Efficient Estimation of Linear Functionals of a Probability Measure P with Known Marginal Distributions

Peter J. Bickel; Ya'Acov Ritov; Jon A. Wellner

The Annals of Statistics, Vol. 19, No. 3 (Sep., 1991), 1316-1346.

Stable URL:
http://links.jstor.org/sici?sici=0090-5364%28199109%2919%3A3%3C1316%3AEEOLFO%3E2.0.CO%3B2-7

The Annals of Statistics is currently published by Institute of Mathematical Statistics.

The Annals of Statistics
1991, Vol. 19, No. 3, 1316–1346

EFFICIENT ESTIMATION OF LINEAR FUNCTIONALS OF A PROBABILITY MEASURE P WITH KNOWN MARGINAL DISTRIBUTIONS

BY PETER J. BICKEL, YA'ACOV RITOV AND JON A. WELLNER[1]

University of California, Berkeley, Hebrew University and University of Washington

Suppose that P is the distribution of a pair of random variables (X, Y) on a product space $\mathbb{X} \times \mathbb{Y}$ with known marginal distributions P_X and P_Y. We study efficient estimation of functions $\theta(h) = \int h\,dP$ for fixed h: $\mathbb{X} \times \mathbb{Y} \to R$ under iid sampling of (X, Y) pairs from P and a regularity condition on P. Our proposed estimator is based on partitions of both $\mathbb{X}$ and $\mathbb{Y}$ and the modified minimum chi-square estimates of Deming and Stephan (1940). The asymptotic behavior of our estimator is governed by the projection on a certain sum subspace of $L_2(P)$, or equivalently by a pair of equations which we call the "ACE equations."

1. Introduction. Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are iid random vectors with joint df H on R^2 and marginal df's H_X and H_Y, respectively. Our goal in this article is to discuss efficient estimation of H (at an arbitrary point, or a fixed linear functional of the form $\int h\,dH$) when the marginal df's H_X and H_Y are known; that is, $H_X = F$ and $H_Y = G$, where F and G are fixed and known. If F and G are continuous, we may, without loss of generality, assume that H is a df on $[0, 1]^2$ with uniform marginals, but our treatment in the following sections will not insist on this.

One justification of the assumption of known marginals is by way of an "auxiliary samples" model as follows [this model was pointed out to J. Wellner by Marshall (1986)]: Suppose $(X_{11}, Y_{11}), \ldots, (X_{1n_1}, Y_{1n_1})$ are iid H as above and, in addition, we also observe independent samples of X's and Y's from the marginals H_X and H_Y of H: $X_{21}, \ldots, X_{2n_2}$ are iid H_X and $Y_{31}, \ldots, Y_{3n_3}$ are iid H_Y. If n_2 and n_3 are very large relative to n_1, we can (at least heuristically) act as if the marginal df's are known and equal to the empirical df's of the n_2 auxiliary X's and n_3 Y's, respectively. This model can, of course, also be viewed as a missing data model: The Y's are missing in the second sample and the X's are missing in the third sample. It is also a submodel of the bivariate censorship model with nonidentically distributed censoring variables [see, e.g., Dabrowska (1988)] and deserves consideration and study in its own right. We intend to do this elsewhere.

Received September, 1988; revised June, 1990.

[1]Research supported by NSF Grant DMS-84-00893 and the John Simon Guggenheim Foundation (and carried out at the Centrum voor Wiskunde en Informatica, Amsterdam and Mathematics Institute, University of Leiden).

AMS 1980 *subject classifications.* Primary 62G05, 60F05; secondary 62G30, 60G44.

Key words and phrases. Marginal distributions, modified minimum chi square, alternating projections, asymptotic normality, efficiency.

Vitale (1979) has studied a regression version of our problem. He also gives motivations and justifications from census problems and from time series.

In the discrete, or contingency table, setting this problem has a long history, apparently beginning with Deming and Stephan (1940), and continuing with the work of Ireland and Kullback (1968). Despite the considerable knowledge and effort devoted to the discrete problem, the continuous version of the problem has not, to our knowledge, received an adequate treatment. In fact, the estimators we study here for the continuous problem are based on the estimators of Deming and Stephan (1940), but with the number of cells tending to infinity with sample size n. However, with increasing number of cells as $n \to \infty$, the number of constraints in the discrete problems also increases to infinity, and this makes the large sample study of estimators much more difficult in the continuous problem which we study here than in the discrete problem with a fixed number of cells as in Deming and Stephan (1940).

There is also a long history of inequalities for bivariate distributions in terms of their marginals beginning with Hoeffding (1940) and Fréchet (1951), and continuing in the more recent work of Whitt (1976) and Cambanis, Simons and Stout (1976). See Marshall and Olkin [(1979), page 381] for a nice treatment. However, these inequalities in themselves apparently do not yield an efficient estimator of F.

Haberman (1984) discusses minimum Kullback–Leibler divergence-type estimators for this and more general problems involving a fixed finite number of constraints. Since our model can be viewed as one with an infinite number of constraints, Haberman's results do not apply. Sheehy (1987, 1988) has extended Haberman (1984) in a study of estimation of probability measures subject to a finite number of constraints. Kullback (1968) and Csiszár (1975) study minimal Kullback–Leibler divergence projections of a given (population) distribution onto the set of distributions with given marginals. The results of Kullback and Csiszár have been extended to other divergence measures by Rüschendorf (1984). [He also indicates difficulties in Csiszár's (1975) Corollaries 3.1 and 3.2.] Their results have apparently not yet been connected with the estimation problem.

One possible explanation for the long-standing lack of a satisfactory solution to the continuous problem is that the influence function of *any* efficient estimator cannot be calculated in closed form, but can only be characterized in terms of a certain pair of equations related to the projection on a certain sum subspace of $L_2(P)$. These equations, which we call the ACE equations because of the alternating conditional expectations algorithm for calculating certain cases of projections of this type, occur repeatedly in this article and form the basis for a large part of our treatment of this model. The alternating projections algorithm for calculating a projection on a sum subspace of a Hilbert space was originated by von Neumann in the early 1930s, but did not appear in print until von Neumann (1949, 1950). It was independently rediscovered by Aronszajn (1950), Nakano (1953) and Wiener (1955). See Deutsch (1985) or Kayalar and Weinert (1988) for recent reviews and further developments.

Appendix A.4 of Bickel, Klaassen, Ritov and Wellner (1991) gives a treatment suited to semiparametric models.

Alternating projection methods have received considerable interest and attention in statistics within the past few years in connection with nonparametric (additive) regression and correlation; see, for example, Breiman and Friedman (1985), Buja (1985) and Buja, Hastie and Tibshirani (1989).

The theory of orthogonal (or spectral) decompositions of bivariate distributions is also closely related to the alternating projection methods and can in fact be used to solve our ACE equations whenever the spectral decomposition is available. Although we will not pursue this direction here, the rich literature concerning spectral decompositions, beginning with Rényi (1959) and continuing with Lancaster (1958, 1963), Eagleson (1964), Venter (1967), Dauxois and Pousse (1975) and Chesson (1976), should be mentioned. Buja (1985) shows the connections between this theory and the work of Breiman and Friedman (1985). Use of spectral decompositions for discrete distributions (sometimes known as *correspondence analysis* or *reciprocal averaging*) goes back even further, to Fisher (1936) among others; see, for example, Schriever [(1986), Chapter 2] for an interesting account.

In fact, the bivariate model with known marginal distributions which we treat here is just one example of a large class of semiparametric models in which the efficient influence function involves a projection on a subspace of a Hilbert space with a sum space structure. When the subspaces involved in forming the sum spaces are orthogonal, explicit formulas are usually possible since the projection on the sum space is then the sum of the individual projections. However, when orthogonality fails (as it does in the present model), explicit formulas are often not available and we are forced to work with the equations defining the projections. As far as we know, this article is the first instance of a complete proof of asymptotic efficiency of an estimator in any model of this type with two honestly infinite-dimensional nonorthogonal subspaces making up the sum space.

2. The estimator. Let $(\mathbb{X} \times \mathbb{Y}, \mathbf{F}_X \times \mathbf{F}_Y)$ be a measurable space and suppose that F and G are given probability measures on $\mathbb{X}$ and $\mathbb{Y}$, respectively. Let $\mathbf{P}$ denote the set of all probability measures on $(\mathbb{X} \times \mathbb{Y}, \mathbf{F}_X \times \mathbf{F}_Y)$ with the given marginal laws F and G. We let (X, Y) be the identity map from $\mathbb{X} \times \mathbb{Y}$ to $\mathbb{X} \times \mathbb{Y}$. Then for $P \in \mathbf{P}$ we have

$$\text{(P1)} \qquad P(A \times \mathbb{Y}) = F(A) \quad \text{for all } A \in \mathbf{F}_X$$

and

$$\text{(P2)} \qquad P(\mathbb{X} \times B) = G(B) \quad \text{for all } B \in \mathbf{F}_Y.$$

In other words, $P(X \in A) = F(A)$ and $P(Y \in B) = G(B)$, while $P((X, Y) \in C) = P(C)$ for $C \in \mathbf{F}_X \times \mathbf{F}_Y$.

For $\alpha \in (0, 1)$, let $\mathbf{P}_\alpha$ denote the subset of $\mathbf{P}$ satisfying, in addition,

$$\text{(P3)} \qquad P(X \in A, Y \in B) \geq \alpha F(A) G(B) \quad \text{for all } A \in \mathbf{F}_X,\ B \in \mathbf{F}_Y.$$

Let $h: \mathbb{X} \times \mathbb{Y} \to R$ be a fixed $\mathbf{F}_X \times \mathbf{F}_Y$-measurable function with $Eh^2(X, Y) < \infty$. We consider estimation of the functional

$$\theta_h(P) = \int\int h(x, y)\, dP(x, y) = Eh(X, Y). \tag{2.1}$$

To introduce and describe our estimator, we first need partitions of $\mathbb{X}$ and $\mathbb{Y}$ as follows: For a given sample size n let

$$\mathbf{A}_n = \{A_{n,1}, \ldots, A_{n,k(n)}\} \equiv \{A_{n1}, \ldots, A_{nk}\}$$

and

$$\mathbf{B}_n = \{B_{n,1}, \ldots, B_{n,m(n)}\} \equiv \{B_{n1}, \ldots, B_{nm}\}$$

be measurable partitions of $\mathbb{X}$ and $\mathbb{Y}$, respectively. (Thus $\bigcup_{i=1}^{k(n)} A_{ni} = \mathbb{X}$ and $A_{ni} \cap A_{nj} = \varnothing$ for $i \neq j$.) Let

$$\mathbf{F}_n \equiv \sigma\{1_{A_{ni} \times B_{nj}}: i = 1, \ldots, k(n), j = 1, \ldots, m(n)\},$$

$$\mathbf{F}_{nX} \equiv \sigma\{1_{A_{ni}}: i = 1, \ldots, k(n)\}$$

and

$$\mathbf{F}_{nY} \equiv \sigma\{1_{B_{nj}}: j = 1, \ldots, m(n)\}.$$

We assume that the partitions are constructed so that

$$\text{(F1)} \qquad \begin{aligned} F(A_{ni}) &\geq \frac{\gamma_n}{\sqrt{n}}, \qquad i = 1, \ldots, k(n), \\ G(B_{nj}) &\geq \frac{\gamma_n}{\sqrt{n}}, \qquad j = 1, \ldots, m(n), \end{aligned}$$

where $\gamma_n^2/\log n \to \infty$ and $\gamma_n/\sqrt{n} \to 0$, the sequences of partitions $\{\mathbf{A}_n\}, \{\mathbf{B}_n\}$ and hence also the sequences of sigma fields $\{\mathbf{F}_{nX}\}, \{\mathbf{F}_{nY}\}$ are nested (or monotone increasing) and this of course entails that $\{\mathbf{A}_n \times \mathbf{B}_n\}$ and $\{\mathbf{F}_n\}$ are also nested:

$$\text{(F2)} \qquad \begin{aligned} &\mathbf{F}_{nX} \subset \mathbf{F}_{(n+1)X}, \\ &\mathbf{F}_{nY} \subset \mathbf{F}_{(n+1)Y} \quad \text{for } n = 1, 2 \ldots, \\ &\mathbf{F}_n \subset \mathbf{F}_{n+1} \end{aligned}$$

and finally, the limiting sigma fields equal the original $\mathbf{F}_X$, $\mathbf{F}_Y$ and $\mathbf{F}_X \times \mathbf{F}_Y$, respectively:

$$\text{(F3)} \qquad \begin{aligned} \mathbf{F}_{\infty X} &\equiv \sigma\left(\bigcup_{n=1}^{\infty} \mathbf{F}_{nX}\right) = \mathbf{F}_X, \\ \mathbf{F}_{\infty Y} &\equiv \sigma\left(\bigcup_{n=1}^{\infty} \mathbf{F}_{nY}\right) = \mathbf{F}_Y, \\ \mathbf{F}_{\infty} &\equiv \sigma\left(\bigcup_{n=1}^{\infty} \mathbf{F}_{n}\right) = \mathbf{F}_X \times \mathbf{F}_Y. \end{aligned}$$

It follows immediately from (F2) that $\bar{h}_n \equiv E(h|\mathbf{F}_n)$ is a martingale with respect to $\{\mathbf{F}_n\}$ and hence from (F3) that $\bar{h}_n \to h$ a.s. and in $L_2(P)$

$$E\big[h(X,Y) - E\big(h(X,Y)|\mathbf{F}_n\big)\big]^2 \to 0 \quad \text{as } n \to \infty. \tag{2.2}$$

When $\mathbb{X}$ and $\mathbb{Y}$ are Euclidean, partitions satisfying F1–F3 can *always* be constructed; see the remarks at the end of this section. When the margins are uniform and the partitions are constructed so that equalities hold in (F1), the growth restrictions on γ_n imply that $m(n) \to \infty$ and $m(n) = o((n/\log n)^{1/2})$.

