EFFICIENT ESTIMATION OF LINEAR FUNCTIONALS OF A BIVARIATE DISTRIBUTION WITH EQUAL, BUT UNKNOWN, MARGINALS: THE LEAST SQUARES APPROACH

By Hanxiang Peng and Anton Schick[1]

University of Mississippi and Binghamton University

In this paper we characterize and construct efficient estimators of linear functionals of a bivariate distribution with equal marginals. An efficient estimator equals the empirical estimator minus a correction term and provides significant improvements over the empirical estimator. We construct an efficient estimator by estimating the correction term. For this we use the least squares principle and an estimated orthonormal basis for the Hilbert space of square-integrable functions under the unknown equal marginal distribution. Simulations confirm the asymptotic behavior of this estimator in moderate sample sizes and the considerable theoretical gains over the empirical estimator.

1. Introduction. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be independent copies of a bivariate random vector (X, Y) with distribution Q. Let ψ be a measurable function from $\mathbb{R}^2$ to $\mathbb{R}$ such that $\int \psi^2 \, dQ < \infty$. We are interested in estimating

$$\theta = \int \psi \, dQ = E(\psi(X, Y)).$$

Special cases are the estimation of mixed moments $E[X^k Y^m]$, which can be used in the estimation of the covariance of X and Y and the correlation coefficient of X and Y. Of interest is also the estimation of moments of transformed variables $Z = h(X, Y)$ such as $Z = X$, $Z = Y$, $Z = X + Y$, $Z = \min(X, Y)$ and $Z = \max(X, Y)$, or the estimation of probabilities such as $P(X < Y)$, $P(X + Y \le t)$, $P(\min(X, Y) > t)$, $P(\max(X, Y) \le t)$ and $P(X \le s, Y \le t)$ for fixed s and t in $\mathbb{R}$.

A natural estimator of θ is the empirical estimator

$$\frac{1}{n} \sum_{j=1}^{n} \psi(X_j, Y_j).$$

This estimator is efficient in the sense of being a least dispersed regular estimator if the distribution Q is completely unknown. There are however better estimators

---

AMS 2000 subject classifications. Primary 62G05, 62G20

Key words and phrases. Least dispersed regular estimator, least squares estimators, efficient influence function, empirical estimator, local asymptotic normality.

[1]Supported in part by NSF Grant DMS 0072174

if additional information about Q is available. For example, if X and Y are independent, a better estimator is given by

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\psi(X_i,Y_j).$$

If X and Y are also identically distributed, an even better estimator is given by the U-statistic based on the pooled sample:

$$\frac{1}{2n(2n-1)}\sum\sum_{1\le i\ne j\le 2n}\psi(Z_i,Z_j)$$

where $Z_i = X_i$ and $Z_{n+i} = Y_i$ for $i = 1,\dots,n$. These estimators are efficient under the minimal assumptions under which they were derived; see Levit (1974).

Improvements are also possible under symmetry considerations. For instance, if the pair of random variables (X,Y) is exchangeable, which means that (Y,X) has the same distribution as (X,Y) and is equivalent to $Q(A\times B) = Q(B\times A)$ for each pair of Borel sets A and B, then the symmetrized empirical estimator

$$\frac{1}{2n}\sum_{j=1}^{n}\left(\psi(X_j,Y_j)+\psi(Y_j,X_j)\right)$$

is better. If (X,Y) is symmetric in the sense that $(-X,-Y)$ has the same distribution as (X,Y), then a better estimator is given by

$$\frac{1}{2n}\sum_{j=1}^{n}\left(\psi(X_j,Y_j)+\psi(-X_j,-Y_j)\right).$$

If (X,Y) is both exchangeable and symmetric, a better estimator is given by

$$\frac{1}{4n}\sum_{j=1}^{n}\left(\psi(X_j,Y_j)+\psi(-X_j,-Y_j)+\psi(Y_j,X_j)+\psi(-Y_j,-X_j)\right).$$

The above are examples of finite group models. In such models $\gamma(X,Y)$ has the same distribution for all members γ of a finite group Γ of, say k, measurable transformations of $\mathbb{R}^2$, and an improved estimator is obtained by averaging over the group:

$$\frac{1}{kn}\sum_{j=1}^{n}\sum_{\gamma\in\Gamma}\psi(\gamma(X_j,Y_j)).$$

Indeed, this estimator is known to be efficient; see for example Bickel et al. (1993, page 231). Thus the above estimators are efficient under the minimal assumptions (exchangeability, symmetry or both) under which they were derived.

Bickel, Ritov and Wellner (1991) considered another situation in which an improvement is possible, namely when the marginal distributions, F of X and G of Y, are known. Using the modified minimum-chi-square estimators of Deming and

Stephan (1940) for contingency tables with fixed marginals and shrinking cells, they constructed an estimator $\hat{\theta}_n$ of θ that satisfies

$$\hat{\theta}_n = \frac{1}{n}\sum_{j=1}^{n}(\psi(X_j, Y_j) - a_*(X_j) - b_*(Y_j)) + o_p(n^{-1/2}) \tag{1.1}$$

as the sample size tends to infinity, where a_* and b_* are the unique (up to equivalence) minimizers of

$$\int (\psi(x, y) - a(x) - b(y))^2 \, dQ(x, y) \tag{1.2}$$

over the set of all measurable functions a and b such that $\int a^2 \, dF + \int b^2 \, dG < \infty$ and $\int a \, dF = \int b \, dG = 0$. The existence of the minimizers a_* and b_* is guaranteed by their assumption (P3) that $Q(A \times B) \geq \eta \, F(A)G(B)$ for all Borel sets A and B and some $\eta > 0$. They also showed that an estimator with the above expansion is efficient for θ.

Let us now shed some additional light on this. Note that, for each F-square-integrable a with $\int a \, dF = 0$ and each G-square-integrable b with $\int b \, dG = 0$,

$$\frac{1}{n}\sum_{j=1}^{n}(\psi(X_j, Y_j) - a(X_j) - b(Y_j)) \tag{1.3}$$

is an unbiased estimator of θ with second moment given by the expression in (1.2) divided by n. Thus the efficient estimator of Bickel, Ritov and Wellner (1991) matches the performance of the best estimator in this class.

Recently, Peng and Schick (2002) proposed an alternative construction of efficient estimators. Their method substitutes estimates of a_* and b_* for a and b in (1.3). The estimates are obtained as follows. Choose an orthonormal basis $v_1, v_2, \ldots$ for the space $L_{2,0}(F) = \{a \in L_2(F) : \int a \, dF = 0\}$ and an orthonormal basis $w_1, w_2, \ldots$ for the space $L_{2,0}(G) = \{b \in L_2(G) : \int b \, dG = 0\}$. Estimate a_* by $\sum_{i=1}^{M} \hat{\alpha}_i v_i$ and b_* by $\sum_{i=1}^{N} \hat{\beta}_i w_i$, where M and N are positive integers that tend to infinity slowly with the sample size n and $\hat{\alpha}_1, \ldots, \hat{\alpha}_M, \hat{\beta}_1, \ldots, \hat{\beta}_N$ are chosen to minimize

$$\frac{1}{n}\sum_{j=1}^{n}\left(\psi(X_j, Y_j) - \sum_{i=1}^{M}\alpha_i v_i(X_j) - \sum_{i=1}^{N}\beta_i w_i(Y_j)\right)^2.$$

Of course, $\hat{\alpha}_1, \ldots, \hat{\alpha}_M, \hat{\beta}_1, \ldots, \hat{\beta}_N$ are simply least squares estimates for the response vector $\Psi = (\psi(X_1, Y_1), \ldots, \psi(X_n, Y_n))^\top$ and the design matrix with j-th row formed by

$$(v_1(X_j), \ldots, v_M(X_j), w_1(Y_j), \ldots, w_N(Y_j))$$

and are easily computed with any standard computer package. The alternative estimator is

$$\frac{1}{n}\sum_{j=1}^{n}(\psi(X_j, Y_j) - \sum_{i=1}^{M}\hat{\alpha}_i v_i(X_j) - \sum_{i=1}^{N}\hat{\beta}_i w_i(Y_j)).$$

Peng and Schick (2002) show that this estimator satisfies (1.1). Their simulations indicate that this estimator compares favorably with the estimator of Bickel, Ritov and Wellner's (1991) in moderate sample sizes.

In this paper we shall pursue this alternative approach in a related problem. We shall study efficient estimation of $\theta = \int \psi\, dQ$ in the case when X and Y have a common, but unknown, distribution. If X and Y are pre- and post-treatment measurements, then the equality of the distributions of X and Y captures the null hypothesis that there is no treatment effect. Thus our results apply to testing this null hypothesis and in particular to the modeling of a control group in which a placebo is administered. Equal marginals can also be a reasonable assumption in situations when data are collected on pairs, such as eyes, kidneys, siblings, etc. Such data are often modeled using exchangeability, see e.g. Wei (1987). Since exchangeability implies equal marginals, the latter is less restrictive and can serve as a competitor to the former. Finally, another situation which can be modeled with equal marginals is a setting where a stationary and ergodic time series $Z_1, Z_2, \ldots$ is only observed at time points $ik, ik+1$, $i = 1, \ldots, n$ resulting in observations $X_i = Z_{ik}$ and $Y_i = Z_{ik+1}$. By stationarity the pairs (X_i, Y_i) have equal marginals, and if k is sufficiently large, these pairs can be treated as if they were independent.

