Ann. Inst. Statist. Math.
Vol. 55, No. 2, 251–263 (2003)

I-PROJECTION ONTO ISOTONIC CONES AND ITS APPLICATIONS TO MAXIMUM LIKELIHOOD ESTIMATION FOR LOG-LINEAR MODELS

WEI GAO[1,2] AND NING-ZHONG SHI[1]

[1] *Department of Mathematics, Northeast Normal University, Changchun 130024, China*
[2] *The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan*

(Received November 12, 2001; revised June 5, 2002)

Abstract. A frequently occurring problem is to find a probability vector, $p \in D$, which minimizes the I-divergence between it and a given probability vector π. This is referred to as the I-projection of π onto D. Darroch and Ratcliff (1972, *Ann. Math. Statist.*, **43**, 1470–1480) gave an algorithm when D is defined by some linear equalities and in this paper, for simplicity of exposition, we propose an iterative procedure when D is defined by some linear inequalities. We also discuss the relationship between I-projection and the maximum likelihood estimation for multinomial distribution. All of the results can be applied to isotonic cone.

Key words and phrases: I-divergence, I-projection, isotonic cone, log-linear models.

1. Introduction

Let $p = (p_1, \ldots, p_k)^T$ and $q = (q_1, \ldots, q_k)^T$ be the probability vectors and I-divergence of p with respect to q, also called the Kullback-Liebler information number, cross-entropy between p and q, information for discrimination, entropy of p relative to q, is given by

$$I(p \mid q) = \sum_{i=1}^{k} p_i \log(p_i/q_i).$$

It is well known that $I(p \mid q)$ has the following results. (See Kullback (1967), Kemperman (1969) and Csiszar (1967).)

LEMMA 1.1. *For $p, q \in P$ (P denote the set of probability vector), then*

$$I(p \mid q) \geq 0, \qquad [2I(p \mid q)]^{1/2} \geq \sum_{i=1}^{k} |p_i - q_i|. \tag{1.1}$$

Thus it is heuristically reasonable to think of $I(p \mid q)$ as representing a "distance" between p and q. If we interpret $I(p \mid q)$ as distance, then the I-projection of the probability vector π onto a set $D \subset P$ is defined as $\hat{\pi} \in D$ such that

$$I(\hat{\pi} \mid \pi) = \min_{p \in D} I(p \mid \pi). \tag{1.2}$$

Minimization problems of the forms (1.2) play a key role in the information theory (Kullback (1959), Good (1963)) and also in statistics for the maximum likelihood estimation (Csiszar (1967, 1975), Agresti (1984), Lemke and Dykstra (1984) and Dykstra (1985)).

Darroch and Ratcliff (1972) considered the problem of I-projection when $p \in D$ satisfies

$$\sum_{i=1}^{k} a_{ij} p_i = h_j, \qquad j = 1, \ldots, t$$

where $a_{ij} (i = 1, \ldots, k; j = 1, \ldots, t)$ and $h_j (j = 1, \ldots, t)$ are given constants. They proposed an algorithm for problem (1.2) and also discussed the relationship between (1.2) and the maximum likelihood estimation of p for multinomial distribution when p belongs to the log-linear models, that is,

$$p_i = \lambda \prod_{j=1}^{t} \lambda_j^{a_{ij}} \tag{1.3}$$

where λ and $\lambda_j (j = 1, \ldots, t)$ are parameters.

When D is defined by some linear inequalities, then for given π, the I-projection of π onto D is important not only in information theory but also in statistics for the maximum likelihood estimation which will be described in Section 3. For example, for some given s, $p \in D$ is defined by

$$\sum_{j=1}^{l} \sum_{i=1}^{k} a_{ij} p_i \le \sum_{j=1}^{l} h_j, \qquad \sum_{j=1}^{s} \sum_{i=1}^{k} a_{ij} p_i = \sum_{j=1}^{s} h_j, \qquad \sum_{i=1}^{k} a_{iu} p_i = h_u \tag{1.4}$$

for $l = 1, \ldots, s-1$; $u = s+1, \ldots, t$.

The relationship between the form (1.3) and the constraints (1.4) may be given in

LEMMA 1.2. *If the probability vector $\bar{p}$ satisfying*

$$\bar{p}_i = \pi_i \bar{\lambda} \prod_{j=1}^{t} \bar{\lambda}_j^{a_{ij}}, \qquad \bar{\lambda}_1 \le \cdots \le \bar{\lambda}_s, \tag{1.5}$$

$$\sum_{j=1}^{l} \sum_{i=1}^{k} a_{ij} \bar{p}_i \le \sum_{j=1}^{l} h_j, \qquad \sum_{j=1}^{s} \sum_{i=1}^{k} a_{ij} \bar{p}_i = \sum_{j=1}^{s} h_j, \qquad \sum_{i=1}^{k} a_{iu} \bar{p}_i = h_u, \tag{1.6}$$

$$\sum_{j=1}^{s} \sum_{i=1}^{k} a_{ij} \bar{p}_i \log \bar{\lambda}_j = \sum_{j=1}^{s} h_j \log \bar{\lambda}_j \tag{1.7}$$

exists $(l = 1, \ldots, s;\ u = s+1, \ldots, t)$, then it minimizes $I(p \mid \pi)$ subject to (1.4) and is unique in doing so.

