COMPONENT PART NOTICE

THIS PAPER IS A COMPONENT PART OF THE FOLLOWING COMPILATION REPORT:

TITLE: Computing Science and Statistics: Proceedings of the Symposium on Interface Critical Applications of Scientific Computing (23rd): Biology, Engineering, Medicine, Speech Held in Seattle, Washington on 21-24 April 1991.

TO ORDER THE COMPLETE COMPILATION REPORT, USE AD-A252 938.

THE COMPONENT PART IS PROVIDED HERE TO ALLOW USERS ACCESS TO INDIVIDUALLY AUTHORED SECTIONS OF PROCEEDING, ANNALS, SYMPOSIA, ETC. HOWEVER, THE COMPONENT SHOULD BE CONSIDERED WITHIN THE CONTEXT OF THE OVERALL COMPILATION REPORT AND NOT AS A STAND-ALONE TECHNICAL REPORT.

THE FOLLOWING COMPONENT PART NUMBERS COMPRISE THE COMPILATION REPORT:

AD#: AD-P007 096 thru AD#: AD-P007 225

AD#: AD#:

AD#: AD#:

Accesion For	
NTIS CRA&I	☒
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution /	
Availability Codes	
Dist	Avail and / or Special
A-1	

Using Gibbs Sampling for Bayesian Inference in Multidimensional Contingency Tables

Leonardo D. Epstein
Department of Biostatistics
The Johns Hopkins University
Baltimore, Md. 21205

Stephen E. Fienberg
Department of Statistics
Carnegie Mellon University
Pittsburgh, Pa. 15213

Abstract

This paper discusses a method suggested by Epstein and Fienberg (1991) for the Bayesian analysis of multidimensional contingency tables in connection with the Gibbs sampler to calculate posterior densities.

The method consists of a two-stage hierarchical prior. The first stage is a Dirichlet distribution with a loglinear reparametrization for its means. The second stage is a multivariate normal distribution on the loglinear parameters. However, other distributions can be used if the Dirichlet-normal combination is not flexible enough to accomodate one's prior beliefs.

These prior distributions are useful when one believes, with uncertainty, in a given loglinear structure for the cell probabilities.

Key words: Contingency tables; Bayesian estimation; Dirichlet prior distribution; Gibbs sampler; Loglinear model; Maximum likelihood estimation of Dirichlet distributions.

1 Introduction

A new Bayesian method for the analysis of multidimensional contingency tables was recently proposed by Epstein and Fienberg (1988) and Epstein (1990). As with many other Bayesian methods, ours uses the posterior means of the cell probabilities to estimate these parameters. The focus on posterior means is in part due to the importance of point estimation and in part due to computational difficulties in drawing further inferences from the posterior. The purpose of this article is to illustrate with an example how to use the Gibbs sampler to compute estimates of the posterior densities that arise from our method. These density estimates are readily integrable to compute posterior probabilities and moments.

We introduce the esentials of the method via a simple example. Suppose our interest is on inferences about the array of cell probabilities $\theta = \{\theta_{ij}\}$ of a 2×2 contingency table and suppose also that given θ, the observed counts $x = \{x_{ij}\}$ follow a multinomial distribution $M(N, \theta)$. We label the two factors by 1 and 2.

When the data follow a multinomial distribution to model prior beliefs it is common to use the conjugate Dirichlet prior $D(K, \eta)$ with density

$$[\theta|K,\eta] = \beta \prod_{ij} \theta_{ij}^{K\eta_{ij}-1},$$

where $\beta = \Gamma(K)/\prod_{ij}\Gamma(K\eta_{ij})$, and $\eta = \{\eta_{ij}\}$.

Before the observation of x we might believe with some uncertainty that the two factors are independent. That is, we might believe that θ satisfies $\theta_{ij} = \theta_{i+}\theta_{+j}, i = 1,2, j = 1,2$, with some degree of uncertainty.

The condition $\theta_{ij} = \theta_{i+}\theta_{+j}$ is equivalent to

$$\log \theta_{ij} = u + u_{1(ij)} + u_{2(ij)}. \qquad (1)$$

with the restriction that the term u in this equation is

$$u = -\log(\sum_{i,j} \exp(u_{1(ij)} + u_{2(ij)}), \qquad (2)$$

so that $\sum_{i,j}\theta_{ij} = 1$. This normalization leads to an equivalent parametrization that uses the multivariate logits, i.e., if

$$\theta_{ij} = \frac{e^{\gamma_{ij}}}{\sum_{ij} e^{\gamma_{ij}}},$$

then the γ_{ij} are the multivariate logits (see Leonard and Novick, 1976). The parametrization (1) and the normalizing condition on u are equivalent to reparametrizing $\{\gamma_{ij}\}$ using

$$\gamma_{ij} = u_{1(ij)} + u_{2(ij)}.$$

Unless neccessary, the remainder of this paper omits explicit reference to the normalizing role of u. Thus, we will simply speak of the loglinear parametrization $\log \theta_{ij} = u + u_{1(ij)} + u_{2(ij)}$.

To see that the parametrization (1) is equivalent to independence, substitute the value of u back in equation (1) to get

$$\theta_{ij} = \frac{e^{u_{1(ij)}}}{\sum_i e^{u_{1(ij)}}} \times \frac{e^{u_{2(ij)}}}{\sum_i e^{u_{2(ij)}}}.$$

Hence

$$\theta_{i+} = \frac{e^{u_{1(ij)}}}{\sum_i e^{u_{1(ij)}}} \text{ and } \theta_{j+} = \frac{e^{u_{2(ij)}}}{\sum_j e^{u_{1(ij)}}}.$$

To incorporate in the prior our uncertain belief in independence, Epstein (1990) and Epstein and Fienberg (1991) proposed using a loglinear parametrization on the Dirichlet means. That is, to reflect the plausibility that the cell probabilities satisfy (1) they suggest using

$$\log \eta_{ij} = u + u_{1(ij)} + u_{2(ij)}, \tag{3}$$

with $u = -\log(\sum_{i,j} \exp(u_{1(ij)} + u_{2(ij)})$.

