The Annals of Statistics
1994, Vol. 22, No. 4, 1976–1992

MINIMUM DISTANCE ESTIMATION IN RANDOM COEFFICIENT REGRESSION MODELS[1]

BY R. BERAN AND P. W. MILLAR

University of California, Berkeley

Random coefficient regression models are important in modeling heteroscedastic multivariate linear regression in econometrics. The analysis of panel data is one example. In statistics, the random and mixed effects models of ANOVA, deconvolution models and affine mixture models are all special cases of random coefficient regression. Some inferential problems, such as constructing prediction regions for the modeled response, require a good nonparametric estimator of the unknown coefficient distribution. This paper introduces and studies a consistent nonparametric minimum distance method for estimating the coefficient distribution. Our estimator translates the difficult problem of estimating an inverse Radon transform into a minimization problem.

1. Introduction. Research in statistics and in econometrics during the past two decades has called increasing attention to random coefficient regression models of the form

$$Y_i = A_i + X_i B_i, \qquad i \geq 1. \tag{1.1}$$

Here Y_i and A_i are $p \times 1$ random vectors, B_i is a $q \times 1$ random vector and X_i is a $p \times q$ random matrix. The triples $\{(A_i, B_i, X_i)\colon i \geq 1\}$ are iid and (A_i, B_i) is independent of X_i. The distribution of (A_i, B_i, X_i) is not known, although it may be restricted further in some applications. The sample S_n that we observe consists of the n pairs $\{(Y_i, X_i)\colon 1 \leq i \leq n\}$.

This model articulates three ideas about the data. First is the assumption that the ith response Y_i depends linearly on the ith set of covariates X_i. Second is the view that the coefficients (A_i, B_i) of the linear response function vary with i. Third is the supposition that the data behaves like a simple random sample from a large population. Thus, Y_i is the response and X_i is the covariate matrix associated with the ith individual in the sample. The first two modeling ideas are expressed by (1.1). The third idea corresponds to the iid assumption on the $\{(Y_i, X_i, A_i, B_i)\}$.

In the statistical literature, several special cases of model (1.1) are well established, under various labels. When each $B_i = b$, an unknown constant vector, then (1.1) becomes a multivariate linear model with random regressors

Received October 1991; revised March 1994.

[1]Supported in part by NSF Grants DMS-90-01710 and DMS-92-24868. Some of the first author's work was done a guest of Sonderforschungsbereich 123 at Universität Heidelberg.

AMS 1991 *subject classifications.* Primary 62G05; secondary 62J05.

Key words and phrases. Radon transform, prediction interval, distribution estimate, weak convergence metric, characteristic function, nonparametric, semiparametric.

and homoscedastic errors. When the $\{X_i\colon 1 \le i \le n\}$ are not observed but the distribution of X_i is known, then (1.1) is an affine mixture model. When each $X_i = x$, a known constant matrix, then (1.1) includes the random effects models of ANOVA [see Scheffé (1959), Chapter 7] and the models studied in nonparametric deconvolution [see Fan (1991) and van Es (1991)].

If the first moments exist and if we write $A_i = a + a_i$, $B_i = b + b_i$ with $a = EA_i$ and $b = EB_i$, then (1.1) can be put into the equivalent form

$$Y_i = a + X_i b + E_i, \tag{1.2}$$

where

$$E_i = a_i + X_i b_i. \tag{1.3}$$

This is a multivariate linear model with heteroscedastic errors possessing the structure (1.3). In the econometric literature, such models have been used to estimate the variances of heteroscedastic regression errors and to test for homoscedasticity. See, for instance, Hildreth and Houck (1968), Goldfeld and Quandt (1972), Chapter 3, and Amemiya (1977). More recent surveys of work on random coefficient regression models, their autoregressive analogs and models combining both features include Raj and Ullah (1981), Chow (1983), Nicholls and Pagan (1985) and Newbold (1988).

Let F_{AB} denote the unknown distribution of (A_i, B_i) in model (1.1). The main topic of this paper is the nonparametric estimation of F_{AB} from the sample $S_n = \{(Y_i, X_i)\colon 1 \le i \le n\}$. This problem is important if we wish to construct prediction regions for response in random coefficient models, such as those used for panel data [see Hsiao (1986)]. For instance, suppose we wish to predict the future observable Y_{n+1} in model (1.1), given the sample S_n and the condition that $X_{n+1} = x$. To simplify the discussion, suppose that Y_{n+1} is scalar ($p = 1$). Let $A_x(\cdot, F_{AB})$ denote the cdf of $A_i + B_i x$ and let $\widehat{F}_{AB,n}$ denote an estimator based on S_n which converges weakly to F_{AB} in probability. Consider the prediction interval $D_{x,n}$ for Y_{n+1} whose lower and upper endpoints are respectively the estimated quantiles $A_x^{-1}[(1-\alpha)/2, \widehat{F}_{AB,n}]$ and $A_x^{-1}[(1+\alpha)/2, \widehat{F}_{AB,n}]$. If F_{AB} is absolutely continuous with strictly positive density, then

$$\Pr\left[Y_{n+1} \in D_{x,n} \,|\, X_{n+1} = x\right] \to \alpha \tag{1.4}$$

as n increases, for every support point x of the distribution of X_{n+1}. The essential reasoning for (1.4) is Proposition 1 in Beran (1990).

To better understand the problem of estimating the coefficient distribution F_{AB}, consider the simplest case, where A_i, B_i, X_i are random scalars and (A_i, B_i) has Lebesgue density f_{AB}. Then the conditional density $f_{Y|X}$ of Y_i given $X_i = x$ is

$$f_{Y|X}(y \,|\, x) = \int f_{AB}(y - \beta x, \beta)\, d\beta, \tag{1.5}$$

the integration being over the real line. The right side of (1.5) is the Radon transform $T(f_{AB})$ of the density f_{AB} [cf. Deans (1983)].

In a seminal paper, Radon (1917) proved that, *under tail conditions on* f_{AB}, there exists an inverse transform T^{-1} such that

$$f_{AB} = T^{-1}(f_{Y|X}) \tag{1.6}$$

and gave an explicit formula for T^{-1}. Radon's result suggests estimating f_{AB} by $T^{-1}(\hat{f}_{Y|X,n})$, where $\hat{f}_{Y|X,n}$ is a consistent nonparametric estimator of the conditional density $f_{Y|X}$. Unfortunately, this plausible approach need not yield a consistent estimator of f_{AB}. Like the differentiation operator, the inverse Radon transform is not continuous in familiar metrics on nonparametric density estimators.