PROOF. Let p be any probability vector satisfying (1.4), $H_j = \sum_{l=1}^{j} h_l$ and $G_j = \sum_{l=1}^{j} \sum_{i=1}^{k} a_{il} p_i$ for $j = 1, \ldots, t$, then

$$G_1 \le H_1, \ldots, G_{s-1} \le H_{s-1}, G_s = H_s, \ldots, G_t = H_t.$$

By the Abel's transformation, for $x_1 \le \cdots \le x_t$, one has

$$\sum_{j=1}^{t} h_j x_j - \sum_{j=1}^{t}\sum_{i=1}^{k} a_{ij} p_i x_j = \sum_{j=1}^{s-1} (H_j - G_j)(x_i - x_{i+1}) \le 0.$$

So

$$\begin{aligned} I(\bar{p} \mid \pi) &= \log \bar{\lambda} + \sum_{j=1}^{t}\sum_{i=1}^{k} a_{ij}\bar{p}_j \log \bar{\lambda}_j = \log \bar{\lambda} + \sum_{j=1}^{t} h_j \log \bar{\lambda}_j \\ &\le \log \bar{\lambda} + \sum_{j=1}^{t}\sum_{i=1}^{k} a_{ij} p_i \log \bar{\lambda}_j = \sum_{i=1}^{k} p_i \log(\bar{p}_i/\pi_i), \end{aligned}$$

and $I(p \mid \pi) - I(\bar{p} \mid \pi) \ge \sum_{i=1}^{k} p_i \log(p_i/\pi_i) - \sum_{i=1}^{k} p_i \log(\bar{p}_i/\pi_i) = I(p|\bar{p}) \ge 0$.

Suppose that there exist two probability vector $\bar{p}$ and $\tilde{p}$ satisfying (1.5)–(1.7). From the above proof, we have $I(\bar{p} \mid \pi) = I(\tilde{p} \mid \pi)$ and $I(\bar{p} \mid \pi) - I(\tilde{p} \mid \pi) = I(\bar{p} \mid \tilde{p}) = 0$ which implies $\bar{p} = \tilde{p}$ by Lemma 1.1.

Remark 1. By similar proof, (1.6) and (1.7) are equivalent to

$$\sum_{j=1}^{s}\sum_{i=1}^{k} a_{ij}\bar{p}_i\mu_j \ge \sum_{j=1}^{s} h_j\mu_j, \quad \sum_{i=1}^{k} a_{iu}\bar{p}_i = h_u, \quad \mu_1 \le \cdots \le \mu_s, \tag{1.6'}$$

$$\text{for} \quad u = s+1, \ldots, t,$$

$$\sum_{j=r}^{v}\sum_{i=1}^{k} a_{ij}\bar{p}_i = \sum_{j=r}^{v} h_j, \quad \bar{\lambda}_{r-1} < \bar{\lambda}_r = \cdots = \bar{\lambda}_v < \bar{\lambda}_{v+1}. \tag{1.7'}$$

LEMMA 1.3. (1.5), (1.6) *and* (1.7) *can be expressed as*

$$\bar{p}_i = \pi_i \prod_{j=1}^{c} \mu_j^{b_{ij}}, \quad \mu_1 \le \cdots \le \mu_s, \tag{1.8}$$

$$\sum_{j=1}^{l}\sum_{i=1}^{k} b_{ij}\bar{p}_i \le \sum_{j=1}^{l} g_j, \quad \sum_{j=1}^{s}\sum_{i=1}^{k} b_{ij}\bar{p}_i = \sum_{j=1}^{s} g_j, \quad \sum_{i=1}^{k} b_{iu}\bar{p}_i = g_u, \tag{1.9}$$

$$\sum_{j=1}^{s}\sum_{i=1}^{k} b_{ij}\bar{p}_i \log \mu_j = \sum_{j=1}^{s} g_j \log \mu_j \tag{1.10}$$

$(l = 1, \ldots, s-1;\ u = s, \ldots, c)$ *where* $b_{ij} \ge 0$ $(i = 1, \ldots, k;\ j = 1, \ldots, c)$, $\sum_{j=1}^{c} b_{ij} = 1$, *and* $\sum_{j=1}^{c} g_j = 1$.

PROOF. Define

$$b_{ij} = e(a_{ij} + u), \quad g_j = e(h_j + u), \quad \text{for} \quad i = 1, \ldots, k; j = 1, \ldots, t$$

where $u \ge 0$, $e > 0$ are chosen to make

$$b_{ij} \ge 0, \quad \text{and} \quad \sum_{j=1}^{t} b_{ij} \le 1.$$

If $\sum_{j=1}^{t} b_{ij} = 1$ for $\forall i$, define $c = t$. Otherwise define $c = t + 1$ and let

$$b_{ic} = 1 - \sum_{j=1}^{t} b_{ij}, \qquad g_c = 1 - \sum_{j=1}^{t} g_j.$$

With these definitions of $\{b_{ij}; i = 1, \ldots, k;$ and $j = 1, \ldots, c\}$ and of $\{g_j; j = 1, \ldots, c\}$, it is clear that the constraints (1.9) are equivalent to (1.6).

To express (1.5) into the form (1.8), define

$$\mu_j = \bar{\lambda}_j^{(1/e)}\delta, \quad j = 1, \ldots, t \quad \text{and} \quad \mu_c = \delta$$

where $\delta = \bar{\lambda} / \prod_{j=1}^{t} \bar{\lambda}_j^{u}$. Also it is obvious that the constraint (1.10) is equivalent to (1.7).

In Section 2, we propose an algorithm to the problem for (1.2) under constraints (1.4). Section 3 is concerned with maximum likelihood estimation.

2. The proposed algorithm

By Lemma 1.3, without loss of generality, we suppose that a_{ij} $(i = 1, \ldots, k;\ j = 1, \ldots, t)$ satisfy

$$a_{ij} \geq 0, \qquad \sum_{j=1}^{t} a_{ij} = 1.$$

Let

$$C = \{x \in R^s; x_1 \leq x_2 \leq \cdots \leq x_s\}$$

and for the weight $w = (w_1, \ldots, w_s)'$, denote the projection of x onto C by $\hat{x} = P_w(x \mid C)$, which satisfies

$$\|x - \hat{x}\|_w^2 = \left\{ \min \sum_{j=1}^{s} (x_j - \mu_j)^2 w_j \quad \text{subject to} \quad \mu \in C \right\}.$$

It can be easily obtained (see Robertson *et al.* (1988)).

