Statistica Sinica **6**(1996), 877-898

THE MINIMUM DISTANCE METHOD IN NONLINEAR RANDOM COEFFICIENT MODELS

Jingou Liu

Ciba-Geigy Pharmaceuticals

Abstract: Nonlinear random coefficient models are used in many different applications, including population pharmacokinetics and econometrics. We describe the minimum distance method of estimating the distributions of the random coefficients, as a substitute for the traditional least-squares type methods. We prove the consistency of the nonparametric minimum distance estimators and the $\sqrt{n}$-consistency of the parametric minimum distance estimators. The applications of the new method in some population pharmacokinetic models are presented as examples. Numerical comparison of the minimum distance method and a least-squares type method is also given.

Key words and phrases: Random coefficient models, mixed effects models, minimum distance, population pharmacokinetics, Radon transform.

1. Introduction

This paper considers the general nonlinear random coefficient model as follows:

$$Y_{ni} = f(X_{ni}, A_{ni}), \quad 1 \le i \le n, \ n = 1, 2, \ldots, \tag{1}$$

where $X_{ni} \in \mathcal{X} \subseteq R^k$, $A_{ni} \in \mathcal{A} \subseteq R^d$, and f a known function from $R^k \times R^d$ to R^m. We assume that $A_{n1}, \ldots, A_{nn}, X_{n1}, \ldots, X_{nn}$ are independent random vectors, $A_{ni} \sim P_{\theta_n}$, $\theta_n \in \Theta$, and $X_{ni} \sim P_{X,n} \in \mathcal{P}_X$ with both $P_{X,n}$ and θ_n unknown. We further assume that Θ is a metric space equipped with its Borel σ-field. The triangular array framework will allow us to study the bootstrap method. This general model covers many special cases that are familiar to statisticians and have broad applications.

Consider the case when f is linear, $Y = CX + B$, where the parameters $C_{m\times k}$ and $B_{m\times 1}$ are random matrices with unknown distributions. Observations are i.i.d. copies of (X, Y). This model was first studied by Beran and Hall (1992) where X, Y, B and C are scalars and B and C are independent. A moment-matching method was suggested to estimate the distribution of the coefficients. Beran (1993) and Beran and Millar (1994) studied the general case and used the minimum distance method to estimate these distributions. The purpose of this paper is to show that this method is still valid when f is nonlinear.

Linear errors-in-variables models have been studied intensively. The literature on nonlinear errors-in-variables models is relatively limited, even though these models are attracting the attention of economists. In addition to maximum likelihood methods, the instrumental variable is also a popular method. This can date back to Wald (1940) and Durbin (1954). Recent work includes Amemiya (1985, 1990). Amemiya and Fuller (1988) and Fuller (1987) give more references. To illustrate the idea, suppose we have observations $(Y_i, X_i) = (y_{0i} + \epsilon_i, x_{0i} + e_i)$ and $y_{0i} = f(x_{0i}, \mu)$ with μ unknown, (ϵ_i, e_i) being independent errors. This can be written as:

$$\begin{cases} Y_i &= f(x_{0i}, \mu) + \epsilon_i \\ X_i &= x_{0i} + e_i. \end{cases}$$

If everything on the right hand side involves unknown parameters, this falls into the form of (1).

Mixed effects models are among the sub-models of (1). If we view a fixed parameter as a random one supported by a single point, mixed effects models become random coefficient models. A very important example is the analysis of repeated measures data, which becomes more and more important in applications. In the pharmaceutical industry, repeated measures data are seen in population pharmacokinetics (PK) and population pharmacodynamics (PD). In a typical population PK study, plasma concentration Y_{ij} of a certain drug is measured for patient i at time t_{ij}, $1 \le i \le I$, $1 \le j \le J$, after having taken the drug in a schematic way. These concentration measurements are modeled as: $Y_{ij} = f(t_{ij}, \mu_{ij}, \eta_i, \epsilon_{ij})$, or in vector form:

$$Y_i = \tilde{f}(t_i, \mu_i, \eta_i, \epsilon_i), \tag{2}$$

where ϵ_{ij} is the measurement error, μ_i is the vector of controlled or observed covariates and η_i is the vector of unobserved variables of patient i that are believed to have effects on the concentration. Although other models are being used, compartment models by far are the most popular ones in PK studies because of their clear pharmacological interpretations. The *one compartment model with first order absorption* (with single dose), for example, can be written as:

$$Y_{ij} = \frac{D_i k_{ia}}{V_i(k_{ia} - k_{ie})}(e^{-k_{ie}t_{ij}} - e^{-k_{ia}t_{ij}}) + \epsilon_{ij}; \quad 1 \le j \le J_i, \quad 1 \le i \le I.$$

Here $\mu_{ij} = D_i$ is the dose for patient i (for multiple doses, this might be a vector), and $\eta_i = (V_i, k_{ia}, k_{ie})$ is the vector for the pharmacokinetic parameters of patient i.

Population PK studies usually involve large inter-individual variations of the PK parameters that we are interested in. Therefore it is necessary to treat η_i as

random effects. This leads us to the assumption that $(\eta_1, \epsilon_1), \ldots, (\eta_I, \epsilon_I)$ are i.i.d. random vectors generated by an unknown distribution P_θ, for some $\theta \in \Theta$. The goal of the statistical analysis is to gain information and make inferences about θ. For further information about the population PK models, see Gibaldi and Perrier (1982), and more recently Grasela and Sheiner (1991a,b). Considering (2) as a model in the form of (1) assumes that the sampling times t_{ij} are also random. Assumptions of this kind are very common in regression analysis. One often finds that a fixed design has the same asymptotic properties under mild conditions.

The methods of estimating θ in nonlinear mixed effects models have been studied in population PK. Beal and Sheiner and their co-authors have done a series of work: Beal (1984), Beal and Sheiner (1988, 1985), Sheiner and Beal (1980, 1981, 1983, 1987) and Sheiner and Ludden (1992). A software package called NONMEM has even been developed (Beal and Sheiner (1989)). The methods suggested by them can be classified into two groups: the two-stage methods and the ELS (Extended Least Square) methods (Sheiner and Beal (1987)). Two-stage methods try to fit each individual patient by classical nonlinear regression techniques to get estimates for $\eta_1, \ldots, \eta_I$ and then estimate θ from these intermediate estimates. These methods yield good estimates when both I and the J_i's are large. They do not use all the information available since they ignore, in the first stage, that the η's come from the same distribution. The ELS methods try to make use of this information. The *first order* (FO) *method*, the simplest of ELS methods, is as follows: 1. Linearize model (2) by a Taylor expansion, which gives us a linear model, called the first-order model; 2. Write down the likelihood of the first-order model as if the random effects follow a multivariate normal distribution; 3. Use maximum likelihood methods to do the estimation. The major advantage of ELS methods over two-stage methods is that they do not require large J_i's (sample sizes on each individual) to give reasonably good estimates. However, one can find two obvious flaws in ELS methods too: 1. These methods generally can only estimate the first two moments of the random effects. 2. More importantly, there is little justification for the linearization step. This means that ELS estimates have bias that cannot be reduced simply by increasing the sample size. Two-stage methods and ELS methods are all variations of least-squares. Least squares is designed for normal variables, and is very sensitive to outliers and poor-quality samples. A more robust method is desirable in applications like population PK/PD.

Other statisticians have studied nonlinear mixed effects models from different aspects under various names, including repeated measures data, growth curve data, longitudinal data. This research includes Dempster (1984), Lindstrom (1984), Lindstrom and Bates (1990), Stiratelli, Laird and Ware (1984), Lange

and Laird (1989), Lipsitz, Laird and Harrington (1990) and Laird (1991). The research concentrated on likelihood-related least-squares type methods involves computational issues, as do many other papers.

In this paper, the minimum distance (MD) method is generalized to the nonlinear model. Section 2 introduces the concepts of identifiability and differentiability and presents the main results. The nonparametric MD estimator is shown to be consistent, and the parametric MD estimator $\sqrt{n}$-consistent under the assumption of differentiability. Section 3 applies these results to discrete models and absolutely continuous models. Section 4 explains the computational difficulties and compares the MD and FO estimators by simulation examples. Section 5 contains some of the proofs.

2. Main Results

The minimum distance (MD) method goes back to Wolfowitz (1953, 1957) and Kac, Kiefer and Wolfowitz (1955). Many statisticians have studied the minimum distance method since then. The author's idea comes mainly from Pollard (1980), Beran (1993) and Beran and Millar (1994). Other related papers include Durbin (1954), Bolthausen (1977), Millar (1984), Donoho and Liu (1988a,b).

Pollard (1980) considered the problem of goodness-of-fit testing via an MD approach. He proved the $\sqrt{n}$-consistency of the MD estimator and found its asymptotic distribution. Although Pollard concentrated on the finite dimensional parametric case, consistency of the MD estimator can be easily generalized to the nonparametric case. Slightly more is needed in our context than in Pollard's. Pollard sought to minimize the distance between $\hat{F}_n$ and F where $\hat{F}_n$, the empirical distribution, is determined completely by the data and F completely specified by the model. We minimize the distance between $\hat{P}_{XY,n}$ and $P(P_X, \theta)$, whose definitions will be given later. Here $\hat{P}_{XY,n}$ is determined by the sample and $P(P_X, \theta)$ depends on both the sample and the unknown parameter. This difference has little impact on the proof of the consistency; but it forces us to adopt another notion of differentiability—an important condition in proving the $\sqrt{n}$-consistency.

