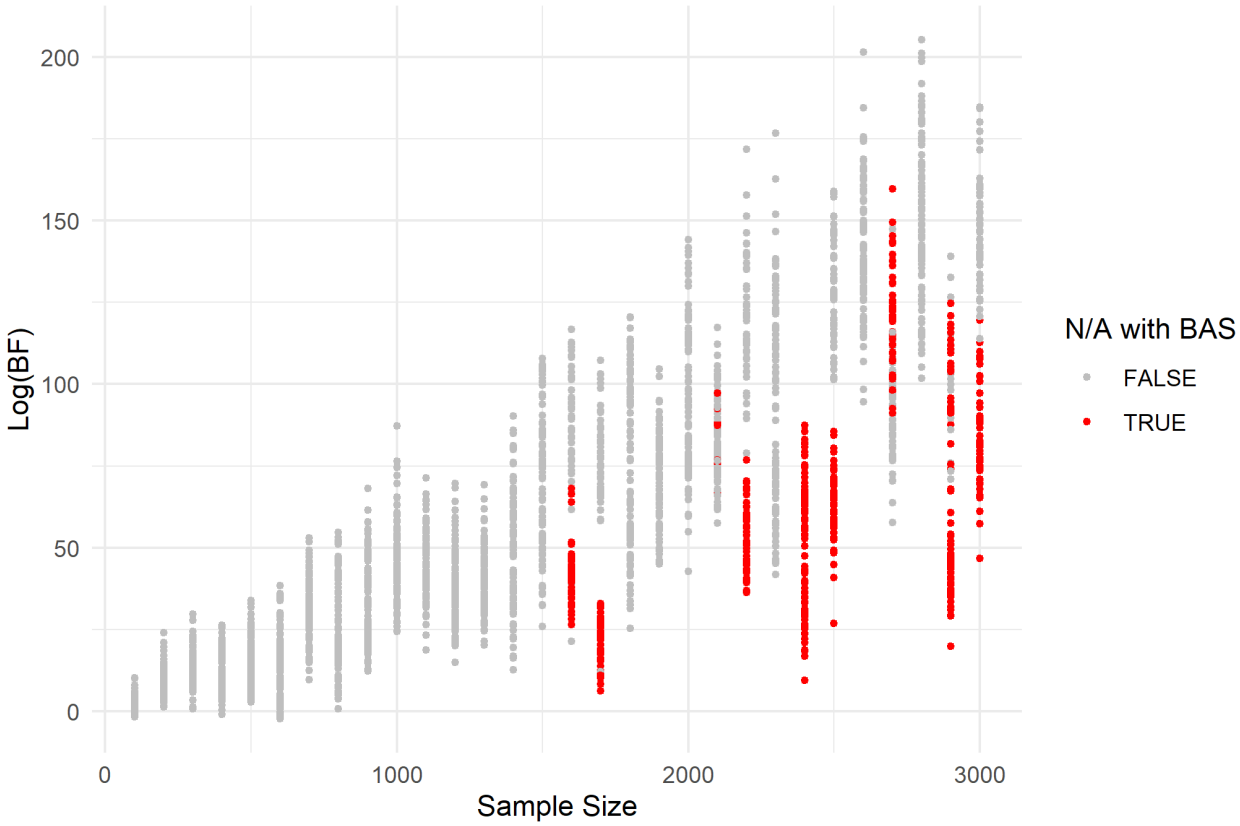


Figure 20: Bayes factors for 100 samples per sample size from 100 to 3000 ($u_{12} = 1.451$) using the modified hyper-g/n Bayes factor function. The red points refer to Bayes factor for which the original function in the BAS package could not return any value.

6 Discussion and recommendations

6.1 General suitability of the g-prior based conditional independence test

For all priors assessed in this thesis, the Bayes factor tests do not completely isolate the test-relevant parameter u_{12}, but instead seem to be affected by the values of the nuisance parameters. While this is not the main criterion to assess the practical relevance of the Bayesian tests, this suggests that either (1) a different testing approach than the default g-priors for GLMs might be needed or (2) that the simulation process might still introduce a bias for the test of independence.

Furthermore, one of the limitations of the Bayesian conditional independence test using the g-priors discussed in this thesis is that - on average - the Bayes factors initially decrease for small effect sizes as n increases. In practical applications where the researcher continuously collects evidence and monitors the Bayes factors, such high chances of observing increasing evidence for the null hypothesis (when in fact there is a small effect) may mislead the researcher to have more confidence in the absence of any effect. For these reasons, future research could focus on developing new priors to address the above-mentioned issues when testing conditional independence in contingency tables. Furthermore, as this simulation study was restricted to $2 \times 2 \times 2$ contingency tables, future simulation studies could extend the simulation to any $I \times J \times K$ design. In addition, future research may also focus on different possible variations of hypotheses for three-dimensional contingency tables. For instance, the Breslow-Day test assesses whether the relationship between two variables is merely the same for each stratum in a contingency table (Breslow, 1980). In this case, one would test if the conditional odds ratio is the same for each k (but not necessarily 1). Regardless, these potential limitations/restrictions should be kept in mind in the following discussion.

6.2 Comparisons of the g-prior variants

Of the g-prior cases used in the simulation study, both a fixed value for g ($g = 100$) and setting $g = n$ are the only variants for which the log-linear and logistic test approaches yield almost identical results. However, our simulation studies show that both tests are not recommended to be used for the conditional independence test in $2 \times 2 \times K$ tables. The obvious limitation of fixing g to an arbitrary value is that the test converges to a constant for data generated under the null hypothesis. In the case of fixing $g = 100$ for the data at hand, it would consequently be impossible to find strong evidence in favour of the null hypothesis (no matter the size of the sample). Compared to the remaining prior variants, the simulation results show that the unit information prior has a much stronger bias towards the null hypothesis for the conditional independence test in contingency tables, which excludes this prior from any further discussion.

For the remaining priors which place a prior distribution on g itself, the differences between the logistic and the log-linear model comparisons can be easily observed. Part of this difference is that the default definition of the sample size differs for both testing approaches. Across all the prior variants, the logistic model comparison favours the null hypothesis systematically more and sometimes results even in opposite conclusions compared to the log-linear model approach. While the definition of the sample size does play a role, modifying the relevant parameter in the hyper-g/n prior indicates that this only plays a small role in the difference of both approaches. Among the assessed prior variants, the one with the highest observed difference between the log-linear and the logistic tests is the Beta-prime prior: for very small sample sizes, the Bayes factor for the log-linear version is higher than 1 for all effect sizes specifications while the logistic version is lower than 1 for all effect size specifications. Compared to the remaining priors, it requires larger sample sizes to detect evidence in favour of the alternative hypothesis for data generated with small effect sizes. Taking these points together, the Beta-prime prior is consequently also not recommended for the conditional independence test in $2 \times 2 \times K$ contingency tables.