Algorithm for (1.2) *under* (1.4)
Let the initial $p_i^{(0)} = \pi_i$ and $\lambda_j^{(0)} = 1$ for $i = 1, \ldots, k$ and $j = 1, \ldots, s$.
step(n):

$$p_i^{(n)} = p_i^{(n-1)} \prod_{j=1}^{s} (\lambda_j^{(n)} / \lambda_j^{(n-1)})^{a_{ij}} \prod_{u=s+1}^{t} \left(h_u / \sum_{i=1}^{k} a_{iu} p_i^{(n-1)} \right)^{a_{iu}}$$

where $\lambda^{(n)} = P_{w^{(n-1)}}(U^{(n-1)} \mid C)$, $w_j^{(n-1)} = \sum_{i=1}^{k} a_{ij} p_i^{(n-1)}$, and $U_j^{(n-1)} = h_j \lambda_j^{(n-1)} / w_j^{(n-1)}$.

Remark 2. From the above algorithm, we can obtain the following results:

$$p_i^{(n)} = \pi_i \prod_{j=1}^{s} (\lambda_j^{(n)})^{a_{ij}} \prod_{u=s+1}^{t} \left[\prod_{l=1}^{n} h_u / \sum_{i=1}^{k} a_{iu} p_i^{(l)} \right]^{a_{iu}}, \tag{2.1}$$

$$\lambda_r^{(n)} = \cdots = \lambda_v^{(n)}, \quad \text{then} \quad \lambda_r^{(n)} = \cdots = \lambda_v^{(n)} = \sum_{j=r}^{v} h_j \lambda_j^{(n-1)} / \sum_{j=r}^{v} \sum_{i=1}^{k} a_{ij} p_i^{(n-1)}, \tag{2.2}$$

$$\sum_{j=1}^{s} h_j \lambda_j^{(n-1)} / \lambda_j^{(n)} = \sum_{j=1}^{s} \sum_{i=1}^{k} a_{ij} p_i^{(n-1)}. \tag{2.3}$$

THEOREM 2.1. *If h_j $(j = 1, \ldots, t)$ are positive, then $\{p^{(n)}\}$ given in the above proposed algorithm converges to the optimal solution of* (1.2) *under* (1.4).

The proof is given in the Appendix.

Consider m sets of constraints each of the form (1.4). Let the r-th set be written

$$\sum_{j=1}^{l} \sum_{i=1}^{k} a_{ij}^{(r)} p_i \le \sum_{j=1}^{l} h_j^{(r)}, \quad \sum_{j=1}^{s_r} \sum_{i=1}^{k} a_{ij}^{(r)} = \sum_{j=1}^{s_r} h_j^{(r)}, \quad \sum_{i=1}^{k} a_{iu} p_i = h_u^{(r)} \tag{2.4}$$

for $l = 1, \ldots, s_r - 1$; $u = s_r + 1, \ldots, t_r$; $r = 1, \ldots, m$. Then we have the following lemma.

LEMMA 2.1. *If the probability vector $\bar{p}$ satisfying*

$$\bar{p}_i = \pi_i \bar{\lambda} \prod_{r=1}^{m} \prod_{j=1}^{t_r} [\bar{\lambda}_j^{(r)}]^{a_{ij}^{(r)}}, \quad \bar{\lambda}_1^{(r)} \le \cdots \le \bar{\lambda}_{s_r}^{(r)}, \tag{2.5}$$

$$\sum_{j=1}^{l} \sum_{i=1}^{k} a_{ij}^{(r)} \bar{p}_i \le \sum_{j=1}^{l} h_j^{(r)}, \quad \sum_{j=1}^{s_r} \sum_{i=1}^{k} a_{ij}^{(r)} \bar{p}_i = \sum_{j=1}^{s_r} h_j^{(r)}, \quad \sum_{i=1}^{k} a_{iu}^{(r)} \bar{p}_i = h_u^{(r)}, \tag{2.6}$$

$$\sum_{j=1}^{s_r} \sum_{i=1}^{k} a_{ij}^{(r)} \bar{p}_i \log \bar{\lambda}_j^{(r)} = \sum_{j=1}^{s_r} h_j^{(r)} \log \bar{\lambda}_j^{(r)} \tag{2.7}$$

exists $(l = 1, \ldots, s_r - 1;\ u = s_r + 1, \ldots, t_r;\ r = 1, \ldots, m)$, *then $\bar{p}$ minimize $I(p \mid \pi)$ subject to* (2.4) *and is unique.*

PROOF. The proof is similar to Lemma 1.2.

Similar to Lemma 1.3, we have the following lemma.