For model (1), let $P(P_X, \theta)$ be the distribution of (X, Y) where $X \sim P_X$, $A \sim P_\theta$ and X and A are independent, $\mathcal{P}_{XY} = \{P(P_X, \theta) : \theta \in \Theta, P_X \in \mathcal{P}_X\}$.

Definition 1. Model (1) is called identifiable if θ, $\theta' \in \Theta$, P_X, $P'_X \in \mathcal{P}_X$, and $P(P_X, \theta) = P(P'_X, \theta')$ imply that $\theta = \theta'$ and $P_X = P'_X$. Let $\overline{\mathcal{P}_X}$ be the collection of all probability measures that are either in $\mathcal{P}_X$ or have finite support on $\mathcal{X}$. Model (1) is called strongly identifiable if for any $\theta_n, \theta \in \Theta$, $P_{X,n} \in \overline{\mathcal{P}_X}$, $P_X \in \mathcal{P}_X$:

$$P(P_{X,n}, \theta_n) \Rightarrow P(P_X, \theta) \text{ if and only if } \theta_n \to \theta \text{ and } P_{X,n} \Rightarrow P_X.$$

Definition 2. Let ρ be a metric on the space of all probability measures on R^{m+k}. Let $\hat{P}_{X,n}$ be the empirical distribution that assigns mass $1/n$ on each X_{ni}, $i = 1, \ldots, n$, $\hat{P}_{XY,n}$ the empirical distribution that puts mass $1/n$ on each of (X_{ni}, Y_{ni}), $1 \leq i \leq n$. We call any measurable sequence $\hat{\theta}_n$ a *minimum distance estimator* (with respect to ρ) if

$$\rho(P(\hat{P}_{X,n}, \hat{\theta}_n), \hat{P}_{XY,n}) = \inf_{\theta \in \Theta} \rho(P(\hat{P}_{X,n}, \theta), \hat{P}_{XY,n}) + o(\frac{1}{\sqrt{n}}),$$

and we call $T_n = \sqrt{n} \inf_{\theta \in \Theta} \rho(P(\hat{P}_{X,n}, \theta), \hat{P}_{XY,n})$ the *minimum distance*.

The following theorem gives an easy but useful relationship between identifiability and strong identifiability.

Theorem 1. *Model* (1) *is strongly identifiable if*
1. *it is identifiable,*
2. Θ *is a countably compact topological space, i.e., any sequence has a cluster point,*
3. *for any* θ_n, $n \geq 1$ *and* θ *in* Θ, $\theta_n \to \theta$ *if and only if* $P_{\theta_n} \Rightarrow P_\theta$, *and*
4. f *is continuous in both arguments.*

Beran (1993) and Beran and Millar (1994) proved identifiability for linear models under mild conditions. Such general results for nonlinear models seem to be impossible. However, Liu (1994) showed that if the random coefficient has a discrete distribution, identifiability can be proved.

Example 1. Under the following assumptions, Model (1) is identifiable.
1. $\mathcal{X} \subseteq R^k$ has non-empty interior,
2. f is continuous in both its arguments and if $a_1, a_2 \in \mathcal{A}$ satisfy $f(x, a_1) = f(x, a_2)$ for all t in some nonempty open set $B \subseteq \mathcal{X}$, then $a_1 = a_2$,
3. any $P_X \in \mathcal{P}_X$ has full support over X,
4. $\{P_\theta : \theta \in \Theta\}$ consists of discrete distributions on $\mathcal{A}$ with finite support.

Beran (1993) showed that strong identifiability implies consistency of the MD estimator for the linear model. This is still true for the nonlinear model.

Theorem 2. *Let* $\hat{\theta}_n$ *be any MD estimator as defined in Definition* 2. *Suppose* ρ *metrizes weak convergence of probability measures and* $\theta_n \to \theta_0 \in \Theta$, $P_{X,n} \Rightarrow P_X$. *Then* $\hat{\theta}_n$ *is consistent, i.e.,* $\hat{\theta}_n \to \theta_0$ *in probability in* Θ.

The proof is given in Liu (1994). Notice that the proof of Beran and Millar (1994) for its linear counterpart could be carried through almost unchanged here. This theorem does not make any assumptions about the form of P_θ; it applies to the nonparametric distribution families. Beran and Millar (1994) also have a consistency result for fitting a discrete model with an increasing number of

supporting points, and this can certainly be generalized to the nonlinear case based on Example 1. The major difficulty for the nonlinear model lies in establishing the strong identifiability. Liu (1994) gave a few techniques to do this. The following example is taken from that dissertation.

Example 2. In population pharmacokinetics, the concentration of an intravenous injection drug is often assumed to follow the model: $Y = Be^{-AX} + \epsilon$, and we observe i.i.d. copies of (X, Y). Let us also assume that A, B, ϵ and X are mutually independent, all nonnegative except ϵ, the support of X is $[0, \infty)$, B is bounded and $E\epsilon = 0$. Then this model is identifiable. Let F and G be two cdf's on R, $F(0) = 0$ and $\int |x| G(dx) < \infty$ and define, for a fixed $M > 0$, distribution families

$$\begin{aligned}
\mathcal{P}_A &= \{P : \ P\{[0,\infty)\} = 1, \ P\{(x,\infty)\} \le 1 - F(x), \text{ for any } x > 0\}, \\
\mathcal{P}_B &= \{P : \ P\{[0,M]\} = 1\}, \\
\mathcal{P}_\epsilon &= \{P : \ \int xP(dx) = 0, \ P\{(x,\infty)\} \le 1 - G(x), \text{ for any } x > 0\}.
\end{aligned}$$

Let us further assume that the distributions of A, B and ϵ belong to $\mathcal{P}_A$, $\mathcal{P}_B$ and $\mathcal{P}_\epsilon$, respectively. Take the topology on $\Theta = \{P_A \times P_B \times P_\epsilon : \ P_A \in \mathcal{P}_A, \ P_B \in \mathcal{P}_B, \ P_\epsilon \in \mathcal{P}_\epsilon\}$ to be that of weak convergence; then strong identifiability holds. The MD estimator defined by any weak-convergence-measuring metric ρ for the distribution of (A, B, ϵ) is consistent. We will give an example for such a ρ later. It also can be shown that this result is true for models like $Y = \sum_1^N B_j e^{-A_j X} + \epsilon$.

To get a further asymptotic result, we restrict ourselves to the parametric case, where differentiability is easily defined. Let H be a separable Hilbert space, Θ a subset of R^p.

Definition 3. Let $\theta_0 \in \Theta^o$ be an inner point of Θ. A sequence of stochastic processes $\varphi_n : \Theta \to H$ is called *asymptotically (norm) differentiable* at θ_0 if there exists a vector $D = D(\theta_0) \in H^p$ such that for any $|h_n| \to 0$ in R^p, $\|\varphi_n(\theta_0 + h_n) - \varphi_n(\theta_0) - h_n' D\| = o_{P_n}(|h_n|)$, as $n \to \infty$, or equivalently

$$\sup_{0 \le |t| \le h_n} \frac{1}{|t|} \|\varphi_n(\theta_0 + t) - \varphi(\theta_0) - t'D\| \xrightarrow{P_n} 0, \tag{3}$$

where P_n is the probability on which φ_n is defined. Call D the *asymptotic derivative* at θ_0.

This definition allows φ_n to be defined on different spaces but requires that the asymptotic derivative D be deterministic.

Definition 4. A vector $D \in H^p$, where H is a Hilbert space, is called *nonsingular* if $\|t'D\| > 0$ for any non-zero t in R^p, or equivalently, the components of D are linearly independent.

From now on, we consider Hilbert spaces H that satisfy:

A1. H contains all probability measures on R^{m+k}, and there exists $C_1 > 0$ such that $\|P\|_H \leq C_1$ for any probability measure P.

A2. Convergence in H for probability measures is equivalent to weak convergence.

A3. If $X \sim P$, then $E\langle \delta_X, h\rangle = \langle P, h\rangle$ for any $h \in H$ and any probability P. Here $\langle \cdot, \cdot \rangle$ is the inner product in H and δ_X the point mass at X.

The metric of any H that fulfills these three conditions will serve for the purpose of Theorem 2, see the discussion in Example 2. One example of such a Hilbert space is as follows.

Example 3. Let Q_0 be a probability measure on R^{k+m} and assume that the support of Q_0 is the full space. Take H to be the collection of all the complex-valued functions on R^{m+k} that are square integrable under Q_0. H is a Hilbert space with the inner product $\langle f, g\rangle_H = \int f(s)\overline{g(s)}Q_0(ds)$. For any probability measure μ on R^{m+k}, its characteristic function $\varphi_\mu(s) = \int e^{is'x}\mu(dx)$ maps it into H. If we identify μ with φ_μ, it can be shown that H satisfies **A1**, **A2** and **A3** with $C_1 = 1$. Both Beran (1993) and Liu (1994) used this space, but the results for the MD estimator do not require any other special properties of H.