The extensive study of tomographic reconstruction [cf. Devaney (1989)] has generated several algorithms for numerical inversion of Radon transforms. These results do not solve the problem of estimating the density f_{AB} consistently, for two reasons:

(a) In tomography, the measurement process provides an accurate discretized version of the function analogous to $f_{Y|X}$. In our problem, we begin with only a sup-norm (say) consistent estimate of $f_{Y|X}$.

(b) The tomographic algorithms handle projections of a two or three-dimensional object. In model (1.1), F_{AB} is a distribution on R^{p+q}, where $p+q$ is often much larger than 2 or 3.

These considerations suggest that estimating F_{AB} consistently differs from tomographic reconstruction and requires new ideas.

If the support of F_{AB} is compact, then the moments of F_{AB} determine the distribution uniquely. Thus, we might estimate r_n moments of F_{AB} from the data, where r_n tends to ∞ slower than n, and then devise an estimate of F_{AB} whose moments match the estimated moments well. Beran and Hall (1992) pursued this strategy in the special case where A_i and B_i are *independent* random scalars, thereby constructing consistent nonparametric estimators of the marginal distributions F_A and F_B. It seems very difficult to extend their moment method and its consistency result to the general case of model (1.1), where A_i and B_i are not necessarily independent and p or q may exceed 1. Too many moments are then required to approximate F_{AB} reasonably.

This paper proposes and studies and entirely different nonparametric estimator for F_{AB} in model (1.1)—a minimum distance estimator that overcomes several of the difficulties just described. Let F_X denote the distribution of X_i, let $P(F_{AB}, F_X)$ denote the distribution of (Y_i, X_i) under model (1.1) and let $\hat{F}_{X,n}$ denote the empirical distribution of the observed $\{X_i: 1 \le i \le n\}$. The key idea is to choose the estimator $\hat{F}_{AB,n}$ so that the *fitted* distribution $P(\hat{F}_{AB,n}, \hat{F}_{X,n})$ under model (1.1) is close to the empirical distribution of the sample S_n. Closeness is measured in any metric d for weak convergence of probabilities on R^{p+pq}.

The consistency of the nonparametric minimum distance estimator $\hat{F}_{AB,n}$ under very general conditions is the main subject of Section 2.1. Section 2.2 narrows the choice of the metric d, on various theoretical and computational grounds, to metrics generated by L_2-norms on characteristic functions. Section 3

presents an explicit numerical algorithm for $\widehat{F}_{AB,n}$. The consistency results in Section 2 finesse the estimation of an inverse Radon transform in dimension $p+q$. The intrinsic difficulty of this task, discussed earlier, does not entirely vanish: the numerical minimization of the distance criterion may encounter many relative minima. Nevertheless, our minimum distance approach translates an unfamiliar problem—consistent nonparametric estimation of F_{AB}—into a minimization problem for which numerical methods, such as simulated annealing, already exist.

Section 4 treats semiparametric estimation of F_{AB}. In the semiparametric version of model (1.1), the unknown distribution F_Y of X_i is unrestricted but the unknown distribution of (A_i, B_i) belongs to a parametric family $\{F_{AB}(\theta)\colon \theta \in \Theta\}$. We give conditions and examples under which the minimum distance estimator of θ is $n^{1/2}$-consistent.

2. Consistent nonparametric estimation. This section defines the minimum distance estimators for the unknown distribution of (A_i, B_i) in nonparametric model (1.1), establishes the consistency of these estimators and then narrows the choice of the distance on computational and theoretical grounds. Proofs are deferred to Section 5.

2.1. *Definitions and consistency.* In model (1.1), let us introduce the following notation:

F_{AB} for the joint distribution of (A_i, B_i), which is restricted to a nonparametric family of distributions $\mathbf{F}_{AB}$ on R^{p+q};
F_X for the distribution of X_i, which is restricted to a nonparametric family of distributions $\mathbf{F}_X$ on R^{pq};
$P(F_{AB}, F_X)$ for the joint distribution of (Y_i, X_i) under model (1.1);
d for any metric that metrizes weak convergence of probability measures on R^{p+pq}.

A sequence of distributions for (A_i, B_i) will be indicated by $\{F_{AB,n}\}$, and similarly for distributions of X_i.

The functional P defined above has two interesting properties that are important for understanding the minimum distance technique to be introduced shortly. The first of these is "continuity."

PROPOSITION 2.1 (Continuity). *Suppose, as* $n \to \infty$,

$$d(F_{AB,n}, F_{AB,0}) \to 0,$$
$$d(F_{X,n}, F_{X,0}) \to 0.$$

Then

$$d\left[P(F_{AB,n}, F_{X,n}), P(F_{AB,0}, F_{X,0})\right] \to 0.$$

The second important property is "strong identifiability." Simple identifiability of F_{AB}, in the usual sense employed in statistical inference, would assert that if $P(F_{AB,1}, F_X) = P(F_{AB,0}, F_X)$, then $F_{AB,1} = F_{AB,0}$. Strong identifiability is a locally uniform version of simple identifiability, described in the next proposition. Let C^* denote the adjoint of any matrix C and let $\text{supp}(G)$ denote the support of any distribution G.

PROPOSITION 2.2 (Strong identifiability). *Assume that:*

(2.1) $\mathbf{F}_{AB}$ *consists of probabilities supported by a fixed compact and* $\{x^*t: x \in \text{supp}(F_{X,0})\}$ *contains an open set in* R^q *for every* $t \neq 0$ *in* R^p.

If

$$d\left[P(F_{AB,n}, F_{X,n}), \quad P(F_{AB,0}, F_{X,0})\right] \to 0,$$

then

$$d(F_{AB,n}, F_{AB,0}) \to 0.$$

When $p = q = 1$, so that Y_i, X_i, A_i, B_i are all real, a stronger variant of Proposition 2.2 is available. Let $\mathbf{F}_B$ denote the possible distributions for B_i when F_{AB} is in the family $\mathbf{F}_{AB}$.

PROPOSITION 2.2′. *Assume that* $p = q = 1$ *and*

(2.2) $\mathbf{F}_{AB}$ *is tight;* $\mathbf{F}_B$ *consists of distributions all supported within a fixed compact; and* $F_{X,0}$ *has a cluster point within its support.*

If

$$d\left[P(F_{AB,n}, F_{X,n}), \quad P(F_{AB,0}, F_{X,0})\right] \to 0,$$

then

$$d(F_{AB,n}, F_{AB,0}) \to 0.$$

To state the minimum distance method and its consistency, we require additional notation. Let:

$\widehat{P}_n$ be the empirical measure of the sample which gives mass n^{-1} to each of the $\{(Y_i, X_i): 1 \leq i \leq n\}$;
$\widehat{F}_{X,n}$ be the empirical measure of the $\{X_i: 1 \leq i \leq n\}$.