When comparing the intrinsic, robust and hyper-g/n priors, the simulation results suggest that there are no big differences between all three approaches for the conditional independence test. With respect to the criteria used to assess the Bayes factor behavior (section

3.2), all three of them perform overall better compared to the other studied prior variants when testing conditional independence in $2 \times 2 \times 2$ contingency tables: For non-informative data, the average Bayes factor is close to 1 and the median Bayes factor based on medium to large effect data stays above 1 and grows rapidly as n increases. In contrast to the fixed g case, the Bayes factors continue to decrease as n increases, allowing researchers to find evidence of the null hypothesis. Furthermore, even though the logistic regression test is more conservative compared to the log-linear version, both tests overall show very similar patterns. For these reasons, we will focus on these three prior variants for the remainder of the discussion. In the next section, some practical recommendations are provided and the (best performing) g-prior cases are compared to the frequentist tests.

6.3 Practical recommendations

By definition, the frequentist case has a constant type I error rate at 0.05 for a rejection threshold at p = 0.05 irrespective of the sample size. Compared to that, the simulation outcomes show that choosing BF = 1 as a decision threshold to accept/reject hypotheses, the resulting type I errors are (unless the sample size is big) higher compared to the type I error rates for the frequentist test. This is in line with the common recommendations for the interpretation of Bayes factors (van Doorn et al., 2021), suggesting that Bayes factors between BF = $\frac{1}{3}$ and BF = 3 should not be interpreted as strong evidence for either hypothesis, but rather as a first indication requiring more data in order to draw any conclusions. Nonetheless, it is noteworthy that for $2 \times 2 \times 2$ contingency tables with $n > 1000$, the probability of falsely rejecting the null hypothesis in favour of the alternative hypothesis is below 5% even when BF = 1 is chosen as the threshold. At $n = 1500$, the type I error is below 0.04 for all three cases.

Which, if any, of the two modelling approaches (logistic or log-linear) should be chosen depends both on the available sample size and the costs of type I and type II errors in the respective research scenario. In research situations where the main goal is to avoid incorrectly concluding that the null hypothesis is false, the Bayes factor test using the hyper-g/n prior can be used with either the logistic regression approach or the log-linear test (when n is large). On the other hand, in situations where the costs of incorrectly finding evidence for the null hypothesis (or missing evidence for the alternative hypothesis) also plays a role, the logistic regression approach is not recommended. In this scenario, even the log-linear test requires a considerably large sample size ($n \geq 1500$) in order to reduce the risk of incorrectly finding moderate evidence (or more) for the null hypothesis in the presence of a small underlying effect size to less than 5%.Irrespective of the general approach (frequentist or Bayesian), small effect data require a very large sample size in order to detect the underlying effects reliably.

Overall, the simulation results show that the Bayesian version using the log-linear model comparison based on the hyper-g/n (or robust/intrinsic) prior to test the conditional independence in $2 \times 2 \times 2$ contingency tables is a valuable alternative compared to the frequentist test for large n. However, especially for small to medium sample sizes ($n \leq 1000$), more research is needed to construct reliable Bayes factor tests for conditional independence problems in contingency tables.

7 Appendix

7.1 Frequentist results

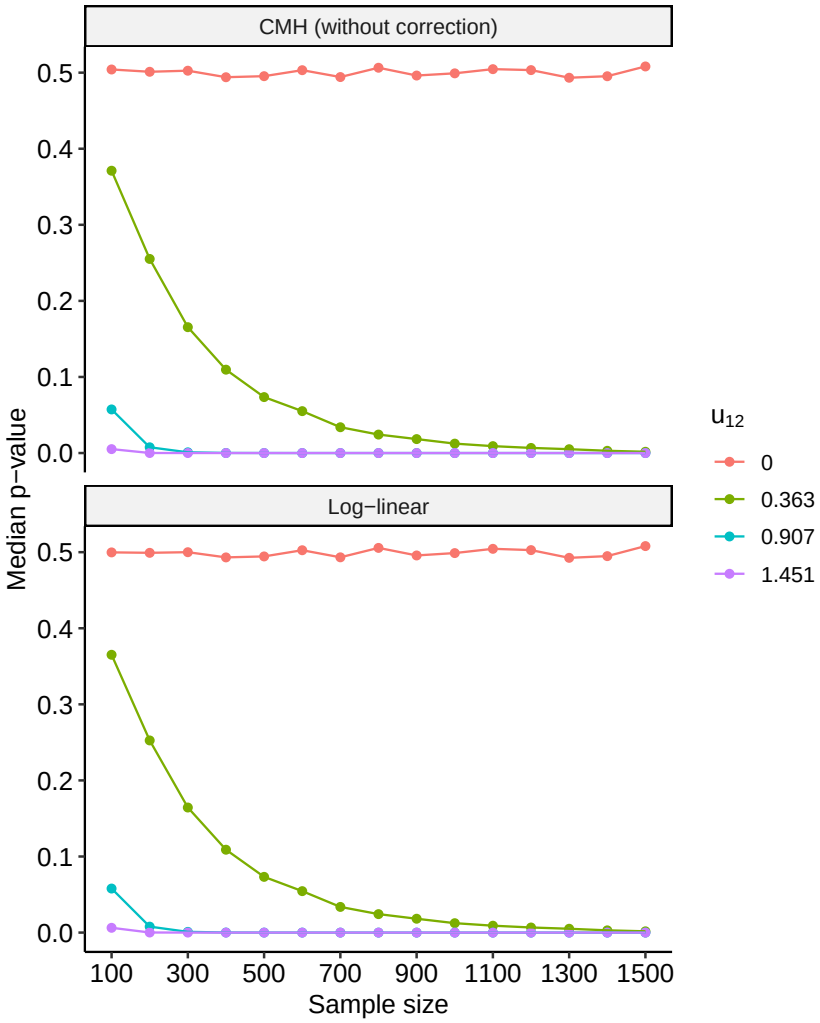


Figure 21: Median p-value of the CMH test (without correction) and the log-linear model test as a function of the sample size different effect size values.

		CMH Test	
		Retain	Reject
Log-Linear Test	Retain	94.9%	0%
	Reject	0.8%	4.2%

Table 7: Agreement (rejecting/retaining the null) between the CMH test and the log-linear test for $u_{12} = 0$ using p = 0.05 as threshold. The percentage points represent the percentage of the total number of tests.