LEMMA 2.2. (2.5), (2.6) *and* (2.7) *can be expressed as*

$$\bar{p}_i = \pi_i \prod_{r=1}^{m} \prod_{j=1}^{t_r'} [\mu_j^{(r)}]^{b_{ij}^{(r)}}, \quad \mu_1^{(r)} \le \cdots \le \mu_{s_r}^{(r)}, \tag{2.8}$$

$$\sum_{j=1}^{l}\sum_{i=1}^{k} b_{ij}^{(r)}\bar{p}_i \le \sum_{j=1}^{l} g_j^{(r)}, \quad \sum_{j=1}^{s_r}\sum_{i=1}^{k} b_{ij}^{(r)}\bar{p}_i = \sum_{j=1}^{s_r} g_j^{(r)}, \quad \sum_{i=1}^{k} b_{iu}^{(r)}\bar{p}_i = g_u^{(r)}, \tag{2.9}$$

$$\sum_{j=1}^{s_r}\sum_{i=1}^{k} b_{ij}^{(r)}\bar{p}_i \log \mu_j^{(r)} = \sum_{j=1}^{s_r} g_j^{(r)} \log \mu_j^{(r)} \tag{2.10}$$

where

$$b_{ij}^{(r)} \ge 0, \quad \sum_{j=1}^{t_r'} b_{ij}^{(r)} = 1.$$

Thus by Lemma 2.2, without loss of generality, we suppose

$$a_{ij}^{(r)} \ge 0, \quad \sum_{j=1}^{t_r} a_{ij}^{(r)} = 1. \tag{2.11}$$

Let

$$C_r = \{x \in R^{s_r}; x_1 \le \cdots \le x_{s_r}\}$$

and we proposed the following algorithm.

Algorithm for (1.2) *under* (2.4)

Let the initial $p_i^{(0)} = \pi_i$ and $\lambda_j^{(0,r)} = 1$ for $i = 1,\ldots,k$; $j = 1,\ldots,s_r$; $r = 1,\ldots,m$.

step$(n,1)$:

$$p_i^{(n,1)} = p_i^{(n-1,m)} \prod_{j=1}^{s_1} (\lambda_j^{(n,1)}/\lambda_j^{(n-1,1)})^{a_{ij}^{(1)}} \prod_{j=s_1+1}^{t_1} \left[h_j^{(1)} / \sum_{i=1}^{k} a_{ij}^{(1)} p_i^{(n-1,m)} \right]^{a_{ij}^{(1)}},$$

step(n,r):

$$p_i^{(n,r)} = p^{(n,r-1)} \prod_{j=1}^{s_r} (\lambda_j^{(n,r)}/\lambda_j^{(n-1,r)})^{a_{ij}^{(r)}} \prod_{j=s_r+1}^{t_r} \left[h_j^{(r)} / \sum_{i=1}^{k} a_{ij}^{(r)} p_i^{(n,r-1)} \right]^{a_{ij}^{(r)}}$$

for $r = 2,\ldots,m$ and where $\lambda^{(n,r)} = P_{w^{(n,r-1)}}(U^{(n,r-1)} \mid C_r)$, $w_l^{(n,r-1)} = \sum_{i=1}^{k} a_{il}^{(r)} p_i^{(n,r-1)}$, and $U_l^{(n,r-1)} = h_l^{(r)} \lambda_l^{(n-1,r)}$ $(l = 1,\ldots,s_r)$.

THEOREM 2.2. *If* $h_j^{(r)}$ $(j = 1,\ldots,s;\ r = 1,\ldots,m)$ *are positive, then* $\{p^{(n,r)}\}$ $(r = 1,\ldots,m)$ *given in the above algorithm converge to the optimal solution of* (1.2) *under* (2.4).

PROOF. The proof is similar to Theorem 2.1.

We have discussed the I-projection as p restricted by the form of (1.4), also can consider other form restrictions and similar results will be obtained. For example, $p \in D$ satisfies

$$\left(\sum_{i=1}^{k} a_{i1} p_i - h_1, \ldots, \sum_{i=1}^{k} a_{it} p_i - h_t \right)^T \in C^* \tag{2.12}$$

where C^* is the duality of C which is a isotonic cone (see Robertson *et al.* (1988)).

3. Application to maximum likelihood estimation

For $2 \times 2 \times k$ contingency table, let n_{ijl} be the observation and p_{ijl} be the corresponding probability, $i, j = 1, 2$; $l = 1, \ldots, k$. Then the log-likelihood function is

$$L(p) = n \sum_{l=1}^{k} \sum_{i=1}^{2} \sum_{j=1}^{2} \hat{p}_{ijl} \log p_{ijl} + c$$

where $n = \sum_{l=1}^{k} \sum_{i=1}^{2} \sum_{j=1}^{2} n_{ijl}$, $\hat{p}_{ijl} = n_{ijl}/n$ and c a constant.

Let $\phi_l = (p_{11l}p_{22l})/(p_{12l}p_{21l})$ $(l = 1, \ldots, k)$, which are usually called the local odds ratios. The maximum likelihood estimation of p_{ijl} is usually considered under

$$\phi_1 \leq \cdots \leq \phi_k. \tag{3.1}$$

See Lemke and Dykstra (1984), McDonald and Diamond (1983, 1990), and Agresti and Coull (1996).

LEMMA 3.1. *Let $\bar{p}$ be the MLE of p under* (3.1), *then $\bar{p}$ satisfies $\bar{p}_{++l} = \hat{p}_{++l}$* $(l = 1, \ldots, k)$.

PROOF. Suppose that the conclusion is not true and let $\beta_l = \hat{p}_{++l}/\bar{p}_{++l}$, $p^*_{ijl} = \beta_l \bar{p}_{ijl}$ $(i, j = 1, 2;\ l = 1, \ldots, k)$. Then

$$\begin{aligned}
\sum_{l=1}^{k} \sum_{i=1}^{2} \sum_{j=1}^{2} \hat{p}_{ijl} \log p^*_{ijl} &= \sum_{l=1}^{k} \sum_{i=1}^{2} \sum_{j=1}^{2} \hat{p}_{ijl} \log \bar{p}_{ijl} + \sum_{l=1}^{k} \hat{p}_{++l} \log \beta_l \\
&= \sum_{l=1}^{k} \sum_{i=1}^{2} \sum_{j=1}^{2} \hat{p}_{ijl} \log \bar{p}_{ijl} + \sum_{l=1}^{k} \hat{p}_{++l} \log(\hat{p}_{++l}/\bar{p}_{++l}) \\
&> \sum_{l=1}^{k} \sum_{i=1}^{2} \sum_{j=1}^{2} \hat{p}_{ijl} \log \bar{p}_{ijl}
\end{aligned}$$

which is contradictory to the fact that $\bar{p}$ is MLE and thus $\hat{p}_{++l} = \bar{p}_{++l}$, $l = 1, \ldots, k$.