Define $\widehat{F}_{AB,n}$ to be the nonparametric minimum distance estimator of F_{AB}—that is, any distribution in $\mathbf{F}_{AB}$ that satisfies

$$\inf_{F_{AB}\in\mathbf{F}_{AB}} d\left[P(F_{AB},\widehat{F}_{X,n}),\widehat{P}_n\right] = d\left[P(\widehat{F}_{AB,n},\widehat{F}_{X,n}),\widehat{P}_n\right] + o_p(n^{-1/2}). \tag{2.3}$$

We shall henceforth write this definition of $\widehat{F}_{AB,n}$ in the following shorter form:

$$\widehat{F}_{AB,n} = \arg\inf_{F_{AB}\in\mathbf{F}_{AB}} d\left[P(F_{AB},\widehat{F}_{X,n}),\widehat{P}_n\right]. \tag{2.4}$$

PROPOSITION 2.3 (Consistency of $\widehat{F}_{AB,n}$). *Assume* (2.1), *or* (2.2) *in the case* $p = q = 1$. *Suppose that the true distributions in model* (1.1) *at sample size* n *are given by* $F_{AB,n}$ *and* $F_{X,n}$, *where* $d(F_{AB,n},F_{AB,0}) \to 0$ *and* $d(F_{X,n},F_{X,0}) \to 0$. *Then*

$$d(\widehat{F}_{AB},F_{AB,0}) \to 0 \quad \textit{in probability.}$$

The tail conditions (2.1) or (2.2) in this proposition are not surprising because Radon's (1917) inversion theorem already requires tail conditions. The triangular array formulation of Proposition 2.3 entails that the convergence in probability of $\widehat{F}_{AB,n}$ to F_{AB} is uniform over every compact subset (in metric d) of $\mathbf{F}_{AB}$. Moreover, by an obvious change in the proof, the pointwise convergence of $\widehat{F}_{AB,n}$ to $F_{AB,0}$ is almost sure.

As it stands the definition of the nonparametric estimator $\widehat{F}_{AB,n}$ via (2.3) appears computationally intractable, because the infimum is taken over a prohibitively large set of measures $\mathbf{F}_{AB}$. Therefore, we next provide a feasible variant of (2.3) this is also consistent. To understand the motivation, recall that the estimation of F_{AB} in model (1.1) is hard because we observe only the $\{(Y_i,X_i)\colon 1 \le i \le n\}$. If we could observe the corresponding $\{(A_i,B_i)\colon 1 \le i \le n\}$ directly, then their empirical distribution would obviously be a consistent estimator of F_{AB}—an estimator that is $n^{1/2}$-consistent in many metrics. This ideal empirical distribution is supported on at most n points—the distinct values among the $\{(A_i,B_i)\colon 1 \le i \le n\}$—with mass at any given support point being an integer multiple of n^{-1}. Perhaps, in constructing a nonparametric minimum distance estimator, we need only minimize over such discrete distributions F_{AB} rather than over the full family $\mathbf{F}_{AB}$.

To set this up rigorously, let $\{m_n\}$ be any sequence of positive integers that goes to ∞ with n and define

$$\begin{aligned}\mathbf{C}(m_n) = \{&\text{all } F_{AB}\in\mathbf{F}_{AB} \text{ that are supported on at most } m_n\\ &\text{points, with mass at each point being an integer multiple}\\ &\text{of } m_n^{-1}\}.\end{aligned} \tag{2.5}$$

Define the discrete nonparametric minimum distance estimator $\widetilde{F}_{AB,n}$ to be

$$\arg\inf_{F_{AB}\in\mathbf{C}(m_n)} d\left[P(F_{AB},\widehat{F}_{X,n}),\widehat{P}_n\right] \tag{2.6}$$

by analogy with (2.4). The minimization in (2.6) is over a space of finite dimension $(p+q)m_n$.

PROPOSITION 2.4 (Consistency of $\widetilde{F}_{AB,n}$). *Under the hypotheses of Proposition* 2.3,

$$d(\widetilde{F}_{AB,n}, F_{AB,0}) \to 0 \quad \textit{in probability.}$$

provided that $m_n \to \infty$.

Note that there is no rate in this proposition on the convergence of m_n to ∞. The heuristic motivating the class $\mathbf{C}(m_n)$ suggests that $m_n = n$ would be sufficiently large. According to the heuristic one might then get $n^{1/2}$-consistency from $\widetilde{F}_{AB,n}$. Unfortunately, $n^{1/2}$-consistency *in general* is impossible in this problem. See, for instance, Fan's (1991) analysis of rates achievable in the submodel for nonparametric deconvolution. On the other hand, Section 4 shows that $n^{1/2}$-consistency is achieved by minimum distance estimators in a semiparametric version of model (1.1).

2.2. *Choice of the metric d.* The theoretical results of Section 2.1 allow enormous freedom in the selection of the metric d that determines the minimum distance method. What is a reasonable specific choice of d? Here are several factors to consider:

(a) To ensure consistency of $\widehat{F}_{AB,n}$ or $\widetilde{F}_{AB,n}$, the distance d must metrize weak convergence of distributions on R^{p+pq}.

(b) For the sake of feasibility, the distance d should be relatively easy to calculate.

(c) To facilitate theoretical investigation of $\widehat{F}_{AB,n}$ beyond consistency, the distance d should be generated by a norm on a nice linear space. This approach involves representing $\widehat{P}_n$ and $P(F_{AB}, F_X)$ as elements of the chosen linear space. [See, for instance, Pollard (1980) and Millar (1984)].

(d) A Hilbertian norm is particularly attractive from the standpoint of both (b) and (c).

These considerations led us to define d through an L_2 norm on characteristic functions. More specifically, suppose that P_1 and P_2 are any two distributions on R^{p+pq}, with characteristic functions $\phi_1(t,u)$ and $\phi_2(t,u)$, respectively, where $t \in R^p$, $u \in R^{pq}$. Define

$$d(P_1,P_2) = \left\{ \int |\phi_1(t,u) - \phi_2(t,u)|^2 \, dQ(t,u) \right\}^{1/2} \tag{2.7}$$
$$= \|\phi_1 - \phi_2\|, \quad \text{say,}$$

where Q is a probability on R^{p+pq} that has full support. Obviously, d so defined metrizes weak convergence.

