The Annals of Statistics
1996, Vol. 24, No. 5, 2178–2198

BAYESIAN MODELS FOR SPARSE PROBABILITY TABLES

BY JIM Q. SMITH AND CATRIONA M. QUEEN

University of Warwick and University of Kent

We wish to make inferences about the conditional probabilities $p(y|x)$, many of which are zero, when the distribution of X is unknown and one observes only a multinomial sample of the Y variates. To do this, fixed likelihood ratio models and quasi-incremental distributions are defined. It is shown that quasi-incremental distributions are intimately linked to decomposable graphs and that these graphs can guide us to transformations of X and Y which admit a conjugate Bayesian analysis on a reparametrization of the conditional probabilities of interest.

1. Introduction. An $n \times m$ matrix of probabilities $\{p(i, j)\}$ needs to be estimated, where $p(i, j) = P(X = x_i,\ Y = y_j)$. Many of these joint probabilities are zero, but the margins of X and Y are nondegenerate, so that

$$P(X = x_i) = \theta_i > 0, \qquad 1 \le i \le n, \qquad \boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$$

and

$$P(Y = y_j) = \psi_j > 0, \qquad 1 \le j \le m, \qquad \boldsymbol{\psi} = (\psi_1, \ldots, \psi_m)^T.$$

A random sample $\mathbf{r} = (r_1, r_2, \ldots, r_m)^T$ of the Y variables is taken, so that the random vector $\mathbf{R}$ associated with $\mathbf{r}$ has a multinomial distribution $Mn(N, \boldsymbol{\psi})$, where $N = \sum_{j=1}^{m} r_j$.

If the conditional distribution of Y given X is fully specified, then interest centers on the margins $\boldsymbol{\theta}$ of X and this becomes an inverse problem [see Grandy (1985) and Vardi and Lee (1993)]. This particular setting was discussed in some detail by Dickey, Jiang and Kadane (1987) and posterior distributions on $\boldsymbol{\theta}$ were found, which under appropriate prior distributions are generalized Dirichlets [Carlson (1977); Dickey (1983)]. In this paper we concentrate on the dual problem comparable to the one above. When $m > n$ it is shown that, even if $\boldsymbol{\theta}$, the margin of X, is unknown it is possible to learn at least something about the conditional probabilities of $Y \mid X$ from the multinomial observation $\mathbf{r}$. Furthermore, we show that it is sometimes possible to perform a conjugate analysis on parameters related to these conditional probabilities.

Everything that is learned about $p(y \mid x)$ and $\boldsymbol{\theta}$ comes from observing $\mathbf{R}$, which is informative about $p(y \mid x)$ and $\boldsymbol{\theta}$ only through its margin $\boldsymbol{\psi}$. Since $\mathbf{R}$ is only m dimensional, a general model to learn about $p(y \mid x)$ and $\boldsymbol{\theta}$ is clearly overparametrized. To overcome this overparametrization, we shall choose to

Received April 1994; revised January 1996.

AMS 1991 *subject classifications*. Primary 62F15; secondary 62H17.

Key words and phrases. Bayesian probability estimation, constraint graph, contingency tables, decomposable graph, generalized Dirichlet distributions, separation of likelihood.

fix the $n \times m$ matrix $Z = \{z_{ij}\}_{1 \le i \le n,\, 1 \le j \le m}$ whose components are defined by

$$z_{ij} = \lambda_j^{-1} P(Y = y_j \mid X = x_i), \qquad 1 \le i \le n,\ 1 \le j \le m, \tag{1.1}$$

where

$$\lambda_j = P(Y = y_j \mid X = x_{i^*(j)}) = \max_{1 \le i \le n} P(Y = y_j \mid X = x_i) > 0, \qquad 1 \le j \le m. \tag{1.2}$$

Note that the strict inequality on $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^T$ is necessary to ensure the strict positivity of the margin $\boldsymbol{\psi}$. Here $i^*(j) = \arg\max_{1 \le i \le n} P(Y = y_j \mid X = x_i)$ indexes a most likely value of X given each observation y_j, $1 \le j \le m$. It follows that for each $1 \le j \le m$, λ_j is the maximum value, over different choices of x_i, of the conditional probability of Y on X. Clearly, from (1.1), $z_{ij} = 0$ whenever $p(i, j) = 0$. So in problems where many of the joint probabilities are known to be zero, many of the entries of the matrix Z are also known to be zero.

Now, the likelihood of our random sample of N copies of Y takes the form

$$L(\boldsymbol{\psi} \mid \mathbf{r}) = \prod_{j=1}^{m} \psi_j^{r_j}, \tag{1.3}$$

where, by the formula for extension, each probability ψ_j satisfies

$$\psi_j = \sum_{i=1}^{n} P(Y = y_j \mid X = x_i) P(X = x_i) = \lambda_j \xi_j(\boldsymbol{\theta}), \tag{1.4}$$

where

$$\xi_j(\boldsymbol{\theta}) = \sum_{i=1}^{n} z_{ij} \theta_i, \qquad 1 \le j \le m. \tag{1.5}$$

Thus for a given Z, as Dickey, Jiang and Kadane (1987) point out, the likelihood L separates in $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$, that is,

$$L(\boldsymbol{\psi}|\mathbf{r}) = L(\boldsymbol{\theta}, \boldsymbol{\lambda}|\mathbf{r}) = L_1(\boldsymbol{\theta}) L_2(\boldsymbol{\lambda}). \tag{1.6}$$

Here

$$L_1(\boldsymbol{\theta}) = \prod_{j=1}^{m} (\xi_j(\boldsymbol{\theta}))^{r_j}, \qquad \sum_{i=1}^{n} \theta_i = 1, \qquad \theta_i > 0,\ 1 \le i \le n, \tag{1.7}$$

where

$$\boldsymbol{\xi}(\boldsymbol{\theta}) = Z^T \boldsymbol{\theta}, \qquad \boldsymbol{\xi}(\boldsymbol{\theta}) = (\xi_1(\boldsymbol{\theta}), \ldots, \xi_m(\boldsymbol{\theta}))^T \tag{1.8}$$

and

$$L_2(\boldsymbol{\lambda}) = \prod_{j=1}^{m} \lambda_j^{r_j}, \tag{1.9}$$

where, for conditional probabilities to sum to unity, we require

$$Z\boldsymbol{\lambda} = \mathbf{1}, \qquad \mathbf{0} < \boldsymbol{\lambda} < \mathbf{1}, \tag{1.10}$$

where $\mathbf{0}$ and $\mathbf{1}$ denote n vectors of 0's and 1's, respectively.

Note that, ignoring awkward positivity constraints, the dimensions of the constrained linear spaces $\boldsymbol{\theta}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\psi}$ are, respectively, $n-1$, $m-n$ and $m-1$, so the dimension of $\boldsymbol{\psi}$ is the same as $(\boldsymbol{\theta}, \boldsymbol{\lambda})$. In this paper we shall consider a class of models called *fixed likelihood ratio models*. These models assume that the matrix Z is given and that the vector $\boldsymbol{\lambda}$ of largest conditional probabilities of $Y \mid X$ and the vector $\boldsymbol{\theta}$ of marginal probabilities on X, are unknowns. Then, by using the likelihood of (1.6), inferences can be made about $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$. An important feature of fixed likelihood ratio models is that they utilize the separation (1.6); that is, if $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ are a priori independent in a Bayesian model, then they will remain independent after observing $\mathbf{r}$. Here we shall concentrate on the inferences that can be made about $\boldsymbol{\lambda}$, a vector of probabilities of Y conditional on X.

The characteristics of fixed likelihood ratio models are determined largely by the pattern of zeros in the joint probability table $\{p(i, j)\}$ and hence in the Z matrix. In Section 2 it is shown how this pattern can be usefully classified in terms of a graph. In Sections 3 and 4 a class of distributions over (X, Y), called quasi-incremental, is introduced, which is shown to admit a useful reparametrization of $\boldsymbol{\lambda}$ that allows for a conjugate Bayesian analysis in terms of mixtures of Dirichlet distributions. Despite the proliferation of numerical methods in Bayesian statistics, conjugate analyses with plausible priors are always interesting because they specify the class of transformations from prior to posterior induced by a particular likelihood and hence give an understanding of *why* we obtain the results we do. In Sections 4 and 5 it is shown how to use the graph of a distribution on (X, Y) to discover a transformation of (X, Y) to $(\tau_1(X), \tau_2(Y))$ which makes this conjugate Bayesian analysis as simple as possible. Sometimes conditions exist when these transformations lead to a particularly simple conjugate analysis in which the components of $\boldsymbol{\lambda}$ can be expressed as products of independent variates. In Section 6 we outline the application of these models in two very different settings. Finally, Section 7 discusses why we believe that fixed likelihood ratio models are most valid when (X, Y) is quasi-incremental.

