The Annals of Statistics
1995, Vol. 23, No. 4, 1160–1174

CONVERGENCE OF THE ITERATIVE PROPORTIONAL FITTING PROCEDURE

By Ludger Rüschendorf

Institut für Mathematische Stochastik, University of Freiburg

The iterative proportional fitting procedure (IPFP) was introduced in 1940 by Deming and Stephan to estimate cell probabilities in contingency tables subject to certain marginal constraints. Its convergence and statistical properties have been investigated since then by several authors and by several different methods. A natural extension of the IPFP to the case of bivariate densities has been introduced by Ireland and Kullback. It has been conjectured that also in the general case the IPFP converges to the minimum discrimination projection on the class of distributions with given marginals. We verify this conjecture under some regularity conditions.

1. Introduction. The adjustment of distributions to a priori known marginals is a problem which has received a lot of attention from a variety of probabilists and statisticians, and it has been realized that this problem has some fundamental applications in various fields. The iterative proportional fitting procedure (IPFP) is an algorithm to construct approximatively an adjustment of this kind. It was introduced by Deming and Stephan (1940) and it has been studied mainly in the finite discrete case. Its justification is closely related to the Kullback–Leibler distance I, and it has been proved that the IPFP converges to the I-projection on the set of distributions with fixed marginals in the finite discrete case by several authors, including Brown (1959), Bishop (1967), Sinkhorn (1967), Kullback (1968), Ireland and Kullback (1968), Bishop and Fienberg (1969), Fienberg (1970), Haberman (1974, 1984) and Csiszar (1975).

Generalizations to the continuous case of the IPFP have been introduced by Ireland and Kullback (1968) and Kullback (1968) and, with modified information on marginal moments, by Haberman (1984). However, a convergence proof has remained an open problem since then.

The IPFP is analogous to the alternating projection algorithm (also called the Gauss–Seidel algorithm or backfitting algorithm), which is one of the fundamental numerical algorithms for solving systems of equations. While the IPFP leads to multiplicative approximations, the alternation algorithm produces additive approximations. It has been studied extensively after its introduction by von Neumann (1950) and Aronszajn (1950) in the framework of Hilbert spaces, and it has been extended to sup-norm and L^p-norm

Received April 1993; revised January 1995.

AMS 1991 *subject classifications*. 60E05, 62B10.

Key words and phrases. Iterative proportional fitting, I-projection, distributions with given marginals, Kullback–Leibler distance, marginal adjustment.

approximation in spaces of continuous functions and L^p-integrable functions, respectively, by Diliberto and Strauss (1951) and many others [cf. the survey in Light and Cheney (1985)].

Both algorithms have found important applications in several fields, such as contingency tables [cf. Haberman (1974)], tomography [cf. Hamaker and Solmon (1978)], in ridge-type regressions models [ACE, cf. Breiman and Friedman (1985), Stone (1985) and Buja, Hastie and Tibshirani (1989)], in connection with Hoeffding's decomposition [cf. Rüschendorf (1985)], restricted least squares estimation [cf. Dykstra (1983), page 838], probabilistic expert systems [cf. Jirousek (1991)] and many others.

The IPFP constructs, in the limit to a probability measure μ, a closest probability measure ν with given marginals ν_1 and ν_2. In the case that μ is unknown there is also a statistical version of this problem: based on a sequence of data $(X_1, Y_1), \dots, (X_n, Y_n)$ with distribution μ, estimate the closest probability ν with marginals ν_i. A natural procedure is to use a (kernel-) estimate $\hat{\mu}_n$ of μ and then project $\hat{\mu}_n$ to the set of distributions with marginals ν_1 and ν_2. Statistical properties of these estimators (consistency, asymptotic normality and efficiency) have been considered in the literature mainly in the discrete case [cf. Haberman (1974)]. In the continuous case an adjustment to marginal moments and its statistical properties were discussed in Haberman (1984).

Use of the Kullback–Leibler distance has been justified in the literature from several viewpoints; for some general arguments see Good (1963). Its use for multiplicative approximations of densities and density estimation, respectively, is discussed in Huber (1985) and Friedman, Stuetzle and Schroeder (1984). Schrödinger bridges have a justification in terms of a large-deviation formula including the I-projection on a marginal class [cf. Rüschendorf and Thomsen (1993)]. It is well known that the Kullback–Leibler distance appears in asymptotics of maximum likelihood tests under a departure from the correct model and is related to maximum likelihood estimators in log-linear models [cf. Haberman (1974)].

The aim of this paper is to prove convergence of the IPFP to the I-projection on the set of distributions with fixed marginals. It is interesting to note that, for the related problem of determining the minimum distance between two convex sets of probability measures w.r.t. I, convergence of the corresponding alternation algorithm has been proved in great generality in an interesting paper by Csiszar and Tusnady (1984) generalizing the classical result of Cheney and Goldstein (1959) in metric spaces. Note that the set of distributions with fixed marginals is the intersection of convex sets where one marginal is fixed. In spite of this similarity the convergence of the IPFP remained an open problem for a long time. For the proof of convergence we use geometric properties of I established in Kullback (1959) and Csiszar (1975) together with some intermediate consideration of a weaker topology, namely, the topology of setwise convergence (τ-topology). The case of unknown μ and the statistical properties of the estimators in this case are not considered in this paper.

