UNIVERSITY OF WISCONSIN-MADISON

SSRI

Workshop Series

ESTIMATION BY THE
ANALOGY PRINCIPLE

Charles F. Manski

8635

SOCIAL SYSTEMS RESEARCH INSTITUTE

Social Systems Research Institute

University of Wisconsin

ESTIMATION BY THE ANALOGY PRINCIPLE

Charles F. Manski

8635

December 1986

This work has been supported under National Science Foundation Grant SES86-05436 to the University of Wisconsin-Madison. I have benefitted from discussions with Gary Chamberlain, Chris Flinn, Art Goldberger, Ariel Pakes, Jim Powell, and Scott Thompson.

ABSTRACT

The analogy principle proposes that population parameters be estimated by sample statistics which make known properties of the population hold as closely as possible in the sample. Applications of the analogy principle are ubiquitous. Nevertheless, estimation theory has not been studied from a consistent analog perspective. This paper makes a start.

TABLE OF CONTENTS

1. INTRODUCTION

ESTIMATION PROBLEMS AND METHODS: Many estimation problems have the following elements. One wants to learn some property of a population probability measure. It is known that the population has certain other properties. A sample of observations drawn at random from the population is available. The problem is to use the known properties of the population and the sample evidence to learn the property of interest.

Once such an estimation problem is specified, consideration of estimation methods becomes possible. The 'analogy principle' offers a means for generating estimators. The analogy principle is instantly recognized. Many authors routinely refer to sample statistics as the 'sample analog' of corresponding population parameters. Nevertheless, the analogy principle is rarely stated explicitly. The essential idea is expressed succinctly in the following quote:

> "the analogy principle of estimation...proposes that population parameters be estimated by sample statistics which have the same property in the sample as the parameters do in the population" (Goldberger,1968,p.4)

This statement needs to be augmented only in that it presumes the existence of a sample statistic having the same property in the sample as the parameters do in the population. More generally, an analog estimate is one chosen so that, in some well-defined sense, the known properties of the population hold as closely as possible in the sample.

SOME APPLICATIONS: Applications of the analogy principle are ubiquitous. Perhaps the oldest are the use of the sample average and median as estimates for the population mean and median. The method of moments (K. Pearson, 1894) applies the analogy principle, as does minimum chi square estimation(Neyman,1949). Maximum likelihood, least squares, and least absolute deviations estimation are analog methods. Econometric contribution to the theory of analog estimation dates back to the development of instrumental variables estimation(Wright,1928; Reiersol, 1941,1945).

Among modern developments, Von Mises(1947) introduced the notion of differentiable statistical functions and studied the local asymptotic behavior of their analog estimates. Wolfowitz(1953,1957) proposed minimum distance estimation, a very general application of the analogy principle to the problem of estimating distribution functions. Most of the literature on robust estimation(Huber,1981) presumes analog estimation. For example, M-estimates(Huber,1967) are analog methods. The term 'Fisher-consistency'(Rao,1973,p.345) refers to analog estimation of a parameter that is a smooth functional of the population distribution. In the recent econometric literature, Burguete, Gallant, and Souza(1982), Hansen(1982), and Manski(1983) have independently proposed analog estimation of a fairly general class of econometric models defined by smooth moment restrictions. These methods subsume the earlier instrumental variables work.

ESTIMATION FROM THE ANALOG PERSPECTIVE: The myriad applications of the analogy principle demonstrate its usefulness as a tool for generating estimators. Consideration of specific applications, however, may not convey the more general value of the analogy principle as a paradigm for

the study of estimation.

I have found that the analogy principle offers an effective framework for teaching estimation. In analog estimation, one begins by asking what he knows about the population. One then treats the sample as if it were the population. Finally, one selects an estimate that makes the known properties of the population hold as closely as possible in the sample. What could be more intuitive?

I have found that the analogy principle disciplines the researcher by encouraging him to focus attention on estimation problems rather than on methods. Much of the statistical literature begins with a method and looks for problems to which it can be applied. It seems more sensible to begin by specifying what it is the researcher wants to learn and then seek applicable methods. The analogy principle forces this mode of thought. Analog estimation follows rather than precedes specification of the estimation problem of interest.

I have, moreover, come to feel that the analogy principle has a certain elegance. Esthetic appeal may not suffice to make a subject worthy of study. It does help though.

PLAN OF THIS PAPER: It is surprising that estimation has not been studied from a consistent analog perspective. This paper makes a start. A monograph under preparation(Manski,1987) will provide further depth and breadth.

Section 2 develops a formal framework and sets out concepts. We pose an abstract estimation problem and define identification. We show that a given estimation problem generally has many alternative representations to which the analogy principle may be applied. Except in special cases, the derived analog estimate depends on the chosen representation. We

discuss informally the consistency and efficiency of analog estimates.

