Ann. Inst. Statist. Math.
Vol. 49, No. 6, 435–446 (1997)

A FENCHEL DUALITY ASPECT OF ITERATIVE I-PROJECTION PROCEDURES

BHASKAR BHATTACHARYA[1] AND RICHARD L. DYKSTRA[2]*

[1] *Department of Mathematics, Southern Illinois University, Carbondale, IL 62901-4408, U.S.A.*
[2] *Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242, U.S.A.*

(Received May 17, 1995; revised October 3, 1996)

Abstract. In this paper we interpret Dykstra's iterative procedure for finding an I projection onto the intersection of closed, convex sets in terms of its Fenchel dual. Seen in terms of its dual formulation, Dykstra's algorithm is intuitive and can be shown to converge monotonically to the correct solution. Moreover, we show that it is possible to sharply bound the location of the constrained optimal solution.

Key words and phrases: Algorithm, convex sets, Fenchel duality, I-projections, iterative, Kullback-Leibler.

1. Introduction

For finite probability vectors (PV's) $\boldsymbol{p} = (p(1), \ldots, p(m))'$ and $\boldsymbol{q} = (q(1), \ldots, q(m))'$ of length m, the *I-divergence* of $\boldsymbol{p}$ with respect to $\boldsymbol{q}$ (also known as the *Kullback-Leibler information number*) is given by

$$I(\boldsymbol{p} \mid \boldsymbol{q}) = \sum_{k=1}^{m} p(k) \ln \left(\frac{p(k)}{q(k)} \right).$$

Since $I(\boldsymbol{p} \mid \boldsymbol{q}) \geq 0$, and equals zero if and only if $\boldsymbol{p} = \boldsymbol{q}$, $I(\boldsymbol{p} \mid \boldsymbol{q})$ is often treated heuristically as a measure of distance or divergence between $\boldsymbol{p}$ and $\boldsymbol{q}$. It is natural to consider the "closest" PV to $\boldsymbol{q}$ which lies within a specified set of PV's $\mathcal{C}$. A PV $\boldsymbol{u} \in \mathcal{C}$ that satisfies

$$I(\boldsymbol{u} \mid \boldsymbol{q}) = \min_{\boldsymbol{p} \in \mathcal{C}} I(\boldsymbol{p} \mid \boldsymbol{q}) < \infty \tag{1.1}$$

is said to be an *I-projection* of $\boldsymbol{q}$ onto $\mathcal{C}$. The I-projection $\boldsymbol{u}$ always exists uniquely if $\mathcal{C}$ is closed and convex (Csiszar (1975)).

* Partial support was provided by National Science Foundation Grant DMS 91-04673.

I-projections play a basic role in the information theoretic approach to statistics (Kullback (1959), Good (1963)). They are also important in the theory of large deviations (Sanov (1957)) and in statistical physics for the maximization of entropy (Rao (1965), Jaynes (1957)). For a duality approach to I-projections for general probability distributions, see Bhattacharya and Dykstra (1995).

Depending on the form of the set $\mathcal{C}$, it may be difficult to find a solution to the I-projection problem in (1.1). Csiszar (1975) has shown that if $\mathcal{C}$ can be expressed as $\bigcap_{i=1}^{t} \mathcal{C}_i$, where each $\mathcal{C}_i$ is a closed, linear set, then the sequence of cyclic iterated I-projections converges to the solution of (1.1). Dykstra (1985*a*) modified Csiszar's procedure to encompass the case where each $\mathcal{C}_i$ is an arbitrary closed, convex set. He showed that the desired I-projection can be obtained as the limit of cyclic I-projections onto the $\mathcal{C}_i$ if the projected vectors are appropriately modified and a mild condition holds. Winkler (1990) showed that the stated condition always holds, and hence the algorithm is always valid. In this paper, we use an extension to a Fenchel duality theorem for $\mathbb{R}^m$ to prove Dykstra's result in a much shorter and more intuitive fashion.

2. Main results

For a closed, proper, convex function f on $\mathbb{R}^m$, the *convex conjugate* f^* defined on $\mathbb{R}^m$ is given by

$$f^*(\boldsymbol{y}) = \sup_{\boldsymbol{x}} \left\{ \sum_{k=1}^{m} x(k)y(k) - f(\boldsymbol{x}) \right\}.$$

It is well known (Rockafellar (1970)) that f^* is also a closed, proper, convex function on $\mathbb{R}^m$ and that $f^{**} = f$. For a convex cone K in $\mathbb{R}^m$, the *dual* cone K^* in $\mathbb{R}^m$ is defined as

$$K^* = \left\{ \boldsymbol{y} \in \mathbb{R}^m : \sum_{k=1}^{m} x(k)y(k) \leq 0, \forall \boldsymbol{x} \in K \right\}.$$