Suppose now that Q has equal marginals and denote the common marginal distribution function by F. Then

$$\frac{1}{n}\sum_{j=1}^{n}(\psi(X_j, Y_j) - a(X_j) + a(Y_j)) \tag{1.4}$$

is an unbiased estimator of θ for each F-square integrable a which we may assume to satisfy $\int a\, dF = 0$. The smallest variance is achieved by a_* which minimizes

$$\int (\psi(x,y) - a(x) + a(y))^2\, dQ(x,y)$$

over the set $L_{2,0}(F)$. The existence of a_* is guaranteed under a mild assumption, see Assumption 1 below. Since a_* is unknown, we shall estimate it. If F were known, we would again have available an orthonormal basis for $L_{2,0}(F)$ and could proceed as outlined above. As F is unknown, we do not know the basis for $L_{2,0}(F)$ and need to estimate it as well. We do this as follows. We assume that F is continuous. Then $F(X)$ and $F(Y)$ are uniform random variables, and an orthonormal basis for $L_{2,0}(F)$ is given by $u_1 \circ F, u_2 \circ F, \ldots$, where $u_1, u_2, \ldots$ is an orthonormal basis for $L_{2,0}(U)$ with U the uniform distribution on $[0,1]$. We take the trigonometric basis given by

$$u_k(x) = \sqrt{2}\cos(\pi k x), \quad 0 \le x \le 1,\ k = 1, 2, \ldots. \tag{1.5}$$

This suggests to estimate the common marginal distribution function F by say $\hat{F}$ and to work with $u_1 \circ \hat{F}, u_2 \circ \hat{F}, \ldots$ in place of the unknown actual orthonormal basis $u_1 \circ F, u_2 \circ F, \ldots$ mentioned above. We take $\hat{F}$ to be the pooled empirical estimator

$$\hat{F}(t) = \frac{1}{2n}\sum_{j=1}^{n}(\mathbf{1}_{\{X_j \le t\}} + \mathbf{1}_{\{Y_j \le t\}}), \quad t \in \mathbb{R}.$$

As estimator of θ we then use

$$\hat{\theta}_n = \frac{1}{n}\sum_{j=1}^{n}\left(\psi(X_j, Y_j) - \sum_{i=1}^{m}\hat{\gamma}_{m,i}\Big[u_i(\hat{F}(X_j)) - u_i(\hat{F}(Y_j))\Big]\right) \tag{1.6}$$

where m tends to infinity slowly with the sample size n and $\hat{\gamma}_{m,1}, \ldots, \hat{\gamma}_{m,m}$ are chosen to minimize

$$\frac{1}{n}\sum_{j=1}^{n}\left(\psi(X_j, Y_j) - \sum_{i=1}^{m}\gamma_i\Big[u_i(\hat{F}(X_j)) - u_i(\hat{F}(Y_j))\Big]\right)^2.$$

These estimates are least squares estimates for the response vector Ψ as before and for the design matrix with j-th row formed by

$$u_1(\hat{F}(X_j)) - u_1(\hat{F}(Y_j)), \ldots, u_m(\hat{F}(X_j)) - u_m(\hat{F}(Y_j)).$$

Thus they can be easily calculated with a standard statistical software package.

We shall show that the proposed estimator matches the performance of the best estimator in the class (1.4) asymptotically in the sense that

$$\hat{\theta}_n = \frac{1}{n}\sum_{j=1}^{n}\left(\psi(X_j, Y_j) - a_*(X_j) + a_*(Y_j)\right) + o_p(n^{-1/2}). \tag{1.7}$$

Moreover, we shall show that this property characterizes efficient (in the sense of being least dispersed and regular) estimators of θ.

Our estimator is a least squares series estimator. For some recent work on series estimators in curve and density estimation see Newey (1997) and Efromovich (1999), and the references therein. These authors use fixed bases, while we use random bases. Thus our work is much closer in spirit to the approach taken by Beran (1974). He used random bases to estimate the score function for location.

Our paper is organized as follows. In Section 2 we shall introduce the assumption on the bivariate distribution that we shall be using in this paper and derive some preliminary results. In particular, we study properties of the minimizer a_* in general. In Section 3 we derive the efficiency theory for our problem. There we describe the tangent space and the canonical gradient and obtain the characterization (1.7) of efficient estimators. We also obtain explicit formulas for the minimizer a_* for some special cases. These are used to discuss the efficiency gains resulting from using an efficient estimator over the empirical estimator. We show that these can be substantial. In the examples considered, the asymptotic variance of an efficient estimator is about 1/3 of that of the empirical estimator or smaller. In Section 4 we shall establish (1.7) and hence the efficiency of our proposed estimator. The results of a simulation study are reported in Section 5. The simulations confirm the theoretical asymptotic results in the moderate sample sizes considered and illustrate considerable possible gains of the efficient estimator over the empirical estimator. For one choice of ψ we observe a variance reduction of at least 95 percent for all distributions considered. We also investigate a data-driven choice for m. Here again the results are very encouraging. Section 6 contains proofs of auxiliary results.

2. Some Preliminaries. In this section we shall assume that Q is a distribution of a bivariate random vector (X, Y) which has equal marginals so that

$$Q(A \times \mathbb{R}) = Q(\mathbb{R} \times A), \quad A \in \mathfrak{B}, \tag{2.1}$$

where $\mathfrak{B}$ denotes the Borel sets of $\mathbb{R}$. For convenience, we assume that X and Y are defined on $\mathbb{R}^2$ by $X(x, y) = x$ and $Y(x, y) = y$, $x, y \in \mathbb{R}$. We denote the common marginal distribution by F. Recall that $L_{2,0}(Q) = \{g \in L_2(Q) : \int g\, dQ = 0\}$ and $L_{2,0}(F) = \{a \in L_2(F) : \int a\, dF = 0\}$. Throughout we assume that the correlation between $a(X)$ and $a(Y)$ is bounded away from 1 and -1 as a ranges over $L_2(F)$.

Assumption 1. *There is a* $\rho < 1$ *such that*

$$|\operatorname{Cov}(a(X), a(Y))| \le \rho \operatorname{Var}(a(X)) \quad \textit{for all } a \in L_2(F). \tag{2.2}$$

Define a linear operator B from $L_{2,0}(F)$ into $L_{2,0}(Q)$

$$Ba = a(X) - a(Y), \quad a \in L_{2,0}(F).$$

Since $\int (Ba)^2\, dQ = 2 \int a^2\, dF - 2E[a(X)a(Y)]$, we see that this operator is bounded:

$$\int (Ba)^2\, dQ \le 2(1 + \rho) \int a^2\, dF, \quad a \in L_{2,0}(F), \tag{2.3}$$

and bounded away from zero:

$$\int (Ba)^2\, dQ \ge 2(1 - \rho) \int a^2\, dF, \quad a \in L_{2,0}(F). \tag{2.4}$$

Actually, the latter is equivalent to Assumption 1. The former holds with $2(1 + \rho)$ replaced by 4 if Assumption 1 is not met.

As B is bounded away from zero, it has a bounded inverse B^{-1}. Hence the range $\{Ba : a \in L_{2,0}(F)\}$ of B is a closed linear subspace of $L_{2,0}(Q)$. Thus the projection of an element g of $L_2(Q)$ onto the range range of B in $L_2(Q)$ exists and is of the form Bg_* for some uniquely determined element g_* of $L_{2,0}(F)$. Note that g_* is determined by the equations

$$\int Bg_* Ba\, dQ = \int gBa\, dQ, \quad a \in L_{2,0}(F).$$

These equations can be written as

$$\int (2g_* - \bar{Q}_X g_* - \bar{Q}_Y g_*) a\, dF = \int (Q_X g - Q_Y g) a\, dF, \quad a \in L_{2,0}(F),$$

where Q_X and Q_Y are the (conditional expectation) operators from $L_2(Q)$ to $L_2(F)$ and $\bar{Q}_X$ and $\bar{Q}_Y$ from $L_{2,0}(F)$ to $L_{2,0}(F)$ defined as follows. For $h \in L_2(Q)$,

$$Q_X h(t) = E(h(X, Y)|X = t) \quad \text{and} \quad Q_Y h(t) = E(h(X, Y)|Y = t), \quad t \in \mathbb{R},$$

and for $k \in L_{2,0}(F)$,

$$\bar{Q}_X k(t) = E(k(Y)|X = t), \quad \text{and} \quad \bar{Q}_Y k(t) = E(k(X)|Y = t), \quad t \in \mathbb{R}.$$

This implies that g_* is determined by the equation

$$2g_* - \bar{Q}_X g_* - \bar{Q}_Y g_* = Q_X g - Q_Y g.$$

With I the identity operator on $L_{2,0}(F)$, this can be written as

$$(2I - \bar{Q}_X - \bar{Q}_Y)g_* = Q_X g - Q_Y g.$$

We were unable to obtain an explicit solution for g_*, but we can represent g_* as an infinite series as shown next. Since $\bar{Q}_Y$ is the adjoint of $\bar{Q}_X$, the operator

$$\bar{Q} = \frac{1}{2}(\bar{Q}_X + \bar{Q}_Y)$$

is self adjoint. Since

$$E[a(X)a(Y)] = \frac{1}{2}(E[a(X)a(Y)] + E[a(Y)a(X)]) = \int a\bar{Q}a\,dF, \quad a \in L_{2,0}(F),$$

Assumption 1 is equivalent to $\bar{Q}$ having operator norm less than 1; see e.g. Theorem 15.9 in Kress (1989). Thus $I - \bar{Q}$ has a bounded inverse given by the Neumann series $\sum_{i=0}^{\infty} \bar{Q}^i$. Upon writing the above equation as $(I - \bar{Q})g_* = (Q_X g - Q_Y g)/2$, we see that g_* can be expressed as

$$\frac{1}{2}\sum_{i=0}^{\infty} \bar{Q}^i (Q_X g - Q_Y g).$$

The next lemma shows that g_* is bounded if g is bounded and if $\bar{Q}$ viewed as an operator on $L_{\infty,0}(F) = \{a \in L_\infty(F) : \int a\,dF = 0\}$ has operator norm less than one. We write $\|\cdot\|_\infty$ for both, the $L_\infty(F)$ and the $L_\infty(Q)$ norm.