We only need to consider p satisfying $p_{++l} = \hat{p}_{++l}$ $(l = 1, \ldots, k)$, thus p can be expressed as the following saturated model:

$$\log p_{11l} = \lambda_l + \lambda^{(r)}_{1l} + \lambda^{(c)}_{1l}, \qquad \log p_{12l} = \lambda_l + \lambda^{(r)}_{1l} + \lambda^{(c)}_{2l}, \tag{3.2}$$

$$\log p_{21l} = \lambda_l + \lambda^{(r)}_{2l} + \lambda^{(c)}_{1l}, \qquad \log p_{22l} = \lambda_l + \lambda^{(r)}_{2l} + \lambda^{(c)}_{2l} + \theta_l \tag{3.3}$$

where $\lambda^{(r)}_{2l} = -\lambda^{(r)}_{1l}$, $\lambda^{(c)}_{2l} = -\lambda^{(c)}_{1l}$, and $\theta_l = \log \phi_l$ for $l = 1, \ldots, k$.

LEMMA 3.2. *$\{\bar{p}_{ijl}\}$ is the MLE of $\{p_{ijl}\}$ under* (3.1) *if and only if*

$$\bar{p}_{i1l} + \bar{p}_{i2l} = \hat{p}_{i1l} + \hat{p}_{i2l}, \qquad \bar{p}_{1jl} + \bar{p}_{2jl} = \hat{p}_{1jl} + \hat{p}_{2jl}, \qquad i, j = 1, 2; l = 1, \ldots, k, \tag{3.4}$$

$$\sum_{l=1}^{s} \bar{p}_{22l} \leq \sum_{l=1}^{s} \hat{p}_{22l}, \qquad \sum_{l=1}^{k} \bar{p}_{22l} = \sum_{l=1}^{k} \hat{p}_{22l}, \qquad s = 1, \ldots, k-1, \tag{3.5}$$

$$\sum_{j=1}^{k} \bar{p}_{22j} \log \bar{\phi}_j = \sum_{j=1}^{k} \hat{p}_{22j} \log \bar{\phi}_j, \qquad \bar{\phi}_1 \leq \cdots \leq \bar{\phi}_k \tag{3.6}$$

where $\bar{\phi}_l = (\bar{p}_{11l}\bar{p}_{22l})/(\bar{p}_{12l}\bar{p}_{21l})$.

PROOF. Suppose that $\bar{p}$ satisfies (3.4)–(3.6), and then for p satisfying (3.1)

$$\begin{aligned}
&\sum_{l=1}^{k}(\hat{p}_{11l}\log p_{11l} + \hat{p}_{12l}\log p_{12l} + \hat{p}_{21l}\log p_{21l} + \hat{p}_{22l}\log p_{22l}) \\
&\quad = \sum_{l=1}^{k}[(\hat{p}_{11l} - \hat{p}_{22l})\log p_{11l} + (\hat{p}_{12l} + \hat{p}_{22l})\log p_{12l} \\
&\qquad\qquad + (\hat{p}_{21l} + \hat{p}_{22l})\log p_{21l} + \hat{p}_{22l}\log \phi_l] \\
&\quad = \sum_{l=1}^{k}[(\bar{p}_{11l} - \bar{p}_{22l})\log p_{11l} + (\bar{p}_{12l} + \bar{p}_{22l})\log p_{12l} \\
&\qquad\qquad + (\bar{p}_{21l} + \bar{p}_{22l})\log p_{21l} + \hat{p}_{22l}\log \phi_l] \\
&\quad \le \sum_{l=1}^{k}(\bar{p}_{11l}\log p_{11l} + \bar{p}_{12l}\log p_{12l} + \bar{p}_{21l}\log p_{21l} + \bar{p}_{22l}\log p_{22l})
\end{aligned}$$

where $\phi_l = (p_{11l}p_{22l})/(p_{12l}p_{21l})$. For $p = \bar{p}$, from the above, $\sum_{l=1}^{k}\sum_{i=1}^{2}\sum_{j=1}^{2}\hat{p}_{ijk} \cdot \log \bar{p}_{ijk} = \sum_{l=1}^{k}\sum_{i=1}^{2}\sum_{j=1}^{2}\bar{p}_{ijk}\log \bar{p}_{ijk}$ and thus

$$\sum_{l=1}^{k}\sum_{i=1}^{2}\sum_{j=1}^{2}\hat{p}_{ijk}\log \bar{p}_{ijk} - \sum_{l=1}^{k}\sum_{i=1}^{2}\sum_{j=1}^{2}\hat{p}_{ijk}\log p_{ijk} \ge \sum_{l=1}^{k}\sum_{i=1}^{2}\sum_{j=1}^{2}\bar{p}_{ijk}\log(\bar{p}_{ijk}/p_{ijk}) \ge 0.$$

Suppose that $\bar{p}$ is the MLE of p under (3.1) and then $\bar{p}$ can be expressed as the form of (3.2) and (3.3). By Kuhn-Tucker Conditions, it may be easily to prove (3.4)–(3.6). This prove the necessity.

