GENERALIZING THE ITERATIVE PROPORTIONAL FITTING PROCEDURE*

by

Michael M. Meyer

Department of Applied Statistics
School of Statistics
University of Minnesota

Technical Report No. 371

April 1980

*This research was supported by ONR contract N00014-78-C-0151 to the University of Minnesota. I wish to thank Stephen Fienberg for many helpful conversations and comments.

Summary

The IPFP can be viewed as a method for maximizing the likelihood for certain loglinear models or equivalently for minimizing the Kullback-Leibler Information between two probability densities. Both of these viewpoints lead to natural generalizations of the classical IPFP. We examine the generalizations suggested by the work of Csiszár (1975), Darroch and Ratcliff (1972), and Haberman (1974) and, with the aid of the theory, explore a practical example of expanding a contingency table.

Key words and phrases: Generalized Iterative Scaling; I-divergence; Kullback-Leibler information number; Contingency tables.

Introduction

There are many ways of calculating maximum likelihood estimates of mean values for loglinear models. The two most popular methods are the Iterative Proportional Fitting Procedure (IPFP) and variants upon Newton's method. Newton's method has many desirable properties, including its quadratic convergence rate near the maximum and the ability to calculate estimates of asymptotic covariance matrices as a by-product of the computations. Its principle disadvantage is a computational one in that the method requires a considerable amount of storage and is thus limited by the size of the design manifold being fitted. In many situations it thus becomes necessary to consider alternatives to Newton's method.

The Iterative Proportional Fitting Procedure is an alternative method for fitting many classes of loglinear models. Although the method is often slow to converge it requires little storage. In our experience it is often the storage requirements of an algorithm, as opposed to the computational time required, that limit the algorithm's usefulness. The classical IPFP (see, e.g., Bishop, Fienberg, and Holland (1975)) is limited in the type of models which may be fitted. As many applications of the loglinear model methodology now use models other than simple factorial situations, we seek generalizations of the IPFP which extend its capabilities to any loglinear model while preserving the desirable properties of the classical IPFP.

There are at least three generalizations of the IPFP. Haberman (1974) shows that the IPFP is really just a special case of the method of cyclic ascent for functional maximization. This observation immediately leads to an algorithm defined for any loglinear model. Csiszár (1975) considers the IPFP as a method for minimizing the Kullback-Leibler information (or I-divergence)

between two probability densities. When specialized to distributions on finite sets, Csiszár's methods yield another type of IPFP. In Section 2, we show that these methods and those of Haberman are closely related and yield equivalent procedures in some situations. A third generalization of the classical IPFP, discussed in Section 3, is the Generalized Iterative Scaling (GIS) method of Darroch and Ratcliff (1972). This generalization is also developed in the setting of minimizing I-divergence but does not appear to be related to the other methods.

The impetus for this report came from an example of expanding a contingency table into a more manageable structure which appeared in Fienberg and Wasserman (1980). Section 4 is devoted to a discussion of this and similar examples.

2. The Results of Haberman and Csiszár on the IPFP

The Iterative Proportional Fitting Procedure is generally considered as a method for obtaining the maximum likelihood estimates for the mean value parameter of a loglinear model for a contingency table. The formulation of the IPFP considered by Csiszár is presented as a problem of minimizing the Kullback-Leibler information number between two Probability Distributions (P.D.'s). Although we shall only use P.D.'s defined on a finite set, it is instructive to outline Csiszár's very general formulation and specialize the results as the need arises.

Haberman (1974, pp. 64-73) noted that the classical IPFP is a version of the cyclic ascent method of functional maximization and suggested the extension to arbitrary loglinear models. The methods of Haberman and Csiszár can be viewed as dual methods although strictly speaking the algorithms dual to Csiszár's encompass a much wider class of maximization techniques. We shall concentrate on stating results, illustrating the ideas with examples. However we should note that Csiszár presents very elegant proofs by developing a "geometry" for the information measure. The geometric ideas have (and were perhaps developed from) a strong analogy with results in finite dimensional Hilbert spaces. We now turn to a detailed discussion of the techniques.

Let N, P, Q, R, S, T denote Probability Distributions on a measure space $(X, \mathcal{X})$. In our applications X will be a finite set and $\mathcal{X}$ the power set of X. If P is absolutely continuous with respect to Q (written as P a.c. Q) we will denote the corresponding density by p_Q. The Kullback-Leibler information number (or I-divergence, or information in P about Q), $I(P||Q)$, is defined to be

$$(2.1)\qquad I(P||Q) = \begin{cases} \int \ln\,(p_Q)dP = \int p_Q \ln\,(p_Q)dQ & \text{if } P \text{ a.c. } Q \\ +\infty & \text{if } P \text{ not a.c. } Q. \end{cases}$$

When P and Q are both a.c. N then (2.1) may be written as

$$(2.2)\qquad I(P||Q) = \int p_N \ln\,(p_N/q_N)dN.$$

In the above formulae we use the conventions that, $\ln(0) = -\infty$, $0 \cdot (\pm\infty) = 0$, and $\ln(r/0) = +\infty$ when $r \in (0,\infty)$.

In the special case where N is the P.D. which assigns equal weight to each point of a finite set X, then all P.D.'s on X are absolutely continuous with respect to N. Unless otherwise indicated (e.g., by the use of some subscript) all densities on finite sets will be with respect to this uniform N and the probability function $p_N(x)$ will be written as $p(x)$. In this situation equation (2.2) becomes:

$$(2.3)\qquad I(P||Q) = \frac{1}{|X|} \cdot \sum_{x\in X} p(x) \ln\,(p(x)/q(x)).$$

We next consider the I-sphere, $\mathcal{S}$, with center R and radius ρ, defined by:

$$(2.4)\qquad \mathcal{S}(R,\rho) = \{P : I(P||R) < \rho\},\quad \rho \in (0,\infty].$$

The I-sphere, $\mathcal{S}$, contains P.D.'s which are close, in the information sense, to a given P.D. If $\mathcal{E}$ is a convex set of P.D.'s such that $\mathcal{E} \cap (R,\infty) \neq \emptyset$, then a P.D., $Q \in \mathcal{E}$, satisfying

$$(2.5)\qquad I(Q||R) = \min_{P\in\mathcal{E}} I(P||R),$$

is called the I-projection of R on $\mathcal{E}$ and will be denoted by $Q = \mathbb{P}_{\mathcal{E}}(R)$. The convex set $\mathcal{E}$ of P.D.'s is called a linear set if when P and Q are in $\mathcal{E}$

and $T = \alpha P + (1-\alpha)Q$, $(\alpha \in \mathbb{R})$ is a P.D. then T is also in $\mathcal{E}$. Csiszár gives conditions under which $P_{\mathcal{E}}(R)$ exists (it is always unique) and develops a geometry of I-divergence by using an analogue of Pythagoras' Theorem.

As our goal is to study maximum likelihood estimation in contingency tables, we turn briefly to the problem of estimating a multinomial probability function with an underlying loglinear model for the probabilities. Consider a multinomial random vector, z(x), of Z counts on the set X, with mean $(m(x) : x \in X)$ where $m(x) = Z \cdot p(x)$ and p(x) is the probability an observation falls in cell $x \in X$. The vector $s(x) = z(x)/Z$ is an observed probability function on X. The log-likelihood of the data, z, given the assumed mean, m, will be denoted by $\ell(m;z)$. A loglinear model for the probability function (equivalently for the mean vector) asserts that the logarithm of the underlying probability function is an element of some linear manifold, $\mathcal{M}$, i.e.,

$$\ln (p(x)) \in \mathcal{M} .$$

It is well known (see e.g., Haberman (1974)) that the maximum likelihood estimator, $\hat{p}$, of p based on the observed probability function, s, satisfies

$$\begin{aligned} &\text{(i)} \quad \ln (\hat{p}) \in \mathcal{M} , \\ &\text{(ii)} \quad \hat{p} - s \in \mathcal{M}^{\perp} , \end{aligned} \tag{2.6}$$

and minus the log-likelihood ratio of $\hat{p}$ compared with s is proportional to

$$\sum_{x \in X} s(x) \ln (s(x)/\hat{p}(x)) = I(S||\hat{P}). \tag{2.7}$$

The problem of minimizing I-divergence (i.e., equation (2.5)) and maximizing the log-likelihood ratio (i.e., equation (2.7)) appear similar. All the same, the relationship between the two methods is not clear. In order to show that the two problems lead to identical estimates we need to

envoke a result of Csiszár (due originally to Kullback (1959)), giving the form of the density of the I-projection. Csiszár's Theorem 3.1, which we state below, gives conditions under which the probability function of the I-projection satisfies a loglinear model. The examples at the end of this section and the discussion after the theorem should help to clarify the notation.