LEMMA 2.1. *Suppose there is a $c < 1$ such that $\|\bar{Q}a\|_\infty \le c\|a\|_\infty$ for all $a \in L_{\infty,0}(F)$. Let $g \in L_\infty(Q)$. Then $g_* \in L_{\infty,0}(F)$ and*

$$\|g_*\|_\infty \le \frac{\|g\|_\infty}{1-c}. \tag{2.5}$$

PROOF. Viewed as an operator on $L_{\infty,0}(F)$, $\bar{Q}$ has operator norm at most c. This shows that $I - \bar{Q}$ viewed as an operator on $L_{\infty,0}(F)$ has a bounded inverse which is given by the Neumann series $\sum_{i=0}^{\infty} \bar{Q}^i$ which has operator norm at most $1/(1-c)$. Let $h = (Q_X g - Q_Y g)/2$. Then $\int h\,dF = 0$ and $\|h\|_\infty \le \|g\|_\infty$. Thus h belongs to $L_{\infty,0}(F)$. Consequently, $g_* = \sum_{i=0}^{\infty} \bar{Q}^i h \in L_{\infty,0}(F)$ and satisfies $\|g_*\|_\infty \le \|g\|_\infty/(1-c)$. This is the desired result. □

Let us now give sufficient conditions for Assumption 1 and for the assumption of the lemma. We have already seen that Assumption 1, B is bounded from below, and $\bar{Q}$ has operator norm less than one, are equivalent. The operator norm of $\bar{Q}$ is bounded by the average of operator norms of $\bar{Q}_X$ and $\bar{Q}_Y$. Since $\bar{Q}_Y$ is the adjoint of $\bar{Q}_X$, $\bar{Q}_Y$ and $\bar{Q}_X$ have the same operator norm. Thus Assumption 1 holds if $\bar{Q}_X$ has operator norm less than one. If $\bar{Q}_X = \bar{Q}_Y$ as is the case when Q is exchangeable, then Assumption 1 is even equivalent to $\bar{Q}_X$ having an operator norm less than one.

Now consider the following condition which is the analogue of (P3) used by Bickel, Ritov and Wellner (1991).

CONDITION 1. *There is an $\eta > 0$ such that for all Borel sets A, B*

$$Q(A \times B) \geq \eta F(A)F(B).$$

The η in the above condition can be at most 1. The case $\eta = 1$ is equivalent to independence of X and Y. It follows from Condition 1 that

$$\int g\, dQ \geq \eta \int g\, d(F \times F)$$

for every non-negative measurable function g on $\mathbb{R}^2$. Taking $g = (Ba)^2$ yields (2.4) with $1 - \rho = \eta$. This shows that Condition 1 implies that B is bounded from below. Thus Condition 1 yields Assumption 1.

Now assume that Q has a density q with respect to the product measure $F \times F$. Then Condition 1 is equivalent to $q \geq \eta$ almost surely $F \times F$. Condition 1, however, does not guarantee the absolute continuity of Q with respect to $F \times F$. [To see this let $Q = (1/2)(U_\Delta + U \times U)$, where U is the uniform distribution on $(0, 1)$ and U_Δ is the uniform distribution on $\{(x, x) : 0 < x < 1\}$. This measure Q has equal marginals $F = U$, but no density with respect to $U \times U$.] Bickel, Ritov and Wellner (1991, page 1331) tacitly assume that Q has a density with respect to the product of its marginal distributions when they use their (P3) to derive that the analogue of our $\bar{Q}_X$ has operator norm less than one, both for the L_2 and L_∞ norms. We shall now generalize their argument.

Since Q has marginals F, we find that, for F-almost all $t \in \mathbb{R}$,

$$\int q(t, y)\, dF(y) = 1 \quad \text{and} \quad \int q(x, t)\, dF(x) = 1.$$

Let now $\bar{q}(x, y) = (q(x, y) + q(y, x))/2$. For $a \in L_{2,0}(F)$, we obtain that for any real μ and for F-almost all $x \in \mathbb{R}$

$$\bar{Q}a(x) = \int a(y)\bar{q}(x, y)\, dF(y) = \int a(y)(\bar{q}(x, y) - \mu)\, dF(y).$$

LEMMA 2.2. *Suppose there is a μ such that $\|h_\mu\|_\infty < 1$, where*

$$h_\mu(x) = \int |\bar{q}(x, y) - \mu|\, dF(y).$$

Then Assumption 1 holds, and so does the assumption of Lemma 2.1.

PROOF. It suffices to show that $\|\bar{Q}a\|_\infty \leq \|h_\mu\|_\infty \|a\|_\infty$ for $a \in L_{\infty,0}(F)$ and $\int (\bar{Q}a)^2\, dF \leq \|h_\mu\|_\infty^2 \int a^2\, dF$ for $a \in L_{2,0}(F)$. The former is immediate, and the latter follows from an application of the the Cauchy–Schwarz inequality, which yields

$$\int (\bar{Q}_X a)^2\, dF \leq \int \int a^2(y)|\bar{q}(x, y) - \mu|\, dF(y) \int |\bar{q}(x, y) - \mu|\, dF(y)\, dF(x),$$

the symmetry of $\bar{q}$, and a change of order of integration. □

Note that under Condition 1 we have $\|h_\eta\|_\infty \le 1-\eta < 1$.

Suppose now that $\int \bar{q}^2\, dF \times F$ is finite. Then $\bar{Q}$ is a compact operator, see Rudin (1973, pg 107). Thus by the spectral theorem for compact self adjoint operators, Kress (1989, Theorem 15.12), we obtain that (in the $L_2(F \times F)$ sense),

$$\bar{q}(x,y) = 1 + \sum_{i=1}^{\infty} c_i v_i(x) v_i(y) \tag{2.6}$$

where $v_1, v_2, \dots$ is an orthonormal basis for $L_{2,0}(F)$ and $c_1, c_2, \dots$ are square summable reals. Then $\bar{Q}a = \sum_{i=1}^{\infty} c_i a_i v_i$ with $a_i = \int a v_i\, dF$ and $\int a\bar{Q}a\, dF = \sum_{i=1}^{\infty} c_i a_i^2$. This shows that the operator norm of $\bar{Q}$ is $\max_i |c_i|$. Hence Assumption 1 holds if $\max_i |c_i| < 1$. Note that $\int \bar{q}^2\, dF \times F = 1 + \sum_{i=1}^{\infty} c_i^2 \ge 1 + \max_i |c_i|^2$. Thus $\int \bar{q}^2\, dF \times F < 2$ is a sufficient condition for Assumption 1. Let us now look at two special cases.

(a) The bivariate normal distribution with standard normal marginals and correlation coefficient ρ in $(-1,1)$ satisfies (2.6) with $c_i = \rho^i$ and $\phi_1, \phi_2, \dots$ standardized and scaled Hermite polynomials; this is known as Mehler's identity; see Szegö (1959, page 377). In this case, $\bar{Q}$ has operator norm $|\rho| < 1$ implying Assumption 1, but Condition 1 is not met in this case.

(b) The uniform distribution on the unit disk $D = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$ has common marginal which has density $f(x) = (2/\pi)\sqrt{1-x^2}\,\mathbf{1}_{(-1,1)}(x)$. Here

$$\bar{q}(x,y) = \frac{\pi}{4\sqrt{1-x^2}\sqrt{1-y^2}} \mathbf{1}_D(x,y).$$

Thus Condition 1 does not hold. However, it is easy to show that $\int \bar{q}^2\, dF \times F = \pi^2/8 < 2$ so that Assumption 1 holds.

3. Efficiency considerations.

In this section we assume that Q is a distribution with equal marginals and satisfies Assumption 1. We shall characterize efficient (more precisely, least dispersed regular) estimators of $\int \psi\, dQ$ via a nonparametric convolution theorem. We begin by deriving the tangent space for our model.