By Lemma 3.1, the MLE of p under (3.1) can be expressed as (3.2) and (3.3), and thus by Lemmas 3.2 and 2.1, the MLE is equivalent to the I-projection for $(1,\ldots,1)^T$ onto

$$D = \left\{ p; p_{1jl} + p_{2jl} = \hat{p}_{1jl} + \hat{p}_{2jl}, p_{i1l} + p_{i2l} = \hat{p}_{i1l} + \hat{p}_{i2l}, \sum_{j=1}^{l} p_{22j} \le \sum_{j=1}^{l}\hat{p}_{22j}, \right.$$
$$\left. \sum_{j=1}^{k} p_{22j} = \sum_{j=1}^{k}\hat{p}_{22j}, i,j = 1,2; l = 1,\ldots,k-1 \right\}$$

which can be obtained by the proposed algorithm given in Section 2.

Algorithm
Let $p_{ijl}^{(0,3)} = 1$ and $a_l^{(0)} = 1$ for $i,j = 1,2$ and $l = 1,\ldots,k$.
Step$(n,1)$:

$$p_{ijl}^{(n,1)} = \frac{\hat{p}_{i+l}}{p_{i+l}^{(n-1,3)}} p_{ijl}^{(n-1,3)}, \qquad i,j = 1,2; l = 1,\ldots,k$$

where $\hat{p}_{i+l} = \hat{p}_{i1l} + \hat{p}_{i2l}$ and $p_{i+l}^{(n-1,3)} = p_{i1l}^{(n-1,3)} + p_{i2l}^{(n-1,3)}$ for $i = 1, 2$ and $l = 1, \ldots, k$.

Step$(n, 2)$:

$$p_{ijl}^{(n,2)} = \frac{\hat{p}_{+jl}}{p_{+jl}^{(n,1)}} p_{ijl}^{(n,1)}, \qquad i, j = 1, 2; l = 1, \ldots, k$$

where $\hat{p}_{+jl} = \hat{p}_{1jl} + \hat{p}_{2jl}$ and $p_{+jl}^{(n,1)} = p_{1jl}^{(n,1)} + p_{2jl}^{(n,1)}$ for $j = 1, 2$ and $l = 1, \ldots, k$.

Step$(n, 3)$:

$$p_{ijl}^{(n,3)} = \begin{cases} p_{ijl}^{(n,2)} & (i,j) = (1,1), (1,2), (2,1); l = 1, \ldots, k \\ p_{ijl}^{(n,2)} \dfrac{a_l^{(n)}}{a_l^{(n-1)}} & (i,j) = (2,2); l = 1, \ldots, k \end{cases}$$

where $(a_1^{(n)}, \ldots, a_k^{(n)})^T$ is the isotonic projection of $(\frac{\hat{p}_{221} a_1^{(n-1)}}{p_{221}^{(n,2)}}, \ldots, \frac{\hat{p}_{22k} a_k^{(n-1)}}{p_{22k}^{(n,2)}})^T$ with the weight $(p_{221}^{(n,2)}, \ldots, p_{22k}^{(n,2)})^T$ onto C,

$$C = \{x \in R^k : x_1 \leq \cdots \leq x_k\}.$$

4. Discussions

The proposed algorithm in Section 2 generalizes the algorithm given by Darroch and Ratcliff (1972). For the application given in Section 4, we have made some computations and $p^{(n)}$ will be convergent at about $n = 50$ once $\hat{p}_{ijl} = n_{ijl}/n$ $(i, j = 1, 2;\ l = 1, \ldots, k)$ are not very small. When s is equal to 1, it degenerates into the case given by Darroch and Ratcliff. When s is equal to t, $p \in D$ satisfying

$$\sum_{j=1}^{l} p_j \leq \sum_{j=1}^{l} h_j, \qquad \sum_{j=1}^{t} p_j = \sum_{j=1}^{t} h_j = 1$$

$(l = 1, \ldots, t)$ is usually defined as p stochastically larger than h and it is a very important relationship between probability vectors (see Robertson *et al.* (1988)).

For $p \in D$, let $y_j = \sum_{i=1}^{k} a_{ij} p_i - h_j$ for $j = 1, \ldots, t$ and

$$C = \left\{ x; \gamma \sum_{j=1}^{t} x_j y_j \leq 0, \quad \text{for any} \quad y, \gamma \geq 0 \right\}.$$

If C is a convex cone induced by some partial ordering in R^t, then the proposed algorithm given in Section 2 is applicable.

Acknowledgements

We wish to thank Professor Kuriki for helpful comments on improvement of earlier manuscript and also thank to referees for their helpful suggestions and comments. Project is supported by the National Natural Science Foundation of China.

Appendix

LEMMA A.1. *For $x \in R^s$ and the weight w, $\hat{x}$ is the projection of x onto C if and only if*

$$\hat{x} \in C,$$
$$\sum_{i=1}^{s}(x_i - \hat{x}_i)\hat{x}_i w_i = 0,$$
$$\sum_{i=1}^{s}(x_i - \hat{x}_i)y_i w_i \le 0, \quad \text{for} \quad \forall y \in C.$$

PROOF. See Theorem 1.3.2 of Robertson *et al.* ((1988), p. 17).

LEMMA A.2. $\sum_{j=1}^{s}\sum_{i=1}^{k} a_{ij}p_i^{(n-1)}\lambda_j^{(n)}/\lambda_j^{(n-1)} \le \sum_{j=1}^{s} h_j$.

PROOF. Let $y_i = -1/\lambda_i^{(n-1)}$ and $y = (y_1, \ldots, y_s)' \in C$. Then by Lemma A.1, we have

$$\sum_{j=1}^{s} w_j^{(n-1)}(\lambda_j^{(n-1)}h_j/w_j^{(n)} - \lambda_j^{(n)})(-1/\lambda_j^{(n-1)}) \le 0$$

which implies the lemma.

LEMMA A.3. $\sum_{i=1}^{k} p_i^{(n)} \le 1$.