The minimum distance application of this distance d requires the characteristic function of $\widehat{P}_n$ and of $P(F_{AB}, \widehat{F}_{X,n})$. The former is just

$$\widehat{\phi}_n(t,u) = n^{-1}\sum_{j=1}^{n} \exp\bigl(i\langle t, Y_j\rangle + i\langle u, X_j\rangle\bigr),$$

where $\langle\cdot,\cdot\rangle$ denotes the appropriate inner product. The characteristic function of $P(F_{AB}, \widehat{F}_{X,n})$ is

$$\widehat{\phi}_{n,AB,X}(t,u) = n^{-1}\sum_{j=1}^{n} \phi_{AB}\bigl(t, X_j^* t\bigr)\exp\bigl(i\langle u, X_j\rangle\bigr),$$

where ϕ_{AB} denotes the characteristic function of F_{AB} and * denotes adjoint. For d as in (2.7), the definition (2.4) of the minimum distance estimator of F_{AB} becomes

$$\widehat{F}_{AB,n} = \underset{F_{AB}\in \mathbf{F}_{AB}}{\arg\inf} \bigl\|\widehat{\phi}_n - \widehat{\phi}_{n,AB,X}\bigr\|. \tag{2.8}$$

Replacing $\mathbf{F}_{AB}$ with $\mathbf{C}(m_n)$ in (2.8) gives the corresponding definition of the discrete minimum distance estimator $\widetilde{F}_{AB,n}$.

3. Calculation of $\widetilde{F}_{AB,n}$. This section describes an algorithm for computing the discretized nonparametric minimum distance estimator $\widetilde{F}_{AB,n}$, whose consistency was established in Proposition 2.4. Since sample size n is fixed in this calculation, we will drop that subscript here. Once $\widetilde{F}_{AB}$ has been found, drawing bootstrap samples from it is a matter of sampling the m support points of $\widetilde{F}_{AB}$ with replacement. Thus, given $\widetilde{F}_{AB}$, the prediction intervals described in Section 1 are easily found.

Calculating $\widetilde{F}_{AB}$ requires three preliminary choices:

Choice of the compact K within which the support of F_{AB} is assumed to lie. For expository simplicity, suppose that $p = q = 1$ in model (1.1). Let $\mu_{j,k}$ denote the (j,k)th moment of F_{AB} and let $\widehat{\mu}_{j,k}$ be a consistent estimation of $\mu_{j,k}$. We suggest defining K to be a rectangle centered at $(\widehat{\mu}_{1,0}, \widehat{\mu}_{0,1})$, the lengths of the sides being 4 or more times the respective estimated standard errors $(\widehat{\mu}_{2,0} - \widehat{\mu}_{1,0}^2)^{1/2}$ and $(\widehat{\mu}_{0,2} - \widehat{\mu}_{0,1}^2)^{1/2}$. The Chebyshev and Bonferroni inequalities are the rationale for this proposal.

For each positive integer r, define the least squares estimates $\{\widehat{\mu}_{r-k,k}\colon 0 \le k \le r\}$ to be the values of the $\{\mu_{r-k,k}\}$ that minimize

$$\sum_{i=1}^{n}\left[Y_i^r - \sum_{k=0}^{r}\binom{r}{k}\mu_{r-k,k}X_i^k\right]^2. \tag{3.1}$$

By the law of large numbers, these least squares estimates are consistent for

each fixed r, whenever the moments of (A_i, B_i) and of X_i are finite. The motivation for (3.1) is the relationship

$$E(Y_i^r) = \sum_{k=0}^{r} \begin{bmatrix} r \\ k \end{bmatrix} \mu_{r-k,k} E(X_i^k). \tag{3.2}$$

Choice of m, the cardinality of the support of $\widetilde{F}_{AB}$. The value of m should be as large as is feasible computationally, in view of Proposition 2.4.

Choice of the distance d that defines $\widetilde{F}_{AB}$. Tractability, both numerical and theoretical, favors taking d to be the $L^2(Q)$ distance on characteristic functions (cf. Section 2.2). The integration with respect to Q can be handled by Monte Carlo methods. Note that the moments of F_{AB} are determined by the derivatives of its characteristic function at the origin. Thus, it is intuitively plausible that Q should have most of its mass near the origin, while retaining full support in R^{p+pq}.

The algorithm. Once K, m and d have been chosen, as discussed above, the algorithm for $\widehat{F}_{AB}$ consists of four steps:

1. Let $S_n = \{(Y_j, X_j)\colon 1 \le j \le n\}$ denote the sample. Write a module to calculate the empirical characteristic function of S_n,

$$\widehat{\phi}(t,u) = n^{-1} \sum_{j=1}^{n} \exp\bigl(i\langle t, Y_j\rangle + i\langle u, X_j\rangle\bigr), \tag{3.3}$$

where $t \in R^p$, $u \in R^{pq}$ and $\langle \cdot, \cdot \rangle$ denotes inner product in these spaces.

2. Let the $\{(a_k, b_k)\colon 1 \le k \le m\}$ be the m candidate support points of $\widetilde{F}_{AB}$, which assigns to each of these the probability m^{-1}. Write a module to calculate the characteristic function of $\widetilde{F}_{AB}$,

$$\widehat{\phi}_{AB}(t,v) = m^{-1} \sum_{k=1}^{m} \exp\bigl(i\langle t, a_k\rangle + i\langle v, b_k\rangle\bigr), \tag{3.4}$$

where $t \in R^p$ and $v \in R^q$. Write a further module to calculate the characteristic function of the estimated distribution for (Y_j, X_j) under model (1.1),

$$\widehat{\phi}_{AB,X}(t,u) = n^{-1} \sum_{j=1}^{n} \widehat{\phi}_{AB}\bigl(t, X_j^* t\bigr) \exp\bigl(i\langle u, X_j\rangle\bigr), \tag{3.5}$$

where $t \in R^p$ and $u \in R^{pq}$.

3. Let $\widehat{Q}_N$ denote the empirical distribution of a pseudo-random sample of size N from the distribution Q. Write a module to calculate the following Monte Carlo approximation to the $L_2(Q)$ distance between $\widehat{\phi}$ and $\widehat{\phi}_{AB,X}$,

$$\widehat{\phi}_{AB}^2 = \int \bigl|\widehat{\phi}(t,u) - \widehat{\phi}_{AB,X}(t,u)\bigr|^2 \, d\widehat{Q}_N(t,u). \tag{3.6}$$

4. Minimize $\hat{d}_{AB}^2$ in step 3 over all possible choices, within the compact K, of the m support points for $\tilde{F}_{AB}$. The uniform distribution on the minimizing support points is the estimator $\hat{F}_{AB}$. Ties are permitted among support points, in which case the uniform probabilities are added together.

REMARKS. (a) Numerical trials by the authors and by Jingou Liu, a student of the first author, indicate that the distance being minimized in step 4 may have many relative minima [see Liu (1994)]. The difficulty of estimating an inverse Radon transform is thus translated into a possibly difficult minimization problem. We found simulated annealing to be more reliable than Nelder and Mead [cf. Press, Flannery, Teukolsky and Vetterling (1992)] but neither was fool-proof in the examples we studied. Considerably more work is needed on the algorithmic aspects of computing the minimum distance estimator of F_{AB}.

