To appear in *Communications in Statistics: Theory and Methods*, Vol 27, Issue 2, 1998.

THE ADMISSIBILITY OF THE MAXIMUM LIKELIHOOD ESTIMATOR FOR DECOMPOSABLE LOG-LINEAR INTERACTION MODELS FOR CONTINGENCY TABLES

Glen Meeden
School of Statistics
University of Minnesota
Minneapolis, MN 55455

Charles Geyer
School of Statistics
University of Minnesota
Minneapolis, MN 55455

Joseph Lang
Department of Statistics
University of Iowa
Iowa City, IA

Eiichiro Funo
Department of Economics
Kanto Gakuin University
Yokohama, Japan

Key Words: exponential families; graphical models; stepwise Bayes

ABSTRACT

It is well known that for certain log-linear interaction models for contingency tables, i.e. those that are decomposable, the maximum likelihood estimator can be found explicitly. In this note we will show that in such cases this estimator is admissible. The proof is based on a stepwise Bayes argument and is a generalization of a proof of the admissibility of the maximum likelihood estimator for the usual unconstrained multinomial model. It is then shown that this result is a special case of a result for discrete exponential families.

1. INTRODUCTION

In log-linear models for contingency tables maximum likelihood is the standard method of estimation. For certain *hierarchical models*, called *decomposable models*, (see Goodman (1970, 1971) and Andersen (1974)) it was shown in Haberman (1974) that the maximum likelihood estimator could be found explicitly. Darroch, Lauritzen and Speed (1980) considered a certain class of graphical models which contains the family of hierarchical models. They showed how the conditional independence of such decomposable hierarchical models was easily interpretable using graphical models. We will follow their notation.

In this note we will show that for these models the maximum likelihood estimator is admissible. We will use a stepwise Bayes argument to prove the result. The stepwise Bayes method was introduced in Johnson (1971), see also Wald and Wolfowitz (1951). It was named in Hsuan (1979). Using this method Alam (1979) proved the admissibility of the maximum likelihood estimator when estimating the standard unconstrained multinomial probabilities. Brown (1981) gave a complete class theorem for this problem using the stepwise Bayes idea. In section 2 we will briefly describe the stepwise Bayes method and show how it will be applied to the present problem. In section 3 we will introduce the necessary notation and then prove the admissibility of the maximum likelihood estimator for decomposable hierarchical models. In section 4 we will prove a more general result for discrete exponential families.

2. THE STEPWISE BAYES METHOD

The stepwise Bayes method is a simple, but powerful technique for proving admissibility. It is particularly useful for problems where the sample space contains a finite number of points. In these arguments a finite sequence of disjoint subsets of the parameter space is selected, where the order of the specified subsets is important. Associated with each subset of the parameter space there is a corresponding subset of the sample space. These subsets of the sample space are also disjoint and their union is the entire sample space. At each stage the subset of the sample space is the collection of sample points which are assigned positive probability by at least one parameter point in the subset of the parameter space for this stage and which have not appeared in any of the subsets of the sample space for the previous stages. For each pair of subsets the probability function can be normalized so that for each point in the subset of the parameter space the renormalized probability function sums to one over the corresponding subset of the sample space. This leads to a sequence of restricted problems. For each restricted problem a prior distribution is defined on the corresponding subset of the parameter space with the property that it assigns positive marginal probability to every point in the corresponding subset of the sample space. Then the Bayes procedure is found for each sample point in the corresponding subset of the sample space. This process yields a well defined procedure which must be admissible. To prove the admissibility of a given procedure one must select the sequence of subsets, their order, and the sequence of priors appropriately.

In what follows we will prove the admissibility of the maximum likelihood estimator for certain models. We will now show how the the sequence of subsets of the parameter space should be selected to facilitate this argument. We begin by introducing some notation. Consider a random variable X, which takes on a finite number of values, say m, where $m \geq 2$. Let $\theta \in \Theta$ be the parameter indexing a family of probability functions for X. Let x denote a typical value of X and let $P(x;\theta)$ denote a typical probability function. To avoid trivialities we assume for each x, there exists a $\theta \in \Theta$ such that $P(x;\theta) > 0$. Furthermore we assume a random sample of size n is to be used to estimate θ. Let $\mathcal{X}$ denote the sample space for this experiment, $\mathbf{x}$ a possible vector of observations and $Pr(\mathbf{x};\theta)$ the joint probability function.

We begin by showing how the argument works for estimating the probabilities in the standard multinomial model where the loss function is just the sum of the squared errors of the individual components. It will be sufficient to just consider the trinomial model.

To this end suppose X takes on three values, say 0, 1 and 2, with corresponding probabilities $1-\theta_1-\theta_2$, θ_1, θ_2. The parameter space is the two-dimensional simplex for $\theta = (\theta_1, \theta_2)$ where $\theta_1 \geq 0$, $\theta_2 \geq 0$ and $\theta_1 + \theta_2 \leq 1$.

Note that there are three distinct vectors which each get assigned probability one by a point in the parameter space. They are the vectors consisting of all 0's, all 1's and all 2's. The corresponding subsets of the parameter space are just the three vertices of the simplex. The vector of all 0's corresponds to the vertex (0,0) and so on. Note that for each restricted problem there is no need to renormalize the probability functions. These three points of Θ are the three subsets Θ which define the restricted problems in the first three steps of the argument. The particular order among the three does not matter however. For each restricted problem there is only one possible prior, the one which puts mass one on the given vertex and the Bayes estimator under this prior is just the maximum likelihood estimator.