Becasuse of this natural link between μ and φ_μ, we write the ch.f. of $P(P_X, \theta)$ as $\varphi(P_X, \theta)(t, s)$ and that of $\hat{P}_{XY,n}$ as $\psi_n(t, s)$. So the MD estimator can be taken as any measurable sequence $\hat{\theta}_n$ such that

$$\|\varphi(\hat{P}_{X,n}, \hat{\theta}_n) - \psi_n\| = \inf_{\theta\in\Theta} \|\varphi(\hat{P}_{X,n}, \theta) - \psi_n\| + o(\frac{1}{\sqrt{n}}),$$

and we denote the minimum distance $\sqrt{n}\inf_{\theta\in\Theta} \|\varphi(\hat{P}_{X,n}, \theta) - \psi_n\|$ by T_n.

Theorem 3. *Let H be a Hilbert space satisfying* **A1**, **A2** *and* **A3**, $\theta_0 \in \Theta \subseteq R^p$ *an inner point of* Θ, $W_n = \sqrt{n}(\varphi(\hat{P}_{X,n}, \theta_n) - \psi_n)$. *If* $|\theta_n - \theta_0| = O(1/\sqrt{n})$ *and* $P_{X,n} \Rightarrow P_X \in \mathcal{P}_X$ *and the following conditions hold*:

1. *model* (1) *is strongly identifiable,*
2. *as a stochastic process in H, $P(\hat{P}_{X,n}, \theta)$ is asymptotically differentiable at θ_0 with derivative $D_0 = D(\theta_0)$,*
3. *the asymptotic derivative D_0 is non-singular.*

Then,

1. *there exists a mean zero Gaussian random element W in H such that $W_n \to W$,*
2. *the minimum distance T_n converges weakly to $\inf_{t\in R^p} \|W + t'D\|$ and $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to $-(\Re\langle D, D'\rangle)^{-1}\Re\langle W, D\rangle$, where $\Re$ represents the real part.*

Besides the $\sqrt{n}$-consistency of the MD estimator $\hat{\theta}_n$, this theorem also gives the asymptotic distribution of the minimum distance T_n, which can be used as a goodness-of-fit statistic. The asymptotic distribution of T_n^2 is a weighted sum of χ^2 distributions whose computation is analytically tedious as shown by Durbin (1954). However, the triangular form of Theorem 3 implies that $\hat{\theta}_n$ and T_n can be bootstrapped. In particular, the following theorem is readily shown to be true.

Theorem 4. *Suppose $P_{X,n} = P_X$, $\theta_n = \theta_0$ and that Conditions* 1-3 *of Theorem* 3 *hold. Let $\hat{\theta}_n$ and T_n be the MD estimator and the minimum distance respectively. Let $S_n^* = \{(X_{n,k}^*, Y_{n,k}^*) : 1 \le k \le n\}$ be i.i.d. samples from the random measure $P(\hat{P}_{X,n}, \hat{\theta}_n)$ and $\hat{\theta}_n^*$ and T_n^* be the MD estimator and minimum distance from S_n^* respectively. Denote G_0 and J_0 as the asymptotic distributions of $\hat{\theta}_n$ and T_n respectively and G_n^* and J_n^* as the distributions of $\hat{\theta}_n^*$ and T_n^* respectively. (G_n^* and J_n^* are random measures.) Then*

$$G_n^* \Rightarrow G_0, \qquad J_n^* \Rightarrow J_0$$

in probability. In particular, for $0 < \alpha < 1$, if U_α and $U_{\alpha,n}^*$ are the upper α-quantiles of J_0 and J_n^* respectively, $U_{\alpha,n}^* \to U_\alpha$ in probability.

3. Further Applications

In this section, we study two direct applications of the general $\sqrt{n}$-consistency result in the last section. We make use of the Hilbert space H defined in Example 3 and deduce the properties of the MD estimator for discrete models and absolutely continuous models.

Model (1), when Θ is parametric, is called absolutely continuous if $\{P_\theta : \theta \in \Theta\}$ consists of absolutely continuous distributions. It is called discrete if $\{P_\theta : \theta \in \Theta\}$ consists of discrete distributions with M supporting points for some known M. We will be particularly interested in two sampling schemes: i.i.d. sampling when $\theta_n = \theta_0$ and $P_{X,n} = P_X$ and standard bootstrap sampling defined as the S_n^* in Theorem 4.

3.1. Discrete models

Discrete models can be used as approximations of nonparametric models. The impact of using an approximation model on the asymptotic behavior of the MD estimator is studied in Liu (1994). Beran (1993) and Beran and Millar (1994) use discrete models as approximations when the parametric form of P_θ is unknown. Discrete models are also useful when the underlying distribution is a mixture. When there is a certain unmeasured biological factor that has significant effects on the pharmacokinetic parameters (e.g., a related disease) patients with

and without that disease may have quite different mean parameters. Fitting discrete models can help to detect this.

Our first task is to parametrize discrete distributions. Define an order in R^d as follows: for $x = (x_1, \ldots, x_d)$, $y = (y_1, \ldots, y_d) \in R^d$, $x > y$ if for some $1 \le i \le d$, $x_i > y_i$ and $x_j = y_j$ for all $j < i$. This is called the alphabetic order. For any $\mathcal{A} \subseteq R^d$, let $\Theta \subseteq R^{M(d+1)}$ be the set of all points $(b_1, \ldots, b_M)$, $b_i = (p_i, a_{i,1}, \ldots, a_{i,d})$ such that $p_i > 0$, $\sum_{i=1}^M p_i = 1$, $a_i = (a_{i,1}, \ldots, a_{i,d}) \in \mathcal{A}$ and $a_1 > \cdots > a_M$. For any $\theta = (b_1, \ldots, b_M) \in \Theta$, P_θ, the probability measure that assigns mass p_i to $(a_{i,1}, \ldots, a_{i,d})$, $1 \le i \le M$, defines a one-to-one correspondence between Θ and all the discrete distributions on $\mathcal{A}$ with exactly M supporting points. When using this parmaetrization, we often write $\theta = \{(p_i, a_i) : \ 1 \le j \le M\}$.

Because of the constraint $\sum_{i=1}^M p_i = 1$, Θ is in a subspace of $R^{M(d+1)}$ with dimension $M(d+1)-1$. When doing computations, we always substitute for $p_{_M}$ with $1 - \sum_{i=1}^{M-1} p_i$. However, this substitution would not affect the differentiability or the nonsingularity of the derivatives. We proceed as if this constraint does not exist. A linear constraint will not change the differentiability by the chain rule. Assume the derivatives w.r.t. $p_1, \ldots, p_{_M}$ are $f_{p_1}, \ldots, f_{p_{_M}}$ respectively; then the derivatives w.r.t. $p_1, \ldots, p_{_{M-1}}$ will be $f_{p_1} - f_{p_{_M}}, \ldots, f_{p_{_{M-1}}} - f_{p_{_M}}$ under the constraint. Therefore, the constrained derivatives exist if the unconstrained ones exist, and the constrained derivatives are linearly independent if the unconstrained ones are.

Fix $\theta_0 = \{(p_{0j}, a_{0j}) : \ 1 \le j \le M\}$ in the interior of Θ. This interior should be considered as an open set in $R^{M(d+1)-1}$, not $R^{M(d+1)}$.

Theorem 5. *Let H and φ_μ be defined as in Example 3, $\hat{\theta}_n$ be the MD estimator and T_n be the minimum distance defined by H. Let $\varphi_n(\theta)(s) = \varphi_{P(\hat{P}_{X,n},\theta)}(s)$, $\psi_n(s) = \varphi_{\hat{P}_{XY,n}}(s)$. Suppose the data are from either i.i.d. sampling or standard bootstrap sampling. In addition, suppose the following conditions hold :*
1. *$\mathcal{X} \subseteq R^k$ has non-empty interior and $\mathcal{A}$ is compact,*
2. *P_X has positive, continuous density almost everywhere on $\mathcal{X}$,*
3. *if a_1, $a_2 \in \mathcal{A}$, and $f(x, a_1) = f(x, a_2)$ for all x in some open ball $B \subseteq \mathcal{X}$, then $a_1 = a_2$,*
4. *$f(x, a)$ is twice differentiable in a,*
5. *Q_0 has finite 4th moments,*
6. *for any $a \in \mathcal{A}$, there exists a $\delta > 0$ such that*

$$E_{P_X} \sup_{|h| \le \delta} \left| \frac{\partial f}{\partial a}(X, a+h) \right|^4 < \infty, \quad E_{P_X} \sup_{|h| \le \delta} \left| \frac{\partial^2 f}{\partial a^2}(X, a+h) \right|^2 < \infty,$$

7. *the columns of the matrix $\frac{\partial f}{\partial a}(x, a_{0j})$ are linearly independent as functions of x for every $1 \le j \le M$, and*

8. $f(x, a_{0j}), 1 \le j \le M$, *are distinct for almost every* $x \in \mathcal{X}$.
For any possible distribution P *of* X *and any* $\theta \in \Theta$, *let* $D(\theta, P)$ *be a* $K = M(d+1) - 1$ *vector with components in* $L^2(Q_0)$ *defined formally as*

$$E_P g(a_j, t, u, X) - E_P g(a_M, t, u, X), \qquad j = 1, \dots, M-1,$$
$$ip_j u' E_P \frac{\partial f}{\partial a}(X, a_j) g(a_j, t, u, X), \qquad j = 1, \dots, M,$$

where $g(a, t, u, x) = \exp\{it'x + iu'f(x, a)\}$ *and* $\{(p_j, a_j) : 1 \le j \le M\}$ *are the coordinates of* θ. *Write* $D_0 = D(\theta_0, P_X)$.