The index "1" in $u_{1(ij)}$ indicates that this u-term depends only on the index i. It is more common to omit the indices on which the u-terms do not depend. Thus often we write $u_{1(i)}$ and $u_{2(j)}$ instead of $u_{1(ij)}$ and $u_{2(ij)}$, but the fact that $\{u_{1(ij)}\}$ and $\{u_{2(ij)}\}$ are arrays of the same dimensions as $\{\eta_{ij}\}$ simplifies many formulas.

To establish the connection with the multidimensional case, we note that parametrization (3) maps an array $\{\gamma_{ij}\}$ belonging to the linear subspace $M = \{\{\gamma_{ij}\} : \gamma_{ij} = u_{1(i)} + u_{2(j)}\}$ into the array $\{\exp(\gamma_{ij})/\sum_{ij} \exp(\gamma_{ij})\}$.

The parametrization (3) implies that

$$\eta_{i+} = \frac{e^{u_{1(ij)}}}{\sum_i e^{u_{1(ij)}}} \text{ and } \eta_{+j} = \frac{e^{u_{2(ij)}}}{\sum_j e^{u_{1(ij)}}}. \tag{4}$$

Thus, $\{u_{1(ij)}\}$ and $\{u_{2(ij)}\}$ parametrize the marginal arrays $\{\eta_{i+}\}$ and $\{\eta_{+j}\}$, respectively.

We follow the notation of Andersen (1974) to represent marginal tables, and the definition will be recalled in section 2 more formally. This notation represents the marginal array with entries η_{i+} by $\eta^Y = \{\eta^Y_{ij}\}$, where $Y = \{1\}$. The set of factor labels Y indicates that η^Y depends only on the index corresponding to factor 1, namely i, and that η was collapsed over the indices corresponding to the factors not in Y, namely j. We will also use products of arrays. Thus, for example, the product of $\eta^{\{1\}}$ and $\eta^{\{2\}}$, denoted by $\eta^{\{1\}}\eta^{\{2\}}$, is the array whose (i,j) entry is $\eta^{\{1\}}_{ij}\eta^{\{2\}}_{ij}$, or, in the usual notation, $\eta_{i+}\eta_{+j}$.

The parametrization $\log \eta_{ij} = u + u_{1(ij)} + u_{2(ij)}$ is equivalent to $\eta_{ij} = \eta^{\{1\}}_{ij}\eta^{\{2\}}_{ij}$ with $0 \le \eta^{\{1\}}_{ij} \le 1$, $0 \le \eta^{\{2\}}_{ij} \le 1$, $\eta^{\{1\}}_{1j} + \eta^{\{1\}}_{2j} = 1$, and $\eta^{\{2\}}_{i1} + \eta^{\{2\}}_{i2} = 1$ (see Albert and Gupta, 1982). However, the loglinear parametrization on the Dirichlet means allowed Epstein and Fienberg (1991) and Epstein (1990) to extend the method to multidimensional tables.

If we feel we cannot specify a value for $\eta^{\{1\}}_{ij}$ and $\eta^{\{2\}}_{ij}$, or, equivalently, for $u_{1(ij)}$ and $u_{2(ij)}$, then the Dirichlet distribution cannot adequately represent our prior beliefs. However, as Albert and Gupta (1982) point out, the Dirichlet distribution may still be used as the first stage of a two-stage prior. With a loglinear parametrization for the Dirichlet means there are two equivalent alternative ways to complete the two-stage prior. One may use distributions on the u-terms or one may prefer to specify distributions on $\eta^{\{1\}}_{ij}$ and $\eta^{\{2\}}_{ij}$ directly.

The loglinear parametrization will be more useful when analyzing tables of higher dimensions where one may consider more complex loglinear structures. As the next section explains, with loglinear parametrizations for n-way tables one can also specify the second-stage in two alternative ways, but to use the second one must determine the generating class of the loglinear parametrization and use the margins of η given by the generator as parameters of the Dirichlet distribution.

The parameter K governs the concentration of the prior distribution about the independence surface

$$\begin{aligned} \mathcal{S} = \{\{\theta_{ij}\} | \theta_{ij} &= \theta^{\{1\}}_{ij}\theta^{\{2\}}_{ij}, \\ &0 \le \theta^{\{1\}}_{ij} \le 1, 0 \le \theta^{\{2\}}_{ij} \le 1, \\ &\theta^{\{1\}}_{11} + \theta^{\{1\}}_{12} = 1, \theta^{\{2\}}_{21} + \theta^{\{2\}}_{22} = 1\}. \end{aligned}$$

In the limit, as $K \to \infty$, the prior, and therefore the posterior, concentrate all of their mass on $\mathcal{S}$.

When we use a two-stage prior we obtain the posterior means

$$\varepsilon(\theta_{ij}|\boldsymbol{x}) = \frac{N}{N+K}\frac{x_{ij}}{N} + \frac{K}{N+K}\varepsilon(\eta^{\{1\}}_{ij}\eta^{\{2\}}_{ij}|\boldsymbol{x}), \tag{5}$$

which we use to estimate $\boldsymbol{\theta}$. The expectation $\varepsilon(\eta^{\{1\}}_{ij}\eta^{\{2\}}_{ij}|\boldsymbol{x})$ is with respect to the distribution induced on $\boldsymbol{\eta}^{\{1\}}$ and $\boldsymbol{\eta}^{\{2\}}$ through equations (4).

In most practical situations, when $K \to 0$ the posterior means $\varepsilon(\theta_{ij}|\boldsymbol{x})$ converge to the observed proportions x_{ij}/N. When $K \to \infty$ not only the posterior distribution concentrates the all of its mass on $\mathcal{S}$, but the posterior mean $\varepsilon(\boldsymbol{\theta}|\boldsymbol{x})$ itself belongs to $\mathcal{S}$. This property translates into

$$\lim_{K\to 0} \varepsilon(\theta_{ij}|\boldsymbol{x}) = \lim_{K\to 0} \varepsilon(\eta^{\{1\}}_{ij}|\boldsymbol{x}) \times \lim_{K\to 0} \varepsilon(\eta^{\{2\}}_{ij}|\boldsymbol{x}).$$

It shows that the estimates corresponding to increasing values of K reflect an increasingly strong prior belief in the plausibility of independence of the two factors by

compromising between estimates obtained under a saturated model and estimates obtained under an independence model. Epstein (1990) showed that this property holds for general loglinear parametrizations.