Here is our estimator of $\theta_h(P)$. Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are iid $P \in \mathbf{P}_\alpha$. For any $k \times m$ array $\{d_{ij}\}$ we write $d_{i\cdot} = \sum_{j=1}^m d_{ij}$ and $d_{\cdot j} = \sum_{i=1}^k d_{ij}$.

$$N_{ij} = \sum_{l=1}^n 1_{A_{ni}\times B_{nj}}(X_l, Y_l) = \#\text{ of } (X_l, Y_l) \text{ pairs in } A_{ni} \times B_{nj} \tag{2.3}$$

for $i = 1, \ldots, k(n)$, $j = 1, \ldots, m(n)$, and let $D \equiv \{(i,j)\colon N_{ij} > 0, i = 1, \ldots, k, j = 1, \ldots, m\}$. Let $\{\hat{p}_{ij}\colon (i,j) \in D\}$ be the unique point minimizing

$$\sum_{i=1}^{k(n)} \sum_{j=1}^{m(n)} \frac{\big(N_{ij} - n\hat{p}_{ij}\big)^2}{N_{ij}} 1_{[N_{ij}>0]} \tag{2.4}$$

subject to the constraints

$$\hat{p}_{i\cdot} = \sum_{j=1}^{m(n)} \hat{p}_{ij} = F(A_{ni}) \equiv p^0_{i\cdot}, \qquad i = 1, \ldots, k(n) \tag{2.5}$$

$$\hat{p}_{\cdot j} = \sum_{i=1}^{k(n)} \hat{p}_{ij} = G(B_{nj}) \equiv p^0_{\cdot j}, \qquad j = 1, \ldots, m(n). \tag{2.6}$$

Since all $p_{ij} > 0$ by (P3) and (F1), all the N_{ij}'s are positive with probability arbitrarily close to 1 for n sufficiently large (see Lemma 2) and hence $D = D_0 \equiv \{(i,j)\colon i = 1, \ldots, k, j = 1, \ldots, m\}$ with high probability for n large.

This is a modified minimum chi-square estimator of the cell probabilities $p_{ij} \equiv P(A_{ni} \times B_{nj})$. As shown by Deming and Stephan (1940) and Ireland and Kullback (1968), the solution $\hat{p}_{ij}$ is of the form

$$\hat{p}_{ij} = \frac{N_{ij}}{n}\big(1 + \hat{a}_i + \hat{b}_j\big),$$

where $\hat{a}_i$ and $\hat{b}_j$ satisfy

$$\hat{a}_i = \frac{np^0_{i\cdot}}{N_{i\cdot}} - 1 - \frac{\sum_{j=1}^m N_{ij}\hat{b}_j}{N_{i\cdot}}, \qquad i = 1, \ldots, k \tag{2.7}$$

and

$$\hat{b}_j = \frac{np^0_{\cdot j}}{N_{\cdot j}} - 1 - \frac{\sum_{i=1}^k N_{ij}\hat{a}_i}{N_{\cdot j}}, \qquad j = 1, \ldots, m. \tag{2.8}$$

Alternatively,

$$\hat{a} = \hat{P}_X(\underline{d} - \hat{b}) \tag{2.9}$$

and

$$\hat{b} = \hat{P}_Y(\underline{d} - \hat{a}) \tag{2.10}$$

where $\underline{d} \equiv \{d_{ij}\}$ is given by

$$d_{ij} \equiv \frac{np_{i\cdot}^0 p_{\cdot j}^0}{N_{ij}} - 1,$$

(note that the d_{ij}'s are computable from the data since $p_{i\cdot}^0$ and $p_{\cdot j}^0$ are known) and for any $\underline{w} \equiv \{w_{ij}\}$,

$$\hat{P}_X\underline{w}(i) \equiv \frac{1}{N_{i\cdot}} \sum_{j=1}^{m} w_{ij}N_{ij}, \qquad i = 1, \ldots, k$$

and

$$\hat{P}_Y\underline{w}(j) = \frac{1}{N_{\cdot j}} \sum_{i=1}^{k} w_{ij}N_{ij}, \qquad j = 1, \ldots, m.$$

Similarly, for vectors $a \in R^k$ and $b \in R^m$,

$$\hat{P}_X b(i) \equiv \frac{1}{N_{i\cdot}} \sum_{j=1}^{m} b_j N_{ij}, \qquad i = 1, \ldots, k$$

and

$$\hat{P}_Y a(j) = \frac{1}{N_{\cdot j}} \sum_{i=1}^{k} a_i N_{ij}, \qquad j = 1, \ldots, m.$$

The coupled pair of equations (2.9) and (2.10) are an example of the ACE equations; similar related equations will reappear repeatedly in the following.

Now set

$$\hat{h}_{ij} = 1_{[N_{ij}>0]} \frac{1}{N_{ij}} \sum_{l=1}^{n} h(X_l, Y_l) 1_{A_{ni} \times B_{nj}}(X_l, Y_l)$$

for $i = 1, \ldots, k(n)$, $j = 1, \ldots, m(n)$. Our estimator of $\theta_h(P)$ is

$$\hat{\theta}_n = \sum_{i=1}^{k} \sum_{j=1}^{m} \hat{P}_{ij}\hat{h}_{ij}. \tag{2.11}$$

To describe the asymptotic behavior of our estimator $\hat{\theta}_n$ given in (2.11), we need to introduce two key functions. Let u: $\mathbb{X} \to R$ and v: $\mathbb{Y} \to R$ be the unique [up to centering and $L_2(P)$ equivalence] solutions of the equations

$$E_P(h(X, Y) - u(X) - v(Y)|X) = 0 \quad \text{a.s.} \tag{2.12}$$

and

$$E_P(h(X, Y) - u(X) - v(Y)|Y) = 0 \quad \text{a.s.} \tag{2.13}$$

Alternatively,

$$u(X) = P_X(h(X,Y) - v(Y)) \tag{2.14}$$

and

$$v(Y) = P_Y(h(X,Y) - u(X)), \tag{2.15}$$

where $P_X \equiv E(\cdot|X)$ and $P_Y = E(\cdot|Y)$. These are the ACE equations again; compare with (2.9) and (2.10). The functions u and v yield the components of the projection of h onto the subspace $\mathbf{H}_X + \mathbf{H}_Y$ of $L_2(P)$ (which is closed under the assumption that $P \in \mathbf{P}_\alpha$; see Lemma 1 in Section 3); here

$$\mathbf{H}_X \equiv \{a = a(X): E_P a^2(X) < \infty\}$$

and

$$\mathbf{H}_Y \equiv \{b = b(Y): E_P b^2(Y) < \infty\}.$$

Thus, letting $\Pi(h|\mathbf{H})$ denote the projection of h onto the subspace $\mathbf{H}$ of the Hilbert space $L_2(P)$,

$$\Pi(h|\mathbf{H}_X + \mathbf{H}_Y) = u(X) + v(Y)$$

or

$$h(X,Y) - u(X) - v(Y) \perp a(X) + b(Y) \tag{2.16}$$

for all $a \in \mathbf{H}_X$, $b \in \mathbf{H}_Y$. The orthogonality relation (2.16) implies that $\tilde{\mathbf{I}}_h(X,Y) \equiv h(X,Y) - u(X) - v(Y) \in \dot{\mathbf{P}}$, the tangent space of $\mathbf{P}$ at P, and hence that our estimator $\hat{\theta}_n$ is efficient. For more on efficiency see Section 4, where we also define and characterize $\dot{\mathbf{P}}$.

THEOREM 1. *Suppose that $P \in \mathbf{P}_\alpha$ for some $\alpha > 0$, that* (F1)–(F3) *hold and $Eh^2(X,Y) < \infty$. Then*

$$\begin{aligned}\sqrt{n}\left(\hat{\theta}_n - \theta_h(P)\right) &= \frac{1}{\sqrt{n}} \sum_{l=1}^{n} \{h(X_l,Y_l) - u(X_l) - v(Y_l)\} + o_p(1) \\ &= \frac{1}{\sqrt{n}} \sum_{l=1}^{n} \tilde{\mathbf{I}}_h(X_l,Y_l) + o_p(1).\end{aligned} \tag{2.17}$$

Hence

$$\sqrt{n}\left(\hat{\theta}_n - \theta_h(P)\right) \to_d N\left(0, E\left(\tilde{\mathbf{I}}_h^2(X,Y)\right)\right) \qquad \text{as } n \to \infty. \tag{2.18}$$

Remarks and further problems. 1. *Alternative estimators.* Two obvious alternatives to modified minimum chi-square estimation of the p_{ij}'s, and hence alternatives to $\hat{\theta}_n$, are (a) minimum Kullback–Leibler divergence estimation, and (b) maximum-likelihood estimation. These estimators are of the forms $\hat{p}_{ij}^{KL} = (n^{-1}N_{ij})\hat{a}_i\hat{b}_j$ and $\hat{p}_{ij}^{ML} = (n^{-1}N_{ij})/(\hat{a}_i + \hat{b}_j)$; see, for example, Ireland and Kullback [(1968), pp. 181 and 180, respectively]. In fact, the minimum Kullback–Leibler estimator $\{\hat{p}_{ij}^{KL}\}$ is easily computed by "iterative proportional fitting" as originally proposed by Deming and Stephan (1940). These alternatives deserve further investigation.

2. *The Assumptions (F1)–(F3) and (P1)–(P3).* When $\mathbb{X}$ and $\mathbb{Y}$ are Euclidean, partitions $\mathbf{A}_n$ of $\mathbb{X}$ and $\mathbf{B}_n$ of $\mathbb{Y}$ can always be constructed so that (F1)–(F3) hold. This is obvious in the special case of $\mathbb{X} = \mathbb{Y} = [0, 1]$ and both F and G uniform on $[0, 1]$; just use the usual diadic partition

$$\mathbf{A}_n \equiv \{(i2^{-r},(i+1)2^{-r}]: i = 1,\ldots,2^r - 1\},$$

with $m(n) = 2^r$, and similarly for $\mathbf{B}_n$, with $m(n)$ and $k(n)$ chosen so that (F1) holds. For general F and G, the partitions can be generated by simply proceeding diadically subject to satisfying (F1). (F1) together with (P3) implies that the expected number of observations in each cell goes to infinity faster than $\log n$. Of course, (P3) alone implies that $\mathbf{H}_X + \mathbf{H}_Y$ is closed; see the proof of Lemma 1 in Section 5. It would be desirable to weaken assumption (P3). A first step would be to try to weaken it to just the hypothesis that $\mathbf{H}_X + \mathbf{H}_Y$ is closed.