In the next stage of the stepwise Bayes argument there are again three similar subsets of the parameter space which yield three restricted problems which can be considered in any order. They are the three edges of the simplex without their vertices. For example consider the subset of the parameter space $\Theta^* = \{ (\theta_1, \theta_2) \,:\, 0 < \theta_1 < 1 \text{ and } \theta_2 = 0 \}$, the edge joining (0,0) and (1,0). For this restriction of the parameter space the set of data vectors which get positive probability under Θ^* and which were not taken care in the first stage are all those vectors which consist of just 0's and 1's and where both 0 and 1 appear at least once. Let Λ denote this collection of data points. For the restricted problem, with sample space Λ and parameter space Θ^*

$$Pr^*(\mathbf{x}; \theta_1, 0) = Pr(\mathbf{x}; \theta_1, 0)/(1 - (1-\theta_1)^n - \theta_1^n) \tag{1}$$

is the renormalized probability function. The prior we take on Θ^* will be

$$\begin{aligned} \pi(\theta_1, 0) &\propto \{ 1 - (1-\theta_1)^n - \theta_1^n \}/\{ \theta_1(1-\theta_1) \} \\ &\propto \sum_{k=1}^{n-1} \binom{n}{k} \theta_1^{k-1} (1-\theta_1)^{n-1-k} \end{aligned} \tag{2}$$

which is a bounded function of θ_1 and so π can made into a density function by the proper choice of a normalizing constant. Now for this restricted problem, it is easy to see, that the posterior mean, for a given $\mathbf{x} \in \Lambda$, of θ_1 under this prior is just the proportion of 1's in $\mathbf{x}$, which is the maximum likelihood estimate of θ_1. In a similar way we see that the posterior mean of θ_2 is just its maximum likelihood estimate. The other two edges are handled in exactly the same way.

The next stage is the final step of the argument and there is just one subset of the parameter space to consider. We take the subset of the parameter space Θ^* to be the interior of the simplex. For this step the sample space for the restricted problem is just the set of all the remaining vectors in sample space, i.e. those where 0, 1 and 2 all appear at least once. Denote this set by Λ. Then for $(\theta_1, \theta_2) \in \Theta^*$ we see that $\lambda(\theta_1, \theta_2|\Lambda) = \sum_{\mathbf{x}\in\Lambda} Pr(\mathbf{x}; (\theta_1, \theta_2))$ is the normalizing constant for the probability function for our restricted problem. Then $\pi(\theta_1, \theta_2) \propto \lambda(\theta_1, \theta_2|\Lambda)/\{\,(1-\theta_1-\theta_2)\theta_1\theta_2\,\}$ is a bounded function over Θ^* and it is easy to check that its posterior mean is just the maximum likelihood estimate for $\theta = (\theta_1, \theta_2)$ for $\mathbf{x} \in \Lambda$.

This is a brief outline of the proof of admissibility of the maximum likelihood estimator for the trinomial problem. It should be emphasized again that both the order in which the subsets of the parameter space are selected and the prior distribution used at each step are important. Suppose for example in the trinomial problem after the first three subsets had been chosen we omitted the next three and then selected the last subset, the interior of the simplex and selected some proper prior distribution for the restricted problem. The resulting estimator would still be admissible but it would not be the maximum likelihood estimator. For proving the admissibility of the maximum likelihood estimator the generalization to an arbitrary multinomial problem is straightforward. Note that the proper sequence of subsets of the parameter space is first the vertices, next the relative interiors of the 1-dimensional faces and then the relative interiors of the 2-dimensional faces and so on, until at the last step we consider the interior of the entire simplex. After the first stage the prior is proportional to the ratio of two factors. In the numerator there is the function which is the renormalizing constant for the probability function for the restricted problem. In the denominator there is the product of the all nonzero θ_i's for the restricted problem. Note that this result is well known and the correct order is clear from the structure of the simplex.

In the two problems to be considered in the next two sections the correct order will be determined by the parameterization and is similar to the order for the multinomial problem described above. Before we begin we need one more bit of notation. Let Λ be a nonempty subset of $\mathcal{X}$. Let

$$\Theta(\Lambda) = \{\, \theta \in \Theta \,:\, \sum_{\mathbf{x} \in \Lambda} Pr(\mathbf{x};\theta) = 1 \,\} \tag{3}$$

and

$$\Theta^{+}(\Lambda) = \{\, \theta \in \Theta(\Lambda) \,:\, Pr(\mathbf{x};\theta) > 0 \text{ for every } \mathbf{x} \in \Lambda \,\} \tag{4}$$

Note that for a given Λ, $\Theta(\Lambda)$ may be empty and $\Theta^{+}(\Lambda)$ may be empty even when $\Theta(\Lambda)$ is not.

3. DECOMPOSABLE LOG-LINEAR MODELS

We will briefly review decomposable log-linear models for contingency tables. Our notation, for the most part, will follow that of Darroch, Lauritzen and Speed (1980). See them for additional details.

