BAYESIAN APPROACHES TO THE PROBLEM OF SPARSE TABLES IN LOG-LINEAR MODELING

Francisca Galindo-Garre
University of Murcia[1]
Jeroen K. Vermunt
Tilburg University[2]
Manuel Ato-García
University of Murcia

This paper presents Bayesian approaches to parameter estimation in the log-linear analysis of sparse frequency tables. The proposed methods overcome the non-estimability problems that may occur when applying maximum likelihood estimation. A crucial point when using Bayesian methods is the specification of the prior distributions for the model parameters. We discuss the various possible priors and assess their influence on the parameter estimates by two empirical examples in which maximum likelihood estimation gives problems. For the practical implementation of the Bayesian estimation methods, we used a Metropolis algorithm.

Key words: Bayesian statistics, log-linear analysis, sampling zeroes, estimation methods, priors, Metropolis algorithm

1 INTRODUCTION

Often the sample size N is not much larger than the number of cells in the contingency table. This occurs when the size of the sample is small or when the number of categories classifying the table is large. In those cases, a substantial number of cells may contain no observations. According to Agresti (1990), a sparse table is a contingency table in which more than about 20 percent of the cells have expected cell counts below 5.

[1] Departamento de Psicología Básica y Metodología. Espinardo (Murcia) Spain. E-mail: fgalindo@fcu.um.es
[2] Department of Methodology. P.O. Box 90153. 5000 LE Tilburg. E-mail: J.K.Vermunt@kub.nl

The analysis of sparse tables can give two types of problems. One class of problems associated with sparse tables is related to the goodness-of-fit testing since the asymptotic approximations of the standard chi-squared statistics tend to be poor for these tables. In certain situations, one might apply exact tests to check the goodness-of-fit (Kim and Agresti 1997), but this is not always the case.

Another class of problems is related to the non-existence of the maximum likelihood (ML) estimates and the asymptotic standard errors for certain log-linear parameters. More precisely, sometimes parameter estimates take on values of plus or minus infinity. In such cases, algorithms like IPF and Newton-Raphson may even fail to converge (Clogg et al., 1991).

A solution that is often used for the latter problem is to add a small constant, say 0.5, to every cell of a sparse table prior to analysis. One of the effects of adding a constant is that the estimates of log-linear parameters smooth toward zero. Another effect is that the sample size is increased. Goodman (1970, 1971) recommended using this procedure for saturated models only. A different approach proposed by Clogg et al. (1991) is to preserve the marginal distribution of the dependent variable when prior observations are divided among cells of the contingency table. Agresti (1990) recommended performing a sensibility analysis repeating the analysis with constants of various sizes. It may even be adequate to add an extremely small constant, such as 10^{-8}, to the empty cells.

From a Bayesian point of view, adding a small constant to every cell is equivalent to using a particular type of prior information about the parameters. However, this just one of the many possible ways to specify prior information on the parameters. In this paper, we explore the various types of priors that can be used in a Bayesian estimation of log-linear models. Our goal is to find a prior that on the one hand prevents the estimation problems associated with sparse tables, but on the other hand influences the parameter as less as possible.

Below, we first introduce two empirical examples in which ML estimation gives problems. Then we present the Bayesian approaches. The paper ends with some conclusions.

2 TWO EMPIRICAL EXAMPLES

In this section, we present two examples that represent two common estimation problems in sparse tables.

2.1 Example 1: Non-existence Problems in the Model of no Three-Factor Interaction

Table 1 gives a 2-by-2-by-2 contingency table that was presented by Clogg et al. (1991).

Table 1. Contingency table with two sampling zeroes

Predictors		Response Variable	
x_1	x_2	Y=1	Y=2
1	1	0	3
-1	1	9	4
1	-1	6	3
-1	-1	5	0
Totals		20	10

Source: Clogg, C. et al. (1991)

The model of interest for this table is the following logit model:

$$\phi_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \ . \qquad (1)$$

This example illustrates the problem of non-existence of ML estimates. It is an interesting case because all group totals and all sufficient statistics are nonzero. As already mentioned by Clogg et al, 1991, existing ML routines based on iterative proportional fitting, Newton-Raphson, or equivalent algorithms have difficulties finding estimates for the log-linear parameters. Table 2 reports the results we obtained by a Newton-Raphson algorithm. As can be seen, the parameters take on very extreme values, and the same applies to their standard errors.