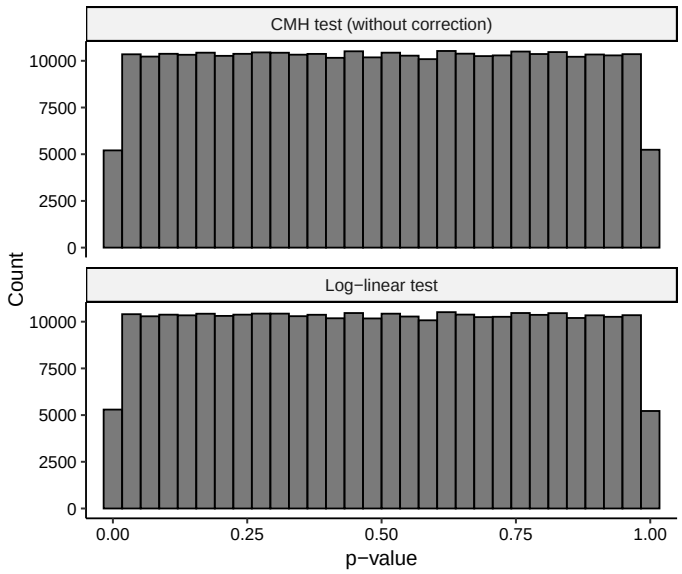


Figure 22: Histogram of p-values for data generated from the null hypothesis ($u_{12} = 0$) for the CMH test and the log-linear test

Table 8: Proportion of rejected null hypotheses for $u_{12} = 0$

Sample size	100	200	500	1000	1500	2000	3000
Log-linear test	0.051	0.053	0.053	0.050	0.052	0.047	0.051
CMH test	0.046	0.051	0.052	0.050	0.052	0.047	0.051

Table 9: Proportion of rejected null hypotheses for $u_{12} = 0.363$

Sample size	100	200	500	1000	1500	2000	3000
Log-linear test	0.118	0.205	0.434	0.705	0.878	0.938	0.985
CMH test	0.114	0.202	0.434	0.705	0.878	0.938	0.985

Table 10: Proportion of rejected null hypotheses for $u_{12} = 0.907$

Sample size	100	200	500	1000	1500	2000	3000
Log-linear test	0.474	0.747	0.965	0.998	1.000	1.000	1.000
CMH test	0.476	0.479	0.968	0.998	1.000	1.000	1.000

Table 11: Proportion of rejected null hypotheses for $u_{12} = 1.451$

Sample size	100	200	500	1000	1500	2000	3000
Log-linear test	0.761	0.933	0.988	1.000	1.000	1.000	1.000
CMH test	0.779	0.944	0.992	1.000	1.000	1.000	1.000