The tangent space H is the set of all $h \in L_2(Q)$ for which there is a sequence $\langle Q_{n,h} \rangle$ of distributions on $\mathfrak{B}^2$ such that $Q_{n,h}$ has equal marginals and has a density $1 + n^{-1/2} h_n$ with respect to Q with $\int (h_n - h)^2\, dQ \to 0$. We refer to the sequence $\langle Q_{n,h} \rangle$ as a *local sequence with tangent* h. For such a sequence $\langle Q_{n,h} \rangle$ we immediately obtain that

$$n^{1/2}\left(\int g\, dQ_{n,h} - \int g\, dQ\right) = \int g h_n\, dQ \to \int g h\, dQ \tag{3.1}$$

for every $g \in L_2(Q)$. If we take $g = 1$, we see that $\int h\, dQ = 0$. If we take $g = Ba$ for some $a \in L_{2,0}(F)$, we obtain from the property of equal marginals that $\int Ba\, dQ_{n,h} = \int Ba\, dQ = 0$, and the latter yields

$$\int hBa\, dQ = 0.$$

Thus we see that H contains only elements in $L_{2,0}(Q)$ that are orthogonal to the range of B. We believe that H consists of all these elements so that

$$H = \{h \in L_{2,0}(Q) : \int hBa\, dQ = 0 \text{ for all } a \in L_{2,0}(F)\}. \tag{3.2}$$

However, we are only able to show this under additional assumptions. Bickel, Ritov and Wellner (1991) derive the corresponding result for their model under stronger assumptions than used here.

For the proof of the identity (3.2) assume also that the assumption of Lemma 2.1 holds. Let K denote the right-hand side of (3.2). Fix a $h \in K$. We need to produce a local sequence $\langle Q_{n,h}\rangle$ with tangent h. If h is bounded, we can choose $h_n = h$ for large n. Indeed, for large enough n, $1 + n^{-1/2}h > 0$ and hence a density as $\int(1 + n^{-1/2}h)dQ = 1$. Moreover, for $A \in \mathfrak{B}$, we can write $\mathbf{1}_A(X) - \mathbf{1}_A(Y) = Ba$ with $a = 1_A - F(A)$ in $L_{2,0}(F)$ so that

$$\int (\mathbf{1}_A(X) - \mathbf{1}_A(Y))(1 + n^{-1/2}h)\, dQ = 0.$$

This establishes $1 + n^{-1/2}h$ as a density of a probability measure with equal marginals. If h is not bounded, we shall first truncate h to $\bar{h}_n = h\mathbf{1}_{\{|h| \le c_n\}}$ with $c_n = cn^{1/4}$ for some positive constant c and then let h_n be the projection of $\bar{h}_n$ onto K so that

$$h_n = \bar{h}_n - \int \bar{h}_n\, dQ - B\chi_n$$

with $B\chi_n$ the projection of $\bar{h}_n$ onto the range of B. It follows from Lemma 2.1 that $B\chi_n$ is bounded by bc_n for some positive b. Thus $1 + n^{-1/2}h_n$ is positive for small c and hence is the density of a probability measure with equal marginals. It is easy to check that $\int(h_n - h)^2\, dQ \to 0$. This completes the proof of (3.2) under the additional assumption of Lemma 2.1. Note that we used the additional assumption only to conclude that the bounded functions in K are dense in K. Thus (3.2) also holds under this weaker property.

Now consider estimation of $\kappa(Q)$ for a functional κ based on independent observations $(X_1, Y_1), \dots, (X_n, Y_n)$ with distribution Q. For this we fix for each $h \in H$ a local sequence $\langle Q_{n,h}\rangle$ with tangent h. We then have a form of local asymptotic normality:

$$\sum_{j=1}^n \log \frac{dQ_{n,h}}{dQ}(X_j, Y_j) = n^{-1/2} \sum_{j=1}^n h(X_j, Y_j) - \frac{1}{2}\int h^2\, dQ + o_p(1)$$

and

$$\mathfrak{L}(n^{-1/2} \sum_{j=1}^n h(X_j, Y_j)|Q) \Rightarrow N(0, \int h^2\, dQ),$$

where $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. We say the functional κ is *differentiable at Q with gradient g* if $g \in L_2(Q)$ and

$$n^{1/2}(\kappa(Q_{n,h}) - \kappa(Q)) \to \int gh\, dQ$$

for every $h \in H$. The gradient g is not unique, but its projection onto H is. This projection is called the *canonical gradient*. We denote it by $g_\#$ and assume that $\int g_\#^2 \, dQ > 0$.

An estimator $\hat{\kappa}_n$ of κ based on the observations $(X_1, Y_1), \dots, (X_n, Y_n)$ is called *regular at* Q if there is a distribution M on $\mathfrak{B}$ such that

$$\mathfrak{L}(n^{1/2}(\hat{\kappa}_n - \kappa(Q_{n,h}))|Q_{n,h}) \Rightarrow M$$

for every $h \in H$, where the left hand side denotes the distribution of $n^{1/2}(\hat{\kappa}_n - \kappa(Q_{n,h}))$ calculated under the assumption that $(X_1, Y_1), \dots, (X_n, Y_n)$ are independent observations with distribution $Q_{n,h}$. It follows from the convolution theorem (see e.g. Pfanzagl and Wefelmeyer (1982), Theorem 9.3.1, pg 158 or Bickel et al. (1993, Theorem 2, pp 63) that the limit distribution M of a regular estimator is a convolution of a centered normal distribution with variance $\sigma_\#^2 = \int g_\#^2 \, dQ$ and some other distribution R

$$M = N(0, \sigma_\#^2) * R$$

and that this other distribution R is point mass at 0 if and only if

$$\hat{\kappa}_n - \kappa(Q) = \frac{1}{n} \sum_{j=1}^{n} g_\#(X_j, Y_j) + o_p(n^{-1/2}). \tag{3.3}$$

Finally, an estimator satisfying (3.3) is regular and hence least dispersed among all regular estimators. Thus we call an estimator satisfying (3.3) efficient.

Of course, we are interested in estimating $\theta = \int \psi \, dQ$. The corresponding functional is differentiable at Q with gradient ψ, see (3.1). The canonical gradient is

$$\psi_\# = \psi - \int \psi \, dQ - Ba_*$$

where a_* minimizes $\int (\psi - Ba)^2 \, dQ$ over $a \in L_{2,0}(F)$. This shows that an efficient estimator $\hat{\theta}_n$ of $\theta = \int \psi \, dQ$ is characterized by (1.7). Let us now summarize this in the following theorem.

THEOREM 3.1. *Suppose Assumption 1 holds, (3.2) is met, and* $\int \psi_\#^2 \, dQ > 0$. *Then an estimator* $\hat{\theta}_n$ *of* $\theta = \int \psi \, dQ$ *is efficient if and only if*

$$\hat{\theta}_n = \frac{1}{n} \sum_{j=1}^{n} (\psi(X_j, Y_j) - a_*(X_j) + a_*(Y_j)) + o_p(n^{-1/2})$$

and a_* *minimizes* $\int (\psi - Ba)^2 \, dQ$ *over* $a \in L_{2,0}(F)$.

It follows from the previous section that a_* is the solution to the equation

$$2a_* - \bar{Q}_X a_* - \bar{Q}_Y a_* = Q_X \psi - Q_Y \psi. \tag{3.4}$$

We were unable to solve this integral equation *explicitly* in general, but we have explicit solutions in some special cases.

EXAMPLE 1. Suppose that X and Y are independent. Then $\bar{Q}_X = \bar{Q}_Y = 0$ and one calculates

$$a_*(t) = \int \frac{1}{2}(\psi(t,x) - \psi(x,t))\,dF(x), \quad t \in \mathbb{R}.$$

Here one also has $\int (Ba_*)^2\,dQ = 2\int a_*^2\,dF$. □

EXAMPLE 2. Suppose that X and Y are exchangeable. Then $\bar{Q}_X = \bar{Q}_Y$. Let us first look at special ψ.

(a) If ψ is *symmetric* in the sense that $\psi(x,y) = \psi(y,x)$ for all $x, y \in \mathbb{R}$, then one finds $Q_X\psi = Q_Y\psi$. In this case $a_* = 0$ and the empirical estimator is already efficient.

(b) If ψ is *antisymmetric* in the sense that $\psi(y,x) = -\psi(x,y)$ for all $x, y \in \mathbb{R}$, then one finds that $Q_Y\psi = -Q_X\psi$. In this case, equation (3.4) becomes $a_* - \bar{Q}_X a_* = Q_X\psi$ and $a_* = \sum_{i=0}^\infty \bar{Q}_X^i Q_X\psi$ as $\bar{Q}_X$ has operator norm less than one.