Theorem 1 (Csiszár (1975), p. 152).

Let $\mathcal{F} = \{f_\gamma : \gamma \in \Gamma\}$ be a set of real valued $\mathcal{X}$ measurable functions on X and $\mathcal{A} = \{a_\gamma : \gamma \in \Gamma\}$ be real constants. Let $\mathcal{E}$ be the (linear) set of all probability distributions, P, on $(X,\mathcal{X})$ for which the integrals, $\int f_\gamma dP$ exist and

$$\int f_\gamma dP = a_\gamma; \qquad \gamma \in \Gamma .$$

Then if a P.D. R has I-projection Q on $\mathcal{E}$, the density of Q with respect to R is of the form

$$(2.8) \qquad q_R(x) = c.\ \exp\,(g(x)) \qquad x \in M$$
$$= 0 \qquad x \notin M$$

where $P(M) = 0$ for every P in $\mathcal{E} \cap \mathcal{M}(R,\infty)$ and g belongs to the closed subspace of $L_1(Q)$ spanned by the f_γ's.

Conversely if a $Q \in \mathcal{E}$ has a density with respect to R of the form (2.8) where g belongs to the linear space spanned by the f_γ's, then Q is the I-projection of R on $\mathcal{E}$. ∎

In our applications of this theorem X will be a finite set. Let S be an observed P.D. on X and consider a set of functions $\mathcal{F} = \{f_\gamma : \gamma \in \Gamma\}$ which span a linear manifold $\mathcal{M}_{\mathcal{F}}$. The set of constants $\mathcal{A} = \{a_\gamma : \gamma \in \Gamma\}$ will be defined by

$$a_\gamma = \int f_\gamma dS = \int f_\gamma s_N dN, \quad \gamma \in \Gamma .$$

That is the a_γ are the "marginal sums" of the observed probability function. We will call the set $\mathcal{A}$, determined by an observed P.D. S and the functions $\mathcal{F}$, the $\mathcal{F}$-margins of S. The set $\mathcal{E}$ of the theorem is the set of all P.D.'s, P, such that $\int f_\gamma dP = a_\gamma = \int f_\gamma dS$ for all γ in Γ. In other words the set $\mathcal{E}$ consists of those P.D.'s which have the same $\mathcal{F}$-margins as the observed P.D. S. This is in turn the same as the set of probability functions, p, such that s-p is in $\mathcal{M}_{\mathcal{F}}^{\perp}$. The conclusion of the theorem (equation (2.8)) says that if $\hat{Q} = P_{\mathcal{E}}(R)$ then $\hat{q}_R$, the density with respect to R, satisfies

$$\ln (\hat{q}_R(x)) \in \mathcal{M}_{\mathcal{F}} .$$

If $\hat{Q}$ a.c. N then this is the same as saying that

$$\ln (\hat{q}_N(x)) \in \mathcal{M}_{\mathcal{F}} + \ln (r_N). \tag{2.9}$$

Equation (2.9) says that the log probabilities of the I-projection lie in an affine subspace of $\mathbb{R}^{|X|}$. Note that if $\ln (r_N)$ is in $\mathcal{M}_{\mathcal{F}}$ then $\mathcal{M}_{\mathcal{F}} + \ln (r_N) = \mathcal{M}_{\mathcal{F}}$ and the log probabilities lie in a linear manifold. When R has a density r_N w.r.t. N and $\ln (r_N)$ is in $\mathcal{M}_{\mathcal{F}}$, then the I-projection is seen to satisfy part (i) of equation (2.6) for the manifold $\mathcal{M}_{\mathcal{F}}$.

In the above development we have restricted our attention to those P.D.'s which had the same $\mathcal{F}$-margins as the observed P.D. S. In other words condition (ii) of equation (2.6), which required that $\hat{p} - s$ be in $\mathcal{M}_{\mathcal{F}}$, is satisfied for all $\hat{p}$

in & and in particular for the I-projection $\hat{q}_N$. The conclusion is that $\hat{q}_N$ satisfies all the conditions required of the M.L.E. and is thus the M.L.E.

An alternative demonstration of this fact comes from an argument due to Darroch and Ratcliff (1972), which shows directly that the likelihood is maximized by the I-projection.

A purely mathematical interpretation of this result is that the problem of minimizing I-divergence subject to linear constraints is the convex (or Fenchel) dual problem to maximizing the likelihood subject to loglinear constraints. For further references to this topic see Rockafeller (1974) or Luenberger (1969). We will use the theorem with many different spaces & to demonstrate the duality between the Iterative Proportional Fitting Procedures of Csiszár and Haberman.

To illustrate the preceeding ideas we present a simple example, where duality and results are well known.

Example 1.

Consider Z observations cross classified according to their response level on two factors, A and B, each with 3 levels. We assume that the data can be considered as Z independent multinomial trials. The observed table of data is

A \ B			
	z_{11}	z_{12}	z_{13}
	z_{21}	z_{22}	z_{23}
	z_{31}	z_{32}	z_{33}

where z_{ij} is the number of observations which have level i of factor A and level j of factor B. The sum of the cell counts, $\sum_{ij} z_{ij}$, equals Z, the total number of observations. We will convert this to an observed probability function by dividing each cell count by the total, Z. Thus the observed probability function, s, is represented by:

$$s = \begin{array}{|c|c|c|} \hline z_{11}/Z & z_{12}/Z & z_{13}/Z \\ \hline z_{21}/Z & z_{22}/Z & z_{23}/Z \\ \hline z_{31}/Z & z_{32}/Z & z_{33}/Z \\ \hline \end{array}$$

We now consider the model of independence of the two responses for the true P.D. P. This corresponds to a loglinear model

$$\ln (p) \in \mathcal{M}_3$$

where $\mathcal{M}_3$ is the manifold spanned by the row and column sum tables;

$$f_R^1 = \begin{array}{|c|c|c|} \hline 1 & 1 & 1 \\ \hline 0 & 0 & 0 \\ \hline 0 & 0 & 0 \\ \hline \end{array} \qquad f_R^2 = \begin{array}{|c|c|c|} \hline 0 & 0 & 0 \\ \hline 1 & 1 & 1 \\ \hline 0 & 0 & 0 \\ \hline \end{array}$$

$$f_R^3 = \begin{array}{|c|c|c|} \hline 0 & 0 & 0 \\ \hline 0 & 0 & 0 \\ \hline 1 & 1 & 1 \\ \hline \end{array} \qquad f_C^1 = \begin{array}{|c|c|c|} \hline 1 & 0 & 0 \\ \hline 1 & 0 & 0 \\ \hline 1 & 0 & 0 \\ \hline \end{array}$$

$$f_C^2 = \begin{array}{|c|c|c|}\hline 0 & 1 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & 1 & 0 \\ \hline \end{array} \qquad f_C^3 = \begin{array}{|c|c|c|}\hline 0 & 0 & 1 \\ \hline 0 & 0 & 1 \\ \hline 0 & 0 & 1 \\ \hline \end{array}$$

The maximum likelihood estimator, $\hat{p}$, of p maximizes $\ell(p;s)$ subject to $\ln(p)$ being in $\mathcal{M}_4$. This is the same as maximizing the log-likelihood ratio, i.e.,

$$\max_{\ln(p) \in \mathcal{M}_4} \sum_{ij} p_{ij} \ln (p_{ij}/s_{ij}) = I(S||\hat{P}).$$

We now turn to the formulation of the problem as a minimum I-divergence problem. Let

$$\mathcal{J} = \{f_R^1, f_R^2, f_R^3, f_C^1, f_C^2, f_C^3\}$$

and

$$\mathcal{A} = \{a_R^1, a_R^2, a_R^3, a_C^1, a_C^2, a_C^3\}$$

where

$$a_\ell^k = \sum_{ij} f_\ell^k(i,j)s_{ij} \quad \text{for } \ell = R,C \text{ and } k = 1,2,3,$$

and define

$$\mathcal{E} = \{\text{P.D.'s P s.t. } \sum_{ij} f_\ell^k(i,j)p_{ij} = a_\ell^k\}.$$

Csiszár's Theorem 3.1 tells us that the M.L.E., $\hat{p}$, for the loglinear model corresponds to $P_{\mathcal{E}}(R)$, the I-projection of R onto $\mathcal{E}$, for any R which has $\ln(r)$ in $\mathcal{M}_4$. A simple R which satisfies this is

$r =$

1/9	1/9	1/9
1/9	1/9	1/9
1/9	1/9	1/9

which assigns the same probability to each cell in the table. ∎

Thus far we have not given any techniques for calculating I-projections. In the following paragraphs we hope to rectify this situation by presenting some forms of the Iterative Proportional Fitting Procedure.