(b) If A_i and B_i are assumed to be independent in model (1.1), steps 2 and 4 should be modified as follows to calculate the discrete minimum distance estimates for the marginal distributions F_A and F_B:

2′. Let the $\{a_k\colon 1 \le k \le m\}$ and the $\{b_k\colon 1 \le k \le m\}$ be m candidate support points for $\tilde{F}_A$ and $\tilde{F}_B$, respectively. $\tilde{F}_A$ assigns probability m^{-1} to each of its support points, as does $\tilde{F}_B$. Write a module to calculate the characteristic functions of $\tilde{F}_A$ and $\tilde{F}_B$,

$$\begin{aligned} \hat{\phi}_A(t) &= m^{-1}\sum_{k=1}^{m}\exp\bigl(i\langle t, a_k\rangle\bigr), \\ \hat{\phi}_B(v) &= m^{-1}\sum_{k=1}^{m}\exp\bigl(i\langle v, b_k\rangle\bigr), \end{aligned} \tag{3.7}$$

where $t \in R^p$ and $v \in R^q$. Put $\hat{\phi}_{AB}(t,v) = \hat{\phi}_A(t)\hat{\phi}_B(v)$ in (3.5).

4′. Minimize $\hat{d}_{AB}^2$ over all possible choices, within the compact K, of the m support points for $\tilde{F}_A$ and the m support points for $\tilde{F}_B$. The uniform distributions on the two sets of minimizing support points are $\tilde{F}_A$ and $\tilde{F}_B$, respectively.

4. Semiparametric models. This section treats more extensively a semiparametric version of model (1.1) in which the unknown distribution F_X of X_i is unrestricted while the unknown distribution of (A_i, B_i) belongs to a parametric family $\{F_{AB}(\theta)\colon \theta \in \Theta\}$. Here Θ is an open subset of R^k. The distance d is taken to be the $L_2(Q)$-distance defined in Section 2.2. We given sufficient conditions under which the minimum distance estimator of θ is $n^{1/2}$-consistent and examples where these conditions hold.

Assumptions. Write $P(\theta, F_X)$ for the distribution of (Y_i, X_i) under the semi parametric model, in place of the earlier notation $P(F_{AB}, F_X)$. Let $\phi(\theta, F_X)$ denote the characteristic function of $P(\theta, F_X)$, let $\|\cdot\|$ be the $L_2(Q)$-norm defined in (2.7) and let $(\theta_0, F_{X,0})$ denote a fixed point in the parameter space of the model. We make the following assumptions.

C1. (Strong identifiability). If $\|\phi(\theta, F_X) - \phi(\theta_0, F_{X,0}\| \to 0$, then $\theta \to \theta_0$.

C2. (Norm differentiability). If $F_X \Rightarrow F_{X,0}$ and $\theta \to \theta_0$, then there exists a $k \times 1$ vector function $\eta_0 = \eta(\theta_0, F_{X,0})$, whose components belong to $L_2(Q)$, such that

$$|\theta - \theta_0|^{-1}\|\phi(\theta, F_X) - \phi(\theta_0, F_X) - \langle \theta - \theta_0, \eta_0\rangle\| \to 0. \tag{4.1}$$

C3. (Nonsingularity). There exists a finite positive constant C such that

$$\|\langle t, \eta_0\rangle\| \geq C|t| \tag{4.2}$$

for every $t \in R^k$.

The hypothesis in C1 implies that $F_X \Rightarrow F_{X,0}$. Convenient sufficient conditions for C1 to C3 are discussed later in this section. The minimum distance estimator $\hat{\theta}_n$ satisfies

$$\hat{\theta}_n = \arg\inf_{\theta \in \Theta} \|\hat{\phi}_n - \phi(\theta, \hat{F}_{X,n})\| \tag{4.3}$$

in the sense of (2.4). The following rate-of-convergence result is proved in Section 5.

PROPOSITION 4.1. *Suppose that conditions* C1 *to* C3 *hold, that* $\{n^{1/2}(\theta_n - \theta_0)\}$ *is bounded and that* $F_{X,n} \Rightarrow F_{X,0}$. *Then*

$$\hat{\theta}_n = \theta_0 + O_p(n^{-1/2}), \tag{4.4}$$

under the sequence of models $\{P(\theta_n, F_{X,n})\}$.

In treating examples, the relatively abstract assumptions C1 to C3 may often be replaced by more convenient sufficient conditions:

Sufficient conditions for C1. By Proposition 2.2, if the parametric family $\{F_{AB}(\theta)\colon \theta \in \Theta\}$ consists of distributions supported on a fixed compact and $\{x^*t\colon x \in \text{supp}(F_{X,0})\}$ contains a nonempty open set in R^q for every $t \neq 0$, then the hypothesis in C1 implies

$$F_{AB}(\theta_n) \Rightarrow F_{AB}(\theta_0). \tag{4.5}$$

Condition C1 is now equivalent to strong identifiability of the parametric family.

Sufficient conditions for C2. The fundamental theorem of calculus and the Cauchy–Schwarz inequality yield the following. Suppose that for every $(t, u) \in R^{p+pq}$ and for every (θ, F_X) in a neighborhood of $(\theta_0, F_{X,0})$ the characteristic function $\phi(t, u;\ \theta, F_X)$ has partial derivatives $\{\eta_{\theta, F_X, j}(t, u)\colon 1 \leq j \leq k\}$ with respect to θ. Suppose as well that these partial derivatives are continuous over a neighborhood of $(\theta_0, F_{X,0})$ and that the convergence $\theta \to \theta_0$, $F_X \Rightarrow F_{X,0}$ implies

$$\|\eta_{\theta, F_X, j}\| \to \|\eta_{\theta, F_{X,0}, j}\|, \qquad 1 \leq j \leq k. \tag{4.6}$$

Then C2 holds with $\eta_0 = \{\eta_{\theta_0, F_{X,0}, j}\colon 1 \leq j \leq k\}$.

Equivalent condition for C3. Because θ is finite dimensional, nonsingularity in the sense of C3 is equivalent to linear independence of the components of η_0 [cf. Pollard (1980)].

Three examples illustrate the usefulness of these sufficient conditions and the scope of Proposition 4.1.

EXAMPLE 1. $F_{AB}(\theta)$ is a discrete distribution supported on r distinct sites $\{s_i\colon 1 \le i \le r\}$ in R^{p+q}. These sites are ordered by their first coordinates, with ties broken by second coordinate ordering and so on. The probability supported on each site s_i is $1/r$. Here $\theta = (s_1, s_2, \ldots, s_r)$ and the dimension of θ is $k = (p+q)r$.