2. The iterative proportional fitting procedure. Let $(E_i, \mathscr{A}_i, \mu_i)$, $i = 1,2$, be two probability spaces; let $\mu \in M(\mu_1, \mu_2)$ be the class of p-measures on the product $(E, \mathscr{A}) = \otimes(E_i, \mathscr{A}_i)$ with marginals μ_i, $i = 1,2$; and let $\nu_i \in M^1(E_i, \mathscr{A}_i)$ be probability measures continuous w.r.t. μ_i and with densities $r_i = d\nu_i/d\mu_i$, $i = 1,2$. Assume finally that $\mu \ll \mu_1 \otimes \mu_2$ with density $h = d\mu/(d\mu_1 \otimes \mu_2)$. In this case we shall also use the notation $\mu = h\mu_1 \otimes \mu_2$.

The aim of the IPFP is to find in the limit an I-projection of μ on $M(\nu_1, \nu_2)$, that is, a closest element of $M(\nu_1, \nu_2)$ to μ w.r.t. Kullback–Leibler distance

$$I(\nu|\mu) = \int \ln \frac{d\nu}{d\mu}\, d\nu.$$

For this purpose we assume generally in this paper (without mentioning it further) that the following holds:

(A0) $\inf\{I(\nu|\mu);\ \nu \in M(\nu_1, \nu_2)\} < \infty$.

Under assumption (A0), Rüschendorf and Thomsen (1993) proved that a unique I-projection $\nu^* \in M(\nu_1, \nu_2)$ exists and is of the form

$$\nu^* = a(x)b(y)\mu, \tag{2.1}$$

for some nonnegative functions $a = a(x) \geq 0$, $b = b(y) \geq 0$. Furthermore, a and b are uniquely determined by the Schrödinger equations [arising from the marginal condition $\nu^* \in M(\nu_1, \nu_2)$]:

$$\begin{aligned} a(x)\int h(x,y)b(y)\mu_2(dy) &= r_1(x)[\,\mu_1], \\ b(y)\int h(x,y)a(x)\mu_1(dx) &= r_2(y)[\,\mu_2]. \end{aligned} \tag{2.2}$$

From assumption (A0) we conclude that

$$I(\nu^*|\mu) = \int (\ln a + \ln b)\, d\nu^* < \infty, \tag{2.3}$$

which implies that

$$\ln a + \ln b \in \mathscr{L}^1(\nu^*). \tag{2.4}$$

At this point it is natural to make the following assumption:

(A1) $F = \mathscr{L}^1(\nu_1) \oplus \mathscr{L}^1(\nu_2) \subset \mathscr{L}^1(\nu^*)$ is closed.

By Kober's criterion for closedness of sum spaces, assumption (A1) is equivalent to

$$\int |f \oplus g|\, d\nu^* \geq c \int |f|\, d\nu_1, \tag{2.5}$$

for some $c > 0$ and all $f \oplus g \in F$, $f \oplus g(x,y) = f(x) + g(y)$ [cf. Rüschendorf and Thomsen (1993)]. In particular, (A1) implies that

$$\ln a \in L^1(\nu_1), \qquad \ln b \in L^1(\nu_2). \tag{2.6}$$

The I-projection in (2.1) is the unique element ν^* in $M(\nu_1, \nu_2)$ which has a product density $a(x)b(y)$ w.r.t. μ. The aim of the IPFP is to construct the I-projection ν^* by alternately matching one marginal distribution to the target marginals ν_1 and ν_2, that is, we construct a sequence $\mu^{(0)}, \mu^{(1)}, \mu^{(2)}, \dots$ such that all elements in this sequence have a product density w.r.t. μ, and $\mu^{(0)}, \mu^{(2)}, \dots$ match the first marginal ν_1, while $\mu^{(1)}, \mu^{(3)}, \dots$ have fitted the second marginal ν_2. [Equivalently, we approximate solutions of equations (2.2) by iteratively solving the first equation of (2.2) in a and then the second equation of (2.2) in b, starting with some initial functions a_0 and b_0.

More explicitly, the IPFP is defined by the following recursion:

$$\begin{aligned} b_0 &= 1, \qquad a_0 = r_1 = \frac{d\nu_1}{d\mu_1}, \\ b_1(y) &= \frac{r_2(y)}{\int h(x,y)a_0(x)\mu_1(dx)}, \\ a_1(x) &= \frac{r_1(x)}{\int h(x,y)b_1(y)\mu_2(dy)} \end{aligned} \tag{2.7}$$

and, generally,

$$\begin{aligned} b_{n+1}(y) &= \frac{r_2(y)}{\int h(x,y)a_n(x)\mu_1(dx)}, \\ a_{n+1}(x) &= \frac{r_1(x)}{\int h(x,y)b_{n+1}(y)\mu_2(dy)}. \end{aligned} \tag{2.8}$$

Note that the recursion is well defined a.s. since $a_0 = r_1 > 0[\nu_1]$ and $h > 0[\mu]$, and, therefore, b_1 is well defined and $b_1 \approx r_2$ (i.e., b_1 and r_2 have the same support). This yields by induction that $a_i \approx r_1$, $b_i \approx r_2$, for all i. Define the sequence of probability measures

$$\mu^{(2n)} := a_n \otimes b_n \mu, \qquad \mu^{(2n+1)} := a_n \otimes b_{n+1} \mu, \qquad n \geq 0, \tag{2.9}$$

where $a_n \otimes b_n(x,y) = a_n(x)b_n(y)$ and define the marginal densities of $a(x)b(y)\mu(dx,dy)$ by

$$\begin{aligned} R(x,a,b) &= a(x)\int h(x,y)b(y)\mu_2(dy), \\ C(y,a,b) &= b(y)\int h(x,y)a(x)\mu_1(dx). \end{aligned} \tag{2.10}$$