Recall that a set of P.D.'s, $\&$, is called linear if when P_1 and P_2 are in $\&$ and $T = \alpha P_1 + (1-\alpha)P_2$, $\alpha \in \mathbb{R}$ is a P.D., then T is also in $\&$. We note that if $\&$ is defined by a set of constraints, $\mathcal{J}$, and corresponding constants, $\mathcal{A}$, then $\&$ is a linear set. In particular the maximum likelihood estimation problem for loglinear models can be posed in terms of linear sets of P.D.'s. Csiszár presents results which enable one to build up the total I-projection onto $\&$ cyclically, by forming the I-projections onto other (and hopefully simpler) spaces, $\&_i$. The statement of Csiszár's Theorem 3.2 follows.

Theorem 2 (Csiszár (1975), p. 155).

Let $\&_1,\ldots,\&_k$ be arbitrary linear sets of P.D.'s on a finite set X with $\& = \bigcap_{i=1}^{k} \&_i \neq \emptyset$, let R be a P.D. for which there exists a $P \in \&$ with P a.c. R and define $Q_1, Q_2, \ldots$ recursively by letting Q_n be the I-projection of Q_{n-1} on $\&_n$, $n = 1,2,3,\ldots$ where $Q_0 = R$ and $\&_n = \&_i$ if $i = n \bmod k$. Then Q_n converges (pointwise) to the I-projection, $Q = P_{\&}(R)$. ∎

Some analogies between this theorem and similar theorems about projection operators on Hilbert spaces are presented in Appendix 1.

Before proceeding to the examples, we make two simple observations. First, if $\mathfrak{F}_1,\ldots,\mathfrak{F}_k$ are sets of constraint vectors with corresponding sets of constants $\mathcal{A}_1,\ldots,\mathcal{A}_k$ which together determine linear sets $\mathfrak{E}_1,\ldots,\mathfrak{E}_k$, then the sets $\bigcup_{i=1}^{k}\mathfrak{F}_i$ and $\bigcup_{i=1}^{k}\mathcal{A}_i$ together determine the linear set $\bigcap_{i=1}^{k}\mathfrak{E}_i$. In other words more constraints (the union of the $\mathfrak{F}_i$) leads to a more restricted or smaller linear set (the intersection of the $\mathfrak{E}_i$).

Our second remark concerns a special case of Csiszár's Theorem 3.2. Let $\mathfrak{F}_1$ and $\mathfrak{F}_2$ be sets of functions on X with corresponding sets of constants, $\mathcal{A}_1$ and $\mathcal{A}_2$. We will assume that the functions in $\mathfrak{F}_1$ have support X_1 while those in $\mathfrak{F}_2$ have support $X_2 = X \setminus X_1$. Let e_1 be the indicator function of X_1 and e_2 the indicator function of X_2. We assume that e_i is in span $(\mathfrak{F}_i)$. Consider $\mathfrak{E}_i$ to be the linear space of positive functions generated by $\mathfrak{F}_i$ and $\mathcal{A}_i$, $(i = 1,2)$, and $\mathfrak{E}$ the space generated by $\{\mathfrak{F}_1, \mathfrak{F}_2\}$ and $\{\mathcal{A}_1, \mathcal{A}_2\}$. For any P.D. Q we define Q_i to be equal to Q on X_i and zero on $X \setminus X_i$.

Corollary to Theorem 2.

Consider $\mathfrak{E}_1$, $\mathfrak{E}_2$ and Q_1, Q_2 as defined above, then:

$$\mathbb{P}_{\mathfrak{E}}(Q) = \mathbb{P}_{\mathfrak{E}_1}(Q_1) + \mathbb{P}_{\mathfrak{E}_2}(Q_2). \blacksquare$$

A rough interpretation of the corollary is that if the constraints can be separated into disjoint pieces, then the I-projection can be similarly separated.

We now turn to some illustrations of the preceeding theorems.

Example 1 (continued). A 3 × 3 table.

Recall that we have a table of observed probabilities, s, and a design manifold $\mathcal{M}_{\mathcal{J}}$ defined by a set of functions, $\mathcal{J} = \{f_R^1,\ldots,f_C^3\}$ and constants $\mathcal{A} = \{a_R^1,\ldots,a_C^3\}$. We consider an arbitrary starting P.D., R with

$$r = \begin{array}{|c|c|c|}\hline r_{11} & r_{12} & r_{13} \\ \hline r_{21} & r_{22} & r_{23} \\ \hline r_{31} & r_{32} & r_{33} \\ \hline \end{array} \quad , \quad \sum_{ij} r_{ij} = 1.$$

The usual IPFP alternately scales the rows and columns of r to have the same margins as the observed table. Specifically, the row adjustment is;

$$r_{ij}^1 = (r_{ij}/\Sigma_j\, r_{ij}) \times a_R^i;\ i,j = 1,2,3,$$

which is followed by the column adjustment,

$$r_{ij}^2 = (r_{ij}^1/\Sigma_i\, r_{ij}^1) \times a_C^j;\ i,j = 1,2,3.$$

This process is repeated until the cell estimates converge. Of course when $\ln(r) \in \mathcal{M}_{\mathcal{J}}$, the iterations converge after just one row and column adjustment. The fitted values then correspond to the M.L.E.'s for the loglinear model, $\ln(p) \in \mathcal{M}_{\mathcal{J}}$. When $\ln(r) \notin \mathcal{M}_{\mathcal{J}}$ the iteration need not converge after one cycle. In this case the fitted values correspond to the M.L.E.'s for the log affine model, $\ln(p) \in \ln(r) + \mathcal{M}_{\mathcal{J}}$.

Let us now investigate how one could use Csiszár's theorem to calculate $\mathbb{P}_{\mathcal{E}}(R)$. Consider two linear spaces of P.D.'s, $\mathcal{E}_1$ and $\mathcal{E}_2$, defined by:

$$\mathcal{E}_1 = \{\text{P.D.'s } P \text{ s.t. } \sum_{ij} f_R^{\ell}(ij)p(i,j) = a_R^{\ell},\ \ell = 1,2,3\}$$

$$\mathcal{E}_2 = \{\text{P.D.'s } P \text{ s.t. } \sum_{ij} f_C^{\ell}(ij)p(i,j) = a_C^{\ell},\ \ell = 1,2,3\}.$$

That is $\mathcal{E}_1$ is the set of P.D.'s whose row sums agree with the observed table and $\mathcal{E}_2$ consists of those P.D.'s whose column sums agree with the observed table. As $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$, the I-projection, $Q = \mathbb{P}_{\mathcal{E}}(R)$ is the limit of

$$Q_1 = \mathbb{P}_{\mathcal{E}_1}(R)$$

$$Q_2 = \mathbb{P}_{\mathcal{E}_2}(Q_1)$$

$$Q_3 = \mathbb{P}_{\mathcal{E}_1}(Q_2)$$

$$Q_4 = \mathbb{P}_{\mathcal{E}_2}(Q_3), \quad \text{etc.}$$

where $Q_n = \mathbb{P}_{\mathcal{E}_i}(Q_{n-1}), \quad i = n \bmod 2.$

Thus we need to be able to calculate the I-projections onto $\mathcal{E}_1$ and $\mathcal{E}_2$. From the definition of I-projections, Q_1 satisfies

$$I(Q_1||R) = \min_{Q \in \mathcal{E}_1} I(Q||R)$$

$$= \min_{Q \in \mathcal{E}_1} \sum_{ij} q_{ij} \ln (q_{ij}/r_{ij}).$$

Expanding this expression leads to the following minimization problem:

$$\min\ q_{11} \ln (q_{11}/r_{11}) + q_{12} \ln (q_{12}/r_{12}) + (a_R^1 - q_{11} - q_{12}) \ln (a_R^1 - q_{11} - q_{12}/r_{13})$$