Define $W_n = \sqrt{n}(\varphi_n(\theta_n) - \psi_n)$, $\tilde{\varphi}(t, u, x) = \sum_{j=1}^{M} p_{0j} \exp\{it'x + iu'f(x, a_{0j})\}$ $= E_{\theta_0} g(A, t, u, x)$ *and let* W *be a mean-zero Gaussian random element in* H *with covariance operator:*

$$E|\langle h, W\rangle|^2 = \int h(t, u)\overline{h(s, v)} S(t, u, s, v) Q_0(dt, du) Q_0(ds, dv),$$

where $S(t, u, s, v) = E_{P_X} \tilde{\varphi}(t - s, u - v, X) - E_{P_X} \tilde{\varphi}(t, u, X)\tilde{\varphi}(-s, -v, X)$. *Then* $W_n \Rightarrow W$ *and*

$$T_n = \inf_{\theta \in \Theta} \|\varphi_n(\theta) - \psi_n\| \Longrightarrow \inf_{t \in R^K} \|W + t'D_0\|,$$
$$\sqrt{n}(\hat{\theta}_n^{(1)} - \theta_n^{(1)}) \Longrightarrow -(\Re\langle D, D'\rangle)^{-1} \Re\langle W, D\rangle,$$

for any MD estimator $\hat{\theta}_n$, *where* $\hat{\theta}_n^{(1)}$ *and* $\theta_n^{(1)}$ *represent the reduced vectors by leaving out the* p_M*-coordinates, so as to make the parameters free.*

The idea of the proof will be sketched later. It helps to understand these conditions. Conditions 1, 2, 3 imply strong identifiability. Condition 3 seems to be the strongest one, but it is satisfied by any f that is analytic in x. Conditions 4, 5, 6 are for the differentiability. Condition 8 is very crude. This theorem by no means looks for the weakest conditions. Conditions 2, 7, 8 will ensure the non-singularity of the derivative.

Example 4. Multi-exponential model $f(x, a) = \sum_1^r b_j e^{-a_j x} + \epsilon$ often appears in pharmacokinetics; here $a = (a_1, b_1, \dots, a_r, b_r, \epsilon)$. Let $\mathcal{X} = [0, \infty)$, and $\mathcal{A}$ be the set of $(a_1, b_1, \dots, a_r, b_r, \epsilon)$ satisfying:
1. $0 < \delta \le a_j, b_j \le M$, $1 \le j \le r$, and $|\epsilon| \le M$, and
2. $a_j + \delta \le a_{j+1}$, $1 \le j \le r - 1$,
for some δ and M. This assumption has a few implications. First, it says the components of a vary over a finite range. It also bounds them away from zero. Moreover, it requires that the exponential rates not be too close to each other. In practice, this means that if two rates were too close, we would not be able to distinguish them, just as if a rate were too close to zero, we would

not be able to distinguish that term from the constant term. Let $\{P_\theta : \theta \in \Theta\}$ be the collection of all discrete distributions on $\mathcal{A}$ with M supporting points parameterized by the alphabetic order, and $\mathcal{P}_\mathcal{X}$ be that of all distributions on $\mathcal{X}$ with positive density. Let Q_0 be a distribution on R^2 with full support such that $\int |x^2 + y^2|^2 Q_0(dx, dy) < \infty$. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. samples from $P(P_X, \theta_0)$ with $P_X \in \mathcal{P}_X$, $\theta_0 \in \Theta$, $\psi_n(t,u) = 1/n \sum_{j=1}^n \exp\{itX_j + iuY_j\}$ and $\varphi_n(t,u;\theta) = 1/n \sum_{j=1}^n \sum_{l=1}^M p_l \exp\{itX_j + iu(\sum_{m=1}^r b_{ml} e^{-a_{ml}X_j} + \epsilon_l)\}$, where $\theta = \{(p_l, a_{1l}, b_{1l}, \ldots, a_{rl}, b_{rl}, \epsilon_l), 1 \le l \le M\}$. Then the MD estimator $\hat{\theta}_n$ defined by

$$\int |\psi_n(t,u) - \varphi_n(t,u;\hat{\theta}_n)|^2 Q_0(dt,du)$$
$$= \inf_{\theta\in\Theta} \int |\psi_n(t,u) - \varphi_n(t,u;\theta)|^2 Q_0(dt,du) + o(\frac{1}{\sqrt{n}})$$

is $\sqrt{n}$-consistent.

Liu (1994) also showed that if A_1, $B_1, \ldots, A_r$, B_r and ϵ are independent, all with discrete distribution and $E\epsilon = 0$, the MD estimator remains $\sqrt{n}$-consistent.

3.2. Absolutely continuous models

Absolutely continuous models are often used in practice. In pharmacokinetics, for example, it is often assumed that the parameters have normal distributions. Write $dP_\theta = h(a,\theta)\eta(da)$, $\theta \in \Theta \subseteq R^p$ for some σ-finite measure η and take θ_0 an inner point of Θ. Unlike the discrete case, where identifiability is guaranteed by Conditions 1, 2 and 3 in Theorem 5, identifiability has to be proved case by case. Liu (1994) established identifiability for the multi-exponential models with normal assumptions.

Example 5. Consider a multi-exponential model $Y = \sum_{j=1}^k B_j e^{-A_j X} + \epsilon$. Assume that $A_1, B_1, \ldots, A_k, B_k, \epsilon$ and X are independent; $A_j \sim N(\mu_j, \sigma_j^2)$, $B_j \sim P_{\lambda_j}$, $\lambda_j \in \Lambda$, $\epsilon \sim P_\omega$, $\omega \in \Omega$, with μ_j, σ_j^2, λ_j and ω unknown. Suppose $X > 0$, the support of X has a convergent point or has a subsequence diverging to $+\infty$, that P_λ has bounded support for any $\lambda \in \Lambda$ and we observe (X, Y). Then this model is identifiable w.r.t. $\theta = (\mu_1, \sigma_1^2, \lambda_1, \ldots, \mu_k, \sigma_k^2, \lambda_k)$ up to the order of $(\mu_j, \sigma_j, \lambda_j)$ in θ provided $E_{P_\omega}\epsilon = 0$ for any $\omega \in \Omega$. This model remains identifiable if we assume $A_j \sim \text{LogNormal}(\mu_j, \sigma_j^2)$ instead of $N(\mu_j, \sigma_j^2)$.

The proof of this example depends on the result of identifiability for linear models established by Beran and Millar (1994). For most applications, identifiability is natural.

Theorem 6. *Under the absolutely continuous model, let H, φ_n, ψ_n, W_n, $\hat{\theta}_n$ and T_n be as in Theorem 5. Suppose data are collected by i.i.d. sampling or standard bootstrap sampling and the following conditions hold* :

1. *the absolutely continuous model is strongly identifiable,*
2. *the second derivative* $(\partial^2 h/\partial\theta^2)(a,\theta)$ *exists around* θ_0 *for every* a,
3. *for some* $\delta > 0$, $\int \sup_{|\theta-\theta_0|\le\delta} |(\partial^j h/\partial\theta^j)(a,\theta)|^2 \eta(da) < \infty$, $j = 1, 2$,
4. P_X *has positive, continuous density over* $\mathcal{X}$, *and*
5. *as functions of* t *and* u *the components of the vector*

$$D = \int\int \frac{\partial h}{\partial\theta}(a,\theta_0)p(x)\exp\{it'x + iu'f(x,a)\}\,\eta(da)dx$$

are linearly independent.

Let $\tilde{\varphi}(t,u,x) = \int h(a,\theta_0)\exp\{it'x + iu'f(x,a_{0j})\}\eta(da) = E_{\theta_0}g(A,t,u,x)$ *and* W *be a mean zero Gaussian random element in* H *with covariance operator*

$$E|\langle h, W\rangle|^2 = \int h(t,u)\overline{h(s,v)}S(t,u,s,v)Q_0(dt,du)Q_0(ds,dv),$$

where $S(t,u,s,v) = E_{P_X}\tilde{\varphi}(t-s,u-v,X) - E_{P_X}\tilde{\varphi}(t,u,X)\tilde{\varphi}(-s,-v,X)$. *Then* $W_n \Rightarrow W$,

$$T_n \Rightarrow \inf_t \|W + t'D\| \quad \textit{and} \quad \sqrt{n}(\hat{\theta}_n - \theta_n) \Rightarrow -(\Re\langle D, D'\rangle)^{-1}\Re\langle W, D\rangle.$$