PROOF.

$$\begin{aligned}\sum_{i=1}^{k} p_i^{(n)} &= \sum_{i=1}^{k} p_i^{(n-1)} \prod_{j=1}^{s}(\lambda_j^{(n)}/\lambda_j^{(n-1)})^{a_{ij}} \prod_{j=s+1}^{t}\left(h_j/\sum_{i=1}^{k} a_{ij}p_i^{(n-1)}\right)^{a_{ij}} \\ &\le \sum_{i=1}^{k}\left[\sum_{j=1}^{s} a_{ij}p_i^{(n-1)}(\lambda_j^{(n)}/\lambda_j^{(n-1)}) + \sum_{j=s+1}^{t} a_{ij}p_i^{(n-1)}\left(h_j/\sum_{i=1}^{k} a_{ij}p_i^{(n-1)}\right)\right] \\ &\le \sum_{j=1}^{s} h_j + \sum_{j=s+1}^{t} h_j = 1.\end{aligned}$$

LEMMA A.4. $\sum_{i=1}^{k} q_i \log\prod_{j=1}^{s}(\lambda_i^{(n)}/\lambda_j^{n-1})^{a_{ij}} \ge (\sum_{j=1}^{s} h_j)\log(\sum_{j=1}^{s} h_j/\sum_{j=1}^{s} \cdot \sum_{i=1}^{k} a_{ij}p_i^{(n-1)})$, *where q satisfies $\sum_{i=1}^{k} a_{ij}q_i = h_j$ for $j = 1, \ldots, s$.*

PROOF.

$$\begin{aligned}\sum_{i=1}^{k} q_i \log\prod_{j=1}^{s}(\lambda_i^{(n)}/\lambda_j^{n-1})^{a_{ij}} &= \sum_{j=1}^{s}\sum_{i=1}^{k} a_{ij}q_i \log(\lambda_j^{(n)}/\lambda_j^{(n-1)}) \\ &= \sum_{j=1}^{s} h_j \log(\lambda_j^{(n)}/\lambda_j^{(n-1)})\end{aligned}$$

$$\geq -\left(\sum_{j=1}^{s} h_j\right) \log \left\{ \sum_{j=1}^{s} h_j \lambda_j^{(n-1)} / \lambda_j^{(n)} / \sum_{j=1}^{s} h_j \right\}$$

$$= -\left(\sum_{j=1}^{s} h_j\right) \log \left(\sum_{i=1}^{k} \sum_{j=1}^{s} a_{ij} p_i^{(n-1)} / \sum_{j=1}^{s} h_j \right)$$

$$= \sum_{j=1}^{s} h_j \log \left(\sum_{j=1}^{s} h_j / \sum_{j=1}^{s} \sum_{i=1}^{k} a_{ij} p_i^{(n-1)} \right).$$

LEMMA A.5. *$\{I(q \mid p^{(n)}\}$ is nonincreasing in n and bounded below by zero.*

PROOF.

$$I(q \mid p^{(n)}) = I(q \mid p^{(n-1)}) - \sum_{i=1}^{k} q_i \log \left\{ \prod_{j=1}^{s} (\lambda_j^{(n)} / \lambda_j^{(n-1)})^{a_{ij}} \prod_{j=s+1}^{t} \left(h_j / \sum_{i=1}^{k} a_{ij} p_i^{(n-1)} \right)^{a_{ij}} \right\}$$

$$= I(q \mid p^{(n-1)}) - \sum_{i=1}^{k} q_i \log \prod_{j=1}^{s} (\lambda_j^{(n)} / \lambda_j^{(n-1)})^{a_{ij}} - \sum_{i=1}^{k} \sum_{j=s+1}^{t} a_{ij} q_i \log \left(h_j / \sum_{i=1}^{k} a_{ij} p_i^{(n-1)} \right)$$

$$\leq I(q \mid p^{(n)}) - \left(\sum_{j=1}^{s} h_j\right) \log \frac{\sum_{j=1}^{s} h_j}{\sum_{j=1}^{s} \sum_{i=1}^{k} a_{ij} p_i^{(n-1)}} - \left(\sum_{j=s+1}^{t} h_j\right) \log \frac{\sum_{j=s+1}^{t} h_j}{\sum_{j=s+1}^{t} \sum_{i=1}^{k} a_{ij} p_i^{(n-1)}}$$

$$\leq I(q \mid p^{(n-1)})$$

by Lemma A.4.

LEMMA A.6. *If h_j $(j = 1, \ldots, s)$ are positive, then $\lim_{n\to\infty} \lambda_j^{(n)} / \lambda_j^{(n-1)} = 1$, for $j = 1, \ldots, s$.*

PROOF. Since $\lambda_j^{(n)} / \lambda_j^{(n-1)}$ are uniformly bounded by (2.3), for $\{n_t\} \subset \{n\}$ we have $\lim_{t\to\infty} \lambda_j^{(n_t)} / \lambda_j^{(n_t-1)} = u_j$ for $j = 1, \ldots, s$.

$$\lim_{t\to\infty} \sum_{i=1}^{k} q_i \log \left\{ \prod_{j=1}^{s} (\lambda_j^{(n_t)} / \lambda_j^{(n_t-1)})^{a_{ij}} \right\} = \sum_{i=1}^{k} q_i \log \left\{ \prod_{j=1}^{s} u_j^{a_{ij}} \right\} = \sum_{j=1}^{s} h_j \log u_j$$

$$\geq -\log \left(\sum_{j=1}^{s} h_j / u_j \right)$$

$$= -\lim_{t\to\infty} \log\left(\sum_{j=1}^{s} h_j \lambda_j^{(n_t-1)}/\lambda_j^{(n_t)}\right)$$

$$= -\lim_{t\to\infty} \log\left(\sum_{i=1}^{k}\sum_{j=1}^{s} a_{ij} p^{(n_t-1)}\right) = 0$$

which implies $u_1 = \cdots = u_s = 1$.

