Generalized Iterative Scaling for Log-Linear Models

J. N. Darroch; D. Ratcliff

GENERALIZED ITERATIVE SCALING FOR
LOG-LINEAR MODELS

By J. N. Darroch and D. Ratcliff

Flinders University of South Australia and C.S.I.R.O. Adelaide

Say that a probability distribution $\{p_i; i \in I\}$ over a finite set I is in "product form" if (1) $p_i = \pi_i \mu \prod_{s=1}^{d} \mu_s^{b_{si}}$ where π_i and $\{b_{si}\}$ are given constants and where μ and $\{\mu_s\}$ are determined from the equations (2) $\sum_{i \in I} b_{si} p_i = k_s$, $s = 1, 2, \cdots, d$; (3) $\sum_{i \in I} p_i = 1$. Probability distributions in product form arise from minimizing the discriminatory information $\sum_{i \in I} p_i \log p_i / \pi_i$ subject to (2) and (3) or from maximizing entropy or maximizing likelihood.

The theory of the iterative scaling method of determining (1) subject to (2) and (3) has, until now, been limited to the case when $b_{si} = 0, 1$. In this paper the method is generalized to allow the b_{si} to be any real numbers. This expands considerably the list of probability distributions in product form which it is possible to estimate by maximum likelihood.

0. Summary. The theory of iterative scaling is generalized and, at the same time, the existing theory is formulated in a general setting and the proof of convergence is shortened. The application to maximum likelihood estimation is discussed, with special reference to truncated distributions, contingency tables and F-independence.

1. Introduction. Let I be a finite set and let $\mathbf{p} = \{p_i; i \in I, p_i \geq 0, \sum_{i \in I} p_i = 1\}$ be a probability function on I. Further let $\boldsymbol{\pi} = \{\pi_i; i \in I, \pi_i > 0, \sum_{i \in I} \pi_i \leq 1\}$ be a positive, "sub-probability" function on I.

"Generalized iterative scaling" is a method for finding a probability function of the form

$$(1) \qquad\qquad p_i = \pi_i \mu \prod_{s=1}^{d} \mu_s^{b_{si}}$$

which satisfies the constraints

$$(2) \qquad\qquad \sum_{i \in I} b_{si} p_i = k_s, \qquad s = 1, 2, \cdots, d, \ \sum_{i \in I} p_i = 1,$$

where $\forall s$, $\exists i \in I$ such that $b_{si} \neq 0$. The constants π_i and b_{si} are given and μ, μ_s are to be found.

We shall describe model (1) as log-*linear* since it says that $\{\log p_i / \pi_i; i \in I\}$ is a linear combination of known vectors $\{l; i \in I\}$ and $\{b_{si}; i \in I\}$, $s = 1, 2, \cdots, d$. Apart from Kullback (1971), "log-linear" has previously been used only for the case when $b_{si} = 0, 1 \ \forall i, s$, (see Bishop (1969), Gokhale (1971)). Model (1) is better known as an example of an *exponential family* of probability functions. For recent work on the general exponential family, with special reference to the multivariate normal, see Dempster (1971). Patil (1966) has worked on a sub-family of the exponential family which he called multivariate, generalized, power-series distributions.

The log-linear form (1) and the constraints (2) are related to each other first through a property of discriminatory information and second, through properties of maximum-likelihood estimation which will be described in Section 3.

The *discriminatory information function* is defined as

$$K[\mathbf{p}, \boldsymbol{\pi}] = \sum_{i \in I} p_i \log \frac{p_i}{\pi_i},$$

where $0 \log 0 = 0$ by definition, and is subject to the basic

LEMMA 1.

$$K[\mathbf{p}, \boldsymbol{\pi}] \geqq 0 \quad \text{and} \quad K[\mathbf{p}, \boldsymbol{\pi}] = 0 \quad \text{if and only if} \quad \mathbf{p} = \boldsymbol{\pi}.$$