7.2 g-prior comparisons

7.2.1 Results for negative effect size values

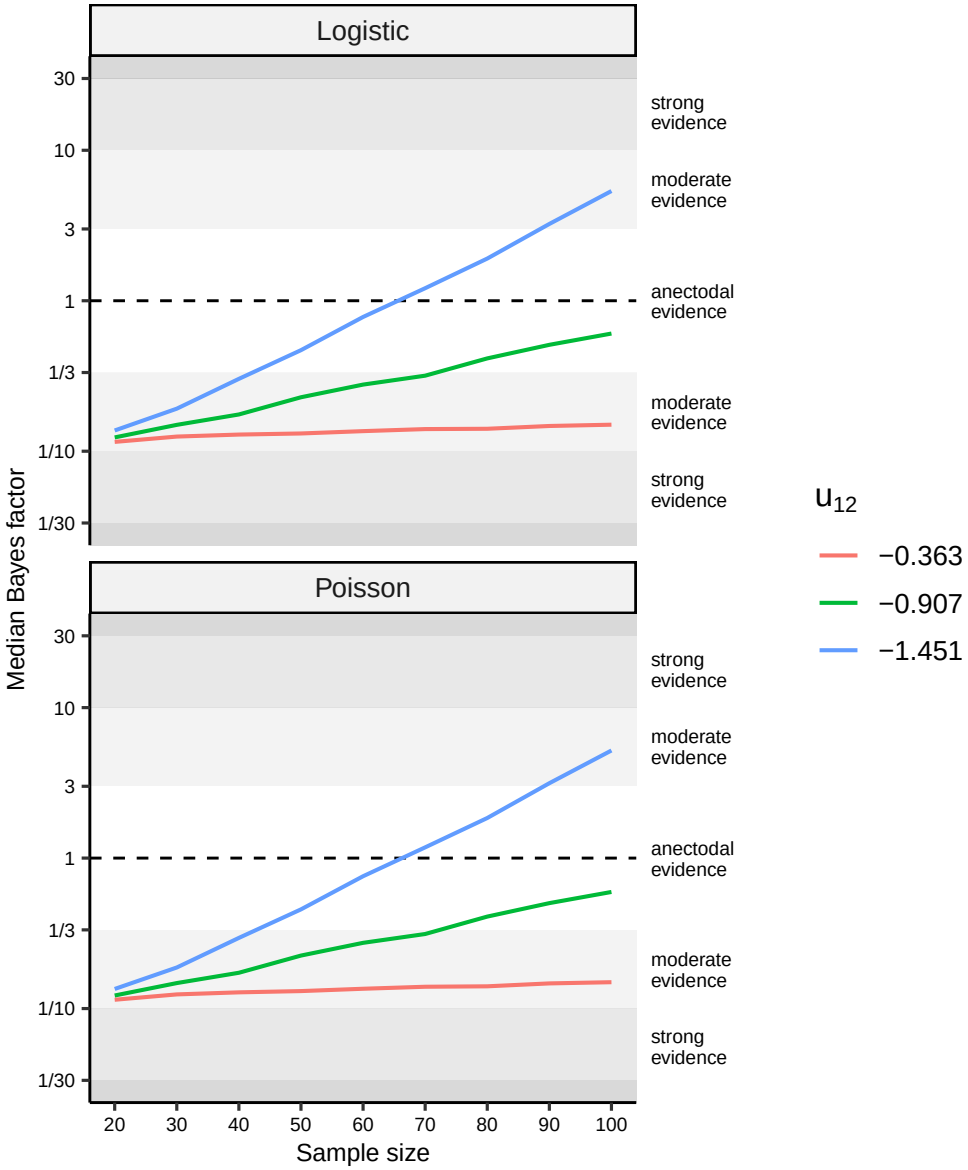


Figure 23: Median Bayes factor using $g = 100$ for negative effect size values

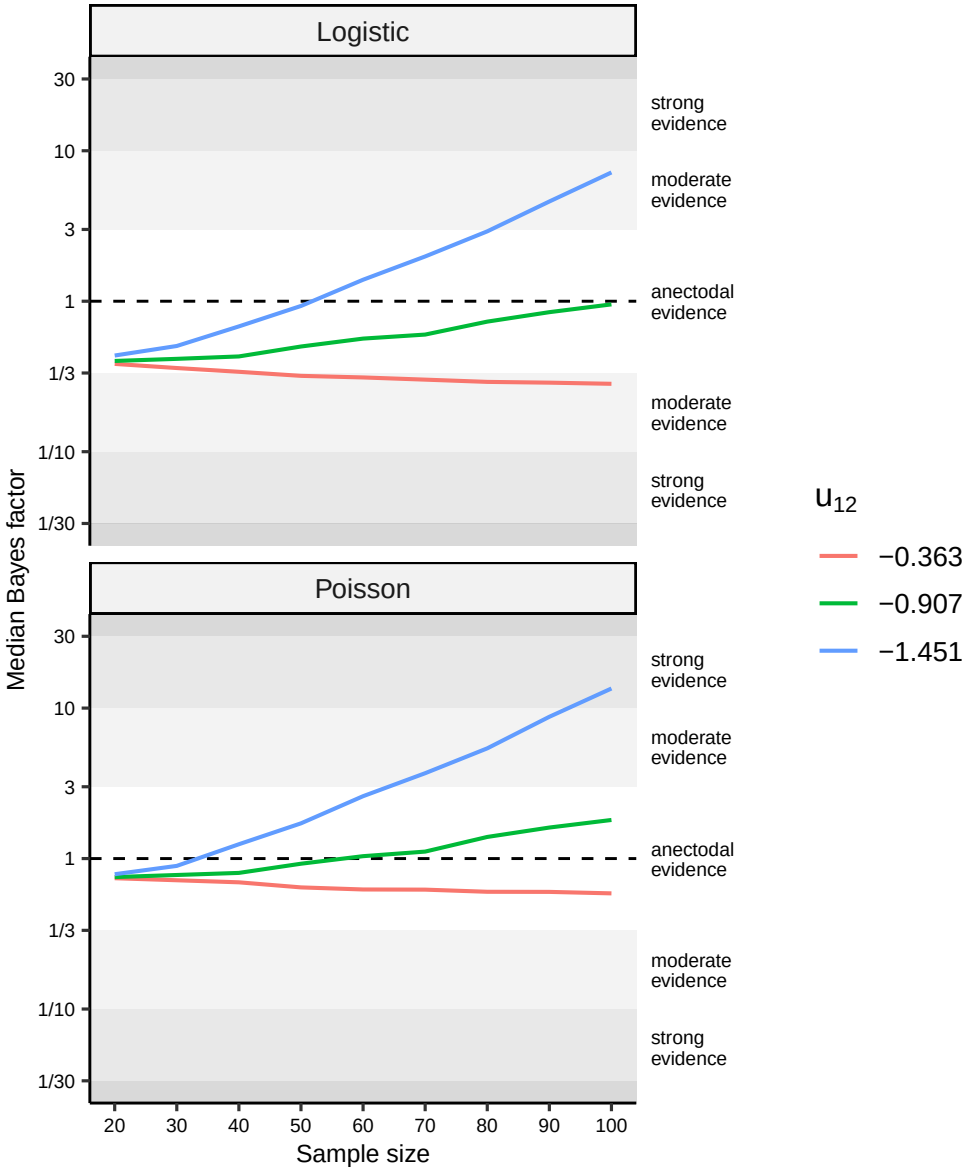


Figure 24: Median Bayes factor using the hyper-g/n prior for negative effect size values

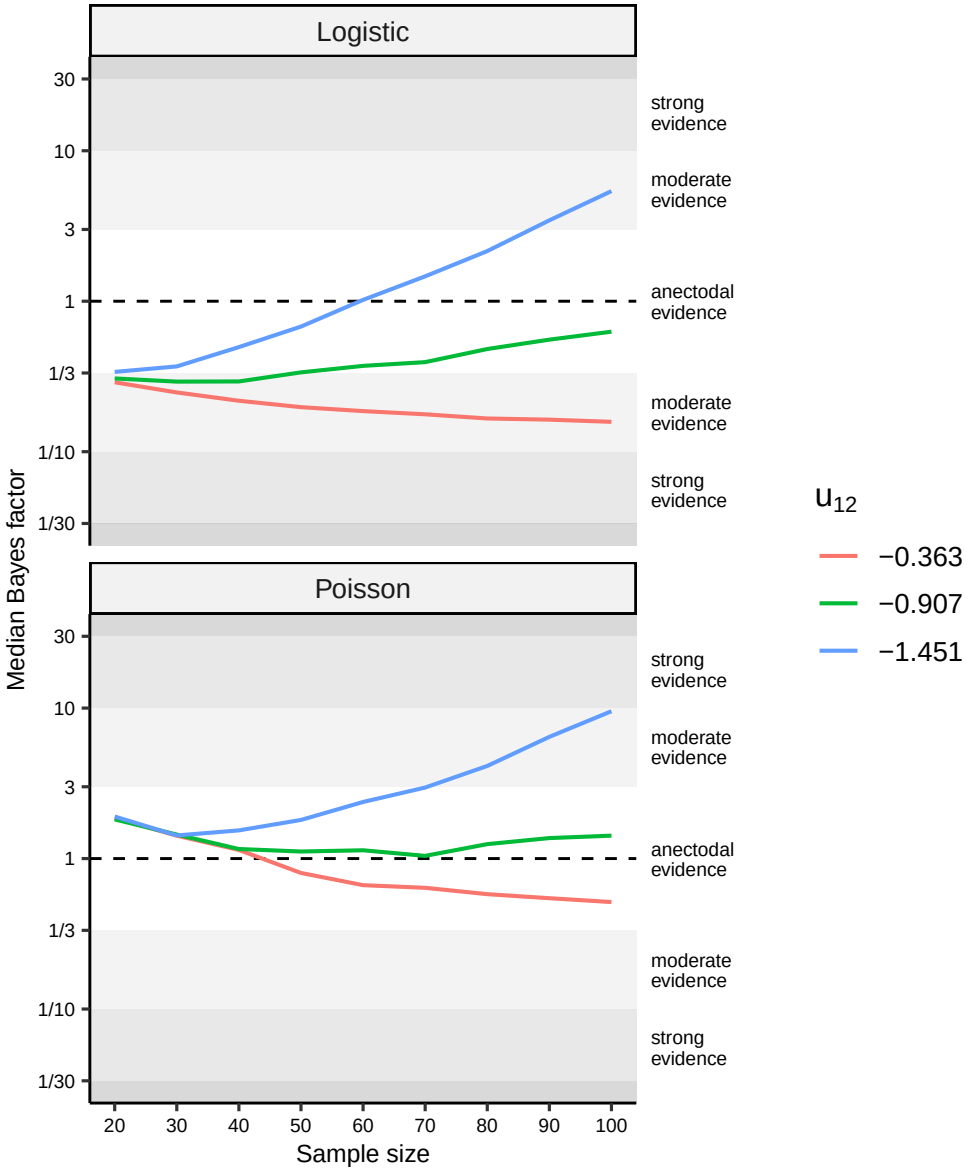


Figure 25: Median Bayes factor using the Beta-prime prior for negative effect size values