In general, ψ can be written as a sum of a symmetric function ψ_+ and an antisymmetric function ψ_-, namely $\psi_+(x,y) = (\psi(x,y) + \psi(y,x))/2$ and $\psi_-(x,y) = (\psi(x,y) - \psi(y,x))/2$, and equation (3.4) simplifies to

$$a_* - \bar{Q}_X a_* = Q_X\psi_-.$$

If $\psi_-(x,y) = h(x) - h(y)$, then $a_* = h - \int h\,dF$. In general, the solution can be expressed as $a_* = \sum_{i=0}^\infty \bar{Q}_X^i Q_X\psi_-$. □

EXAMPLE 3. Suppose that Q has a density q with respect to $F \times F$ of the form

$$q(x,y) = 1 + \alpha r(x,y), \quad x, y \in \mathbb{R},$$

for some constant $\alpha \in (-1,1)$ and some antisymmetric function r that is bounded by 1 and satisfies $\int r(x,y)\,dF(y) = 0$. Then Condition 1 holds with $\eta = 1 - |\alpha|$. In this case, $\bar{Q}_X a_* + \bar{Q}_Y a_* = 0$ and $a_* = \frac{1}{2}(Q_X\psi - Q_Y\psi)$. One calculates

$$a_*(t) = \frac{1}{2}\int \Big[\psi(t,x) - \psi(x,t) + \alpha r(t,x)(\psi(t,x) + \psi(x,t))\Big]\,dF(x), \quad t \in \mathbb{R},$$

and finds $\int (Ba_*)^2\,dQ = 2\int a_*^2\,dF$. Note that if $\alpha = 0$, then X and Y are independent and a_* is as in the first example. □

EXAMPLE 4. Assume that Q has a density q with respect to $F \times F$ that is of the form

$$q(x,y) = 1 + \alpha v(x)w(y), \quad x, y \in \mathbb{R},$$

with v, w elements of $L_{2,0}(F)$ both bounded by 1 and $\alpha \in (-1,1)$ so that Condition 1 holds with $\eta = 1 - |\alpha|$. Then the equation (3.4) simplifies to

$$2a_*(s) - \alpha\int a_* w\,dF v(s) - \alpha\int a_* v\,dF w(s) = \bar{\psi}(s), \quad s \in \mathbb{R}, \tag{3.5}$$

where

$$\bar{\psi}(s) = \int (\psi(s,t) - \psi(t,s))\, dF(t) + \alpha v(s) \int \psi(s,t) w(t)\, dF(t)$$
$$- \alpha w(s) \int \psi(t,s) v(t)\, dF(t), \quad s \in \mathbb{R}.$$

This suggests to try

$$2a_*(s) = \bar{\psi}(s) + c_1 \alpha v(s) + c_2 \alpha w(s), \quad s \in \mathbb{R},$$

with constants c_1 and c_2. Substituting this into (3.5), we find this to be a solution if c_1 and c_2 are chosen to satisfy the linear system

$$2c_1 = \int \bar{\psi} w\, dF + c_1 \alpha \int vw\, dF + c_2 \alpha \int w^2\, dF$$
$$2c_2 = \int \bar{\psi} v\, dF + c_1 \alpha \int v^2\, dF + c_2 \alpha \int vw\, dF$$

which has a unique solution as $(2 - \alpha \int vw\, dF)^2 > \alpha^2 \int v^2\, dF \int w^2\, dF$ in view of the fact that v^2 and w^2 are bounded by 1. The solutions are

$$c_1 = \frac{(2 - \alpha \int vw\, dF) \int \bar{\psi} w\, dF + \alpha \int w^2\, dF \int \bar{\psi} v\, dF}{(2 - \alpha \int vw\, dF)^2 - \alpha^2 \int v^2\, dF \int w^2\, dF}$$

and

$$c_2 = \frac{(2 - \alpha \int vw\, dF) \int \bar{\psi} v\, dF + \alpha \int v^2\, dF \int \bar{\psi} w\, dF}{(2 - \alpha \int vw\, dF)^2 - \alpha^2 \int v^2\, dF \int w^2\, dF}.$$

One also has

$$\int (Ba_*)^2\, dQ = 2 \int a_*^2\, dF - 2\alpha \int a_* v\, dF \int a_* w\, dF.$$

There are simplifications if $v = w$. In this case,

$$c_1 = c_2 = c = \frac{\int \bar{\psi} v\, dF}{2 - 2\alpha \int v^2\, dF}$$

and

$$a_*(s) = (1/2)\bar{\psi}(s) + c\alpha v(s), \quad s \in \mathbb{R}.$$

Since $\int a_* v\, dF = c$, we find that

$$\int (Ba_*)^2\, dQ = (1/2) \int \bar{\psi}^2\, dF + 2\alpha c^2 \left(1 - \alpha \int v^2\, dF\right).$$

□

Example 5. Efficiency Gains. To see how much we can gain by using an efficient estimator instead of the empirical estimator, let us now calculate the asymptotic relative efficiency for the choice

$$\psi(x,y) = \mathbf{1}[x \le y], \quad x, y \in \mathbb{R},$$

under three (parametric) families of distributions for which we can calculate a_*. In the three families the common marginal distribution F is the uniform distribution on $[-1,1]$. The parameter is α and takes values in $(-1,1)$. It is chosen such that Condition 1 holds with $\eta = 1-|\alpha|$. We shall describe the distributions by describing their densities on $[-1,1]\times[-1,1]$. The first family is of the type described in Example 3:

$$q_{1,\alpha}(x,y) = 1+\alpha(x-y-\operatorname{sign}(x-y)), \quad -1\le x,y\le 1,$$

while the second and the third are of the type described in Example 4:

$$q_{2,\alpha}(x,y) = 1+\alpha xy, \quad -1\le x,y\le 1,$$

and

$$q_{3,\alpha}(x,y) = 1+\alpha x\operatorname{sign}(y), \quad -1\le x,y\le 1.$$

For the first family we find $\theta = 1/2+\alpha/6$ and calculate $a_*(t) = -t/2$. The (asymptotic) variance of the empirical estimator is $(9-\alpha^2)/36$, while that of the efficient estimator is $(9-\alpha^2)/36 - 1/6$. Hence the asymptotic relative efficiency as a function of α is

$$\mathrm{ARE}(\alpha) = \frac{3-\alpha^2}{9-\alpha^2}, \quad |\alpha|<1.$$

The range of this function is $(1/4, 1/3]$. The largest value $1/3$ occurs at $\alpha=0$, while values of α close to 1 and -1 yields asymptotic relative efficiencies close to $1/4$.

For the second family, we have $\theta = 1/2$. Using the results of Example 4 with $v(s)=w(s)=s$, $s\in[-1,1]$, we calculate

$$a_*(s) = \frac{15\alpha-30-3\alpha^2}{60-20\alpha}s - \frac{\alpha}{4}s^3, \quad -1\le s\le 1,$$

and

$$\mathrm{ARE}(\alpha) = \frac{525-280\alpha-25\alpha^2+26\alpha^3-2\alpha^4}{175(3-\alpha)^2}, \quad |\alpha|<1.$$

The ARE attains the approximate maximum 0.35135 at $\alpha = 0.745$ and gets close to the approximate minimum 0.26857 as α approaches -1.

For the third family, $\theta = 1/2$ and utilizing the results of Example 4 with $v(s)=s$ and $w(s)=\operatorname{sign}(s)$, $s\in[-1,1]$, straightforward calculations yield

$$\bar\psi(s) = (\alpha/4)\operatorname{sign}(s) + (\alpha/2-1)s - (3\alpha/4)s|s|, \quad s\in[-1,1],$$

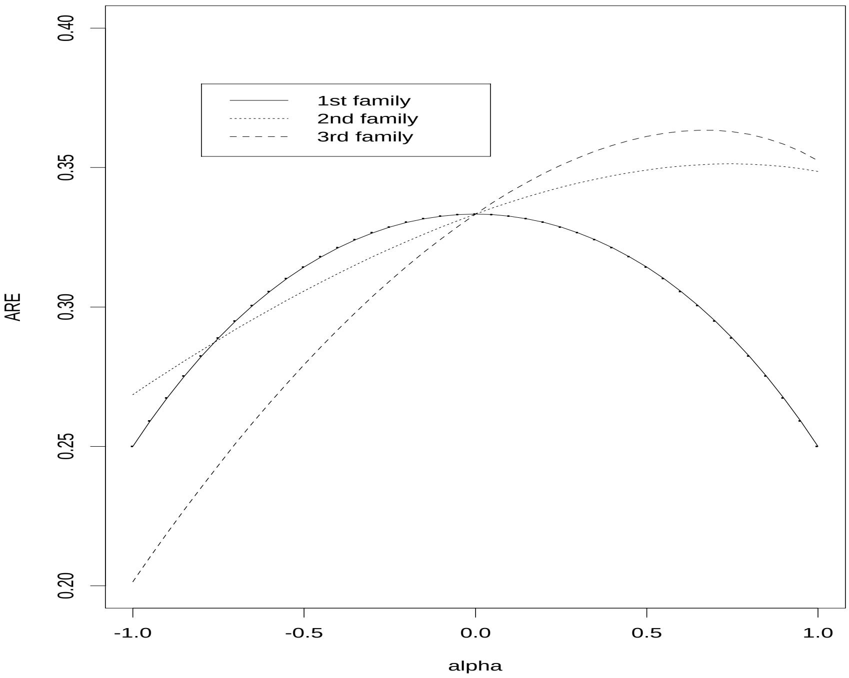
and

$$a_*(s) = \frac{\alpha+4\alpha c_2}{8}\operatorname{sign}(s) + \frac{\alpha+2\alpha c_1-2}{4}s - \frac{3\alpha}{8}s|s|, \quad s\in[-1,1],$$

where

$$c_1 = \frac{-48+20\alpha-\alpha^2}{192-96\alpha-4\alpha^2}, \quad c_2 = \frac{-64+20\alpha+3\alpha^2}{384-192\alpha-8\alpha^2}.$$