Corollary 2.1, from Rockafellar ((1970), p. 335), will be used to identify the dual problem. By "ri" we mean relative interior as defined in Rockafellar (1970)

COROLLARY 2.1. *Let f be a closed, proper, convex function on $\mathbb{R}^m$, and let K be a nonempty, closed, convex cone in $\mathbb{R}^m$. Then*

$$\text{(I)} \quad \inf_{\boldsymbol{x} \in K} f(\boldsymbol{x}) = \sup_{\boldsymbol{y} \in K^*} -f^*(-\boldsymbol{y}) \quad \text{(II)} \tag{2.1}$$

provided $\operatorname{ri}(\operatorname{dom} f) \bigcap \operatorname{ri}(K) \neq \emptyset$, *and moreover, the supremum on the right side is attained. In general, $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ are solutions to these problems which satisfy*

$$f(\hat{\boldsymbol{x}}) = \inf_{\boldsymbol{x} \in K} f(\boldsymbol{x}) = \sup_{\boldsymbol{y} \in K^*} -f^*(-\boldsymbol{y}) = -f^*(-\hat{\boldsymbol{y}})$$

if and only if

$$\begin{aligned}
&\text{(i)} && -\hat{\boldsymbol{y}} \text{ is a subgradient of } f \text{ at } \hat{\boldsymbol{x}},\\
&\text{(ii)} && \hat{\boldsymbol{x}} \in K,\\
&\text{(iii)} && \hat{\boldsymbol{y}} \in K^*, \quad \text{and}\\
&\text{(iv)} && \sum_{k=1}^{m} \hat{x}(k)\hat{y}(k) = 0.
\end{aligned}$$

For a given convex set $\mathcal{C}$ of PV's, the I-projection problem (1.1) can be expressed as the left side of (2.1), if we define

$$f(\boldsymbol{x}) = \begin{cases} \displaystyle\sum_{k=1}^{m} x(k) \ln\left(\frac{x(k)}{q(k)}\right), & \text{if} \quad x(k) > 0,\ \forall k,\ \displaystyle\sum_{k=1}^{m} x(k) = 1 \\ +\infty, & \text{otherwise,} \end{cases} \tag{2.2}$$

and K by

$$K = \{\alpha \boldsymbol{x} : \boldsymbol{x} \in \mathcal{C}, \alpha \geq 0\}.$$

Dykstra (1985*b*) has shown that the convex conjugate of the function in (2.2) is given by

$$f^*(\boldsymbol{y}) = \ln\left(\sum_{k=1}^{m} q(k) \exp(y(k))\right)$$

($\boldsymbol{q}$ need not sum to 1). Note that f^* gives the cumulant generating function of $\boldsymbol{q}$ (with probability on a particular set of values $z(1), z(2), \ldots, z(m)$) if f^* is evaluated along the one-dimensional path $y \cdot (z(1), z(2), \ldots, z(m))'$. Thus minimizing f^* over the (one-dimensional) region $K^* = \{\boldsymbol{v} \in \mathbb{R}^m : \boldsymbol{v} = \alpha \boldsymbol{z}, \alpha \geq 0\}$ is equivalent to minimizing $I(\boldsymbol{p} \mid \boldsymbol{q})$ over the set $-K = \{\boldsymbol{x} \in \mathbb{R}^m : \sum_{i=1}^{m} x(i)z(i) \geq 0\}$, since K^* is the dual of K. Restricting $-K$ to the PV's, the duality theorem then shows that minimizing the cumulant generating function over nonnegative values of y is equivalent to finding the I-projection of $\boldsymbol{q}$ onto the set of distributions over $(z(1), z(2), \ldots, z(m))$ with a nonnegative mean.

Dykstra (1985*b*) has also shown that if $\hat{\boldsymbol{y}}$ solves the dual problem (II) in (2.1), then the vector $\hat{\boldsymbol{p}} = (\hat{p}(1), \hat{p}(2), \ldots, \hat{p}(m))'$ given by

$$\hat{p}(k) = \frac{q(k) \exp(-\hat{y}(k))}{\sum_{s=1}^{m} q(s) \exp(-\hat{y}(s))}, \qquad k = 1, 2, \ldots m, \tag{2.3}$$

solves the I-projection problem (1.1) if $\mathcal{C}$ is closed and $\operatorname{ri}(\operatorname{dom} f) \cap \operatorname{ri}(K) \neq \emptyset$.

In some problems $\mathcal{C}$ may constrain some of the $x(k)$ to always be zero, which causes the condition $\operatorname{ri}(\operatorname{dom} f) \cap \operatorname{ri}(K) \neq \emptyset$ to be violated. It can be shown that the dual formulation above is still valid if the domain of f^* is expanded to be $\mathbb{R}_1^* \times \mathbb{R}_2^* \times \cdots \times \mathbb{R}_m^*$, where $\mathbb{R}_i^* = \mathbb{R} \cup \{\pm\infty\}$. The results hold as before with the interpretation that $e^{-\infty} = 0$ and $0 \cdot \infty = 0$.