This model does *not* induce a classically regular semiparametric model in the sense of Begun, Hall, Huang and Wellner (1983) because the support of $F_{AB}(\theta)$ depends on θ. However, this semiparametric model is regular from the viewpoint of our minimum distance estimate. For simplicity, suppose that $p = q = 1$ so that the sites in $\theta = (s_1, \ldots, s_r)$ have the form $s_j = (a_j, b_j)$, where a_j and b_j are real. Here the characteristic function $\phi(\theta, F_X)$ reduces to

$$\phi(t,u,\theta,F_X) = r^{-1}\sum_{j=1}^{r} \exp(ita_j + ixtb_j)\exp(iux)\,dF_X(x). \tag{4.7}$$

Suppose that $\int t^2\,dQ(t)$ is finite and that $\mu(F_X) = \int |x|\,dF_X(x)$ is finite and weakly continuous in F_X. The components of $\eta_{\theta,F_X}(t,u)$ are the $k = 2r$ elements

$$\begin{aligned}\frac{\partial\phi(t,u)}{\partial a_j} &= r^{-1}\int it\,\exp(ita_j + ixtb_j)\exp(iux)\,dF_X(x),\\ \frac{\partial\phi(t,u)}{\partial b_j} &= r^{-1}\int ixt\exp(ita_j + ixtb_j)\exp(iux)\,dF_X(x),\end{aligned} \tag{4.8}$$

where $1 \le j \le r$. The sufficient conditions for C2 hold by dominated convergence.

Since $F_{AB}(\theta)$ puts mass $1/r$ in each of the distinct sites $\{s_j\colon 1 \le j \le r\}$, the strong identifiability of the model $\{F_{AB}(\theta)\colon \theta \in \Theta\}$ is apparent. Consequently, C1 holds provided the support of $F_{X,0}$ contains a nonempty open set and the sites $\{s_i\}$ lie within a given compact. Finally, C3 holds because the partial derivatives in (4.8) are linearly independent.

EXAMPLE 2. $\{F_{AB}(\theta)\colon \theta \in \Theta\}$ is a canonical exponential family model supported on a fixed compact and Θ is the interior of the natural parameter space. Unlike Example 1, this model $F_{AB}(\theta)$ can be continuous. The induced semiparametric model satisfies conditions C1 to C3, by reasoning similar to that for Example 1. Moreover, this semiparametric model is classically regular, in the sense of Begun, Hall, Huang and Wellner (1983). The robustness of the minimum distance estimate $\hat{\theta}_n$ against small departures from the semiparametric model is an attractive feature of $\hat{\theta}_n$ when compared with possible likelihood-type estimators for this example [cf. Millar (1984)].

EXAMPLE 3. The support of $F_{AB}(\theta)$ is discrete as in Example 1. The probability supported on the site s_i is now p_i, where $p_i > 0$ and $\sum_{i=1}^{r-1} p_i < 1$. In this semiparametric model, $\theta = (s_1, s_2, \ldots, s_r, p_1, \ldots, p_{r-1})$ and the dimension of θ is $k = (p+q+1)r - 1$. This generalization of Example 1 is also not classically regular but satisfies conditions C1 to C3 for Proposition 4.1.

In Example 2, Proposition 4.1 and the differentiability in θ of $F_{AB}(\theta)$ imply that $F_{AB}(\hat{\theta}_n) = F_{AB}(\theta) + O_p(n^{-1/2})$ in supremum norm over any Vapnik–Cervonenkis class. This reasoning breaks down in Examples 1 and 3, for lack of differentiability. In Example 3, $F_{AB}(\theta)$ is the probability measure on R^{p+q} whose support points and probabilities are given by the appropriate elements of θ. Define the *double-variation* norm (DV-norm) of $F_{AB}(\theta)$ by

$$\|F_{AB}(\theta)\|_{\mathrm{DV}} = |\theta|, \tag{4.9}$$

where $|\cdot|$ is the Euclidean norm on R^w. The corresponding DV-distance between two probabilities $F_{AB}(\theta)$, $F_{AB}(\theta')$, is then $\|F_{AB}(\theta) - F_{AB}(\theta')\|_{\mathrm{DV}}$. Evidently, if $\|F_{AB}(\theta_n) - F_{AB}(\theta_0)\|_{\mathrm{DV}} \to 0$, then the sites and corresponding probabilities of $F_{AB}(\theta_n)$ converge uniformly to those of $F_{AB}(\theta_0)$; and the DV-distance metrizes weak convergence under hypothesis (4.1). Moreover, $F_{AB}(\hat{\theta}_n) = F_{AB}(\theta) + O_p(n^{-1/2})$ in DV-distance by Proposition 4.1.

Convergence in DV-norm is weaker than convergence in the usual variation norm $\|\cdot\|_V$. That is, $\|F_{AB}(\theta_n) - F_{AB}(\theta_0)\|_V \to 0$ implies $\|F_{AB}(\theta_n) - F_{AB}(\theta_0)\|_{\mathrm{DV}} \to 0$, but the converse need not hold. However, if the support points of $F_{AB}(\theta_n)$ coincide with those of $F_{AB}(\theta_0)$, then the two probability metrics are equivalent. Further properties of the DV-metric will be described in Propositions 4.2 and 4.3. Note that replacing Euclidean norm in (4.9) with an equivalent norm generates a norm on probabilities that is equivalent to the DV-norm.

We conclude this section by relating the DV-metric to more familiar metrics. Let μ_n, $n = 0, 1, \ldots$, be discrete probability measures with sites $c_{n,1}, \ldots, c_{n,r}$ and with probabilities $p_{ni} = \mu_n(\{c_{ni}\})$. For every n, the $\{c_{n,i}\}$ are restricted to the common compact set K.

PROPOSITION 4.2.

(a) *Let* $\|\cdot\|_{\mathrm{BL}}$ *denote the bounded Lipschitz norm on probability measures. If* $n^{1/2}\|\mu_n - \mu_0\|_{\mathrm{DV}}$ *is bounded, then so is* $n^{1/2}\|\mu_n - \mu_0\|_{\mathrm{BL}}$.

(b) *Let* $\|\cdot\|_V$ *denote variation norm on probability measures. If* $n^{1/2}\|\mu_n - \mu_0\|_{\mathrm{DV}}$ *is bounded and if* $c_{ni} = c_i$ *for every* n, $1 \le i \le r$, *then* $n^{1/2}\|\mu_n - \mu_0\|_V$ *is bounded.*

PROPOSITION 4.3. *Suppose that the discrete probability measures* $\{\mu_n\}$ *are supported on* R^1 *and have cdf's* $\{\mu_n(t)\}$.