With this introductory example it is now easy to see how our approach extends to tables of higher dimension. If we believe, with uncertainty, in a given loglinear structure for the cell probabilities, we use a two-stage prior. In the first stage use a Dirichlet distribution with means having the same loglinear structure. In the second stage use distributions, Gaussian for example, on the u-terms of the loglinear parametrization.

In the introductory example we speak of independence being a plausible structure for the cell probabilities to indicate that we believe in independence only to a certain degree. In general, we will speak of a plausible loglinear structure to indicate that we believe in that structure only to a certain degree.

In the multidimensional case, as $K \to \infty$, the prior and therefore the posterior concentrate all their mass in the subset of arrays $\boldsymbol{\eta}$ defined by the loglinear parametrization. Epstein (1990) studied properties of the posterior means as estimators when the loglinear parametrization on $\boldsymbol{\eta}$ is hierarchical.

The next section reviews the extension of the method for multidimensional tables and the basic elements of loglinear parametrizations.

Section 3 presents our implementation of the Gibbs sampler. The implementation requires finding maximum likelihood estimates for Dirichlet means under a loglinear parametrization. Subsection 3.1 describes the use of the projection gradient method to compute these maximum likelihood estimates. Additionally, section 3 discusses a rejection-acceptance scheme to draw deviates from a posterior distribution that does not require the marginal (predictive) distribution. Section 4 illustrates the implementation of the Gibbs sampler and the method of Epstein and Fienberg (1991) with simple sociological example concerning student politics and family structure.

2 A Bayesian Method for Multidimensional Tables

In this section we review the method proposed by Epstein (1990) and Epstein and Fienberg (1991) for multidimensional tables. We refer the reader to Epstein (1990) for proofs and a detailed discussion of this section's results.

Following the notation of Andersen (1974), consider n factors or treatments labeled $1, 2, \ldots, n$, with factor i having r_i levels. Define $\bar{r}_i = \{1, \ldots, r_i\}$ and call it the set of levels of factor i. The set $I = \bar{r}_1 \times \cdots \times \bar{r}_n$, is usually refered to as the index set or the set of cells.

A selection of levels $\iota = (i_1, i_2, \ldots, i_n)$, a generic element in I, is often referred to as the $(i_1, i_2, \ldots, i_n)$-cell. One obtains a $r_1 \times \cdots \times r_n$ contingency table $\boldsymbol{x} = \{x_\iota, \iota \in I\}$ when N individuals are examined and cross-classified according to the levels of each of the factors.

We shall assume that $\boldsymbol{x} = \{x_\iota, \iota \in I\}$ has a multinomial $M(N, \{\theta_\iota\})$ distribution, where θ_ι is the probability of an individual being classified in cell ι. However, the method easily adapts to other sampling distributions, such as Poisson and product multinomial (Bishop, Fienberg, and Holland, 1975).

In the first stage use a Dirichlet $D(K, \boldsymbol{\eta})$ distribution with density

$$[\boldsymbol{\theta}|K, \boldsymbol{\eta}] = \beta \prod_{\iota \in I} \theta_\iota^{K\eta_\iota - 1}, \tag{6}$$

indexed by $\boldsymbol{\eta} = \{\eta_\iota, \iota \in I\}$, and where $\beta = \frac{\Gamma(K)}{\prod_{\iota \in I} \Gamma(K\eta_\iota)}$. The parameter $K > 0$ is prespecified. Thus, $\mathcal{E}(\theta_\iota|K, \boldsymbol{\eta}) = \eta_\iota$ for $\iota \in I$.

Let $w \subset \bar{n}$, i.e., w is a set of factor labels. We shall denote $\boldsymbol{u}_w$ the interaction parameter among the factors in w. More specifically, the interaction $\boldsymbol{u}_w$ is the $r_1 \times \ldots \times r_n$ array

$$\boldsymbol{u}_w = \{u_{w(i_1, \ldots, i_n)}\},$$

where the entries $u_{w(i_1, \ldots, i_n)}$ of $\boldsymbol{u}_w$ depend only upon the indices i_j with $j \in w$. Often the interactions are taken to satisfy the usual ANOVA constraints, i.e., the sum of the entries $u_{w(i_1, \ldots, i_n)}$ over the levels of any factor $j \notin w$ is zero. These constraints achieve identifiability of the parametrization. The Bayesian approach does not require identifiable parametrizations and therefore we need not use constraints. Their use, however, is not precluded. One should use them whenever they facilitate producing a prior distribution reflecting one's beliefs.

Loglinear parametrizations are usually used for the multinomial parameters. The model defined by

$$\log \boldsymbol{\theta} = \sum_{w \subset \bar{n}} \boldsymbol{u}_w, \tag{7}$$

is the saturated or unrestricted model. Whenever a vector, $\boldsymbol{x}$ say, appears as the argument of a real function of one variable, f say, then $f(\boldsymbol{x})$ shall stand for the vector $(f(x_1), \ldots, f(x_l))^t$.

The entries of the array $\boldsymbol{u}_\emptyset$, where $\emptyset$ is the empty set, are all the same. The term $\boldsymbol{u}_\emptyset$ is usually referred to as the constant term.

In the general case we can use the multivariate logits γ_ι by writing:

$$\theta_\iota = \frac{e^{\gamma_\iota}}{\sum_{\iota \in I} e^{\gamma_\iota}}.$$

The parametrization (7) is equivalent to

$$\gamma = \sum_{w \subset \bar{n}, w \neq \emptyset} u_w.$$

We obtain submodels by including only some interactions in the formula above. To specify which interactions we include in a submodel, we use a class of subsets of $\bar{n}$ which we call $\mathcal{A}$. For example, we write

$$\log \theta = \sum_{w \in \mathcal{A}} u_w.$$

to specify a parametrization which only includes the interactions among factors in w, with $w \in \mathcal{A}$.