$$+ q_{21} \ln (q_{21}/r_{21}) + q_{22} \ln (q_{22}/r_{22}) + (a_R^2 - q_{21} - q_{22}) \ln (a_R^2 - q_{21} - q_{22}/r_{23})$$

$$+ q_{31} \ln (q_{31}/r_{31}) + q_{32} \ln (q_{32}/r_{32}) + (a_R^3 - q_{31} - q_{32}) \ln (a_R^3 - q_{31} - q_{32}/r_{33}),$$

where q_{11}, q_{12}, q_{21}, q_{22}, q_{31}, q_{32} are allowed to vary freely over (0,1). If one takes partial derivatives of this expression with respect to $q_{11}, q_{12}, \ldots, q_{32}$ and equates each derivative to zero one obtains the equations:

$$\text{(i)} \quad \ln (q_{11}/r_{11}) = \ln (q_{12}/r_{12}) = \ln (a_R^1 - q_{11} - q_{12}/r_{13})$$

$$\text{(ii)} \quad \ln (q_{21}/r_{21}) = \ln (q_{22}/r_{22}) = \ln (a_R^2 - q_{21} - q_{22}/r_{23})$$

$$\text{(iii)} \quad \ln (q_{31}/r_{31}) = \ln (q_{32}/r_{32}) = \ln (a_R^3 - q_{31} - q_{32}/r_{33}).$$

By removing the logarithms one sees that these are just linear equations whose solution is found by scaling the rows so that the marginal sums are correct. Analogous equations result for the I-projection onto $\mathcal{E}_2$. Thus for this partition of the space, $\mathcal{E}$, the Csiszár algorithm reduces to the usual IPFP.

The particular subdivision of the space, $\mathcal{E}$, is not the only one possible. Consider dividing the space $\mathcal{E}$ into more linear spaces, namely:

$$\mathcal{E}_1 = \{\text{P.D.'s P s.t. } \int f_R^1 dP = \int f_R^1 dS = a_R^1\}$$

$$\mathcal{E}_2 = \{\text{P.D.'s P s.t. } \int f_R^2 dP = \int f_R^2 dS = a_R^2\}$$

$$\vdots$$

$$\mathcal{E}_6 = \{\text{P.D.'s P s.t. } \int f_C^3 dP = \int f_C^3 dS = a_C^3\}.$$

As $\mathcal{E} = \bigcap_{i=1}^{6} \mathcal{E}_i$, Csiszár's theorem tells us we can find the I-projection, Q_∞, onto $\mathcal{E}$ by cyclically projecting onto $\mathcal{E}_1, \ldots, \mathcal{E}_6$. Again we need to calculate each of the elementary I-projections. For example we need Q_1 such that $I(Q_1||R) = \min_{Q \in \mathcal{E}_1} I(Q||R)$. If we write the expression out more fully we require

the:

$$\min \; q_{11} \ln (q_{11}/r_{11}) + q_{12} \ln (q_{12}/r_{12}) + (a_R^1 - q_{11} - q_{12}) \ln (a_R^1 - q_{11} - q_{12}/r_{13})$$
$$+ q_{21} \ln (q_{21}/r_{21}) + q_{22} \ln (q_{22}/r_{22}) + q_{23} \ln (q_{23}/r_{23}) + q_{31} \ln (q_{31}/r_{31})$$
$$+ q_{32} \ln (q_{32}/r_{32}) + ([1 - a_R^1] - q_{21} - q_{22} - q_{23} - q_{31} - q_{32})$$
$$\ln ([1 - a_R^1] - q_{21} - q_{22} - q_{23} - q_{31} - q_{32}/r_{33})$$

where the minimization is over q_{11}, q_{12}, q_{21}, q_{22}, q_{23}, q_{31} and q_{32} all in (0,1). We can obtain the minimizing Q_1 as we did before. The procedure is,

(i) scale the first row of r so that it has the correct margin

(ii) scale the rest of the table so that the sum of all the cells is again one.

The full algorithm then cycles through rows and columns one at a time.

This algorithm is not as efficient as the earlier procedure, as we need to adjust the entire table at each iteration and only one of the (row) margins is necessarily satisfied at the end of each iteration. In this approach we have ignored the corollary to the theorem whereas in the earlier approach the corollary was implicitly invoked. ■

Example 2. Goodman's Association Models for Tables with Ordered Categories.

A recent article by Goodman (1979) presents a class of models for I × J contingency tables with ordered categories. This class of models postulates a structure for the odds ratio (association) in the 2 × 2 subtables formed by cells in adjacent rows and columns of the table. (It is well known that the odds ratio for any 2 × 2 subtable can be recovered from the odds ratios in the adjacent subtables.) Goodman presents two classes of models for the table, but we will consider only the loglinear version, which Goodman denotes as Model I.

Consider the odds ratio, θ_{ij}, of the 2 × 2 subtable formed by the intersection of rows i and i+1 with columns j and j+1, i.e.,

$$\theta_{ij} = \frac{P_{i,j} \cdot P_{i+1,j+1}}{P_{i,j+1} \cdot P_{i+1,j}} .$$

The model we wish to consider asserts that θ_{ij} can be written as the product of a row effect and a column effect, i.e.,

$$\theta_{ij} = \theta_{i.} \cdot \theta_{.j}$$

or $\ln(\theta_{ij}) = \ln(\theta_{i.}) + \ln(\theta_{.j})$. When written as a loglinear model for the expected probabilities this becomes:

$$\ln(p_{ij}) = \alpha_i + \beta_j + j\gamma_i + i\delta_j .$$

We now describe a spanning set and some of the calculations required to fit this model for the special case of a 3 × 3 table.

The linear manifold, $\mathcal{M}$, for this model is spanned by a set of tables, f_R^i, f_{OR}^i, f_C^j and f_{OC}^j; i,j = 1,2,3. The subscripts R, OR, C and OC indicate that the vector corresponds to Row, Ordered Row, Column or Ordered Column parts of the model, while the superscript indicates the row or column number, e.g.,

$$f_R^2 = \begin{array}{|ccc|}\hline 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \\ \hline \end{array} \qquad f_{OR}^2 = \begin{array}{|ccc|}\hline 0 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \\ \hline \end{array}$$

$$f_C^1 = \begin{vmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{vmatrix}, \qquad f_{OC}^3 = \begin{vmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 2 \end{vmatrix}.$$

The general structure is that f_R^i (or f_C^j) is a table of zeros except for the ith row (jth column) which contains ones, i.e.,

$$f_R^i(k,\ell) = \begin{cases} 1 & k = i \\ 0 & k \neq i. \end{cases}$$

Similarly, for the ordered row and column tables, the general form is

$$f_{OC}^j(k,\ell) = \begin{cases} k-1 & \ell = j \\ 0 & \ell \neq j. \end{cases}$$

We now group the spanning tables into sets of related constraints. Let

$$\mathcal{F}_R = \{f_R^i, f_{OR}^i : i = 1,2,3\}$$

and

$$\mathcal{F}_C = \{f_C^j, f_{OC}^j : j = 1,2,3\}.$$

We also need the corresponding sets of constants, $\mathcal{A}_R$ and $\mathcal{A}_C$. Generally these are determined from some observed table of probabilities. The linear spaces of P.D.'s corresponding to these constraints and constants are:

$$\mathcal{E}_R = \{\text{P.D.'s P s.t. } \int f_A^i dP = a_A^i;\ A = 0,OR;\ i = 1,2,3\}$$

$$\mathcal{E}_C = \{\text{P.D.'s P s.t. } \int f_B^j dP = a_B^j;\ B = C,OC;\ j = 1,2,3\}.$$

In order to find the M.L.E.'s of cell probabilities for this model we need to be able to compute, $\mathbb{P}_{\mathcal{E}}(R)$ for some suitable R and $\mathcal{E} = \mathcal{E}_R \cap \mathcal{E}_C$. The theory tells us that this I-projection can be obtained by cyclically projecting onto $\mathcal{E}_R$ and $\mathcal{E}_C$. It is these projections which we now compute. The first observation we should make is that each of the spaces is generated by three pairs of functions with disjoint support. For example, $\mathcal{E}_R$ is generated by the pair of functions (f_R^1, f_{OR}^1), (f_R^2, f_{OR}^2) and (f_R^3, f_{OR}^3) and the support for these functions is respectively the first, second, and third rows of the space of tables. Thus we can apply the corollary to this estimation problem.