Example 6. Consider the multi-exponential model for the absolutely continuous case. Take $f(x,a) = \sum_{j=1}^r b_j e^{-a_j x} + \epsilon$, $a = (a_1, b_1, \ldots, a_r, b_r, \epsilon)$ and assume:
1. $\mathcal{X} = [0, +\infty)$, $\mathcal{P}_X = \{\text{distributions on } \mathcal{X} \text{ with positive density}\}$.
2. $A_1, B_1, \ldots, A_r, B_r, \epsilon$ are mutually independent, and all normally distributed: $A_j \sim N(\mu_j, \sigma_j^2)$, $B_j \sim N(\eta_j, \omega_j^2)$ and $\epsilon \sim N(0, \lambda^2)$.
3. The parameter space Θ is defined as all the $(\mu_1, \sigma_1^2, \eta_1, \omega_1^2, \ldots, \mu_r, \sigma_r^2, \eta_r, \omega_r^2, \lambda^2)$ satisfying: $\mu_j > 0$, $\eta_j > 0$; the sequence of pairs $(\mu_1, \sigma_1^2), \ldots, (\mu_r, \sigma_r^2)$ is strictly increasing in alphabetic order; and each μ_i, η_i, σ_i^2, ω_i^2 and λ is bounded.
Let $S_n = \{(X_k, Y_k) : 1 \le k \le n\}$ be an i.i.d. sample of size n from $P(P_X, P_{\theta_0})$, $\theta_0 \in \Theta$ and $\psi_n(t,u) = 1/n\sum_{k=1}^n \exp\{itX_k + iuY_k\}$, $\varphi_n(t,u;\theta) = 1/n\sum_{k=1}^n E_{P_\theta}\exp\{itX_k + iuf(X,A)\}$ for any $\theta \in \Theta$. Let Q_0 be a distribution on R^2 with full support such that $\int |x^2 + y^2|^2 Q_0(dx,dy) < \infty$. Then the MD estimator defined by

$$\int |\psi_n(t,u) - \varphi_n(t,u;\hat{\theta}_n)|^2 Q_0(dt,du)$$
$$= \inf_{\theta\in\Theta}\int |\psi_n(t,u) - \varphi_n(t,u;\theta)|^2 Q_0(dt,du) + o(\frac{1}{\sqrt{n}})$$

is $\sqrt{n}$-consistent.

4. Simulations

Estimating the distribution of the random coefficients nonparametrically is very difficult. Beran and Millar (1994) and Liu (1994) argued that, for the simple

linear regression model, it is more difficult than the problem of the inverse Radon transform, which has been known in applications such as computerized tomography, Deans (1983). It seems that the more feasible method is to use parametric approximations. Discrete models are therefore studied for their simplicity and the ability to approximate arbitrary distributions. Beran and Hall (1992) also used discrete models though their estimating criterion is moment matching.

Liu (1994) gave two numerical examples. One of these was for discrete models, and he showed that multiple minimal points were a very general phenomenon for even the simple models and concluded that some good optimization techniques had to be employed in order to get around this. It is worth knowing that this is also true for least-squares type estimators implemented in NONMEM.

The second example tried to compare the MD method with the simplest ELS method — the FO method. Consider the model $Y_i = \exp\{-A_i X_i\} + \epsilon_i$, $1 \le i \le n$. The assumptions on the distributions are: $(Y_j, X_j, A_j, \epsilon_j)$, $1 \le j \le n$ are i.i.d.; X_j, A_j and ϵ_j are mutually independent; $A_j \sim N(\mu, \sigma_A^2)$, $\epsilon_j \sim N(0, \sigma_\epsilon^2)$, $X_j \sim P_X$, $0 < \mu, \sigma_A^2, \sigma_\epsilon^2 \le M$, support$(P_X) \subseteq [0, M]$ and has a cluster point. Suppose $(X_1, Y_1) = (x_1, y_1), \ldots, (X_n, Y_n) = (x_n, y_n)$ are observed and we want to estimate $\theta = (\mu, \sigma_A^2, \sigma_\epsilon^2)$.

We use the Hilbert space as in Theorem 6 with the kernel probability Q_0 having independent normal marginals $N(0, \sigma_t^2)$ and $N(0, \sigma_u^2)$. Thus σ_t and σ_u can be used as parameters for the MD method.

For fixed θ_0 and P_X, $\{(y_j, x_j) : 1 \le j \le n\}$ were generated by the mechanism described in the model; then both FO and MD estimates were computed. To observe the effect of the kernel Q_0 on the MD estimator, four combinations of σ_t and σ_u were used: $(1,1)$, $(1,3)$, $(3,1)$ and $(3,3)$. Thus five estimators were obtained, one from the FO method, four from the MD method with different choices of (σ_t, σ_u). In order to compare these estimates, this process was repeated m times, each run resulting in five estimates. Then the m FO estimates were used to compute an average $\overline{\theta}_{FO}$ vector and the vector of their standard deviations of its components, and similarly we computed $\overline{\theta}_{MD,(1,1)}, \ldots, \overline{\theta}_{MD,(3,3)}$ and their corresponding SD vector. Some results are as follows:

1. The underlying model was taken as: $\theta_0 = (1, 0.2^2, 0.1^2)$, $P_X = \text{Uniform}(0,1)$, $n = 20$, and $m = 100$. To compare under the same scale, we used the estimates for σ_A and σ_ϵ rather than their squares in the table. The seven columns are for estimating methods, corresponding average estimates and the estimated standard deviations. Table 1 shows the results.
2. The underlying model was taken as: $\theta_0 = (2, 0.2^2, 0.1^2)$, $P_X = \text{Uniform}(0,1)$, $n = 20$, and $m = 100$. The output of the simulation is given in Table 2.
3. The underlying model was taken as: $\theta_0 = (1, 0.2^2, 0.1^2)$, $P_X = \text{Uniform}(0,1)$, $n = 100$, and $m = 20$. The output of the simulation is given in Table 3.

From these simulations and our understanding of the two methods, a few conclusions can be drawn:
1. Both of the methods can estimate the mean of the unknown parameter, μ, reasonably well. Estimates for the variance of the additive error become good when the sample size is very large ($n = 100$ in this case). Estimation for the variance of the unknown rate parameter is very difficult. Neither of the methods gives satisfactory results. Fortunately, μ is the most interesting parameter in most applications.
2. In the first two cases above, the FO method outperforms the MD method regardless of which kernel is used. The MD method also consumes considerably more computer time than the FO method does. Therefore, the FO method should be preferred when one has a strong belief in the model and the sample size is not too large. The reason is that if the sample size gets very large, we know that the FO method would lead to wrong answers because of the linear approximation. This is seen in Example 3 when the sample size is 100. The FO estimate of the mean has relatively larger bias. On the other hand, the MD estimate converges to the true parameter. The FO method may still be favored for its simplicity but in comparing it with the MD estimate we note its inconsistency.
3. The choice of kernel in estimating μ is an important factor for the MD method. For example, in Table 1, the estimates from the kernels with $\sigma_u = 1$ are always better than those with $\sigma_u = 3$. But in Table 2, we find that using $\sigma_u = 3$ would give us more accurate results. It is not clear how the kernel affects the estimations. It is suggested that a couple of kernels be tried.

Robustness of the two methods were investigated by contaminated data. The computer generated data just as before except that the distribution of ϵ was "contaminated". That is, ϵ follows a distribution of $pN(0, \sigma_1^2) + (1-p)N(0, \sigma_2^2)$, whose density is defined as

$$\frac{1}{\sqrt{2\pi}\sigma_1} \exp\Big\{ -\frac{x^2}{2\sigma_1^2}\Big\} p + \frac{1}{\sqrt{2\pi}\sigma_2} \exp\Big\{ -\frac{x^2}{2\sigma_2^2}\Big\}(1-p)$$

with $0 < p < 1$. The two methods are used to estimate the three parameters. Some results are given below. In Table 4 we present the results of estimating μ, using sample size $n = 20$, $m = 1000$ repetitions. In these simulations, the contamination percentage p is always taken to be 0.1 and $\sigma_1 = .1$, $\sigma_2 = .3$ and $\sigma_A = 0.2$.

These results show great robustness of the MD method. When contamination is present, the FO method performs very poorly. In applications like population pharmacokinetics, one often is interested in A, whose distribution depends on the parameters μ and σ_A in this example. Parameters involved in the distribution of ϵ are usually nuisance parameters. It is not desirable for a statistical procedure

to give very poor estimates for the interesting parameters when the assumptions about the nuisance parameters are violated. In nonlinear models that are derived through many approximations, these violations of the assumptions are very likely to happen.

Table 1. Simulation results with parameters $\theta_0 = (1, 0.2^2, 0.1^2)$, $n = 20$ and $m = 100$.

Method		$\overline{\mu}$	SD_μ	$\overline{\sigma}_A$	SD_{σ_A}	$\overline{\sigma}_\epsilon$	$\mathrm{SD}_{\sigma_\epsilon}$
FO		0.98931	0.09205	0.17215	0.15923	0.08473	0.03346
MD	$(\sigma_t, \sigma_u) = (1, 1)$	1.05148	0.17552	0.27380	0.35650	0.06444	0.05645
	$(\sigma_t, \sigma_u) = (1, 3)$	1.06688	0.26612	0.26692	0.37247	0.08225	0.03834
	$(\sigma_t, \sigma_u) = (3, 1)$	1.03714	0.13584	0.25399	0.30258	0.06371	0.05329
	$(\sigma_t, \sigma_u) = (3, 3)$	1.11055	0.32855	0.32984	0.46398	0.08406	0.04015

Table 2. Simulation results with parameters $\theta_0 = (2, 0.2^2, 0.1^2)$, $n = 20$ and $m = 100$.