We are concerned with making inferences when we feel it is plausible that

$$\log \theta = \sum_{w \in \mathcal{A}} u_w, \tag{8}$$

where $\mathcal{A}$ is a strict subset of $\bar{n}$. To incorporate this belief into the prior we suggest that instead of using the loglinear parametrization on θ we use it on the Dirichlet means, that is,

$$\log \eta = \sum_{w \in \mathcal{A}} u_w. \tag{9}$$

This restricts $\log \eta$ to lie in a linear subspace M of $\Re^I$. To ensure that the parametrization is such that $\sum_{\iota \in I} \eta_\iota = 1$, it is necessary to assume that M contains the array $\mathbf{1}$ whose entries are all 1. In the introductory example the class $\mathcal{A}$ is $\{\{1\},\{2\}\}$ and therefore equation (9) becomes

$$\log \eta_{ij} = u + u_{1(ij)} + u_{2(ij)}.$$

In the parametrization (9) the term $u_\emptyset = \{u\}$ must satisfy

$$u = -\log\Big(\sum_{\iota \in I} \exp\Big(\sum_{w \in \mathcal{A}, w \neq \emptyset} u_{w(\iota)}\Big)\Big),$$

so that $\sum_{\iota \in I} \eta_\iota = 1$. The term $u_\emptyset$ in (8) must satisfy this restriction as well. The restriction on $u_\emptyset$ will remain implicit whenever we refer to parametrizations such as those in (8) and (9).

In summary, we suggest a two-stage hierarchical prior. The first stage consists of setting $\theta \sim D(K, \eta)$ where η is parametrized using (9). The second stage consists of setting distributions on the u-terms in (9).

As a consequence of using the parametrization (9) one can specify a value for η by specifying values for some margins of η. For example, if $\log \eta_{ij} = u + u_{1(ij)} + u_{2(ij)}$, then $\eta_{ij} = \eta_{i+}\eta_{+j}$. In this fashion we specify the rc values η_{ij} by specifying values for η_{i+} and η_{+j}, a total of only $r + c$ values.

This result extends to the general case. When the loglinear parametrization (9) is hierarchical then η is totally specified by the value of the margins $\eta^{Y_1}, \ldots, \eta^{Y_T}$, where $\{Y_1, \ldots, Y_T\}$ is the generating class of the loglinear parametrization. Therefore, we can implement the second stage either by using distributions on the u-terms or by using distributions on the margins $\eta^{Y_1}, \ldots, \eta^{Y_T}$. For $Y \subset \bar{n}$ the Y-margin η^Y is defined as being the array whose entries are

$$\eta^Y_{i_1 \cdots i_n} = \sum_{\bar{n} \setminus Y} \eta_{i_1 \cdots i_n} = \sum_{i_j \in r_j,\, j \in \bar{n} \setminus Y} \eta_{i_1 \cdots i_n}.$$

3 Implementation of the Gibbs Sampler

This section describes the specifics of the implementation of the Gibbs sampler for calculating the posterior densities of the cell probabilities.

We start with a brief review of the Gibbs sampler and refer the reader to Gelfand *et al.* (1990) and Gelfand and Smith (1990) for a detailed description of the use of Gibbs sampling in Bayesian inference.

Suppose that one wishes to estimate the density $[X]$ of the random variable X assuming it is possible to draw deviates from the conditional densities $[X|Y]$ and $[Y|X]$, where Y is another random variable.

The algorithm consists of iteratively repeating a two-step cycle. Before starting one draws a deviate $X^{(0)}$ from an arbitrary density $[X]_0$. Step one of the cycle is to draw a deviate $Y^{(1)}$ from $[Y|X^{(0)}]$. Step two is to draw $X^{(1)}$ from $[X|Y^{(1)}]$. Then one first replaces $X^{(0)}$ by $X^{(1)}$ and proceeds with the second cycle. A succession of cycles produces a sequence $(X^{(1)}, Y^{(1)}), (X^{(2)}, Y^{(2)}), \ldots, (X^{(i)}, Y^{(i)}), \ldots$. The sequence $X^{(i)}$ converges in distribution to $X \sim [X]$ and $Y^{(i)}$ converges in distribution to $Y \sim [Y]$.

Gelfand and Smith (1990) suggest building an estimate of the density $[X]$ as follows. Using the Gibbs sampler, obtain m independent replicates $(X_1^{(t)}, Y_1^{(t)}), \ldots, (X_m^{(t)}, Y_m^{(t)})$. With these deviates obtain the density estimate

$$\widehat{[X]} = m^{-1} \sum_{j=1}^{m} [X|Y_j^{(t)}]. \tag{10}$$

We use Gibbs sampling to estimate the posterior distribution $[\boldsymbol{\theta}|\boldsymbol{x}]$. For the two-stage prior $\boldsymbol{\theta} \sim [\boldsymbol{\theta}|\boldsymbol{\eta}]$ and $\boldsymbol{\eta} \sim [\boldsymbol{\eta}]$. We identify X with $\boldsymbol{\theta}$ and Y with $\boldsymbol{\eta}$ and use the following Gibbs scheme to draw deviates from the posterior $[\boldsymbol{\theta}|\boldsymbol{x}]$:

```
do j = 1,m
   Set θ^(0)
   do i = 1,t
      Step 1: draw η from [η|θ^(0)]
      Step 2: draw θ^(1) from [θ|η, x]
      θ^(0) ← θ^(1)
   end do
   η_j^(t) ← η
   θ_j^(t) ← θ^(0).
end do
```

On exit, this process has generated m independent deviates $\boldsymbol{\eta}_j^{(t)} \sim [\boldsymbol{\eta}^{(t)}], j = 1, \ldots, m$ and m independent deviates $\boldsymbol{\theta}_j^{(t)} \sim [\boldsymbol{\theta}^{(t)}], j = 1, \ldots, m$.

With these deviates, the density estimate in equation (10) is a finite mixture of Dirichlet densities,

$$\widehat{[\boldsymbol{\theta}|\boldsymbol{x}]} = m^{-1} \sum_{j=1}^{m} [\boldsymbol{\theta}|\boldsymbol{\eta}_j^{(t)}, \boldsymbol{x}].$$

It is particularly simple to evaluate the marginal density estimate of a cell probability. For example, in the situation of the introductory example,

$$\widehat{[\theta_{11}|\boldsymbol{x}]} = m^{-1} \sum_{j=1}^{m} [\theta_{11}|\boldsymbol{\eta}_j^{(t)}, \boldsymbol{x}], \tag{11}$$

where $[\theta_{11}|\boldsymbol{\eta}, \boldsymbol{x}]$ is the beta$(K^*\eta_{11}^*, K^*(1 - \eta_{11}^*))$ density with $K^* = N + K$, $\eta_{11}^* = \alpha\eta_{11} + (1 - \alpha)x_{11}/N$, and $\alpha = K/(N + K)$.