3. *The asymptotic variance* $E[\tilde{\mathbf{I}}_h^2(X, Y)] \equiv \sigma_h^2$. The asymptotic variance of our estimator is not easily calculated because it involves a projection on $\mathbf{H}_X + \mathbf{H}_Y$; see Section 4 for some efficiency comparisons via inequalities. It is, however, consistently estimated under the same conditions as used in Theorem 1 by the estimator

$$\hat{\sigma}_n^2 \equiv \sum_{i=1}^{k}\sum_{j=1}^{m} \hat{p}_{ij}\left(\hat{h}_{ij} - \tilde{u}_i - \tilde{v}_j\right)^2,$$

where $\tilde{u}_i + \tilde{v}_j = \mathrm{ACE}(\hat{h}|\hat{p})(i, j)$; that is, $\tilde{u}_i$ and $\tilde{v}_j$ solve

$$\sum_j \hat{p}_{ij}\left(\hat{h}_{ij} - \tilde{u}_i - \tilde{v}_j\right) = 0, \qquad i = 1,\ldots,k(n)$$

and

$$\sum_i \hat{p}_{ij}\left(\hat{h}_{ij} - \tilde{u}_i - \tilde{v}_j\right) = 0, \qquad j = 1,\ldots,m(n).$$

4. *Nonlinear functionals and the estimator as a process.* We have only considered one fixed function h in Theorem 1. Of course this generalizes immediately to finitely many functions h since a sequence of random vectors converges in probability to zero if and only if each coordinate thereof converges in probability to zero. This allows for treatment of nonlinear functionals of the form $g(\int h_1\, dP,\ldots,\int h_r\, dP)$, where g is a fixed differentiable map from R^r to R^1, such as the correlation coefficient, via the delta method. To handle more general nonlinear functionals, it would be useful to deal with the estimator as a process indexed by $h \in \mathbb{H} \subset L_2(P)$, for example, for $\mathbb{X} = \mathbb{Y} = [0, 1]$, $\mathbb{H} \equiv \{1_{[0,s]\times[0,t]}: 0 \le s \le 1,\ 0 \le t \le 1\}$. Once a result concerning convergence of the entire process is obtained, then many more nonlinear functionals can be treated via the delta method; see, for example, Gill (1989) and Wellner (1989).

5. *Local regularity of* $\hat{\theta}_n$. In Theorem 1 we have not only established asymptotic normality of the estimator at a fixed $P \in \mathbf{P}_\alpha$, but something slightly stronger, namely, asymptotic linearity. It follows from contiguity

theory (Le Cam's third lemma) and pathwise differentiability of $\theta_h(P)$ that the estimator $\hat{\theta}_n$ is locally regular.

6. *The case of one fixed margin.* If only one marginal distribution is known, say $P_X = F$, then efficient estimation of $\theta_h(P)$ is somewhat easier. Condition (P3) is no longer needed, and condition (F1) can be relaxed to $\gamma_n \to \infty$ and $n^{-1/2}\gamma_n \to 0$. The estimator is just

$$\tilde{\theta}_n \equiv \sum_{i=1}^{k} p_{i\cdot}^0 \hat{h}_i,$$

where

$$\hat{h}_i \equiv \frac{1}{N_{i\cdot}} \sum_{l=1}^{n} h(X_l, Y_l) 1_{A_{ni}}(X_l).$$

When the distribution of X is known and discrete, estimation of $\theta_h(P) = EY$ [i.e., $\mathbb{X} = \mathbb{Y} = R$ and $h(x,y) = y$] has been considered in both finite and random sampling contexts by Jagers, Odén and Trulsson (1985).

7. *Higher-dimensional versions of the problem.* In more than two dimensions ($d > 2$) there are many different variants of the problem we consider here, depending on which lower-dimensional marginal distributions are assumed known. In the simplest variant in which only the collection of one-dimensional marginal distributions are assumed known, condition (F1) would be replaced by $F_i(A_{nj}) \geq \gamma_n n^{-1/d}$, $i = 1, \ldots, d$, $j = 1, \ldots,$ $k = k(n)$, where $\gamma_n^d/(\log n) \to \infty$ and $\gamma_n^d/n \to 0$.

3. Proof of Theorem 1. First we introduce some useful notation. For $P \in \mathbf{P}_\alpha$ let

$$p_{ij} \equiv P(A_{ni} \times B_{nj}). \tag{3.1}$$

We write, in the usual way, $N_{i\cdot}$, $N_{\cdot j}$, $\hat{p}_{i\cdot}$, $\hat{p}_{\cdot j}$ and $p_{i\cdot}$, $p_{\cdot j}$ for the sums over j or i, respectively, of the corresponding N_{ij}'s, $\hat{p}_{ij}$ or p_{ij}'s.

Set

$$\begin{aligned} \bar{h}(X,Y) &\equiv E[h(X,Y)|\mathbf{F}_n] \\ &= \sum_{i,j} p_{ij}^{-1} \iint_{A_{ni}\times B_{nj}} h(x,y)\, dP(x,y) 1_{A_{ni}\times B_{nj}}(X,Y), \end{aligned} \tag{3.2}$$

so that $\bar{h}(x,y) = h_{ij} \equiv p_{ij}^{-1} \iint_{A_{ni}\times B_{nj}} h(x,y)\, dP(x,y)$ on $A_{ni} \times B_{nj}$. Now let $\overline{\mathbf{H}}_X$ and $\overline{\mathbf{H}}_Y$ be the subspaces of $L_2(P)$ consisting of functions which are $\mathbf{F}_{nX}$ and $\mathbf{F}_{nY}$ measurable, respectively:

$$\overline{\mathbf{H}}_X = L_2(\mathbb{X}, \mathbf{F}_{nX}, F) = \{u \in \mathbf{H}_X : u \text{ is } \mathbf{F}_{nX} \text{ measurable}\},$$

$$\overline{\mathbf{H}}_Y = L_2(\mathbb{Y}, \mathbf{F}_{nY}, G) = \{v \in \mathbf{H}_Y : v \text{ is } \mathbf{F}_{nY} \text{ measurable}\}.$$

Similarly, let $\hat{\mathbf{H}}_X$ and $\hat{\mathbf{H}}_Y$ be the subspaces of $L_2(\mathbb{P}_n)$ consisting of functions

which are $\mathbf{F}_{nX}$ and $\mathbf{F}_{nY}$ measurable, respectively:

$$\hat{\mathbf{H}}_X = L_2\big(\mathbb{X}, \mathbf{F}_{nX}, \mathbb{P}_n^X\big) = \big\{u \in L_2\big(\mathbb{P}_n^X\big)\colon u \text{ is } \mathbf{F}_{nX} \text{ measurable}\big\},$$

$$\hat{\mathbf{H}}_Y = L_2\big(\mathbb{Y}, \mathbf{F}_{ny}, \mathbb{P}_n^Y\big) = \big\{v \in L_2\big(\mathbb{P}_n^Y\big)\colon v \text{ is } \mathbf{F}_{nY} \text{ measurable}\big\},$$

where $\mathbb{P}_n^X$ and $\mathbb{P}_n^Y$ are the marginal empirical measures of X and Y, respectively.

Corresponding to u and v, the components of the projection of h onto $\mathbf{H}_X + \mathbf{H}_Y$, let $\bar{u}(X), \bar{v}(Y)$ denote the components of the projection of $\bar{h}$ onto $\bar{\mathbf{H}}_X + \bar{\mathbf{H}}_Y$ in $L_2(P)$. Because $\bar{h}, \bar{u}$ and $\bar{v}$ are constant on elements of the partition, $\bar{u}(x) = \sum_{i=1}^k \bar{u}_i 1_{A_{ni}}(x)$, $\bar{v}(y) = \sum_{j=1}^m \bar{v}_j 1_{B_{nj}}(y)$, this projection problem reduces to just solving a finite system of linear equations governed by the cell probabilities p_{ij}: The coordinates $\bar{u}_i$ and $\bar{v}_j$ of $\bar{u}$ and $\bar{v}$ must satisfy

$$\sum_{j=1}^m p_{ij}\big(\bar{h}_{ij} - \bar{u}_i - \bar{v}_j\big) = 0 \quad \text{for} \quad i = 1, \ldots, k, \tag{3.3}$$

$$\sum_{i=1}^k p_{ij}\big(\bar{h}_{ij} - \bar{u}_i - \bar{v}_j\big) = 0 \quad \text{for} \quad j = 1, \ldots, m. \tag{3.4}$$

Similarly, we let $\hat{u}(X), \hat{v}(Y)$ denote the components of the projection of $\bar{h}$ onto $\hat{\mathbf{H}}_X + \hat{\mathbf{H}}_Y$ in $L_2(\mathbb{P}_n)$. Again, since $\bar{h}$, $\hat{u}$ and $\hat{v}$ are constant on elements of the partition, this projection problem reduces to a finite system of equations, governed now by the empirical cell probabilities $N_{ij}/n \equiv \mathbb{P}_n(A_{ni} \times B_{nj})$. The coordinates $\hat{u}_i$ and $\hat{v}_j$ of $\hat{u}$ and $\hat{v}$ must satisfy

$$\sum_{j=1}^m N_{ij}\big(\bar{h}_{ij} - \hat{u}_i - \hat{v}_j\big) = 0 \quad \text{for } i = 1, \ldots, k, \tag{3.5}$$

$$\sum_{i=1}^k N_{ij}\big(\bar{h}_{ij} - \hat{u}_i - \hat{v}_j\big) = 0 \quad \text{for } j = 1, \ldots, m. \tag{3.6}$$

These equations are, of course, just discrete analogues of the ACE equations (2.14) and (2.15).

We will prove Theorem 1 by way of seven lemmas, which we now state. We will then show how the lemmas yield the theorem. Proofs of the lemmas are deferred to Section 5.

The seven lemmas.

LEMMA 1. *If $P \in \mathbf{P}_\alpha$, then $u(X) - Eu(X) \in \mathbf{H}_X$ and $v(Y) - Ev(Y) \in \mathbf{H}_Y$ are uniquely defined [up to equivalence in $L_2(P_X)$ and $L_2(P_Y)$, respectively].*

LEMMA 2. *Suppose that* $P \in \mathbf{P}_\alpha$ *and* (F1) *holds. Then*

(i) *For any* $\varepsilon > 0$,

$$\Pr\left(\max_{i,j} \frac{p_{ij}}{N_{ij}/n} \geq 1 + \varepsilon \right) \to 0 \quad \text{as } n \to \infty,$$

$$\Pr\left(\max_{i,j} \frac{N_{ij}/n}{p_{ij}} \geq 1 + \varepsilon \right) \to 0 \quad \text{as } n \to \infty$$

and

$$\Pr\left(\max_{i,j} \left| \frac{N_{ij}/n}{p_{ij}} - 1 \right| \geq \varepsilon \right) \to 0 \quad \text{as } n \to \infty.$$

(ii) *For any* $\varepsilon > 0$,

$$\Pr\left(\max_{i=1,\ldots,k} \left| \frac{N_{i\cdot}/n}{p_{i\cdot}} - 1 \right| > \varepsilon \right) \to 0 \quad \text{as } n \to \infty$$

and

$$\Pr\left(\max_{j=1,\ldots,m} \left| \frac{N_{\cdot j}/n}{p_{\cdot j}} - 1 \right| > \varepsilon \right) \to 0 \quad \text{as } n \to \infty.$$

(iii) *For any* $0 < \gamma < \alpha$

$$\Pr\left(\max_{i,j} \frac{(N_{i\cdot}/n)(N_{\cdot j}/n)}{N_{ij}/n} > \frac{1}{\gamma} \right) \to 0 \quad \text{as } n \to \infty.$$

LEMMA 3. *Suppose that* $P \in \mathbf{P}_\alpha$ *and* (F1) *holds. Then*

(i) $\hat{p}_{ij} = \dfrac{N_{ij}}{n}\left(1 + \hat{a}_i + \hat{b}_j\right)$ *for* $i = 1, \ldots, k(n)$ *and* $j = 1, \ldots, m(n)$.

(ii) $$\max_{1 \leq i \leq k} |\hat{a}_i| + \max_{1 \leq j \leq m} \left|\hat{b}_j\right| \to_p 0 \quad \text{as } n \to \infty.$$

LEMMA 4. *Suppose that* $P \in \mathbf{P}_\alpha$ *and* (F1)–(F3) *hold. Then*

$$\sum_{i,j} \hat{p}_{ij}\left(\hat{h}_{ij} - \bar{h}_{ij}\right) = o_p\left(n^{-1/2}\right).$$

LEMMA 5. *If* $\{\hat{p}_{ij}\}$ *is the solution of the minimization problem* (2.3)–(2.5), *then*

$$\begin{aligned} \sum_{i,j} \hat{p}_{ij}\bar{h}_{ij} &= \theta_h(P) + \frac{1}{n}\sum_{i,j} N_{ij}\left(\bar{h}_{ij} - \bar{u}_i - \bar{v}_j\right) \\ &\quad + \sum_i \left(p_{i\cdot} - \frac{N_{i\cdot}}{n}\right)\left(\hat{u}_i - \bar{u}_i\right) + \sum_j \left(p_{\cdot j} - \frac{N_{\cdot j}}{n}\right)\left(\hat{v}_j - \bar{v}_j\right). \end{aligned} \tag{3.7}$$

LEMMA 6. *Suppose that* $P \in \mathbf{P}_\alpha$ *and* (F1)–(F3) *hold. Then*

$$nC_n^2 \equiv n\,\mathrm{Var}\left\{\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{m} N_{ij}\left(\overline{h}_{ij} - \overline{u}_i - \overline{v}_j\right) - \int (h(x,y) - u(x) - v(y))\, d\mathbb{P}_n(x,y)\right\}$$

$$\leq \iint \left(h(x,y) - \overline{h}(x,y)\right)^2 dP(x,y) \to 0 \quad \text{as } n \to \infty.$$

LEMMA 7. *Suppose that* $P \in \mathbf{P}_\alpha$ *and* (F1)–(F3) *hold. Then both*

$$\sqrt{n}\sum_i (p_{i\cdot} - N_{i\cdot}/n)(\hat{u}_i - \overline{u}_i) = O_p(\gamma_n^{-1}) = o_p(1)$$

and

$$\sqrt{n}\sum_j (p_{\cdot j} - N_{\cdot j}/n)(\hat{v}_j - \overline{v}_j) = O_p(\gamma_n^{-1}) = o_p(1).$$

Proof of the asymptotic linearity and normality theorem. Now we show how Theorem 1 follows from Lemmas 1–7.