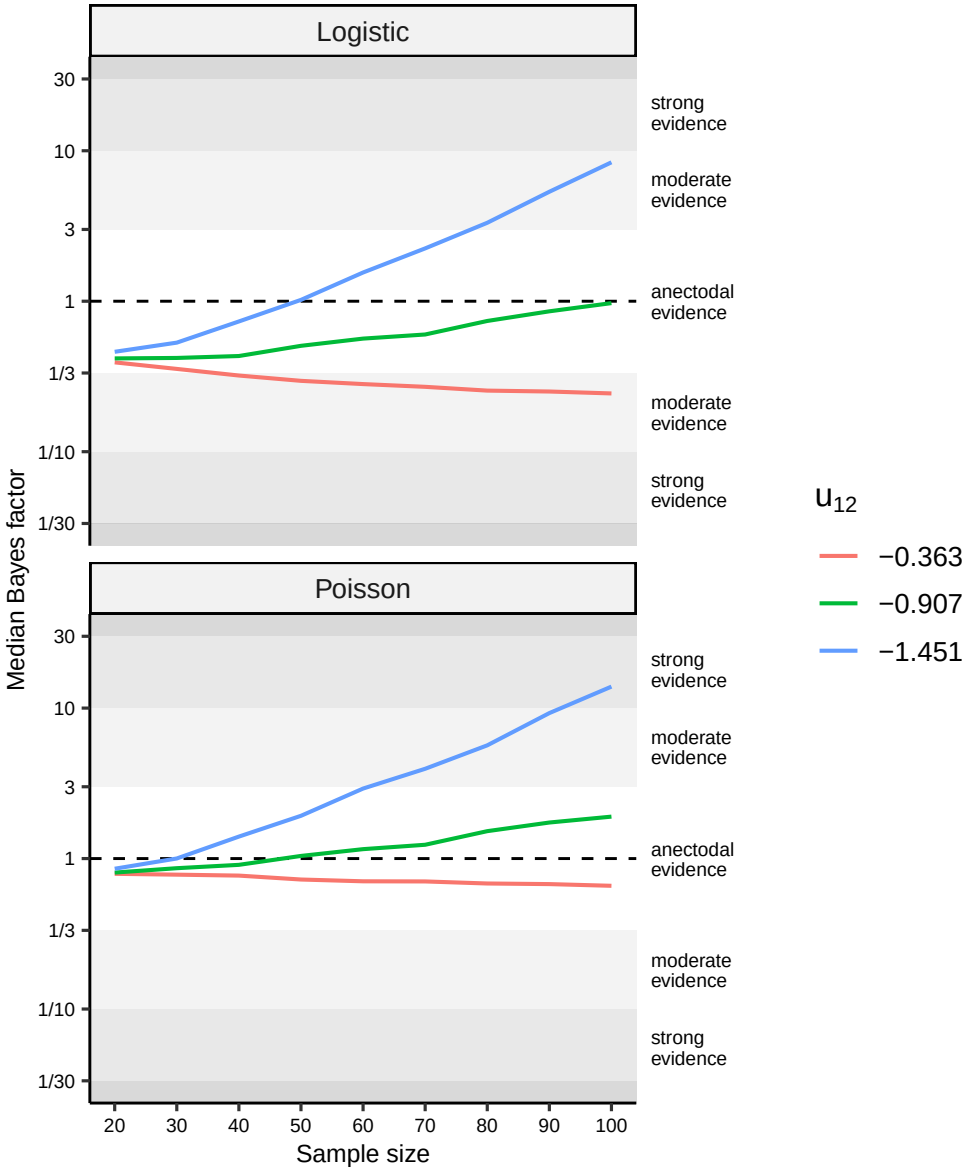


Figure 26: Median Bayes factor using the intrinsic prior for negative effect size values

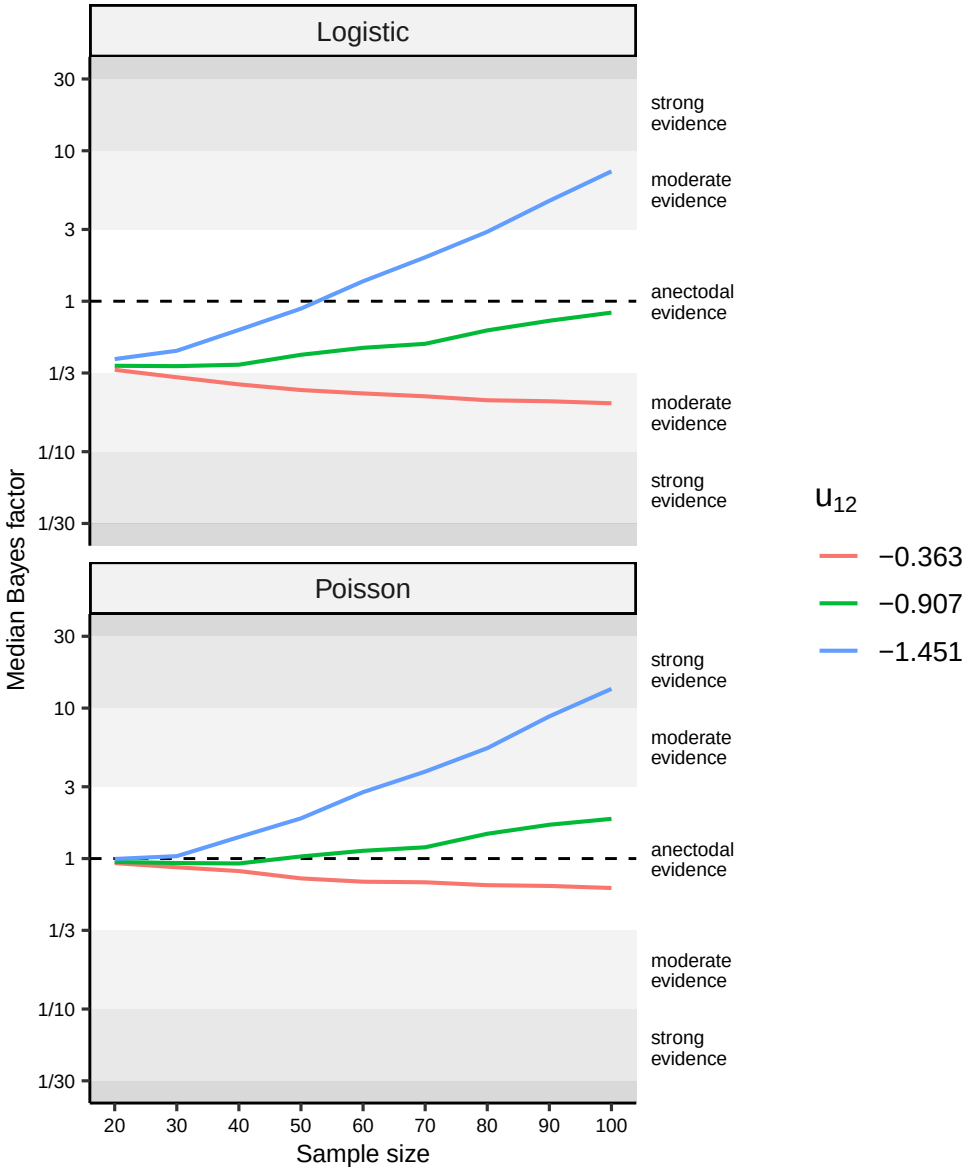


Figure 27: Median Bayes factor using the robust prior for negative effect size values