FIG. 1. *ARE curves for estimating $P(X \le Y)$ for $q_{1,\alpha}, q_{2,\alpha}, q_{3,\alpha}$ (1st, 2nd, 3rd family).*



Since the variance of the empirical estimator is 1/4, the asymptotic relative efficiency is

$$\mathrm{ARE}(\alpha) = 1 - 8\Big(\int a_*^2 \, dF - \alpha \int a_*(x) x \, dF(x) \int a_*(y) \,\mathrm{sign}(y) \, dF(y) \Big), \quad |\alpha| < 1.$$

The ARE attains the approximate maximum 0.36342 at $\alpha = 0.67$, and gets close to the approximate minimum 0.20150 as α approaches -1. Graphs of the above three AREs are given in Fig. 1. □

4. Asymptotic behavior of the proposed estimator. Throughout this section we shall assume that $(X_1, Y_1), \dots, (X_n, Y_n)$ are independent bivariate random vectors with a common distribution Q and equal marginals. We now let F denote the common marginal distribution function. We shall study the asymptotic behavior of the estimator proposed in (1.6) with $u_1, u_2, \dots$ chosen to be the trigonometric basis defined in (1.5). We shall show that this estimator satisfies (1.7) which establishes the efficiency of this estimator for estimating $\theta = \int \psi \, dQ$. Recall that a_* minimizes $\int (\psi - Ba)^2 \, dQ$ over $a \in L_{2,0}(F)$.

THEOREM 4.1. *Suppose Assumption 1 holds, the common distribution function F is continuous, and m tends to infinity slowly with n in the sense that $m \to \infty$ but $m^5/n \to 0$. Then the estimator $\hat{\theta}_n$ defined in (1.6) with $u_1, u_2, \dots$ given in (1.5) satisfies (1.7)*

$$\hat{\theta}_n = \frac{1}{n}\sum_{j=1}^{n} (\psi(X_j, Y_j) - a_*(X_j) + a_*(Y_j)) + o_p(n^{-1/2})$$

and hence is efficient under the assumptions of Theorem 3.1.

There exists a well established theory for the construction of efficient estimates in semiparametric models. Early constructions (Bickel (1982), Klaassen (1987) and Schick (1986)) use sample splitting techniques and call for appropriate estimates of the influence function. The papers by Klaassen (1987) and Schick (1986) provide necessary and sufficient conditions for the existence of efficient estimators in terms of the existence of appropriate estimators of the influence functions. Schick (1987) shows that sample splitting can be avoided under stronger conditions on the estimators of the influence function. These stronger conditions are verified in Schick (1993) and Schick (1994) in homoscedastic and heteroscedastic regression models. See also Forrester at al (2003) for weaker conditions under additional structural assumptions. As all the above constructions call for appropriate estimators of the efficient influence function, they are easier to implement when the influence function is available in closed form. Here we could apply Schick's (1987) approach directly to verify Theorem 4.1. But we found it more convenient to use a slightly different approach. Still, we heavily draw on the basic ideas of Schick (1987) in the proof of Theorem 4.1.

A critical part for the proof of this theorem is the appropriate asymptotic behavior of the least squares estimates $\hat{\gamma}_{m,1}, \dots, \hat{\gamma}_{m,m}$ which we shall formulate as a separate result next. For notational convenience we set

$$v_k = u_k \circ F \quad \text{and} \qquad \hat{v}_k = u_k \circ \hat{F}, \quad k = 1, 2, \dots$$

Since $v_1, v_2, \dots$ is an orthonormal basis for the domain of B and B has a bounded inverse, $Bv_1, Bv_2, \dots$ form a basis for the range of B. Thus $Bv_1, \dots, Bv_m$ are linearly independent. This shows that there are uniquely determined coefficients $\gamma_{m,1}, \dots, \gamma_{m,m}$ such that

$$\gamma_{m,1}Bv_1 + \cdots \gamma_{m,m}Bv_m$$

is the projection of ψ onto the linear span of $Bv_1, \dots, Bv_m$.

LEMMA 4.1. *Under the assumptions of the previous theorem,*

$$m\sum_{k=1}^{m} (\hat{\gamma}_{m,k} - \gamma_{m,k})^2 = o_p(1) \tag{4.1}$$

and

$$\sum_{k=1}^{m} \hat{\gamma}_{m,k}^2 = O_p(1). \tag{4.2}$$

We shall defer the proof of this lemma to Section 6. Another important fact in our proof of Theorem 4.1 is the Lipschitz-continuity of the trigonometric basis. More precisely, for $k = 1, 2, \dots$, one has

$$|u_k(t) - u_k(s)| \le \sqrt{2}\pi k|t - s|, \quad s, t \in \mathbb{R}. \tag{4.3}$$

Let now $\hat{F}_j$ denote the pooled empirical of F constructed without the observation pair (X_j, Y_j) so that

$$\hat{F}_j(t) = \frac{1}{2(n-1)}\Big(2n\hat{F}(t) - \mathbf{1}_{\{X_j \le t\}} - \mathbf{1}_{\{Y_j \le t\}}\Big)$$

and let $\hat{F}_{i,j}$ denote the pooled empirical of F constructed without the observation pairs (X_i, Y_i) and (X_j, Y_j) with $i \ne j$ so that

$$\hat{F}_{i,j}(t) = \frac{1}{2(n-2)}\Big(2n\hat{F}(t) - \mathbf{1}_{\{X_i \le t\}} - \mathbf{1}_{\{Y_i \le t\}} - \mathbf{1}_{\{X_j \le t\}} - \mathbf{1}_{\{Y_j \le t\}}\Big), \quad t \in \mathbb{R}.$$

Easy calculations show that for $k = 1, 2, \dots$ and all $t \in \mathbb{R}$

$$\max_{1 \le j \le n} |u_k(\hat{F}_j(t)) - u_k(\hat{F}(t))| \le 2\sqrt{2}\pi k/(n-1) \tag{4.4}$$

and

$$\max_{i \ne j} |u_k(\hat{F}_{i,j}(t)) - u_k(\hat{F}_j(t))| \le 2\sqrt{2}\pi k/(n-2). \tag{4.5}$$

Thus the influence of any pair (X_j, Y_j) of observations on the estimator $\hat{v}_k$ is small.

Now we are ready to give the proof of Theorem 4.1. Let

$$\hat{a}_m = \sum_{k=1}^{m} \hat{\gamma}_{m,k}\hat{v}_k \quad \text{and} \quad a_m = \sum_{k=1}^{m} \gamma_{m,k} v_k.$$

Then we can write the estimator defined in (1.6) as

$$\hat{\theta}_n = \frac{1}{n}\sum_{j=1}^{n} \psi(X_j, Y_j) - \hat{a}_m(X_j) + \hat{a}_m(Y_j).$$

We need to show that

$$n^{-1/2}\sum_{j=1}^{n}(\hat{a}_m(X_j) - \hat{a}_m(Y_j) - a_*(X_j) + a_*(Y_j)) = o_p(1). \tag{4.6}$$

Recall that Ba_* is the projection of ψ onto the range of B and note that

$$Ba_m = \gamma_{m,1}Bv_1 + \cdots + \gamma_{m,m}Bv_m$$

is the projection of ψ onto the linear span of $Bv_1, \dots, Bv_m$. Since $Bv_1, Bv_2, \dots$ is a basis for the range of B, the projection Ba_m converges in $L_2(Q)$ to the projection Ba_*:

$$\int (Ba_m - Ba_*)^2 \, dQ \to 0.$$

As $\int (Ba_m - Ba_*)\,dQ = \int B(a_m - a_*)\,dQ = 0$, this immediately implies that

$$n^{-1/2}\sum_{j=1}^{n}(a_m(X_j) - a_m(Y_j) - a_*(X_j) + a_*(Y_j)) = o_p(1). \tag{4.7}$$

Thus it suffices to show that

$$n^{-1/2}\sum_{j=1}^{n}(\hat a_m(X_j) - \hat a_m(Y_j) - a_m(X_j) + a_m(Y_j)) = o_p(1). \tag{4.8}$$

With the aid of the Cauchy–Schwarz Inequality we can bound the square of the left hand side in (4.8) by

$$2W_1\sum_{k=1}^{m}(\hat\gamma_{m,k} - \gamma_{m,k})^2 + 2W_2\sum_{k=1}^{m}\hat\gamma_{m,k}^2$$

where

$$W_1 = \sum_{k=1}^{m}\left(n^{-1/2}\sum_{j=1}^{n}[v_k(X_j) - v_k(Y_j)]\right)^2$$

and

$$W_2 = \sum_{k=1}^{m}\left(n^{-1/2}\sum_{j=1}^{n}\left[\hat v_k(X_j) - \hat v_k(Y_j) - v_k(X_j) + v_k(Y_j)\right]\right)^2.$$

