Statistica Sinica **17**(2007), 1355-1370

ALGEBRAIC BAYESIAN ANALYSIS OF CONTINGENCY TABLES WITH POSSIBLY ZERO-PROBABILITY CELLS

Guido Consonni and Giovanni Pistone

Università di Pavia and Politecnico di Torino

Abstract: In this paper we consider a Bayesian analysis of contingency tables allowing for the possibility that cells may have probability zero. In this sense we depart from standard log-linear modeling that implicitly assumes a positivity constraint. Our approach leads us to consider mixture models for contingency tables, where the components of the mixture, which we call model-instances, have distinct support. We rely on ideas from polynomial algebra in order to identify the various model instances. We also provide a method to assign prior probabilities to each instance of the model, and we describe methods for constructing priors on the parameter space of each instance. We illustrate our methodology through a 5×2 table involving two structural zeros, as well as a zero count. The results we obtain show that our analysis may lead to conclusions that are substantively different from those that would obtain in a standard framework, wherein the possibility of zero-probability cells is not explicitly accounted for.

Key words and phrases: Algebraic statistics, Bayes factor, compatible priors, exponential family, log-linear model, model-instance, positivity constraint, structural zero, toric model.

1. Introduction

The analysis of contingency tables has a well-established tradition, both in the frequentist and Bayesian setting. A typical framework for this analysis is represented by the exponential family representation of the sampling distribution, together with the log-linear, or more generally log-affine, model for the expected cell count, see Lauritzen (1996, Chap. 4) for a rigorous treatment. Under multinomial sampling, this approach presupposes implicitly that cell-probabilities, equivalently cell-expected counts, are strictly positive. On the other hand, this assumption is not particularly justified from a substantive viewpoint; indeed, as we argue below, it might well hide some interesting aspects of modeling.

Typically, the positivity constraint is viewed as problematic when performing Maximum Likelihood Estimation (MLE) in a log-linear framework if there are some cells having zero counts, see for instance the discussion in Christensen (1997, Chap. 8). One usually distinguishes between *structural* (or "fixed") zeros,

and *random* (or "sampling") zeros. The former arise when the cells are logically forced to have a zero-count. Consider for instance a cross-classification for people where the personal highest educational attainment (Less than high school, High school, College, Postgraduate) is recorded at a given time, and five years later. Clearly it is impossible for someone to have a highest attainment of College on the first time point, and Less than high school or High school five years later; in general every cell that corresponds to lower attainment at the second time period compared to the first time period is a structural zero. On the other hand, random zeros typically occur because the sample size or the corresponding cell probability, or both, are "small".

Structural zeros are typically dealt with by removing them from the analysis. One way to do this is through regression models on effect codings, see e.g., Simonoff (2003, Sec. 6.4). Random zeros, on the other hand, require special handling. Essentially one should first identify those cells for which the regular MLE of the cell-probability does not exist, i.e., is zero (this requires special care as such cells need not coincide with those having zero counts), and then remove them from the analysis. In any case the computation of the degrees of freedom for model testing must be done on a case by case basis, and this requires some ingenuity. Another difficulty generated by the presence of random zeros is that asymptotic arguments may effectively break down because of the small-sample size, although some computer programs may still provide MLEs when they actually do not exist. For an informative account of the above problems, see Haberman (1974), Bishop, Fienberg and Holland (1975, Sec. 5) and Christensen (1997, Sec. 8.3). Recently Eriksson, Fienberg, Rinaldo and Sullivant (2006) have provided a polyhedral description of the conditions for the existence of the MLE for a hierarchical log-linear model, together with an algorithm for determining if the MLE exists.

In this paper we take the view that the modeling of contingency tables should allow explicitly for the possibility of zero-probability cells, not only to deal with structural zeros, but also with zero-counts whose nature is undecided in the sense that their occurrence may be consistent with either a zero probability or a positive probability. We call the latter cells *possibly zero-probability* cells.

An early paper that takes a similar view is Lauritzen (1975), although the techniques used there are quite different from the ones that we employ here.

From a modeling perspective, we contend that, for each given model, the usual exponential-family/log-linear representation of the sampling distribution is simply one *instance* of such model, while several other instances are conceptually consistent with the assumed model, each being essentially a log-linear model with a restricted support. The identification of such instances represent a crucial aspect in the implementation process, and typically is of high complexity.

In our work we rely on ideas, from polynomial algebra and the related geometric and combinatorial structure, which have been recently applied to the analysis of some classes of (finitely) discrete statistical models. In particular, Eriksson, Fienberg, Rinaldo and Sullivant (2006) deals with hierarchical log-linear models, Geiger, Meek and Sturmefels (2006) discusses graphical models.

Our approach falls broadly under the heading of *Algebraic Statistics*, see Pistone, Riccomagno and Wynn (2001) for an early general account, as well as the pioneering work of Diaconis and Sturmfels (1998). The field is now growing at an impressive speed both in terms of theoretical contributions and applications, see for example the recent monograph by Pachter and Sturmfels (2005). Further useful references are Geiger, Heckerman, King and Meek (2001), who develop the concept of stratified exponential families, as well as Garcia, Stillman and Sturmfels (2005), who carry out the analysis of Bayesian networks from an algebraic statistical perspective. Rapallo (2006) discusses some basic algebraic statistical tools that deal explicitly with models for contingency tables and is a simple and useful introduction to this paper. Our interest in the use of algebraic methodology for statistical purposes was stimulated by the availability of various symbolic computational software: here we use CoCoA, developed and maintained at the University of Genova, Italy. Another option is the softare 4ti2.