Let C denote a finite set of *classification criteria* or *factors*. For each $\gamma \in C$ let I_γ be the set of *levels* of the criterion or factor γ. $I = \prod_{\gamma \in C} I_\gamma$ is the set of *cells* in our table and a particular cell will be denoted $\mathbf{i} = (i_\gamma, \gamma \in C)$. Let $P(\mathbf{i})$ denote a probability function defined on I. Our sample space $\mathcal{X}$ is just the product space I^n. Given a random sample of size n from this distribution we let the *counts* $n(\mathbf{i})$ be the number of observations that fall in cell $\mathbf{i}$. Under this model the distribution of the counts becomes a multinomial distribution given by

$$Pr\{\, N(\mathbf{i}) = n(\mathbf{i}),\, \mathbf{i} \in I \,\} = \begin{pmatrix} n \\ n(\mathbf{i}),\ \mathbf{i} \in I \end{pmatrix} \prod_{\mathbf{i} \in I} P(\mathbf{i})^{n(\mathbf{i})} \tag{5}$$

For $a \subseteq C$, we consider the *marginal counts* $n(\mathbf{i}_a)$, where $n(\mathbf{i}_a)$ is the number of observations in the marginal cell $\mathbf{i}_a = (\, i_\gamma,\ \gamma \in a)$. That is $n(\mathbf{i}_a)$ is the count in the *marginal table*, where observations are only classified according to the criteria in a. Similarly we let $P(\mathbf{i}_a)$ denote the probability that any given observation belongs to the marginal cell $\mathbf{i}_a$.

It remains to describe how $P(\mathbf{i})$ is to be parameterized. One of the advantages of decomposable hierarchical models is that they can be parametrized in a very convenient fashion. Let $\mathcal{C} = \{\, a_1, \ldots, a_k \,\}$ be the *generating class* for some decomposable hierarchical model. We assume that $C = \cup_{a \in \mathcal{C}} a$ and without loss of generality that the model is connected. We may assume that the elements are ordered so that for $t = 2, \ldots, k$

$$a_t \cap (a_1 \cup \cdots \cup a_{t-1}) = a_t \cap a_{r_t} \tag{6}$$

for some $r_t \in \{\, 1, \ldots, t-1 \,\}$. It follows that

$$b_t = a_t \setminus (a_1 \cup \cdots \cup a_{t-1}) = a_t \setminus a_{r_t} \neq \emptyset$$

for $t = 2, \ldots, k$. Letting

$$c_t = a_t \setminus b_t = a_t \cap a_{r_t}$$

which is also nonempty since the model is assumed to be connected, it follows that

$$P(\mathbf{i}) = P(\mathbf{i}_{a_1}) \prod_{t=2}^{k} P(\mathbf{i}_{b_t} | \mathbf{i}_{c_t}) \tag{7}$$

which expresses the conditional independence of the model. There are, generally, many orderings satisfying (6). Hence the parameterization of $P(\mathbf{i})$ given in (7) is not unique. Often the terminology "log-linear" indicates that the cell probabilities are all positive. We however are considering the completion of the parameter space. Note that the above parameterization still makes sense for the completed parameter space.

In what follows we will assume that for the model given by $\mathcal{C}$ the probability function $P(\mathbf{i})$ is parameterized as in equation (7). In terms of our earlier notation $P(\mathbf{i};\theta) = P(\mathbf{i})$ where θ just represents all the marginal and conditional probability functions on the right hand side of (7). This is some subset of the simplex which is the parameter space for the usual unconstrained multinomial problem.

Darroch, Lauritzen and Speed gave an explicit formula for the maximum likelihood estimate (mle) $\hat{P}(\mathbf{i})$ of $P(\mathbf{i})$ based on n observations which did not depend on the parameterization of equation (7). However for the given parameterization it is easy to see that

$$\hat{P}(\mathbf{i}) = (n(\mathbf{i}_{a_1})/n) \prod_{t=2}^{k} n(\mathbf{i}_{b_t \cup c_t})/n(\mathbf{i}_{c_t}) \tag{8}$$

when $n(\mathbf{i}_{c_t}) > 0$ for $t = 2, \ldots, k$ and is 0 otherwise. We will now prove the admissibility of this estimator.

Theorem 1 *Consider the problem of using a random sample of n observations to estimate the cell probabilities of a decomposable hierarchical log-linear model, defined by the generating class $\mathcal{C} = \{ a_1, \ldots, a_k \}$, with $C = \cup_{a \in \mathcal{C}} a$. If the loss function is the sum of the component squared error losses, then the maximum likelihood estimator of the cell probabilities is admissible.*

Proof. Let the model be defined by the generating class $\mathcal{C} = \{ a_1, \ldots, a_k \}$. It suffices to consider only connected models since various connected components correspond to independent sets of factors and their probabilities as well as their estimates multiply. Then the parameter space, Θ, is defined by the right hand side of (7) and the maximum likelihood estimate is given in (8).

We will use a stepwise Bayes argument with the correct ordering of restricted problems. The argument is by induction and is similar to the one for the standard multinomial problem given earlier. That is, we first consider the vertices, next the 1-dimensional edges or faces, then the 2-dimensional faces and so on until all possible sample points have been considered. To help understand the general argument it will useful to keep in mind a special case. Suppose the model C contains four factors and each factor has three levels. Let α_i, β_i, γ_i and λ_i for $i = 1, 2, 3$ denote the possible levels of each of the four factors. Let $\mathcal{C} = \{\, \{1, 2\}, \{2, 3\}, \{1, 4\} \,\}$ be the generating class for the model. Then for a typical cell $(\alpha_{i_1}, \beta_{i_2}, \zeta_{i_3}, \lambda_{i_4})$ the parameterization of (7) becomes

$$P(\alpha_{i_1}, \beta_{i_2}, \zeta_{i_3}, \lambda_{i_4}) = P(\alpha_{i_1}, \beta_{i_2}) P(\zeta_{i_3} | \beta_{i_2}) P(\lambda_{i_4} | \alpha_{i_1})$$

We need to introduce some additional notation. Let $\mathbf{x}$ denote a fixed vector of observations of length n. Let $I(\mathbf{x})$ denote the set of cells that appear at least once in the vector $\mathbf{x}$. Let $\mathcal{C}$ be the generating class which leads to the parameterization given in (7). For any $a \in C$ let

$$I_a(\mathbf{x}) = \{\, \mathbf{i}_a \,:\, \mathbf{i} \in I(\mathbf{x}) \,\}$$

Let Λ be a specified set of vectors $\mathbf{x}$ then

$$I_a(\Lambda) = \{\, I_a(\mathbf{x}) \,:\, \mathbf{x} \in \Lambda \,\}$$

Let $b \subset C$ be some subset of factors. We now let

$$Z_b(\mathbf{i}) = \mathbf{i}_b$$

be just the coordinate projection of the whole table onto the factors of b. More generally if $b \subset a \subset C$ then $Z_b(\mathbf{i}_a) = \mathbf{i}_b$ is well defined as well.