The standard solution to the estimation problems that were encountered is to add a small constant, say 0.5, to every cell in the contingency table. Note that since the sample size is very small (N=30), even with a small number like 0.5 quite a lot of non-observed information is added with such procedure. As can be seen from the results reported in Table 2, adding 0.5 to each cell smoothes the estimates of log-linear parameters toward zero and gives estimates for all s.e.'s.

Table 2. Maximum likelihood estimates for example 1

	ML		ML after adding 0.5	
Predictor	Parameter	s.e.	Parameter	s.e.
Constant	0.75	0.46	0.62	0.40
x_1	-144.90	685.93	-1.15	0.58
x_2	-144.84	685.93	-1.07	0.58
Model fit	L^2= 0.00, df= 1		L^2= 0.16, df = 1	

2.2 Example 2: Non-existence Problems in a Model with a Sufficient Statistic zero

Table 3 gives a 2-by-2-by-2-by-2 contingency table that was presented by Fahrmeir and Tutz (1994) showing data on infection following birth by Caesarean section. The response variable is the occurrence or non-occurrence of infection. Three dichotomous covariates were considered.

Table 3. Contingency table with zero sufficient statistics

	Caesarean planned		Not planned	
	Infection		Infection	
	yes	No	Yes	No
Antibiotics				
Risk factors	1	17	11	87
No risk factors	0	2	0	0
No antibiotics				
Risk factors	28	30	23	3
No risk factors	8	32	0	9

Source: Fahrmeir, L. and Tutz, G. (1991)

The model considered is a logit model with an interaction effect between Antibiotic and Risk factors,

$$\phi_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i2} x_{i3}. \quad (2)$$

This example illustrates the non-existence of ML estimates because zeroes in the sufficient statistics. In such cases, ML routines based on iterative proportional fitting, Newton-Raphson, or equivalent algorithms have difficulties to estimate the parameters concerned, in this case, the interaction parameter.

Table 4 presents results obtained with Newton-Raphson. We can observe that our program has problems to estimate the parameter whose sufficient statistics contains a zero. If a small constant (0.5) is added to each cell, also the interaction parameter can be estimated and the other parameters are smoothed towards zero. The effect of adding a constant does not seem to be very large since the ratio between the parameter values and the standard deviations remains approximately the same.

Table 4. Maximum likelihood estimates for example 2

	ML		ML after adding 0.5	
Predictor	Parameter	s.e.	Parameter	s.e.
Constant	-9.03	5.30	-2.16	0.33
x_1	6.31	5.29	-0.12	0.31
x_2	8.00	5.30	1.22	0.31
x_3	-1.99	0.32	-1.64	0.26
x_2 x_3	-7.29	5.30	-0.59	0.31
Model Fit	L^2= 0.95, df = 3		L^2= 8.30, df = 3	

3 BAYESIAN APPROACH

The main difference between ML and Bayesian approaches is that in the latter the likelihood function is combined with a prior. The objective is to provide information of the posterior distribution of the parameters. The posterior distribution can be obtained from the likelihood and the priors by the Bayes rule; that is,

$$p(\theta / y) = \frac{p(y/\theta)p(\theta)}{\int p(y/\theta)p(\theta)d\theta} \propto p(y/\theta)p(\theta), \qquad (3)$$

where θ is the parameter vector, y represents the observed data, $p(y \mid \theta)$ is the likelihood function, $p(\theta)$ denotes the prior distribution of the parameters, and $p(\theta / y)$ denotes the posterior distribution of the parameters.

When analysing the sparse contingency tables, the prior information can be to make sure that estimates of the parameter can be calculated. However, there are many possible choices for the prior distribution. In some cases, an informative prior may be more adequate because there are some theories about the model. But, in other cases, no a priori information about the model is available or the objective is to get the same estimates that would be obtained by a classical approach.