Automatically monitoring convergence is still an open issue; at present the best one can do is to prespecify the total number of iterations t, say.

The distribution $[\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{x}]$, which is used in step two of a Gibbs sampler cycle, is Dirichlet with concentration parameter $K^* = (N + K)$ and cell means $\eta_i^* = K/(N + K)\eta_i + 1/(N + K)x_i$. Drawing deviates from the distribution $[\boldsymbol{\theta}|\boldsymbol{\eta}, \boldsymbol{x}]$ is straightforward. We chose to generate these deviates by independently generating $\gamma_i \sim$ Gamma$(p_i, 1)$, $i \in I$ with, $p_i = K^*\eta_i^*$, and then setting $\theta_i = \gamma_i / \sum_{i' \in I} \gamma_{s'}$. The joint distribution of $\{\theta_i\}$ is Dirichlet $D(K^*, \{q_i\})$ with $q_i = p_i/K^*$.

However, drawing deviates from the distribution $[\boldsymbol{\eta}|\boldsymbol{\theta}^{(0)}] = [\boldsymbol{\theta}^{(0)}|\boldsymbol{\eta}]\ [\boldsymbol{\eta}]/[\boldsymbol{\theta}^{(0)}]$, used in step one of a cycle, is not straightforward.

We suggest the following adaptation of the rejection method to sample from $[\boldsymbol{\eta}|\boldsymbol{\theta}^{(0)}]$. This adaptation uses deviates $\boldsymbol{\eta}$ from $[\boldsymbol{\eta}]$, which are easy to generate, to obtain deviates from $[\boldsymbol{\eta}|\boldsymbol{\theta}^{(0)}] = [\boldsymbol{\theta}^{(0)}|\boldsymbol{\eta}]\ [\boldsymbol{\eta}]/[\boldsymbol{\theta}^{(0)}]$.

```
accept ← false
do while ( not( accept ) )
   generate a deviate η from [η]
   generate a deviate v from U[0, B]
   if ( v ≤ [θ^(0)|η] ) then accept ← true
end do
```

Above, B is such that $B \geq [\boldsymbol{\theta}^{(0)}|\boldsymbol{\eta}]$ for all $\boldsymbol{\eta}$ in its domain. It is simple to show that an accepted $\boldsymbol{\eta}$ is a deviate from $[\boldsymbol{\eta}|\boldsymbol{\theta}^{(0)}]$. An important feature of this approach is that it does not require the calculation of $[\boldsymbol{\theta}^{(0)}]$, or an estimate of it, as is sometimes necessary in some implementations of the Gibbs sampler (see for example Gelfand and Smith, 1990).

A geneneralized rejection method that uses an enveloping function $B(\boldsymbol{\eta})$ for $\boldsymbol{\eta} \rightarrow [\boldsymbol{\eta}|\boldsymbol{\theta}]$ may increase the speed of this algorithm. At present, however, we will content ourselves with a boxed envelop, the main advantage being the ease of programming. Obtaining a good value for B is crucial for a good performance of the rejection method. The ideal choice is to find $\hat{\boldsymbol{\eta}}$ such that

$$[\boldsymbol{\theta}^{(0)}|\hat{\boldsymbol{\eta}}] = \max\{[\boldsymbol{\theta}^{(0)}|\boldsymbol{\eta}] : \log \boldsymbol{\eta} = \sum_{w \in \mathcal{A}} \boldsymbol{u}_w\},$$

and then take $B = [\boldsymbol{\theta}^{(0)}|\hat{\boldsymbol{\eta}}]$. Observe that $\boldsymbol{\eta} \rightarrow [\boldsymbol{\theta}^{(0)}|\boldsymbol{\eta}]$, is the Dirichlet likelihood function given the data $\boldsymbol{\theta}^{(0)}$.

The next subsection introduces a maximization procedure to find B. The procedure appears to be fast enough to use it in combination with the Gibbs sampler.

Observe that under the loglinear parametrization $\boldsymbol{\gamma} = \log \boldsymbol{\eta} = \sum_{w \in \mathcal{A}} \boldsymbol{u}_w$, $\hat{\boldsymbol{\gamma}} = \log \hat{\boldsymbol{\eta}}$ is the maximum likelihood estimate of $\boldsymbol{\gamma}$.

3.1 Maximizing the Dirichlet Likelihood

In this section we briefly describe the "gradient projection method" and apply it to maximize the Dirichlet loglikelihood. In addition to being easy to implement, various features of the Dirichlet likelihood and loglinear parametrizations make the gradient projection method preferable to other methods. We discuss the advantages of the gradient projection method after introducing additional definitions.

Recall that a loglinear parametrization for $\boldsymbol{\eta}$ restricts $\log \boldsymbol{\eta}$ to lie in a linear subspace M. The usual form of writing a loglinear parametrization with u-terms expresses $\boldsymbol{\gamma} \in M$ in terms of a basis matrix of M, i.e., a matrix B whose columns form a basis for M. When expressing a vector $\boldsymbol{\gamma} \in M$ in terms of the unique $\boldsymbol{u}$ such that $\boldsymbol{\gamma} = B\boldsymbol{u}$, the coordinates of $\boldsymbol{u}$ are the u-terms.

To avoid technical complications that the restriction $\sum_{\iota \in I} \eta_\iota = 1$ introduces, we redefine some functions of $\boldsymbol{\eta}$ as functions of $\eta_\iota, \iota \neq (r_1, \ldots, r_n)$ only. To this effect, for $\boldsymbol{\eta}$ given we define $\bar{\boldsymbol{\eta}}$ as $\bar{\eta}_\iota = \eta_\iota$, $\iota \in \bar{I}$ with $\bar{I} = I - \{(r_1, \ldots, r_n)\}$ as the index set for the vectors $\bar{\boldsymbol{\eta}}$.