A specific feature of this paper is the combination of methods from algebraic statistics with the Bayesian approach. Specifically, we deal with issues like the assignment of a prior on model space, prior elicitation on the parameter space under each model, or instances thereof, together with model choice using the Bayes factor, see Kass and Raftery (1995) for a review.

The paper is organized as follows: Section 2 contains some basic tools from algebraic statistics that are used in the paper; in Section 3 such tools are applied to a data-set; Section 4 is the core of the paper, presenting a Bayesian approach to testing quasi-independence in two-way contingency tables using a mixture of model-instances, thus accounting for the possible presence of zero-probability cells. Finally, Section 5 summarizes the paper and presents some points for discussion.

2. Algebraic Statistical Models

Consider a finite state space $\mathcal{Q}$ and a probability distribution on $\mathcal{Q}$, which we write as $\{p(x),\, x \in \mathcal{Q}\}$, with $p(x) \geq 0$ and $\sum_{x\in\mathcal{Q}} p(x) = 1$. In particular, we deal with multi-way contingency tables identified by a collection of factors $X = \{X_1, \ldots, X_F\}$. If $\mathcal{I}_f$ denotes the set of levels for the factor X_f, $f = 1, \ldots, F$, the state space is a product space, i.e., $\mathcal{Q} = \times_{f=1}^{F} \mathcal{I}_f$.

A log-linear model assumes that $p(x) > 0$ and that $\log p(x)$ belongs to a linear subspace H of $L = \mathbb{R}^{\mathcal{Q}}$, where $\mathbb{R}^{\mathcal{Q}}$ denotes the vector space of real-valued

functions on $\mathcal{Q}$. If H is spanned by $\{T_1, \dots, T_s\}$, where the T_j's are integer valued functions, we can write the log-linear model as

$$\log p(x) = \sum_{j=1}^{s} (\log \zeta_j) T_j(x), \tag{2.1}$$

with $\sum_x p(x) = 1$. Recall that (2.1) assumes strict positivity of $p(x)$. However, the latter is no longer needed if we rewrite (2.1) as

$$q(x) = \zeta_1^{T_1(x)} \cdots \zeta_s^{T_s(x)}, \quad \zeta_j \geq 0, \quad j = 0, \dots, s, \tag{2.2}$$

where $q(x)$ is the un-normalized probability, so that the parameters $\zeta_1, \dots, \zeta_s$ are only subject to non-negativity constraints. Notice that (2.2) is, for each $x \in \mathcal{Q}$, a (monic) monomial in the indeterminates $\zeta_1, \dots, \zeta_s$. When x scans $\mathcal{Q}$, we get a system of binomial equations and so (2.2) could also be called a parametric *toric* model, borrowing terminology from commutative algebra, see Sturmfels (1996), as suggested in Pistone, Riccomagno and Wynn (2001).

When the cell probabilities are assumed to be strictly positive, then the log-linear model (2.1) and the toric model (2.2) can be easily shown to be equivalent. A third expression of the same model can be derived by elimination of the indeterminates $\zeta_1, \dots, \zeta_s$ in the monomial parameterization of (2.2). In fact, if $M = \left[T_1(x) \cdots T_s(x)\right]_{x \in \mathcal{Q}}$ is the design matrix of the log-linear model (2.1), the orthogonal space of its range can be generated by integer valued vectors with zero sum $K = \left[k_1 \cdots k_r\right]$, and (2.2) gives for each $j = 1, \dots, r$,

$$\prod_x q(x)^{k_j(x)} = \prod_x \left(\zeta_1^{T_1(x)} \cdots \zeta_s^{T_s(x)}\right)^{k_j(x)} = \zeta_1^{T_1(x) \cdot k_j(x)} \cdots \zeta_s^{T_s(x) \cdot k_j(x)} = 1, \tag{2.3}$$

where the dot symbol "$\cdot$"denotes scalar product.

As the sum of the elements of each k_j, $j = 1, \dots, r$, is zero, the sum of the elements of both the positive part k_j^+ and the negative part k_j^- are equal, so that we could write equation (2.3) as

$$\prod_x q(x)^{k_j(x)^+} - \prod_x q(x)^{k_j(x)^-} = 0, \qquad j = 1, \dots, r. \tag{2.4}$$

It follows that the toric model (2.2) implies a set of r binomial and homogeneous equations in the un-normalized probabilities $q(x)$, $x \in \mathcal{Q}$.

If the probabilities are assumed to be strictly positive, then the three descriptions, i.e., log-linear (2.1), toric (2.2), and implicit binomial (2.4), are equivalent. We remark that while (2.1) and (2.2) are parametric models, the nature of (2.4) is essentially non-parametric. When the positivity assumption is relaxed, a non-

trivial situation occurs. The basic fact is that different toric parameterizations can lead to the same implicit binomial, because they are equivalent only on the strictly positive part of the model. However, the implicit binomial equations are satisfied by all limits of the positive cases; thus the implicit binomial is the best expression of the so-called extended exponential model, i.e., the exponential model plus all its limits.

We summarize here a few basic facts of the theory of toric statistical models. Given a log-linear model and all its limit points, a specific set of configurations of zero-probability cells arises. This set cannot be recovered by setting to zero some parameters in a generic toric parametric representation, because most of the equivalent toric representations will not produce all possible probabilities of the model in (2.4). However, there exists a "maximal" parametric toric representation such that all configurations of zero-probability cells compatible with, i.e., limit of, the initial model are obtained by letting some parameters be zero. Such a representation results from the following steps.