2. Graphs and the depiction of a probability model. There are various graphs which can be usefully employed to depict a zero–nonzero configuration in Z. One that is particularly useful for analyzing the conditional probability vector $\boldsymbol{\lambda}$ is the *primal-constraint graph* or briefly *graph*, $G(Z)$, of Z [Dechter and Pearl (1987); Dechter, Dechter and Pearl (1990)]. The "constraints" here are those imposed by condition (1.10).

The m nodes of $G(Z)$ are labelled by the m possible values of Y. Two nodes j_1 and j_2 are joined by an undirected edge iff there exists a value $i(j_1, j_2)$ such that both $p(i, j_1) > 0$ and $p(i, j_2) > 0$. Equivalently, j_1, j_2, are joined

by an edge iff there is a row (the ith) such that z_{ij_1} and z_{ij_2} are both strictly positive.

Notice that all joint distributions with the same pattern of nonzero and zero joint probabilities have the same graph. Let the matrix $Z^* = \{z^*_{ij}\}$ be defined by

$$z^*_{ij} = \begin{cases} 0, & \text{if } z_{ij} = 0 \\ 1, & \text{if } z_{ij} > 0, \end{cases} \tag{2.1}$$

where $Z = \{z_{ij}\}$. The *differential noninformedness hypothesis* (dnh) used by Rubin (1976) and Dawid and Dickey (1977) gives a model for which $Z = Z^*$. In a sense described in these papers, models satisfying the dnh are models which are most conservative—or noninformative—about the relationship between X and Y among all those which respect the zeros in the probability table. In general, for any graph G there is a set of probability distributions satisfying the dnh, and a set of Z^*, called the *G-set*, whose graph is G.

There are several reasons why these graphs of probability models are important. The first is that, unlike the matrix Z itself, $G(Z)$ is invariant under 1–1 transformations of the margins of X and Y. Thus, if $X' = \tau_1(X)$ and $Y' = \tau_2(Y)$, τ_1, τ_2 are bijections and Z' corresponds to the conditional probability ratios of X' against Y', then $G(Z) = G(Z')$. The second is that they emphasize in an evocative way the fundamental structure of the problem via the zeros of the joint mass function. The third is that they can be used to find convenient reparametrizations of the vector $\boldsymbol{\lambda}$ so that a conjugate analysis can be performed.

A subgraph of a graph G is said to be *complete* iff there is an edge connecting each pair of its nodes. The *cliques* of a graph are those of its complete subgraphs which are not properly contained in any other complete subgraph. Let

$$\mathscr{Y}(i) = \{j\colon\ p(i, j) > 0,\ 1 \le j \le m\}$$

be called the *range sets* of Y given X. Note that in the notation of Section 1,

$$\mathscr{Y}(i) = \{j\colon\ z_{ij} > 0,\ 1 \le j \le m\}.$$

We shall call the joint distribution of (X, Y) *graphical* if $\mathscr{Y}(i)$, $1 \le i \le n$, form the cliques of $G(Z^*)$. It will be shown later that many interesting joint probability models are in fact graphical.

A graph is called *decomposable* if its n' cliques $C(1), \ldots, C(n')$ can be indexed in a *compatible order* so that

$$S(i) = C(i) \cap \left\{ \bigcup_{\ell=1}^{i-1} C(\ell) \right\} \subseteq C(p(i)), \qquad 2 \le i \le n', \tag{2.2}$$

for some $p(i)$, $1 \le p(i) \le i-1$. Probability distributions on (X, Y) which have decomposable graphs will be central to this paper.

Decomposable graphs have been studied for some time and many of their properties are well known. One important result [see Lauritzen, Speed and

Vijagan (1984)] is that a graph is decomposable iff it contains no chordless cycle of length greater than or equal to 4. This enables condition (2.2) to be checked by eye. There are also quick ways of finding a compatible ordering of cliques from a given graph, for example, maximum cardinality search [see Tarjan and Yannakakis (1984); Lauritzen and Spiegelhalter (1988)]. It will be shown in Sections 4 and 5 that these results concerning graphical models enable the identification of the useful reparametrization of $\boldsymbol{\lambda}$ mentioned above.

Figure 1 gives a selection of graphs of graphical joint distributions. By the result above it is easy to check that all but G_1 and G_2 are decomposable. G_1 is not decomposable because it has the chordless four cycle $\{1, 2, 3, 4, 1\}$ while G_2 is not decomposable because of the chordless five cycle $\{1, 2, 3, 4, 6, 1\}$.

3. Quasi-incremental joint distributions. In this section a useful class of joint distributions on (X, Y) is considered which contains models in which, in the very weak sense defined below, X is increasing in Y. First some notation: let the *residuals* $\bar{\mathscr{Y}}(i)$ of Y on X be defined by

$$\begin{aligned} &\bar{\mathscr{Y}}(1) = \mathscr{Y}(1), \\ &\bar{\mathscr{Y}}(i) = \mathscr{Y}(i) \backslash \bigcup_{\ell=1}^{i-1} \mathscr{Y}(\ell), \qquad 2 \le i \le n, \end{aligned} \tag{3.1}$$

where for two sets A and B, $A\backslash B$ denotes the set of elements of A which are not in B and where $\{\mathscr{Y}(i)\}$ are the range sets of Y given X defined in Section 2. If $\mathscr{Y} = \{y_1, \ldots, y_m\}$, the set of possible values of Y, then clearly $\{\bar{\mathscr{Y}}(1), \ldots, \bar{\mathscr{Y}}(n)\}$ form a partition of $\mathscr{Y}$.

DEFINITION 3.1. Say Y is *recursive* in X if for some $p(i) < i$ the following statements hold:

(i) $\bar{\mathscr{Y}}(i) \neq \phi,\ 1 \le i \le n$;
(ii) $\mathscr{Y}(i) \backslash \bar{\mathscr{Y}}(i) \subseteq \mathscr{Y}(p(i)),\ 2 \le i \le n$.

[Henceforth let $p(i)$ denote the least index for i with the property above.]

(iii) Say Y is *quasi-incremental* in X if it is recursive in X and for each $y_j \in \mathscr{Y}(i) \backslash \bar{\mathscr{Y}}(i)$, $1 \le j \le m$, $2 \le i \le n$,

$$P(Y = y_j \mid X = x_{p(i)}) \ge P(Y = y_j \mid X = x_i)$$

with strict inequality for some value of $y_j \in \mathscr{Y}(p(i))$.

The notion of Y being quasi-incremental in X is pertinent when the range $\mathscr{Y}$ of Y is larger than the range of X and the joint probability table of (X, Y) is sparse in nonzero entries.

Condition (i) demands that as the value x_{i-1} is increased to x_i, $2 \le i \le n$, at least one "higher" value of Y becomes a possibility. Condition (ii) requires that the collection of values in $\mathscr{Y}(i) \backslash \bar{\mathscr{Y}}(i)$ be contained in a single lower indexed range set $\mathscr{Y}(p(i))$. Thus, informally as x_{i-1} is increased to x_i the range $\mathscr{Y}(p(i))$, $1 \le p(i) \le i - 1$, is shifted up to $\mathscr{Y}(i)$.