PROOF. The proof of this result is well known (see eg. Kullback (1959)) for the case $\sum_{i \in I} \pi_i = 1$. If $\sum_{i \in I} \pi_i < 1$ define $\sigma = \sum_{i \in I} \pi$. Then

$$K(\mathbf{p}, \boldsymbol{\pi}) = K\left(\mathbf{p}, \frac{1}{\sigma} \boldsymbol{\pi}\right) - \log \sigma$$

which is clearly nonnegative and equal to zero if and only if $\mathbf{p} = \sigma^{-1}\boldsymbol{\pi}$ and $\sigma = 1$, i.e. $\mathbf{p} = \boldsymbol{\pi}$. $\square$

The property connecting (1) and (2) is given in

LEMMA 2. *If a positive, probability function* $\mathbf{p}$ *of the form* (1) *satisfying* (2) *exists then it minimizes* $K(\mathbf{p}, \boldsymbol{\pi})$ *subject to* (2) *and is unique in doing so.*

PROOF. Let $\mathbf{q}$ be any probability function satisfying (2). Then

$$
\begin{aligned}
K(\mathbf{p}, \boldsymbol{\pi}) &= \sum_{i \in I} p_i[\log \mu + \sum_{s=1}^d b_{si} \log \mu_s] \\
&= \log \mu[\sum_{i \in I} p_i] + \sum_{s=1}^d \log \mu_s[\sum_{i \in I} b_{si} p_i] \\
&= \log \mu[\sum_{i \in I} q_i] + \sum_{s=1}^d \log \mu_s[\sum_{i \in I} b_{si} q_i] \\
&= \sum_{i \in I} q_i \log \frac{p_i}{\pi_i}.
\end{aligned}
$$

Hence

$$K(\mathbf{q}, \boldsymbol{\pi}) - K(\mathbf{p}, \boldsymbol{\pi}) = K(\mathbf{q}, \mathbf{p})$$

which is nonnegative and equal to zero if and only if $\mathbf{q} = \mathbf{p}$. $\square$

Kullback and Khairat (1966) provide a much more general version of Lemma 2. Note that Lemma 2 is almost the same as the following result (see Good (1963)) concerning the maximization of entropy.

LEMMA 3. *If there exists a positive probability function of the form*

$$p_i = \mu \prod_{s=1}^d \mu_s^{b_{si}}$$

satisfying (2) *then it maximizes the entropy* $H(\mathbf{p}) = -\sum_{i \in I} p_i \log p_i$ *and is unique in doing so.*

PROOF. Consider $H(\mathbf{p}) - H(\mathbf{q})$ and proceed in the same way as for Lemma 2. $\square$

DEFINITION. The constraints (2) are *consistent* if $\{\mathbf{q}; q_i > 0, \mathbf{q}$ satisfies (2)$\}$ is nonempty.

As a last preliminary result before discussing iterative scaling we express (1) and (2) in a more convenient form.

LEMMA 4. *Given that the constraints* (2) *are consistent, then* (1) *and* (2) *are expressible as*

(1')
$$p_i = \pi_i \prod_{r=1}^{c} \lambda_r^{a_{ri}}$$

(2')
$$\sum_{i \in I} a_{ri} p_i = h_r , \qquad\qquad r = 1, 2, \cdots, c$$

where
$$a_{ri} \geqq 0 , \qquad \sum_{r=1}^{c} a_{ri} = 1 , \qquad h_r > 0 , \qquad \sum_{r=1}^{c} h_r = 1 .$$

PROOF. Define
$$a_{si} = t_s(u_s + b_{si}) , \quad \forall i , \qquad h_s = t_s(u_s + k_s) , \quad s = 1, 2, \cdots, d$$

where $u_s \geqq 0$, $t_s > 0$ are chosen to make

$$a_{si} \geqq 0 \quad \forall i \in I \qquad \text{and} \qquad \sum_{s=1}^{d} a_{si} \leqq 1 .$$

If $\sum_{s=1}^{d} a_{si} = 1 \; \forall i$, define $c = d$. Otherwise define $c = d + 1$ and let

$$a_{ci} = 1 - \sum_{s=1}^{d} a_{si} , \qquad h_c = 1 - \sum_{s=1}^{d} h_s .$$

With these definitions of $\{a_{ri}; r = 1, 2, \cdots, c, i \in I\}$ and of $\{h_r = 1, 2, \cdots, c\}$ it is clear that the constraints (2') are equivalent to (2). By the assumption that $\forall s, s = 1, 2, \cdots, d$ not all $b_{si} = 0$, $i \in I$, it follows that $\forall r, r = 1, 2, \cdots, c$, at least one $a_{ri} > 0$. Therefore, by the assumption that the constraints (2) are consistent, it follows that $h_r > 0$.