Consider a starting table (which may already be the result of several iterations),

$$r = \begin{array}{|c|c|c|} \hline r_{11} & r_{12} & r_{13} \\ \hline r_{21} & r_{22} & r_{23} \\ \hline r_{31} & r_{32} & r_{33} \\ \hline \end{array}$$

Then $\mathbb{P}_{\mathcal{E}_R}(R)$ is the probability function, p, which minimizes

$$(2.10) \qquad \sum_{ij} p_{ij} \ln (p_{ij}/r_{ij})$$

subject to p being in $\mathcal{E}_R$. By applying the corollary we can separately minimize (2.10) for each i (i.e., row) and combine the results. For i = 1 we need to minimize,

$$(2.11) \qquad (a_R^1 - a_{OR}^1 + p_{13}) \ln (a_R^1 - a_{OR}^1 + p_{13}/r_{11}) + (a_{OR}^1 - 2p_{13}) \ln (a_{OR}^1 - 2p_{13}/r_{12}) + p_{13} \ln (p_{13}/r_{13}).$$

subject to $p_{13} \in (0,1)$.

By taking the derivative of (2.11) with respect to p_{13} and equating it with zero we obtain the equation:

$$(2.12) \quad \ln (a_R^1 - a_{OR}^1 + p_{13}/r_{11}) - 2 \ln (a_{OR}^1 - 2p_{13}/r_{12}) + \ln (p_{13}/r_{13}) = 0,$$

which can be written as

$$\frac{(a_R^1 - a_{OR}^1 + p_{13}) \cdot (p_{13}) \cdot (r_{12})^2}{(a_{OR}^1 - p_{13})^2 \, r_{11} \cdot r_{13}} = 1,$$

or equivalently as,

$$p_{13}^2(1 - \frac{r_{11} \cdot r_{13}}{r_{12}^2}) + p_{13}(a_R^1 - a_{OR}^1 + 2a_{OR}^1 \frac{r_{11} \cdot r_{13}}{r_{12}^2}) - \frac{r_{11} \cdot r_{13}}{r_{12}^2} (a_{OR}^1)^2 = 0.$$

This equation is relatively easy to solve for p_{13}. The estimates for p_{11} and p_{12} are derived by solving the constraint equations. The equations from the second and third rows and the columns are analogous.

If we consider the same class of models for I × J tables where one of I or J is greater than 3, the resulting equations are systems of higher order polynomials. Clearly, solving such systems may themselves be a difficult task.

In the next section we will show another algorithm that can be used for this problem. ▪

The preceeding examples have used an I-divergence approach to the IPFP. We now consider the approach discussed by Haberman (1974, p. 64). That discussion uses the method of co-ordinate cyclic ascent to directly maximize the likelihood. A fixed set of vectors which span the model space, $\mathcal{M}$, is chosen and the likelihood is maximized along each of these directions in

turn. Specifically, consider a set of vectors $\mathcal{F} = \{f_\gamma : \gamma \in \Gamma\}$ which span $\mathcal{M}$, denote the log-likelihood by $\ell(p;s)$ and consider an initial estimate p^0 with $\ln(p^0)$ in $\mathcal{M}$. The algorithm proceeds by finding p^i such that

$$\ln(p^i) = \ln(p^{i-1}) + \alpha_i f_\gamma; \quad i = \gamma \bmod |\Gamma|,$$

where α_i is determined so that

$$|\alpha_i \cdot \langle \dot{\ell}(p^{i-1};s), f_\gamma \rangle| \leq b \cdot |\ell(p^i;s) - \ell(p^{i-1};s)|$$

for some fixed b in (0,1). Generally we would find α_i by attempting to maximize

$$(2.13) \qquad \ell(\exp\{\ln(p^{i-1}) + \alpha_i f_\gamma\}).$$

This is a one-dimensional maximization problem in the fixed direction, f_γ. This problem can be re-expressed as:

$$(2.14) \qquad \text{maximize } \ell(p^i),$$

subject to,

$$\ln(p^i) \in \ln(p^{i-1}) + \text{span}(f_\gamma),$$

which is a maximum likelihood problem for a log affine model. Csiszár's Theorem 3.1 showed that this has a dual representation as a minimum I-divergence problem. But the solution, p^i, to (2.14) is not necessarily a P.D., even though $P^\infty = \lim_{i\to\infty} p^i$ must be a P.D. To rectify this situation we consider the related problem:

(2.15) maximize $\ell(p^i)$,

subject to,

$$\ln(p^i) \in \ln(p^{i-1}) + \text{span}(f_\gamma, e),$$

where e is the vector of all ones. The dual representation of this problem is to consider $\mathcal{E}_\gamma = \{\text{P.D.'s P s.t. } \int f_\gamma dP = \int f_\gamma dP^{i-1}\}$ and then form p^i as $P_{\mathcal{E}_\gamma}(P^{i-1})$. Thus, in a certain sense, the co-ordinate cyclic ascent methods are conjugate dual problems to the algorithms of Csiszár. It appears that the cyclic ascent method is easier to work with as it does not require the result of each iteration to be a P.D.

If we use the duality result the other way around, we can describe the I-projections in an alternate form. For example, consider the linear space $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$ generated by $\mathcal{F}_1 \cup \mathcal{F}_2$ and suppose we wish to calculate $P_{\mathcal{E}_i}(Q)$ for an arbitrary P.D. Q. The dual to this problem is to

maximize $\ell(p)$,

subject to,

$$\ln(p) \in \ln(q) + \text{span}(\mathcal{F}_i, e).$$

Thus another type of IPFP maximizes the likelihood, not in a set of fixed directions, but in a set of planes spanned by $(\mathcal{F}_i, e)$. If R is a starting vector such that $\ln(r) \in \mathcal{M}$ and $\mathcal{M}$ is equal to span $(\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_k)$ where the $\mathcal{F}_i$ are not necessarily 1 dimensional subspaces then the following algorithm converges. Form $\mathcal{G}_i$ = span $(\mathcal{F}_i, e)$ and $P^0 = R$. Then p^{i+1} is the p which maximizes $\ell(p)$ subject to $\ln(p)$ is in span $(\mathcal{G}_i)$, where $\mathcal{G}_i = \mathcal{G}_j$ if $i = j \bmod k$. If each of these problems is easy to do then this may form a useful algorithm.

3. The Darroch and Ratcliff Generalized Iterative Scaling

The preceeding section has shown how Csiszár's Theorems and the "dual" theorems of Haberman may be cyclically applied to compute I-projections and maximum likelihood estimates. A paper by Darroch and Ratcliff (1972) attacks the same problem, again by looking at it from the information theory point of view; however their generalization of the IPFP is different from those we have thus far considered. Darroch and Ratcliff (D & R) succeed in calculating the total I-projection without necessarily calculating any of the marginal I-projections. In the "usual" case, i.e., where the space, &, is generated by vectors containing only zeros and ones their generalization also reduces to the conventional IPFP algorithm.

The D & R algorithm, or Generalized Iterative Scaling (GIS), ensures that at each iteration equation (2.8) is satisfied (i.e., $q_R(x) = c.\ \exp(g(x))$, where g is in the linear space spanned by the constraints). When the algorithm has converged one is able to show that the fitted values also satisfy the marginal constraints. This should be contrasted to the algorithms we have discussed earlier. The algorithms of Csiszár and Haberman alternately satisfy the marginal constraints, with only the final fitted values necessarily satisfying equation (2.8). We now consider the D & R algorithm in more detail.