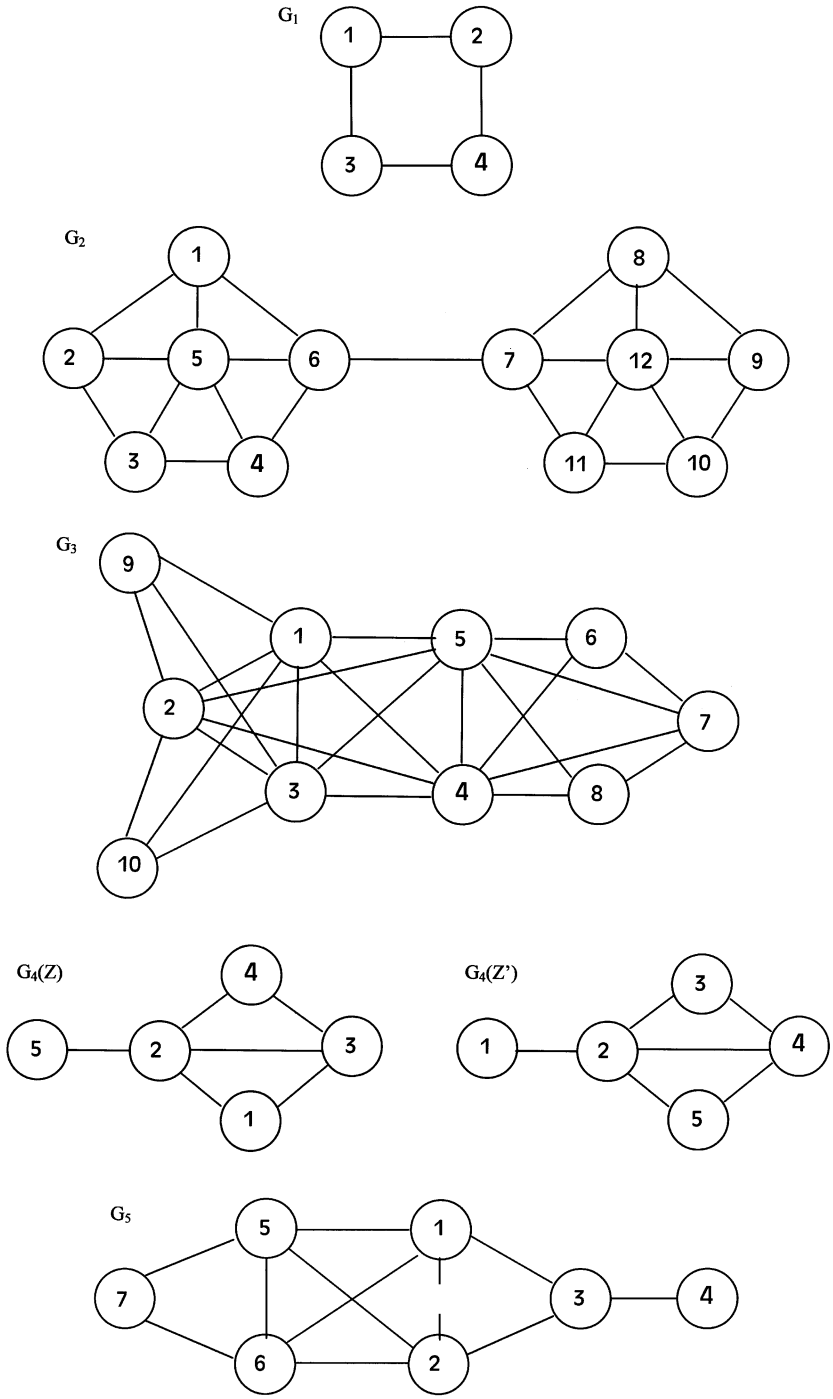

FIG. 1. *Some graphical models.*

The final condition (iii) concerns the magnitude of probabilities of observing $y_j \in \mathscr{Y}(p(i)) \cap \mathscr{Y}(i)$ given $x_{p(i)}$ compared with x_i. It states that were one to believe that x_i and $x_{p(i)}$ were equally probable, then one would believe $(y_j, x_{p(i)})$, the pair of values associated with the lower value of X, to be at least as probable as (y_j, x_i). In this sense the relationship between (X, Y) could be said to be positively skewed. Notice that by fixing the matrix Z, the ratios of the two conditional probabilities on either side of the inequalities in (iii) are fixed. It can also be seen that, under the dnh, by dividing (iii) by the expression for λ_j in (1.2), condition (iii) is automatically satisfied provided that the containment condition in (ii) is always strict.

It is easy to check whether a matrix Z corresponds to a quasi-incremental distribution on (X, Y). Explicitly, if $Z = \{z_{ij}\}$, then z_{ij} should have the following properties.

1. There exist values $t(1), \ldots, t(n)$, with $t(\ell) > 0, 1 \le \ell \le n$ and $\sum_{\ell=1}^{n} t(\ell) = m$, such that $z_{1j} = 1$ for $1 < j \le t(1)$, $z_{1j} = 0$ for $t(1) < j \le m$.
2. In addition, for i, $2 \le i \le n$,
   (a) there exists a row $p(i)$, $1 < p(i) < i$, such that

$$0 \le z_{ij} \le z_{p(i),\,j} \le 1, \qquad 1 \le j \le \sum_{\ell=1}^{i-1} t(\ell)$$

   with strict inequality for some j, such that $\sum_{\ell=1}^{p(i)-1} t(\ell) < j < \sum_{\ell=1}^{p(i)} t(\ell)$;
   (b) $z_{ij} = 1$ for j such that $\sum_{\ell=1}^{i-1} t(\ell) < j \le \sum_{\ell=1}^{i} t(\ell)$;
   (c) $z_{ij} = 0$ for $\sum_{\ell=1}^{i} t(\ell) < j \le m$.

   Here $t(\ell)$, $1 \le \ell \le n$, is just the number of values in $\bar{\mathscr{Y}}(\ell)$.
   Of course, when X is a function of Y,

$$\mathscr{Y}(i) = \bar{\mathscr{Y}}(i), \qquad 1 \le i \le n,$$

so Y is trivially quasi-incremental in X. Another way of interpreting this class of models is that they allow some skewed noise into a functional relationship between X and Y.

The following result shows that the more interesting recursive distributions are graphical and furthermore that there is a very close link between these joint distributions and decomposable graphs.

THEOREM 3.1. *Suppose no range set of Y contains any other and Y is recursive in X. Then the following statements hold*:

(i) *The joint distribution of X, Y is graphical.*
(ii) *The graph of $Z(X, Y)$ is decomposable.*

Furthermore, given $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$, every decomposable graph is the graph of a unique dnh distribution on (X, Y).

PROOF. (i) Go by contradiction and suppose G is not graphical. Since, by hypothesis, no range set is contained in any other, this supposes that G has

a clique which is not itself a range set. So, in particular, there must exist a value of the index i, $1 \le i \le n$, and two possible values of Y, $y^{(1)}$ and $y^{(2)}$, say, such that

$$y^{(1)} \in \bar{\mathscr{Y}}(i) \quad \text{and} \quad y^{(2)} \in \bigcup_{\ell=1}^{i-1} \mathscr{Y}(\ell) \backslash \mathscr{Y}(i)$$

with $y^{(1)}$ and $y^{(2)}$ connected by an edge in G. However, for $y^{(1)}$ and $y^{(2)}$ to be connected by an edge, they must lie in some range set $\mathscr{Y}(i')$, $i' > i$. Choose i' to be the smallest index with this property. Then it is easily seen that property (ii) of Definition 3.1 is violated for $\mathscr{Y}(i')$, so G must be graphical.

(ii) Note that $(\boldsymbol{\theta}, \boldsymbol{\lambda})$ and the range sets of Y on X define a unique dnh distribution. The result is now immediate from comparing (2.2) and Definition 3.1(ii). □

It will be shown in Section 5 that the link between recursive distributions and decomposable graphs contained in Theorem 3.1 proves useful in defining a straightforward conjugate Bayesian analysis for $\boldsymbol{\lambda}$.

4. Reparametrizing quasi-incremental distributions. In the Introduction it was mentioned that $\boldsymbol{\lambda}$ could often be reparametrized to allow a simple conjugate Bayesian analysis. Here is a useful reparametrization of a quasi-incremental distribution.

Notice that under conditions (i)–(iii) of Definition 3.1, if $y_j \in \bar{\mathscr{Y}}(i)$, then

$$P(Y = y_j \mid X = x_i) = \lambda_j, \qquad 1 \le j \le m. \tag{4.1}$$

Write

$$\dot{\lambda}_i = P(Y \in \bar{\mathscr{Y}}(i) \mid X = x_i), \tag{4.2}$$

$$\rho_k(i) = P(Y = y_j \mid X = x_i, y_j \in \bar{\mathscr{Y}}(i)), \tag{4.3}$$

where

$$j = \begin{cases} \sum_{\ell=1}^{i-1} t(\ell) + k, & \text{if } 2 \le i \le n, \\ k, & \text{if } i = 1, \end{cases} \tag{4.4}$$

such that $t(i) = \#\bar{\mathscr{Y}}(i)$ is the number of elements in $\bar{\mathscr{Y}}(i)$. Note that $\dot{\lambda}_i > 0$, $\rho_k(i) > 0$, $k = 1, \ldots, t(i)$, and $\sum_{k=1}^{t(i)} \rho_k(i) = 1$, $i = 1, \ldots, n$. Write

$$\boldsymbol{\rho} = (\boldsymbol{\rho}(1), \ldots, \boldsymbol{\rho}(n))^T \quad \text{where } \boldsymbol{\rho}(i) = (\rho_1(i), \ldots, \rho_{t(i)}(i)).$$

By the rules of probability we have that $\boldsymbol{\lambda}$ and $\boldsymbol{\rho}$ are related by the equations

$$\lambda_j = \dot{\lambda}_i \rho_k(i) \quad \text{for } y_j \in \bar{\mathscr{Y}}(i),\ 1 \le j \le m, \tag{4.5}$$

where the indices k and j are related as in (4.4). Substituting (4.1) into the probability constraint (1.10) gives $\dot{\lambda}_i$ as a function of $\boldsymbol{\rho}$. Thus

$$\dot{\lambda}_1 = 1$$

and

$$\dot{\lambda}_i = 1 - \sum_{j:\ y_j \in \bigcup_{\ell=1}^{i-1} \mathscr{Y}(\ell)} z_{ij}\lambda_j, \qquad 2 \le i \le n. \tag{4.6}$$

Equation (4.6) can be written as an explicit function of $\boldsymbol{\rho}$ by repeated substitution of λ_j by (4.5) and (4.6). In fact $\dot{\lambda}_i(\boldsymbol{\rho})$ is a polynomial function of $(\boldsymbol{\rho}(1), \ldots, \boldsymbol{\rho}(i-1))$.