As $v_1, v_2, \dots$ is an orthonormal basis for $L_{2,0}(F)$, we find with the help of (2.3) that

$$E(W_1) = \sum_{k=1}^{m}\int (Bv_k)^2\,dQ \le 4\sum_{k=1}^{m}\int v_k^2\,dF = 4m.$$

In view of this and Lemma 4.1 it suffices to show that $W_2 = o_p(1)$. In view of (4.4) and $m^3/n \to 0$, we have

$$\sum_{k=1}^{m}\left(n^{-1/2}\sum_{j=1}^{n}[\hat v_{k,j}(X_j) - \hat v_k(X_j)]\right)^2 = o_p(1)$$

and

$$\sum_{k=1}^{m}\left(n^{-1/2}\sum_{j=1}^{n}[\hat v_{k,j}(Y_j) - \hat v_k(Y_j)]\right)^2 = o_p(1)$$

where $\hat v_{k,j} = u_k \circ \hat F_j$. Consequently, the desired $W_2 = o_p(1)$ follows if we show that

$$W_3 = \sum_{k=1}^{m}\left(n^{-1/2}\sum_{j=1}^{n}[\hat v_{k,j}(X_j) - \hat v_{k,j}(Y_j) - v_k(X_j) + v_k(Y_j)]\right)^2 = o_p(1).$$

We shall show the stronger $E(W_3) \to 0$. To this end we let

$$D_{k,j} = \hat v_{k,j}(X_j) - \hat v_{k,j}(Y_j) - v_k(X_j) + v_k(Y_j)$$

denote the j-th summand in the inner sum of W_3. Then we can write

$$E(W_3) = \sum_{k=1}^{m} \frac{1}{n}\Big(\sum_{j=1}^{n} E(D_{k,j}^2) + 2 \sum_{1\le i<j\le n} E(D_{k,i}D_{k,j})\Big)$$
$$= \sum_{k=1}^{m}\Big(E(D_{k,1}^2) + (n-1)E(D_{k,1}D_{k,2})\Big)$$

Since $v_k = u_k \circ F$, $\hat{v}_{k,1} = u_k \circ \hat{F}_1$, we obtain from (4.3) that

$$\sum_{k=1}^{m} E(D_{k,1}^2) \le \sum_{k=1}^{m} 4\pi^2k^2(E(\hat{F}_1(X_1) - F(X_1))^2) + E(\hat{F}_1(Y_1) - F(Y_1))^2)$$
$$\le 8\pi^2m^3/(n-1) \to 0.$$

To deal with the cross product term $E(D_{k,1}D_{k,2})$ let us set

$$\bar{D}_{k,j} = u_k(\hat{F}_{1,2}(X_j)) - u_k(\hat{F}_{1,2}(Y_j)) - v_k(X_j) + v_k(Y_j), \quad k = 1,\dots,m, j = 1,2.$$

Since $E(D_{k,i}|Z_i) = 0$ and $E(\bar{D}_{k,i}|Z_i) = 0$ for $i = 1,2$, where Z_i is obtained from the full sample $(X_1, Y_1, \dots, X_n, Y_n)$ by deleting the i-th pair (X_i, Y_i), and since $\bar{D}_{k,1}$ is independent of (X_2, Y_2) and $\bar{D}_{k,2}$ is independent of (X_1, Y_1), we obtain that

$$E(D_{k,1}D_{k,2}) = E(D_{k,1} - \bar{D}_{k,1})(D_{k,2} - \bar{D}_{k,2})$$

so that by (4.5)

$$|E(D_{k,1}D_{k,2})| \le 32\pi^2k^2/(n-2)^2.$$

This shows that

$$(n-1)\sum_{k=1}^{m} |E(D_{k,1}D_{k,2})| \le 32\pi^2m^3(n-1)/(n-2)^2 \to 0.$$

The above show that $E(W_3) \to 0$. This completes the proof of Theorem 4.1.

5. Simulations. To study the performance of our estimator in moderate sample sizes we carried out a small simulation study. Simulations were run for one member of each of the three parametric families introduced in Example 5, for four different choices of functions ψ, for two sample sizes, namely $n = 100$ and $n = 200$, and for different values of m, namely $m = 1,\dots,5$. The densities chosen were the density $q_{1,\alpha}$ with $\alpha = -1/3$:

$$q_{1,-1/3}(x,y) = 1 - (1/3)\big(x - y - \operatorname{sign}(x-y)\big), \quad -1 \le x,y \le 1;$$

the density $q_{2,\alpha}$ with $\alpha = 1/2$:

$$q_{2,1/2}(x,y) = 1 + (1/2)xy, \quad -1 \le x,y \le 1;$$

the density $q_{3,\alpha}$ with $\alpha = -1/2$:

$$q_{3,-1/2}(x,y) = 1 - (1/2)x\operatorname{sign}(y), \quad -1 \le x,y \le 1.$$

TABLE 1
Simulated MSEs (times 10^3) based on $N = 20000$ repetitions

	$\psi \setminus m$	0	1	2	3	4	5	True
$n = 100$								
$q_{1,-1/3}$	ψ_1	1.124	1.144	1.138	1.112	1.099	1.086	1.085
	ψ_2	0.666	0.481	0.474	0.468	0.462	0.457	0.474
	ψ_3	2.457	0.882	0.997	1.117	1.297	1.508	0.802
	ψ_4	2.021	0.059	0.053	0.049	0.048	0.047	0.047
$q_{2,1/2}$	ψ_1	1.082	1.094	1.090	1.077	1.063	1.052	1.080
	ψ_2	0.668	0.553	0.546	0.539	0.535	0.530	0.555
	ψ_3	2.486	0.982	1.106	1.224	1.436	1.695	0.873
	ψ_4	1.794	0.052	0.046	0.044	0.043	0.042	0.043
$q_{3,-1/2}$	ψ_1	1.042	1.072	1.068	1.069	1.069	1.071	1.037
	ψ_2	0.679	0.397	0.391	0.373	0.370	0.367	0.382
	ψ_3	2.480	0.770	0.889	1.062	1.273	1.526	0.700
	ψ_4	2.470	0.063	0.059	0.048	0.046	0.045	0.046
$n = 200$								
$q_{1,-1/3}$	ψ_1	0.561	0.561	0.559	0.552	0.550	0.546	0.543
	ψ_2	0.331	0.241	0.239	0.237	0.234	0.233	0.237
	ψ_3	1.236	0.428	0.456	0.482	0.528	0.580	0.401
	ψ_4	1.027	0.027	0.026	0.024	0.024	0.024	0.023
$q_{2,1/2}$	ψ_1	0.548	0.552	0.550	0.547	0.543	0.540	0.540
	ψ_2	0.333	0.276	0.274	0.272	0.271	0.270	0.277
	ψ_3	1.265	0.476	0.507	0.522	0.577	0.641	0.436
	ψ_4	0.911	0.023	0.022	0.021	0.021	0.021	0.021
$q_{3,-1/2}$	ψ_1	0.520	0.528	0.526	0.527	0.526	0.526	0.519
	ψ_2	0.335	0.197	0.196	0.188	0.187	0.186	0.191
	ψ_3	1.248	0.365	0.394	0.434	0.489	0.553	0.350
	ψ_4	1.236	0.030	0.029	0.024	0.023	0.023	0.023

We considered the following four choices of ψ:

$$\psi_1(x,y) = xy, \qquad \psi_2(x,y) = xy^2, \qquad \psi_3(x,y) = \mathbf{1}[x \le y],$$

and

$$\psi_4(x,y) = \frac{x-y}{1+x^2+y^2}.$$

For each choice of distribution Q, we generated 20,000 random samples of size n and then calculated the empirical estimator and our proposed estimator for the above choices of m.

Table 1 gives the simulated mean square errors (multiplied by 10^3) of the empirical estimator ($m = 0$) and the efficient estimator for the choices $m = 1, \dots, 5$. The standard errors of these simulated mean square errors are 1 percent of the stated values. For comparison we give in the last column the values suggested by the asymptotic theory for the efficient estimator. We see that for all three densities there are significant improvements over the empirical estimator for the choices ψ_2, ψ_3 and ψ_4. The improvements for ψ_4 are particularly impressive. For the function ψ_1, there is essentially no detectable improvement. For the functions ψ_1, ψ_2 and ψ_4 we are already at the value suggested by the asymptotic theory. For ψ_3 we are still between 5 to 10 percent higher even for the best m.