Maximizing the Dirichlet likelihood $l(\boldsymbol{\eta}|\boldsymbol{\theta})$ is equivalent to maximizing

$$E(\bar{\boldsymbol{\eta}}) = K < \boldsymbol{\eta}, \boldsymbol{\lambda} > - \sum_{\iota \in I} \log \Gamma(K\eta_\iota),$$

where $\boldsymbol{\lambda} = \log \boldsymbol{\theta}$ and $\boldsymbol{\eta}$ is given by $\eta_\iota = \bar{\eta}_\iota, \iota \in \bar{I}$ and $\eta_{(r_1, \ldots, r_n)} = 1 - \sum_{\iota \in \bar{I}} \bar{\eta}_\iota$. Except for an additive constant, $E(\boldsymbol{\eta})$ is $\log l(\boldsymbol{\eta}|\boldsymbol{\theta})$.

The Dirichlet means corresponding to the multivariate logits $\boldsymbol{\gamma}$ are given by $\boldsymbol{\eta} = H(\boldsymbol{\gamma})$ with $\bar{\eta}_\iota = \exp(\gamma_\iota) / \sum_{\iota' \in I} \exp(\gamma_{\iota'}), \iota \in \bar{I}$. To use the parametrization with the multivariate logits it is convenient to define

$$G(\boldsymbol{\gamma}) = E(H(\boldsymbol{\gamma})). \tag{12}$$

Observe that if we use the parametrization with the u-terms, then we may find $\boldsymbol{u}$, the m.l.e. of $\boldsymbol{u}$, by maximizing $U(\boldsymbol{u}) = G(B\boldsymbol{u})$.

Roughly speaking, there are three classes of alternative methods to maximize U. One possibilty is to solve the equation $JU(\boldsymbol{u}) = 0$, where JU stands for the array of partial derivatives of U. Typically, iterative procedures to solve this equation require updating an estimate of the Hessian of U after some iterations. On the one hand, it is difficult to obtain formulas for the second derivatives of U and on the other, computing second derivatives numerically is in general expensive and roundoff errors are difficult to control. Since $U(\boldsymbol{u}) = G(B\boldsymbol{u})$, this approach poses the additional difficulty of explicitly requiring a basis matrix for M.

An alternative is to use a steepest ascent method where at each step there is a unidimensional search along the direction $JU(\boldsymbol{u})$. This alternative also requires a basis matrix. In fact, any method that uses $\boldsymbol{u}$ as the variable of the objective function, will require a basis matrix.

The gradient projection method is preferable to these alternatives because it does not require estimating Hessians or a basis matrix of M. Moreover, the gradient projection method allows us to take advantage of the ANOVA-type parametrization for $\boldsymbol{\gamma}$ to perform certain computations more efficiently.

To use the gradient projection method we view the problem of maximizing the Dirichlet likelihood as the problem of finding $\boldsymbol{\gamma} \in M$ such that

$$G(\boldsymbol{\gamma}) = \max\{G(\boldsymbol{\gamma}), \boldsymbol{\gamma} \in M\},$$

which is a constrained maximization problem. A point $\boldsymbol{\gamma} \in M$ is refered to as a "feasible point". The gradient projection method projects the gradient of the objective function onto M to increase the value of $G(\boldsymbol{\gamma})$ and to maintain feasibility at the same time.

The following is a summary of the gradient projection to solve the above maximization problem:

Step 1 Initialization: Choose $\boldsymbol{\gamma}_0 \in M$
Let $v_0 = G(\boldsymbol{\gamma}_0)$
Step 2 Compute $\boldsymbol{d}_0 = JG(\boldsymbol{\gamma}_0)$
Step 3 Compute $\boldsymbol{\delta}_0 = P_M \boldsymbol{d}_0$
Step 4 Unidimensional maximization:
Find $\hat{\alpha} > 0$ such that
$G(\boldsymbol{\gamma}_0 + \hat{\alpha}\boldsymbol{\delta}_0)) = \max_{\alpha > 0} G(\boldsymbol{\gamma}_0 + \alpha\boldsymbol{\delta}_0))$
Set $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_0 + \hat{\alpha}\boldsymbol{\delta}_0$
Step 5 Convergence test:
Let $v_1 = G(\boldsymbol{\gamma}_1)$
If $(v_1 - v_0)/v_0 < \epsilon$ then stop
else $\boldsymbol{\gamma}_0 \leftarrow \boldsymbol{\gamma}_1$
$v_0 \leftarrow v_1$
go to Step 2.

On exit, $\boldsymbol{\gamma}_1$ is such that $v_1 = G(\boldsymbol{\gamma}_1)$ is an estimate of the maximum value of G. Therefore $l(H(\boldsymbol{\gamma}_1)|\boldsymbol{\theta})$ is an estimate of the maximum value of $l(\boldsymbol{\eta}|\boldsymbol{\theta})$.

In Step 2, $JG(\boldsymbol{\gamma}_0)$ stands for the array of partial derivatives $\{\partial G(\boldsymbol{\gamma}_0)/\partial\gamma_\iota, \iota \in I\}$. It follows from (12) that, for $\iota = (i_1, \ldots, i_n) \in \bar{I}$ and $c = (r_1, \ldots, r_n)$,

$$\frac{\partial G}{\partial \gamma_\iota}(\boldsymbol{\gamma}) = K\eta_\iota[(\lambda_\iota - \lambda_c - \{\psi(K\eta_\iota) - \psi(K\eta_c)\})(1 - \eta_\iota) + \sum_{\iota' \in I} (\lambda_{\iota'} - \lambda_c - \{\psi(K\eta_{\iota'}) - \psi(K\eta_c)\})\eta_{\iota'}],$$

and,

$$\frac{\partial G}{\partial \gamma_c}(\boldsymbol{\gamma}) = K\eta_c[-\sum_{\iota' \in I} (\lambda_{\iota'} - \lambda_c - \{\psi(K\eta_{\iota'}) - \psi(K\eta_c)\})\eta_{\iota'}],$$

where ψ is the digamma function and $\lambda_\iota = \log\theta_\iota$, $\iota \in I$.