7.2.2 Effect of the nuisance parameters

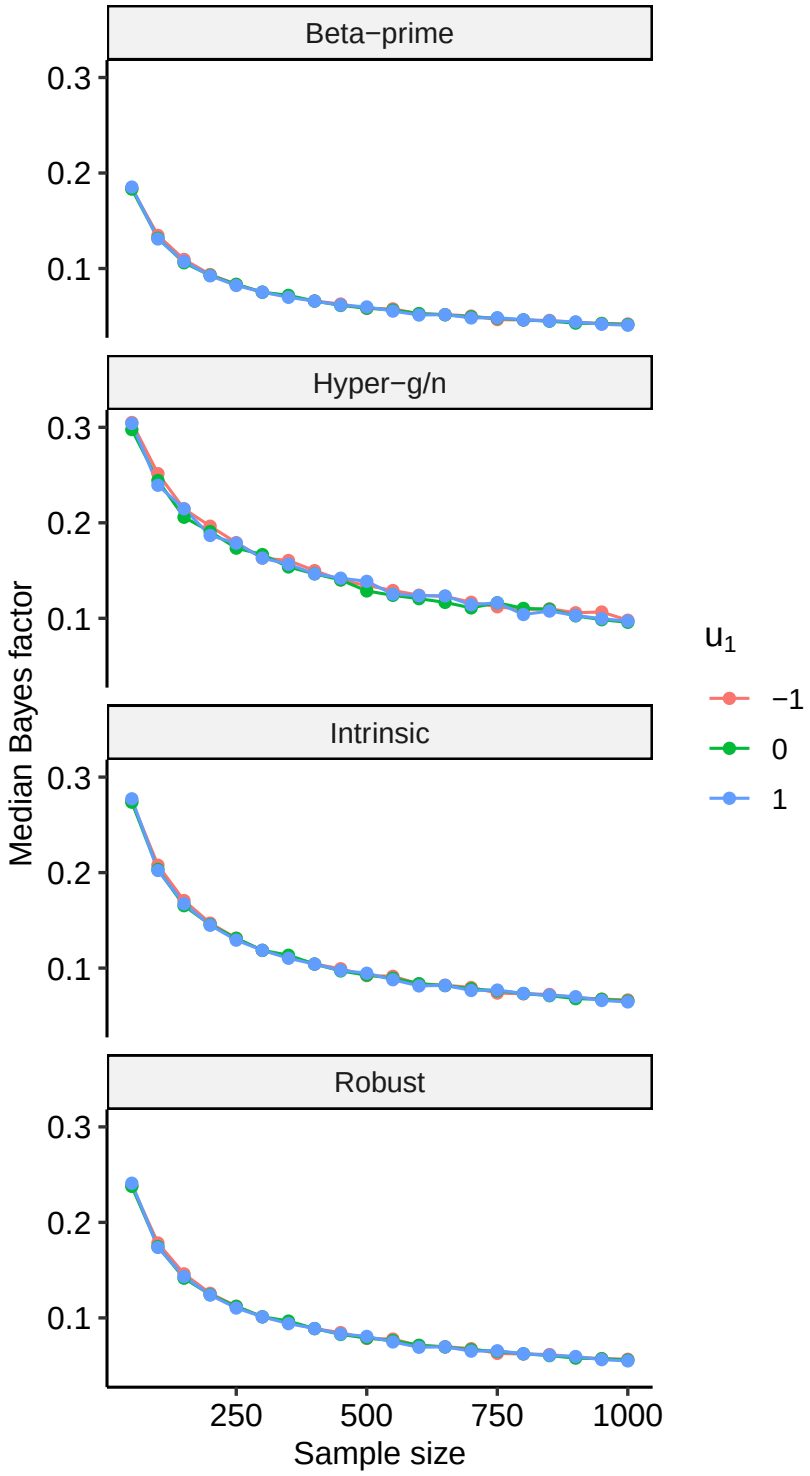


Figure 28: Median Bayes factor for different u_1 values for the Beta-prime,hyper-g/n, intrinsic and robust prior ($u_{12} = 0$, logistic model)