To put (1) in the form (1'), define

$$\lambda_s = \nu \mu_s^{1/t_s} , \qquad\qquad s = 1, 2, \cdots, d , \quad \lambda_c = \nu$$

where
$$\nu = \mu \prod_{s=1}^{d} \mu_s^{-u_s} . \qquad\qquad\qquad \square$$

The main results about generalized iterative scaling are given in Section 2. Section 3 is concerned with maximum likelihood estimation. In Section 4 the relationship of this paper to the existing literature is discussed and possible applications to new problems are suggested.

2. Generalized iterative scaling.

THEOREM 1. *Consider the sequence* $\{\mathbf{p}^{(n)}; n = 0, 1, 2, \cdots\}$ *where* $\mathbf{p}^{(n)} = \{p_i^{(n)}; i \in I\}$ *defined by*

$$p_i^{(0)} = \pi_i$$
$$p_i^{(n+1)} = p_i^{(n)} \prod_{r=1}^{c} \left(\frac{h_r}{h_r^{(n)}}\right)^{a_{ri}} , \qquad\qquad n = 0, 1, 2, \cdots$$

where
$$h_r^{(n)} = \sum_{i \in I} a_{ri} p_i^{(n)} .$$

Provided that the constraints (2') are consistent the sequence $\{\mathbf{p}^{(n)}\}$ converges to a solution of (1') which is unique and positive.

PROOF. We begin by proving that $\sum_{i \in I} p_i^{(n)} \leqq 1$ for all $n = 1, 2, \cdots$. Now, for $n \geqq 1$,

$$\sum_{i \in I} p_i^{(n)} = \sum_{i \in I} p_i^{(n-1)} \prod_{r=1}^{c} \left(\frac{h_r}{h_r^{(n-1)}} \right)^{a_{ri}},$$

and, by the inequality between the generalized arithmetic and geometric means (or, equivalently, by the convexity of the logarithmic function),

$$(3) \qquad \prod_{r=1}^{c} \left(\frac{h_r}{h_r^{(n-1)}} \right)^{a_{ri}} \leqq \sum_{r=1}^{c} a_{ri} \left(\frac{h_r}{h_r^{(n-1)}} \right).$$

Hence,

$$\sum_{i \in I} p_i^{(n)} \leqq \sum_{r=1}^{c} \frac{h_r}{h_r^{(n-1)}} \sum_{i \in I} a_{ri} p_i^{(n-1)} = \sum_{r=1}^{c} \frac{h_r}{h_r^{(n-1)}} h_r^{(n-1)}.$$

Since $\sum_{i \in I} p_i^{(n)} \leqq 1$, it follows that $\sum_{r=1}^{c} h_r^{(n)} \leqq 1$. Also, it is clear that, for all n, $p_i^{(n)} > 0$ and therefore $h_r^{(n)} > 0$. Thus, by Lemma 1,

$$K(\mathbf{h}, \mathbf{h}^{(n)}) = \sum_{r=1}^{c} h_r \log \frac{h_r}{h_r^{(n)}} \geqq 0.$$

Now let $\mathbf{q}$ denote any positive probability function satisfying (2'). Such a $\mathbf{q}$ exists by the assumption of the consistency of (2'). Then

$$K(\mathbf{q}, \mathbf{p}^{(n+1)}) = K(\mathbf{q}, \mathbf{p}^{(n)}) - \sum_{i \in I} q_i \log \prod_{r=1}^{c} \left[\frac{h_r}{h_r^{(n)}} \right]^{a_{ri}}$$

$$= K(\mathbf{q}, \mathbf{p}^{(n)}) - \sum_{r=1}^{c} \log \frac{h_r}{h_r^{(n)}} \sum_{i \in I} a_{ri} q_i$$

$$= K(\mathbf{q}, \mathbf{p}^{(n)}) - K(\mathbf{h}, \mathbf{h}^{(n)}).$$

It follows that the sequence

$$\{K(\mathbf{q}, \mathbf{p}^{(n)}); n = 1, 2, \cdots\}$$

is decreasing and, by Lemma 1 is bounded below by zero. Therefore it has a limit and this implies that

$$K(\mathbf{h}, \mathbf{h}^{(n)}) \to 0 \qquad\qquad \text{as} \quad n \to \infty.$$