For c_2 it follows that

$$I_{c_2}(\mathbf{x}) = \{\, Z_{c_2}(\mathbf{i}_{a_1}) \,:\, \mathbf{i}_{a_1} \in I_{a_1}(\mathbf{x}) \,\}$$

since $a_2 \cap a_1 = c_2 \subseteq a_1$. Now assuming $I_{c_2}(\mathbf{x})$ is nonempty we let $\mathbf{i}^*_{c_2}$ be a fixed member of this set and let

$$I_{b_2|\mathbf{i}^*_{c_2}}(\mathbf{x}) = \{\, Z_{b_2}(\mathbf{i}_{a_2}) \,:\, \mathbf{i}_{a_1 \cup a_2} \in I_{a_1 \cup a_2}(\mathbf{x}) \text{ and } Z_{c_2}(\mathbf{i}_{a_1}) = \mathbf{i}^*_{c_2} \,\}$$

For a fixed $\mathbf{i}^*_{c_2}$, $I_{b_2|\mathbf{i}^*_{c_2}}(\mathbf{x})$ is the set of levels which appear for the set of factors belonging to b_2 for some observation belonging to $\mathbf{x}$ whose levels for the factors c_2 are just the levels $i^*_{c_2}$. In the same way we can define $I_{c_t}(\mathbf{x})$ and $I_{b_t|\mathbf{i}^*_{c_t}}(\mathbf{x})$ for $t = 3, \ldots, k$. Finally we let $|I_{a_1}(\mathbf{x})|$ be the number of marginal cells belonging to $I_{a_1}(\mathbf{x})$. We define $|I_{c_1}(\mathbf{x})|$ and $|I_{b_t|\mathbf{i}^*_{c_t}}(\mathbf{x})|$ similarly. Let $m(\mathbf{x}) = (|I_{a_1}(\mathbf{x})|, |I_{c_2}(\mathbf{x})|, \ldots, |I_{c_k}(\mathbf{x})|)$ and for $t = 2, \ldots, k$ let $m_t(\mathbf{x})$ be the vector of length $|I_{c_t}(\mathbf{x})|$ consisting of the values $|I_{b_t|\mathbf{i}^*_{c_t}}(\mathbf{x})|$, corresponding to the distinct values $\mathbf{i}^*_{c_t}$ belong to $I_{c_t}(\mathbf{x})$, arranged in increasing order.

We now begin the admissibility proof. In the first stage we take care of all those samples where all the observations fall into one cell. For such samples $m(\mathbf{x})$ is just a vector of ones and the mle is the Bayes estimator of a prior which puts mass one on the appropriate point of the sample space. The proof will be completed by induction.

Let v be a vector of positive integers of length k where at least one value is greater than or equal to two. Consider all such v's where the set

$$\Lambda(v) = \{\mathbf{x} : m(\mathbf{x}) = v\}$$

is not empty. A typical stage of the stepwise Bayes argument will consider all the sample points in such a $\Lambda(v)$ where the order of the stages is determined by the usual lexicographical order of the v's. Now within each stage there will usually be many different restricted problems and we must still determine the proper order for them. For a given vector v let $v_2, \ldots, v_k$ be vectors of nondecreasing positive integers where the length of v_t is $v[t]$, the tth coordinate of v. Let $\Lambda(v, \{v_t\})$ be the subset of $\Lambda(v)$ for which $m_t(\mathbf{x}) = v_t$ for $t = 2, \ldots, k$. Let $\{v_t\}$ and $\{v_t'\}$ be two collection of vectors which are consistent with v. Then $\Lambda(v, \{v_t\})$ will precede $\Lambda(v, \{v_t'\})$ in the argument if v_2 precedes v_2' in the lexicographical ordering, or if $v_2 = v_2'$ and v_3 precedes v_3' in the lexicographical ordering or if the first two vectors are identical and v_4 precedes v_4' in the lexicographical ordering and so on. In summary the lexicographical ordering of the vector v gives the main stages of the argument while the ordering defined by $\{v_t\}$ gives the ordering within a stage. Hence for the inductive step we will assume that we are considering some $\Lambda(v, \{v_t\})$.