We now let $\boldsymbol{q}$ be a fixed PV of length m. We wish to find the I-projection of $\boldsymbol{q}$ onto the nonempty set $\mathcal{C} = \bigcap_{i=1}^{t} \mathcal{C}_i$, where each $\mathcal{C}_i$ is a closed, convex set of PV's of length m. We only assume that there exists an $\boldsymbol{r} \in \mathcal{C}$ such that $I(\boldsymbol{r} \mid \boldsymbol{q}) < \infty$. Winkler (1990) has shown that each of the I-projections required for the algorithm exists. We can use the Fenchel duality theorem in each step of the algorithm to identify a corresponding dual problem. Our point is that matters are much simpler and more intuitive in the dual formulation.

We present the algorithm below incorporating the primal and the dual formulations separately. Multiplication and division of vectors is done coordinatewise. To implement the algorithm for the primal problem at the i-th cycle and j-th step, we first form $\boldsymbol{s}_{ij}$, the vector to be projected; the I-projection obtained by projecting $\boldsymbol{s}_{ij}$ onto $\mathcal{C}_j$ is denoted by $\boldsymbol{p}_{ij}$ (although $\boldsymbol{s}_{ij}$ may not be a PV, we can define the I-projection $\boldsymbol{p}_{ij}$ of $\boldsymbol{s}_{ij}$ in the same way, for further details see Dykstra (1985a)); and the solution to the corresponding dual problem is $\boldsymbol{y}_{ij}$ (used in the dual formulation of the algorithm). We note that it is straightforward to show that at the i-th cycle, j-th step, $p_{ij}(k) = 0$ if and only if $s_{ij}(k) = 0$ or there exists v such that $p(k) = 0$, for all $\boldsymbol{p} \in \mathcal{C}_v$ (Winkler (1990)).

We first state algorithm in the primal form (Dykstra (1985a)) and then reformulate it in terms of the much more intuitive dual problem. The key point is that the intersection constraints in the primal problem translate to direct sum constraints in the dual problem. This is what allows the dual formulation of the algorithm to be phrased as a cyclic, descent algorithm which successively minimizes over one vector at a time while the rest are held fixed. We use the duality structure to give a short proof that the algorithm must work correctly. It will be convenient to let $\pi_i(\boldsymbol{s})$ denote the I-projection of the vector $\boldsymbol{s}$ onto the set $\mathcal{C}_i$.

Primal formulation of the algorithm.

(1) *Initialization.* Set $\boldsymbol{s}_{0,i} = \boldsymbol{p}_{0,i} = \boldsymbol{q}$, and begin with $n = 1$, $i = 1$.

(2) *Implementation.*

(i) For $i = 1$, set $\boldsymbol{s}_{n,1} = \boldsymbol{p}_{n-1,t}/(\boldsymbol{p}_{n-1,1}/\boldsymbol{s}_{n-1,1})$; for $2 \le i \le t$, set $\boldsymbol{s}_{n,i} = \boldsymbol{p}_{n,i-1}/(\boldsymbol{p}_{n-1,i}/\boldsymbol{s}_{n-1,i})$; (we assume that $0/0 = 0$).

(ii) Let $\boldsymbol{p}_{n,i} = \pi_i(\boldsymbol{s}_{n,i})$.

(iii) If $i < t$, increment i by 1 and repeat (2). If $i = t$, increment n by 1, set $i = 1$ and repeat (2).

Of course, the key point is that $\boldsymbol{p}_{n,i}$ must converge to the I-projection of $\boldsymbol{q}$ onto $\mathcal{C} = \bigcap_{i=1}^{t} \mathcal{C}_i$. However, the proof (Dykstra (1985a)) is quite complicated and involved. It is much cleaner in the dual formulation.

The dual problem is equivalent to

$$\inf_{\boldsymbol{y} \in (\bigcap_1^t K_i)^*} \sum_{k=1}^{m} q(k) e^{-y(k)} - \inf_{\boldsymbol{y} \in cl(K_1^* + \cdots + K_t^*)} \sum_{k=1}^{m} q(k) e^{-y(k)}$$

$$= \inf_{\boldsymbol{y}_i \in K_i^*, 1 \le i \le t} \sum_{k=1}^{m} q(k) e^{-y_1(k) - \cdots - y_t(k)}.$$

Typically the dual constraint region would be the direct sum $K_1^* + \cdots + K_t^*$. However, the direct sum of closed sets need not be closed (Hestenes (1975)) and

hence our dual constraint region is the closure of the direct sum. This possible lack of closure of the direct sum complicates the proof substantially.