Then by definition we obtain for the marginals $\mu_i^{(n)} = \pi_i(\mu^{(n)})$, where π_i are the projections on the ith component,

$$\begin{aligned} \mu_1^{(2n)} &= \pi_1(\mu^{(2n)}) = R(\cdot, a_n, b_n)\mu_1 = r_1\mu_1 = \nu_1, \\ \mu_2^{(2n+1)} &= \nu_2, \\ \mu_1^{(2n+1)} &= q_{2n+1}\mu_1, \\ \mu_2^{(2n)} &= p_{2n}\mu_2, \end{aligned} \tag{2.11}$$

where $q_{2n+1} = a_n \int b_{n+1}(y)h(\cdot, y)\mu_2(dy) = R(\cdot, a_n, b_{n+1})$ and $p_{2n} = b_n \int a_n(x)h(x, \cdot)\mu_1(dx) = C(\cdot, a_n, b_n)$. This implies in particular that $\mu^{(2n)}$ has correct first marginal ν_1, while $\mu^{(2n+1)}$ has correct second marginal ν_2.

From the construction it is clear that $\mu^{(0)}$ is the I-projection of μ on $M(\nu_1)$, the measures with (correct) first marginal ν_1, $\mu^{(1)}$ is the I-projection of $\mu^{(0)}$ on $M(\nu_2)$, $\mu^{(2)}$ is the I-projection of $\mu^{(1)}$ on $M(\nu_1)$ and so on.

In the finite discrete case $E_1 = \{1, \dots, n\}$, $E_2 = \{1, \dots, m\}$, $\mu\{(i, j)\} = \mu_{i,j}$, $\nu\{(i, j)\} = \nu_{i,j}$, the problem and the IPFP-algorithm specializes to the familiar case of contingency tables. If $h = 1$ (i.e., $\mu = \mu_1 \otimes \mu_2$), then $\mu^{(0)} = r_1 \mu$, $\mu^{(1)} = r_1 \otimes r_2 \mu$ and so we are done already after two steps. For examples like $\mu_i = \nu_i$ the uniform distribution on $[0, 1]$ and $h_1(x, y) = c_1 e^{-xy}$, $h_2(x, y) = c_2(1 + xy)$, we see using Maple or Mathematica that after about six steps the marginal densities are close to the uniform distribution (cf. the final remarks in Section 4). Note, however, some critical remarks in Haberman (1974) indicating that the "speed" of convergence may be slow already in the discrete case.

It is relatively easy to see that the marginals of $\mu^{(n)}$ converge to the correct ones.

PROPOSITION 2.1. *The following hold*:

(a) $I(\mu_1^{(n)}|\nu_1) \to 0$, $I(\mu_2^{(n)}|\nu_2) \to 0$;

(b) $\|\mu_1^{(n)} - \nu_1\| \to 0$, $\|\mu_2^{(n)} - \nu_2\| \to 0$, *where* $\| \ \|$ *is the total variation distance.*

PROOF. Since $d\mu^{(2n)}/d\mu^{(2n-1)} = a_n/a_{n-1}$,

$$I(\mu^{(2n)}|\mu^{(2n-1)}) = \int \ln \frac{a_n}{a_{n-1}} \, d\mu^{(2n)} = \int \ln \frac{a_n}{a_{n-1}} \, d\nu_1. \tag{2.12}$$

Similarly, $d\mu^{(2n+1)}/d\mu^{(2n)} = b_{n+1}/b_n$ and

$$I(\mu^{(2n+1)}|\mu^{(2n)}) = \int \ln \frac{b_{n+1}}{b_n} \, d\mu^{(2n+1)} = \int \ln \frac{b_{n+1}}{b_n} \, d\nu_2. \tag{2.13}$$

From the "Pythagorean law" for the Kullback–Leibler distance I [cf. Csiszar (1975), Theorems 2.3 and 3.14], we obtain

$$I(\nu^*|\mu) = I(\nu^*|\mu^{(n)}) + \sum_{i=0}^{n} I(\mu^{(i)}|\mu^{(i-1)}), \tag{2.14}$$

where $\mu^{(-1)} := \mu$ and ν^* is the I-projection of μ on $M(\nu_1, \nu_2)$.

(a) From (2.14) we obtain that

$$\sum_{i=0}^{\infty} I(\mu^{(i)}|\mu^{(i-1)}) \le I(\nu^*|\mu) < \infty, \tag{2.15}$$

and, therefore,

$$I(\mu^{(i)}|\mu^{(i-1)}) \to 0. \tag{2.16}$$

This implies by the monotonicity theorem [cf. Liese and Vajda (1987) Corollary 1.29] that

$$I(\mu_1^{(2n+1)}|\nu_1) = I(\pi_1(\mu^{(2n+1)})|\pi_1(\mu^{(2n)})) \le I(\mu^{(2n+1)}|\mu^{(2n)}) \to 0.$$

Similarly, $I(\mu_2^{(2n)}|\nu_2) \to 0$.