But

$$K(\mathbf{h}, \mathbf{h}^{(n)}) \geqq \tfrac{1}{2} \sum_{r=1}^{c} h_r (h_r - h_r^{(n)})^2$$

(see eg. Rao (1965)) and hence

$$(4) \qquad\qquad \mathbf{h}^{(n)} \to \mathbf{h}$$

because $h_r > 0$, $r = 1, 2, \cdots, c$.

Since $\{\mathbf{p}^{(n)}\}$ is a bounded sequence it possesses at least one limit point. Let $\mathbf{p}'$, $\mathbf{p}''$ both be limit points. We know, from (4), that both satisfy the constraints (2'). Also both are of the form (1') since every $\mathbf{p}^{(n)}$ is of this form and $\mathbf{p}'$, $\mathbf{p}''$

are each the limits of convergent subsequences of $\{\mathbf{p}^{(n)}\}$. Further, $\mathbf{p}'$ and $\mathbf{p}''$ are both positive probability functions because, if this were not so, $\{K(\mathbf{q}, \mathbf{p}^{(n)})\}$ would have to diverge to $+\infty$ whereas it converges downwards to a finite limit. Finally, by Lemma 2, there is at most one positive $\mathbf{p}$ satisfying $(1')$ and $(2')$ and therefore $\{\mathbf{p}^{(n)}\}$ converges to it. $\square$

COROLLARY 1. *If the constraints* (2) *are consistent there exists a* (*unique, positive*) *solution to* (1) *and* (2).

PROOF. This result is part of Theorem 1, but is separated for emphasis. $\square$

COROLLARY 2. *Consider k sets of constraints each of the form* $(2')$. *Let the mth set,* $m = 0, 1, \cdots, k - 1$, *be written*

(5) $$\sum_{i \in I} a_{ri}^{(m)} p_i = h_r^{(m)} , \qquad\qquad r = 1, 2, \cdots, c^{(m)}$$

where

$$\sum_{r=1}^{c^{(m)}} a_{ri}^{(m)} = 1 , \qquad \sum_{r=1}^{c^{(m)}} h_r^{(m)} = 1 ,$$

and

$$a_{ri}^{(m)} \geqq 0 , \qquad h_r^{(m)} > 0 .$$

Then, provided that the constraints (5) *are consistent, there exists a unique positive $\mathbf{p}$ which satisfies them and is of the form*

(6) $$p_i = \pi_i \prod_{m=0}^{k-1} \prod_{r=1}^{c^{(m)}} [\lambda_r^{(m)}]^{a_{ri}^{(m)}} .$$

Moreover $\mathbf{p}$ can be obtained as the limit of a "cyclical" iterative scaling process. Precisely, $\mathbf{p}$ is the limit of the sequence $\{\mathbf{p}^{(n)}\}$ defined by

$$p_i^{(0)} = \pi_i$$

and

$$p_i^{(n+1)} = p_i^{(n)} \prod_{r=1}^{c^{(n_k)}} \left(\frac{h_r^{(n_k)}}{h_r^{(n_k, n)}} \right)^{a_{ri}^{(n_k)}}$$

where n_k denotes the remainder after dividing n by k and

$$h_r^{(m, n)} = \sum_{i \in I} a_{ri}^{(m)} p_i^{(n)} .$$

PROOF. The proof is easily obtained from the proof of Theorem 1. $\square$

The class of iterative scaling processes previously discussed in the literature (see Section 4) is mainly distinguished by the fact that

$$a_{ri}^{(m)} = 0, 1 .$$

In this case it is appropriate to use a different notation and Corollary 3 is a re-description of Corollary 2 in this notation.