It is easy but tedious to verify that condition (iii) of Definition 3.1 is necessary and sufficient to ensure that the probability constraints are equivalent to (4.6) and do not impose further constraints on the simplicies $(\boldsymbol{\rho}(1), \ldots, \boldsymbol{\rho}(n))$ [Queen (1991); Smith and Queen (1992)].

From (1.9), the likelihood of the reparametrization $\boldsymbol{\rho}$ of $\boldsymbol{\lambda}$ can now be written as

$$L_2(\boldsymbol{\rho}) = \prod_{i=1}^{n} \dot{\lambda}_i^{\dot{r}_i} \prod_{k=1}^{t(i)} [\rho_k(i)]^{r_k(i)} \tag{4.7}$$

such that

$$r_k(i) = r_j,$$

where k and j are related as in (4.4) and

$$\dot{r}_i = \sum_{k=1}^{t(i)} r_k(i),$$

where $\dot{\lambda}_i$ are defined as functions of $\boldsymbol{\rho}$ from (4.6) and $\boldsymbol{\rho}(i)$ satisfy the simplex constraints

$$\sum_{k=1}^{t(i)} \rho_k(i) = 1, \qquad \rho_k(i) > 0,\ 1 \le k \le t(i),\ 1 \le i \le n.$$

The likelihood above is functionally more complicated than the original. However, it is familiar, being a discrete mixture of multinomial likelihoods. The simple constraints on $\boldsymbol{\rho}$ allow common prior distributions like independent Dirichlets to be used in a Bayesian analysis. Furthermore, $\boldsymbol{\rho}$ is just a vector of conditional probabilities so that it is at least plausible to set a proper prior distribution over these.

The obvious choice of prior density $p_2(\boldsymbol{\rho})$ on $\boldsymbol{\rho}$ would be of the form

$$p_2(\boldsymbol{\rho}) \propto \prod_{i=1}^{n} \dot{\lambda}_i^{\dot{\alpha}_i} \prod_{k=1}^{t(i)} [\rho_k(i)]^{\alpha_k(i)},$$

where $\alpha_k(i) > 0$, $1 \le k \le t(i)$, $1 \le i \le n$, and

$$\dot{\alpha}_i = \sum_{k=1}^{t(i)} \alpha_k(i).$$

The posterior density after observing $\mathbf{r} = (r_1, \ldots, r_m)$ would then clearly take the same form with $\dot{\alpha}_i$ and $\alpha_k(i)$ replaced by $\dot{\alpha}_i + \dot{r}_i$ and $\alpha_k(i) + r_k(i)$, respectively, where $\dot{r}_i$ and $r_k(i)$ are defined above. We shall call this family *nested generalized Dirichlet* densities. Their moments are straightforward to calculate for moderate sizes of $N = \sum_{j=1}^{m} r_j$. For large N, since the posterior density is log concave, the posterior mode and its associated matrix of second derivatives of the log density are easy to calculate numerically.

Thus by specifying a matrix of probability ratios Z of the form above, it is possible to perform a Bayesian analysis on a family of distributions on (X, Y) consistent with this Z which is conjugate in the unknown probability vectors $(\boldsymbol{\theta}, \boldsymbol{\lambda})$. Two examples of how this conjugate analysis works out on the vector $\boldsymbol{\lambda}$ are given below.

EXAMPLE 4.1. Assume X takes three values (x_1, x_2, x_3) and Y the five values $(y_1, y_2, y_3, y_4, y_5)$ and that the matrix Z of probability ratios is given by

$$Z = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0.5 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}.$$

Here $\boldsymbol{\rho} = (\boldsymbol{\rho}(1), \boldsymbol{\rho}(2), \boldsymbol{\rho}(3))^T$, where $\boldsymbol{\rho}(1) = (\rho_1(1), \rho_2(1))$, $\rho_1(2) = 1$, $\boldsymbol{\rho}(3) = (\rho_1(3), \rho_2(3))$ and from (4.2) and (4.6),

$$(\dot{\lambda}_1, \dot{\lambda}_2, \dot{\lambda}_3) = (1, 1 - 0.5\rho_2(1), 0.5\rho_2(1))$$

and

$$(\dot{r}_1, \dot{r}_2, \dot{r}_3) = (r_1 + r_2, r_3, r_4 + r_5).$$

So by (4.7) we find that the likelihood $L_2(\boldsymbol{\rho})$ separates into

$$L_2(\boldsymbol{\rho}) = L_2^{(1)}(\boldsymbol{\rho}(1)) L_2^{(2)}(\boldsymbol{\rho}(3)),$$

where

$$L_2^{(1)}(\boldsymbol{\rho}(1)) = [\rho_1(1)]^{r_1}(1 - \rho_1(1))^{r_2 + r_4 + r_5}(1 - 0.5\rho_2(1))^{r_3}, \qquad 0 < \rho_1(1) < 1,$$

$$L_2^{(2)}(\boldsymbol{\rho}(3)) = \rho_1(3)^{r_4}(1 - \rho_1(3))^{r_5}, \qquad 0 < \rho_1(3) < 1.$$

The nested Dirichlet conjugate density sets $\boldsymbol{\rho}(1)$ and $\boldsymbol{\rho}(3)$ are a priori independent with $\boldsymbol{\rho}(3)$ having a beta density and $\boldsymbol{\rho}(1)$ having a generalized Dirichlet density [see Dickey, Jiang and Kadane (1987)]. A posteriori $\boldsymbol{\rho}(1)$ and $\boldsymbol{\rho}(3)$ remain independent, the beta and generalized Dirichlet density being updated in the usual fashion. The Bayesian analysis of the conditional probabilities

associated with y_4 and y_5 as governed by $\boldsymbol{\rho}(3)$ is particularly simple, since the associated values of Y are only possible if X takes the single value x_3.

EXAMPLE 4.2. This time we consider a family of joint probabilities on (X, Y) with range dimensions 4 and 7, respectively, and Z given by

$$Z = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 1 \end{bmatrix}.$$

The conjugate Bayesian analysis on $\boldsymbol{\lambda}$ proceeds as follows.

Here

$$\boldsymbol{\rho} = (\rho_1(1), \rho_2(1), \rho_1(2), \rho_2(2), \rho_1(3), \rho_2(3), 1)^T$$

and

$$(\dot{\lambda}_1, \dot{\lambda}_2, \dot{\lambda}_3, \dot{\lambda}_4) = (1, \rho_1(1), 1 - (1 - 0.5\rho_1(2))\rho_1(1), 1 - 0.5\rho_2(3)\dot{\lambda}_3)$$

$$L_2(\boldsymbol{\rho}) = L_2^{(1)}(\boldsymbol{\rho}(1))L_2^{(2)}(\boldsymbol{\rho}(2)|\boldsymbol{\rho}(1))L_2^{(3)}(\boldsymbol{\rho}(3)|\boldsymbol{\rho}(1)),$$

where

$$L_2^{(1)}(\boldsymbol{\rho}(1)) = [\rho_1(1)]^{r_1+r_3+r_4}[\rho_2(1)]^{r_2}, \qquad \rho_1(1) + \rho_2(1) = 1,$$

$$L_2^{(2)}(\boldsymbol{\rho}(2)|\boldsymbol{\rho}(1)) = \rho_1(2)^{r_3}\rho_2(2)^{r_4}\dot{\lambda}_3^{r_5+r_6}, \qquad \rho_1(2) + \rho_2(2) = 1,$$

$$L_2^{(3)}(\boldsymbol{\rho}(3)|\boldsymbol{\rho}(1), \boldsymbol{\rho}(2)) = \rho_1(3)^{r_5}\rho_2(3)^{r_6}[1 - 0.5\dot{\lambda}_3\rho_2(3)]^{r_7}, \qquad \rho_1(3) + \rho_2(3) = 1.$$