Method		$\overline{\mu}$	SD_μ	$\overline{\sigma}_A$	SD_{σ_A}	$\overline{\sigma}_\epsilon$	$\mathrm{SD}_{\sigma_\epsilon}$
FO		2.00530	0.14385	0.23794	0.27905	0.086945	0.12098
MD	$(\sigma_t, \sigma_u) = (1, 1)$	2.12508	0.22941	0.46389	0.50995	0.05735	0.04888
	$(\sigma_t, \sigma_u) = (1, 3)$	2.05437	0.15120	0.27743	0.28812	0.06964	0.04294
	$(\sigma_t, \sigma_u) = (3, 1)$	2.12481	0.22922	0.46165	0.51139	0.057047	0.04921
	$(\sigma_t, \sigma_u) = (3, 3)$	2.05702	0.15562	0.28571	0.29268	0.06823	0.04378

Table 3. Simulation results with parameters $\theta_0 = (1, 0.2^2, 0.1^2)$, $n = 100$ and $m = 100$.

Method		$\overline{\mu}$	SD_μ	$\overline{\sigma}_A$	SD_{σ_A}	$\overline{\sigma}_\epsilon$	$\mathrm{SD}_{\sigma_\epsilon}$
FO		0.96947	0.04265	0.14028	0.09598	0.11751	0.02594
MD	$(\sigma_t, \sigma_u) = (1, 1)$	0.98904	0.03756	0.16253	0.16169	0.08815	0.03517
	$(\sigma_t, \sigma_u) = (1, 3)$	0.98312	0.03572	0.16396	0.12256	0.10285	0.01704
	$(\sigma_t, \sigma_u) = (3, 1)$	0.99028	0.03793	0.17129	0.17347	0.08397	0.03594
	$(\sigma_t, \sigma_u) = (3, 3)$	0.98212	0.03500	0.16354	0.12324	0.10327	0.01688

Table 4. Robust simulation with parameters $\mu = 1.0$, $\sigma_A = 0.2$, $n = 20$ and $m = 1000$.

		μ=1.0		μ=2.0	
Method		$\overline{\mu}$	SD_μ	$\overline{\mu}$	SD_μ
FO		0.72954	0.27003	1.47927	0.43534
MD	$(\sigma_t, \sigma_u) = (1, 1)$	0.96185	0.25921	1.94327	0.34026
	$(\sigma_t, \sigma_u) = (1, 3)$	1.17453	0.40106	2.08274	0.27145
	$(\sigma_t, \sigma_u) = (3, 1)$	0.96924	0.26033	1.96326	0.35178
	$(\sigma_t, \sigma_u) = (3, 3)$	1.12756	0.37694	2.09218	0.28099

The MD estimator is much more difficult to compute. This is due to the fact that each evaluation of the distance function requires $O(n^2)$ computations while the number for FO estimate is $O(n)$. It may be possible by choosing a better distance and algorithm to reduce the computational complexity for MD procedure. However, that is not the purpose of this paper. The two factors in choosing between FO and MD are computing time and data-model quality. When the data-model is of high quality the FO method is certainly more appealing. But when data and model have moderate discrepencies, one may have to sacrifice speed in order to produce a more robust result.

5. Proofs

Proof of Theorem 1. If $\theta_n \to \theta_0$ and $P_{X,n} \Rightarrow P_X$, then $P_{\theta_n} \Rightarrow P_\theta$ by Condition 3 and $P(P_{X,n}, \theta_n) \Rightarrow P(P_X, \theta_0)$ by the continuity of f.

Suppose $P(P_{X,n}, \theta_n) \Rightarrow P(P_X, \theta_0)$, i.e. we have $P_{X,n} \Rightarrow P_X$ as marginal distributions. By Condition 2, let $\tilde{\theta}$ be any cluster point of θ_n; then there exists a subsequence $\theta_{n'}$ such that $\theta_{n'} \to \tilde{\theta}$. So, from the first part of the proof, $P(P_{X,n'}, \theta_{n'}) \Rightarrow P(P_X, \tilde{\theta})$. Identifiability then forces $\tilde{\theta} = \theta_0$ and this proves $\theta_n \to \theta_0$.

Proof for Example 1. Let $X \sim P_X$, $A_1 \sim P_A$ and $A_2 \sim Q_A$ such that $f(X, A_1) \stackrel{d}{=} f(X, A_2)$. Suppose P_A has support $\{c_i : i \geq 1\}$ and Q_A has support $\{d_i : i \geq 1\}$. To prove P_A and Q_A have the same support, it suffices to show that from each c_n, there exists j such that $c_n = d_j$. Pick any closed ball $B \subseteq \mathcal{X}$ with positive radius and let

$$B_{nj} = \{x \in B : f(x, c_n) = f(x, d_j)\}.$$

Then B_{nj} is closed and $\cup_{j \geq 1} B_{nj} = B$. Therefore one of the B_{nj}'s contains a non-empty ball. Otherwise, one can find a sequence of closed balls $C_1 \supseteq C_2 \supseteq \cdots$ in B such that radius(C_i) goes to zero and $C_i \cap (\cup_{j \leq i} B_{nj}) = \emptyset$. These closed balls have a non-empty intersection which is not covered by any B_{nj}; this is a contradiction. Thus Condition 2 implies $c_i = d_j$ for some j and P_A and Q_A have the same support. This sequence of closed balls can be constructed as follows: C_1 exists because B_{n1} does not cover B and B_{n1} is closed. When $C_1 \supseteq C_2 \supseteq \cdots \supseteq C_k$ have been constructed, in C_k find a point x that is not in $B_{n,k+1}$. This is possible because $B_{n,k+1}$ does not cover C_k. Then there exists a closed ball C_{k+1} around x with positive radius and $C_{k+1} \cap B_{n,k+1} = \emptyset$ because $B_{n,k+1}$ is closed. We can also make sure that $C_{k+1} \subseteq C_k$ and radius(C_{k+1})<radius$(C_k)/2$.

Let $P_A(c_i) = p_i$, $Q_A(c_i) = q_i$, $i \geq 1$. Assume $\{c_i : 1 \leq i \leq N\}$ are distinct and define closed sets $C_{ij} = \{x \in B : f(x, c_i) = f(x, c_j)\}$, $j \geq i$. Since for any

$j \neq i$, C_{ij} does not contain any ball (Condition 2), there is a point $x \in B$ and a neighborhood B_x of x such that $B_x \cap C_{ij} = \emptyset$ for any $j \neq i$. This gives

$$P(Y = f(x, c_i), X \in B_x) = p_i P_X(X \in B_x) = q_i P_X(X \in B_x).$$

Thus $p_i = q_i$ because P_X has full support.

Proof for Example 2. We need only prove identifiability. The strong identifiability claim will then be obvious from Theorem 1 and the fact that $\Theta = \{P_A \times P_B \times P_\epsilon : \ P_A \in \mathcal{P}_A, \ P_B \in \mathcal{P}_B, \ P_\epsilon \in \mathcal{P}_\epsilon\}$ is countably compact.

The conditional expectation of Y given $X = x$ is $E(Y|X=x) = (EB)(Ee^{-Ax})$, so $Ee^{-Ax} = E(Y|X = x)/E(Y|X = 0)$. Therefore P_{XY} uniquely determines $q(x) = Ee^{-Ax}$ on $[0, \infty)$. But $q(x)$ uniquely determines the distribution of A by Proposition 8.5.1 of Breiman (1968). Now we can take e^{-AX} as our new X and the model reduces to a linear one, from which identifiability can be concluded by the result of Beran and Millar (1994).

Proof for Theorem 3. Let δ_x be the point mass at x, we then have

$$W_n = \sqrt{n}(\varphi_{P(\hat{P}_{X,n},\theta_n)} - \varphi_{\hat{P}_{XY,n}}) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (\varphi_{P(\delta_{X_{nj}},\theta_n)} - \varphi_{\delta_{(X_{nj},Y_{nj})}}).$$

This is a mean zero triangular array sum, each term is bounded by 2 and the summands converge in distribution to $\varphi_{P(\delta_X,\theta_0)} - \varphi_{\delta_{X,Y}}$ where $(X, Y) \sim P(P_X, \theta_0)$. So by the triangular array version of the CLT in Hilbert space, W_n converges weakly to a Gaussian variable W.