The remainder of the paper gives applications. Section 3 presents short case studies of five leading classes of estimation problems. These are the finite dimensional moment problems, nonparametric density problems, smooth statistical functions, index problems, and separable econometric models. In each case, we define the estimation problem, develop alternative representations of that problem, and obtain analog estimates by applying the analogy principle to these representations.

Sections 4 and 5 study analog estimation of regressions. Section 4 introduces an abstract class of regression problems, examines the method of moments approach to the estimation of moment regressions, and discusses method of moment estimation of best conditional predictors. Section 5 develops analog methods for the estimation of general regression functions. Here, we introduce the smallest neighborhood method for nonparametric estimation of regressions.

DISCLAIMERS: Writing this paper, I have had to struggle to achieve an appropriate balance of abstraction and concreteness, of formal analysis and heuristics. To make the task manageable, I have decided to forego treatment of some rather important topics.

First, we shall consider no sampling process other than random sampling from a fixed probability measure. As an idealization, random sampling is central to statistics, much as competitive behavior is to economics. It seems essential to understand the analogy principle in the random sampling setting before giving attention to more complex sampling processes.

Second, we maintain the assumption that all prior information is exact and correct. Thus, we do not consider application of the analogy

principle to misspecified models. Nor do we consider probabilistic prior information, as in Bayesian analysis.

Third, we do not investigate the researcher's choice among estimation problems. To apply the analogy principle, the researcher must specify what he knows and what he wants to learn. These logically necessary requirements are undeniably burdensome in practice. We often have difficulty eliciting our information sets and making our objectives explicit. Nevertheless, we shall assume that a coherent estimation problem has specified and that the researcher is now prepared to proceed with estimation.

My forthcoming monograph does give attention to these topics, particularly to the third.

2. THE ANALOGY PRINCIPLE

2.1. THE ESTIMATION PROBLEM

MAINTAINED ASSUMPTIONS: Throughout this paper, the sample space Z is a measurable subset of a finite dimensional real space, endowed with the Borel σ-algebra. The population probability measure P is known to be a member of Π, a specified space of probability measures on Z. An observable random variable distributed P is denoted z. The parameter space B is a metric space. Additional structure is assumed as needed.

Let $T(*,*)$ be a given function mapping $\Pi \times B$ into Υ, where Υ is a vector space. We are concerned with estimation problems of the following type. It is known that some $b \in B$ solves the equation

$$T(P,b) = 0. \tag{2.1}$$

A random sample of N realizations from P, that is N observations of z, is drawn. The problem is to combine the sample data with the knowledge that $b \in B$, $P \in \Pi$, and $T(P,b)=0$ so as to estimate b.

We shall maintain the assumption that the estimation problem is properly specified. The data really are a random sample from P, the space Π contains P, and there exists a $b \in B$ solving equation (2.1). Given that the specification is proper, the spaces Π and B can be restricted to feasible probability measures and parameter values. That is, $Q \in \Pi$ implies that $T(Q,a)=0$ for some $a \in B$ and $c \in B$ implies that $T(Q,c)=0$ for some $Q \in \Pi$.

IDENTIFICATION: In analyzing a specified estimation problem, one should first ask whether the parameter b could be learned if P were known. After all, knowledge of P makes sample data superfluous.

The parameter b is said to be identified relative to $(P,\mathcal{B})$ if $T(P,*)$ has a unique zero in $\mathcal{B}$. We say that b is uniformly identified relative to $(\Pi,\mathcal{B})$ if for every $Q\in\Pi$, $T(Q,*)$ has a unique zero in $\mathcal{B}$. In practice, we can be sure that b is identified only if it is uniformly identified. The reason, of course, is that we do not know P and sample data cannot reveal P with certainty.

If a parameter is uniformly identified with respect to $(\Pi,\mathcal{B})$, then there exists a function $t:\Pi\to\mathcal{B}$ such that for all $(Q,a)\in\Pi\times\mathcal{B}$,

$$T(Q,a) = 0 \iff a = t(Q). \tag{2.2}$$

In particular, $b = t(P)$. It is possible to think of $t(*)$ as defining the parameter of interest as a function of the population probability measure.

EXTENSION OF THE DOMAIN OF $T(*,*)$ TO THE SPACE OF EMPIRICAL MEASURES: Let P_N be the empirical measure in a sample of size N. The analogy principle suggests that to estimate b, one should substitute P_N for P in $T(P,*)$ and isolate the subset of $\mathcal{B}$ on which $T(P_N,*)$ is as close as possible to zero, in some sense. But the domain of $T(*,*)$ is the space $\Pi\times\mathcal{B}$. If $P_N\notin\Pi$, then $T(P_N,*)$ is not defined.