(a) *The convergence* $\|\mu_n - \mu_0\|_{\mathrm{DV}} \to 0$ *does not imply convergence in the Kolmogorov metric. [That is,* $\sup_t |\mu_n(t) - \mu_0(t)| \not\to 0$ *in general.]*

(b) *The convergence* $\|\mu_n - \mu_0\|_{\mathrm{DV}} \to 0$ *does imply convergence in the Skorokhod metric or in any other metric for weak convergence.*

(c) *Let* $\|\cdot\|_p$ *denote the* L_p*-norm on cdf's*:

$$\|\mu_n\|_p = \left\{ \int_K |\mu_n(t)|^p \, dt \right\}^{1/p}.$$

If $n^{1/2}\|\mu_n - \mu_0\|_{\mathrm{DV}}$ *is bounded, then so is* $n^{1/2}\|\mu_n - \mu_0\|_1$. *The implication fails if* $p > 1$.

These results, together with Propositions 2.3 and 4.1, are one indication that supremum norms over Vapnik–Cervonenkis classes may not be appropriate in studying the convergence of estimators for F_{AB}, in general.

5. Proofs.

PROOF OF PROPOSITION 2.1. It suffices to show that the chf of $P(F_{AB,n}, F_{X,n})$ converges to that of $P(F_{AB,0}, F_{X,0})$. To set this up, let A_n, B_n have distribution $F_{AB,n}$ and abbreviate $F_{X,n}$ by F_n. We then wish to show that

$$\int e^{i\langle u,x\rangle} E \exp\{i\langle t, A_n\rangle + i\langle t, xB_n\rangle\} F_n(dx)$$

converges to

$$\int e^{i\langle u,x\rangle} E \exp\{i\langle t, A_0\rangle + i\langle t, xB_0\rangle\} F_0(dx)$$

for all $t \in R^p$ and all $u \in R^{pq}$. For fixed (t,u) defined for $n = 1, 2, \ldots$ and $n = 0$, let

$$f_n(x) = e^{i\langle u,x\rangle} E \exp\{i\langle t, A_n\rangle + i\langle t, xB_n\rangle\}.$$

With this notation, we then must show that

$$\int f_n(x) F_n(dx) \to \int f_0(x) F_0(dx).$$

Note that $f_n(x), f_0(x)$ are continuous, uniformly bounded (by 1), and that f_n converges to f_0 uniformly on x-compacts. Since F_n converges weakly to F_0, the tightness implies that there is a compact K_ε carrying all but ε of the mass of $\{F_n\}$ and F_0. Thus

$$\begin{aligned}\left| \int f_n \, dF_n - \int f_0 \, dF_0 \right| &\le 2\varepsilon + \left| \int_{K^\varepsilon} f_n \, dF_n - \int_{K^\varepsilon} f_0 \, dF_0 \right| \\ &\le 2\varepsilon + \int_{K^\varepsilon} |f_n - f_0| \, dF_n + \left| \int_{K^\varepsilon} f_0 \, d(F_n - F_0) \right|.\end{aligned}$$

By the uniform convergence in compacts, the first integral on the right is less than ε for all sufficiently large n, while the second goes to 0 by the definition of weak convergence. This completes the proof. □

PROOF OF PROPOSITIONS 2.2 AND 2.2′. Let us first establish the case for real A, B, X, as in Proposition 2.2′. The hypothesis implies that the chf of $P(F_{AB,n}, F_{X,n})$ converges to that of $P(F_{AB,0}, F_{X,0})$. Let A_0, B_0 have distribution $F_{AB,0}$. Since $\mathbf{F}_{AB}$ is tight, $F_{AB,n}$ has a subsequence converging weakly; let A_1, B_1 denote random variables with this limiting distribution. The convergence of the chf's and of $F_{X,n}$ then implies

$$\int e^{iux} E e^{itA_0 + ixtB_0} F_0(dx) = \int e^{iux} E e^{itA_1 + ixtB_1} F_0(dx)$$

for all $u, t \in R^1$. This in turn implies that for all x in the support of F_0,

$$Ee^{itA_0 + ixtB_0} = Ee^{itA_1 + ixtB_1}. \tag{5.1}$$

Since the possible distributions for B_i are, by hypothesis, all concentrated on a compact, the left and right sides of (5.1) are both analytic as function of x. Hence (5.1) holds for all real x because $\operatorname{supp}(F_0)$ contains a cluster point; it already held for all t. Hence (A_0, B_0) and (A_1, B_1) have the same joint characteristic functions, so their distributions are identical. We conclude that every weakly convergent subsequence of $F_{AB,n}$ has the same limit, namely $F_{AB,0}$. This is equivalent to the convergence of $F_{AB,n}$ to $F_{AB,0}$, proving Proposition 2.2′. The proof of Proposition 2.2 (the vector-valued case) is similar. Proceed in the same way as above to see that

$$E \exp\{i\langle A_1, t\rangle + i\langle xB_1, t\rangle\} = E \exp\{i\langle A_0, t\rangle + i\langle xB_0, t\rangle\}$$

for all t and all $x \in \operatorname{supp}(F_0)$. Writing $\langle xB_i, t\rangle = \langle B_i, x^*t\rangle$ and applying the hypothesis that $\{x^*t\colon x \in \operatorname{supp}(F_0)\}$ contains an open set for every $t \neq 0$ then implies that (A_1, B_1), (A_0, B_0) have the same characteristic function, and so the same distribution. □

PROOF OF PROPOSITIONS 2.3 AND 2.4. To prove Proposition 2.3, note first that

$$\begin{aligned} d(F_{X,n}, \widehat{F}_{X,n}) &\xrightarrow{p} 0, \\ d\big[P(F_{AB,n}, F_{X,n}), \widehat{P}_n\big] &\xrightarrow{p} 0 \end{aligned} \tag{5.2}$$

by Kiefer's (1961) inequality, applied to the triangular array here. Next, by the foregoing display, continuity (Proposition 2.1) and the triangle inequality

$$d\big[P(F_{AB,n}, \widehat{F}_{X,n}), \widehat{P}_n\big] \xrightarrow{p} 0.$$

Third, note that the definition (2.3) of the minimum distance estimate $\widehat{F}_{AB,n}$ then forces

$$d\big[P(\widehat{F}_{AB,n}, \widehat{F}_{X,n}), \widehat{P}_n\big] \xrightarrow{p} 0.$$

Applying the triangle inequality with the last display and (5.2) shows

$$d\big[P(\widehat{F}_{AB,n}, \widehat{F}_{X,n}),\ P(F_{AB,n}, F_{X,n})\big] \xrightarrow{p} 0.$$

Hence, by continuity (Proposition 2.1),

$$d\left[P(\widehat{F}_{AB,n},\widehat{F}_{X,n}),\ P(F_{AB,0},F_{X,0})\right] \xrightarrow{p} 0.$$

Strong identifiability (Proposition 2.2 or 2.2′) now implies

$$d(\widehat{F}_{AB,n},F_{AB,0}) \xrightarrow{p} 0,$$

proving the proposition. Almost sure convergence holds if the triangular array formulation here is dispensed with.