1. All toric models compatible with the implicit binomial model (2.4) are characterized by a string of T exponents, see (2.2), which is a non-negative integer vector orthogonal to the basis $[k_1 \dots k_r]$ of the orthogonal space of the initial design matrix M.
2. The lattice of non-negative integer vectors $t \in \mathbb{N}_+^{\mathcal{Q}}$, such that the condition $t \cdot k_j = 0$ holds for each $j = 1, \dots, r$, has a finite number of generators that can be computed with symbolic software. Here "generator" means that all such vectors are component-wise sums of a finite number of generators, possibly repeated. The minimal set of generators is called the minimal Hilbert basis.
3. If the generators are $S_1, \dots, S_u$, then the "maximal" toric model is

$$q(x) = \zeta_1^{S_1(x)} \cdots \zeta_u^{S_u(x)} \quad x \in \mathcal{Q}. \tag{2.5}$$

Here "maximal" means that (2.5) is a (possibly non-identifiable) parameterization of the full implicit binomial model, i.e., the extended model. All members of the implicit model (2.4) with zero-cell probabilities are obtained by letting some ζ_j's be zero. For example, let $\zeta_1 = 0$. Then the support of the resulting probability will be the set $\mathcal{Q}_1 = \{x \in \mathcal{Q} : S_1(x) = 0\}$. On such a restricted support, the model will again be toric:

$$q(x) = \zeta_2^{S_2(x)} \cdots \zeta_u^{S_u(x)}, \quad x \in \mathcal{Q}_1,$$

or exponential if all the other parameters $\zeta_2, \dots, \zeta_u$ are assumed to be strictly positive. In this sense, we say that each toric model is a union of exponential models with different supports. Each one of these models is called an *instance* of the model.

Current symbolic software allows one to compute, for a given parametric model, the set of corresponding implicit binomial descriptions. Moreover, the collection of allowable models obtained by setting some cell probabilities equal to zero can be identified in terms of the functions $T_j(x)$, see Geiger, Meek and Sturmefels (2006) and Rapallo (2006).

3. Example: New Cancer Incidence and Gender

We now turn to the discussion of an example involving both structural and random zeros. Our analysis aims primarily at illustrating the main features of our method.

The Division of Cancer Prevention and Control of the National Cancer Institute in the United States provides (estimates of) counts of new cases of cancer classified according to various demographic and geographic factors, see Simonoff (2003, p.226). The following table reports data for different types of cancer, separated by gender, for Alaska in the year 1989.

Type of cancer	Female	Male	Total
Lung	38	90	128
Melanoma	15	15	30
Ovarian	18	*	18
Prostate	*	111	111
Stomach	0	5	5
Total	71	221	292

Clearly cells $(3,2)$ and $(4,1)$ are structural zeros, while we regard the zero count corresponding to the combination (Stomach, Female) as a possibly zero-probability cell. An assumption of interest in this case is that of *quasi-independence* (QI), corresponding to the standard independence assumption for all cells, excluding those having a structural zero. For this hypothesis, Simonoff (2003, p. 228) finds a p-value between 2% and 3%, depending on the method that is employed. Using a conventional frequentist interpretation, the data thus seem to provide significant evidence against the QI-model, although this evidence is not very strong.

Let $I = \{1,2,3,4,5\}$, $J = \{1,2\}$ denote the set of levels for the rows and columns respectively, and consider the two-way table with cells in the set $A = I \times J \setminus \{(3,2),(4,1)\}$.

Under the QI-model the un-normalized cell probabilities q_{ij} are given by

$$q_{ij} = \rho_i \psi_j, \quad (i,j) \in A. \tag{3.1}$$

If the probabilities are strictly positive, one can write $\log q_{i,j} = \alpha_i + \beta_j$, $(i,j) \in A$, with $\alpha_i = \log \rho_i$, $\beta_j = \log \psi_j$. Accordingly the design matrix M, together with a

suitable choice of an orthogonal matrix K, as described in Step 1 of Section 2, are

$$M = \begin{array}{c} \\ 11 \\ 21 \\ 31 \\ 51 \\ 12 \\ 22 \\ 42 \\ 52 \end{array} \begin{array}{c} \begin{array}{ccccccc} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & \beta_1 & \beta_2 \end{array} \\ \left[\begin{array}{ccccccc} 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{array}\right] \end{array} \qquad K = \begin{array}{c} \\ 11 \\ 21 \\ 31 \\ 51 \\ 12 \\ 22 \\ 42 \\ 52 \end{array} \begin{array}{c} \begin{array}{cc} k_1 & k_2 \end{array} \\ \left[\begin{array}{cc} 1 & 0 \\ -1 & -1 \\ 0 & 0 \\ 0 & 1 \\ -1 & 0 \\ 1 & 1 \\ 0 & 0 \\ 0 & -1 \end{array}\right] \end{array}.$$

One can check that, under the condition $q_{ij} > 0$, $(i,j) \in A$, the model of quasi-independence in (3.1) is equivalent to the implicit binomial model given by the two constraints

$$\begin{cases} q_{11}q_{22} - q_{21}q_{12} = 0 \\ q_{51}q_{22} - q_{21}q_{52} = 0. \end{cases} \tag{3.2}$$

The above equations are the standard conditions for independence in the two 2×2 tables with rows $\{1,2\}$, respectively $\{2,5\}$. This is equivalent to the independence of the sub-table $\{1,2,5\} \times \{1,2\}$, since independence for an $R \times C$-table is equivalent to its 2×2 minors being zero.