Nowadays, it is becoming more and more common to use noninformative priors. There is, however, a controversy with respect the definition of noninformative prior distributions in multinomial models, The introduction of covariates, as in our examples, makes the problem even more complicated.

Four kinds of priors have been proposed for Bayesian estimation of logit models: natural conjugate priors (Koop and Poirier, 1993, 1994, and 1995), normal priors (Koop and Poirier, 1993), the Jeffreys' prior (Ibrahim and Laud, 1991), and uniform priors for the logit coefficients (Koop and Poirier 1993, 1994, 1995, and Ibrahim and Laud 1991).

The estimation of the marginal posterior densities of the parameters reported below was performed with a Metropolis algorithm. We used independent univariate normals as jump functions. The jump functions have means equal to the previous estimates and variances that are set after the burning in. The algorithm generates initial runs of 1000 iterations as a "burn-in" in order to reach convergence. After that, it performs iterations until a maximum of 100.000 iterations or until the convergence criterion was less than 1.001. We monitor convergence of the iterative simulation by the factor:

$$\sqrt{\hat{R}} = \sqrt{\frac{\widehat{\text{var}}^{+}(\psi / y)}{W}}, \tag{4}$$

where,

$$\widehat{\text{var}}^{+}(\psi / y) = \frac{n-1}{n}W + \frac{1}{n}B. \tag{5}$$

Here, B and W are the between- and within-sequence variances. This statistic decline to 1 as $n \rightarrow \infty$. We used three chains to calculate the convergence factor. When this factor is around one[3], we have reasons to believe that we are simulating from the posterior distribution. The program we made also calculates posterior modes by a Newton-Raphson.

In order to assess the effect of different priors on the parameters, we use the ratio between means (or modes) and standard errors. The main objective is to get stable estimates of the significance of the parameters changes when different kinds of priors are used. As we will show, when priors are less informative both the parameter estimates and standard deviations tend to increase.

3.1 Uniform prior

Using a uniform prior is equivalent to estimation without prior information. In this case, the prior density is constant, where the constant is typically set to 1. The uniform prior was popularised by Laplace (1812).

[3] Gelman, et al. (1995) suffested that values bellow 1.2 are acceptable.

Table 5. Results with uniform prior

Coeffic.	Mean	s.e.	z	Mode	s.e.	z
Example 1						
β_0	-66.49	0.49	1.73	0.75	0.46	1.62
β_1	-66.49	63.33	-1.05	-144.90	685.93	-0.21
β_2	-66.44	63.33	-1.05	-144.84	685.93	-0.21
Example 2						
β_0	-9.03	5.30	-1.70	-9.53	3.16	-3.01
β_1	6.31	5.29	1.19	6.91	3.16	2.18
β_2	8.00	5.30	1.51	8.52	3.16	2.69
β_3	-1.99	0.32	-6.20	-1.91	0.30	-6.36
β_4	-7.29	5.30	-1.38	-7.83	3.16	-2.48

We observe from Table 5 that the values of the parameters and standard deviations are extremely high in the first example. Another problem is that the Metropolis algorithm does not converge. The value of convergence is 1.44 in example 1 and 1.46 in example 2, which indicates that after 100000 iteration the simulations are not from the target distribution.

It seems that with sparse tables this type of prior information is not very useful because it has the same problems as ML estimation.

3.2 Jeffreys' prior

Another interesting prior is the Jeffreys' prior. This prior exhibits many nice features making it an attractive reference prior. One of its properties is that it is parameterization invariance. Jeffrey's prior also has the property of being approximately noninformative in the sense of Box and Tiao (1973), who motivated Jeffreys' prior by introducing the notion of data translated likelihood. The Jeffreys' prior is defined by

$$p(\theta) = \left|I(\theta)\right|^{\frac{1}{2}} , \qquad (6)$$

where $|\ .|$ denotes the determinant, and $I(\theta)$ is the expected Fisher information matrix calculated using the log-likelihood function. A common approach, assuming that "ignorance" is consistent with "independence ", is to obtain a noninformative prior for each parameter individually, and calculate the joint prior as the product of these individual prior (Schafer, 1995). For an unrestricted multinomial distribution this prior is a flattering prior with constant value c=1/2. When there are no covariates, the prior obtained by using this method corresponds to a natural conjugate prior.