Recall that in the argument for admissibility of the mle for the general multinomial problem at each stage there were several restricted problems which could be considered in any particular order, i. e. all the faces of the same dimensional could be considered in any particular order. The same thing is happening here. To see this assume for a moment that $v[2] = 1$ and $v_2[1] < \prod_{\gamma \in b_2} |I_\gamma|$, i. e. the number of marginal cells we are considering for i_{b_2} is strictly less then maximum number of such cells. Let $\Lambda_1 \neq \Lambda_2$ be two sets each of which contain exactly $v_2[1]$ of the marginal cells i_{b_2}. Now the set of probability distributions which assign marginal probability one to the cells in Λ_1 is a face which is disjoint from the face of probability distributions which assign marginal probability one to the cells in Λ_2. This is just the earlier multinomial argument applied to the distribution $P(\cdot|\mathbf{i}_{c_2}^*)$, where $\mathbf{i}_{c_2}^*$ is the one cell that appears in the marginal table of c_2, but keeping in mind that it is multiplied by the other conditional distributions in the definition of $P(\mathbf{i})$ in equation (7). Note that these remarks remain true if more than one cell appears in the marginal table of c_2 for the sample points we are considering. Furthermore the conditional distributions for the cells that appear in c_t for $t = 3, \ldots, k$ are handled in exactly the same manner.

We now introduce some more notation that will allow us to identify the faces that will define a particular subset of the sample space which we will want to consider at a particular step in the argument. For each $a \subseteq C$ and every positive integer $j \leq |I_a| = \prod_{\gamma \in a} |I_\gamma|$ let $l(a, j)$ denote a set of j elements of $I_a = \prod_{\gamma \in a} I_\gamma$, the marginal cells for the factors a. Let v and $\{v_t\}$ be as above and fixed such that $\Lambda(v, \{v_t\})$ is nonempty. We now use l to define a particular subset of this set.

$$\begin{aligned}\Lambda(v, \{v_t\}, l) = \{ & \mathbf{x} \in \Lambda(v, \{v_t\}) : I_{a_1}(x) = l(a_1, v[1]), \\ & I_{c_t}(\mathbf{x}) = l(c_t, v[t]) \text{ for } t = 2, \ldots, k \text{ and} \\ & I_{b_t | i_{c_{t_j}}}(\mathbf{x}) = l(b_t, v_t[j]) \; t = 2, \ldots, k \text{ and } j = 1, \ldots, v[t] \, \}\end{aligned}$$

Note that l does not need to be defined for every $a \subseteq C$ just the subsets that appear in the definition. If l and l^* are two such functions which differ for some a and j in the definition then $\Lambda(v, \{v_t\}, l)$ and $\Lambda(v, \{v_t\}, l^*)$ will be disjoint subsets of $\Lambda(v, \{v_t\})$. Moreover $\Lambda(v, \{v_t\})$ can be partitioned by considering sufficiently many such l. This partition will consistent of all the faces which make up the set $\Lambda(v, \{v_t\})$ and they can be considered in any order at this stage of the argument. We are now ready to complete the inductive argument where we are considering the face $\Lambda(v, \{v_t\}, l)$ for some fixed v, $\{v_t\}$ and l.

$I_{a_1}(\Lambda(v, \{v_t\}, l)$ must contain at least one cell in the marginal table defined by a_1. Let $\mathbf{x}$ be a fixed sample vector of cells which belongs to $\Lambda(v, \{v_t\}, l)$. For $i^*_{a_1} \in I_{a_1}(\Lambda(v, \{v_t\}, l)$ we have that $n_{\mathbf{x}}(i^*_{a_1})$, the number of times the cell $i^*_{a_1}$ appears in the cells making up $\mathbf{x}$, must be greater than or equal to 1. If not this $\mathbf{x}$ would have been taken care of at an earlier stage of the stepwise Bayes argument. If $I_{a_1}(\Lambda(v, \{v_t\}, l)$ contains only one cell we take as our prior over $P(\mathbf{i}_{a_1})$ the distribution which assigns mass one to the parameter point which assigns mass one to this cell. If it contains more than one cell then this is just some multinomial model and we proceed just as we did in the last section for the multinomial problem. That is our prior is just proportional to the ratio of the renormalizing constant for the restricted problem to the product of the probabilities for each of the cells.

For c_2 it follows that

$$I_{c_2}(\Lambda(v,\{v_t\},l)) = \{\, Z_{c_2}(\mathbf{i}_{a_1}) \,:\, \mathbf{i}_{a_1} \in I_{a_1}(\Lambda(v,\{v_t\},l))\,\}$$

since $a_2 \cap a_1 = c_2 \subseteq a_1$. Now $I_{c_2}(\Lambda(v,\{v_t\},l))$ is nonempty so we let $\mathbf{i}^*_{c_2}$ be a fixed member of this set. For each $\mathbf{x} \in \Lambda(v,\{v_t\},l)$ we must have that $n_{\mathbf{x}}(i^*_{c_2}) \geq 1$ because if not it would have been taken care of in an earlier stage of the argument. Let

$$\begin{aligned} I_{b_2|\mathbf{i}^*_{c_2}}(\Lambda(v,\{v_t\},l)) =\{\, & Z_{b_2}(\mathbf{i}_{a_2}) \,:\, \mathbf{i}_{a_1\cup a_2} \in I_{a_1\cup a_2}(\Lambda(v,\{v_t\},l)) \\ & \text{and } Z_{c_2}(\mathbf{i}_{a_1}) = \mathbf{i}^*_{c_2}\,\} \end{aligned}$$

For a fixed $\mathbf{i}^*_{c_2}$, $I_{b_2|\mathbf{i}^*_{c_2}}(\Lambda(v,\{v_t\},l))$ is the set of levels which appear for the set of factors belonging to b_2 for some $\mathbf{x} \in \Lambda(v,\{v_t\},l))$ whose levels for the factors c_2 are just the levels $i^*_{c_2}$. Note that for each $\mathbf{x} \in \Lambda$ and $i^*_{b_2} \in I_{b_2|\mathbf{i}^*_{c_2}}(\Lambda(v,\{v_t\},l))$ we have $n_{\mathbf{x}}(i^*_{b_2\cup c_2}) \geq 1$ because if not it would have been taken care of in an earlier stage of the argument. For each $i^*_{c_2}$ the set $I_{b_2|\mathbf{i}^*_{c_2}}(\Lambda(v,\{v_t\},l))$ must contain at least one member. If it contains just one member our prior puts mass one on the parameter point which assigns probability one to this one cell. If it contains more than one cell then this is just some multinomial problem and we proceed exactly as before. Now we must repeat this process for each $i^*_{c_2} \in I_{c_2}(\Lambda(v,\{v_t\},l))$ and get a collection of priors one for each member of $I_{c_2}(\Lambda(v,\{v_t\},l))$.