The maximal design matrix $M_{\max}$ and the model in monomial form, see (2.5), are

$$M_{\max} = \begin{array}{c} \\ 11 \\ 21 \\ 31 \\ 51 \\ 12 \\ 22 \\ 42 \\ 52 \end{array} \begin{array}{c} \begin{array}{ccccccc} \zeta_1 & \zeta_2 & \zeta_3 & \zeta_4 & \zeta_5 & \zeta_6 & \zeta_7 \end{array} \\ \left[\begin{array}{ccccccc} 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{array}\right] \end{array} \qquad \begin{cases} q_{11} = \zeta_5\zeta_7 \\ q_{21} = \zeta_3\zeta_7 \\ q_{31} = \zeta_1 \\ q_{51} = \zeta_4\zeta_7 \\ q_{12} = \zeta_5\zeta_6 \\ q_{22} = \zeta_3\zeta_6 \\ q_{42} = \zeta_2 \\ q_{52} = \zeta_4\zeta_6 \end{cases}. \tag{3.3}$$

Notice that the cells associated to a structural zero in the same row are parameterized independently from the rest of the table. If we take out these cells, we simply get the full independence model on the sub-table with rows $\{1,2,5\}$.

The instances for the QI-model are computed by considering the $(2^3-1)(2^2-1) = 21$ instances corresponding to independence in the 3×2 sub-table, times

the $2^2 = 4$ instances of the two free cells, plus the $(2^2 - 1)$ instances where the 3×2 sub-table is zero. The total is 87.

4. Testing Quasi-Independence in the New Cancer Data

We provide a Bayesian analysis of these data using the methodology developed in the previous sections. We refer to the model which imposes no restriction on the cell probabilities, save the zero-probability cells $(3, 2)$ and $(4, 1)$, as the Structural Zero model and label it as SZ. Since the table has ten probability cells, of which two are fixed to be zero, the number of SZ-instances is equal to $2^8 - 1 = 255$, corresponding to all possible combinations of "+" and "0" in the 8 free cells, excluding the trivially impossible case of all "0".

Moreover, only two of the above SZ-instances are logically consistent with the observed data: that giving a positive probability to all eight free cells, and that giving zero-probability to cell $(5, 1)$ only. We label these instances SZ_0 and SZ_1, where the subscript refers to the number of zero-probability cells, corresponding to the tables

	SZ_0		SZ_1	
Type of cancer	Female	Male	Female	Male
Lung	+	+	+	+
Melanoma	+	+	+	+
Ovarian	+	0	+	0
Prostate	0	+	0	+
Stomach	+	+	0	+

Similarly, for the given data, it is not difficult to realize that there exists only one logically consistent instance of the quasi-independence model, i.e., that having all positive cell-probabilities (except for the two cells corresponding to structural zeros), which we label QI_0; it is schematically equivalent to SZ_0 above.

4.1. Conventional approach

We test the model of quasi-independence against the structural-zero model using a Bayesian approach. In a "conventional setting", wherein no particular provision for zero-probability cells is envisaged, we would simply consider one instance for each of the above two models, namely SZ_0 and QI_0.

Given the cell counts $n = (n_{ij})$, a typical analysis would involve the computation of the Bayes factor, see Kass and Raftery (1995), of QI_0 *versus* SZ_0, i.e.,

$$\mathrm{BF}(QI_0 : SZ_0) = \frac{\int f_{QI_0}(n|\theta_{QI_0})\pi_{QI_0}(\theta_{QI_0})d\theta_{QI_0}}{\int f_{SZ_0}(n|\theta_{SZ_0})\pi_{SZ_0}(\theta_{SZ_0})d\theta_{SZ_0}} = \frac{m_{QI_0}(n)}{m_{SZ_0}(n)}, \tag{4.1}$$

where

- f_{SZ_0} is the multinomial sampling distribution under SZ_0, with cell-probabilities $\theta_{SZ_0} = (\theta_{ij})$, $(i,j) \in A$, and similarly for f_{QI_0} under the quasi-independence model, whose cell-probabilities are denoted by θ_{QI_0};
- π_{SZ_0} and π_{QI_0} are the prior densities for θ_{SZ_0}, respectively θ_{QI_0};
- m_{SZ_0} denote the marginal distribution of n under SZ_0, and similarly for m_{QI_0}.

To obtain the posterior probability of model QI_0 one should provide, in addition, its prior probability $p_{QI_0} = \Pr(QI_0)$, leading to

$$\Pr(QI_0|n) = \frac{p_{QI_0}\mathrm{BF}(QI_0 : SZ_0)}{p_{QI_0}\mathrm{BF}(QI_0 : SZ_0) + p_{SZ_0}}, \tag{4.2}$$

where $p_{SZ_0} = \Pr(SZ_0) = 1 - p_{QI_0}$.

A Bayesian analysis of this problem might take the prior π_{SZ_0} to be Dirichlet, i.e.,

$$\theta_{SZ_0} \sim \mathrm{Di}(\alpha), \tag{4.3}$$

with $\alpha = (\alpha_{ij})$ and $\alpha_{ij} > 0$, see e.g., Bernardo and Smith (1994, pp.134-135 and 441)) and O'Hagan and Forster (2004, Chap.12). As a consequence, $m_{SZ_0}(n)$ is a Multinomial-Dirichlet with distribution

$$m_{SZ_0}(n) = \frac{N!}{\prod_{(i,j)\in A} n_{ij}!} \times \frac{H_A(\alpha)}{H_A(\alpha^*)}, \qquad n = (n_{ij}), \quad n_{ij} = 0, 1, \ldots, N, \quad \sum_{i,j} n_{ij} = N, \tag{4.4}$$

where

$$H_T(y) = \frac{\Gamma(\sum_{t\in T} y_t)}{\prod_{t\in T}\Gamma(y_t)},$$

and $\alpha^* = \alpha + n$.