The formulas to compute the projection $P_M \boldsymbol{d}_0$ in Step 3 are derived in a similar fashion to the formulas to compute fitted values of the cell means in ANOVA. However, these formulas are not the same because the parametrization for $\boldsymbol{\gamma}$ does not involve the constant term of ANOVA parametrizations.

The existence of $\hat{\alpha}$ in Step 4 is guaranteed by the concavity of the Dirichlet likelihood. We used routine e04abf from the *NAg* library for the unidimensional maximizations. Although it would take more programming, perhaps an algorithm that uses the derivative of $G(\boldsymbol{\gamma}_0 + \alpha\boldsymbol{\delta}_0)$

with respect to α would be more efficient for the unidimensional maximizations.

It is possible to use other convergence tests in Step 5. Since our interest here is not on the maximizer $\hat{\gamma}$, but on the maximum value $G(\hat{\gamma})$, it is appropriate to use the test in Step 5 to ensure that on exit γ_1 provides a function value v_1 sufficiently close to $G(\hat{\gamma})$.

4 Illustrative Example

In this section we reanalyze the 2×2 table given in Table 1 which classifies college students with respect to their political affiliation and their family structure (from Braungart 1971, and analyzed in Bishop, Fienberg and Holland, 1975, pp 379-380), and by Albert and Gupta (1984). We use this data to estimate the cell probabilities using the prior belief that the two variables under study are plausibly independent. This is the situation described in the introduction. For illustrative purposes we use normal distributions on the u-terms in the parametrization

$$\log \eta_{ij} = u + u_{1(i)} + u_{2(j)}. \tag{13}$$

More precisely, we use

Stage I: $\theta|K, \eta \sim d(K, \eta)$, with η_{ij} reparametrized according to equations (13).

Stage II: The $u_{1(i)}$ are independent, $i = 1, 2$. The $u_{2(j)}$ are independent, $j = 1, 2$, and also independent of the $u_{1(i)}, i = 1, 2$. The distribution of $u_{1(i)}$ is $N(\mu_{1(i)}, \sigma^2_{1(i)})$ and the distribution of $u_{2(j)}$ is $N(\mu_{2(j)}, \sigma^2_{2(j)})$, $i, j = 1, 2$.

To use this prior density one first specifies the parameter vectors $\boldsymbol{\mu}_1 = (\mu_{1(1)}, \mu_{1(2)})$, and $\boldsymbol{\sigma}_1 = (\sigma_{1(1)}, \sigma_{1(2)})$, reflecting the user's prior knowledge about the proportion of students in the two political affiliations and, $\boldsymbol{\mu}_2 = (\mu_{2(1)}, \mu_{2(2)})$, and $\boldsymbol{\sigma}_2 = (\sigma_{2(1)}, \sigma_{2(2)})$, reflecting the user's prior knowledge about the proportion of students in the two family structures.

In this example we set $\boldsymbol{\mu}_1 = (.5; .5), \boldsymbol{\sigma}_1 = (2.0; 2.0)$ and $\boldsymbol{\mu}_2 = (.5; .5), \boldsymbol{\sigma}_2 = (2.0; 2.0)$, reflecting a rather imprecise belief about the u-terms. Second, one specifies a value for the parameter K.

Table 1: Parental decision making and political affiliation. Source: Braungart(1971).

		Political Affiliation	
		SDS	YAF
Parental Decision Making	Authoritarian	29	33
	Democratic	131	78

Albert and Gupta (1982) and Epstein and Fienberg (1991) computed the posterior means (5) for this table but they used different distributions to reflect uncertain prior beliefs about independence. In both articles the posterior expectation of the η's were estimated using a Monte Carlo method.

Table 2 reports the computed values for the posterior means of each of the cell probabilities for several values of K (the column headed by $K = \infty$ actually corresponds to a very large, but finite, value of K). The estimates corresponding to finite values of K reflect the uncertain prior belief in independence by compromising between estimates obtained under a saturated model and estimates obtained under an independence model

Figure 1 reports reports estimates of the marginal posterior densities for each of the cell probabilities. These estimates were obtained using formula (11) for the posterior density of θ_{11} and with the obvious modifications for the other cell probabilies. We used $m = 20$ independent replicates and each of the replicates was generated with $t = 20$ cycles of the Gibbs sampler. In addition we computed these density estimates using different values of m and t. On a plot the resulting estimates appeared to be fairly similar for values of t and m as low as 10.

Table 2: Computed values of posterior means for different values of K

K	0	100	200	400	600	1000	2000	∞
$\hat{\theta}_{11}$	.107	.115	.119	.125	.130	.126	.135	.133
$\hat{\theta}_{12}$	.122	.115	.110	.105	.102	.098	.100	.093
$\hat{\theta}_{21}$	.483	.474	.471	.469	.468	.463	.453	.459
$\hat{\theta}_{22}$	.288	.296	.300	.302	.299	.313	.312	.316

5 Discussion

This article reports on an implementation of the Gibbs sampler to estimate the full posterior density of the array of cell probabilities of n-way contingency tables using the method proposed by Epstein (1990) and Epstein and Fienberg (1991). One easily obtains estimates of the posterior distributions of the individual cell probabilities as a finite mixture of beta densities.

Gelfand and Smith (1990) proposed the Gibbs sampler as an easy to implement algorithm to generate deviates from posterior distributions. An expeditious implementation requires that all necessary distributions be available for sampling. This was not the case in this article