Remark A. From Lemma A.5, A.6 and (1.1), we have the following results:

$$\lim_{n\to\infty} \sum_{i=1}^{k} a_{ij} p_i^{(n-1)} = h_j, \quad \text{for} \quad j = s+1, \ldots, t, \tag{A.1}$$

$$\lim_{n\to\infty} \sum_{j=1}^{s}\sum_{i=1}^{k} a_{ij} p_i^{(n-1)} = \sum_{j=1}^{s} h_j, \tag{A.2}$$

$$\lim_{n\to\infty} \sum_{j=1}^{k} p_i^{(n)} = 1, \tag{A.3}$$

$$\lim_{n\to\infty} (p_i^{(n)} - p_i^{(n-1)}) = 0. \tag{A.4}$$

PROOF OF THEOREM 2.1. $\{p^{(n)}\}$ is uniformly bounded, so for any subsequence, there exists $\{n_r\}$ which satisfies $\lim_{r\to\infty} p_i^{(n_r)} = p_i^*$ for $i = 1, \ldots, k$. From (2.1), p^* can be expressed into $p_i^* = \pi_i^* \mu^* \prod_{j=1}^{t} (\lambda_j^*)^{a_{ij}}$, $i = 1, \ldots, k$ and by (A.1), (A.2), (A.3) and Lemma 1.2, we only need to prove p^* satisfying (1.6) and (1.7) or (1.6′) and (1.7′).

Suppose there exists $1 = f_0 < f_1 < \cdots < f_{l-1} < f_l = s$, which satisfies

$$\lambda_1^* = \cdots = \lambda_{f_1}^* < \lambda_{f_1+1}^* = \cdots = \lambda_{f_2}^* < \cdots < \lambda_{f_{l-1}+1}^* = \cdots = \lambda_{f_l}^*. \tag{A.5}$$

From the algorithm given in Section 2 and (A.5), for sufficient large n_r, $(\lambda_{f_{l-1}+1}^{(n_r)}, \ldots, \lambda_{f_l}^{(n_r)})'$ is the solution of

$$\min\left\{\sum_{j=f_{l-1}+1}^{f_l} w_j^{(n_r)} (\lambda_j^{(n_r-1)} h_j / w_j^{(n_r)} - \mu_j)^2; \quad \text{subject to} \quad \mu_{f_{l-1}+1} \le \cdots \le \mu_{f_l}\right\}$$

where $w_j^{(n_r)} = \sum_{i=1}^{k} a_{ij} p_i^{(n_r-1)}$. Thus for $f_{l-1} + 1 \le m \le f_l$, by Lemma A.1 we have

$$\sum_{j=f_{l-1}+1}^{m}\sum_{i=1}^{k} a_{ij} p_i^* = \lim_{r\to\infty} \sum_{j=f_{l-1}+1}^{m}\sum_{i=1}^{k} a_{ij} p_i^{(n_r)} (\lambda_j^{(n_r)}/\lambda_j^{(n_r-1)}) \le \sum_{j=f_{l-1}+1}^{m} h_j,$$

$$\sum_{j=f_{t-1}+1}^{f_t} h_j = \lim_{r\to\infty} \sum_{j=f_{t-1}+1}^{f_t} h_j (\lambda_j^{(n_r-1)}/\lambda_j^{(n_r)})$$

$$= \lim_{r\to\infty} \sum_{j=f_{t-1}+1}^{f_t}\sum_{i=1}^{k} a_{ij} p_i^{(n_r-1)} = \sum_{j=f_{t-1}+1}^{f_t}\sum_{i=1}^{k} a_{ij} p_i^*.$$

Thus p^* satisfies (1.6′)and (1.7′) and By Lemma 1.2 and (A.4) this implies the theorem.

REFERENCES

Agresti, A. (1984). *Analysis of Ordinal Categorical*, Wiley, New York.

Agresti, A. and Coull, B. (1996). Order-restricted test for stratified comparisons of binomial proportions, *Biometrics*, **52**, 1103–1111.

Csiszar, I. (1967). Information-type measure of difference of probability distributions and indirect observations, *Studia Sci. Math. Hungar.*, **2**, 299–318.

Csiszar, I. (1975). I-divergence geometry of probability distributions and minimization problems, *Ann. Probab.*, **3**, 146–159.

Darroch, J. N. and Ratcliff, D. (1972). Generalized iterative scaling for log-linear models, *Ann. Math. Statist.*, **43**, 1470–1480.

Dykstra, R. L. (1985). An iterative procedure for obtaining I-projections onto the intersection of convex sets, *Ann. Probab.*, **13**, 975–984.

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

Kemperman, J. H. B. (1969). On the optimum rate of transmitting information, *Ann. Math. Statist.*, **40**, 2156–2177.

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

Kullback, S. (1967). An extension of information-theoretic derivation of certain limit relations for a Markov chain, *SIAM J. Control.*, **5**, 51–53.

Lemke, J. H. and Dykstra, R. L. (1984). An algorithm for multinomial maximum likelihood estimation with multiple cone restrictions. Tech. Report, 84-1, Department of Preventive Medicine, University of Iowa.

McDonald, J. W. and Diamond, I. (1983). Fitting generalized linear models with linear inequality parameter constraints, *The GLIM Newsletter*, **8**, 29–36.

McDonald, J. W. and Diamond, I. (1990). On the fitting of generalized linear models with nonnegativity parameter constraints, *Biometrics*, **46**, 201–206.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*, Wiley, New York.