Consider now the quasi-independence model QI_0 and, in particular, the choice of the prior π_{QI_0}. This presents some conceptual and practical challenges that we now try to elucidate. Although, in principle, priors under distinct models need not be related, as they express prior beliefs conditionally on different states of information, it is nevertheless desirable that they should be related at least when models are nested within an encompassing model. Pragmatically, this would simplify the elicitation task, since one would only assign a prior on the parameter under the latter model, and then derive the corresponding priors under each of the remaining models from this single prior. This procedure should also

achieve some sort of internal "compatibility" among prior specifications. A general discussion of strategies for building compatible priors under several related models is contained in Dawid and Lauritzen (2001). Further discussion, elaboration and references may be found in Consonni, Gutiérrez-Peña and Veronese (2007), and in Consonni and Veronese (2006).

Before turning to model QI_0, it is expedient to rewrite the joint distribution of the counts n_{ij}, $(i,j) \in A$, for the SZ_0-model as

$$f_{SZ_0}(n|\theta) = f_{SZ_0,1}(n_{(1)}|\theta) \times f_{SZ_0,2}(n_{(2)}|n_{(1)},\theta), \tag{4.5}$$

where $n_{(1)} = (n_{31}, n_{42}, N - n_{31} - n_{42})$ and $n_{(2)} = (n_{ij} : (i,j) \in A \setminus \{(3,1),(4,2)\})$. Since, for $(i,j) \in A$, the joint distribution of $n = (n_{ij})$, under SZ_0, is multinomial with size N and vector of probabilities $\theta = (\theta_{ij})$, written $\mathrm{Mu}(N;\theta)$, it is easy to check that $f_{SZ_0,1}(n_{(1)}|\theta)$ is a $\mathrm{Mu}(N;\lambda)$ with $\lambda_1 = \theta_{31}$, $\lambda_2 = \theta_{42}$, $\lambda_3 = 1 - \lambda_1 - \lambda_2$, while $f_{SZ_0,2}(n_{(2)}|n_{(1)},\theta)$ is given by $\mathrm{Mu}(N - n_{31} - n_{42};\gamma)$, where

$$\gamma_{ij} = \frac{\theta_{ij}}{1 - \theta_{31} - \theta_{42}} = \frac{\theta_{ij}}{\sum_{(i,j)\in A\setminus\{((3,1),(4,2))\}}}, \quad (i,j) \in A \setminus \{((3,1),(4,2))\}.$$

The parameters λ and γ are variation independent, i.e., their joint range is the product of the two individual ranges.

Under QI_0 we must have

$$\gamma_{ij} = \gamma_{i+}\gamma_{+j}, \quad (i,j) \in A \setminus \{(3,1),(4,2)\},$$

where $\gamma_{i+} = \gamma_{i1} + \gamma_{i2}$, $i = 1,2,5$, $\gamma_{+j} = \gamma_{1j} + \gamma_{2j} + \gamma_{5j}$, $j = 1,2$.

Let γ_R denote the collection of γ_{i+}, and γ_C that of γ_{+j}. Then the distribution of the counts n under QI_0 can be written as

$$f_{QI_0}(n|\lambda,\gamma_R,\gamma_C) = f_{QI_0,1}(n_{(1)}|\lambda) f_{QI_0,2}(n_{(2)}|n_{(1)};\gamma_R,\gamma_C), \tag{4.6}$$

where $f_{QI_0,1}(n_{(1)}|\lambda)$ is $\mathrm{Mu}(N;\lambda_1,\lambda_2,\lambda_3)$, and so coincides with the expression of $f_{SZ_0,1}(n_{(1)}|\theta)$ in (4.5), while $f_{QI_0,2}(n_{(2)}|n_{(1)},\gamma_R,\gamma_C)$ is given by

$$\begin{aligned} &f_{QI_0,2}(n_{(2)}|n_{(1)};\gamma_R,\gamma_C) \\ &\quad = \frac{(N - n_{31} - n_{42})!}{\prod_{(i,j)\in A\setminus\{(3,1),(4,2)} n_{ij}!\}} \times \gamma_{1+}^{n_{1+}} \gamma_{2+}^{n_{2+}} \gamma_{5+}^{n_{5+}} \times \gamma_{+1}^{\tilde{n}_{+1}} \gamma_{+2}^{\tilde{n}_{+2}}, \end{aligned} \tag{4.7}$$

where $\tilde{n}_{+j} = n_{1j} + n_{2j} + n_{5j}$.

One sees that under QI_0 the joint distribution factors into three terms, one involving λ, one involving γ_R and one involving γ_C.

Consider now the prior distribution. Given that $\theta_{SZ_0} \sim \mathrm{Di}(\alpha_{SZ_0})$ we first remark that λ and γ, are independent, because of (ii) of Lemma 1, see the

Appendix; as a consequence we also get that λ is independent of the pair (γ_R, γ_C). Furthermore $\gamma \sim Di(\alpha_{ij},\ (i,j) \in A \setminus \{(3,1),(4,2)\})$, so that $\gamma_R \sim \mathrm{Di}(\alpha_R)$ and $\gamma_C \sim \mathrm{Di}(\alpha_C)$, where α_R and α_C are defined in accordance with γ_R and γ_C, respectively. Assuming independence of γ_R and γ_C makes the computation of the marginal distribution $m_{QI_0}(n)$ straightforward since we can separately integrate the three terms in (4.6), see also (4.7), each integral being, up to the multinomial coefficient, of Multinomial-Dirichlet type.