Ibrahim and Laud (1991) developed a general formulation to obtain the Jeffreys prior for the family of generalised linear models. In the case of a multinomial model with covariates, the Jeffreys' prior can be obtained as follows:

$$p(\theta) \propto \left| \sum_{n=1}^{N} \sum_{j=1}^{J} p_{nj}(\theta) \cdot \left[x_{nj} - \bar{x}_n(\theta) \right] \cdot \left[x_{nj} - \bar{x}_n(\theta) \right]' \right|^{\frac{1}{2}}, \quad (7)$$

where x_{nj} is and element of the design matrix and $\bar{x}_n(\theta)$ is a weighted average of x_n characteristics for observation n.

Table 6. Results with Jeffreys' priors

Coefficient	Mean	s.d.	z	Mode	s.d.	z
Example 1						
β_0	0.76	0.46	1.63	0.67	0.43	1.54
β_1	-2.15	1.18	-1.82	-1.45	0.77	-1.89
β_2	-2.10	1.19	-1.77	-1.38	0.76	-1.78
Example 2						
β_0	-2.81	0.56	-4.99	-2.49	0.42	-5.87
β_1	0.19	0.53	0.36	-0.04	0.40	-0.09
β_2	1.80	0.54	3.30	1.50	0.40	3.72
β_3	-1.91	0.30	-6.39	-1.84	0.29	-6.40
β_4	-1.11	0.54	-2.06	-0.84	0.40	-2.09

The results in Table 6 show that in comparison with the uniform prior the Jeffreys'prior is quite informative. It yields guarantees that one obtains stable estimates of the parameters as well as their standard errors. The advantage of this prior is that it is calculated in a standard way and no ad hoc choice has to make about the amount of information that has to be added. We will take this prior as a reference to compare with the others priors presented below. A disadvantage of the Jeffrey's prior that it is takes much more time to calculate it than other priors, which can become problematic when using simulation methods like the metropolis algorithm.

3.3 Dirichlet prior

One of the most interesting characteristics of the Dirichlet distribution is that it is a member of the same family as the multinomial distribution. More precisely, these two distributions are conjugate since their kernels are of the same form. This is attractive in sensitivity analyses since it is easy to quantify the impact of the prior on the estimates. The consequence of the multinomial and Dirichlet distributions being conjugate is that when they are combined one

obtains another Dirichlet distribution with parameters equal to the sum of the multinomial parameters and the Dirichlet prior parameters. If we have little prior information then it may be appropriate to take the Dirichlet parameters equal to a common value. However, there is not a unique choice for this value that clearly represents a state of ignorance about the parameters. An interesting option is described by Clogg et. al., (1991): they proposed using a Dirichlet that preserves the marginal distribution of the dependent variable, and that takes number of cells of the contingency table into account. Koop and Poirier (1993, 1994, and 1995) used similar types of conjugate priors in the estimation of logit models.

When a constant (c) is added to every cell, the information equivalent to c multiplied by the size of the table is introduced. This implies that that if the sample size is small, the information that is added can be even larger than the actual sample size. In the absence of strong prior beliefs about the parameters, it is probably unwise to add prior information that amounts to more than about 10-20% of the actual sample size (see Schafer, 1997)

Therefore, it can be appropriate to consider other types of priors in which a certain number of prior observations is added. Also the way of splitting them into the prior information can be chosen by the researcher. Clogg et al. (1991) advocated a strategy in which prior observations are divided among cells of the contingency table in such a way that the marginal distribution of one of the variables is preserved. In the same way, it is possible to use a prior that preserves the marginal distributions of all the variables in the contingency table. These approaches prevent smoothing parameters towards a uniform model, which can distort inference about parameters when the marginal distributions of the variables are far from uniform.

As can be seen from Table 7, in the first example, the equiprobability prior with c=0.5 smoothes the estimates more towards zero that the Jeffreys' prior, which is not surprising given that the constant added is quite high for this sample size (N=30). When c=0.1,there is on the one hand less smoothing, but on the other hand because the standard deviations increase even more, effects which were significant with c=0.5 are no longer significant.