Let $h_n = \theta_n - \theta_0$, $\hat{h}_n = \hat{\theta}_n - \theta_0$ and $R(t) = \varphi_{P(\hat{P}_{X,n},\theta_0+t)} - \varphi_{P(\hat{P}_{X,n},\theta_0)} - t'D_0 = o_{P_n}(|t|)$ (differentiability). Apply this expansion to $\hat{\theta}_n$ and θ_n:

$$\begin{aligned}
\varphi_{P(\hat{P}_{X,n},\theta_0)} - \varphi_{\hat{P}_{XY,n}} &= \frac{W_n}{\sqrt{n}} - (\varphi_{P(\hat{P}_{X,n},\theta_n)} - \varphi_{P(\hat{P}_{X,n},\theta_0)}) = \frac{W_n}{\sqrt{n}} - h_n' D_0 - R(h_n), \\
\varphi_{P(\hat{P}_{X,n},\hat{\theta}_n)} - \varphi_{P(\hat{P}_{X,n},\theta_0)} &= \hat{h}_n' D_0 - R(\hat{h}_n), \\
\varphi_{P(\hat{P}_{X,n},\hat{\theta}_n)} - \varphi_{\hat{P}_{XY,n}} &= (\varphi_{P(\hat{P}_{X,n},\hat{\theta}_n)} - \varphi_{P(\hat{P}_{X,n},\theta_0)}) + (\varphi_{P(\hat{P}_{X,n},\theta_0)} - \varphi_{\hat{P}_{XY,n}}) \\
&= \hat{h}_n' D_0 - R(\hat{h}_n) + \frac{W_n}{\sqrt{n}} - h_n' D_0 - R(h_n).
\end{aligned}$$

It is not hard to see that the nonsingularity of D_0 is equivalent to the existence of a constant $C > 0$ such that $\|t'D_0\| \geq C|t|$ for any $t \in R^p$, then

$$\begin{aligned}
&\|\varphi_{P(\hat{P}_{X,n},\hat{\theta}_n)} - \varphi_{\hat{P}_{XY,n}}\| - \|\varphi_{P(\hat{P}_{X,n},\theta_0)} - \varphi_{\hat{P}_{XY,n}}\| \\
&\geq \|\hat{h}_n' D_0\| - \|R(\hat{h}_n)\| - 2\frac{W_n}{\sqrt{n}} - 2\|h_n' D_0\| - 2\|R(h_n)\| \\
&\geq C|\hat{h}_n'| - o_{P_n}(|\hat{h}_n|) - 2\frac{W_n}{\sqrt{n}} - 2\|D_0\|\|h_n'\| - 2o_{P_n}(|h_n|).
\end{aligned}$$

The fact that this quantity is non-positive and $|h_n| = O_{P_n}(1/\sqrt{n})$, $\|W_n\| = O_{P_n}(1)$ imply that $|\hat{h}_n| = O_{P_n}(1/\sqrt{n})$.

Let $\hat{t}_n = \sqrt{n}(\hat{h}_n - h_n) = \sqrt{n}(\hat{\theta}_n - \theta_n)$; then it follows that

$$\sqrt{n}[\varphi_{P(\hat{P}_{X,n},\hat{\theta}_n)} - \varphi_{\hat{P}_{XY,n}}] = W_n + \hat{t}_n' D_0 + o_{P_n}(1). \tag{4}$$

Let us compare the two functions $\sqrt{n}\|\varphi(P_{X,n}, \theta_0 + t/\sqrt{n}) - \varphi_{\hat{P}_{XY,n}}\|$ and $\|W_n + (t - \sqrt{n}(\theta_n - \theta_0))' D_0\|$. Since $\hat{t}_n = O_{P_n}(1)$, both of them, with high probability, attain their minimum in a finite range. On the other hand, for any finite $N > 0$,

$$\begin{aligned}
&\sqrt{n} \sup_{|t|\le N} \left| \|\varphi_{P(P_{X,n},\theta_0+t/\sqrt{n})} - \varphi_{\hat{P}_{XY,n}}\| - \|\frac{W_n}{\sqrt{n}} + (\frac{t}{\sqrt{n}} - (\theta_n - \theta_0))' D_0\| \right| \\
&\le N \sup_{|t|\le N} \frac{1}{N/\sqrt{n}} \|\varphi_{P(P_{X,n},\theta_0+t/\sqrt{n})} - \varphi_{P(P_{X,n},\theta_n)} - (\frac{t}{\sqrt{n}} - (\theta_n - \theta_0))' D_0\| \\
&\le N \sup_{0<|t|\le N} \frac{1}{t/\sqrt{n}} \|\varphi_{P(P_{X,n},\theta_0+t/\sqrt{n})} - \varphi_{P(P_{X,n},\theta_0)} - \frac{t'}{\sqrt{n}} D_0\| + \sqrt{n}\|R(\theta_n - \theta_0)\| \\
&= o_{P_n}(1), \quad \text{by (3)}.
\end{aligned}$$

This means the two functions are uniformly close to each other over any finite region. Thus we conclude:

$$\begin{aligned}
T_n &= \sqrt{n} \inf_{\theta\in\Theta} \|\varphi_{P(P_{X,n},\hat{\theta}_n)} - \varphi_{\hat{P}_{XY,n}}\| \\
&= \inf_{t\in R^p} \|W_n + (t - \sqrt{n}(\theta_n - \theta_0))' D_0\| + o_{P_n}(1) \\
&= \inf_{t\in R^p} \|W_n + t' D_0\| + o_{P_n}(1) \Longrightarrow \inf_{t\in R^p} \|W + t' D_0\|.
\end{aligned}$$

Let $t_n = -(\Re\langle D_0, D_0'\rangle)^{-1}\Re\langle W_n, D_0\rangle$, $t_n \Rightarrow -(\Re\langle D_0, D_0'\rangle)^{-1}\Re\langle W, D_0\rangle$. By (4), $T_n = \|W_n + \hat{t}_n' D_0\| + o_{P_n}(1)$. But we also know that $T_n = \inf_{t\in R^p} \|W_n + t' D_0\| + o_{P_n}(1)$. By a simple quadratic form calculation, it can be seen that this inf is attained by t_n, therefore: $\|W_n + \hat{t}_n' D_0\| = \|W_n + t_n' D_0\| + o_{P_n}(1)$, and

$$2\hat{t}_n'\Re\langle W_n, D_0\rangle + \hat{t}_n'\Re\langle D_0, D_0'\rangle\hat{t}_n = 2t_n'\Re\langle W_n, D_0\rangle + t_n'\Re\langle D_0, D_0'\rangle t_n + o_{P_n}(1).$$

Let $\tilde{t}_n = \hat{t}_n - t_n$ to get

$$2\tilde{t}_n'\Re\langle W_n, D_0\rangle + \tilde{t}_n'\Re\langle D_0, D_0'\rangle\tilde{t}_n + 2\tilde{t}_n'\Re\langle D_0, D_0'\rangle t_n = o_{P_n}(1).$$

Substitute $t_n = -(\Re\langle D_0, D_0'\rangle)^{-1}\Re\langle W_n, D_0\rangle$,

$$\tilde{t}_n'\Re\langle D_0, D_0'\rangle\tilde{t}_n = o_{P_n}(1).$$

By the non-singularity of D_0, this gives $\hat{t}_n = t_n + o_{P_n}(1)$ which completes the proof.

Proof for Theorem 5. We only give the main idea of the proof (details are given in Liu (1994)). We check the three conditions for Theorem 3. Strong identifiability is a consequence of Example 1 and the fact that $\mathcal{A}$ is compact. Conditions 2, 7 and 8 guarantee the differentiability, by making the exchange of integration and differentiation valid. So the asymptotic derivatives are given, for the unconstrained parameters, by

$$\int h(x) \exp\{it'x + iu'f(x, a_j)\}dx,$$

and

$$ip_j u' \int h(x) \frac{\partial f}{\partial a}(x, a_j) \exp\{it'x + iu'f(x, a_j)\}dx, \quad 1 \le j \le M,$$

where $h(x)$ is the density of P_X. It follows from the discussion preceding Theorem 5 that, for nonsingulaity, we need only show these functions to be linearly independent. Suppose they are not linearly independent as functions of t and u; then it can be argued by Fejér's theorem that

$$h(x) \exp\{iu'f(x, a_j)\}, \quad 1 \le j \le M,$$

$$u'h(x) \frac{\partial f}{\partial a}(x, a_j) \exp\{iu'f(x, a_j)\}, \quad 1 \le j \le M,$$

are linearly dependent in x and u. Here we have left out the non-zero factor ip_j. We can further leave out $h(x)$. Therefore there exist constants $c_1, \ldots, c_M$ and constant vectors $\alpha_1, \ldots, \alpha_M$ such that

$$\sum_{k=1}^{M} c_k \exp\{iu'f(x, a_k)\} = \sum_{k=1}^{M} \exp\{iu'f(x, a_k)\}\, u' \Big[\frac{\partial f}{\partial a_k}\Big] \alpha_k,$$

or equivalently,

$$\sum_{k=1}^{M} c_k \exp\{i\lambda u_0' f(x, a_k)\} = \lambda \sum_{k=1}^{M} \exp\{i\lambda u_0' f(x, a_k)\}\, u_0' \Big[\frac{\partial f}{\partial a_k}\Big] \alpha_k,$$

for any x, λ and $|u_0| = 1$. This implies, when $\lambda \to \infty$, $\sum_{k=1}^{M} \exp\{i\lambda u_0' f(x, a_k)\}$ $u_0'[\partial f/\partial a_k]\alpha_k \to 0$. For any fixed x and u_0, this function is a sum of periodic functions; hence it has to be zero in order to vanish at infinity. So

$$\sum_{k=1}^{M} \exp\{i\lambda u_0' f(x, a_k)\}\, u_0' \Big[\frac{\partial f}{\partial a_k}\Big] \alpha_k = 0, \tag{5}$$

and $\sum_{j=1}^{M} c_j \exp\{i\lambda u_0' f(x, a_j)\} = 0$, for any x, λ and $|u_0| = 1$.