One of the drawbacks to iterative methods of optimization are the difficulties in specifying the distance of the current estimate from the actual solution. Small changes in the objective function between successive steps of the iterative procedure are no guarantee that the actual optimal value is close by.

However, consider the situation where $\boldsymbol{p}^*$ is a feasible PV (i.e. $\boldsymbol{p}^* \in \bigcap_1^t K_i$) which is close to $\boldsymbol{p}_{n,t}$. We let $\hat{\boldsymbol{p}}$ denote the actual, constrained solution to our problem. We note that $\boldsymbol{p}^*$ may be obtained by a least square projection onto a subset of $\bigcap_{i=1}^t K_i$, by a cyclic descent method if the constraint region can also be written as a direct sum of convex sets, or by some other procedure. However, since

$$\sum_{k=1}^{m} \hat{p}(k) \ln\left(\frac{\hat{p}(k)}{q(k)}\right) \geq \sum_{i=1}^{t}\sum_{k=1}^{m} p_{n,i}(k) \ln\left(\frac{p_{n,i}(k)}{s_{n,i}(k)}\right), \tag{2.4}$$

with the difference between the two sides of (2.4) converging monotonically to zero as $n \to \infty$, we can estimate $I(\hat{\boldsymbol{p}} \mid \boldsymbol{q})$ to arbitrary accuracy by choosing n sufficiently large. Moreover, since

$$\begin{aligned} I(\boldsymbol{p}^* \mid \boldsymbol{q}) - \sum_{i=1}^{t} I(\boldsymbol{p}_{n,i} \mid \boldsymbol{s}_{n,i}) &\geq I(\boldsymbol{p}^* \mid \boldsymbol{q}) - I(\hat{\boldsymbol{p}} \mid \boldsymbol{q}) \\ &\geq I(\boldsymbol{p}^* \mid \hat{\boldsymbol{p}}) \geq \frac{1}{2}\left(\sum_{k=1}^{m} |p^*(k) - \hat{p}(k)|\right)^2 \end{aligned} \tag{2.5}$$

(Csiszar (1975)), we can specify an upper bound on both the I-divergence distance and the variation distance (L_1-norm) between the vector $\boldsymbol{p}^*$ and the true solution $\hat{\boldsymbol{p}}$.

In similar fashion, we can expand $I(\boldsymbol{p}^* \mid \hat{\boldsymbol{p}})$ using a Taylor series expansion to obtain the bound

$$I(\boldsymbol{p}^* \mid \hat{\boldsymbol{p}}) \geq \frac{1}{2}\sum_{k=1}^{m}\left(1 - \frac{p^*(k) \wedge \hat{p}(k)}{p^*(k) \vee \hat{p}(k)}\right)^2 p^*(k) \tag{2.6}$$

where "$\wedge$" ("$\vee$") indicates infimum (supremum). Thus, for example, if $I(\boldsymbol{p}^* \mid \hat{\boldsymbol{p}})$ is bounded above by ϵ, it is straightforward to show that for each i,

$$a_i p^*(i) < \hat{p}(i) < \frac{p^*(i)}{a_i}$$

where $a_i = 1 - \sqrt{2\epsilon/p^*(i)}$. For ϵ sufficiently small, this gives a bound on how far $\hat{\boldsymbol{p}}$ can fall from the vector $\boldsymbol{p}^*$. Of course one can also find an approximation region for possible values of $\hat{\boldsymbol{p}}$ by numerically checking whether candidate $\boldsymbol{p}$'s (plugged in for $\hat{\boldsymbol{p}}$) satisfy (2.5) and (2.6). However this becomes rather intractable for large m.

The efficiency of the algorithm depends largely on the nature of the K_i's. If all the K_i's are orthogonal with each other (in terms of I-divergence) a single pass through each constraint will suffice.

The algorithm stated in terms of the individual dual problems is just the following simple, cyclic, descent procedure. In the dual formulation, the algorithm amounts to just sequentially minimizing the objective function with all $\boldsymbol{y}$'s but one held fixed, and then updating that $\boldsymbol{y}$.

Dual formulation of the algorithm.

(1) *Initialization.* Set $\boldsymbol{y}_{0,i} \equiv \mathbf{0}$, and begin with $n = 1$, $i = 1$.

(2) *Implementation.*

(i) Let $\boldsymbol{y}_{n,i}$ denote the solution to

$$\inf_{y \in K_i^*} \sum_{k=1}^{m} q(k) e^{-y_{n,1}(k) - \cdots - y_{n,i-1}(k) - y(k) - y_{n-1,i+1}(k) - \cdots - y_{n-1,t}(k)}.$$

(ii) If $i < t$, increment i by 1 and repeat (2). If $i = t$, increment n by 1, set $i = 1$ and repeat (2).