Let $\mathcal{f} = \{f_\gamma : \gamma \in \Gamma\}$ be a set of constraints with corresponding constants $\mathcal{A} = \{a_\gamma : \gamma \in \Gamma\}$ and consider the linear set of P.D.'s

$$\& = \{\text{P.D.'s } P : \int f_\gamma dP = a_\gamma; \quad \gamma \in \Gamma\}.$$

As before, & is just the set of P.D.'s whose $\mathcal{f}$-margins are correct. It is always possible, as D & R show, to find a set of vectors $\mathcal{g} = \{g_\delta : \delta \in \Delta\}$ whose span is the same as span ($\mathcal{f}$) and which satisfy

(3.1) $\quad g_\delta \geq 0 \ \forall \delta \quad$ and $\quad \sum_{\delta \in \Delta} (g_\delta) = e.$

In this formulation,

$$\mathcal{E} = \{\text{P.D.'s } P : \int g_\delta dP = b_\delta;\ \delta \in \Delta\}$$

where the b_δ are determined from the a_γ. We now consider the problem of finding the I-projection of some R onto $\mathcal{E}$, i.e., we wish to find a $P \in \mathcal{E}$ such that

$$p = r \cdot \exp\{\sum_{\delta \in \Delta} \lambda_\delta \cdot g_\delta\}$$

where the λ_δ are to be determined. D & R show that the following algorithm converges to the M.L.E.:

(i) set $p^0 = r$

(ii) set $p^{n+1} = p^n \cdot \prod_{\delta \in \Delta} \{(b_\delta / \langle g_\delta, p^n \rangle)^{g_\delta}$

$$= p^n \exp\{\sum_{\delta \in \Delta} g_\delta \cdot \ln (b_\delta / \langle g_\delta, p^n \rangle)$$

where $\delta = n \bmod |\Delta|$. The algorithm, as given, adjusts for all of the marginal constraints at once. However, it is possible to adjust for several sets of simultaneous constraints, one set at a time using partitions not unlike those those discussed in Section 2.

Consider two linear spaces of P.D.'s, $\mathcal{E}_1$ and $\mathcal{E}_2$, defined by constraint sets $\mathcal{F}_1$ and $\mathcal{F}_2$, each of which satisfies equation (3.1) on its support. Csiszar's results suggest that to calculate the I-projection of $Q_0 = R$ onto $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2$ one should successively form

$$Q_n = P_{\mathcal{E}_i}(Q_{n-1}), \quad i = n \bmod 2.$$

The GIS algorithm would be one way of calculating these elementary I-projec-

tions. Darroch and Ratcliff suggest an alternative approach which does not necessarily involve calculating the individual I-projections. The idea is to perform one iteration of the previous algorithm on the space $\mathcal{E}_1$, then one iteration on $\mathcal{E}_2$ and continue cycling. If we let $\mathcal{G}_1 = \{g^1_\delta;\ \delta \in \Delta_1\}$ and $\mathcal{G}_2 = \{g^2_\delta;\ \delta \in \Delta_2\}$, and let $\{b^1;\ \delta \in \Delta_1\}$ and $\{b^2;\ \delta \in \Delta_2\}$ be the associated constraint spaces then the algorithm would proceed as follows:

(i) set $p^0 = r$

(ii) set $p^{n+1} = p^n \cdot \prod_{\delta=1}^{\Delta_i} \{(b^i/\langle g^i, p^n\rangle)^{g^i_\delta}\}$, where $i = n \bmod 2$.

To illustrate the ideas presented here we reconsider Example 2.

Example 2 (continued).

We illustrate one of the ways that the GIS algorithm can be used to find the M.L.E.'s for Goodman's association model in a 3×3 table. Recall that the constraints came in natural pairs (e.g., f^2_R and f^2_{OR}) of a row (column) and ordered row (column) function. These pairs do not satisfy equation (3.1) on their support, but we can convert them into:

$$g^j_R = f^j_R - \tfrac{1}{2} f^j_{OR} \quad \text{and} \quad g^j_{OR} = \tfrac{1}{2} f^j_{OR}$$

which still span the same space. We also need to make a similar adjustment to the constants, viz:

$$b^j_R = a^j_R - \tfrac{1}{2} a^j_{OR} \quad \text{and} \quad b^j_{OR} = \tfrac{1}{2} a^j_{OR}.$$

Analogous transformations are made to the columns. As all the pairs of constraints are similar we will concentrate on the pair corresponding to the first row and consider only one step of the algorithm. We note that

$$g_R^1 = \begin{array}{|c|c|c|}\hline 1 & 1/2 & 0 \\ \hline 0 & 0 & 0 \\ \hline 0 & 0 & 0 \\ \hline \end{array} \qquad \text{and}$$

$$g_{OR}^1 = \begin{array}{|c|c|c|}\hline 0 & 1/2 & 1 \\ \hline 0 & 0 & 0 \\ \hline 0 & 0 & 0 \\ \hline \end{array}$$

If Q^n is our current approximation to $\mathbb{P}_{\mathcal{E}}(R)$ then the GIS algorithm would form as its next estimate,

$$q_{ij}^{n+1} = q_{ij}^n \cdot \left(\frac{\int g_R^1 dQ^n}{b_R^1}\right)^{g_R^1(i,j)} \cdot \left(\frac{\int g_{OR}^1 dQ^n}{b_{OR}^1}\right)^{g_{OR}^1(i,j)} .$$

The algorithm continues by considering each of the constraint sets in turn. In this example we sometimes need to take square roots of the ratio of the observed margin to the expected margin. In a more typical situation we would take arbitrary powers rather than just square roots.

Note that applying the same adjustment to the new table (i.e., <u>not</u> cycling through the pairs of constraints) produces another new table. If we were to continue with the same pair of constraints we would arrive at the I-projection onto that constraint space. Thus in many respects GIS just combines the first steps of an algorithm to compute I-projections. ∎

The GIS algorithm is a method which is conceptually easy to compute and guaranteed to converge. Unfortunately the algorithm is also known to converge very slowly in some situations. In contrast, the Csiszár approach is appealing as it maximizes along a fixed space at each step but it has the disadvantage that the elementary I-projections may themselves be difficult to compute or require iteration. Which procedure (or combination) is better may depend on the problem under investigation but certainly requires further study.

4. Methods for Maximum Likelihood Estimation in Special Cases

In this section we shall use the ideas of the preceeding sections to study some problems in which the constraints, $\mathcal{F}$, have a special structure. We consider as examples the ordered categories model for a 3 × 3 table introduced earlier and a special situation considered in Fienberg and Wasserman (1980). In both examples we will find it edifying to expand the table (i.e., increase the number of cells) and fit a transformed model to the larger table. Clearly we will need some conditions on the model and how we "expand" the table. The following "theorem" is a collection of conditions which we will need to verify in the examples. In general verifying the conditions may itself be a difficult task.

Theorem 3

Let g be a one to one mapping of the P.D.'s on a set X into the P.D.'s on a set X*. If & is a linear set of P.D.'s on X, then define $g(\&) = \{g(P) : P \in \&\}$. Let &* be a linear set of P.D.'s on X* such that $g(\&) \subset \&^*$. If g is such that

$$(4.1) \qquad I(P||Q) = k \cdot I(g(P)||g(Q)) \text{ for } P,Q \in \&$$

and $\mathbb{P}_{\&^*}(g(R)) \in g(\&)$ then

$$\mathbb{P}_{\&}(R) = g^{-1}(\mathbb{P}_{\&^*}(g(R))). \quad \blacksquare$$

This theorem allows us, under certain conditions, to calculate the I-projections in a transformed problem and then invert the transformation to obtain the I-projection in the original setting. There are at least two ways of using the theorem. In some situations it may be possible to define the linear set &* so that $g(\&) = \&^*$. This is the easier case and it essen-

tially just relabels the problem. However even such simple relabeling can be helpful if it helps one to interpret the model or recognize, say, a model in the transformed space for which closed form estimates are known to exist. The second application of the theorem requires more work to verify the conditions, but is also more generally applicable. Here we take a linear set $\&^*$ which is much larger than $g(\&)$, we need to prove that $P_{\&^*}(g(r)) \in g(\&)$. In other words, even though $\&^*$ contains $g(\&)$ we need to show that for any $g(R)$, the I-projection onto $\&^*$ is always an element of $g(\&)$. For a particular set of data it may be easy to verify this condition. All we need do is fit the transformed model and see if the I-projection is in $g(\&)$. To prove this type of result for a general class of R's and &'s is much more difficult. These ideas are best illustrated in the context of two examples.

Example 2 (Continued)

We have previously shown that the row and column constraints can be considered in pairs and each of the pairs of constraints can be individually fit. Thus if (w_1,w_2,w_3) are the current fitted values for, say, the first row, we need to adjust this triple so that its row and ordered row margins match some specified constants.