PROOF OF THEOREM 1. We first use Lemma 5 to write

$$\begin{aligned}
&\sqrt{n}\left(\hat{\theta}_n - \theta_h(P)\right) \\
&\quad = \sqrt{n}\left(\sum_{i,j} \hat{p}_{ij}\overline{h}_{ij} - \theta_h(P)\right) + \sqrt{n}\left(\sum_{i,j}\hat{p}_{ij}\left(\hat{h}_{ij} - \overline{h}_{ij}\right)\right) \\
&\quad = \frac{1}{\sqrt{n}}\sum_{i,j} N_{ij}\left(\overline{h}_{ij} - \overline{u}_i - \overline{v}_j\right) + \sqrt{n}\sum_i (p_{i\cdot} - N_{i\cdot}/n)(\hat{u}_i - \overline{u}_i) \\
&\qquad + \sqrt{n}\sum_j (p_{\cdot j} - N_{\cdot j}/n)(\hat{v}_j - \overline{v}_j) + \sqrt{n}\sum_{i,j}\hat{p}_{ij}\left(\hat{h}_{ij} - \overline{h}_{ij}\right) \quad \text{(by Lemma 5)} \\
&\quad = \sqrt{n}\iint (h(x,y) - u(x) - v(y))\, d\mathbb{P}_n(x,y) \\
&\qquad + \sqrt{n}\iint \left[\overline{h} - \overline{u} - \overline{v} - h + u + v\right] d\mathbb{P}_n + \sqrt{n}\sum_i (p_{i\cdot} - N_{i\cdot}/n)(\hat{u}_i - \overline{u}_i) \\
&\qquad + \sqrt{n}\sum_j (p_{\cdot j} - N_{\cdot j}/n)(\hat{v}_j - \overline{v}_j) + \sqrt{n}\sum_{i,j}\hat{p}_{ij}\left(\hat{h}_{ij} - \overline{h}_{ij}\right) \\
&\quad \equiv \sqrt{n}\iint (h(x,y) - u(x) - v(y))\, d\mathbb{P}_n(x,y) + \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV}.
\end{aligned}$$

But $\mathrm{I} = o_p(1)$ by Lemma 6 and F2, $\mathrm{II} = o_p(1)$ and $\mathrm{III} = o_p(1)$ by Lemma 7 and $\gamma_n \to \infty$ and $\mathrm{IV} = o_p(1)$ by Lemma 4. □

4. Efficiency. Here we do two things. We first show that the estimator $\hat{\theta}_n$ is asymptotically efficient. In view of general asymptotic estimation theory [e.g., Levit (1978), van der Vaart (1988), Pfanzagl (1982) or Bickel, Klaassen, Ritov and Wellner (1991)], since the tangent set $\dot{\mathbf{P}}^0$ is linear it suffices to show that the influence function $\tilde{\mathbf{I}}_h$ of our estimator $\hat{\theta}_n$ is the projection of the (pathwise) derivative of the functional $\theta_h(P)$ onto the tangent space $\dot{\mathbf{P}}$ of the model $\mathbf{P}_\alpha$. We then make some efficiency comparisons with inefficient, but simpler, estimators.

The tangent space $\dot{\mathbf{P}}$. The tangent space $\dot{\mathbf{P}}$ of the model $\mathbf{P}$ at $P_0 \in \mathbf{P}$ is defined to be the closure in $L_2(P_0)$ of the linear span of all score functions of regular parametric submodels through P_0. We claim that

$$\dot{\mathbf{P}} = (\mathbf{H}_X + \mathbf{H}_Y)^\perp . \tag{4.1}$$

To show that (4.1) holds, suppose that $\mathbf{P}_0 = \{P_\theta\colon\ \theta \in \Theta \subset R\} \subset \mathbf{P}$ is a regular parametric submodel through $P_0 \in \mathbf{P}$. Fix $\theta_0 \in \Theta$ so that $P_{\theta_0} = P_0$. Then for any bounded function $a = a(x) \in L_2(P_0)$, since $P_\theta \in \mathbf{P}$ for all $\theta \in \Theta$,

$$\iint a(x)\, dP_\theta(x,y) = \int a(x)\, dF(x) \quad \text{for all } \theta \in \Theta, \tag{4.2}$$

where the right side does not depend on θ since F is known. Hence differentiation across (4.2) yields

$$\iint a(x)\dot{\mathbf{l}}_{\theta_0}(x,y)\, dP_{\theta_0}(x,y) = 0 \tag{4.3}$$

(where $\dot{\mathbf{l}}_{\theta_0}$ denotes the score function for θ in $\mathbf{P}_0 \equiv \{P_\theta\}$), or

$$\dot{\mathbf{l}}_{\theta_0} \perp a \in \mathbf{H}_X. \tag{4.4}$$

By a symmetric argument,

$$\dot{\mathbf{l}}_{\theta_0} \perp b \in \mathbf{H}_Y. \tag{4.5}$$

It follows from (4.4) and (4.5) that

$$\dot{\mathbf{P}} \subset (\mathbf{H}_X + \mathbf{H}_Y)^\perp .$$

To prove the reverse inclusion, let $h \in (\mathbf{H}_X + \mathbf{H}_Y)^\perp$ and $P \in \mathbf{P}_\alpha$. Then there is a uniformly bounded function $\tilde{h} \in (\mathbf{H}_X + \mathbf{H}_Y)^\perp$ which is arbitrarily close to h in $L_2(P_0)$; a detailed proof of this claim is given at the end of this section. Then the parametric model $\mathbf{P}_0 = \{P_\theta\colon\ |\theta| < \theta_0\}$ defined (for some $\theta_0 > 0$) by

$$\frac{dP_\theta}{dP_0}(x,y) = 1 + \theta\tilde{h}(x,y) \tag{4.6}$$

has tangent $\tilde{h}$ at $\theta = 0$. Since the uniformly bounded functions are dense in $(\mathbf{H}_X + \mathbf{H}_Y)^\perp$, (4.1) holds.

Now the linear functional $\theta_h(P)$ has pathwise derivative $\dot{\theta}_h(P)(g) = \int hg\, dP$ at $g \in \dot{\mathbf{P}}$, so the "canonical gradient" (or adjoint of the map $\dot{\theta}_h\colon \dot{\mathbf{P}} \to R$) is just

$\Pi(h|\dot{\mathbf{P}}) = h - \Pi(h|\mathbf{H}_X + \mathbf{H}_Y) = \tilde{\mathbf{l}}_h$. Since the estimator $\hat{\theta}_n$ constructed in Section 2 is locally regular with influence function $\tilde{\mathbf{l}}_h$, it is asymptotically efficient; see, for example, Pfanzagl [(1982), page 158] or Bickel, Klaassen, Ritov and Wellner [(1991), Section 3.3].

Efficiency comparisons. A simple inefficient estimator of $\theta_h(P)$ is the "empirical estimator" given by

$$\theta_h(\mathbb{P}_n) = \iint h(x,y)\, d\mathbb{P}_n(x,y) = \frac{1}{n}\sum_{l=1}^{n} h(X_l, Y_l) \tag{4.7}$$

and, of course, since $Eh^2(X,Y) < \infty$, it follows from the central limit theorem that

$$\sqrt{n}\,(\theta_h(\mathbb{P}_n) - \theta_h(P)) \to_d N(0, \mathrm{Var}_P[h(X,Y)]) \tag{4.8}$$

as $n \to \infty$. Note that this simple estimator does not take advantage of the known marginal distributions.

It follows that the asymptotic relative efficiency of $\theta_h(\mathbb{P}_n)$ with respect to the efficient estimator $\hat{\theta}_n$ is

$$\varepsilon_h(P) \equiv \frac{E_P\big(\tilde{\mathbf{l}}_h^2(X,Y)\big)}{\mathrm{Var}_P(h(X,Y))} = \frac{E_P h^2 - E_P(u+v)^2}{E_P h^2 - (E_P h)^2}. \tag{4.9}$$

Because of the difficulty in calculating the projection on $\mathbf{H}_X + \mathbf{H}_Y$ involved in $\tilde{\mathbf{l}}_h$, we are able to calculate $\varepsilon_h(P)$ explicitly for only a few special P's. In view of these difficulties, it is of interest to examine $\varepsilon_h(P)$ for special P's and to look for inequalities.

PROPOSITION 1. *Suppose that $\varepsilon_h(P)$ is the asymptotic relative efficiency of the empirical estimator $\theta(\mathbb{P}_n)$ with respect to the efficient estimator $\hat{\theta}_n$ given in* (4.3). *Then*

$$\varepsilon_h(P_{\text{indep}}) = \frac{E_P\{h(X,Y) - E(h|X) - E(h|Y) + E(h)\}^2}{E_P\{h(X,Y) - Eh\}^2}$$

$$= 1 - \frac{E\{(h - Eh)(E(h|X) + E(h|Y))\}}{E(h - Eh)^2}.$$

PROOF. The assertion follows immediately from (4.3) by noting that the subspaces $\mathbf{H}_X^0$ and $\mathbf{H}_Y^0$ are orthogonal under independence ($P = P_{\text{indep}}$) and hence projection on $\mathbf{H}_X^0 + \mathbf{H}_Y^0$ is given by the sum of the projections onto $\mathbf{H}_X^0$ and $\mathbf{H}_Y^0$ and these are just the conditional expectation operators. □

While our Theorem 1 does not apply to P's concentrated on curves, it is perhaps instructive to consider the expression for the variance $\sigma_h^2 = E[\tilde{\mathbf{l}}_h^2(X,Y)]$ for P's of this type. If $Y = \phi(X)$ a.s. $P \equiv P_\phi$ for some function ϕ

then $\tilde{\mathbf{I}}_h = 0$ a.s. and $\sigma_h^2 = 0$, suggesting that the efficiency of $\hat{\theta}_n$ relative to the empirical estimator $\theta(\mathbb{P}_n)$ increases with increasing dependence.

Now we specialize to the case $\mathbb{X} = \mathbb{Y} = [0,1]$, $F = G =$ uniform$(0,1)$ and $h(x,y) = 1_{[0,s]\times[0,t]}(x,y) \equiv h_{s,t}(x,y)$, so that

$$\theta_h(P) = P(X \le s, Y \le t) \equiv F(s,t).$$

Under independence, $P = P_{\text{indep}} \equiv F \times G$, we can calculate $\varepsilon_h(P_{\text{indep}}) \equiv \varepsilon_{st}(P_{\text{indep}})$ in (i) of Proposition 1 explicitly: Straightforward calculations yield

$$\varepsilon_{s,t}(P_{\text{indep}}) = \frac{st(1-s)(1-t)}{st(1-st)} = \frac{(1-s)(1-t)}{1-st}. \tag{4.10}$$

In particular,

$$\varepsilon_{t,t}(P_{\text{indep}}) = \frac{(1-t)^2}{(1-t^2)} = \frac{1-t}{1+t} \quad \text{and} \quad \varepsilon_{1/2,1/2}(P_{\text{indep}}) = \frac{1}{3}, \tag{4.11}$$

so the empirical df estimator of $\theta_h(P) = P(X \le 1/2, Y \le 1/2)$ has three times the asymptotic variance of the efficient estimator $\hat{\theta}_n$ at $P = P_{\text{indep}} = F \times G$.

In the bivariate df case with uniform marginals, yet another competing (but inefficient) estimator is the empirical copula function defined by

$$\mathbb{C}_n(s,t) = \mathbb{H}_n\bigl(\mathbb{F}_n^{-1}(s), \mathbb{G}_n^{-1}(t)\bigr),$$

where $\mathbb{H}_n$ is the (joint) empirical df of the (X,Y)'s and $\mathbb{F}_n, \mathbb{G}_n$ are the marginal empirical df's of the X's and Y's, respectively; see, for example, Stute [(1984), page 370]. The limit process for this estimator is

$$\mathbb{X}(s,t) - H_1(s,t)\mathbb{X}(s,1) - H_2(s,t)\mathbb{X}(1,t),$$

where (assuming that the partial derivatives exist)

$$H_1(s,t) \equiv \frac{\partial}{\partial s}H(s,t), \qquad H_2(s,t) \equiv \frac{\partial}{\partial t}H(s,t)$$

and

$$\operatorname{Cov}[\mathbb{X}(s,t), \mathbb{X}(s',t')] = H(s \wedge s', t \wedge t') - H(s,t)H(s',t').$$

It is easy to calculate that the asymptotic variance of this estimator is

$$\begin{aligned} &p(1-p) + H_1^2(s,t)s(1-s) + H_2^2(s,t)t(1-t) - 2H_1(s,t)p(1-s) \\ &\quad - 2H_2(s,t)p(1-t) + 2H_1(s,t)H_2(s,t)(p-st); \end{aligned}$$

this reduces to $st(1-s)(1-t)$ under independence.