If $k = 2$ then we would be done with this stage of the stepwise Bayes argument, since we assume that all these priors are independent and independent of the prior defined over $P(\mathbf{i}_{a_1})$. Then by multiplying all these priors together we have defined the appropriate prior over $P(\mathbf{i}) = P(\mathbf{i}_{a_1})P(\mathbf{i}_{b_2}|\mathbf{i}_{c_2})$ for this stage of the argument. It is now an easy calculation to check that under this prior the Bayes estimate of $P(\mathbf{i})$ is just the maximum likelihood estimate given in (8). In fact this is just the same posterior mean calculation done for the multinomial model in the previous section.

If $k > 2$ then for each $t = 3, \ldots, k$ we consider the sets $I_{c_t}(\Lambda(v, \{v_t\}, l)$ and $I_{b_t|\mathbf{i}^*_{c_t}}(\Lambda(v, \{v_t\}, l)$ and proceed as in the above. We then multiply all the priors together and the same conclusion follows.

This concludes the proof for this stage and by induction completes the proof of the theorem.

In the next section we will prove a more general result for exponential families which contains the admissibility results of both the multinomial problem of section two and the result of this section. We have included this proof since we find it to be instructive. The next proof just uses certain facts about exponential families which in some sense makes it simplier and more transparent at the "cost" of being more abstract.

4. EXPONENTIAL FAMILIES WITH FINITE SAMPLE SPACE

Let $\mathcal{X}$ be a finite subset of d-dimensional Euclidean space, $\mathbb{R}^d$. Let ν be a positive measure on $\mathcal{X}$. Define

$$\lambda(\theta) = \int_{\mathcal{X}} \exp(\theta \cdot x)\, d\nu(x) \tag{9}$$

and note that λ is finite for all $\theta \in \mathbb{R}^d$ because the integral is a finite sum. Let

$$M(\theta) = \log \lambda(\theta)$$

and define

$$P(x; \theta) = \exp(\theta \cdot x - M(\theta))$$

for $\theta \in \mathbb{R}^d$. This family of probability functions is called a d-dimensional standard exponential family and will be denoted by $\mathcal{P}$. $\mathcal{P}$ can be thought of as a subset of the usual m-dimensional multinomial simplex because each $P(\cdot;\theta)$ is associated with the unique vector $(\,P(x^1;\theta),\ldots,P(x^m;\theta)\,)$ where $x^1,\ldots,x^m$ are the distinct points making up $\mathcal{X}$.

Let X be a random variable taking values in $\mathcal{X}$ with $\mathcal{P}$, the family of possible probability functions for X. Then

$$E_\theta(X) = \nabla M(\theta) = \tau(\theta)$$

where ∇ denotes the gradient. We consider the problem of using X to estimate $\tau(\theta)$ with the sum of the squared errors as the loss function. Because of the convexity of the loss function we can reduce to a sufficient statistic (Lehmann (1983)) and assume that we have just one observation, X, rather than a random sample. Our goal is to prove the admissibility of the maximum likelihood estimator for this problem. We need to recall various facts about exponential families. Barndorff-Nielsen (1978) and Brown (1986) are good references. In particular the facts about maximum likelihood estimation can be found in section 3 of chapter 9 of Barndorff-Nielsen (1978).

Let $\mathcal{H}$ be the convex hull of $\mathcal{X}$. We assume that the dimension of $\mathcal{H}$ is d so that the family $\mathcal{P}$ is minimal. By (9) the family $\mathcal{P}$ is regular and hence steep. It follows that the range of $\tau(\cdot)$ is the interior of $\mathcal{H}$. Moreover if $X = x$ is observed and x is in the interior of $\mathcal{H}$ then x is the maximum likelihood estimate of $\tau(\theta)$. When x is in the boundary of $\mathcal{H}$ the situation is a bit more complicated since the maximum likelihood estimate is not well defined. However $\mathcal{P}$ can be enlarged in a natural way so that the maximum likelihood estimator becomes well defined. We now briefly outline how this can be done. It will be this extended estimator whose admissibility will be demonstrated.

Recall that θ is called the natural parameter for the exponential family $\mathcal{P}$. Since $\tau(\cdot)$ defines a homeomorphism between $\mathbb{R}^d$ and the interior of $\mathcal{H}$, $\mathcal{P}$ can also be parameterized by $\tau(\theta)$, the mean value parameterization. In what follows it will be convenient to use both of these parameterizations.