Differentiate (5) l times w.r.t. λ:

$$\sum_{k=1}^{M} \exp\{i\lambda u_0' f(x, a_k)\} (iu_0' f(x, a_k))^l u_0' \Big[\frac{\partial f}{\partial a_k}\Big] \alpha_k = 0, \quad 0 \le l \le s-1. \tag{6}$$

Let $b_j = iu_0' f(x, a_j)$, $d_k(x, \lambda, u_0) = \exp\{i\lambda u_0' f(x, a_k)\} u_0' [\partial f / \partial a_k] \alpha_k$. Define the $s \times s$ matrix B by $b_{jk} = b_k^{j-1}$ and $s \times 1$ vector $d(x, \lambda, u_0) = (d_1, \ldots, d_s)'$, then (6) can be written as $Bd = 0$. It is a well known result of linear algebra that if $b_1, \ldots, b_M$ are distinct, B is invertible and $d = 0$. From Condition 3, $b_j(x, u_0)$, $j = 1, \ldots, M$ are distinct for almost every x and u_0, so $d(x, \lambda, u_0) = 0$ for almost every x and u_0. By continuity, $d(x, \lambda, u_0) = 0$ for any x, λ and u_0. Therefore $[\partial f / \partial a_k] \alpha_k = 0$ for any $1 \le k \le s$, any x, which again, by Condition 3, implies $\alpha_k = 0$, any k.

A similar argument shows that $c_j = 0$, $j = 1, \ldots, M$. Therefore the components of D are linearly independent.

The only thing left to show is the covariance function of W and this is not difficult to check.

Proof of Example 4. We check the conditions for Theorem 5. Conditions 1, 2, 5, 6 and 7 clearly hold. Condition 3 holds because $f(x, a)$ is analytic in x. Now

$$\frac{\partial f}{\partial a} = (-b_1 x e^{-a_1 x}, e^{-a_1 x}, \ldots, -b_r x e^{-a_r x}, e^{-a_r x}, 1), \tag{7}$$

$$\Big(\frac{\partial^2 f}{\partial a^2}\Big) = \operatorname{diag}\Big(\begin{pmatrix} b_1 x^2 e^{-a_1 x} & -x e^{-a_1 x} \\ -x e^{-a_1 x} & 0 \end{pmatrix}, \ldots, \begin{pmatrix} b_r x^2 e^{-a_r x} & -x e^{-a_r x} \\ -x e^{-a_1 x} & 0 \end{pmatrix}, 0\Big), \tag{8}$$

where $\operatorname{diag}(A_1, \ldots, A_k)$ denotes the quasi-diagonal matrix with diagonal blocks $A_1, \ldots, A_k$. The linear independence for $(\partial f / \partial a)$ as functions of x is obvious. Condition 8 holds because all the elements in the derivatives are bounded by constants not depending on x.

Proof of Example 5. The proof is similar to that of Example 2, using the result of Beran and Millar (1994). To prove $E(Y|X = x)$ for $x \in \text{support}(X)$ uniquely determines the parameter in the distributions of A_j, $j = 1, \ldots, r$, we use some elementary calculus as in proof of Theorem 5.

Theorem 6 and Example 6 are self evident and the proofs are not given in this paper.

Acknowledgement

This research was partly supported by NSF Grant DMS 9224868. The author wants to thank Professor Rudolf Beran for his advice and Professors Warwick

Millar and Deborah Nolan for helpful discussions. The author is also grateful to Blake Witten and the referees for carefully reading early drafts and correcting many errors.

References

Amemiya, Y. (1985). Instrument variable estimator for the nonlinear errors-in-variables model. *J. Econom.* **28**, 273–289.

Amemiya, Y. (1990). Two-stage instrumental variable estimators for the nonlinear errors-in-variables model. *J. Econom.* **44**, 311-332.

Amemiya Y. and Fuller, W. A. (1988). Estimation for the nonlinear functional relationship. *Ann. Statist.* **16**, 147-160.

Beal, S. L. (1984). Population pharmacokinetics data and parameter estimation based on their first two statistical moments. *Drug Metabolism Reviews* **15**, 173-193.

Beal, S. and Sheiner, L. (1985). Methodology of population pharmacokinetics. In *Drug Fate and Metabolizm* (Edited by E. R. Garrett and J. L. Hirtz), **5**, 135-183. M. Dekker, New York.

Beal, S. L. and Sheiner, L. B. (1988). Heteroscedastic nonlinear regression. *Technometrics* **30**, 327-338.

Beal, S. L. and Sheiner, L. B. (1989). *NONMEM Users Guides.* NONMEM Project Group, UCSF, San Francisco, 1989.

Beran, R. and Hall, P. (1992). Estimating coefficient distributions in random coefficient regressions. *Ann. Statist.* **20**, 1970-1984.

Beran, R. (1993). Semi-parametric random coefficient regression models. *Ann. Inst. Statist. Math.* **45**, 639-654.

Beran, R. and Millar, W. (1994). Minimum distance estimation in random coefficient regression models. *Ann. Statist.* **22**, 1976-1992.

Bolthausen, E. (1977). Convergence in distribution of minimum distance estimators. *Metrika* **24**, 215-227.

Breiman, L. (1968). *Probability.* Addison-Wiley.

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

Dempster, A. et al. (1984). Statistical and computational aspects of mixed model analysis. *Appl. Statist.* **33**, 203-214.

Donoho, D. and Liu, R. (1988a). The automatic robustness of minimum distance functionals. *Ann. Statist.* **16**, 552-586.

Donoho, D. and Liu, R. (1988b). Pathologies of some minimum distance estimators. *Ann. Statist.* **16**, 587-608.

Durbin, W. E. (1954). Errors in variables. *Internat. Statist. Rev.* **22**, 23-32.

Fuller, W. A. (1987). *Measurement Error Models.* John Wiley, New York.

Gibaldi, M. and Perrier, D. (1982). *Pharmacokinetics*, 2nd edition. M. Dekker, New York.

Grasela, T. H. and Sheiner, L. B. (1991a). An introduction to mixed effect modeling – concepts, definitions, and justification. *Journal of Pharmacokinetics and Biopharmaceutics* **19**, S11-S24.

Grasela T. H. and Sheiner, L. B. (1991b). Pharmacostatistical Modeling for observational data. *Journal of Pharmacokinetics and Biopharmaceutics* **19**, S25-S36.

Kac, M., Kiefer, J. and Wolfowitz, J. (1955). On tests of normality and other tests of goodness of fit based on distance methods. *Ann. Math. Statist.* **26**, 189-211.

Laird, N. (1991). Topics in likelihood-based methods for longitudinal data analysis. *Statist. Sinica* **1**, 33-50.

Lange, N. and Laird, N. (1989). The effect of covariance structure on variance estimation in balanced growth-curve models with random parameters. *J. Amer. Statist. Assoc.* **84**, 241-247.

Lindstrom, M. (1984). Estimation of population pharmacokinetics parameters using destructively obtained experimental data: A simulation study of the one-compartment open model. *Drug Metabolism and Reviews* **15**, 195-264.

Lindstrom, M. and Bates, D. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics* **46**, 673-687.

Lipsitz, S., Laird, N. and Harrington, D. (1990). Using the Jackknife to estimate the variance of regression estimators from repeated measures studies. *Comm. Statist.* **19**, 821-845.

Liu, J. (1994). *Minimum Distance Procedures in Nonlinear Random Coefficient Models.* PhD dissertation, University of California at Berkeley.

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

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

Sheiner, L. B. and Beal, S. L. (1980). Evaluation of methods for estimating population pharmacokinetic parameters. i. Michaelis-Menten model: Clinical pharmacokinetic data. *Journal of Pharmacokinetics and Biopharmaceutics* **8,** 553-571.

Sheiner, L. B. and Beal, S. L. (1981). Evaluation of methods for estimating population pharmacokinetic parameters. ii. Biexponential model and experimental pharmacokinetic data. *Journal of Pharmacokinetics and Biopharmaceutics* **9**, 635-651.

Sheiner, L. and Beal, S. (1983). Evaluation of methods for estimating population pharmacokinetic parameters. iii. Monoexponential model: routine clinical pharmacokinetic data. *Journal of Pharmacokinetics and Biopharmaceutics* **11**, 303-319.

Sheiner, L. and Beal, S. (1987). A note on confidence intervals with extended least squares parameter estimates. *Journal of Pharmacokinetics and Biopharmaceutics* **15**, 93-98.

Sheiner, L. and Ludden, T. (1992). Methodology of population pharmacikinetics/pharmacodynamics. *Annual Review of Pharmacological Toxicology* **32**, 185-209.

Stiratelli, R., Laird, N. and Ware, J. (1984). Random-effects models for serial observations with binary response. *Biometrics* **40**, 961-971.

Wald, A. (1940). Fitting of straight lines if both variables are subject to error. *Ann. Math. Statist.* **11**, 284-300.

Wolfowitz, J. (1953). Estimation by the minimum distance method. *Ann. Inst. Statist. Math.* **5**, 9-23.

Wolfowitz, J. (1957). The minimum distance method. *Ann. Math. Statist.* **28**, 75-88.

Biostatistics Department, Ciba Pharmaceuticals, 556 Morris Ave., Summit, NJ 07901, U.S.A.

(Received October 1994; accepted November 1995)