TABLE 2
MSE (times 10^3) for $m = 0, \dots, 5$ and the data driven choice $\hat{m}$; $n = 100, N = 1000, B = 200$

	$\psi \setminus m$	0	1	2	3	4	5	$\hat{m}$
$q_{1,-1/3}$	ψ_1	1.113	1.121	1.115	1.090	1.083	1.066	1.086
	ψ_2	0.693	0.487	0.477	0.472	0.465	0.459	0.458
	ψ_3	2.454	0.844	0.945	1.064	1.245	1.469	0.848
	ψ_4	2.075	0.057	0.052	0.048	0.047	0.046	0.046
$q_{2,1/2}$	ψ_1	1.134	1.149	1.143	1.127	1.117	1.103	1.124
	ψ_2	0.671	0.562	0.555	0.549	0.546	0.539	0.533
	ψ_3	2.542	1.008	1.130	1.244	1.473	1.729	1.013
	ψ_4	1.794	0.052	0.047	0.045	0.043	0.041	0.042
$q_{3,-1/2}$	ψ_1	1.047	1.067	1.064	1.070	1.063	1.061	1.110
	ψ_2	0.692	0.397	0.390	0.375	0.371	0.367	0.362
	ψ_3	2.522	0.735	0.852	1.006	1.222	1.506	0.735
	ψ_4	2.512	0.061	0.058	0.047	0.045	0.045	0.045

Choice of m. The above simulations show that the proposed estimator is somewhat sensitive to the choice of m. This raises the question of how to choose m. Here is a possibility. For a given sample, estimate the mean square variance of the estimator for various choices of m using the bootstrap mean square error. Then select the estimator belonging to the m with smallest bootstrap mean square error. We studied the behavior of this data driven choice $\widehat{m}$ of m via simulations for the three given densities and the four choices of ψ. We took $n = 100$ and bootstrap sample size $B = 200$. The results for $N = 4000$ repetitions are reported in Table 2. The table gives the mean square errors (multiplied by 10^3) for $m = 0, \dots, 5$ and the data driven choice $\widehat{m}$. In each case, the mean square error of the estimator based on the data driven choice $\widehat{m}$ is very close to the minimal mean square error among the estimators with fixed $m = 0, \dots, 5$. The standard errors of the reported mean square errors are around 2 percent of the reported values. We see that this data driven method is quite successful.

6. Proof of Lemma 4.1.

Let us write $\|A\|$ for the Euclidean norm of the $p \times q$ matrix A and $\|A\|_o$ for its operator (or spectral) norm so that

$$\|A\|^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} A_{ij}^2$$

and

$$\|A\|_o = \sup\{\|Ax\| : x \in \mathbb{R}^q, \|x\| = 1\}$$

is the square root of the largest eigen value of $A^\top A$. We have $\|A\|_o \le \|A\|$.

The vector $\gamma_m = (\gamma_{m,1}, \dots, \gamma_{m,m})^\top$ satisfies the normal equation

$$S_m \gamma_m = T_m$$

where S_m is the symmetric $m \times m$ matrix whose (i, j)-entry is $\int Bv_i Bv_j \, dQ$ and T_m is the m-dimensional column vector whose i-th entry is $\int \psi B v_i \, dQ$, $i, j = 1, \dots, m$. Note that

$$x^T S_m x = \int (B(\sum_{i=1}^m x_i v_i))^2 \, dQ, \quad x = (x_1, \dots, x_m)^\top \in \mathbb{R}^m.$$

It follows from this, (2.3) and (2.4), that the eigen values of S_m fall into the interval $[2(1-\rho), 2(1+\rho)]$. Thus the matrix S_m is invertible with an inverse S_m^{-1} that has eigen values in the interval $[1/(2+2\rho), 1/(2-2\rho)]$. This yields that

$$\|S_m^{-1}\|_o \le 1/(2-2\rho). \tag{6.1}$$

Since $\int \psi B v_i \, dQ = \int B^* \psi v_i \, dF$ is the i-th Fourier coefficient of $B^*\psi$ with respect to the basis $v_1, v_2, \dots$, where B^* is the adjoint of B. This shows that

$$\|T_m\|^2 \le \int (B^*\psi)^2 \, dF \le (2+2\rho) \int \psi^2 \, dQ. \tag{6.2}$$

Consequently,

$$\sum_{k=1}^m \gamma_{m,k}^2 \le \|S_m^{-1}\|_o^2 \|T_m\|^2 \le \frac{(1+\rho)^2}{(1-\rho)^2} \int \psi^2 \, dQ.$$

Thus we only need to show the first part of Lemma 4.1.

The random vector $\hat\gamma_m = (\hat\gamma_{m,1}, \dots, \hat\gamma_{m,m})^\top$ satisfies the normal equation

$$\hat S_m \hat\gamma_m = \hat T_m$$

where $\hat S_m$ is the symmetric $m \times m$ matrix whose (i, j)-entry is

$$\frac{1}{n} \sum_{r=1}^n (\hat v_i(X_r) - \hat v_i(Y_r))(\hat v_j(X_r) - \hat v_j(Y_r))$$

and $\hat T_m$ is the m-dimensional column vector whose i-th entry is

$$\frac{1}{n} \sum_{r=1}^n \psi(X_r, Y_r)(\hat v_i(X_r) - \hat v_i(Y_r))$$

for $i, j = 1, \dots, m$. Finally, let $\bar S_m$ be the $m \times m$ matrix whose (i, j)-entry is

$$\frac{1}{n} \sum_{r=1}^n (v_i(X_r) - v_i(Y_r))(v_j(X_r) - v_j(Y_r))$$

and $\bar{T}_m$ be the m-dimensional column vector whose i-th entry is

$$\frac{1}{n}\sum_{r=1}^{n}\psi(X_r,Y_r)(v_i(X_r)-v_i(Y_r)).$$

Since u_k is bounded by $\sqrt{2}$, it is easy to check that

$$E(\|\bar{T}_m - T_m\|^2) \le \frac{8m}{n}\int \psi^2\,dQ \quad \text{and} \quad E(\|\bar{S}_m - S_m\|^2 \le \frac{64m^2}{n}.$$

It follows from (4.3) and the Cauchy–Schwarz Inequality that

$$\|\hat{T}_m - \bar{T}_m\|^2 \le \frac{1}{n}\sum_{r=1}^{n}\psi^2(X_r,Y_r)8\pi^2 m^3 \sup_{t\in\mathbb{R}}|\hat{F}(t)-F(t)|^2 = O_p(m^3/n).$$

Similarly, one obtains

$$\|\hat{S}_m - \bar{S}_m\|^2 = O_p(m^4/n).$$

Combining the above we obtain in view of $m^5/n \to 0$ that

$$m\|\hat{T}_m - T_m\|^2 = o_p(1) \quad \text{and} \quad m\|\hat{S}_m - S_m\|^2 = o_p(1).$$

The second statement holds also in the operator norm and implies that $\hat{S}_m$ is invertible on an event whose probability tends to one. Moreover, on this event $m\|\hat{S}_m^{-1} - S_m^{-1}\|_o^2 = o_p(1)$ in view of (6.1). The desired (4.1) is now immediate from this and (6.2).

REFERENCES

Beran, R. (1974). Asymptotically efficient adaptive rank estimates in location models. *Ann. Statist.*, **2**, 63–74.

Bickel, P.J. (1982). On adaptive estimation. *Ann. Statist.*, **10**, 647–671.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1993). *Efficient and adaptive estimation in semiparametric models.* Johns Hopkins Univ. Press, Baltimore.

Bickel, P. J., Ritov, Y. and Wellner, J.A. (1991). Efficient estimation of linear functionals of a probability measure P with known marginal distributions *Ann. Statist.*, **19**, 1316–1346.

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.

Efromovich, S. (1999). *Nonparametric curve estimation: Methods, Theory, and Applications* Springer, New York.

Forrester, J., Hooper, W., Peng, H., and Schick, A. (2003). On the construction of efficient estimators in semiparametric models. *Statist. Decisions*, **21**, 109–138.

Klaassen, C.A.J. (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Statist.*, **15**, 1548–1563.

Kress, R. (1989). *Linear integral equations.* Springer-Verlag, Berlin.

Levit, B.Y. (1974). On optimality of some statistical estimates. In: *Proceedings of the Prague Symposium on Asymptotic Statistics*, Vol. 2 (J. Hájek, ed), 215-238, Charles University, Prague.

Newey, W. K. (1997) Convergence rates and asymptotic normality for series estimators. *J. Econometrics*, **79**, 147–168.

Peng, H. and Schick, A. (2002). On efficient estimation of linear functionals of a bivariate distribution with known marginals. *Statist. Probab. Lett.*, **59**, 83–91.

Pfanzagl, J. and Wefelmeyer, W. (1982): Contributions to a general asymptotic statistical theory. *Lecture Notes in Statistics*, **13,** Springer, New York.

Rudin, W. (1973). *Functional analysis.* McGraw-Hill, New York.

Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.*, **14**, 1139–1151.

Schick, A. (1987). A note on the construction of asymptotically linear estimators. *J. Statist. Planning Inference*, **16**, 89-105. Correction, **22** (1989), 269–270.

Schick, A. (1993). On efficient estimation in regression models. *Ann. Statist.*, **21**, 1486–1521. Correction and addendum, **23** (1995), 1862–1863.

Schick, A. (1994). Efficient estimation in regression models with unknown scale functions. *Math. Meth. Statist.*, **3**, 171–212.

Szegö, G. (1959). Orthogonal polynomials. *American Mathematical Society Colloquium Publications*, Vol. 23, Fourth Edition.

Wei, L. J. (1987). Two-sample problem with bivariate exchangeable observations. *J. R. Statist. Soc. B*, **49**, 40–45.

Hanxiang Peng
Department of Mathematics
University of Mississippi
University, MS 38677-1848
USA

Anton Schick
Department of Mathematical Sciences
Binghamton University
Binghamton, NY 13902-6000
USA