Let $\mathcal{F}$ be the family of non-empty faces of $\mathcal{H}$. Recall that $\mathcal{H}$ is a face of itself, hence an element of $\mathcal{F}$. Any $F \in \mathcal{F}$ is the convex hull of the set Λ^F of points of $\mathcal{X}$ belonging to F. For any $F \in \mathcal{F}$ and $P \in \mathcal{P}$ let $P(\cdot|F)$ denote the conditional distribution obtained by conditioning on the event F, which assigns probability zero to the complement of F and is proportional to $P(\cdot)$ on F. For $F \in \mathcal{F}$ let

$$\mathcal{P}^F = \{\, P(\cdot|F) : P \in \mathcal{P} \,\},$$

and define

$$\overline{\mathcal{P}} = \bigcup_{F \in \mathcal{F}} \mathcal{P}^F. \tag{10}$$

Each $\mathcal{P}^F$ is a regular exponential family and can be parametrized by the mean values. Let

$$\Phi(P) = E_P(X), \qquad P \in \overline{\mathcal{P}}.$$

Then Φ is a homeomorphism from $\overline{\mathcal{P}}$ to $\mathcal{H}$ where $\overline{\mathcal{P}}$ has its usual topology (since the state space is finite the topology of convergence in distribution is the same as the topology of pointwise convergence of the multinomial probabilities). As the notation suggests, $\overline{\mathcal{P}}$ is the completion of $\mathcal{P} = \mathcal{P}^{\mathcal{H}}$. The point of (10) is that the completion is a union of standard exponential families, one for each face of the convex support.

For an observation $x \in \mathcal{H}$, the maximum likelihood estimate in $\overline{\mathcal{P}}$ of the mean value parameter is just x. To see this, note that by Theorem 18.2, on page 164, of Rockafellar (1970) the collection of all relative interiors of the non-empty faces of $\mathcal{H}$ is a partition of $\mathcal{H}$. So there is exactly one face F containing x in its relative interior. The mean value mapping for $\mathcal{P}^F$ maps onto the relative interior of F so there is a unique $P \in \mathcal{P}^F$ that satisfies $E_P(X) = x$.

As we noted before this must maximize the likelihood over this face F and we denote this estimate by $\hat{P}(x|F)$. Let G be any other face which strictly contains F. Then for any $P \in \mathcal{P}$

$$P(x|G) < P(x|F) \leq \hat{P}(x|F).$$

On the other hand if G is a face which which does not contain F then $P(x|G) = 0$ and so $\hat{P}(x|F)$ maximizes the likelihood function over $\overline{\mathcal{P}}$.

These facts makes it easy to identify the correct order for the stepwise Bayes argument. Before starting the argument we need one more bit of notation. If F is a face of $\mathcal{H}$ let Λ_0^F be those members of Λ^F which also belong to the relative interior of F. Note that Λ_0^F may be empty.

In this argument, for the first stage we consider all the vertices of $\mathcal{H}$, which are its 0-dimensional faces. As before in each stage the particular order we use within a stage is not important. That is the faces to be considered at each stage can be handled in any order. At the next stage we consider all 1-dimensional faces F for which Λ_0^F is not empty. At the next stage we consider all 2-dimensional faces F for which Λ_0^F is not empty. We will continue in this way, where at each step we consider faces with one more dimension than those considered in the previous step, until all the points of $\mathcal{X}$ have been taken care of.

The first stage is easy. Let x be a member of $\mathcal{X}$ which is a vertex of $\mathcal{H}$. Then in $\overline{\mathcal{P}}$ there is a distribution, say P_x, which assigns mass one to x. We take as our restricted problem the one whose sample space is x and parameter space is P_x and the prior which puts mass one on P_x. For this restricted problem the Bayes estimate of the mean is just x which is the maximum likelihood estimate as well.

The proof will be completed by induction. Let F be a face of dimension d^*, where $1 \leq d^* \leq d$, for which Λ_0^F is not empty. For this step we consider the restricted problem with sample space Λ_0^F and parameter space $\mathcal{P}^F$ with the probability function renormalized to be a probability function over Λ_0^F. Let x be a member of Λ_0^F, if $X = x$ is observed, then by an earlier argument we know that x is the maximum likelihood estimator of the mean. It remains to show that x is also the Bayes estimate for this problem for a suitably chosen prior.

Now $\mathcal{P}^F$ can be parameterized not only by its mean value parameter, but by its natural parameter. By linear change of natural parameter and natural statistic to obtain a minimal representation we may assume the natural parameter is $\vartheta \in \mathbb{R}^{d^*}$ and that F is a set of full dimension in $\mathbb{R}^{d^*}$. We will use the natural parameter to define the prior for this stage.

Let

$$\lambda(\vartheta|\Lambda_0^F) = \int_{\Lambda_0^F} \exp(\vartheta \cdot x)\, d\nu(x)$$

and

$$\lambda(\vartheta|\Lambda^F) = \int_{\Lambda^F} \exp(\vartheta \cdot x)\, d\nu(x)$$

Then for $\vartheta \in \mathbb{R}^{d^*}$ the prior is given by

$$\pi(\vartheta) \propto \frac{\lambda(\vartheta|\Lambda_0^F)}{\lambda(\vartheta|\Lambda^F)} \tag{11}$$

There remain two things to be checked. The first is to verify that $\pi(\vartheta)$ is indeed a proper prior over $\mathbb{R}^{d^*}$. The second is that it gives x as the Bayes estimate of the mean. We begin by proving the first.

Since Λ_0^F contains only a finite number of points it is enough to show that

$$\exp(\vartheta \cdot x)/\lambda(\vartheta|\Lambda^F) \tag{12}$$

is integrable over $\mathbb{R}^{d^*}$, whenever $x \in \Lambda_0^F$. We also show that (12) is bounded, a fact used later.

In proving this, we may assume without loss of generality that x is the origin. Hence we must show that

$$1/\lambda(\vartheta|\Lambda^F) \tag{13}$$

is integrable when the interior of F contains the origin.