Let $\&_S$ be the set of positive triples which satisfy the row and ordered row constraints for the first row, i.e.,

$$\&_S = \{\text{positive triples, } q : 2q_1 + q_2 = 2a_R^1 - a_{OR}^1 = a_3 \text{ and } q_2 + 2q_3 = a_{OR}^1 = a_4\}.$$

As the vector $e = (1,1,1)$ is in the span of the space of constraints which defines $\&_S$, we can apply the corollary of Csiszár's Theorem 3-2 and just work with $P_{\&_S}(W)$. Now consider the function

$$g : w \rightarrow \begin{array}{|c|c|} \hline w_1 & \frac{1}{2} w_2 \\ \hline \frac{1}{2} w_2 & w_3 \\ \hline \end{array}$$

and define

$$\mathcal{E}^* = g(\mathcal{E}_S)$$

$$= \{2 \times 2 \text{ tables } \begin{array}{|c|c|} \hline a & b \\ \hline c & d \\ \hline \end{array} \text{ such that } a + b = a + c = \frac{1}{2} a_3 \text{ and } d + c = d + d = \frac{1}{2} a_4\}.$$

Note that the constraints on $\mathcal{E}^*$ imply that b equals c which means that g^{-1} is well defined on $\mathcal{E}^*$. It is not a difficult calculation to verify that $I(Q||W) = I(g(Q)||g(W))$. Our theorem now allows us to calculate $\mathbf{P}_{\mathcal{E}_S}(W)$ as $g^{-1}\mathbf{P}_{\mathcal{E}^*}(g(W))$.

The constraints which define $\mathcal{E}^*$ are just simple row and column margins. Thus the I-projection, $\mathbf{P}_{\mathcal{E}^*}(g(W))$, can be calculated by the usual IPFP (i.e., adjusting row and column margins), or, as it in a 2×2 table, by direct calculation. As the logarithms of the starting values, w, do not necessarily satisfy any model, the IPFP will in general require several iterations to converge. Thus to obtain the I-projection, $\mathbf{P}_{\mathcal{E}_R}(Q_n)$, where $\mathcal{E}_R$ is the space of P.D.'s which satisfy all of the row constraints, we could transform each row of the 3×3 table into a 2×2 table, calculate with the 2×2 table and then use g^{-1} to return a triple of fitted values. The approach for the columns would be similar.

There is another g, which transforms the entire 3 × 3 table into a 2 × 2 × 2 × 2 table. In this case &* = g(&) becomes the model of no fourth order interaction for the 2^4 table. Specifically,

g :

a	b	c
d	e	f
g	h	i

→

a	$\frac{1}{2}$ b
$\frac{1}{2}$ d	$\frac{1}{4}$ e

$\frac{1}{2}$ b	c
$\frac{1}{4}$ e	$\frac{1}{2}$ f

$\frac{1}{2}$ d	$\frac{1}{4}$ e
g	$\frac{1}{2}$ h

$\frac{1}{4}$ e	$\frac{1}{2}$ f
$\frac{1}{2}$ h	i

It is not difficult to check that the model of no fourth order interaction corresponds to g(&) and that I(P||Q) = I(g(P)||g(Q)). Therefore the usual IPFP, with starting values g(e) and the model of no fourth order interaction applied to $g(Q_n)$ will yield a 2^4 table of fitted values which can in turn be transformed (by g^{-1}) into a 3 × 3 table for the original problem. ∎

Both applications of the theorem in the previous example used an &* which was equal to g(&). The following example gives a situation where &* is much larger than g(&). Here we need some trickery to show that the I-projection of g(R) onto &* is in g(&).

Example 3.

Fienberg and Wasserman (1980) describe a class of loglinear models for some multivariate directed graphs. Their paper considers as an example a set of data concerning the interrelationships between 73 organizations in a small community. Three types of relations were observed

for each of the pairs of organizations, but for simplicity we restrict our attention to two of these criteria, support and money. For each criterion the organizations were asked to respond to the questions:

(i) to which organizations do you give support (money)?

(ii) from which organizations do you receive support (money)?

A particular directed relationship (i.e., giving or receiving) is regarded to be present if either or both the organizations in a pair perceived the relationship. For each pair of organizations it is possible to construct a four-vector of zeros and ones indicating the presence or absence of (support out, support in, money out, money in). Consider for a moment just the support relationship. A pair of organizations are said to have a Mutual relationship if they support each other (i.e., (support out, support in) = (1,1)), a Null relationship if neither supports the other (i.e., (0,0)), or an Asymmetric relationship if support is unreciprocated (i.e., (0,1) or (1,0)). If we aggregate over all $\binom{73}{2}$ = 2628 pairs of organizations there are ten distinguishable support-money relationships, namely:

MM	with four vector	(1,1,1,1)
MA		(1,1,0,1) or (1,1,1,0)
MN		(1,1,0,0)
AM		(0,1,1,1) or (1,0,1,1)
AA		(0,1,0,1) or (1,0,1,0)
$A\bar{A}$		(0,1,1,0) or (1,0,0,1)
AN		(0,1,0,0) or (1,0,0,0)
NM		(0,0,1,1)
NA		(0,0,1,0) or (0,0,0,1)
NN		(0,0,0,0)

Notice that when both relationships are asymmetric there are two different cases, corresponding to whether the relationships flow the in the same or in different ways. We denote the table of observed probabilities by Z where for example z_{MM} is the number of mutual-mutual relationships divided by $\binom{73}{2}$. The table is represented by

$Z =$

SUPPORT \ MONEY	M	A	N
M	z_{MM}	z_{MA}	z_{MN}
A	z_{AM}	z_{AA} / $z_{A\bar{A}}$	z_{AN}
N	z_{NM}	z_{NA}	z_{NN}

An alternate, though somewhat deceptive, description of the data is to consider four-vectors for each of the $\binom{73}{2} \times 2$ ordered pairs of organizations and to aggregate this into a 2^4 table, $Y = y_{ijk\ell}$, $ijk\ell = 1,2$, where a 1 indicates the presence of a flow and a 2 indicates the absence of a flow. Thus y_{1111} is the number of mutual-mutual relationships divided by 5256. The Y table duplicates certain relationships and gives double weight to certain others. The Y-table has the form,

	money out	1		2	
	money in	1	2	1	2
supp out	supp in				
1	1	y_{1111}	y_{1112}	y_{1121}	y_{1122}
	2	·	·	·	·
2	1	·	·	·	·
	2	·	·	·	y_{2222}

We now consider one of the models for the Z-table considered by Fienberg and Wasserman. (The same arguments work for all of their models.) The model takes as a linear space, &, of P.D.'s the set of tables, S, which have margins S_{a+} and S_{+b}, a,b = M,A,N, which are the same as the corresponding margins for the Z-table. For example we require

$$S_{a+} = s_{AM} + s_{AA} + s_{A\bar{A}} + s_{AN} = z_{AM} + z_{AA} + z_{A\bar{A}} + z_{AN} = Z_{A+}.$$

This model can be fit directly to the Z-table using the methods of the previous sections. As the model space can be spanned by vectors consisting of only 0's and 1's, both the D & R and Csiszár algorithms reduce to the same simple scaling algorithm which takes an initial table of all 1's and successively adjusts the row and column "margins" to match those in the observed table. This algorithm is easy to do by hand, but because the Z-table is not square (i.e., it has 10 cells rather than the 9 one would expect), and consequently has an extended interpretation of marginal totals, standard IPFP computer programs would not be able to analyze this table. Moreover, for some of the models considered by Fienberg and Wasserman

the models are not so simple and the computations on the Z-table require the full power of the generalized IPFP's. For this reason we prefer to work with a transformed problem, where the sufficient statistics for the models can be represented by simple marginal totals.

Consider

$$g\ ;\ Z \to \frac{1}{2}\begin{array}{|c|c||c|c|}\hline 2z_{MM} & z_{MA} & z_{MA} & 2z_{MN} \\ \hline z_{AM} & z_{AA} & z_{A\bar{A}} & z_{AN} \\ \hline\hline z_{AM} & z_{A\bar{A}} & z_{AA} & z_{AN} \\ \hline 2z_{NM} & z_{NA} & z_{NA} & 2z_{NN} \\ \hline\end{array}$$

$$= Y$$

which maps the Z-table into the 2^4 Y-table. We denote the factors support (out, in), money (out, in) by the numbers 1, 2, 3, and 4. It is now easy to see that the marginal sums considered for the Z-table can all be found (twice) in the [12] and [34] margins of the Y-table. Also note that the Y-table has a strong symmetry, $y_{ijk\ell} = y_{ji\ell k}$ $\forall ijk\ell$. Now g(&) is just the set of tables which have (i) the correct [12] and [34] margins and (ii) preserve the observed symmetry in the Y-table. Consider just the first of these conditions ignoring the symmetry constraint. It is this model which we shall consider to be &*. As we have relaxed some conditions it is clear that $g(\&) \subset \&^*$.