(b) This follows from part (a) and the well-known inequality

$$\|P - Q\| \le (2I(P|Q))^{1/2}. \tag{2.17}$$ □

Remark. From Proposition 2.1 it is clear that $I(\nu|\mu^{(n)}) \to 0$ for some $\nu \in M^1(E, \mathscr{A})$ implies that $\nu \in M(\nu_1, \nu_2)$, and from the Pythagorean law

$$I(\nu|\mu^{(n)}) = I(\nu|\nu^*) + I(\nu^*|\mu^{(n)}) \tag{2.18}$$

(ν^* is the I-projection of $\mu^{(n)}$ too!); therefore, $\nu = \nu^*$. So the problem of convergence of $\mu^{(n)}$ to the I-projection is equivalent to proving convergence of $\mu^{(n)}$. At this point the paper of Kullback (1968) is incomplete, since his formula (2.30) only yields (in our terminology)

$$\|\mu^{(N+m)} - \mu^{(N)}\| \to 0 \quad \forall\, m, \text{as } N \to \infty, \tag{2.19}$$

which is not enough to imply convergence

$$\|\mu^{(n)} - \nu\| \to 0 \quad \text{for some } \nu \in M^1(E, \mathscr{A}). \tag{2.20}$$

[For this remark cf. also the paper of Csiszar (1975).]

3. Convergence of the IPFP. Convergence of the IPFP is established in this section under various conditions. We introduce the following boundedness assumption on h and ν_i:

(B1) For some $0 < c < \infty$, $\int h(x, y)\nu_1(dx) \ge c r_2(y)[\mu_2]$.

Condition (B1) is satisfied if, for example, for some function u_1 with $0 < c := \int u_1\, d\nu_1 < \infty$,

$$h(x, y) \ge u_1(x) r_2(y)[\mu]. \tag{3.1}$$

In particular, (B1) is satisfied if

$$h/r_2 \ge c > 0[\mu]. \tag{3.2}$$

Theorem 3.1. *If condition* (B1) *holds and if* (ln b_n) $\subset L^1(\nu_2)$ *is uniformly integrable, then the following hold*:

(a) $I(\mu^{(n)}|\mu) \to I(\nu^*|\mu)$;

(b) $\|\mu^{(n)} - \nu^*\| \to 0$ *and* $I(\nu^*|\mu^{(n)}) \to 0$.

The uniform integrability condition on (ln b_n) $\subset L^1(\nu_2)$ is ensured by the following conditions:

(B2) $0 < c \le a/r_1 \le C < \infty$ for some $0 < c < C < \infty$;

(B3) $0 < c \le h(x, y)/r_2(y) \le C < \infty$ for some $0 < c < C < \infty$.

PROPOSITION 3.2.

(a) *Condition* (B3) *implies* (B2).

(b) *The conditions* $\ln b \in L^1(\nu_2)$ *and* (B2) *imply that* $(\ln b_n) \subset L^1(\nu_2)$ *is uniformly integrable.*

Note that by (2.6) the condition $\ln b \in L^1(\nu_2)$ is implied by condition (A1). A direct criterion for condition (A1) is given in the following proposition (for some further criteria cf. Rüschendorf and Thomsen (1993)], using the following condition:

(B4) $h(x, y) \geq c(r_1(x)/a(x))v(y)[\mu_1 \otimes \mu_2]$, for some function $v \geq 0$ with $\int v \, d\mu_2 > 0$.

PROPOSITION 3.3.

(a) *Condition* (B4) *implies* (A1).

(b) *Condition* (B3) *implies* (B4) *and* (B1).

Altogether, we have established the following sequence of relations:

$$(\mathrm{B3}) \to (\mathrm{B4}) \to (\mathrm{A1}) \to (\ln b \in L^1);$$
$$(\mathrm{B3}) \to (\mathrm{B1}), (\mathrm{B2}) \text{ and } (\mathrm{B3}) \to ((\ln b_n) \text{ is u.i.}).$$

As corollary we obtain the following convergence result.

COROLLARY 3.4. *Under condition* (B3) *or under conditions* (A1), (B1) *and* (B2), *the IPFP* $(\mu^{(n)})$ *is convergent to the I-projection* ν^* *in total variation and also* $I(\nu^* | \mu^{(n)}) \to 0$.

A modification in the steps of the proofs allows us to obtain a convergence result under a condition weaker than (B3). The statement of Theorem 3.1 and the following propositions remains useful, since they allow us to modify the assumptions and are also used in the proof of the following theorem.

THEOREM 3.5. *If* $h/r_2 \geq c > 0[\mu]$, *then the IPFP* $(\mu^{(n)})$ *converges to the I-projection* ν^* *in total variation and also* $I(\nu^* | \mu^{(n)}) \to 0$.

4. Proofs and final remarks. For the proof of Theorem 3.1 we need to bound the "variation" of b_n and a_n, respectively.

LEMMA 4.1. *For all $n \in \mathbb{N}$, the following hold*:

(a) $\int \ln a_n \, d\nu_1 + \int \ln b_n \, d\nu_2 \le I(\nu^*|\mu) < \infty$;
(b) $\int (\ln ab - \ln a_n b_n) \, d\nu^* \le I(\nu^*|\mu) < \infty$;
(c) *under condition* (A1),

$$\sup_n \int |\ln b_n| \, d\nu_2 < \infty, \qquad \sup_n \int |\ln a_n| \, d\nu_1 < \infty. \tag{4.1}$$

PROOF. (a) From (2.14) we conclude that

$$\sum_{i=0}^{n} I(\mu^{(i)}|\mu^{(i-1)}) \le I(\nu^*|\mu) \quad \forall\, n \in \mathbb{N}. \tag{4.2}$$

Therefore,

$$\begin{aligned}\sum_{i=0}^{2n} I(\mu^{(i)}|\mu^{(i-1)}) &= \sum_{i=0}^{n} \left[\int \ln \frac{a_i}{a_{i-1}} \, d\nu_1 + \int \ln \frac{b_i}{b_{i-1}} \, d\nu_2 \right] \\ &= \int \ln a_n \, d\nu_1 + \int \ln b_n \, d\nu_2 \\ &\le I(\nu^*|\mu) < \infty.\end{aligned}$$

(b) This follows from (2.14).