Specifically we get

$$m_{QI_0}(n) = \frac{N!}{\prod_{(i,j)\in A} n_{ij}!} \times \frac{H(\alpha_{31}, \alpha_{42}, \alpha_+ - \alpha_{31} - \alpha_{42})}{H(\alpha^*_{31}, \alpha^*_{42}, \alpha^*_+ - \alpha^*_{31} - \alpha^*_{42})} \times \frac{H(\alpha_{1+}, \alpha_{2+}, \alpha_{5+})}{H(\alpha^*_{1+}, \alpha^*_{2+}, \alpha^*_{5+})} \times \frac{H(\tilde{\alpha}_{+1}, \tilde{\alpha}_{+2})}{H(\tilde{\alpha}^*_{+1}, \tilde{\alpha}^*_{+2})}, \quad (4.8)$$

where $\alpha_+ = \sum_{(i,j)\in A} \alpha_{ij}$, $\tilde{\alpha}_{+j} = \alpha_{1j} + \alpha_{2j} + \alpha_{5j}$, $\tilde{\alpha}^*_{+j} = \alpha_{1j} + n_{1j} + \alpha_{2j} + n_{2j} + \alpha_{5j} + n_{5j}$.

4.2. Allowing for zero-probability cells

Philosophically, we stress the view that each instance of a model must be assigned a positive probability *a-priori*; in this sense we adhere to the principle that Lindley (1985, p.104) names the "Cromwell's rule". This leads us naturally to the idea of regarding a model $\mathcal{M}$ as a finite *mixture* of its instances. This aspect represents a characterizing feature of our approach to the analysis of contingency tables.

We can thus write the mixture representation of $\mathcal{M}$ as

$$f_{\mathcal{M}}(n|\theta_{\mathcal{M}}) = \sum_h q_{\mathcal{M}_h} f_{\mathcal{M}_h}(n|\theta_{\mathcal{M}_h}), \quad (4.9)$$

where $\theta_{\mathcal{M}}$ is the collection of all instance-specific parameters $\theta_{\mathcal{M}_h}$ and $q_{\mathcal{M}_h}$ is the prior probability attached to instance $\mathcal{M}_h$.

Specializing (4.9) to the SZ and QI model, and then computing the marginal distribution of the data under each model, leads to the Bayes factor

$$\mathrm{BF}(QI : SZ) = \frac{q_{QI_0} m_{QI_0}(n)}{q_{SZ_0} m_{SZ_0}(n) + q_{SZ_1} m_{SZ_1}(n)}. \quad (4.10)$$

We consider in detail the computations that are needed for the evaluation of $\mathrm{BF}(QI : SZ)$. Let $\xi \in (0,1)$ be the chance that a cell has zero probability, and assume that the allocation of zero probability to each cell takes place

independently. Then, we can derive q_{SZ_0} and q_{SZ_1}, and obtain

$$q_{SZ_0} = \frac{(1-\xi)^8}{1-\xi^8}, \tag{4.11}$$

$$q_{SZ_1} = \frac{\xi(1-\xi)^7}{1-\xi^8}. \tag{4.12}$$

Consider now the assignment of q_{QI_0}. We recall that we have 87 instances with total probability $C(\xi)$, so

$$q_{QI_0} = \frac{(1-\xi)^8}{C(\xi)}. \tag{4.13}$$

Table 4.1 reports the value of $q_{SZ_0}, q_{SZ_1}, q_{QI_0}$ for selected choices of ξ (for values of ξ above 0.5, the values are zero to two decimal places).

Table 4.1. Prior probabilities q_{SZ_0}, q_{SZ_1}, q_{QI_0} for selected values of ξ.

ξ	q_{SZ_0}	q_{SZ_1}	q_{QI_0}
0.1	0.43	0.05	0.78
0.2	0.17	0.04	0.51
0.3	0.06	0.03	0.23
0.4	0.02	0.01	0.07
0.5	0.00	0.00	0.01

Consider the marginal distribution of the data under the SZ_1-instance. The conditioning method of Lemma 1, item (ii), leads immediately to $\theta_{SZ_1} \sim \mathrm{Di}(\alpha_{SZ_1})$, where $\alpha_{SZ_1} = (\alpha_{ij},\, (i,j) \in A \setminus \{(5,1)\})$, whence m_{SZ_1} has an expression analogous to that of m_{SZ_0}, the only difference being that now the set over which the indexes vary is $A \setminus \{(5,1)\}$.

For given ξ and α, the Bayes factor $\mathrm{BF}(QI : SZ)$ can be computed using (4.10). Notice that the multiplicative term $N!/\prod_{(ij)\in A} n_{ij}!$ appears both in the numerator and denominator of (4.10), and so cancels out (strictly speaking the product for the instance SZ_1 is over a set that does not contain $(5,1)$; however, since $n_{51} = 0$, the result is the same whether this value appears or not).

Consider first the assignment of ξ, which represents the chance that a cell has probability zero. Save for the case of a structural zero, it seems reasonable that we assign a low value to ξ, since the corresponding event should be regarded *a priori* as a rather unusual circumstance. In view of Table 4.1, setting $\xi = 0.1$ seems a sensible choice. Indeed, while the prior probability of model QI is higher than that of SZ, nevertheless the discrepancy between the two values (0.78 against $0.48 = 0.43 + 0.05$) is less pronounced for this choice of ξ than for other choices. Thus, the comparison between the two models is fairer.