COROLLARY 3. *Consider k partitions of I and denote the mth, $m = 0, 1, \cdots, k - 1$, by*

$$I_1^{(m)}, I_2^{(m)}, \cdots, I_{c^{(m)}}^{(m)}$$

where each subset is nonempty. Then provided the constraints

$$\sum_{i \in I_r^{(m)}} p_i = h_r^{(m)} , \quad h_r^{(m)} > 0 , \quad r = 1, 2, \cdots, c^{(m)} , \quad m = 0, 1, \cdots, k - 1$$

are consistent, there is a unique $\mathbf{p}$ *satisfying them of the form*

(7)
$$p_i = \pi_i \prod_{m=0}^{k-1} \lambda_{r_i^{(m)}}^{(m)}$$

where $r_i^{(m)}$ *is the value of* r *for which* $i \in I_r^{(m)}$. *The iterative scaling process is defined by*

$$p_i^{(n+1)} = p_i^{(n)} \frac{h_{r_i(n_k)}^{(n_k)}}{h_{r_i(n_k)}^{(n_k,n)}}$$

where

$$h_r^{(m,n)} = \sum_{i \in I_r^{(m)}} p_i^{(n)} \, .$$

PROOF. To see that Corollary 3 is a special case of Corollary 2 define

$$\begin{aligned} a_{ri}^{(m)} &= 1 \quad && \text{if } i \in I_r^{(m)} \\ &= 0 \quad && \text{if } i \notin I_r^{(m)} \, . \end{aligned} \qquad \square$$

Here we note two points concerning the relationship of Corollaries 2 and 3 to the parent theorem. The first is that, when $a_{ri} = 0, 1$, inequality (3) in the proof of Theorem 1 reduces to a trivial equality. The second is that, while it is clear that the structure of iterative scaling is essentially revealed when $k = 1$ (the proof of Corollary 2 is little more than the proof of Theorem 1) it is unlikely that this observation would have been made in the context of $a_{ri} = 0, 1$. For, when $a_{ri} = 0, 1$ and $k = 1$, the problem of solving (1′) and (2′) is trivial and does not require iterative scaling. Thus, if the constraints are written

$$\sum_{i \in I_r} p_i = h_r \, , \qquad\qquad r = 1, 2, \cdots, c$$

where $I_1, I_2, \cdots, I_c$ is a partition of I, then the solution of (1′) is $p_i = h_{r_i}/|I_{r_i}|$, where r_i is the value of r for which $i \in I_r$, and $|I_{r_i}|$ is the number of elements in I_{r_i}.

COROLLARY 4. *The iterative process in Corollary 2 may be commenced with* $\mathbf{p}^{(0)}$ *of the form*

(8)
$$p_i^{(0)} = \pi_i \prod_{m=0}^{k-1} \prod_{r=1}^{c^{(m)}} [\alpha_r^{(m)}]^{a_{ri}^{(m)}} \, .$$

Further, there is no requirement that $\sum_{i \in I} p_i^{(0)} \leqq 1$.

PROOF. This is immediate. $\square$

We note that all of the above theory is trivially adaptable to the case when there is no factor π_i in (6). The general $p_i^{(0)}$ is of the form (8) with π_i replaced by 1. However we comment that the simplest and most natural choice is for all the $p_i^{(0)}$ to be equal. The actual common value taken is immaterial since it does not affect $p_i^{(1)}$; for it is clearly true that if $\mathbf{p}^{(n+1)}(\mathbf{p}^{(n)})$ denotes $\mathbf{p}^{(n+1)}$ expressed as a function of $\mathbf{p}^{(n)}$ then

$$\mathbf{p}^{(n+1)}(\theta \mathbf{p}^{(n)}) = \mathbf{p}^{(n+1)}(\mathbf{p}^{(n)}) \, .$$

Finally, we comment that the iterative scaling process produces the probabilities p_i but that, if it is desired to find the parameters $\{\lambda_r\}$, this can be done

in two ways. The first is simply to compile λ_r as a product of terms $h_r/h_r^{(n)}$. The particular solution obtained in this way will of course depend on the initial approximation $\mathbf{p}^{(0)}$. Alternatively, we can use the fact that $\{\log p_i; i \in I\}$ is a linear combination of known vectors $\{a_{ri}; i \in I\}, r = 1, 2, \cdots, c$, with coefficients $\{\log \lambda_r\}$. Therefore the parameters $\{\log \lambda_r\}$ can be found by solving c linear equations.