Careful inspection of the two algorithms together with (2.3) will reveal that

$$p_{n,i}(k) = \frac{q(k) e^{-y_{n,1}(k) - \cdots - y_{n,i}(k) - y_{n-1,i+1}(k) - \cdots - y_{n-1,t}(k)}}{\sum_{s=1}^{m} q(s) e^{-y_{n,1}(s) - \cdots - y_{n,i}(s) - y_{n-1,i+1}(s) - \cdots - y_{n-1,t}(s)}}$$

for $1 \leq k \leq m$.

If it should happen that

$$\boldsymbol{y}_{n,1} + \cdots + \boldsymbol{y}_{n,i} + \boldsymbol{y}_{n-1,i+1} + \cdots + \boldsymbol{y}_{n-1,t} \to \hat{\boldsymbol{y}}$$

for some vector $\hat{\boldsymbol{y}}$ as $n \to \infty$, then it would easily follow by continuity that

$$p_{n,i}(k) \to \frac{q(k) e^{-\hat{y}(k)}}{\sum_{s=1}^{m} q(s) e^{-\hat{y}(s)}} = \hat{p}(k), \qquad k = 1, \ldots, m,$$

as $n \to \infty$. Moreover, if $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{y}}$ should satisfy conditions (i)–(iv) of Corollary 2.1, then $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{y}}$ would have to solve (1.1) and its dual, respectively, which would establish the validity of both the primal and dual forms of the algorithm. That this is indeed the case is the gist of the following theorem.

THEOREM 2.1. *Let the sets C_i, $1 \leq i \leq t$, be closed, convex sets of PV's of length m and let $\boldsymbol{q}$ be a positive PV such that there exists $\boldsymbol{r} \in \mathcal{C} = \bigcap_{i=1}^{t} C_i$ with $I(\boldsymbol{r} \mid \boldsymbol{q}) < \infty$. Then $\boldsymbol{p}_{ni} \to \hat{\boldsymbol{p}}$ as $n \to \infty$, for $1 < i < t$, where $\hat{\boldsymbol{p}}$ is the unique I-projection of $\boldsymbol{q}$ onto $\mathcal{C}$ and the $\boldsymbol{p}_{ni}$ are obtained as in the algorithm.*

PROOF. It is easily shown that

$$\sum_{k=1}^{m} q(k) e^{-y_{n,1}(k) - \cdots - y_{n,i}(k) - y_{n-1,i+1}(k) - \cdots - y_{n-1,t}(k)}$$

is nonincreasing in n and i. Moreover, it easily follows that every subsequence of $\{n\}_{n=1}^{\infty}$ has a further subsequence $\{n_j\}_{j=1}^{\infty}$ such that

$$\boldsymbol{y}_{n_j,1} + \cdots + \boldsymbol{y}_{n_j,t} \to \hat{\boldsymbol{y}}$$

where $\hat{\boldsymbol{y}}$ may have coordinates that are $+\infty$. Clearly, $\hat{\boldsymbol{y}} \in cl(K_1^* + \cdots + K_t^*) = (\bigcap_{i=1}^t K_i)^* = K^*$.

It easily follows by continuity that

$$p_{n_j,t}(k) = \frac{q(k)e^{-y_{n_j,1}(k)-\cdots-y_{n_j,t}(k)}}{\sum_{s=1}^m q(s)e^{-y_{n_j,1}(s)-\cdots-y_{n_j,t}(s)}}$$
$$\to \frac{q(k)e^{-\hat{y}(k)}}{\sum_{s=1}^m q(s)e^{-\hat{y}(s)}} = \hat{p}(k), \qquad k = 1, \ldots, m.$$

Dykstra (1985a) has shown that $I(\boldsymbol{p}_{n,i} \mid \boldsymbol{s}_{n,i}) - I(\boldsymbol{p}_{n-1,i} \mid \boldsymbol{s}_{n-1,i}) \geq I(\boldsymbol{p}_{n,i} \mid \boldsymbol{p}_{n,i-1}) \to 0$. Then using the inequality that

$$\sum_{s-1}^{m} |q(s) - r(s)| \leq [2I(\boldsymbol{q} \mid \boldsymbol{r})]^{1/2}$$

for any two vectors $\boldsymbol{q}$, $\boldsymbol{r}$, it follows easily that $\boldsymbol{p}_{n_j,i}$ is arbitrarily close to $\boldsymbol{p}_{n_j,t}$ if j is sufficiently large, for any i. Then it must be the case that $\hat{\boldsymbol{p}} \in \bigcap_{i=1}^t K_i = K$.

It follows from Dykstra ((1985b), Corollary 2.1) that $-\hat{\boldsymbol{y}}$ is a subgradient of the f from (2.2) at $\hat{\boldsymbol{p}}$.