PROOF OF $\dot{\mathbf{P}}(P) \supset (\mathbf{H}_X + \mathbf{H}_Y)^{\perp}$ FOR $P \in \mathbf{P}_\alpha$. Suppose that $P \in \mathbf{P}_\alpha$. Let h^* be a bounded function arbitrarily close to h in $L_2(P)$ and set

$$\tilde{h} \equiv h^* - \Pi(h^* | \mathbf{H}_X + \mathbf{H}_Y).$$

Then

$$\|\tilde{h} - h\|_2 \le \|h^* - h\|_2$$

is as small as we please, and it remains only to show that $\tilde{h}$ is bounded. Since h^* is bounded, this will hold if and only if $\Pi(h^*|\mathbf{H}_X + \mathbf{H}_Y)$ is bounded. To prove this, the key step is to show that the operator $P_X P_Y$ on $\mathbf{H}_X^0$ satisfies

$$\|P_X P_Y\|_\infty \leq (1^* - \alpha)^2 < 1, \quad \text{for } P \in \mathbf{P}_\alpha. \tag{4.12}$$

We then use the following explicit form of this projection obtained from the pair of equations (2.14) and (2.15) characterizing the projection $u + v = \Pi(h^*|\mathbf{H}_X + \mathbf{H}_Y)$. Substitution of (2.15) into (2.14) and vice versa yields

$$(I - P_X P_Y)u = P_X(I - P_Y)h^*$$

and

$$(I - P_Y P_X)v = P_Y(I - P_X)h^*$$

or

$$u = (I - P_X P_Y)^{-1} P_X(I - P_Y)h^*$$

and

$$v = (I - P_Y P_X)^{-1} P_Y(I - P_X)h^*,$$

where $w_X \equiv P_X(I - P_Y)H^* \in \mathbf{H}_X^0$ and $w_Y \equiv P_Y(I - P_X)h^* \in \mathbf{H}_Y^0$. This basic representation will also be used in the proof of Lemma 7 in Section 5. Thus it suffices to show that $u = (I - P_X P_Y)^{-1} w$ is bounded if $w \in \mathbf{H}_X^0$ is bounded, and similarly for v. This clearly holds if we show that there is a constant $0 < K < \infty$ such that

$$\left\|(I - P_X P_Y)^{-1} w\right\|_\infty < K\|w\|_\infty \quad \text{for } w \in \mathbf{H}_X^0. \tag{4.13}$$

Let

$$r(x, y) \equiv \frac{dP}{d(F \times G)}(x, y) \geq \alpha.$$

But for $A \in \mathbf{F}_Y$, $P \in \mathbf{P}_\alpha$ and $w \in \mathbf{H}_X^0$,

$$\begin{aligned}\int_{\mathbb{X}\times A} w(x)\, dP(x, y) &= \int_{\mathbb{X}\times A} w(x)\{r(x, y) - \alpha\}\, dF(x)\, dG(y) \\ &\leq \|w\|_\infty \int_{\mathbb{X}\times A} \{r(x, y) - \alpha\}\, dF(x)\, dG(y) \\ &\leq \|w\|_\infty G(A)(1 - \alpha).\end{aligned}$$

Hence

$$\text{ess.sup}\, E(w(X)|\mathbf{F}_Y) \leq (1 - \alpha)\|w\|_\infty$$

and, by symmetry,

$$\text{ess.sup}\, E(w(Y)|\mathbf{F}_X) \leq (1 - \alpha)\|w\|_\infty$$

for $w \in \mathbf{H}_Y^0$. Thus (4.12) holds and

$$\begin{aligned}\left\|(I - P_X P_Y)^{-1} w\right\|_\infty &\le \sum_{i=0}^{\infty} \left\|(P_X P_Y)^i w\right\| \\ &\le \sum_{i=0}^{\infty} (1-\alpha)^{2i} \|w\|_\infty \\ &= \frac{1}{1-(1-\alpha)^2} \|w\|_\infty,\end{aligned}$$

proving (4.13). The corresponding inequality for v holds by symmetry. Since $\tilde{h}$ is bounded, the construction (4.6) is valid for sufficiently small θ, and hence $\dot{\mathbf{P}}(P) \supset (\mathbf{H}_X + \mathbf{H}_Y)^\perp$ for $P \in \mathbf{P}_\alpha$. □

5. Proofs of the lemmas. Throughout this section (X, Y) will denote a random vector with distribution P which is independent of the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$.

PROOF OF LEMMA 1. Let

$$\mathbf{H}_X^0 \equiv \{a(X): E_P a^2(X) < \infty,\ E_P a(X) = 0\},$$

$$\mathbf{H}_Y^0 \equiv \{b(Y): E_P b^2(Y) < \infty,\ E_P b(Y) = 0\}$$

with the $L_2(P)$ norm. Then $u(X) + v(Y) - Eu(X) - Ev(Y)$ is the projection of $h(X, Y) - Eh(X, Y)$ on the closure of $\mathbf{H}_X^0 + \mathbf{H}_Y^0$. The result will follow if $\mathbf{H}_X^0 \cap \mathbf{H}_Y^0 = \{0\}$ and $\mathbf{H}_X^0 + \mathbf{H}_Y^0$ is closed. Let $a \in \mathbf{H}_X^0$ and $b \in \mathbf{H}_Y^0$. Then

$$\begin{aligned}\int\int [a(x) - b(y)]^2 \, dP(x, y) &\ge \alpha \int\int [a(x) - b(y)]^2 \, dF(x) \, dG(y) \\ &= \alpha \left\{ \int a^2(x) \, dP_X(x) + \int b^2(y) \, dP_Y(y) \right\}\end{aligned}$$

and hence

$$\int\int a(x) b(y) \, dP(x, y) \le 1 - \alpha < 1$$

for all $a \in \mathbf{H}_X^0$ and $b \in \mathbf{H}_Y^0$ with $\|a\| = \|b\| = 1$. It follows that $\mathbf{H}_X^0 \cap \mathbf{H}_Y^0 = \{0\}$ and

$$\rho(\mathbf{H}_X^0, \mathbf{H}_Y^0) \equiv \sup\{\langle a, b \rangle: \|a\| = 1, \|b\| = 1\} \le 1 - \alpha < 1.$$

Hence, by Kober (1939) or Kato [(1976), Theorem 4.2, page 219], $\mathbf{H}_X^0 + \mathbf{H}_Y^0$ is closed, and the equations (2.14) and (2.15) uniquely determine the projection onto $\mathbf{H}_X^0 + \mathbf{H}_Y^0$. □

PROOF OF LEMMA 2. (i) Shorack and Wellner [(1986), inequality 10.3.2, page 415]. If $B_n \cong \text{binomial}(n, p)$, then

$$\text{(a)} \qquad P\left(\frac{B_n}{np} \geq \lambda\right) \leq \exp(-nph(\lambda)), \qquad \lambda \geq 1$$

and

$$\text{(b)} \qquad P\left(\frac{np}{B_n} \geq \lambda\right) \leq \exp\left(-nph\left(\frac{1}{\lambda}\right)\right), \qquad \lambda \geq 1,$$

where $h(\lambda) = \lambda(\log \lambda - 1) + 1 > 0$ for $\lambda > 1$ or $\lambda < 1$. Hence

$$\begin{aligned}
&\Pr\left\{\max_{i,j} \frac{p_{ij}}{N_{ij}/n} \geq 1 + \varepsilon\right\} \\
&\quad \leq \sum_{i=1}^{k}\sum_{j=1}^{m} P\left(\frac{p_{ij}}{N_{ij}/n} \geq 1 + \varepsilon\right) \\
&\quad \leq \sum_{i=1}^{k}\sum_{j=1}^{m} \exp\left(-np_{ij}h\left(\frac{1}{1+\varepsilon}\right)\right) \quad \text{[by (b)]} \\
&\quad \leq k(n)m(n)\exp\left(-n\alpha\frac{\gamma_n^2}{n}h\left(\frac{1}{1+\varepsilon}\right)\right) \\
&\quad \leq \frac{n}{\gamma_n^2}\exp\left(-\alpha h\left(\frac{1}{1+\varepsilon}\right)\frac{\gamma_n^2}{\log n}\log n\right) \\
&\quad \leq \frac{n}{\gamma_n^2}n^{-M} \quad \left[\text{for } n \text{ so large that } \alpha h\left(\frac{1}{1+\varepsilon}\right)\frac{\gamma_n^2}{\log n} \geq M \geq 2, \text{ say}\right] \\
&\quad \to 0.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\Pr\left\{\max_{i,j} \frac{N_{ij}/n}{p_{ij}} \geq 1 + \varepsilon\right\} \\
&\quad \leq \sum_{i=1}^{k}\sum_{j=1}^{m} \exp(-np_{ij}h(1+\varepsilon)) \quad \text{[by (a)]} \\
&\quad \leq k(n)m(n)\exp\left(-n\alpha\frac{\gamma_n^2}{n}h(1+\varepsilon)\right) \\
&\quad \leq \frac{n}{\gamma_n^2}\exp\left(-\alpha h(1+\varepsilon)\frac{\gamma_n^2}{\log n}\log n\right) \\
&\quad \leq \frac{n}{\gamma_n^2}n^{-M} \quad \left[\text{for } n \text{ so large that } \alpha h(1+\varepsilon)\frac{\gamma_n^2}{\log n} \geq M \geq 2, \text{ say}\right] \\
&\quad \to 0.
\end{aligned}$$

(ii) As in the proof of (i), but now using both (a) and (b), for any $\varepsilon > 0$,

$$P\left(\max_{i=1,\ldots,k}\left|\frac{N_i./n}{p_i.} - 1\right| \geq \varepsilon\right)$$

$$\leq \sum_{i=1}^{k}\left\{\exp(-np_i.h(1+\varepsilon)) + \exp\left(-np_i.h\left(\frac{1}{1+\varepsilon}\right)\right)\right\}$$

$$\leq \frac{n^{1/2}}{\gamma_n} 2\exp(-n^{1/2}\gamma_n c(\varepsilon)) \quad \left[\text{where } c(\varepsilon) \equiv h(1+\varepsilon)h\left(\frac{1}{1+\varepsilon}\right) > 0\right]$$

$$\to 0 \quad \text{as } n \to \infty.$$

The argument for the $N_{.j}$ margins follows by symmetry.