The proof of Proposition 2.4 is nearly the same. Indeed, if $m_n \to \infty$, the distributions in $\mathbf{C}(m_n)$ will approximate those in $\mathbf{F}_{AB}$, and so the argument just given applies with just one more use of the triangle inequality. □

PROOF OF PROPOSITION 4.1. By the triangular array weak law of large numbers and Proposition 2.1,

$$\left\|\widehat{\phi}_n - \phi(\theta_0,F_{X,0})\right\| \xrightarrow{p} 0, \tag{5.3}$$

under the models $\{P(\theta_n,F_{X,n})\}$. Hence, because of C1, $\widehat{\theta}_n$ converges in probability to θ_0.

By the definition (4.3) of $\widehat{\theta}_n$ and the triangle inequality,

$$\begin{aligned}&\|\phi(\widehat{\theta}_n,\widehat{F}_{X,n}) - \phi(\theta_0,\widehat{F}_{X,n})\| \\ &\le \|\phi(\theta_n,\widehat{F}_{X,n}) - \phi(\theta_0,\widehat{F}_{X,n})\| + \|\widehat{\phi}_n - \phi(\widehat{\theta}_n,\widehat{F}_{X,n})\| + \|\widehat{\phi}_n - \phi(\theta_n,\widehat{F}_{X,n})\| \\ &\le 2\|\widehat{\phi}_n - \phi(\theta_n,\widehat{F}_{X,n})\| + \|\phi(\theta_n,\widehat{F}_{X,n}) - \phi(\theta_0,\widehat{F}_{X,n})\| + o\left(n^{-1/2}\right).\end{aligned} \tag{5.4}$$

The first term on the right side of (5.4) is bounded above by

$$2\|\widehat{\phi}_n - \phi(\theta_n,F_{X,n})\| + 2\|\phi(\theta_n,\widehat{F}_{X,n}) - \phi(\theta_n,F_{X,n})\| \tag{5.5}$$

and is thus $O_p(n^{-1/2})$ by the central limit theorem in $L_2(Q)$. The second term on the right side of (5.4) is also $O_p(n^{-1/2})$ by C2 because $\{n^{1/2}(\theta_n - \theta_0)\}$ is bounded. Hence the left side of (5.4) is $O_p(n^{-1/2})$.

On the other hand, C2 and C3 imply that the left side of (5.4) is bounded from below by $C|\widehat{\theta}_n - \theta_0| + o(|\widehat{\theta}_n - \theta_0|)$. In view of the previous paragraphs, the proposition follows. □

PROOFS OF PROPOSITIONS 4.2 AND 4.3. Immediate from the definitions and standard properties of the other probability metrics. The counterexamples needed can be based on the two-site distributions

$$\begin{aligned}\mu_n\left(\left\{1 - n^{-1/2}\right\}\right) &= 1/4, \qquad \mu_n(\{2\}) = 3/4, \\ \mu_0(\{1\}) &= 1/4, \qquad \mu_0(\{2\}) = 3/4.\end{aligned}$$

□

Acknowledgment. The first author wishes to thank Andrey Feuerverger for stimulating conversations on the link with the inverse Radon transform.

REFERENCES

AMEMIYA, T. (1977). A note on a heteroscedastic model. *J. Econometrics* **6** 365–370.

BEGUN, J. M., HALL, W. J., HUANG, W.-M. and WELLNER, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11** 432–452.

BERAN, R. (1990). Calibrating prediction regions. *J. Amer. Statist. Assoc.* **85** 715–723.

BERAN, R. and HALL, P. (1992). Estimating coefficient distributions in random coefficient regression. *Ann. Statist.* **20** 1110–1119.

CHOW, G. C. (1983). Random and changing coefficient models. In *Handbook of Econometrics* (Z. Griliches and M. D. Intriligator, eds.) 1213–1245. North-Holland, Amsterdam.

DEANS, S. R. (1983). *The Radon Transform and Some of Its Applications*. Wiley, New York.

DEVANEY, A. J. (1989). The limited-view problem in diffraction tomography. *Inverse Problems* **5** 501–521.

FAN, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* **19** 1257–1272.

GOLDFELD, S. M. and QUANDT, R. E. (1972). *Nonlinear Methods in Econometrics*. North-Holland, Amsterdam.

HILDRETH, C. and HOUCK, J. P. (1968). Some estimators for a linear model with random coefficients. *J. Amer. Statist. Assoc.* **63** 584–595.

HSIAO, C. (1986). *Analysis of Panel Data*. Cambridge Univ. Press.

KIEFER, J. (1961). On large deviations of the empiric D.F. of vector chance variables and a law of iterated logarithm. *Pacific J. Math.* **11** 649–660.

LIU, J. (1994). Minimum distance approach in nonlinear mixed effect models. Ph.D. dissertation, Univ. California, Berkeley.

MILIAR, P. W. (1984). A general approach to the optimality of minimum distance estimators. *Trans. Amer. Math. Soc.* **286** 377–418.

NEWBOLD, P. (1988). Some recent developments in time series analysis. III. *Internat. Statist. Rev.* **56** 17–29.

NICHOLLS, D. F. and PAGAN, A. R. (1985). Varying coefficient regression. In *Handbook of Statistics* (E. J. Hannan, P. R. Krishnaiah and M. M. Rao, eds.) **5** 413–449. North-Holland, Amsterdam.

POLLARD, D. (1980). The minimum distance method of testing. *Metrika* **27** 43–70.

PRESS, W. H. FLANNERY, B. P., TEUKOLSKY, S. A. and VETTERLING, W. T. (1992). *Numerical Recipes in FORTRAN: The Art of Scientific Computing*, 2nd ed. Cambridge Univ. Press.

RADON, J. (1917). Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten. *Berichte Saechsische Akademie der Wissenschaften* **69** 262–277.

RAJ, B. and ULLAH, A. (1981). *Econometrics, A Varying Coefficients Approach*. Croom-Helm, London.

SCHEFFÉ, H. (1959). *The Analysis of Variance*. Wiley, New York.

VAN ES, A. J. (1991). Uniform deconvolution: nonparametric maximum likelihood and inverse estimation. In *Nonparametric Functional Estimation and Related Topics* (G. Roussas, ed.) 191–198. Kluwer, Dordrecht.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720