Thus if,

$$\sum_{k=1}^{m} \hat{y}(k)\hat{p}(k) = 0, \tag{2.7}$$

the four conditions of Corollary 2.1 will be met and $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{y}}$ will be the solutions to the primal and dual problems. Since every subsequence of $\{\boldsymbol{p}_{n,i}\}_{n=1}^{\infty}$ has a sub-subsequence which converges to $\hat{\boldsymbol{p}}$, the entire sequence must converge to $\hat{\boldsymbol{p}}$.

To establish (2.7), recall that

$$\boldsymbol{s}_{n,i} = \boldsymbol{p}_{n,i-1}/(\boldsymbol{p}_{n-1,i}/\boldsymbol{s}_{n-1,i})$$

(for $2 \leq i \leq t$). It follows that (suppressing the index of summation k) for $2 \leq i \leq t$,

$$\begin{aligned} I(\boldsymbol{p}_{n_j,i} \mid \boldsymbol{s}_{n_j,i}) &= \sum_k \boldsymbol{p}_{n_j,i} \ln(\boldsymbol{p}_{n_j,i}/\boldsymbol{s}_{n_j,i}) \\ &= \sum_k \boldsymbol{p}_{n_j,i} \ln(\boldsymbol{p}_{n_j,i}/\boldsymbol{p}_{n_j,i-1}) + \sum_k \boldsymbol{p}_{n_j,i} \ln(\boldsymbol{p}_{n_j-1,i}/\boldsymbol{s}_{n_j-1,i}) \\ &= \sum_{m=n_h+1}^{n_j} \sum_k \boldsymbol{p}_{n_j,i} \ln(\boldsymbol{p}_{m,i}/\boldsymbol{p}_{m,i-1}) \\ &\quad + \sum_k (\boldsymbol{p}_{n_j,i} - \boldsymbol{p}_{n_h,i}) \ln(\boldsymbol{p}_{n_h,i}/\boldsymbol{s}_{n_h,i}) \\ &\quad + \sum_k \boldsymbol{p}_{n_h,i} \ln(\boldsymbol{p}_{n_h,i}/\boldsymbol{s}_{n_h,i}). \end{aligned}$$

A similar expression can be derived when $i = 1$.

Dykstra (1985a) has shown that the left side is nondecreasing in j and bounded above. Then, since the left side limit exists as $j \to \infty$, as does the limit of the last two summations on the right side, the limit of the first right side summation must also exist as $j \to \infty$.

Then summing over i and letting $h \to \infty$, we obtain

$$
\begin{aligned}
0 &= \lim_{h\to\infty} \sum_{i=1}^{t} \sum_{k} (\boldsymbol{p}_{n_h,i} - \hat{\boldsymbol{p}}) \ln(\boldsymbol{p}_{n_h,i}/\boldsymbol{s}_{n_h,i}) \\
&= \lim_{h\to\infty} \sum_{i=1}^{t} \sum_{k} (\boldsymbol{p}_{n_h,i} - \hat{\boldsymbol{p}})(-\boldsymbol{y}_{n_h,i}) \\
&= \lim_{h\to\infty} \sum_{i=1}^{t} \sum_{k} \hat{\boldsymbol{p}} \boldsymbol{y}_{n_h,i} \left(\text{since } \sum_{k} \boldsymbol{p}_{n,i} \boldsymbol{y}_{n,i} = 0, \forall n, i \right) \\
&= \lim_{h\to\infty} \sum_{k} \hat{\boldsymbol{p}} \sum_{i=1}^{t} \boldsymbol{y}_{n_h,i} \\
&= \sum_{k} \hat{\boldsymbol{p}} \hat{\boldsymbol{y}}
\end{aligned}
$$

which is condition (2.7). □

Csiszar (1975) showed that sequentially projecting the I-projections onto the individual constraint regions gives a sequence of probability vectors that converges to the true I-projection of $\boldsymbol{q}$ if all the $\mathcal{C}_i$ are linear sets (i.e., $p_1, p_2 \in \mathcal{C}_i \Rightarrow \alpha p_1 + (1-\alpha)p_2 \in \mathcal{C}_i$, for *all* α for which $\alpha p_1 + (1-\alpha)p_2$ is a PV). However, this algorithm will typically not work if the $\mathcal{C}_i$ are not linear sets and the reason is clear from the dual formulation. If the $\mathcal{C}_i$ are linear sets, then each projection (in the dual formulation) allows each point in K_i^* as a possible value of $\boldsymbol{y}_{ni}$ (thus all of K_i^* is feasible) since $K_i^* = \{\boldsymbol{y} + \boldsymbol{z} : \boldsymbol{y} \in K_i^*\}$ for every $\boldsymbol{z} \in K_i^*$. This won't be true, however, if the $\mathcal{C}_i$'s are not linear sets. Although Csiszar's procedure is simpler in the primal formulation, it is more complex in the dual form (and, of course, only works if the constraint regions are linear sets).