(iii) The probability in question equals

$$P\left(\max_{i,j}\frac{(N_i./n)(N_{.j}/n)}{N_{ij}/n} > \frac{1}{\gamma}\right)$$

$$= P\left(\max_{i,j}\frac{(N_i./np_i.)(N_{.j}/np_{.j})}{N_{ij}/np_{ij}}\frac{p_i.p_{.j}}{p_{ij}} > \frac{1}{\gamma}\right)$$

$$\leq P\left(\max_{i,j}\frac{(N_i./np_i.)(N_{.j}/np_{.j})}{N_{ij}/np_{ij}} > \frac{\alpha}{\gamma}\right) \quad \text{[by (P3)]}$$

$$\leq P\left(\max_{i}\frac{N_{i.}}{np_i.} > \left(\frac{\alpha}{\gamma}\right)^{1/3}\right) + P\left(\max_{j}\frac{N_{.j}}{np_{.j}} > \left(\frac{\alpha}{\gamma}\right)^{1/3}\right)$$

$$+ P\left(\max_{i,j}\frac{np_{ij}}{N_{ij}} > \left(\frac{\alpha}{\gamma}\right)^{1/3}\right)$$

$$\to 0 \quad \text{as } n \to \infty$$

by (i) and (ii) since $\gamma < \alpha$. □

PROOF OF LEMMA 3. (i) First not that $\{\hat{p}_{ij}\}$ is the point minimizing a strictly convex function over a convex set. A simple argument with Lagrange multipliers $\hat{a}_1, \ldots, \hat{a}_k$ and $\hat{b}_1, \ldots, \hat{b}_m$ shows that (i) holds and that the $\hat{a}_i$'s and $\hat{b}_j$'s satisfy

(a) $$p_{i.}^0 = \frac{N_{i.}}{n} + \frac{N_{i.}}{n}\hat{a}_i + \frac{1}{n}\sum_{j=1}^{m} N_{ij}\hat{b}_j, \qquad i = 1, \ldots, k$$

and

(b) $$p_{.j}^0 = \frac{N_{.j}}{n} + \frac{N_{.j}}{n}\hat{b}_j + \frac{1}{n}\sum_{i=1}^{k} N_{ij}\hat{a}_i, \qquad j = 1, \ldots, m.$$

Equivalently, (a) and (b) can be rewritten as

$$\text{(c)} \qquad \hat{a}_i = \frac{np_{i\cdot}^0}{N_{i\cdot}} - 1 - \frac{\sum_{j=1}^m N_{ij}\hat{b}_j}{N_{i\cdot}}, \qquad i = 1, \ldots, k$$

and

$$\text{(d)} \qquad \hat{b}_j = \frac{np_{\cdot j}^0}{N_{\cdot j}} - 1 - \frac{\sum_{i=1}^k N_{ij}\hat{a}_i}{N_{\cdot j}}, \qquad j = 1, \ldots, m.$$

Note that $\{\hat{a}_i\}$ can be centered arbitrarily; but multiplying across (c) by $N_{i\cdot}$ and summing over i yields

$$\text{(e)} \qquad \sum_{i=1}^k N_{i\cdot}\hat{a}_i = -\sum_{j=1}^m N_{\cdot j}\hat{b}_j$$

and without loss of generality we can assume that the sum is 0:

$$\text{(f)} \qquad \sum_{i=1}^k N_{i\cdot}\hat{a}_i = -\sum_{j=1}^m N_{\cdot j}\hat{b}_j = 0.$$

(ii) First, by Lemma 2(iii), for $0 < \gamma < \alpha$ it follows that

$$\text{(g)} \qquad \frac{N_{ij}}{n} - \gamma\frac{N_{i\cdot}}{n}\frac{N_{\cdot j}}{n} > 0$$

for all $i = 1, \ldots, k$, $j = 1, \ldots, m$ with probability converging to 1. Suppose (g) holds (for all i, j). Then from (c) it follows that

$$\text{(h)} \qquad \begin{aligned} \hat{a}_i &= \frac{np_{i\cdot}^0}{N_{i\cdot}} - 1 - \frac{\sum_{j=1}^m \left[N_{ij} - (\gamma/n)N_{i\cdot}N_{\cdot j}\right]\hat{b}_j}{N_{i\cdot}} - \frac{\gamma}{n}\frac{\sum_{j=1}^m N_{i\cdot}N_{\cdot j}\hat{b}_j}{N_{i\cdot}} \\ &= \frac{np_{i\cdot}^0}{N_{i\cdot}} - 1 - \frac{1}{N_{i\cdot}}\sum_{j=1}^m \left(N_{ij} - \frac{\gamma}{n}N_{i\cdot}N_{\cdot j}\right)\hat{b}_j \quad \text{by (f)} \end{aligned}$$

for $i = 1, \ldots, k$. Hence

$$\begin{aligned} |\hat{a}_i| &\le \left|\frac{np_{i\cdot}^0}{N_{i\cdot}} - 1\right| + \max_j |\hat{b}_j| \frac{1}{N_{i\cdot}}\sum_{j=1}^m \left(N_{ij} - \frac{\gamma}{n}N_{i\cdot}N_{\cdot j}\right) \\ &= \left|\frac{np_{i\cdot}^0}{N_{i\cdot}} - 1\right| + \max_j |\hat{b}_j|(1-\gamma) \end{aligned}$$

or

$$\text{(i)} \qquad \max_i |\hat{a}_i| \le \max_i \left|\frac{np_{i\cdot}^0}{N_{i\cdot}} - 1\right| + (1-\gamma)\max_j |\hat{b}_j|.$$

Similarly

$$\text{(j)} \qquad \max_j |\hat{b}_j| \le \max_j \left| \frac{np^0_{\cdot j}}{N_{\cdot j}} - 1 \right| + (1-\gamma) \max_i |\hat{a}_i|.$$

Summing (i) and (j) and algebra yield

$$\text{(k)} \qquad \gamma \left\{ \max_i |\hat{a}_i| + \max_j |\hat{b}_j| \right\} \le \max_i \left| \frac{np^0_{i\cdot}}{N_{i\cdot}} - 1 \right| + \max_j \left| \frac{np^0_{\cdot j}}{N_{\cdot j}} - 1 \right| \to_p 0 \quad \text{as } n \to \infty$$

by Lemma 2 (ii). □

PROOF OF LEMMA 4. We first compute conditionally on $\underline{N}$:

$$E\left\{ \left(\sum_{i,j} \hat{p}_{ij} \left(\hat{h}_{ij} - \overline{h}_{ij} \right) \right)^2 \middle| \underline{N} \right\}$$

$$= \sum_{i,j} \hat{p}^2_{ij} E\left\{ \left(\hat{h}_{ij} - \overline{h} \right)^2 \middle| \underline{N} \right\}$$

$$\text{(a)} \qquad = \sum_{i,j} \frac{\hat{p}^2_{ij}}{N_{ij} p_{ij}} \int\int_{A_i \times B_j} \left(h(x,y) - \overline{h}(x,y) \right)^2 dP(x,y) 1_{[N_{ij} > 0]}$$

$$= \left(1 + o_p(1) \right) \sum_{i,j} \frac{N_{ij}}{n^2 p_{ij}} \int\int_{A_i \times B_j} \left(h(x,y) - \overline{h}(x,y) \right)^2 dP(x,y)$$

by Lemma 3(ii), but

$$E\left(\sum_{i,j} \frac{N_{ij}}{n^2 p_{ij}} \int\int_{A_i \times B_j} \left(h(x,y) - \overline{h}(x,y) \right)^2 dP(x,y) \right)$$

$$\text{(b)} \qquad = \frac{1}{n} \int\int \left(h(x,y) - \overline{h}(x,y) \right)^2 dP(x,y)$$

$$= o\left(\frac{1}{n} \right) \quad \text{by (F2).}$$

Combining (a) and (b) yields the claim. □

PROOF OF LEMMA 5. First note that

$$\text{(a)} \quad \theta_h(P) = \int\int h(x,y)\, dP(x,y) = \int\int \overline{h}(x,y)\, dP(x,y) = \sum_{i,j} \overline{h}_{ij} p_{ij}.$$

Then

$$\sum_i \left(p_{i\cdot} - \frac{N_{i\cdot}}{n}\right)(\hat{u}_i - \bar{u}_i) + \sum_j \left(p_{\cdot j} - \frac{N_{\cdot j}}{n}\right)(\hat{v}_j - \bar{v}_j)$$

$$\text{(b)} \quad = \frac{1}{n}\sum_{i,j} N_{ij}(\bar{u}_i + \bar{v}_j - \hat{u}_i - \hat{v}_j) - \sum_{i,j} p_{ij}(\bar{u}_i + \bar{v}_j) + \sum_i p_{i\cdot}\,\hat{u}_i + \sum_j p_{\cdot j}\hat{v}_j$$

$$= \frac{1}{n}\sum_{i,j} N_{ij}(\bar{u}_i + \bar{v}_j - \hat{u}_i - \hat{v}_j) - \sum_{i,j} p_{ij}(\bar{u}_i + \bar{v}_j) + \sum_{i,j} \hat{p}_{ij}(\hat{u}_i + \hat{v}_j),$$

since $\hat{p}_{i\cdot} = p_{i\cdot} = p^0_{i\cdot}$ and $\hat{p}_{\cdot j} = p_{\cdot j} = p^0_{\cdot j}$. Substitution of (a) and (b) into the right side of (4.7) yields

$$\sum_{i,j} p_{ij}\bar{h}_{ij} + \frac{1}{n}\sum_{i,j} N_{ij}(\bar{h}_{ij} - \bar{u}_i - \bar{v}_j) + \frac{1}{n}\sum_{i,j} N_{ij}(\bar{u}_i + \bar{v}_j - \hat{u}_i - \hat{v}_j)$$
$$- \sum_{i,j} p_{ij}(\bar{u}_i + \bar{v}_j) + \sum_{i,j} \hat{p}_{ij}(\hat{u}_i + \hat{v}_j)$$

$$= \sum_{i,j} p_{ij}(\bar{h}_{ij} - \bar{u}_i - \bar{v}_j) + \frac{1}{n}\sum_{i,j} N_{ij}(\bar{h}_{ij} - \hat{u}_i - \hat{v}_j) + \sum_{i,j} \hat{p}_{ij}(\hat{u}_i + \hat{v}_j)$$

$$= 0 + \frac{1}{n}\sum_{i,j} N_{ij}(\bar{h}_{ij} - \hat{u}_i - \hat{v}_j) + \sum_{i,j} \hat{p}_{ij}(\hat{u}_i + \hat{v}_j) \quad \text{[by (4.3) or (4.4)]}$$

$$= \frac{1}{n}\sum_{i,j} N_{ij}(1 + \hat{a}_i + \hat{b}_j)(\bar{h}_{ij} - \hat{u}_i - \hat{v}_j) + \sum_{i,j} \hat{p}_{ij}(\hat{u}_i + \hat{v}_j)$$

[by (4.5) or (4.6)]

$$= \sum_{i,j} \hat{p}_{ij}\bar{h}_{ij} \quad \text{[by Lemma 3(i)]},$$

thus completing the proof. □

PROOF OF LEMMA 6. Note that

$$\text{(a)} \qquad \bar{l} \equiv \sum_{i=1}^{k}\sum_{j=1}^{m} 1_{A_{n,j}\times B_{n,j}}(\bar{h}_{ij} - \bar{u}_i - \bar{v}_j)$$

is the projection of $\bar{h}$ onto $\{E(w|\mathbf{F}_n)\colon w \in L_2(P)\} \cap (\mathbf{H}_X + \mathbf{H}_Y)^{\perp}$ and therefore, with $l \equiv h - u - v$,

$$\text{(b)} \qquad \begin{aligned} nC_n^2 &= n\,\mathrm{Var}\left(\int\int (\bar{l} - l)\, d\mathbb{P}_n\right) \\ &= \int\int (\bar{l} - l)^2\, dP \\ &\le \int\int (\bar{h} - h)^2\, dP \\ &\to 0 \end{aligned}$$

since $\bar{h} = E(h|\mathbf{F}_n)$ is an L_2-bounded martingale [recall (2.2)]. □

PROOF OF LEMMA 7. By Cauchy–Schwarz,

$$\text{(a)}\quad \left\{\sum_i \left(p_{i\cdot} - \frac{N_{i\cdot}}{n}\right)(\hat{u}_i - \bar{u}_i)\right\}^2 \le \sum_{i=1}^k \frac{(p_{i\cdot} - N_{i\cdot}/n)^2}{p_{i\cdot}} \sum_{i=1}^k p_{i\cdot}(\hat{u}_i - \bar{u}_i)^2.$$

The first factor can be bounded easily since

$$\text{(b)}\quad E\left\{\sum_{i=1}^k \frac{(p_{i\cdot} - N_{i\cdot}/n)^2}{p_{i\cdot}}\right\} = \frac{1}{n}\sum_{i=1}^k (1 - p_{i\cdot}) = \frac{k(n) - 1}{n} \le \frac{1}{\gamma_n n^{1/2}}.$$

To bound the second factor on the right side in (a), we first define operators

$$\bar{P}_X\colon L_2(\mathbb{X}\times\mathbb{Y}, \mathbf{F}_n, P) \to \bar{\mathbf{H}}_X,$$
$$\bar{P}_Y\colon L_2(\mathbb{X}\times\mathbb{Y}, \mathbf{F}_n, P) \to \bar{\mathbf{H}}_Y,$$
$$\hat{P}_X\colon L_2(\mathbb{X}\times\mathbb{Y}, \mathbf{F}_n, \mathbb{P}_n) \to \hat{\mathbf{H}}_X,$$
$$\hat{P}_Y\colon L_2(\mathbb{X}\times\mathbb{Y}, \mathbf{F}_n, \mathbb{P}_n) \to \hat{\mathbf{H}}_Y$$

as follows: For $w(x,y) = \sum_{ij} w(i,j) 1_{A_{ni}\times B_{nj}}(x,y)$, let

$$\text{(c)}\quad \bar{P}_X w(x) = E_{XY}(w(X,Y)|\mathbf{F}_{nX})(x) = \sum_{i=1}^k \left(\sum_{j=1}^m p_{ij} w(i,j)/p_{i\cdot}\right) 1_{A_{ni}}(x),$$

$$\text{(d)}\quad \bar{P}_Y w(y) = E_{XY}(w(X,Y)|\mathbf{F}_{nY})(y) = \sum_{j=1}^m \left(\sum_{i=1}^k p_{ij} w(i,j)/p_{\cdot i}\right) 1_{B_{nj}}(y),$$

$$\text{(e)}\quad \hat{P}_X w(x) = \mathbb{E}_n(w(X,Y)|\mathbf{F}_{nX})(x) = \sum_{i=1}^k \left(\sum_{j=1}^m N_{ij} w(i,j)/N_{i\cdot}\right) 1_{A_{ni}}(x),$$

$$\text{(f)}\quad \hat{P}_Y w(y) = \mathbb{E}_n(w(X,Y)|\mathbf{F}_{nY})(y) = \sum_{j=1}^m \left(\sum_{i=1}^k N_{ij} w(i,j)/N_{\cdot i}\right) 1_{B_{nj}}(y),$$

where $E_{XY}, \mathbb{E}_n$ denote expectation with respect to $P, \mathbb{P}_n$, respectively. Note that we can represent these operators as matrices: In an obvious notation [as in (2.9) and (2.10)],

$$\bar{P}_X(w)(i) = \sum_{j=1}^m p_{ij} w(i,j)/p_{i\cdot},$$
$$\bar{P}_Y(w)(j) = \sum_{i=1}^k p_{ij} w(i,j)/p_{\cdot j};$$
$$\hat{P}_X(w)(i) = \sum_{j=1}^m N_{ij} w(i,j)/N_{i\cdot},$$
$$\hat{P}_Y(w)(j) = \sum_{i=1}^k N_{ij} w(i,j)/N_{\cdot j}.$$

For the remainder of this proof we go back and forth between the operators themselves and their representations in terms of matrices as needed.