Compared to the equiprobability priors, the independence priors shrink the intercept less to zero. However, the to effects are quite similar between equiprobability and independence prior. When using the independence prior with small constants the results are most similar to the results with a Jeffreys' prior. When using smaller constants, the posterior-mode seems to be more stable than the posterior mean. However, the ratios between means/modes and standard deviations are approximately the same in all the cases.

For example 2, we get similar results, but changes are less important in this example than in the first one. When c=0.5, changes in parameter values and standard deviations are smaller because the sample size is larger than in example 1. When c=0.1, changes are also smaller because the contingency table has a larger number of cells and the amount of information added is higher than in the other example.

3.4 Univariate normal priors

Another possible specification of the priors is to assume the log-linear parameters to come from independent univariate normal distributions. For each parameter, we have to specify the mean and the variance. The mean of the normal prior will typically be set to 0. The size of the variance affects to the amount of prior information that is added. A noninformative prior is obtained by setting variances to a very large value.

In sparse tables, this procedure helps to get estimates of the log-linear parameters because it permits to calculate the value of log-linear parameters even when they are close to the boundary.

Table 7. Results with Dirichlet prior

Coefficient	Mean	s.e.	z	Mode	s.e.	z
Example 1						
Equiprobability (c= 0.5)						
β_0	0.68	0.44	1.54	0.62	0.40	1.53
β_1	-1.43	0.70	-2.04	-1.15	0.58	-1.99
β_2	-1.36	0.71	-1.91	-1.07	0.58	-1.84
Equiprobability (c= 0.1)						
β_0	0.75	0.48	1.58	0.72	0.45	1.60
β_1	-3.65	2.36	-1.55	-1.92	1.16	-1.65
β_2	-3.59	2.36	-1.52	-1.86	1.17	-1.59
Independence (c_1= 0.5, c_2= 0.25)						
β_0	0.79	0.45	1.77	0.73	0.42	1.73
β_1	-1.66	0.81	-2.04	-1.29	0.65	-1.97
β_2	-1.60	0.82	-1.95	-1.23	0.66	-1.86
Independence (c_1= 0.3, c_2= 0.1)						
β_0	0.85	0.47	1.80	0.77	0.44	1.73
β_1	-2.36	1.32	-1.79	-1.59	0.85	-1.87
β_2	-2.31	1.33	-1.74	-1.54	0.86	-1.79
Example 2						
Equiprobability (c= 0.5)						
β_0	-2.34	0.40	-5.85	-2.16	0.33	-6.48
β_1	-0.01	0.37	-0.01	-0.12	0.31	-0.40
β_2	1.38	0.38	3.65	1.22	0.31	3.93
β_3	-1.68	0.27	-6.31	-1.64	0.26	-6.36
β_4	-0.74	0.37	-1.97	-0.59	0.31	-1.92
Equiprobability (c= 0.1)						
β_0	-3.14	0.85	-3.68	-2.72	0.60	-4.48
β_1	0.48	0.84	0.57	0.17	0.59	0.30
β_2	2.12	0.84	2.52	1.72	0.59	2.90
β_3	-1.93	0.31	-6.25	-1.85	0.29	-6.38
β_4	-1.42	0.84	-1.69	-1.05	0.59	-1.77
Independence (c_1= 0.5, c_2= 0.25)						
β_0	-2.33	0.39	-5.92	-2.15	0.33	-6.44
β_1	-0.04	0.36	-0.11	-0.14	0.31	-0.45
β_2	1.38	0.37	3.74	1.22	0.31	3.93
β_3	-1.71	0.27	-6.24	-1.66	0.26	-6.34
β_4	-0.73	0.37	-1.99	-0.59	0.31	-1.92
Independence (c_1= 0.3, c_2= 0.1)						
β_0	-2.59	0.47	-5.56	-2.36	0.40	-5.95
β_1	0.09	0.44	0.21	-0.06	0.37	-0.15
β_2	1.61	0.45	3.62	1.40	0.37	3.73
β_3	-1.81	0.28	-6.41	-1.75	0.28	-6.36
β_4	-0.94	0.44	-2.12	-0.74	0.37	-2.00