Now for simplicity assume that a priori $\rho_1(1)$, $\rho_1(2)$ and $\rho_1(3)$ have independent beta distributions so that their joint density takes the product form $f_1(\boldsymbol{\rho}(1)), f_2(\boldsymbol{\rho}(2)), f_3(\boldsymbol{\rho}(3))$. Then, the posterior joint density of $\boldsymbol{\rho}$ takes the form

$$f_3(\boldsymbol{\rho}(3)|\boldsymbol{\rho}(1), \boldsymbol{\rho}(2), \mathbf{r}) = I_3^{-1}L_2^{(3)}(\boldsymbol{\rho}(3)|\boldsymbol{\rho}(1), \boldsymbol{\rho}(2))f_3(\boldsymbol{\rho}(3)),$$

$$f_2(\boldsymbol{\rho}(2)|\boldsymbol{\rho}(1), \mathbf{r}) = I_2^{-1}I_3L_2^{(2)}(\boldsymbol{\rho}(2)|\boldsymbol{\rho}(1))f_2(\boldsymbol{\rho}(2)),$$

$$f_1(\boldsymbol{\rho}(1)|\mathbf{r}) = I_1^{-1}I_2L_2^{(1)}(\boldsymbol{\rho}(1))f(\boldsymbol{\rho}(1)),$$

where I_3, I_2 and I_1 are the proportionality constants that ensure $f_3(\cdot|\mathbf{r})$, $f_2(\cdot|\mathbf{r})$ and $f_1(\cdot|\mathbf{r})$ integrate to unity, I_2 being a function of $\boldsymbol{\rho}(1)$ and $\mathbf{r}$ and I_3 being a function of $\boldsymbol{\rho}(1)$, $\boldsymbol{\rho}(2)$ and r.

Note that $f_3(\boldsymbol{\rho}(3)|\boldsymbol{\rho}(1), \boldsymbol{\rho}(2), \mathbf{r})$ is a generalized Dirichlet density. Furthermore, since $\dot{\lambda}_3$ is a polynomial in $\boldsymbol{\rho}(1)$ of degree 2, I_3 is a polynomial in $\boldsymbol{\rho}(1)$ of degree $2r_7$, making f_2 a discrete mixture of generalized Dirichlet densities.

Similarly I_2 is a polynomial of degree $2(r_5+r_6+r_7)$ so a posteriori $f(\boldsymbol{\rho}(1)|\mathbf{r})$ is a mixture of generalized Dirichlets.

Note that the posterior margins on the space $\boldsymbol{\rho}$ in such examples are often algebraically complex. However, it is straightforward, if tedious, to calculate the posterior moments of $\boldsymbol{\rho}$ (and hence $\boldsymbol{\lambda}$) explicitly [see Geng and Asano (1989) for an explicit demonstration of such methods in an analogous problem]. Alternatively, since under this reparametrization the posterior density on $\boldsymbol{\rho}$ is log concave, numerical methods for calculating various margins via Monte Carlo techniques [e.g., see Tanner and Wong (1987)] can be modified so that they converge very quickly [see, e.g., Gilks (1992)].

5. Choosing the simplest compatible ordering. In Section 2 we mentioned that the class of dnh models had been used by other authors to model conservatively the relationship between X and Y. If the dnh property holds and Y is quasi-incremental in X, then the graph G of (X, Y) can be used to discover a transformation $(X, Y) \to (\tau_1(X), \tau_2(Y))$ which gives rise to the most tractable reparametrization of the vector $\boldsymbol{\lambda}$ of conditional probabilities. We now show that these reparametrizations assign to $\boldsymbol{\lambda}$ a family of distributions which contains the class of hyper-Dirichlet distributions [Dawid and Lauritzen (1993)] as a special case. Tarjan and Yannakakis' (1984) algorithm implies that if a graph is decomposable, there are at least n compatible orderings associated with it, where n is the number of its cliques. A compatible ordering of nodes in G can therefore be chosen which defines a transformation (τ_1, τ_2) on (X, Y) which makes $(\tau_1(X), \tau_2(Y))$ recursive through the equivalence of (2.2) and Definition 3.1(ii). This ordering can be found quickly from G by using a maximal cardinality search or in simple problems, by eye. The reparametrization given in Section 4 may not only depend upon $G(Z)$, but also on Z as well.

For example, the matrix Z given below gives rise to a quasi-incremental distribution on (X, Y); however, so does Z', and both Z and Z' define the same class of models.

$$Z = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix}, \qquad Z' = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \end{bmatrix}.$$

The graphs $G_4(Z)$ and $G_4(Z')$ are given in Figure 1. The transformation $(X, Y) \to (\tau_1(X), \tau_2(Y))$ is given by

$$(\tau_1(x_1), \tau_1(x_2), \tau_1(x_3)) = (x_3, x_2, x_1),$$

$$(\tau_2(y_1), \tau_2(y_2), \tau_2(y_3), \tau_2(y_4), \tau_2(y_5)) = (y_5, y_2, y_4, y_3, y_1).$$

Following Section 4, the transformations of $\boldsymbol{\lambda}$ to $\boldsymbol{\rho}$ and $\boldsymbol{\rho}'$ with respect to compatible orderings associated with Z and Z', respectively, are different, namely,

$$L_2(\boldsymbol{\rho}) = [\rho_1(1)]^{r_1+r_4}[\rho_2(1)]^{r_2}[\rho_3(1)]^{r_3}(\rho_1(1) + \rho_3(1))^{r_5},$$

$$\rho_1(1) + \rho_2(1) + \rho_3(1) = 1,$$

and

$$L_2(\boldsymbol{\rho}') = \rho_1'(1)^{r_1+r_3+r_4+r_5}\rho_2'(1)^{r_2}\rho_1'(2)^{r_3+r_5}\rho_2'(2)^{r_4},$$

$$\rho_1'(1)+\rho_2'(1)=1,\ \ \rho_1'(2)+\rho_2'(2)=1.$$

Unless there is some good reason to the contrary, the second reparametrization should be preferred to the first. The fact that it leads to a reparametrization in terms of *independent* beta processes not only makes the interpretation of the model easier, but also simplifies the math.

The next result gives sufficient conditions for when a conjugate analysis on $\boldsymbol{\lambda}$ can be performed in which each component of $\boldsymbol{\lambda}$ is expressed as a product of terms in $\boldsymbol{\rho} = (\boldsymbol{\rho}(1), \ldots, \boldsymbol{\rho}(n))$ and where each $\boldsymbol{\rho}(i)$ are a priori independent of each other. In this case $\boldsymbol{\rho}(i)$, $1 \le i \le n$, have a Dirichlet distribution both a priori and a posteriori.

First a definition. Suppose a decomposable graph G has cliques $(C(1), C(2), \ldots, C(n))$, $S(i)$ and $p(i)$ are defined in (2.2) and

$$R(i) = C(i)\backslash S(i).$$

DEFINITION 5.1. Call a compatible ordering of a decomposable graph G *simple* if for each clique $C(i)$, $2 \le i \le n$, one of the following conditions holds:

(i) $\#(S(i)) = \#C(p(i)) - 1$, $2 \le i \le n$;
(ii) $C(p(i))\backslash S(i) = R(p(p(i)))$, $2 \le p(p(i)) < p(i)$;
(iii) $C(p(i))\backslash S(i) = R^*$ or $C(1)\backslash R^*$, where the set R^* does not depend on i, $2 \le i \le n$, and $\phi \subset R^* \subset C(1)$.

The reparametrization of $\boldsymbol{\lambda}$ we advocate below uses the compatible ordering above on the graph of (X, Y), with $\mathscr{Y}(i) = C(i)$, $\bar{\mathscr{Y}}(i) = R(i)$ and $\mathscr{Y}(i)\backslash\bar{\mathscr{Y}}(i) = S(i)$, $1 \le i \le n$, following exactly the reparametrization given in Section 4, except that we replace $\boldsymbol{\rho}(1)$ by writing

$$\rho_k(1) = \begin{cases} \tau\sigma_k, & \text{if } k \in R^* \\ (1-\tau)\omega_k, & \text{if } k \in C(1)\backslash R^*, \end{cases}$$

where

$$\sum_{k\in R^*} \sigma_k = 1, \qquad \sigma_k > 0,\ k \in R^*,\ \boldsymbol{\sigma} = \{\sigma_k\colon k \in R^*\},$$

$$\sum_{k\in C(1)\backslash R^*} \omega_k = 1, \qquad \omega_k > 0,\ k \in C(1)\backslash R^*,\ \boldsymbol{\omega} = \{\omega_k\colon k \in R^*\}$$

and $0 < \tau < 1$. Clearly $\boldsymbol{\rho}(1) \to (\boldsymbol{\tau}, \boldsymbol{\sigma}, \boldsymbol{\omega})$ is invertible and hence so is $\boldsymbol{\lambda} \to (\boldsymbol{\tau}, \boldsymbol{\sigma}, \boldsymbol{\omega}, \boldsymbol{\rho}(2), \ldots, \boldsymbol{\rho}(n))$.