We now take into consideration the choice of α. Unless there exists substantive prior information allowing one to discriminate *a-priori* between cells, we choose the same value $\bar{\alpha}$ for each α_{ij}; low values of $\bar{\alpha}$ are recommended when prior information is weak. Natural choices are represented by $\bar{\alpha} = 0.5$, corresponding to the Jeffreys prior, or $\bar{\alpha} = 1$, corresponding to a uniform prior on the simplex.

We now provide a method for the choice of $\bar{\alpha}$, using the technique of the *imaginary training sample*. This method has been implemented for instance by Spiegelhalter and Smith (1980) to deal with model choice using improper priors. We believe, however, that the idea can be usefully applied also in the context of proper priors, see Consonni, Gutiérrez-Peña and Veronese (2007) for elaboration.

Consider for simplicity only the models SZ_0 and QI_0. Suppose we can identify a *minimal imaginary training sample* that provides *maximal* support (irrespective of the prior) to model QI_0. Then it is reasonable to require that the Bayes factor for these fictitious data should be approximately 1, i.e., the models are "equally likely" in terms of the empirical evidence. To see why this should be the case, notice that on the one hand the data actually support QI_0 very strongly and, on the other hand, the sample size is so small that the evidence in favor of either model should be roughly the same. The condition that the Bayes factor should be 1 can be employed to select reasonable values for the hyper-parameters of the prior distribution.

Consider the situation in which we have one observation in each cell, for a total of eight observations. It is straightforward to verify that this table is perfectly consistent with the QI_0-model: in particular the actual and fitted counts (the latter based on ML estimates) coincide. If we fix $\xi = 0.1$ as suggested above, the value $\bar{\alpha} = 1$ provides a Bayes factor equal to 1.03, which is quite satisfactory; on the other hand $\bar{\alpha} = 0.5$ would give a BF equal to 0.67. We also experimented with other values of $\bar{\alpha}$ and did not get values of BF close to 1.

Having set $\xi = 0.1$ and $\bar{\alpha} = 1$, we now proceed to the analysis of the cancer data. The Bayes factor of QI against SZ is 0.17, which clearly does not support the hypothesis of quasi-independence. To better assess this value, it is useful to derive the Bayes factor *against* QI, which is merely the reciprocal of the above, and to further transform it using the logarithm in base 10. In this way we can make use of the scale developed by Jeffreys, see Kass and Raftery (1995) and Robert (2001, p.228), for the interpretation of the evidence provided by a Bayes factor. Specifically, the evidence *against* QI is

- *poor* if $0 < \log_{10} \mathrm{BF}(SZ : QI) < 0.5$,
- *substantial* if $0.5 < \log_{10} \mathrm{BF}(SZ : QI) < 1$,
- *strong* if $1 < \log_{10} \mathrm{BF}(SZ : QI) < 2$,
- *decisive* if $\log_{10} \mathrm{BF}(SZ : QI) > 2$,

where $\mathrm{BF}(SZ : QI) = 1/\mathrm{BF}(QI : SZ)$. We find $\log_{10}(1/0.17) = 0.77$, which thus represents *substantial* evidence against QI, essentially in accord with the frequentist answer which states a p-value between 2% and 3%. It is instructive to verify what would have been the result of a conventional Bayesian analysis based on the positive-cell models SZ_0 and QI_0, as opposed to the model based on mixtures developed in this paper. Recall that, in the standard case, the BF would simply be the ratio $m_{QI_0}(n)/m_{SZ_0}(n)$. In this case the BF takes the value 0.55, which is appreciably higher than the value 0.17 obtained with our analysis. More interestingly, when translated to the Jeffreys scale, we obtain $\log_{10}(1/0.55) = 0.26$ which only represents *poor* evidence against QI, an order of magnitude lower on the Jeffreys scale than the one we obtained with our analysis.

5. Discussion

We have presented a new methodology for the Bayesian analysis of contingency tables that allows explicitly for the possibility of zero-probability cells.

In order to apply our algebraic Bayesian approach to large and sparse contingency tables, we realize that a purely "automated" approach can be expected to run into serious computational issues. Still, technology is rapidly evolving in this area, as evidenced for instance, within the field of Maximum Likelihood Estimation in the recent paper by Eriksson, Fienberg, Rinaldo and Sullivant (2006); see also Patcher and Sturmfels (2005) for a variety of high-dimensional applications. A careful choice of prior distribution is often the only sensible way to make the analysis viable, see for instance Diaconis and Rolles (2006) in the context of Markov chains with forced zeros. We therefore believe that a blend of computational algebraic methods and prior information on the set of possibly-zero probability cells is likely to be the best option for the analysis of moderate to large multi-way tables.

Appendix

We summarize below some useful facts about the Dirichlet distribution, see e.g., Bernardo and Smith (1994, pp.134-135), but notice that our notation is slightly different from theirs.

Lemma 1. *Let* $\theta = (\theta_1, \ldots, \theta_s)$, *with* $0 < \theta_k < 1$, $k = 1, \ldots, s$, *and* $\sum_{k=1}^{s} \theta_k = 1$. *Assume that* $\theta \sim \mathrm{Di}(\alpha)$, *with* $\alpha = (\alpha_1, \ldots, \alpha_s)$ *and* $\alpha_k > 0$.