Table 8. Results with univariate normal priors

Coefficient	Mean	s.e.	z	Mode	s.e.	z
Example 1						
μ= 0, σ^2=10						
β_0	0.78	0.47	1.67	0.72	0.45	1.60
β_1	-2.66	1.31	-2.03	-1.92	1.06	-1.81
β_2	-2.61	1.33	-1.97	-1.86	1.07	-1.74
μ= 0, σ^2=100						
β_0	0.81	0.49	1.66	0.75	0.46	1.62
β_1	-6.50	4.18	-1.55	-2.89	2.74	-1.05
β_2	-6.44	4.18	-1.54	-2.83	2.75	-1.03
Example 2						
μ= 0, σ^2=10						
β_0	-3.30	0.79	-4.20	-2.80	0.64	-4.38
β_1	0.63	0.78	0.80	0.23	0.63	0.37
β_2	2.27	0.78	2.93	1.80	0.63	2.87
β_3	-1.95	0.31	-6.31	-1.87	0.29	-6.40
β_4	-1.58	0.77	-2.03	-1.12	0.63	-1.79
μ= 0, σ^2=100						
β_0	-6.48	2.87	-2.26	-3.33	1.68	-1.98
β_1	3.76	2.86	1.31	0.72	1.68	0.43
β_2	5.44	2.86	1.90	2.32	1.68	1.38
β_3	-1.99	0.31	-6.32	-1.91	0.30	-6.36
β_4	-4.74	2.86	-1.66	-1.64	1.68	-0.98

From the results reported in Table 8 we can observe that in the first example the parameters are quite dependent on the prior distribution. However, again the posterior modes are less dependent on the specification of the prior than the posterior means. In addition, the z values are much less influenced by the choice of the prior than the parameter estimates themselves. In both examples, normal priors with a variance of 10 seem to work best. In addition, the results obtained with these priors are similar to the ones obtained with Jeffreys' prior.

3.5 Multivariate normal priors

Since log-linear parameters are usually related, it is may be better to consider a multivariate normal prior distribution for the parameter vector. Ibrahin and Laud (1991) propose using a normal distribution with mean 0 and covariance matrix $k(X'IX)^{-1}$, where k is a known positive constant. Here, the value of k represents the amount prior information that the researcher wishes to add. If k is near to zero, then the prior will be near to noninformative. However, in our examples (X'IX) is a diagonal matrix with the same number on each

diagonal prior. This means that it yields the same prior as the set of independent univariate normal priors presented above.

Another option is to approximate the covariance matrix using the observed cell frequencies plus a small number, say 0.1, to circumvent the observed zeroes. More precisely, we propose using a multivariate normal distribution with covariance matrix equal to k times the approximate covariance matrix of the beta parameters. Here, k is again a constant influencing the effect of the prior. Table 9, presents the results of k=1 and k=0.01.

Table 9. Results with multivariate normal priors

Coefficient	Mean	s.e.	z	Mode	s.e.	z
Example 1						
k=1						
β_0	0.38	0.31	1.23	0.37	0.31	1.20
β_1	-1.56	0.69	-2.27	-1.34	0.63	-2.12
β_2	-1.53	0.69	-2.22	-1.30	0.63	-2.07
k=0.01						
β_0	0.81	0.48	1.70	0.74	0.46	1.61
β_1	-10.37	6.96	-1.49	-3.29	4.14	-0.79
β_2	-10.32	6.96	-1.48	-3.23	4.14	-0.78
Example 2						
k=1						
β_0	-1.85	0.38	-4.87	-1.69	0.34	-5.02
β_1	0.42	0.38	1.11	0.27	0.33	0.83
β_2	1.28	0.37	3.41	1.12	0.33	3.39
β_3	-1.06	0.17	-6.27	-1.05	0.17	-6.28
β_4	-0.90	0.37	-2.41	-0.75	0.33	-2.28
k=0.01						
β_0	-5.53	2.42	-2.28	-3.45	2.18	-1.58
β_1	2.85	2.41	1.18	0.86	2.18	0.39
β_2	4.50	2.42	1.86	2.45	2.18	1.12
β_3	-1.96	0.31	-6.39	-1.89	0.30	-6.38
β_4	-3.81	2.42	-1.58	-1.77	2.18	-0.81