In terms of the operators $\bar{P}_X$ and $\bar{P}_Y$, (3.3) and (3.4) can be written as

(g) $$\bar{P}_X(\bar{h} - \bar{u} - \bar{v}) = 0$$

and

(h) $$\bar{P}_Y(\bar{h} - \bar{u} - \bar{v}) = 0$$

or

(i) $$\bar{u} = \bar{P}_X(\bar{h} - \bar{v}),$$

(j) $$\bar{v} = \bar{P}_Y(\bar{h} - \bar{u}).$$

Substitution of (j) into (i) and rearranging yields

(k) $$(I - \bar{P}_X\bar{P}_Y)\bar{u} = \bar{P}_X(I - \bar{P}_Y)\bar{h},$$

where I denotes the identity operator. Reversing this process gives

(l) $$(I - \bar{P}_Y\bar{P}_X)\bar{v} = \bar{P}_Y(I - \bar{P}_X)\bar{h}.$$

As an operator from $\bar{H}_X$ to $\bar{H}_X$, $\bar{A} \equiv I - \bar{P}_X\bar{P}_Y = I - \bar{P}_X\bar{P}_Y\bar{P}_X$ is self-adjoint, has range $\mathbf{R}(\bar{A}) \subset \bar{\mathbf{H}}_X^0 \equiv \bar{\mathbf{H}}_X \cap \{\text{constants}\}^\perp$ and has null space equal to the constant functions. Moreover, (P3) implies that on $\bar{\mathbf{H}}_X^0$,

(m) $$\left\| \bar{P}_X\bar{P}_Y\bar{P}_X \right\| \leq (1 - \alpha)^2 < 1 - \alpha,$$

which will be proved below. Hence the minimal eigenvalue of $\bar{A}$ as an operator from $\bar{\mathbf{H}}_X^0$ to $\bar{\mathbf{H}}_X^0$ is bounded below by α. It follows that $\bar{A}^{-1} = (I - \bar{P}_X\bar{P}_Y\bar{P}_X)^{-1}$ exists and

(n) $\bar{A}^{-1} = \left(I - \bar{P}_X\bar{P}_Y\bar{P}_X\right)^{-1}$ is bounded (uniformly in k, m) on $\bar{\mathbf{H}}_X^0$.

By symmetry, the same is true for $\bar{B} \equiv I - \bar{P}_Y\bar{P}_X\bar{P}_Y$ as an operator on $\bar{\mathbf{H}}_Y^0$:

(o) $\bar{B}^{-1} = \left(I - \bar{P}_Y\bar{P}_X\bar{P}_Y\right)^{-1}$ is bounded (uniformly in k, m) on $\bar{\mathbf{H}}_Y^0$.

To prove (m, let $w \in \bar{\mathbf{H}}_X^0$, so that $w(x) = \sum_{i=1}^k w_i 1_{A_{ni}}(x)$ with $\sum_{i=1}^k w_i p_{i\cdot} = 0$. Then

$$\bar{P}_Y w(y) = \sum_{j=1}^m \left\{ \sum_{i=1}^k \frac{p_{ij}}{p_{\cdot j}} w_i \right\} 1_{B_{nj}}(y)$$

or

$$\bar{P}_Y w(j) = \sum_{i=1}^k \frac{p_{ij}}{p_{\cdot j}} w_i = \sum_{i=1}^k \frac{(p_{ij} - \alpha p_{i\cdot} p_{\cdot j})}{p_{\cdot j}} w_i$$

and hence [using (P3) implies $p_{ij} - \alpha p_{i\cdot}p_{\cdot j} \geq 0$]

$$\begin{aligned}
\|\bar{P}_Y w\|_2^2 &= \sum_{j=1}^{m} \left\{ \sum_{i=1}^{k} \frac{(p_{ij} - \alpha p_{i\cdot}p_{\cdot j})}{p_{\cdot j}} w_i \right\}^2 p_{\cdot j} \\
&\leq \sum_{j=1}^{m} p_{\cdot j} \left\{ \sum_{i} \frac{(p_{ij} - \alpha p_{i\cdot}p_{\cdot j})}{p_{\cdot j}} w_i^2 \right\} \left\{ \sum_{i} \frac{(p_{ij} - \alpha p_{i\cdot}p_{\cdot j})}{p_{\cdot j}} \right\} \\
&= \sum_{i,j} (p_{ij} - \alpha p_{i\cdot}p_{\cdot j}) w_i^2 (1 - \alpha) \\
&= \|w\|_2^2 (1 - \alpha)^2.
\end{aligned}$$

Since $\bar{P}_Y w \in \bar{\mathbf{H}}_Y^0$ if $w \in \bar{\mathbf{H}}_X^0$, we can repeat this argument with $\bar{P}_Y$ replaced by $\bar{P}_X$ and w replaced by $\bar{P}_Y w$ to obtain

$$\|\bar{P}_X \bar{P}_Y w\|_2^2 \leq \|\bar{P}_Y w\|_2^2 (1 - \alpha)^2 \leq \|w\|_2^2 (1 - \alpha)^4$$

and this proves (m).

Now we show that

(p) $$\|\hat{P}_X - \bar{P}_X\| \to_p 0 \quad \text{as } n \to \infty$$

and

(q) $$\|\hat{P}_Y - \bar{P}_Y\| \to_p 0 \quad \text{as } n \to \infty,$$

where $\|\cdot\|$ denotes the operator norm. But, for $w \in \{E(w|\mathbf{F}_n)\colon w \in L_2(P)\}$,

$$\begin{aligned}
(\hat{P}_X - \bar{P}_X) w(x) &= \sum_{i=1}^{k} \left\{ \sum_{j=1}^{m} w_{ij} \frac{N_{ij}/N_{i\cdot} - p_{ij}/p_{i\cdot}}{(p_{ij}/p_{i\cdot})} \frac{p_{ij}}{p_{i\cdot}} \right\} 1_{A_{ni}}(x) \\
&= \sum_{i=1}^{k} \left\{ \sum_{j=1}^{m} w_{ij} D_{ij} \frac{p_{ij}}{p_{i\cdot}} \right\} 1_{A_{ni}}(x),
\end{aligned}$$

where

$$D_{ij} \equiv \frac{N_{ij}/np_{ij}}{N_{i\cdot}/np_{i\cdot}} - 1.$$

Therefore

$$\begin{aligned}
\|(\hat{P}_X - \bar{P}_X) w\|_2^2 &= \sum_{i=1}^{k} \left\{ \sum_{j=1}^{m} w_{ij} D_{ij} \frac{p_{ij}}{p_{i\cdot}} \right\}^2 p_{i\cdot} \\
&= E_{XY} E_{XY} \{ \bar{w} \bar{D} | \mathbf{F}_{nX} \}^2 \\
&\leq E_{XY} (\bar{w}^2 \bar{D}^2) \\
&\leq \max_{i,j} D_{ij}^2 E_{XY} (\bar{w}^2) \\
&= o_p(1) \|w\|_2^2 \quad \text{by Lemma 2,}
\end{aligned}$$

which proves (p), and (q) follows by symmetry.

It follows from (p), (q) and the fact that $\bar{A}$ has minimal eigenvalue bounded below by α that $\hat{A} \equiv I - \hat{P}_X\hat{P}_Y\hat{P}_X$ (as an operator on $\bar{\mathbf{H}}_X^0$) has minimal eigenvalue bounded below by $\alpha/2$ with probability converging to 1. Hence with probability arbitrarily close to 1 for n large, $\hat{A}^{-1} = (I - \hat{P}_X\hat{P}_Y\hat{P}_X)^{-1}$ exists and is bounded.

Now note that

$$w_X \equiv \bar{P}_X(I - \bar{P}_Y)\bar{h} \in \bar{\mathbf{H}}_X^0,$$

$$w_Y \equiv \bar{P}_Y(I - \bar{P}_X)\bar{h} \in \bar{\mathbf{H}}_Y^0.$$

Hence the solutions $\bar{u}, \bar{v}$ of (k) and (l) can be written as

$$\bar{u} = \left(I - \bar{P}_X\bar{P}_Y\bar{P}_X\right)^{-1}\left(\bar{P}_X\bar{h} - \bar{P}_X\bar{P}_Y\bar{h}\right) + \sum_{i,j} p_{ij}\bar{h}_{ij}, \tag{r}$$

$$\bar{v} = \left(I - \bar{P}_Y\bar{P}_X\bar{P}_Y\right)^{-1}\left(\bar{P}_Y\bar{h} - \bar{P}_Y\bar{P}_X\bar{h}\right). \tag{s}$$

Similarly, with probability converging to 1, the solutions $\hat{u}, \hat{v}$ of the analogous equations corresponding to (3.5) and (3.6) can be written (with probability arbitrarily close to 1 for large n) as

$$\hat{u} = \left(I - \hat{P}_X\hat{P}_Y\hat{P}_X\right)^{-1}\left(\hat{P}_X\bar{h} - \hat{P}_X\hat{P}_Y\bar{h}\right) + \frac{1}{n}\sum_{i,j} N_{ij}\bar{h}_{ij}, \tag{t}$$

$$\hat{v} = \left(I - \hat{P}_Y\hat{P}_X\hat{P}_Y\right)^{-1}\left(\hat{P}_Y\bar{h} - \hat{P}_Y\hat{P}_X\bar{h}\right). \tag{u}$$

Therefore we can write

$$\begin{aligned}
\hat{u} - \bar{u} &= \left(I - \hat{P}_X\hat{P}_Y\hat{P}_X\right)^{-1}\left(\left(\hat{P}_X - \bar{P}_X\right)\bar{h}\right) \\
&\quad + \left(I - \hat{P}_X\hat{P}_Y\hat{P}_X\right)^{-1}\left(\left(\bar{P}_X\bar{P}_Y - \hat{P}_X\hat{P}_Y\right)\bar{h}\right) \\
&\quad + \left\{\left(I - \hat{P}_X\hat{P}_Y\hat{P}_X\right)^{-1} - \left(I - \bar{P}_X\bar{P}_Y\bar{P}_X\right)^{-1}\right\}\left(\bar{P}_X\bar{h} - \bar{P}_X\bar{P}_Y\bar{h}\right) \\
&\quad + \sum_{i,j}\left(\frac{N_{ij}}{n} - p_{ij}\right)\bar{h}_{ij} \\
&\equiv R_1 + R_2 + R_3 + R_4.
\end{aligned} \tag{v}$$

Now the constant term R_4 makes no contribution to the left side of (a) and hence (a) can be replaced by

$$\begin{aligned}
&\left\{\sum_i\left(p_{i\cdot} - \frac{N_{i\cdot}}{n}\right)(\hat{u}_i - \bar{u}_i)\right\}^2 \\
&\qquad \le \sum_{i=1}^k \frac{(p_{i\cdot} - N_{i\cdot}/n)^2}{p_{i\cdot}} \sum_{i=1}^k p_{i\cdot}\{(R_1 + R_2 + R_3)(i)\}^2,
\end{aligned} \tag{w}$$

where

$$\text{(x)}\qquad \begin{aligned}&\sum_{i=1}^{k} p_{i\cdot}\{(R_1 + R_2 + R_3)(i)\}^2 \\ &\quad \le 4\left\{\sum_{i=1}^{k} p_{i\cdot}R_1^2(i) + \sum_{i=1}^{k} p_{i\cdot}R_2^2(i) + \sum_{i=1}^{k} p_{i\cdot}R_3^2(i)\right\} \\ &\quad \equiv A + B + C.\end{aligned}$$

To handle the terms A–C, we first show that

$$\text{(y)}\qquad \sum_i p_{i\cdot}\left\{\left(\hat{P}_X w - \bar{P}_X w\right)(i)\right\}^2 = O_p\left(\frac{1}{\gamma_n\sqrt{n}}\right)$$

and

$$\text{(z)}\qquad \sum_j p_{\cdot j}\left\{\left(\hat{P}_Y w - \bar{P}_Y w\right)(j)\right\}^2 = O_p\left(\frac{1}{\gamma_n\sqrt{n}}\right)$$

for nonrandom functions w and, for any $\varepsilon > 0$,

$$\text{(aa)}\qquad \sum_i p_{i\cdot}\left\{\hat{P}_X w(i)\right\}^2 \le (1+\varepsilon)^2 \sum_{i,j} p_{ij} w^2(i,j)$$

with probability converging to 1 as $n \to \infty$ for random or nonrandom functions w.