(c) From (2.14), $\ln ab \in \mathscr{L}^1(\nu^*)$ (ab is synonymous for $a \otimes b$), and from part (b) we conclude that $\ln ab - \ln a_n b_n \in \mathscr{L}^1(\nu^*)$ and $\int |\ln ab - \ln a_n b_n| \, d\nu^* \le I(\nu^*|\mu) + c_0$, for some constant c_0 [cf. the corresponding argument in Liese (1975)]. Therefore,

$$\begin{aligned}\int |\ln a_n b_n| \, d\nu^* &\le \int |\ln ab| \, d\nu^* + \int |\ln ab - \ln a_n b_n| \, d\nu^* \\ &\le 2I(\nu^*|\mu) + 2c_0.\end{aligned} \tag{4.3}$$

From assumption (A1) we conclude that

$$\int |\ln a_n| \, d\nu_1 \le c\big(I(\nu^*|\mu) + c_0\big),$$

$$\int |\ln b_n| \, d\nu_1 \le c\big(I(\nu^*|\mu) + c_0\big). \qquad \square$$

LEMMA 4.2. *For all $n \in \mathbb{N}$, the following hold*:

(a) $I(\mu^{(2n)}|\mu) = \int \ln a_n \, d\nu_1 + \int \ln b_n (b_n / b_{n+1}) \, d\nu_2$.
(b) $I(\mu^{(2n+1)}|\mu) = \int \ln a_n (a_n / a_{n+1}) \, d\nu_1 + \int \ln b_n \, d\nu_2$.

PROOF.

$$I(\mu^{(2n)}|\mu) = \int \ln(a_n b_n)\, d\mu^{(2n)} = \int \ln a_n \, d\mu_1^{(2n)} + \int \ln b_n \, d\mu_2^{(2n)}$$

$$= \int \ln a_n \, d\nu_1 + \int \ln b_n \frac{b_n}{b_{n+1}} \, d\nu_2,$$

by (2.12) and (2.13).

(b) The proof of part (b) is similar. □

Next we have

LEMMA 4.3. *Under condition* (B1) *the following hold* ν_i*-a.s.*:

$$\frac{d\mu_1^{(2n-1)}}{d\nu_1} \geq \frac{1}{c}, \qquad \frac{d\mu_2^{(2n)}}{d\nu_2} \leq c \quad \textit{for all } n. \tag{4.4}$$

PROOF. Condition (B1) is equivalent to $b_0/b_1 \leq c$ a.s. Observe that

$$\frac{b_n}{b_{n+1}} = \frac{\int h a_n \, d\mu_1}{\int h a_{n-1} \, d\mu_1} \quad \text{and} \quad \frac{a_{n-1}}{a_n} = \frac{\int h b_n \, d\mu_2}{\int h b_{n-1} \, d\mu_2}.$$

Therefore,

$$\frac{a_1}{a_0} = \frac{\int h b_1(b_0/b_1) \, d\mu_2}{\int h b_1 \, d\mu_2} \leq c \quad \text{and} \quad \frac{b_1}{b_2} = \frac{\int h a_0(a_1/a_0) \, d\mu_1}{\int h a_0 \, d\mu_1} \leq c.$$

By induction we obtain, for all $n \in \mathbb{N}$,

$$\frac{b_n}{b_{n+1}} \leq c, \qquad \frac{a_n}{a_{n-1}} \leq c \quad \text{a.s.} \tag{4.5}$$

This implies, by (2.10),

$$\frac{d\mu_1^{(2n-1)}}{d\mu_1^{(2n)}} = \frac{d\mu_1^{(2n-1)}}{d\nu_1} = \frac{a_{n-1}}{a_n} \geq \frac{1}{c}$$

and

$$\frac{d\mu_2^{(2n)}}{d\mu_2^{(2n+1)}} = \frac{d\mu_2^{(2n)}}{d\nu_2} = \frac{b_n}{b_{n+1}} \leq c. \qquad \square$$

As a consequence of condition (B1) we formulate the following crucial uniform integrability property.

LEMMA 4.4. *If condition* (B1) *holds and if* $\sup_n \int |\ln b_n| \, d\nu_2 < \infty$, *then* $(a_n \otimes b_n) \subset L^1(\mu)$ *is uniformly integrable.*

PROOF. Consider the continuous, convex function $\varphi(x) = x \ln x$. Since $\lim_{x \to \infty}(\varphi(x)/x) = \infty$, by the criterion of Valle Poussain [in the modified form of Liese (1975)] it is enough to establish that

$$\sup_n \int \varphi(a_n b_n)\, d\mu < \infty. \tag{4.6}$$

From Lemma 4.2 we have

$$\begin{aligned}\int \varphi(a_n b_n)\, d\mu &= I(\mu^{(2n)}|\mu)\\ &= \int \ln a_n\, d\nu_1 + \int \ln b_n\, d\nu_2 + \int \ln b_n \left(\frac{b_n}{b_{n+1}} - 1 \right) d\nu_2\\ &\le I(\nu^*|\mu) + (c+1) \int |\ln b_n|\, d\nu_2\\ &\le I(\nu^*|\mu) + (c+1) \sup_n \int |\ln b_n|\, d\nu_2. \end{aligned}$$

□

Now we are able to prove Theorem 3.1.