In the first example we see that k=1 smoothes the parameters towards a model of equiprobability: the effect is similar to the effect of using Dirichlet equiprobability prior with c=0.5. Compared to the first example 1, in the second example the effect of different values of k is less important. This is comparable to what to found with the Dirichlet priors

3.6 The Maximal Data Information Prior (MDIP)

This prior was developed in Zellner (1971), based on information theoretic arguments. It is given by $\pi(\theta)=\exp\{\int p(x/\theta)\log p(x/\theta)dx\}$, where $p(x/\theta)$ is the data density function. In our examples with sparse tables, this prior was not useful because it gave convergence problems.

4 CONCLUSIONS

In this paper we showed that Bayesian methods may be used solve the estimation problems associated with maximum likelihood estimation when the contingency table contains ampling zeroes. The problem is, however, to find the most appropriate prior given the sample size and the number of cells in the contingency table.

We saw that the less informative the priors, the more extreme the parameter values. On the other hand, the significance of parameters (the z values) seem to be quite stable under different priors because the standard errors tend to change in the same direction as the parameter estimates when the priors become less informative.

The posterior mode estimates seem less dependent of the amount of prior information than the posterior mean estimates. This effect has been observed with all priors used in our paper.

When using too noninformative priors, such as uniform priors or normal priors with large variances, we encountered the same problems as with maximum likelihood estimation. Our conclusion is that irrespective of the type of prior that is used - Dirichlet, univariate normal or multivariate normal, there is a kind of equilibrium at which the amount of prior information is just enough to obtain stable estimates of both parameters and standard errors.

The Jeffreys' prior proved useful in our examples because the contingency tables and the number of parameters were not very large. However, in larger problems it is more difficult to apply because of its computational intensity.

Simulation studies in which one starts from a known population distribution should be performed to get more insight in the bias introduced by each of the types of the priors. Another point of future research is the comparison of posterior mean and posterior mode estimation.

REFERENCES

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

Berger, J.O. and Bernardo, J.M. (1992). Ordered group reference priors with application to the multinomial problem. *Biometrika*, 79, 25-37.

Bernardo, J.M. (1979). Reference posterior distributions for Bayes inference. *Journal of the Royal Statistical Society B*, 41, 113-147.

Clogg, C.C. and Eliason, S.R. (1987). Some common problems in Log-linear analysis. *Sociological Methods and Research*, 16, 8-44.

Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B., and Widman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association*, 86, 68-78.

Fahrmeir, L. and Tutz, G. (1994) *Multivariate Statistical Modelling Based on Generalized Linear Models*. Berlin: Springer.

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

Goodman, L.A. (1970). The multivariate analysis of qualitative data: interactions among multiple classifications. *Journal of the American Statistical Association,* 65,225-256.

Goodman, L.A. (1971). The analysis of multidimansional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics*, 13, 33-61.

Ibrahim, J.G. and Laud, P.W. (1991). On Bayesian analysis of general linear models using Jeffreys'prior. *Journal of the American Statistical Association*, 86,981-986.

Kim, D. and Agresti, A. (1997) Nearly exact tests of conditional independence and marginal homogeneity for sparse contingency tables. *Computational Statistics and Data Analysis,* 24, 89-104.

Koop, G. and Poirier, D.J. (1993). Bayesian analysis of logit models using natural conjugate priors. *Journal of Econometrics*, 56,323-340.

Koop, G. and Poirier, D.J. (1994). Rank-ordered logit models: An empirical analysis of Ontario voter preferences, *Journal of Applied Econometrics,* 9,369-388.

Koop, G. and Poirier, D.J. (1995). An empirical investigation of Wagner's hypothesis by using a model occurrence framework, *Journal of the Royal Statistical Society A,* 58,123-141.

Poirier, D. (1994). Jeffreys'prior for logit models. *Journal of Econometrics*, 63,327-339.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Vermunt, J.K. (1997). LEM: *a general program for the analysis of categorical data.* Tilburg: Tilburg University.

Yang, R. and Berger, J.O. (1996). A catalog of noninformative priors. *Technical report.*