3. **Application to maximum-likelihood estimation.** The practical problem of solving (1) and (2) arises when estimating probability functions rather than finding true ones. The appropriate theory to cover Kullback's (1959) minimum discriminatory information method of estimation is given in Lemma 2 and Corollary 1 to Theorem 1. The theory relevant to maximum likelihood estimation is given in

LEMMA 4. *Let I be a finite set and let $\{p_i\}$ be an unknown probability function on I of the form* (1), *namely*

$$p_i = \pi_i \mu \prod_{s=1}^{d} \mu_s^{b_{si}}$$

where $\{\pi_i\}$ is known, $\{b_{si}\}$ is known but μ and $\{\mu_s\}$ are unknown. Given a sample of size N with positive frequencies, $\{f_i; i \in I, f_i > 0, \sum_{i \in I} f_i = N\}$, the maximum likelihood estimate $\{\hat{p}_i\}$ of $\{p_i\}$ is the unique probability function of the form (1) *which satisfies*

$$(9) \qquad \sum_{i \in I} b_{si} p_i = \sum_{i \in I} b_{si} \frac{f_i}{N}, \qquad s = 1, 2, \cdots, d, \qquad \sum_{i \in I} p_i = 1.$$

Therefore $\{\hat{p}_i\}$ can be obtained by iterative scaling.

PROOF. The constraints (9) are consistent because they are satisfied when $p_i = f_i/N$. Therefore there exists a unique $\{\hat{p}_i\}$ of the form

$$\hat{p}_i = \pi_i \hat{\mu} \prod_{s=1}^{d} \hat{\mu}_s^{b_{si}}$$

which satisfies them and which can be obtained by iterative scaling. To show that $\{\hat{p}_i\}$ uniquely maximizes the log-likelihood $\sum_{i \in I} f_i \log p_i$, consider

$$\sum_{i \in I} f_i \log \hat{p}_i - \sum_{i \in I} f_i \log p_i = \sum_{i \in I} f_i \log \frac{\hat{p}_i}{p_i}$$

$$= N \sum_{i \in I} \frac{f_i}{N} \left[\log \frac{\hat{\mu}}{\mu} + \sum_{s=1}^{d} b_{si} \log \frac{\hat{\mu}_s}{\mu_s} \right]$$

$$= N \sum_{i \in I} \hat{p}_i \left[\log \frac{\hat{\mu}}{\mu} + \sum_{s=1}^{d} b_{si} \log \frac{\hat{\mu}_s}{\mu_s} \right] \qquad \text{by (9)}$$

$$= N \sum_{i \in I} \hat{p}_i \log \frac{\hat{p}_i}{p_i}$$

which, by Lemma 1, is nonnegative and equal to zero if and only if $p_i = \hat{p}_i$, $\forall i \in I$. □

Thus, whereas the property of discriminatory information given in Lemma 2

takes us from the linear constraints to the log-linear model, the property of maximum likelihood given in Lemma 4 takes us in the opposite direction. Campbell (1970) proved a converse to Lemma 4 which leads from the constraints (9) to the log-linear form. Before stating Campbell's theorem in the present context of finite I note that, if the random variable T_s is defined by $T_s(i) = b_{si}$, then (9) may be written

$$(9') \qquad\qquad E[T_s] = \frac{1}{N} \sum_{j=1}^{N} t_s^{(j)}, \qquad\qquad s = 1, 2, \cdots, d,$$

where $\{t_s^{(1)}, t_s^{(2)}, \cdots, t_s^{(N)}\}$ is the sample of N observed values of T_s.

CAMPBELL'S THEOREM. *Choose* **p** *according to "Gauss's principle": that maximum-likelihood estimation is to be equivalent to first-moment estimation given by* (9'). *Then* **p** *is necessarily of log-linear form* (1).

4. Discussion. Kullback (1971) has recently applied the theory of this paper to a problem in which some of the $a_{ri}^{(m)}$ are $\frac{1}{2}$. Ireland, Ku and Kullback (1969) applied iterative scaling to tests of marginal homogeneity in $r \times r$ contingency tables for which some of the $b_{si}^{(m)}$ are -1. Apart from this the literature on the theory and application of iterative scaling relates to probability functions of the form (7) with $a_{ri}^{(m)} = 0, 1$, viz.

$$p_i = \pi_i \prod_{m=1}^{k} \lambda_{r_i(m)}^{(m)}$$

which, with its associated constraints, was the subject matter of Corollary 3. Moreover, virtually all of the applications relate in some way to contingency tables.