PROOF OF THEOREM 3.1. (a) Uniform integrability of $(\ln b_n) \subset L^1(\nu_2)$ implies, by Lemma 4.4, uniform integrability of $(a_n \otimes b_n) \subset L^1(\mu)$. Therefore, the sequence $(\mu^{(2n)}) = (a_n b_n \mu)$ is relatively compact in the τ-topology [i.e., with respect to the weak topology $\sigma(L^1(\mu), L^\infty(\mu))$]. Since $I(\mu^{(2n+1)}|\mu^{(2n)}) \to 0$ this implies that also $(\mu^{(n)})$ is relatively τ-compact and τ-sequentially compact.

Let $(\mu^{(m)})$ be a τ-convergent subsequence, $\mu^{(m)} \to_\tau \nu$. We can assume w.l.o.g. that $(m) \subset 2\mathbb{N}$. Then from the lower semicontinuity of I w.r.t. τ-convergence we obtain

$$\begin{aligned} I(\nu|\mu) &\le \liminf I(\mu^{(m)}|\mu)\\ &\le \limsup I(\mu^{(m)}|\mu)\\ &= \limsup \int \ln(a_k b_k) a_k b_k\, d\mu, \qquad m = 2k. \end{aligned} \tag{4.7}$$

Since

$$\frac{d\nu_2}{d\mu_2^{(2n)}} = \frac{d\mu_2^{(2n+1)}}{d\mu_2^{(2n)}} = \frac{d\mu^{(2n+1)}}{d\mu^{(2n)}} = \frac{b_{n+1}}{b_n}, \tag{4.8}$$

we conclude (from the proof of) Proposition 2.1,

$$I(\nu_2|\mu_2^{(2n)}) = \int \ln \frac{b_{n+1}}{b_n}\, d\nu_2 \to 0.$$

Therefore,

$$\int\left|\frac{d\mu_2^{(2n)}}{d\nu_2}-1\right|d\nu_2=\int\left|\frac{b_{n+1}}{b_n}-1\right|d\nu_2\to 0,$$

which implies that $b_{n+1}/b_n\to 1$, ν_2-stochastically.

Uniform integrability of $(\ln b_n)\subset L^1(\nu_2)$ and boundedness of b_k/b_{k+1} implies

$$\limsup\int \ln b_k\frac{b_k}{b_{k+1}}\,d\nu_2=\limsup\int \ln b_k\,d\nu_2. \tag{4.9}$$

From (4.7) and Lemmas 4.1 and 4.2, therefore, it follows that

$$I(\nu|\mu)\le\limsup\left(\int \ln a_k\,d\nu_1+\int \ln b_k\,d\nu_2\right)\le I(\nu^*|\mu). \tag{4.10}$$

This implies, by the uniqueness of the I-projection, that $\nu=\nu^*$. So ν^* is the only limit point of $(\mu^{(n)})$ in the τ-topology and, therefore, $\mu^{(n)}\to_\tau \nu^*$. By (4.7) and (4.10) we obtain $I(\mu^{(n)}|\mu)\to I(\nu^*|\mu)$.

(b) Since $I(\mu^{(n)}|\mu)\to I(\nu^*|\mu)=\inf\{I(\nu|\mu);\ \nu\in M(\nu_1,\nu_2)\}$, we conclude from the proof of Theorem 2.1 in Csiszar (1975) that $\|\mu^{(n)}-\nu^*\|\to 0$. From

$$I(\mu^{(2n)}|\mu)=\int \ln a_n\,d\nu_1+\int \ln b_n\frac{b_n}{b_{n+1}}\,d\nu_2\to I(\nu^*|\mu)$$

and

$$\left|\int \ln b_n\left(\frac{b_n}{b_{n+1}}-1\right)d\nu_2\right|\to 0,$$

we finally obtain, using (2.14), that $I(\nu^*|\mu^{(n)})\to 0$. □

PROOF OF PROPOSITION 3.2. (a) Under condition (B3),

$$\frac{r_1}{a}=\int bh\,d\mu_2=\int b\frac{h}{r_2}\,d\nu_2\begin{cases}\le C\int b\,d\nu_2,\\ \ge c\int b\,d\nu_2.\end{cases} \tag{4.11}$$

On the other hand,

$$\int b\,d\nu_2\le\frac{1}{c}\int b\frac{h}{r_2}\,d\nu_2=\frac{1}{c}\int bh\,d\mu_2=\frac{1}{c}\frac{r_1}{a} \tag{4.12}$$

and $\int b\,d\nu_2\ge(1/C)(r_1/a)$, that is, $0<\int b\,d\nu_2<\infty$. This proves part (a).

(b) From $a_0=r_1$ we obtain, by (2.2),

$$\frac{b_1}{b}=\frac{\int ha\,d\mu_1}{\int ha_0\,d\mu_1}=\frac{\int ha_0(a/a_0)\,d\mu_1}{\int ha_0\,d\mu_1}\quad\text{and}\quad\frac{a_1}{a}=\frac{\int hb_1(b/b_1)\,d\mu_2}{\int hb_1\,d\mu_2}.$$

Therefore, $c \le b_1/b \le C$ and $1/C \le a_1/a \le 1/c$. By induction this implies, for all $n \in \mathbb{N}$,

$$c \le \frac{b_n}{b} \le C, \qquad \frac{1}{C} \le \frac{a_n}{a} \le \frac{1}{c}. \tag{4.13}$$

Therefore, $|\ln b_n| \le |\ln b| + c'$ and $(\ln b_n)$ is uniformly integrable. □

PROOF OF PROPOSITION 3.3. (a) Consider any $f \in \mathscr{L}^1(\nu_1)$ and $g \in \mathscr{L}^1(\nu_2)$ with the normalization that the median of f is zero, $\operatorname{med}_{\nu_1} f = 0$. Then

$$\begin{aligned}\int |f(x) + g(y)|\, d\nu^* &\ge c \int \left[\int |f(x) + g(y)| b(y) v(y) \mu_2(dy) \right] r_1(x) \mu_1(dx) \\ &= c \int \left[\int |f(x) + g(y)| r_1(x) \mu_1(dx) \right] b(y) v(y) \mu_2(dy) \\ &\ge c' \int |f|\, d\nu_1.\end{aligned}$$

The last inequality follows from the assumption that $\operatorname{med}_{\nu_1} f = 0$. This inequality implies closedness of F by Kober's criterion [cf. Rüschendorf and Thomsen (1993)].