Furthermore, the new reparametrization has a simple interpretation in terms of the conditional distribution of Y on X. Thus

$$\tau = P(y_j \in R^* \mid x = x_1)$$

and, for each $y_j \in R^*$,

$$\sigma_j = P(y_j \mid x = x_1,\ y_j \in R^*)$$

and, for each $y_j \in Y(1)\backslash R^*$,

$$\omega_j = P(y_j \mid x = x_1,\ y_j \in Y(1)\backslash R^*).$$

THEOREM 5.1. *If the graph G of a joint distribution of (X, Y) is decomposable and admits a simple compatible ordering and (X, Y) satisfies the dnh, then the conditional probability vector $\boldsymbol{\lambda}$ can be reparametrized to $(\tau, \boldsymbol{\sigma}, \boldsymbol{\omega}, \boldsymbol{\rho}(2), \ldots, \boldsymbol{\rho}(n))$. If each of the $(n+2)$ parameter vectors is given an independent Dirichlet distribution, then posterior to observing $\mathbf{r}$, $(\tau, \boldsymbol{\sigma}, \boldsymbol{\omega}, \boldsymbol{\rho}(2), \ldots, \boldsymbol{\rho}(n))$ will remain independent and each will still have a Dirichlet distribution. Each component of $\boldsymbol{\lambda}$ will be expressed as a product of single components from $(\tau, \boldsymbol{\sigma}, \boldsymbol{\omega}, \boldsymbol{\rho}(2), \ldots, \boldsymbol{\rho}(n))$.*

Furthermore, two families of these prior distributions associated with different compatible parametrizations of G are equivalent in the sense that they give an identical family of distributions over $\boldsymbol{\lambda}$.

PROOF. It is sufficient to show that under the conditions of the theorem, $\dot{\lambda}_i$, $1 \le i \le n$, as defined in (4.2), is a product of terms in $(\tau, 1-\tau, \boldsymbol{\sigma}, \boldsymbol{\omega}, \boldsymbol{\rho}(2), \ldots, \boldsymbol{\rho}(n))$, since $\boldsymbol{\rho}(1)$ is clearly a product of terms in $(\tau, 1-\tau, \boldsymbol{\sigma}, \boldsymbol{\omega})$. Go by induction on i. The theorem is clearly true for $i = 1$ since $\dot{\lambda}_i = 1$, so suppose it is true for all ℓ, $1 \le \ell \le i-1$.

From Definition 5.1: if (i), then $\dot{\lambda}_i = \lambda_{j(i)}$, where $j(i)$ is the only element in $C(p(i))\backslash S(i)$ and so $\dot{\lambda}_i = \dot{\lambda}_s \rho_k(s)$ for some $s < i-1$ such that $j(i) = \sum_{r=1}^{s-1} t(r) + k$; if (ii), then $\dot{\lambda}_i = \dot{\lambda}_{p(p(i))}$, $p(p(i)) < i$; if (iii), then $\dot{\lambda}_i = \tau$ or $(1-\tau)$. In all these cases under the inductive hypothesis, $\dot{\lambda}_i$ takes the right product form and so the first part of the theorem is proved.

To prove uniqueness, suppose G admits a simple compatible ordering. In this ordering, let $\mathscr{Y}(i)\backslash\bar{\mathscr{Y}}(i)$ have $q(i)$ elements, $\bar{\mathscr{Y}}(i)$ have $t(i)$ elements and write $\boldsymbol{\lambda}(i) = (\boldsymbol{\lambda}^{(1)}(i), \boldsymbol{\lambda}^{(2)}(i))$, where for $2 \le i \le n$, $\boldsymbol{\lambda}_k^{(1)}(i)$, $1 \le k \le q(i)$, is the kth lowest indexed component of $\boldsymbol{\lambda}$ lying in $\mathscr{Y}(i)\backslash\bar{\mathscr{Y}}(i)$ and $\boldsymbol{\lambda}_\ell^{(2)}(i)$, $1 \le \ell \le t(i)$, is the ℓth lowest indexed component of $\bar{\mathscr{Y}}(i)$.

By the construction above we can set, a priori, $\boldsymbol{\lambda}^{(2)}(i)$ to be dependent on $\boldsymbol{\lambda}^{(1)}(i)$ only through the *sum* of the components of $\boldsymbol{\lambda}^{(1)}(i)$, that is,

$$\boldsymbol{\lambda}^{(2)}(i) | \boldsymbol{\lambda}^{(1)}(i) = \boldsymbol{\lambda}^{(2)}(i) | \left(\sum_{k=1}^{q(i)} \lambda_k^{(1)}(i) = 1 - \dot{\lambda}_i \right), \qquad 2 \le i \le n, \tag{5.1}$$

and we can also set $\boldsymbol{\lambda}^{(2)}(i) | \boldsymbol{\lambda}^{(1)}(i)$ to have an arbitrary Dirichlet distribution a priori. It is immediate from the properties of the Dirichlet distribution [Johnston and Kolz (1972)] that under the construction above the vector $\boldsymbol{\lambda}(1)$ of components in $\mathscr{Y}(1)$ has been set to have an arbitrary Dirichlet distribution as a prior. By induction on i, it follows that, since $\boldsymbol{\lambda}^{(1)}(i)$ is a subvector of $\boldsymbol{\lambda}(p(i))$, $2 \le i \le n$ and so $(1-\dot{\lambda}_i)^{-1}\boldsymbol{\lambda}^{(1)}(i)$ is Dirichlet [Johnson and Kotz

(1972)] and (5.1) holds that $\boldsymbol{\lambda}(i)$ has a Dirichlet distribution [Johnson and Kotz (1972)] $2 \leq i \leq n$.

It follows from Dawid and Lauritzen [(1993) page 1304] that our prior family over $(\tau, \boldsymbol{\sigma}, \mathbf{w}, \boldsymbol{\rho}(2), \ldots, \boldsymbol{\rho}(n))$ gives the (uniquely specified) hyper-Dirichlet family of prior distributions over $\boldsymbol{\lambda}$. Since this is true for any simple compatible ordering, the uniqueness is proven.

EXAMPLE 5.1. It is easy to check that G_5 of Figure 1 admits a simple compatible ordering. Thus transform (X, Y) so that $\mathscr{Y}(1) = \{y_3, y_4\}$, $\mathscr{Y}(2) = \{y_1, y_2, y_3\}$, $\mathscr{Y}(3) = \{y_1, y_2, y_5, y_6\}$ and $\mathscr{Y}(4) = \{y_5, y_6, y_7\}$. Under this ordering $C(2)$ and $C(3)$ satisfy (i) and $C(4)$ satisfies (ii) of Definition 5.1:

$$L_2(\boldsymbol{\lambda}) = [\rho_1(2)]^{r_1}[\rho_2(2)]^{r_2}[\rho_1(3)]^{r_5}[\rho_2(3)]^{r_6}\tau^{(r_1+r_2+r_4+r_7)}(1-\tau)^{(r_3+r_5+r_6)},$$

where

$$(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6, \lambda_7) = (\rho_1(2)\tau, \rho_2(2)\tau, 1-\tau, \tau, \rho_1(3)(1-\tau), \rho_2(3)(1-\tau), \tau).$$

Here σ and ω are simply 1.

Sometimes it is rather difficult to check by eye for a compatible ordering of a graph which has this property. The following construction is often useful.

DEFINITION 5.2. Call a graph H the *contraction* of a graph G if it has the following properties.

(a) The nodes of H are maximal complete subsets of nodes of G which share the same neighbors (other than themselves) in G.

(b) An edge between nodes J_1 and J_2 in H exists iff there is an edge $(j_1, j_2) \in G$ for every $j_1 \in J_1$ and $j_2 \in J_2$.

In Figure 2, H_3 is the contraction of the decomposable graph G_3 of Figure 1. A contraction H is useful because: (1) H is no more complicated than G and (2) under the obvious mapping, the cliques of H correspond to the cliques of G, as do their intersections and complements.

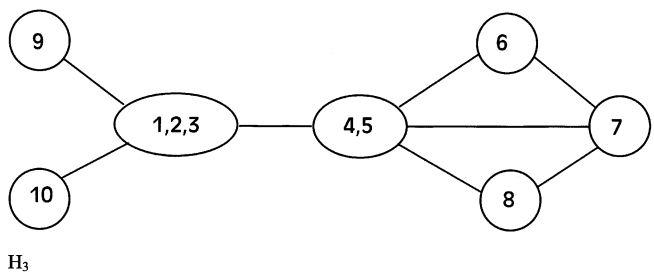

FIG. 2. *A contraction graph of* G_3.