It is convenient now to explicitly define the space &* and the conditions we need to verify to show that $\mathbb{P}_{\&^*}(g(R))$ is in g(&). Consider

$\mathcal{f} = \{f_1,\ldots,f_8\}$ where

$f_1 =$

1	1	1	1
0	0	0	0
0	0	0	0
0	0	0	0

$\ldots$ $f_4 =$

0	0	0	0
0	0	0	0
0	0	0	0
1	1	1	1

$f_5 =$

1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0

$\ldots$ $f_8 =$

0	0	0	1
0	0	0	1
0	0	0	1
0	0	0	1

and constants $\mathcal{a} = \{a_1,\ldots,a_8\}$ where $a_j = \langle f_j, g(Z)\rangle$. Note that $a_2 = a_3$ and $a_6 = a_7$. We define &* to be the space of P.D.'s defined by $\mathcal{f}$ and $\mathcal{a}$. Now consider the symmetry transformation:

$$h : y_{ijk\ell} \to y_{ji\ell k}.$$

For $\mathbb{P}_{\&*}(g(R))$ to be in $g(\&)$ we require

$$h(\mathbb{P}_{\&*}(g(R))) = \mathbb{P}_{\&*}(g(R)).$$

It is possible to assert this because the space &* is invariant under h. Specifically $h(f_i) = f_i$ for $i = 1,4,5,8$ and $h(f_2) = f_3$, $h(f_3) = f_2$, $h(f_7) = f_6$ and $h(f_6) = f_7$. Because $a_2 = a_3$ and $a_6 = a_7$ the linear space $h(\&*)$ generated by $h(\mathcal{f})$ and $h(\mathcal{a})$ is the same as &*. We also note that $h(g(R)) = g(R)$, because of the nature of g function. That is the starting values necessarily satisfy the symmetry constraints. Now let

$$\hat{Q} = \mathbb{P}_{\&^*}(g(R)) \quad \text{and}$$

$$\tilde{Q} = \mathbb{P}_{h(\&^*)}(h(g(r))) = \mathbb{P}_{\&^*}(g(R)).$$

But note that $\tilde{Q} = h(\hat{Q})$ as all we have done is relabeled the co-ordinates. Thus

$$\hat{Q} = \tilde{Q} = h(\hat{Q}),$$

i.e., the fitted P.D. is (i) invariant under h and (ii) is in &*. Thus $\hat{Q}$ is in g(&) and $g^{-1}(\hat{Q})$ is the fitted P.D. in the space of Z-tables.

For any of the other models considered by Fienberg and Wasserman, it is easy to show that the space, &*, is invariant under h and thus the above argument still works. ■

Both the examples of this section have shown situations where, for reasons of computational ease, it was desirable to transform a contingency table into a related but larger table. In the transformed table it was possible to fit a model using the standard IPFP whereas in the original table the corresponding model would have required a more complicated algorithm. This approach of using transformed tables is especially important in practice as versions of the standard IPFP are widely available and easy to use. An additional bonus which can sometimes be found in the transformed table is the existence of closed form maximum likelihood estimates. The theory about when closed form estimates exist in complete tables with factorial models is well known and such situations are easily recognized. On the contrary, when a table is incomplete or has a more complicated structure, very little is known about the existence of closed form estimates. The insight gained from looking at the transformed table may also assist in interpreting the models.

Appendix 1. Analogies with Hilbert Space

In this appendix we discuss some of the analogies between the IPFP and methods for cyclically calculating projections in Hilbert spaces.

Consider $(V,\langle\cdot,\cdot\rangle)$ to be a finite dimensional Hilbert space and let $\&_1,\ldots,\&_k$ be linear subspaces of V with corresponding orthogonal projections $E_1,\ldots,E_k$. In other words the orthogonal projection of a vector $v \in V$ onto $\&_i$ will be denoted by $E_i v$. The following theorem can be shown to be true.

Theorem A1.1.

If $(V,\langle\cdot,\cdot\rangle)$ is a finite dimensional Hilbert Space and $\&_1,\ldots,\&_k$ are linear subspaces of V then the orthogonal projection of any $v \in V$ onto $\& = \bigcap_{i=1}^{k} \&_i$ is equal to

$$\lim_{m\to\infty} \{(E_k \cdot E_{k-1} \cdot \ldots \cdot E_1)^m v\}. \blacksquare$$

A simple extension of this result states that if Q_n is defined to be the projection of Q_{n-1} onto $\&_n$, where $\&_n = \&_i$ when $i = n \bmod k$, and $Q_0 = v$ then Q_n converges to the projection of v onto &. This is a direct analogue of Csiszár's Theorem 3.2. I am not sure if the above theorem is always true when $(V,\langle\cdot,\cdot\rangle)$ is an infinite dimensional Hilbert space, but it is true when any of the $\&_i$ are finite dimensional. There is however a version of cyclically projecting onto subspaces which is always true (for a proof see Von Neumann (1950)).

Theorem A1.2.

If $(V,\langle\cdot,\cdot\rangle)$ is a Hilbert space, the orthogonal projection of any $v \in V$ onto $\& = \bigcap_{i=1}^{k} \&_i$ is equal to

$$\lim_{m\to\infty} (E_1 \cdot E_2 \cdot E_3 \cdot \ldots \cdot E_{k-1} \cdot E_k \cdot E_{k-1} \cdot \ldots \cdot E_2 \cdot E_1)^m v. \blacksquare$$

In this version of the theorem we use a symmetric form of the operator. Again it is true that the piecewise projections (in the correct order) converge. The advantage of Theorem A1.2 is that powers of symmetric operators generally converge more quickly than do powers of unsymmetric ones.

The proof of Csiszár's Theorem 3.2 can easily be modified to prove the symmetric version of that theorem. Arguing by analogy with Hilbert spaces we conjecture that a symmetric form of the IPFP may converge more quickly than the usual version. This conjecture needs to be numerically investigated.

References

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

Csiszár, I. (1975), "I-Divergence geometry of probability distributions and minimization problems," The Annals of Probability, 3:146-158.

Darroch, J.N. and D. Ratcliff (1972), "Generalized iterative scaling for log-linear models," The Annals of Mathematical Statistics, 43:1470-1480.

Fienberg, S.E. and S. Wasserman (1980), "Methods for the analysis of data from multivariate directed graphs," to appear in Proceedings of Conference on Recent Developments in Statistical Methods and Applications, Institute of Mathematics, Academia Sinica, Taipai, Taiwan, Republic of China.

Goodman, L.A. (1979), "Simple models for the analysis of association in cross-classifications having ordered categories," Journal of the American Statistical Association, 74:537-52.

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

Kullback, S. (1959), Information Theory and Statistics. New York: Wiley.

Luenberger, D.G. (1969), Optimization by Vector Space Methods. New York: Wiley.

Rockafeller, R.T. (1974), Conjugate Duality and Optimization, Regional Conference Series in Applied Mathematics, Number 16. Philadelphia: Society for Industrial and Applied Mathematics.

Von Neumann, J. (1950), Functional Operators, Volume II: The Geometry of Orthogonal Spaces, Annals of Mathematical Studies, Number 22. Princeton: Princeton University Press.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Technical Report No. 371	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): Generalizing the Iterative Proportional Fitting Procedure	5. TYPE OF REPORT & PERIOD COVERED
6. PERFORMING ORG. REPORT NUMBER	
7. AUTHOR(s): Michael M. Meyer	8. CONTRACT OR GRANT NUMBER(s): N00014-78-C-0151
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Applied Statistics School of Statistics, University of Minnesota 1994 Buford Avenue, St. Paul, MN 55108	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research 800 N. Quincy Street Arlington, VA 22217	12. REPORT DATE: April 1980 13. NUMBER OF PAGES: 42
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report) 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Generalized Iterative Scaling; I-divergence; Kullback-Leibler Information Number; Contingency Tables

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The IPFP can be viewed as a method for maximizing the likelihood for certain loglinear models or equivalently for minimizing the Kullback-Leibler Information between two probability densities. Both of these viewpoints lead to natural generalizations of the classical IPFP. We examine the generalizations suggested by the work of Csiszár (1975), Darroch and Ratcliff (1972), and Haberman (1974) and, with the aid of the theory, explore a practical example of expanding a contingency table.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)