(i)
$$\left(\theta_1, \ldots, \theta_r, \sum_{l=r+1}^{s} \theta_l\right) \sim \mathrm{Di}\left(\alpha_1, \ldots, \alpha_r, \sum_{l=r+1}^{s} \alpha_l\right), \quad r < s.$$

(ii) *Let* $\theta'_m = \theta_m / \sum_{q=1}^{r} \theta_q$, $m = 1, \ldots, r$, $r < s$. *Then* $(\theta'_1, \ldots, \theta'_r) \sim \mathrm{Di}(\alpha_1, \ldots, \alpha_r)$, *and* $(\theta'_1, \ldots, \theta'_r)$ *is independent of* $(\theta_{r+1}, \ldots, \theta_s)$.

(iii) *Let* $\theta_1^* = \theta_1 + \ldots + \theta_{i_1}, \ \ldots, \ \theta_t^* = \theta_{i_{t-1}} + \ldots + \theta_s, \quad 1 \leq t < s$. *Then* $(\theta_1^*, \ldots, \theta_t^*) \sim \mathrm{Di}(\alpha_1^*, \ldots, \alpha_t^*)$, $\alpha_1^* = \alpha_1 + \ldots + \alpha_{i_1}, \ldots, \alpha_t^* = \alpha_{i_{t-1}} + \ldots + \alpha_s$.

Acknowledgement

Work partially supported by MIUR, Rome, under the projects PRIN 2003138887 and PRIN 2005132307, by the University of Pavia, the University of Genova and Politecnico of Torino. We thank Simplice Dossou-Gbété and Laboratoire de Mathématiques Appliqués UMR CNRS 5142 at Université de Pau et des Pays de l'Adour for providing hospitality and support while part of this article was written. The second author thanks H.P. Wynn for many discussions and suggestions. A special thanks goes to Persi Diaconis who provided us with thoughtful feedback. Finally, the careful reading and comments by two referees are gratefully acknowledged.

References

4ti2 team. 4ti2 – A software package for algebraic, geometric and combinatorial problems on linear spaces. http://www.4ti2.de.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Wiley, Chichester.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis*. MIT Press. Cambridge, MA.

Christensen, R. (1997). *Log-linear Models and Logistic Regression*. Springer, New York.

CoCoATeam. CoCoA: a system for doing Computations in Commutative Algebra. http://cocoa.dima.unige.it.

Consonni, G. and Veronese, P. (2006). Prior specifications for the comparison of linear models. Submitted.

Consonni, G., Gutiérrez-Peña, E. and Veronese, P. (2007). Compatible priors for Bayesian model comparison with an application to the Hardy-Weinberg equilibrium model. *Test* DOI 10.1007/s11749-007-0057-7.

Dawid, A. P. and Lauritzen, S. L. (2001). Compatible prior distributions. In *Bayesian Methods with Applications to Science, Policy and Official Statistics* (Edited by E. George). Monographs of Official Statistics, 109-118. Office for official publications of the European Communities: Luxembourg. http://www.stat.cmu.edu/ISBA/index.html.

Diaconis, P. and Rolles, S. W. W. (2006). Bayesian analysis for reversible Markov chains. *Ann. Statist.* **34**, 1270-1292.

Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.* **26**, 363-397.

Eriksson, N., Fienberg, S. E., Rinaldo, A. and Sullivant, S. (2006). Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models. *J. Symbolic Comput.* **41**, 222-233.

Garcia, Stillman and Sturmfels (2005). Algebraic geometry of Bayesian networks. *J. Symbolic Comput.* **39**, 331-355.

Geiger, D., Heckerman, D., King, H. and Meek, Ch. (2001). Stratified exponential families: graphical models and model selection, *Ann. Statist.* **29**, 505-529.

Geiger, D, Meek, C. and Sturmefels, B. (2006). On the toric algebra of graphical models. *Ann. Statist.* **34**, 1463-1492.

Haberman, S. J. (1974). *The Analysis of Frequency Data*. University of Chicago Press, Chicago.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773-795.

Lauritzen, S. L. (1975). General exponential models for discrete observations. *Scand. J. Statist.* **2**, 23-33.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press, New York.

Lindley, D. V. (1985). *Making Decisions*. 2nd edition. Wiley, London.

O'Hagan, A. and Forster, J. (2004). *Kendall's Advanced Theory of Statistics*, **2B**. *Bayesian Inference*. 2nd edition. Arnold, London.

Patcher, L. and Sturmfels, B. (2005). *Algebraic Statistics for Computational Biology*. Cambridge University Press, Cambridge, UK.

Pistone, G., Riccomagno, E. and Wynn, H. P. (2001). *Algebraic Statistics: Computational Commutative Algebra in Statistics*. Chapman & Hall, London.

Rapallo, F. (2006). Toric statistical models: parametric and binomial representations. *AISM* DOI 10.1007/s10463-006-0079-z.

Robert, C. P. (2001). *The Bayesian Choice*. 2nd edition. Springer, New York.

Simonoff, J. S. (2003). *Analyzing Categorical Data*. Springer, New York

Spiegelhalter, D. J. and Smith, A. F. M. (1980). Bayes factors and choice criteria for linear models. *J. Roy. Statist. Soc. Ser. B* **42**, 215-220.

Sturmfels, B. (1996). *Gröbner bases and convex polytopes*. American Mathematical Society, Providence, RI.

Dipartimento di Economia Politica e Metodi Quantitativi, Via S. Felice, 7, 27100 Pavia, Italy.

E-mail: guido.consonni@unipv.it

Politecnico di Torino, DIMAT, Corso Duca degli Abruzzi 24, 10129 Torino Italy.

E-mail: giovanni.pistone@polito.it

(Received June 2006; accepted May 2007)