3. Example

We consider the problem of finding the maximum likelihood estimates of the probabilities for a two-way classification with multinomial sampling under the constraints that the local odds ratios ($\theta_{ij} = p_{ij}p_{i+1,j+1}/p_{i+1,j}p_{i,j+1}$) are all at least 1 without any additional model assumptions. As is well known $\theta_{ij} = 1, \forall i, j$ implies the two classifications are stochastically independent and $\theta_{ij} \geq 1, \forall i, j$ implies there is a positive association between the ordinal variables.

The problem we consider can be expressed as

$$
\sup_{\ln \boldsymbol{p} \in \bigcap_{i=1}^{r} \bigcap_{j=1}^{c} K_{ij}} \prod_{j=1}^{c} \prod_{i=1}^{r} p_{ij}^{n_{ij}}, \tag{3.1}
$$

where $\boldsymbol{p} \in P$ (the class of all PV's of length rc), the n_{ij}'s are the observed frequencies and the closed convex cones K_{ij} are defined as

$$K_{ij} = \{\boldsymbol{x} : x_{ij} - x_{i,j+1} - x_{i+1,j} + x_{i+1,j+1} \geq 0\}$$

for $i = 1, \ldots, r-1$, $j = 1, \ldots, c-1$. Let $\boldsymbol{g}$ denote the $(r \times c)$ matrix of the empirical distribution. Dykstra and Lemke (1988) have shown that (3.1) has the same solution as the I-projection problem

$$\inf_{\boldsymbol{p} \in \bigcap_{i=1}^{r} \bigcap_{j=1}^{c} (\boldsymbol{g} - H_{ij})} I(\boldsymbol{p} \mid \boldsymbol{q}) \tag{3.2}$$

where

$$H_{ij} = \left\{ \boldsymbol{x} : \sum_{k=1}^{i} \sum_{m=1}^{j} x_{km} \leq 0 \right\}$$

when $i = 1, \ldots, r-1$, $j = 1, \ldots, c-1$, and

$$H_{ij} = \left\{ \boldsymbol{x} : \sum_{k=1}^{i} \sum_{m=1}^{j} x_{km} = 0 \right\}$$

when $i = r$ or $j = c$, and $\boldsymbol{q}$ is the uniform PV which has values $q_{ij} = (rc)^{-1}$. In words, this amounts to the sum of every upper-left corner of cell (i, j) of $\boldsymbol{p}$ being as large as the same of $\boldsymbol{g}$ and equality holds whenever $i = r$ or $j = c$. The (i, j)-th constraint $(1 \leq i \leq r-1,\ 1 \leq j \leq c-1)$ can be expressed as

$$\boldsymbol{g} - H_{ij} = \{\boldsymbol{p} : \boldsymbol{p} \text{ is a PV and } \boldsymbol{a}_{ij}' \boldsymbol{p} \leq 0\}$$

where the (k, m)-th term of the vector $\boldsymbol{a}_{ij}$ is given by

$$\sum_{s=1}^{i} \sum_{t=1}^{j} g_{st} - \mathrm{I}(1 \leq k \leq i, 1 \leq m \leq j),$$

where I is the indicator function. The dual cone $(\boldsymbol{g} - H_{ij})^*$ is then given by

$$(\boldsymbol{g} - H_{ij})^* = \{\alpha \boldsymbol{a}_{ij} : \alpha \geq 0\}$$

when $i = 1, \ldots, r-1$, $j = 1, \ldots, c-1$, and

$$(\boldsymbol{g} - H_{ij})^* = \{\alpha \boldsymbol{a}_{ij} : \alpha \in \mathbb{R}\}$$

when $i = r$ or $j = c$.

Solving (3.1) is thus equivalent to solving the problem

$$\inf \sum_{k} \sum_{m} e^{-\sum_{i=1}^{r} \sum_{j=1}^{c} \alpha_{ij} a_{ij}(k,m)}$$

Table 1. Cross-classification of job satisfaction by income.

	Job Satisfaction			
Income (US$)	VD	LD	MS	VS
<6000	20	24	80	82
6000–15,000	22	38	104	125
15,000 25,000	13	28	81	113
>25,000	7	18	54	92

(VD = Very Dissatisfied; LD = Little Dissatisfied; MS = Moderately Satisfied; VS = Very Satisfied).

Table 2. Maximum likelihood estimates of data in Table 1 subject to the constraints $\theta_{ij} \geq 1$ $\forall i, j$ (values in first parentheses are the unrestricted male and values in second parentheses are the expected counts after smoothing).