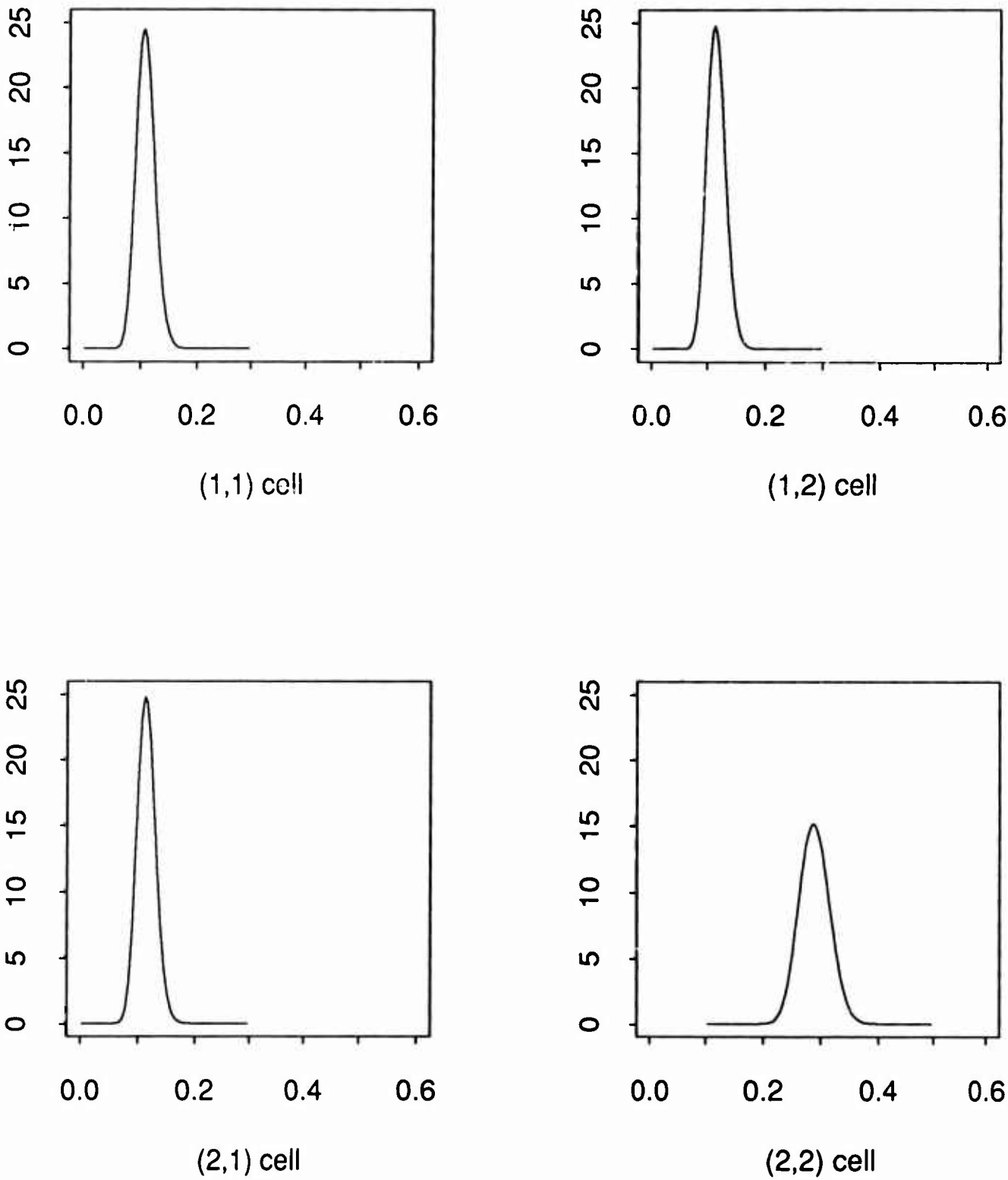

Figure 1:
Estimated posterior densities, K=150

and we expended some efforts to generate deviates from $[\eta|\theta]$.

To sample from $[\eta|\theta]$ we used a scheme that does not requires the marginal density $[\theta]$, which is often the main obstacle to compute $[\eta|\theta]$. The scheme uses the facts that $[\eta]$ is available for sampling, that $[\theta|\eta]$ as a function of η can be viewed as a concave likelihood function with a unique maximum. This maximum provides the height of a box for a rejection sampling method. The gradient projection method proved to be fast and very easy to program. We are currently investigating its use in maximum likelihood estimation for generalized linear models and will report on this work elsewhere.

Our scheme to sample from $[\eta|\theta]$ can be used to implement the Gibbs sampler for a variety of other problems involving two-stage priors where the first stage is the conjugate prior for the sampling distribution and the second stage distribution is available for sampling.

Furthermore, we feel that the simplicity of the Gibbs sampler warrants exploring new algorithms to generate deviates from distributions that thus far have not been available for sampling. For clarity we used a simple 2×2 example to illustrate our implementation.

In higher dimensional tables, it makes special sense to utilize the structure of η in terms of its marginals as part of the algorithm and to set up a cycle involving steps for the conditional densities for each of the marginals of η instead of a single step for $[\eta|\theta]$. We hope to report on the details of such an algorithm at a future date.

References

[1] ALBERT, A. H. & GUPTA, A. K. (1982), Mixtures of Dirichlet Distributions and Estimation in Contingency Tables. *Ann. Statist.*, **10** , *No. 4, 61-68.*

[2] ANDERSEN, A. H. (1974), Multidimensional Contingency Tables. *Scand. J. Statist.*, **1**, *115-127.*

[3] BISHOP, Y. M. M., FIENBERG, S. E. & HOLLAND, P. (1975). Discrete Multivariate Analysis: Theory and Practice. Cambridge, Mass: M.I.T. Press.

[4] BRAUNGART, R. G. (1971). Family status, socialization and student politics: a multivariate analysis. Cambridge, Mass: M.I.T. Press.

[5] EPSTEIN, L. D. (1990), Bayesian Estimation in Multidimensional Contingency Tables. Ph.D. Thesis. Department of Statistics, Carnegie Mellon University.

[6] EPSTEIN, L. D. & FIENBERG, S. E. (1991), Bayesian Estimation in Multidimensional Contingency Tables. *Bayesian Inference in Statistics and Econometrics: Proceedings of the Indo-US Workshop, 1988.* Lecture Notes in Statistics, Springer-Verlag New York. (To appear.)

[7] GELFAND, A. E., HILLS, S. E., RACINE-POON, A., & SMITH, A. F. M. (1990), Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling. *J. Am. Statist. Assoc.*, **85**,*No. 412, 972-985.*

[8] GELFAND, A. E. & SMITH, A. F. M. (1990), Sampling Based Approaches to Calculating Marginal Densities. *J. Am. Statist. Assoc.*, **85** , *398-409.*

[9] LEONARD, T. & NOVICK, M. R. (1986), Bayesian Full Rank Marginalization for Two-Way Contingency Tables. *J. Ed. Statist.*, **11** , *No. 1, 33-56.*