Let $\epsilon > 0$ be chosen so that the closed ϵ ball centered at the origin lies in the interior of F. Since this ball is equal to the the intersection of all the half spaces containing it, for a fixed nonzero ϑ there is an $x \in F$ such that $\vartheta \cdot x > \epsilon|\vartheta|$ where $|\vartheta|$ is the norm of ϑ. If not F must lie in the closed ϵ ball centered at the origin which is a contradiction. From this it follows that

$$\sup_{y \in F} \vartheta \cdot y \geq \epsilon|\vartheta|$$

where ϵ does not depend on ϑ. Then

$$\begin{aligned} 1/\lambda(\vartheta|\Lambda^F) &= \{\int_{\Lambda^F} \exp(\vartheta \cdot x)\, d\nu(x)\}^{-1} \\ &\leq \{\exp(\epsilon|\vartheta|) \min_{x \in \mathcal{X}} \nu(x)\}^{-1} \\ &\leq K \exp(-\epsilon|\vartheta|) \end{aligned}$$

for some constant K. From this it follows that (13) is an integrable function over $\mathbb{R}^{d^*}$ and hence the function in (11) is indeed a proper prior.

We will now verify for the restricted problem that the Bayes estimator of the mean of X is just x when $X = x$ is observed. To this end we need a result that follows from Theorem 2 of Diaconis and Ylvisaker (1979). For $i = 1, \ldots, d$ let x_i be the ith coordinate of x and

$$\tau_i(\vartheta) = \frac{\partial M(\vartheta)}{\partial \vartheta_i} = \frac{\partial \lambda(\vartheta|\Lambda^F)/\partial \vartheta_i}{\lambda(\vartheta|\Lambda^F)}$$

be the expected value of the ith coordinate. Then they show, their equation (2.10), for any $n_0 > 0$ and any x_0 in the interior of F (in particular any

$x_0 \in \Lambda_0^F$) that

$$\frac{\displaystyle\int \tau_i(\vartheta) \frac{\exp([n_0 x_0 + x] \cdot \vartheta)}{\lambda(\vartheta|\Lambda^F)^{n_0+1}} \, d\vartheta}{\displaystyle\int \frac{\exp([n_0 x_0 + x] \cdot \vartheta)}{\lambda(\vartheta|\Lambda^F)^{n_0+1}} \, d\vartheta} = \frac{n_0 x_{0,i} + x_i}{n_0 + 1} \tag{14}$$

Note that we would be done if equation (14) were valid with $n_0 = 0$, which it is since both sides of the equation are continuous functions of n_0 as it decreases to zero. This is obvious for the right hand side.

For the left hand side, this follows from the Lebesgue dominated convergence theorem. To see this rewrite the fractions in the numerator and denominator of (14) as

$$\left(\frac{\exp(x_0 \cdot \vartheta)}{\lambda(\vartheta|\Lambda^F)} \right)^{n_0} \frac{\exp(x \cdot \vartheta)}{\lambda(\vartheta|\Lambda^F)}$$

The first factor is bounded, the second integrable, as we showed in the argument about (12). Also $\tau_i(\vartheta)$ is bounded because it is the expectation of X_i, which is bounded.

This concludes the proof of admissibility of the maximum likelihood estimator. The result is stated formally in the following theorem.

Theorem 2 *Consider the problem of estimating the mean vector in the completion of a full exponential family with finite sample space. If the loss function is the sum of the component squared error losses then the maximum likelihood estimator is admissible.*

ACKNOWLEDGEMENTS

Glen Meeden's research was supported in part by NSF Grant SES 9201718 and Charles Geyer's research was supported in part by NSF Grant DMS 9007833.

BIBLIOGRAPHY

Alam, K. (1979). "Estimation of multinomial probabilities." *Annals of Statistics* 7, 282-283.

Andersen, A. H. (1974). "Multidimensional contingency tables." *Scandinavian Journal of Statistics*, 1, 115-127.

Barndorff-Nielsen, O. (1978) *Information and Exponential Families in Statistical Theory*, Wiley: New York.

Brown, L. D. (1981). "A complete class theorem for statistical problems with finite sample space." *Annals of Statistics* 9, 1289-1300.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families*, IMS monographs: Hayward, CA.

Darroch, J. N., Lauritzen, S. L. and Speed, T. P. (1980). "Markov fields and log-linear interaction models for contingency tables." *Annals of Statistics* 8, 522-539.

Diaconis, P. and Ylvisaker, D. (1979). "Conjugate priors for exponential families." *Annals of Statistics* 7, 269-281.

Goodman, L. A. (1970). "The multivariate analysis of qualitative data: Interaction among multiple classifications." *Journal of the American Statistical Association* 65, 226-256.

Goodman, L. A. (1971). "Partition of chi-squared, analysis of marginal contingency tables, and estimation of expected frequencies in multidimensional contingency tables." *Journal of the American Statistical Association* 66, 339-344.

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

Hsuan, F. C. (1979). "A stepwise Bayes procedure." *Annals of Statistics* 7, 860-868.

Johnson, B. M. (1971). "On admissible estimators for certain fixed sample binomial problems." *Annals of Mathematical Statistics*, 42, 1579-1587.

Lehmann, E. L. (1983). *Theory of Point Estimation*, Wiley: New York.

Rockafellar, R. T. (1970). *Convex Analysis*, Princeton University Press: Princeton, New Jersey.

Wald, A. and Wolfowitz, J. (1951). "Characterization of the minimal complete class of decision functions when the number of distributions and decisions is finite." *Proc. Second Berk. Symp. Math. Statist. and Prob.*, 149-157.