	Job Satisfaction			
Income (US$)	VD	LD	MS	VS
<6000	0.0222	0.0293	0.0862	0.0910
	(0.0222)	(0.0266)	(0.0888)	(0.0910)
	(20.00)	(26.37)	(77.63)	(82.00)
6000–15,000	0.0244	0.0400	0.1176	0.1387
	(0.0244)	(0.0422)	(0.1154)	(0.1387)
	(22.00)	(36.00)	(106.00)	(125.00)
15,000–25,000	0.0144	0.0307	0.0903	0.1254
	(0.0144)	(0.0311)	(0.0899)	(0.1254)
	(13.00)	(27.63)	(81.37)	(113.00)
>25,000	0.0078	0.0200	0.0599	0.1021
	(0.0078)	(0.0200)	(0.0599)	(0.1021)
	(7.00)	(18.00)	(54.00)	(92.00)

(VD = Very Dissatisfied; LD = Little Dissatisfied; MS = Moderately Satisfied; VS = Very Satisfied).

where the infimum is taken over the set

$$\{\alpha_{ij} \geq 0, 1 \leq i \leq r-1, 1 \leq j \leq c-1 \text{ and } \alpha_{ij} \in \mathbb{R}, i = r \text{ or } j = c\}.$$

However, equivalently, we can consider each I-projection problem at hand, find its dual problem, and the corresponding dual solution and then relate the primal solution to this dual solution.

We illustrate the procedure using the data described in Table 1 taken from Agresti ((1990), p. 21), also in 1984 General Social Survey (see Norušis (1988)). Thus we wish to solve the problem in (3.1) using the n_{ij} in Table 1. Since $\theta_{12} = 0.82$, some constraints will have to be active. When written in terms of (3.2) there

are essentially fifteen constraints (in addition to the always present probability vector constraint). For every constraint, the dual problem is solved by using a one-dimensional Newton-Raphson method. The maximum likelihood estimates subject to the constraints $\theta_{ij} \geq 1\ \forall i, j = 1, 2, 3$ are listed in Table 2.

We calculated the bound given in (2.5) for this example. We obtained $\boldsymbol{p}^*$ by changing some elements of the solution given in Table 2 ($p^*_{12} = 0.0291$, $p^*_{13} = 0.0863$, $p^*_{32} = 0.0308$, $p^*_{33} = 0.0903$). It is easy to check that this $\boldsymbol{p}^*$ is feasible. Then the left side of (2.5) is zero upto five decimal places (for large n).

Patefield (1982) considered $r \times c$ contingency tables and tests of hypotheses $H_0 : \theta_{ij} = 1,\ \forall i, j$ versus $H_1 : \theta_{ij} \geq 1,\ \forall i, j$. The likelihood ratio statistic was calculated using numerical routines (for small values of r and c). The restricted estimates as derived in this paper would prove to be useful for this purpose.

4. Discussion

The iterative proportional fitting procedures, starting with Deming and Stephan (1940), are used widely in many facets of statistical applications. We have used the Fenchel duality theorem to show that Dykstra's iterative proportional fitting procedure amounts to a sequential coordinatewise minimization procedure in the dual space. This algorithm is simple and helps to explain why Dykstra's algorithm should work in the primal (I-projection) problem.

Acknowledgements

The authors thank the referees for helpful and knowledgeable comments.

References

Agresti, A. (1990). *Categorical Data Analysis*, Wiley, New York.

Bhattacharya, B. and Dykstra, R. L. (1995). A general duality approach to I-projections, *J. Statist. Plann. Inference*, **3**, 146–159.

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

Deming, W. E. and Stephan, F. F. (1940). On a least square adjustment of a sampled frequency table when the expected marginal totals are known, *Ann. Math. Statist.*, **11**, 427–444.

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

Dykstra, R. L. (1985*b*). Computational aspects of I-projections, *J. Statist. Comput. Simulation*, **21**, 265–274.

Dykstra, R. L. and Lemke, J. H. (1988). Duality of I-projections and maximum likelihood estimation for log-linear models under cone constraints, *J. Amer. Statist. Assoc.*, **402**, 546–554.

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

Hestenes, M. R. (1975). *Optimization Theory*, Wiley, New York.

Jaynes, E. T. (1957). Information theory and statistical mechanics, *Phys. Rev.*, **106**, 620–630.

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

Norušis, M. J. (1988). *SPSSX Advanced Statistics Guide*, 2nd ed., McGraw-Hill, New York.

Patefield, W. M. (1982). Exact tests for trends in ordered contingency tables, *J. Roy. Statist. Soc. Ser. C*, **31**, 32–43.

Rao, C. R. (1965). *Linear Statistical Inference and Its Applications*, Wiley, New York.
Rockafellar, R. T. (1970). *Convex Analysis*, Princeton University Press, New York.
Sanov, I. N. (1957). On the probability of large deviations of random variables, *Mat. Sb.*, **42**, 11–44
Winkler, W. (1990). On Dykstra's iterative proportional fitting procedure, *Ann. Probab.*, **18**, 1410–1415.