(b) Condition (B3) implies condition (B2) (i.e., c $\le h/r_1 \le C$) using Proposition 3.2. Furthermore, with $v(y) = r_2(y)$, $\int v\, d\mu_2 = 1$ and condition (B4) holds. From (3.2) we finally infer (B1) using condition (B3). □

PROOF OF THEOREM 3.5. From $h/r_2 \ge c$ we conclude (as in Proposition 3.2) that $r_1/a \ge c \int b\, d\nu_2 =: 1/C > 0$ (since $b > 0[\nu_2]$). This implies (as in Proposition 3.2) that $b_n \le Cb \;\forall\; n \in \mathbb{N}$. Since $\int b\, d\nu_2 < \infty$ we conclude that $(b_n) \subset L^1(\nu_2)$ is uniformly integrable.

We then use the following lemma, which is an extension of Lemma 3.1 of Csiszar (1975).

LEMMA 4.5. *Let (f_n) be measurable; let $(e^{|f_n|}) \subset L^1(Q)$ be uniformly integrable; and let $I(P_n|Q) \to 0$. Then $|\int f_n\, dP_n - \int f_n\, dQ| \to 0$.*

PROOF. We follow the idea of the proof of Csiszar (1975). Let $p_n = dP_n/dQ$ and $A_n = A_{n,K} = \{|f_n| \le K\}$. Then $\|P_n - Q\| \to 0$ and, therefore, $|\int_{A_n} f_n\, d(P_n - Q)| \to 0$. Since $(|f_n|) \subset L^1(Q)$ is uniformly integrable, it is sufficient to show that, for any $\varepsilon > 0$,

$$\limsup \int_{A_n^c} |f_n|\, dP_n = \limsup \int_{A_n^c} |f_n| p_n\, dQ < \varepsilon \quad \text{for some } K = K(\varepsilon).$$

From the inequality $ab < a \log a + e^b$, $a, b \ge 0$, we infer

$$\int_{A_n^c} |f_n| p_n\, dQ \le \int_{A_n^c} p_n \ln p_n\, dQ + \int_{A_n^c} e^{|f_n|}\, dQ.$$

Furthermore, $\limsup \int_{A_n^c} e^{|f_n|}\,dQ < \varepsilon/2$ for $K \geq K(\varepsilon)$ by uniform integrability of $(e^{|f_n|})$, and, finally, $\lim \int_{A_n^c} p_n \ln p_n\,dQ = 0$ by Fatou's lemma and the assumption that $\int p_n \ln p_n\,dQ \to 0$ [cf. the argument in Csiszar (1975)]. □

Continuing the proof of Theorem 3.5, we apply Lemma 4.4 to $f_n = \ln b_n$, $P_n = \mu_2^{(2n)}$, $Q = \nu_2$. By our assumption, $\exp|f_n| = \exp|\ln b_n| = \exp(\ln_+ b_n)\exp(\ln_- b_n) = |b_n \vee 1||b_n \wedge 1| = |b_n|$ is uniformly integrable. Therefore,

$$\left|\int \ln b_n\,d\mu_2^{(2n)} - \int \ln b_n\,d\nu_2\right| \to 0.$$

This implies that (cf. the proof of Proposition 3.2)

$$\begin{aligned}\limsup I(\mu^{(2n)}|\mu) &= \limsup\left[\int \ln a_n\,d\nu_1 + \int \ln b_n\,d\mu_2^{(2n)}\right]\\ &= \limsup\left[\int \ln a_n\,d\nu_1 + \int \ln b_n\,d\nu_2\right] \leq I(\nu^*|\mu) < \infty.\end{aligned}$$

In particular, $(a_n b_n) \subset L^1(\mu)$ is uniformly integrable and (as in the proof of Theorem 3.1) any limit point of $(\mu^{(n)})$ is ν^*, that is, $\mu^{(n)} \to \nu^*$ in total variation and $I(\nu^*|\mu^{(n)}) \to 0$. □

Remarks.

(a) Some simple examples can be calculated using Maple or Mathematica. Consider the case $\mu_i = \nu_i$ the uniform distribution on $[0,1]$ and the density $h(x,y) = \frac{4}{5}(1 + xy)$, that is, we want to fit the measure $\mu = h\mu_1 \otimes \mu_2$ with density h to uniform marginals. Then the marginals of μ have Lebesgue densities $h_1(x) = \frac{4}{5}(1 + x/2)$ and $h_2(y) = \frac{4}{5}(1 + y/2)$.