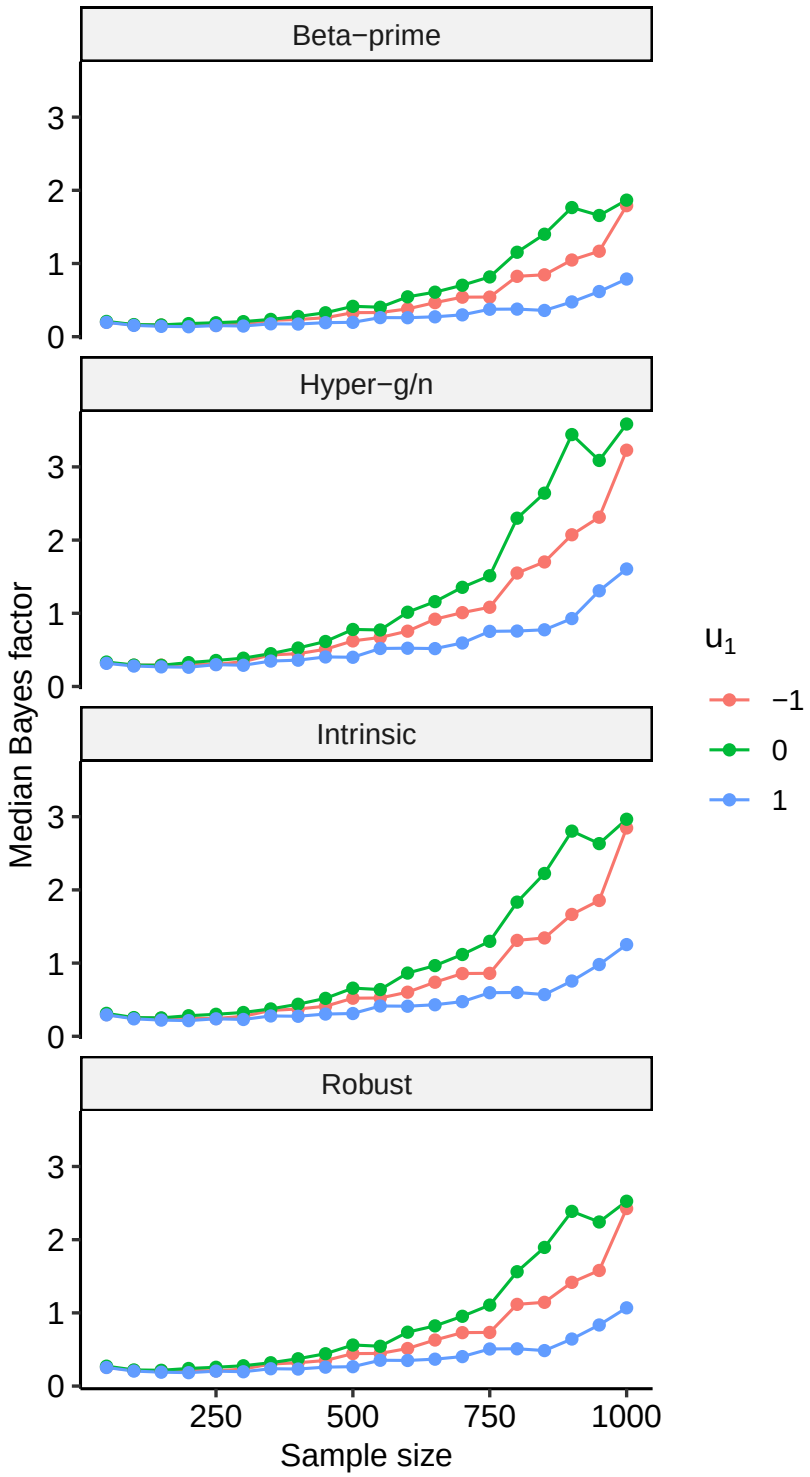


Figure 29: Median Bayes factor for different u_1 values for the Beta-prime, hyper-g/n, intrinsic and robust prior ($u_{12} = 0.4$, logistic model)

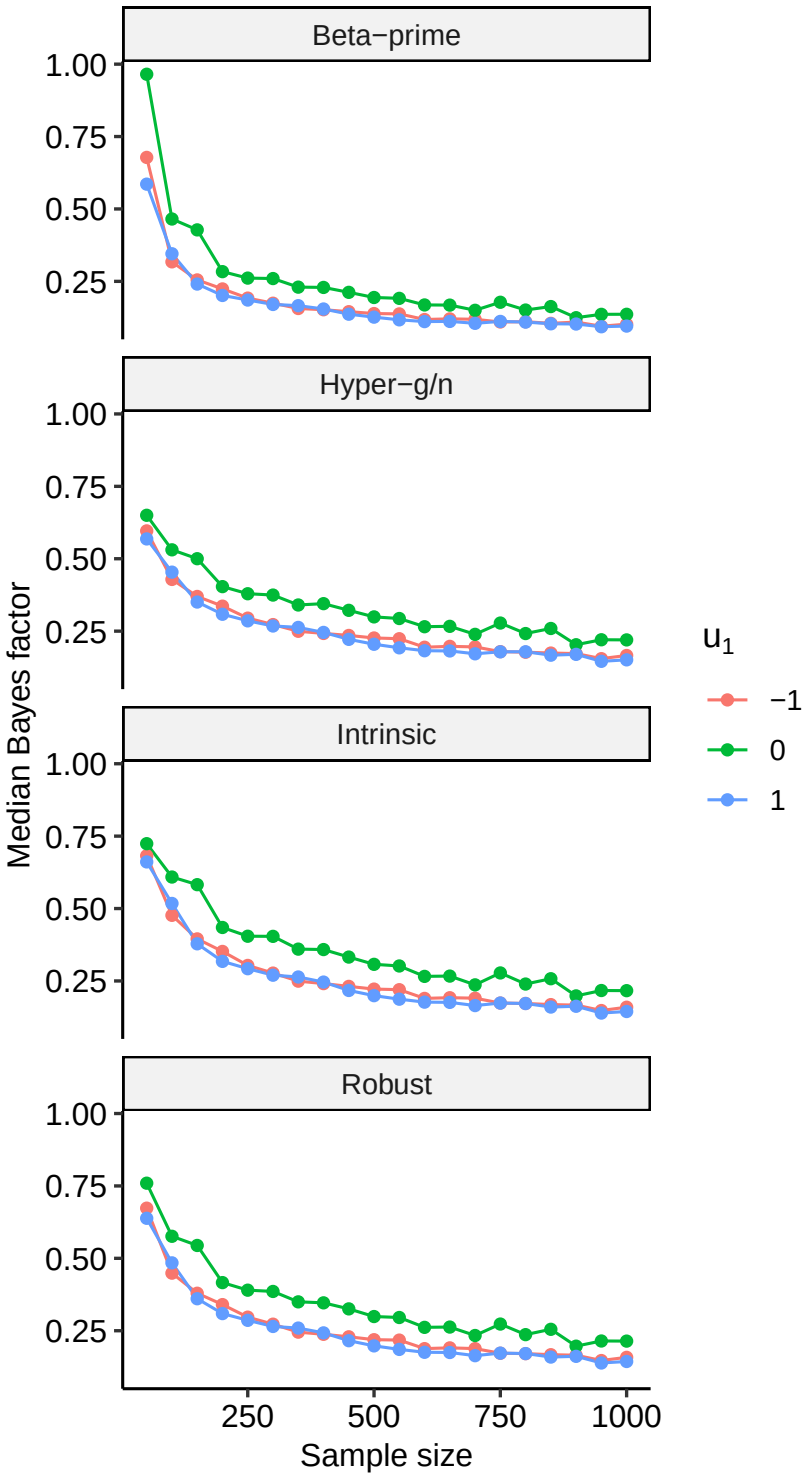


Figure 30: Median Bayes factor for different u_1 values for the Beta-prime, hyper-g/n, intrinsic and robust prior ($u_{12} = 0$, Poisson model)