Because of the property (b) it is possible to use H to identify an independent Dirichlet breakdown of probabilities of (X, Y) with an associated G, even when G is quite complicated.

EXAMPLE 5.2. Once H_3 is constructed from G_3 of Figure 1 it is clear that the clique $\{1, 2, 3, 4, 5\}$ of G_3 $[\{\{1, 2, 3\}, \{4, 5\}\}$ of $H_3]$ breaks G_3 into two components. Thus in the construction of Theorem 5.1 set $R^* = \{1, 2, 3\}$,

$$C(1) = \{1, 2, 3, 4, 5\}, \qquad C(2) = \{4, 5, 6, 7\}, \qquad C(3) = \{4, 5, 7, 8\},$$

$$C(4) = \{1, 2, 3, 9\}, \qquad C(5) = \{1, 2, 3, 10\}.$$

Then

$$L_2(\boldsymbol{\lambda}) = \sigma_1^{r_1}\sigma_2^{r_2}\sigma_3^{r_3}\omega_4^{r_4}\omega_5^{r_5}[\rho_1(2)]^{r_6+r_8}[\rho_2(2)]^{r_7}\tau^{r_1+r_2+r_3+r_6+r_7+r_8}(1-\tau)^{r_4+r_5+r_9+r_{10}},$$

where $\sum_{j=1}^{3}\sigma_j = 1$, $\sum_{j=4}^{5}\omega_j = 1$ and $\sum_{j=1}^{2} p_j(2) = 1$, $0 \le \tau \le 1$. The transform of $\boldsymbol{\lambda}$ is

$$\boldsymbol{\lambda} = (\tau\sigma_1, \tau\sigma_2, \tau\sigma_3, (1-\tau)\omega_4, (1-\tau)\omega_5, \tau\rho_1(2), \tau\rho_2(2), \tau\rho_1(2), (1-\tau), (1-\tau)).$$

6. Two examples of the use of fixed likelihood ratio models. One of the main features of fixed likelihood models is that all odds ratios $\mathrm{OR}(i, i', j, j')$, $1 \le i, i' \le n$, $1 \le j, j' \le m$,

$$\mathrm{OR}(i, i', j, j') = \frac{P(Y = y_j|X = x_i)}{P(Y = y_j|X = x_{i'})}\,\frac{P(Y = y_{j'}|X = x_i)}{P(Y = y_{j'}|X = x_{i'})}, \tag{6.1}$$

are functions of the Z matrix, so all fixed likelihood ratio models have fixed odds ratios. Typically in the applications we have studied it is reasonable, at least in the first instance, to assume that Z is fixed in time but that both the margin $\boldsymbol{\theta}$ of X and the selected conditional probabilities of $Y|X$ encoded in $\boldsymbol{\lambda}$ move as a time series. The separation of (1.6) allows these two time series to be analyzed independently of one another.

The issue we address in this section is the scope of applicability of such models. We shall concentrate our attention on two areas. The first is derived from the statistical analysis of certain marketing models. This was the motivating example for this work and is now well studied [Queen (1994); Queen, Smith and James (1994)]. The second relates to the area of probability estimation in medical probabilistic expert systems.

EXAMPLE 6.1 (Statistics of marketing). In this application, Y labels a set of competing brands and X labels categories of purchasers of these brands. The distribution of X varies slowly from week to week reflecting the changing demography, aspirations, affluence and so on of the potential customers. However, brand sales are typically volatile, being subject to many stochastic covariates, which include promotions directed to retail outlets, who are rewarded financially for displaying a brand more prominently, and promotions

directed at the general customer such as TV advertising, money-off offers or larger pack offers.

Within this setting, notice first that there is some flexibility not only in the choice of Y—through which brands are included in the study—but also considerable modelling choice in how X, the category list of customers, is defined. So, in practice, X can be chosen so that many of the Z entries are zeros; indeed, this is recommended by various market researchers [see, e.g., Grover and Srinivasan (1987)]. For example, X could be defined in terms of the needs of the customer, where a brand will either satisfy the list of needs of the purchaser or not [see, e.g., Queen (1994)]. For the fixed likelihood ratio model to be valid it is necessary that, within the period of study, promotions which affect brand sales act in an even-handed way across types of purchaser in the sense that the odds ratio of (6.1) remains unchanged. In the types of markets we have studied, the empirical evidence available suggests that this invariance holds, at least approximately, and is plausible unless the promotions employed by a brand target a specific type of customer [Queen, Smith and James (1994)]. Again in a large number of markets it appears empirically that a good working hypothesis is that the joint distribution is quasi-incremental. The index i seems to be related to the sophistication of the brand required by that category of customer.

Despite these arguments, we are still left with the practical problem of how to specify the nonzero elements of Z. There are essentially two ways to do this: the first is empirical; the second is to introduce new modelling assumptions. For some products, at infrequent periods, large sample surveys are performed which allow for the direct measurement of the nonzero elements z_{ij} of Z so that

$$\hat{z}_{ij} = \frac{\#(x_i, y_j)}{\#(x_{i^*(j)}, y_i)} \left[\frac{\#(x_i)}{\#(x_{i^*(j)})} \right]^{-1}, \qquad 1 \le i \le n, 1 \le j \le m,$$

where $x_{i^*(j)}$ is the type of purchaser most likely to buy brand y_j and $\#(x_i, y_j)$ is the number of purchases made by customer type x_i of brand y_j and so on. The vectors $(\boldsymbol{\lambda}, \boldsymbol{\theta})$ can then be allowed to develop as a time series while Z is held fixed to allow prediction of future brand sales.

When such sampling data are not available it is possible to derive Z from a behavioral model like the one outlined below. A customer walks into a shop and with probability θ_i she is of type i. She first picks one of the brands ($\tilde{Y} = y_j$) with probability $\tilde{\psi}_j$ which does not depend on i, where $\sum_{j=1}^{m} \tilde{\psi}_j = 1$. After looking at the selected brand, if she decides it is appropriate [i.e., $y_j \in \mathscr{Y}(i)$], she buys it; if it is inappropriate, she replaces it on the shelf and picks another brand $y_{j'}$ with probability $\tilde{\psi}_{j'}(1 - \tilde{\psi}_j)^{-1}$. She repeats this process until she has bought a brand.

Under this model, X and $\tilde{Y}$ are independent. If $y_j \notin \mathscr{Y}(i)$,

$$z_{ij} = \frac{P(Y = y_j | X = x_i)}{P(Y = y_j | X = x_{i^*(j)})} = 0.$$

On the other hand, if $y_j \in \mathscr{Y}(i)$,

$$z_{ij} = \frac{P(Y = y_j | X = x_i)}{P(Y = y_j | X = x_{i^*(j)})}$$
$$= \frac{P(\tilde{Y} = y_j | X = x_i)}{P(\tilde{Y} \in \mathscr{Y}(i) | X = x_i)} \frac{P(\tilde{Y} \in \mathscr{Y}(i^*(j)) | X = x_{i^*(j)})}{P(\tilde{Y} = y_j | X = x_{i^*(j)})},$$

which, since X and $\tilde{Y}$ are independent,

$$= \frac{P(\tilde{Y} \in \mathscr{Y}(i^*(j)))}{P(\tilde{Y} \in \mathscr{Y}(i))}$$
$$= \frac{\pi(i^*(j))}{\pi(i)}, \qquad j \in \mathscr{Y}(i), 1 \le i \le n,$$

where $(1 - \pi(i))$ is the probability a customer of type i replaces her first chosen brand on the shelf. This probability can be estimated, at least in principle, directly from an experiment. If we are completely ignorant about these replacement probabilities, it seems reasonable to set them all equal and this then gives the dnh model discussed in detail in Section 5. It can be shown that when these replacement probabilities are not equal, for quasi-incremental distributions at least, it is sometimes possible to redefine the categories i so that they are (at least approximately) equal. This trick is not always available for other fixed likelihood ratio models, however. Finally, if $\boldsymbol{\pi} = (\pi(1), \ldots, \pi(n))$ is deemed uncertain, then its distribution can be updated in the light of $\mathbf{r}$ in the usual way using the conjugate predictive probabilities of $\mathbf{r}|\boldsymbol{\pi} = \mathbf{r}|Z$. However, it should be noted that there are technical reasons why the distribution of $\mathbf{r}|\boldsymbol{\pi}$ depends only weakly on $\boldsymbol{\pi}$ and in practice the posterior distribution of $\boldsymbol{\pi}|\mathbf{r}$ often tends to be very close to the prior.