The arguments for (y) and (z) are the same by symmetry, so it suffices to prove (y). To prove (y), we argue conditionally on $N_{i\cdot}$. First note that

$$((N_{i1}, \ldots, N_{im})|N_{i\cdot}) \cong \text{mult}_m\left(N_{i\cdot}; \left(\frac{p_{i1}}{p_{i\cdot}}, \ldots, \frac{p_{im}}{p_{i\cdot}}\right)\right)$$

for $i = 1, \ldots, k$. Therefore

$$\begin{aligned}&E\left\{\left(\hat{P}_X w(i) - \bar{P}_X w(i)\right)^2 \middle| N_{i\cdot}\right\} \\ &\quad = \frac{1}{N_{i\cdot}^2}\text{Var}\left[\sum_j N_{ij} w(i,j) \middle| N_{i\cdot}\right] \\ &\quad = \frac{1}{N_{i\cdot}}\left\{\sum_j \frac{p_{ij}}{p_{i\cdot}}\left(1 - \frac{p_{ij}}{p_{i\cdot}}\right)w^2(i,j) - \sum_{j \ne j'} \frac{p_{ij}}{p_{i\cdot}}\frac{p_{ij'}}{p_{i\cdot}} w(i,j)w(i,j')\right\}; \qquad \text{(bb)} \\ &\quad = \frac{1}{N_{i\cdot}}\left\{\sum_j \frac{p_{ij}}{p_{i\cdot}} w^2(i,j) - \left(\sum_j \frac{p_{ij}}{p_{i\cdot}} w(i,j)\right)^2\right\} \\ &\quad \le \frac{1}{N_{i\cdot}}\sum_j \frac{p_{ij}}{p_{i\cdot}} w^2(i,j).\end{aligned}$$

Hence

$$\sum_i p_{i\cdot}\left(\hat{P}_X w(i) - \bar{P}_X w(i)\right)^2$$

$$= O_p\left(\sum_i \frac{1}{N_{i\cdot}} \sum_{j=1}^{m} p_{ij} w^2(i,j)\right)$$

(cc)
$$= O_p\left(\max_i \frac{np_{i\cdot}}{N_{i\cdot}}\right)\left(\max_i \frac{1}{np_{i\cdot}} \sum_{i,j} p_{ij} w^2(i,j)\right)$$

$$\leq \frac{\sqrt{n}}{n\gamma_n} O_p(1) \quad \text{[by Lemma 2(ii)]}$$

$$\leq \frac{1}{\gamma_n \sqrt{n}} O_p(1)$$

and (y) holds. Finally, (aa) follows from Lemma 2 and the fact that $\hat{P}_X$ is a contraction in $L_2(\mathbb{P}_n)$.

Now we have

(dd)
$$A = \sum_i p_{i\cdot} R_1^2(i) = O_p\left(\frac{1}{\gamma_n \sqrt{n}}\right)$$

by (n) and (y);

$$R_2 = \left(I - \hat{P}_X \hat{P}_Y \hat{P}_X\right)^{-1}\left(\left(\bar{P}_X \bar{P}_Y - \hat{P}_X \hat{P}_Y\right)\bar{h}\right)$$

(ee)
$$= \left(I - \hat{P}_X \hat{P}_Y \hat{P}_X\right)^{-1}\left\{\left(\bar{P}_X - \hat{P}_X\right)\bar{P}_Y \bar{h} + \hat{P}_X\left(\bar{P}_Y - \hat{P}_Y\right)\bar{h}\right\}$$

$$\equiv R_{21} + R_{22},$$

where

(ff)
$$B_1 \equiv \sum_i p_{i\cdot} R_{21}^2(i) = O_p\left(\frac{1}{\gamma_n \sqrt{n}}\right)$$

by (y) and $P(\|\hat{A}^{-1}\| > 2/\alpha) \to 0$ and

(gg)
$$B_2 \equiv \sum_i p_{i\cdot} R_{22}^2(i) = O_p\left(\frac{1}{\gamma_n \sqrt{n}}\right)$$

by (z), (aa) and $P(\|\hat{A}^{-1}\| > 2/\alpha) \to 0$. To handle C, write

$$R_3 = \left\{\left(I - \hat{P}_X\hat{P}_Y\hat{P}_X\right)^{-1} - \left(I - \bar{P}_X\bar{P}_Y\bar{P}_X\right)^{-1}\right\}\left(\bar{P}_X\bar{h} - \bar{P}_X\bar{P}_Y\bar{h}\right)$$

$$= \left\{\left(I - \hat{P}_X\hat{P}_Y\hat{P}_X\right)^{-1} - \left(I - \bar{P}_X\bar{P}_Y\bar{P}_X\right)^{-1}\right\}w$$

(hh) $$\left[\text{say, where } w \equiv \bar{P}_X\left(I - \bar{P}_Y\right)\bar{h} \in \bar{\mathbf{H}}_X^0\right]$$

$$= \left(I - \hat{P}_X\hat{P}_Y\hat{P}_X\right)^{-1}\left(\hat{P}_X\hat{P}_Y - \bar{P}_X\bar{P}_Y\right)\left(I - \bar{P}_X\bar{P}_Y\bar{P}_X\right)^{-1}w$$

$$= \left(I - \bar{P}_X\hat{P}_Y\hat{P}_X\right)^{-1}\left(\hat{P}_X\hat{P}_Y - \bar{P}_X\bar{P}_Y\right)w' \quad \text{say,}$$

so that

(ii) $$C = \sum_i p_{i\cdot} R_3^2(i) = O_p\left(\frac{1}{\gamma_n\sqrt{n}}\right)$$

by the same argument as for B.

Combining (a), (b), (v) and (dd)–(ii) completes the proof. □

Acknowledgments. We owe thanks to Andreas Buja for help with the literature on spectral decompositions of bivariate distributions and to Richard Olshen for help with references leading to the work of von Neumann and the early history of ACE. The third author's work on this article was conducted while visiting the Mathematics Institute of the University of Leiden and the Centrum voor Wiskunde en Informatica, Amsterdam, thanks to the cordial hospitality of W. van Zwet and R. Gill.

REFERENCES

ARONSZAJN, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** 337–404.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1991). *Efficient and Adaptive Estimation in Semiparametric Models*. Johns Hopkins Univ. Press. To appear.

BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–598.

BUJA, A. (1985). Theory of bivariate ACE, Technical Report 74, Dept. Statistics, Univ. Washington.

BUJA, A., HASTIE, T. and TIBSHIRANI, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17** 453–510.

CAMBANIS, S., SIMONS, G. and STOUT, W. (1976). Inequalities for $Ek(X, Y)$ when the marginals are fixed. *Z. Wahrsch. Verw. Gebiete.* **36** 285–294.

CHESSON, P. L. (1976). The canonical decomposition of bivariate distributions. *J. Multivariate Anal.* **6** 526–537.

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

DABROWSKA, D. (1988). Kaplan–Meier estimation on the plane. *Ann. Statist.* **16** 1475–1489.

DAÜXOIS, J. and POUSSE, A. (1975). Une estension de l'analyse canonique; quelques applications. *Ann. Inst. H. Poincaŕe Sect. B* (*N.S.*) **11** 335–379.

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** 423–444.

DEUTSCH, F. (1985). Rate of convergence of the method of alternating projections. In *Parametric Optimization and Approximation* (B. Brosowski and F. Deutsch, eds.) 96–107. Birkhäuser, Basel.

EAGLESON, G. K. (1964). Polynomial expansions of bivariate distributions. *Ann. Math. Statist.* **35** 1208–1215.

FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7** 179–188.

FRÉCHET, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon Sect. A* **14** 53–77.

GILL, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part I). *Scand. J. Statist.* **16** 97–128.

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

HOEFFDING, W. (1940). Maassstabinvariante korrelations theorie. *Schr. Math. Inst. Angewandte Mathematik der Univ. Berlin* **5** 179–233.

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

JAGERS, P., ODÉN, A. and TRULSSON, L. (1985). Post-stratification and ratio estimation: Usages of auxiliary information in survey sampling and opinion polls. *Internat. Statist. Rev.* **53** 221–238.

KATO, T. (1976). *Perturbation Theory of Linear Operators*, 2nd ed. Springer, New York.

KAYALAR, S. and WEINERT, H. L. (1988). Error bounds for the method of alternating projections. *Math. Control Signals Systems* **1** 43–59.

KOBER, H. (1939). A theorem on Banach spaces. *Compositio Math.* **7** 135–140.

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

LANCASTER, H. O. (1958). The structure of bivariate distributions. *Ann. Math. Statist.* **29** 719–736.

LANCASTER, H. O. (1963). Correlations and canonical forms of bivarite distributions. *Ann. Math. Statist.* **34** 532–538.

LEVIT, B. YA. (1978). Infinite-dimensional informational inequalities. *Theory Probab. Appl.* **23** 371–377.

MARSHALL, A. (1986). Personal communication.

MARSHALL, A. and OLKIN, I. (1979). *Inequalities: Theory of Majorization and Its Applications.* Academic, New York.

NAKANO, H. (1953). *Spectral Theory in Hilbert Space.* Japanese Society for the Promotion of Science, Tokyo.

PFANZAGL, J. (with W. WEFELMEYER) (1982). *Contributions to a General Asymptotic Statistical Theory. Lecture Notes in Statist.* **13**. Springer, New York.

RÉNYI, A. (1959). On measures of dependence. *Acta Math. Acad. Sci. Hungar.* **10** 441–451.

RÜSCHENDORF, L. (1984). On the minimum discrimination information theorem. *Statist. Decisions Suppl.* **1** 263–283.

SCHRIEVER, B. F. (1986). Order dependence. *CWI Tract* **20**. Centrum voor Wiskunde en Informatica, Amsterdam.

SHEEHY, A. (1987). Constrained estimation and clustering based on minimum Kullback–Leibler information methods. Ph.D. dissertation, Dept. Statistics, Univ. Washington.

SHEEHY, A. (1988). Kullback–Leibler constrained estimation of probability measures. Report, Dept. Statistics, Stanford Univ.

SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York.

STUTE, W. (1984). The oscillation behavior of empirical processes: The multivariate case. *Ann. Probab.* **12** 361–379.

VAN DER VAART, A. W. (1988). Statistical estimation in large parameter spaces. *CWI Tract* **44**. Centrum voor Wiskunde en Informatica, Amsterdam.

VENTER, J. H. (1967). Probability measures on product spaces. *South African Statist. J.* **1** 3–20.
VITALE, R. A. (1979). Regression with given marginals. *Ann. Statist.* **7** 653–658.
VON NEUMANN, J. (1949). On rings of operators. Reduction theory. *Ann. of Math.* **50** 401–485.
VON NEUMANN, J. (1950). Functional Operators, Volume II: The Geometry of Orthogonal Spaces. *Annals of Mathematics Studies* **22**. Princeton Univ. Press.
WELLNER, J. A. (1989). Discussion of "Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part I)" by R. D. Gill. *Scand. J. Statist.* **16** 124–127.
WHITT, W. (1976). Bivariate distributions with given marginals. *Ann. Statist.* **4** 1280–1289.
WIENER, N. (1955). On the factorization of matrices. *Comment. Math. Helv.* **29** 97–110.

PETER J. BICKEL
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

YA'ACOV RITOV
DEPARTMENT OF STATISTICS
HEBREW UNIVERSITY
JERUSALEM 91905
ISRAEL

JON A. WELLNER
DEPARTMENT OF STATISTICS, GN-22
B320 PADELFORD HALL
UNIVERSITY OF WASHINGTON,
SEATTLE, WASHINGTON 98195