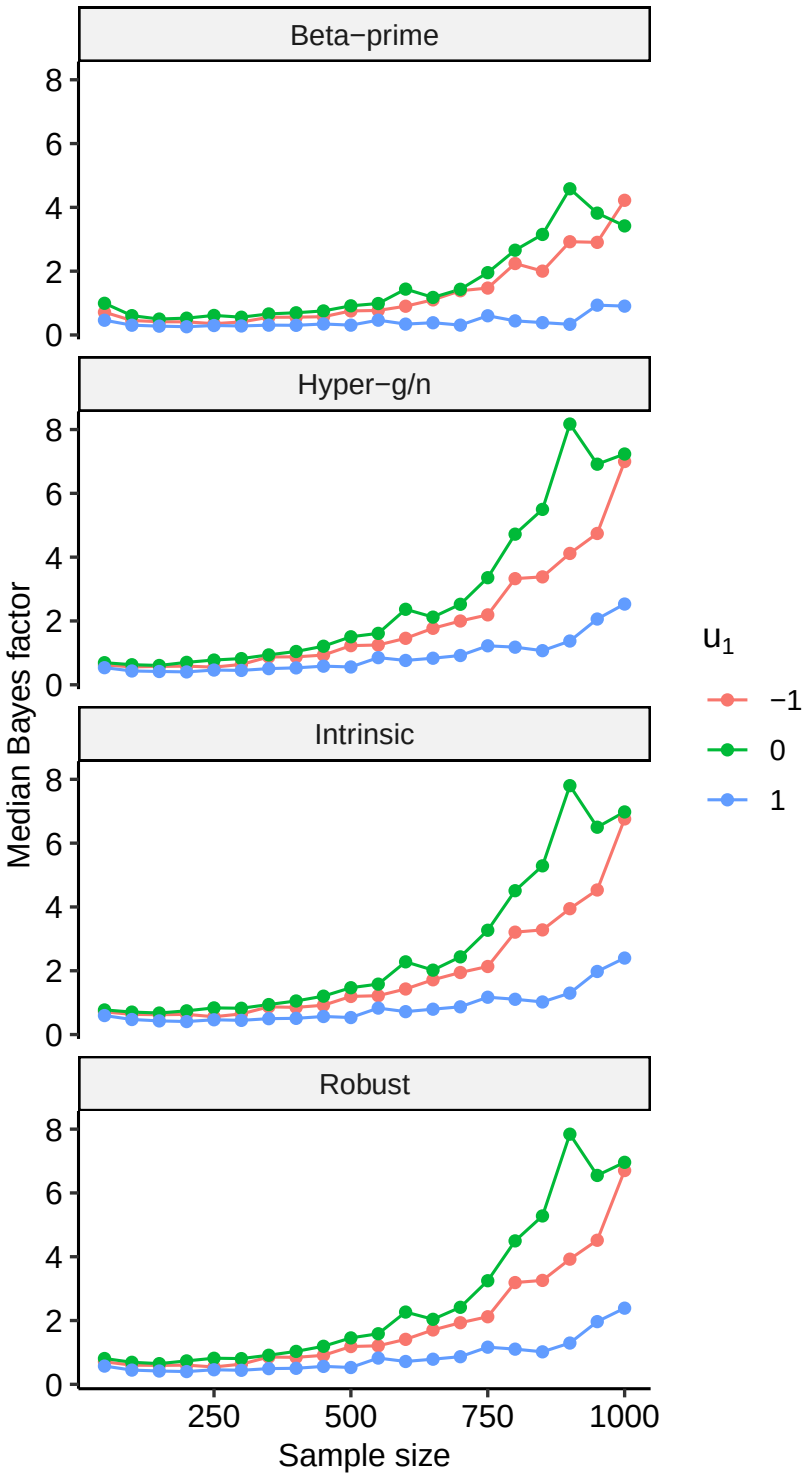


Figure 31: Median Bayes factor for different u_1 values for the Beta-prime, hyper-g/n, intrinsic and robust prior ($u_{12} = 0.4$, Poisson model)

References

Alan Agresti. *Categorical data analysis*, volume 482. John Wiley & Sons, 2003.

Alan Agresti. *An introduction to categorical data analysis*. John Wiley & Sons, 2018.

Casper J Albers. Dutch research funding, gender bias, and simpson's paradox. *Proceedings of the National Academy of Sciences*, 112(50):E6828–E6829, 2015.

Maria J Bayarri, James O Berger, Anabel Forte, and Gonzalo García-Donato. Criteria for bayesian model choice with application to variable selection. *The Annals of statistics*, 40 (3):1550–1577, 2012.

Richard Becker. *The new S language*. CRC Press, 2018.

James O Berger and Luis R Pericchi. The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122, 1996.

NS Breslow. The analysis of case-control studies. *Statistical methods in cancer research*, 1, 1980.

Henian Chen, Patricia Cohen, and Sophie Chen. How big is a big odds ratio? interpreting the magnitudes of odds ratios in epidemiological studies. *Communications in Statistics—simulation and Computation®*, 39(4):860–864, 2010.

Ronald Christensen. Log-linear models and logistic regression. 1997.

Merlise Clyde. Package 'bas'. *Bernoulli*, 8:1, 2015.

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

NE Day and DP Byar. Testing hypotheses in case-control studies-equivalence of mantel-haenszel statistics and logit score tests. *Biometrics*, pages 623–630, 1979.

Brian S Everitt. *The analysis of contingency tables*. Chapman and Hall/CRC, 2019.

Quentin F Gronau, Alexander Ly, and Eric-Jan Wagenmakers. Informed bayesian t-tests. *The American Statistician*, 2019.

Allard Hendriksen, Rianne de Heide, and Peter Grünwald. Optional stopping with bayes factors: a categorization and extension of folklore results, with an application to invariant situations. *Bayesian Analysis*, 16(3):961–989, 2021.

Pierre Humbert. Some extensions of pincherle's polynomials. *Proceedings of the Edinburgh Mathematical Society*, 39:21–24, 1920.

Tahira Jamil, Alexander Ly, Richard D Morey, Jonathon Love, Maarten Marsman, and Eric-Jan Wagenmakers. Default "gunel and dickey" bayes factors for contingency tables. *Behavior Research Methods*, 49(2):638–652, 2017.

Kaggle.com. Kaggle titanic data set. `https://www.kaggle.com/c/titanic/data`. Accessed: 2021-10-15.

Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

Robert E Kass and Larry Wasserman. A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the american statistical association*, 90(431):928–934, 1995.

Yingbo Li and Merlise A Clyde. Mixtures of g-priors in generalized linear models. *Journal of the American Statistical Association*, 113(524):1828–1845, 2018.

Feng Liang, Rui Paulo, German Molina, Merlise A Clyde, and Jim O Berger. Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008.

Alexander Ly, Josine Verhagen, and Eric-Jan Wagenmakers. Harold jeffreys's default bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, 72:19–32, 2016.

Yuzo Maruyama and Edward I George. Fully bayes factors with a generalized g-prior. *The Annals of Statistics*, 39(5):2740–2765, 2011.

Peter McCullagh and John A Nelder. *Generalized linear models*. Routledge, 2019.

Mary L McHugh. The chi-square test of independence. *Biochemia medica*, 23(2):143–149, 2013.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL `https://www.R-project.org/`.

Jeffrey N Rouder. Optional stopping: No problem for bayesians. *Psychonomic bulletin & review*, 21(2):301–308, 2014.

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

Johnny van Doorn, Don van den Bergh, Udo Böhm, Fabian Dablander, Koen Derks, Tim Draws, Alexander Etz, Nathan J Evans, Quentin F Gronau, Julia M Haaf, et al. The jasp guidelines for conducting and reporting a bayesian analysis. *Psychonomic Bulletin & Review*, 28(3):813–826, 2021.

Alexander von Eye and Alka Indurkhya. Log-linear representations of the mantel-haenszel and the breslow-day tests. *Methods of Psychological Research Online*, 5:13–30, 2000.

Eric-Jan Wagenmakers, Maarten Marsman, Tahira Jamil, Alexander Ly, Josine Verhagen, Jonathon Love, Ravi Selker, Quentin F Gronau, Martin Šmíra, Sacha Epskamp, et al. Bayesian inference for psychology. part i: Theoretical advantages and practical ramifications. *Psychonomic bulletin & review*, 25(1):35–57, 2018.

Thomas D Wickens. *Multiway contingency tables analysis for the social sciences*. Psychology Press, 2014.

Brian S Yandell. *Practical data analysis for designed experiments*. Routledge, 2017.

Arnold Zellner. On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques*, 1986.