EXAMPLE 6.2 (Environmental medicine). Here we let Y label the (set of) symptom(s) first reported to a doctor and X the patient's disease category. One reason why our class of models is suitable in this case is that there often exists a partial logical relationship between symptoms and diseases; that is, by definition a disease cannot be observed unless certain symptoms appear. Another reason is that, for elicitation purposes, the conditioning in our parametrization of symptom, given disease, is the right way round.

Consider the following hypothetical study on the effects of air pollution on health. Assume that various disease conditions make the individual more susceptible to certain subsets of pollutants. Suppose the first reported data are symptoms such as ulcers, headaches, dizziness, dyspnea, diarrhea and eczema, and that exposure to particular pollutants causes one or more of the symptoms in susceptible individuals but not in the insusceptible. As in the last example, $\boldsymbol{\lambda}$ will be treated as an unknown and stochastic; that is, the probability that a most susceptible individual exhibits certain symptoms will change in time (due to differing effects of pollution, in combination with other factors,

over time), in location and in category of patient. Furthermore, the probabilities $\boldsymbol{\theta}$ of different diseases being observed is also modelled as depending on time, location and category of individual. However, we do assume that Z is fixed. In particular, this will imply that the *relative* toxicity of two pollutants measured over different categories of disease, as measured by the odds ratio, will be invariant—in this context a reasonable assumption at least in the first instance.

As in the last example, the fixed likelihood ratio model can be seen as the product of censoring—in this case the censoring occurs when individuals exposed to a potentially toxic substance are immune and so are not counted. Thus let $\tilde{Y} = y_j$ correspond to an individual's first exposure to a pollutant which can cause symptoms y_j, $1 \leq j \leq m$. Assuming $\tilde{Y}$ is independent of X and that each individual will be exposed at random to a succession of values of y_j, the arguments in the last example give us a fixed likelihood ratio model provided that $\pi(i)$—the probability that an individual with disease i is exposed first to a pollutant to which she is susceptible—is known. Again, but perhaps less plausibly, if diseases are classified in a way which makes $\pi(i) = \pi(1)$, $2 \leq i \leq n$, then we have a dnh model.

7. Graphical dnh joint distributions without a simple ordering. When a graphical distribution on (X, Y) is not quasi-incremental, the constraints (1.10) can become more active. Strange implications can then arise which suggest that fixed likelihood ratio models might be dubious. Consider the graphical dnh models defined by graphs G_1 and G_2 of Figure 1.

Using the reparametrization of $\boldsymbol{\lambda}$ to $\boldsymbol{\rho}$ as defined in Sections 4 and 5, a reparametrization of G_1 gives $\boldsymbol{\rho} = (\tau, 1-\tau, 1, 1)$. This is one dimensional rather than the expected $m - n = 0$ dimensional solution space and has $\boldsymbol{\lambda}$ of the form

$$\lambda = (\tau, 1-\tau, 1-\tau, \tau).$$

It is easily checked that the additional dimension in λ arises because the probability vector $\boldsymbol{\theta}$ on the margins of X is unidentifiable. Notice, however, that a conjugate beta analysis can be performed on $\boldsymbol{\lambda}$ if this dnh model is considered appropriate.

Even stranger, it can easily be shown that no dnh model on (X, Y) can *exist* which has range sets defined by G_2. By using the reparametrization of Section 4, where $C(1) = \{6, 7\}$, $C(2) = \{1, 5, 6\}$, $C(3) = \{1, 2, 5\}$, $C(4) = \{2, 3, 5\}$, $C(5) = \{3, 4, 5\}$, $C(6) = \{4, 5, 6\}$, $C(7) = \{7, 8, 12\}$, $C(8) = \{8, 9, 12\}$, $C(9) = \{9, 10, 12\}$, $C(10) = \{10, 11, 12\}$ and $C(11) = \{7, 11, 12\}$, set

$$\boldsymbol{\rho} = \{\rho_1(1), \rho_2(1), \rho_1(2), \rho_2(2), 1, 1, 1, \rho_1(7), \rho_2(7), 1, 1, 1\},$$

where $\rho_1(1), \rho_1(2), \rho_1(7) < 1$ and $\sum_{j=1}^{2} \rho_j(2) = \sum_{j=1}^{2} \rho_j(7) = 1$.

Unfortunately though, since (X, Y) is not quasi-incremental there are two extra constraints from (1.10), namely,

$$\rho_2(1)\rho_1(1) = \rho_1(1)$$

and

$$\rho_1(1)\rho_1(7) = \rho_2(1).$$

Estimating $\boldsymbol{\rho}(1)$ from these equations gives

$$(1 + \rho_1(2))^{-1} + (1 + \rho_1(7))^{-1} = 1,$$

which has no solution if $\rho_1(2)$ and $\rho_1(7)$ are strictly positive. Other examples can be constructed when the dnh is not assumed where under fixed likelihood ratio models such non-existence problems arise.

These examples illustrate that fixed likelihood ratio models may have undesirable modelling implications. Such models seem best suited to be used in conjunction with quasi-incremental distributions.

Acknowledgment. We would like to thank the referees for their very helpful comments on an earlier version of this paper.

REFERENCES

CARLSON, B. C. (1977). *Special Functions of Applied Mathematics*. Academic Press, New York.

DAWID, A. P. and DICKEY, J. M. (1977). Likelihood and Bayesian inference from selectively reported data. *J. Amer. Statist. Assoc.* **72** 845–850.

DAWID, A. P. and LAURITZEN, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21** 1272–1317.

DECHTER, R., DECHTER, A. and PEARL, J. (1990). Optimisation in constraint networks. In *Influence Diagrams, Belief Nets and Decision Analysis* (R. M. Oliver and J. Q. Smith, eds.) 411–425. Wiley, New York.

DECHTER, R. and PEARL, J. (1987). Network-based heuristics for constraint satisfaction problems. *Artificial Intelligence* **34** 1–38.

DICKEY, J. M. (1983). Multiple hypergeometric functions: probabilistic interpretations of statistical uses. *J. Amer. Statist. Assoc.* **78** 628–637.

DICKEY, J. M., JIANG, J. and KADANE J. B. (1987). Bayesian methods for censored categorical data. *J. Amer. Statist. Assoc.* **82** 773–781.

GENG, Z. and ASANO, C. (1989). Bayesian estimation methods for catagorical data with misclassification. *Comm. Statist. Theory Methods* **18** 2935–2954.

GILKS, W. R. (1992). Bayes derivative-free adaptive rejection sampling for Gibbs sampling. In *Bayesian Statistics* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 641–649. Oxford Science Publications.

GRANDY, W. T., JR. (1985). Incomplete information in generalised inverse problems. In *Maximum Entropy and Bayesian Methods in Inverse Problems* (C. R. Smith and W. T. Grandy, Jr., eds.) 41–72. Reidel, Boston.

GROVER, R. and SRINIVASAN, V. (1987). A simultaneous approach to market segmentation and market structuring. *Journal of Marketing Research* **3** 139–153.

JOHNSON, N. L. and KOTZ, S. (1972). *Distributions in Statistics, Continuous Multivariate Distributions*. Wiley, New York.

LAURITZEN, S. L., SPEED, T. P. and VIJAGAN, K. (1984). Decomposable graphs and hypergraphs. *J. Austral. Math. Soc. Ser. A* **36** 12–29.

LAURITZEN, S. and SPIEGELHALTER, D. J. (1988). Local computations with probabilities on graphical structures and their applications to expert systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **50** 157–224.

QUEEN, C. M. (1991). Bayesian graphical forecasting models for business time series. Ph.D. dissertation, Univ. Warwick.

QUEEN, C. M. (1994). Using the multi-regression dynamic model to forecast brand sales in a competitive product market. *The Statistician* **43** 87–98.

QUEEN, C. M., SMITH, J. Q. and JAMES, D. M. (1994). Bayesian forecasts in markets with overlapping structures. *International Journal of Forecasting* **10** 209–233.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592.

SMITH, J. Q. and QUEEN, C. M. (1992). Bayesian models of partially segmented markets. Research Report 232, Univ. Warwick.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–550.

TARJAN, R. E. and YANNAKAKIS, M. (1984). Simple linear time algorithms to test chordality of graphs, test acyclicity of hypergraphs and selectively reduce acyclic hypergraphs. *SIAM J. Comput.* **13** 566–579.

VARDI, Y. and LEE, D. (1993). From image deblurring to optimal investments: maximum likelihood solutions for positive linear inverse problems (with discussion). *J. Roy. Statist. Soc. Ser. B* **55** 569–612.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WARWICK
COVENTRY CV4 7AL
ENGLAND

INSTITUTE OF MATHEMATICS AND STATISTICS
UNIVERSITY OF KENT
CANTERBURY
KENT CT2 7NF
ENGLAND
E-MAIL: c.m.queen@ukc.ac.uk