Ireland and Kullback (1968) gave a full discussion of iterative scaling and they chose, for convenience, the following context. Find a table P of probabilities which has prescribed marginal totals and which is closest to the table II in the sense of minimizing discriminatory information. Here

$$I = \{i = (i_1, i_2); i_1 = 1, 2, \cdots, c^{(1)}, i_2 = 1, 2, \cdots, c^{(2)}\}$$

and, by Lemma 2,

$$(10) \qquad\qquad p_i = \pi_i \lambda_{i_1}^{(1)} \lambda_{i_2}^{(2)}.$$

The constraints are

$$\sum_{i \in I_r^{(1)}} p_i = h_r^{(1)}, \quad r = 1, 2, \cdots, c^{(1)}, \qquad \sum_{i \in I_r^{(2)}} p_i = h_r^{(2)}, \quad r = 1, 2, \cdots, c^{(2)},$$

where

$$I_r^{(1)} = \{i = (i_1, i_2); i_1 = r\}, \qquad I_r^{(2)} = \{i = (i_1, i_2); i_2 = r\}.$$

Ireland and Kullback gave a proof of convergence which is based on the incomplete proof given by Brown (1959). Apart from the generalization to $a_{ri}^{(m)} \neq 0, 1$, our proof is essentially the same as Ireland and Kullback's except that it is shorter in the final stages.

Fienberg (1970) gave a geometric proof of convergence applicable to contingency tables. Also he provided a full bibliography of the literature to date on iterative scaling.

Darroch (1962) applied iterative scaling to finding the "no three-factor inter-action" probabilities in a three-factor contingency table given the two-factor marginal totals. Here

$$I = \{i = (i_1, i_2, i_3); i_j = 1, 2, \cdots, t_j, j = 1, 2, 3\}$$

and

$$p_i = \lambda^{(1)}_{i_2 i_3} \lambda^{(2)}_{i_1 i_3} \lambda^{(3)}_{i_1 i_2},$$

which is uniquely determined by constraints of the form

$$\sum^{t_1}_{i_1=1} p_i = h^{(1)}_{i_2 i_3}, \qquad \sum^{t_2}_{i_2=1} p_i = h^{(2)}_{i_1 i_3}, \qquad \sum^{t_3}_{i_3=1} p_i = h^{(3)}_{i_1 i_2}.$$

A new application of the $a^{(m)}_{ri} = 0, 1$ type is to the calculation of "F-inde-pendent" probabilities, discussed in Darroch and Ratcliff (1972). In the bivariate case the F-independent probability function $\mathbf{p}$ is defined on the set

$$I = \{i = (i_1, i_2); i_1 \geqq 0, i_2 \geqq 0, i_1 + i_2 \leqq n\}$$

by

$$p_i = \lambda^{(1)}_{i_1} \lambda^{(2)}_{i_2} \lambda^{(3)}_{n-i_1-i_2},$$

and is uniquely determined by constraints of the form

$$\sum_{i:i_1=r} p_i = h_r^{(1)}, \qquad \sum_{i:i_2=r} p_i = h_r^{(2)}, \qquad \sum_{i:i_1+i_2=n-r} p_i = h_r^{(3)}.$$

Now we consider some possible applications of iterative scaling to problems in which $0 \leqq a^{(m)}_{ri} \leqq 1$ or, in terms of (1) and (2) in which the b_{si} are arbitrary real numbers. First, we note that any of the above-mentioned applications in-volving k sets of constraints can be converted into a problem involving one set of constraints by multiplying each constraint of the mth set by θ_m, $\sum^k_{m=1} \theta_m = 1$. Thus, for the no three-factor interaction probability function, the three sets of constraints may be replaced by one set in which they are all multiplied by $\frac{1}{3}$. This offers an alternative, noncyclic, iterative process for solving the problem. However, we do not wish to advocate it as numerical evidence strongly indicates that it is inferior to the cyclic process.