We obtain the following for the marginal densities of the IPFP:

$$f^{(3)}(x) = \frac{2.14307 + 1.00312x}{2.14453 + x}, \qquad f^{(4)}(y) = \frac{2.136 + 0.999825y}{2.13592 + y},$$

$$f^{(5)}(x) = \frac{2.1364 + 1.00001x}{2.1364 + x}, \qquad f^{(6)}(y) = \frac{2.1638 + 0.999999y}{2.13639 + y};$$

and we obtain the following for the joint density:

$$f^{(5)}(x,y) = 5.56419\frac{1 + xy}{(2.1364 + x)\cdot(2.13639 + y)},$$

$$f^{(6)}(x,y) = 5.56411\frac{1 + xy}{(2.13638 + x)\cdot(2.13638 + y)}.$$

So after few steps one has a good approximation to the joint density, and the marginals are nearly identical to the target marginals equal to constant 1.

For the calculation of complicated examples of the IPFP in the continuous case one has to use numerical integration or simulation since the number of terms in exact integration formulas explodes after few terms. Simulation methods are used also in the related projection pursuit algorithm and its implementation (in S-PLUS).

Our convergence result suggests that it should be possible also to use discrete approximations to continuous densities, that is, to do the calculations in an approximate contingency table. For contingency tables, examples and empirical evidence of convergence are reported in Haberman (1974).

(b) Some extensions of the convergence results can be given to the multivariate case, in particular to the case $(E, \mathscr{A}) = \otimes_{i=1}^{k} (E_i, \mathscr{A}_i)$, $\mu \in M(\mu_1, \dots, \mu_k)$ a measure with marginals μ_i on E_i. The problem is to find the I-projection of μ on $M(\nu_1, \dots, \nu_k)$, where $\nu_i \ll \mu_i$, $\mu \ll \mu_1 \otimes \cdots \otimes \mu_k$. Some of the techniques of this paper can be extended also to the case of multivariate joint marginals. Details will be given in a separate paper.

(c) It would be of interest also to have a convergence result for the associated statistical problem with unknown distribution μ and to study the asymptotic properties of the corresponding estimators.

Acknowledgments. The author would like to thank an Associate Editor and two reviewers for their constructive comments on a first version of this paper.

REFERENCES

ARONSZAJN, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** 377–404.

BISHOP, Y. M. M. (1967). Multidimensional contingency tables: cell estimates. Ph.D. dissertation, Harvard Univ.

BISHOP, Y. M. and FIENBERG, S. E. (1969). Incomplete two dimensional contingency tables. *Biometrics* **25** 119–128.

BREIMAN, L. and FRIEDMAN, J. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–619.

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

BUJA, A., HASTIE, T. J. and TIBSHIRANI, R. J. (1989). Linear smoothers and additive models. *Ann. Statist.* **17** 453–555.

CHENEY, E. W. and GOLDSTEIN, A. A. (1959). Proximity maps for convex sets. *Proc. Amer. Math. Soc.* **10** 448–450.

CSISZAR, I. (1975). *I*-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **3** 146–158.

CSISZAR, I. and TUSNADY, G. (1984). Information geometry and alternating minimization procedures. *Statist. Decisions* **1** 205–237.

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

DILIBERTO, S. P. and STRAUSS, E. G. (1951). On the approximation of functions of several variables by the sum of functions of fewer variables. *Pacific J. Math.* **1** 195–210.

DYKSTRA, R. L. (1983). An algorithm for restricted least squares regression. *J. Amer. Statist. Assoc.* **78** 837–842.

FIENBERG, S. E. (1970). An iterative procedure for estimation in contingency tables. *Ann. Math. Statist.* **41** 907–917.

FRIEDMAN, J. H., STUETZLE, W. and SCHROEDER, A. (1984). Projection pursuit density estimation. *J. Amer. Statist. Assoc.* **79** 599–608.

GAFFKE, N. and MATHAR, R. (1989). A cyclic projection algorithm via duality. *Metrika* **36** 29–54.

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

HABERMAN, S. J. (1974). *The Analysis of Frequency Data*. Univ. Chicago Press.

HABERMAN, S. J. (1984). Adjustment by minimum discriminant information. *Ann. Statist.* **12** 971–988.

HAMAKER, C. and SOLOMON, D. C. (1978). The angles between the null spaces of x-rays. *J. Math. Anal. Appl.* **62** 1–23.

HUBER, P. (1985). Projection pursuit. *Ann. Statist.* **13** 435–475.

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

JIROUSEK, R. (1991). Solution of the marginal problem and decomposable distributions. *Kybernetika* **27** 1–9.

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

KULLBACK, S. (1968). Probability densities with given marginals. *Ann. Math. Statist.* **39** 1236–1243.

LIESE, F. (1975). On the existence of f-projections. *Colloq. Math. Soc. János Bolyai* **16** 431–446.

LIESE, F. and VAJDA, I. (1987). *Convex Statistical Distances*. Teubner, Leipzig.

LIGHT, W. A. and CHENEY, E. W. (1985). *Approximation Theory in Tensor Product Spaces. Lecture Notes in Math.* **1169**. Springer, Berlin.

RÜSCHENDORF, L. (1985). Projections and iterative procedures. In *Multivariate Analysis VI* (P. R. Krishnaiah, ed.) 485–493. North-Holland, Amsterdam.

RÜSCHENDORF, L. and THOMSEN, W. (1993). Note on the Schrödinger equation and I-projections. *Statist. Probab. Lett.* **17** 369–375.

SINKHORN, R. (1967). Diagonal equivalence to matrices with prescribed row and column sums. *Amer. Math. Monthly* **74** 402–405.

STONE, C. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705.

VON NEUMANN, J. (1950). *Functional Operators* **2**. Princeton Univ. Press.

INSTITUT FÜR MATHEMATISCHE STOCHASTIK
HEBELSTRASSE 27
D-79189 FREIBURG
GERMANY