As a second field of application, we consider truncated probability distribu-tions. An example is the truncated multinomial which may be expressed in the form (1) by defining

(11) $$p_i = \pi_i \mu \prod^d_{s=1} \mu_s^{i_s}, \qquad\qquad\qquad i \in I$$

where

$$i = (i_1, i_2, \cdots, i_d), \qquad \sum^d_{s=1} i_s = n, \qquad \pi_i = \frac{1}{d^n} \frac{n!}{\prod^d_{s=1} i_s!}$$

and where I is a subset of $\{i; i_s \geqq 0, \sum^d_{s=1} i_s = n\}$. Consider the problem of finding the maximum likelihood estimates of μ, $\{\mu_s\}$, that is, of solving (11) subject to

(12) $$\sum_{i \in I} i_s p_i = \sum_{i \in I} i_s \frac{f_i}{N}, \qquad s = 1, 2, \cdots, d, \qquad \sum_{i \in I} p_i = 1.$$

Of course, if the subset I is such that the "normalizing parameter" μ is a function of $\mu_1, \mu_2, \cdots, \mu_s$ with manageable derivatives, the usual method of maximizing the likelihood by differentiation can be used. (See eg. Patil (1962).) The advantage of the iterative-scaling method of solving (11) and (12) is that it does not require any knowledge of the function $\mu = \mu(\mu_1, \mu_2, \cdots, \mu_d)$ and therefore is insensitive to the nature of the truncation subset I.

Finally, in addition to the estimation of truncated versions of standard log-linear distributions, iterative scaling opens the way to using new distributions in product form, all of which may be generated by the "principle of minimum discrimination information." (See Good (1963) for a discussion of the almost equivalent "principle of maximum entropy.") The list of standard distributions in product form is a very short one, primarily because it is implicitly required that μ be a simple function of $\mu_1, \mu_2, \cdots, \mu_d$. However, this property is unnecessary for us to understand the shape of the probability function and we now see that it is unnecessary for maximum likelihood estimation of the unknown parameters.

REFERENCES

BROWN, D. T. (1959). A note on approximations to discrete probability distributions. *Information and Control* **2** 386–392.

BISHOP, Yvonne, M. M. (1969). Full contingency tables, logits, and split contingency tables. *Biometrika* **25** 383–339.

CAMPBELL, L. L. (1970). Equivalence of Gauss's principle and minimum discrimination information estimation of probabilities. *Ann. Math. Statist.* **41** 1011–1015.

DARROCH, J. N. (1962). Interaction in multifactor contingency tables. *J. Roy. Statist. Soc., Ser. B* **24** 251–263.

DARROCH, J. N. and RATCLIFF, D. (1972). Test for F-independence, with application to Waite's finger-print data. In preparation.

DEMPSTER, A. P. (1971). Covariance selection. Research report 5–12, Harvard Univ.

FIENBERG, S. E. (1970). An iterative procedure for estimation in contingency tables. *Ann. Math. Statist.* **41** 907–915.

GOKHALE, D. V. (1971). An iterative procedure for analysing log-linear models. *Biometrics* **27** 681–687.

GOOD, I. J. (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Ann. Math. Statist.* **34** 911–934.

IRELAND, C. T. and KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika* **55** 179–188.

IRELAND, C. T., KU, H. H., and KULLBACK, S. (1969). Symmetry and marginal homogeneity of an $r \times r$ contingency table. *J. Amer Statist. Assoc.* **64** 1323–1341.

KULLBACK, S. (1959). *Information Theory and Statistics.* Wiley, New York.

KULLBACK, S. (1971). Estimating and testing interaction parameters in the log-linear model. Unpublished typescript.

PATIL, G. P. (1962). Maximum likelihood estimation for generalized power series distributions and its application to a truncated binomial distribution. *Biometrika* **49** 227–237.

PATIL, G. P. (1966). On multivariate generalized power series distribution and its application to the multinomial and negative multinomial. *Sankhyā Ser A.* **28** 225–238.

RAO, C. R. (1965). *Linear Statistical Inference and its Applications.* Wiley, New York.

SCHOOL OF MATHEMATICAL SCIENCES
FLINDERS UNIVERSITY OF S.A.
BEDFORD PARK 5042 (S.A.)
AUSTRALIA