To apply the analogy principle in estimation problems where the empirical measure is not a feasible population measure, we must extend the domain of $T(*,*)$ so that $T(P_N,*)$ is defined. The smallest space

certain to contain P_N for all finite N is Φ, the space of probability measures on Z having finite support. So henceforth, T(*,*) is assumed defined on $(\Pi \cup \Phi) \times B$.

Of course, the extension of T(*,*) to $\Phi \times B$ is not unique. Application of the analogy principle requires that a definition of T(*,*) on $\Phi \times B$ be chosen. In some contexts, there is a natural way to define T(*,*) on its enlarged domain. For example, if T(P,a) has the form $H[\int g(z,a)dP]$ for some measurable $g:Z \times B \to R^1$ and $H:R^1 \to T$, then it is natural to define $T(P_N,a)$ to be $H[\int g(z,a)dP_N]$. In other settings, it may not be obvious how the extension should be accomplished. For example, how should one define $T(P_N,a)$ when Π is a space of measures having densities with respect to Lebesgue measure and T(P,a) depends on P through its density.

One general method for extending T(*,*) to $\Phi \times B$ is to replace T(*,*) with $T[\pi(*),*]$, where $\pi:\Pi \cup \Phi \to \Pi$ is such that $Q \in \Pi \Rightarrow \pi(Q)=Q$. Consider $T[\pi(*),*]$ as a function on $(\Pi \cup \Phi) \times B$. For $Q \in \Pi$, $T[\pi(Q),*]=T(Q,*)$. For $Q \in \Phi$, $T[\pi(Q),*]$ takes some value in T; hence, $T[\pi(P_N),*]$ is well-defined. Thus, replacement of T(*,*) by $T[\pi(*),*]$ leaves the estimation problem unchanged and makes analog estimation possible. We shall use this method later to derive nonparametric analog estimates of density and regression functions.

2.2. ALTERNATIVE REPRESENTATIONS OF THE ESTIMATION PROBLEM

Given Π and B, one can construct alternative representations of the knowledge that b solves (2.1). Let T' be any vector space. Let $T'(*,*):(\Pi \cup \Phi) \times B \to T'$ be such that for all $(Q,a) \in \Pi \times B$,

(2.3) $T(Q,a) = 0 \iff T'(Q,a) = 0.$

It follows that whatever the true measure $P \in \Pi$ may be,

(2.4) $T(P,b) = 0 \iff T'(P,b) = 0.$

Hence, given Π and B, one can replace (2.1) by the right hand side of (2.4) and leave the estimation problem unchanged.

When $P_N \in \Pi$, the choice of representation of an estimation problem has no consequences for analog estimation. Here, there exists a non-empty subset of B on which $T(P_N,*)$ equals zero. By (2.3), $T'(P_N,*)$ equals zero on the same subset. Hence, the analog estimate is invariant with respect to representation of the estimation problem.

When $P_N \notin \Pi$, selection of a representation is critical to application of the analogy principle. Depending on the chosen $T'(*,*)$, there may or may not exist a subset of B on which $T'(P_N,*)$ equals zero. If $T'(P_N,*)$ has no zero on B, the analog estimate of b generally depends both on $T'(*,*)$ and on the sense in which one makes $T'(P_N,*)$ close to zero.

It does not seem possible to characterize exhaustively the set of all representations of a given estimation problem. It will suffice to call attention to three classes of representations which are commonly used in the development of analog estimates.

THE STATISTICAL FUNCTION REPRESENTATION: We noted earlier that if b is uniformly identified with respect to (Π,B), then there exists a $t:\Pi \to B$ such that equation (2.2) holds. Thus, $b-t(P)=0$ is an alternative representation of (2.1).

Thus far, $t(*)$ has been defined only on the domain Π. Now extend the

domain of $t(*)$ to Φ. Then application of the analogy principle yields $t(P_N)$ as an estimate of b. Following convention, we shall refer to $t(*)$ defined on $\Pi \cup \Phi$ as a 'statistical function'.

REPRESENTATIONS FORMED FROM ORIGIN-PRESERVING TRANSFORMATIONS OF T:
A class of representations can be obtained by taking 'origin-preserving transformations' of the space T. Let Ω be an arbitrary space and T' be a vector space. Then $r:(T \times \Omega) \rightarrow T'$ is said to be an origin-preserving transformation of T if, for all $\omega \in \Omega$, $T=0 \Leftrightarrow r(T,\omega)=0$.

Let $r(*,*)$ be origin-preserving. Then for all $(Q,a,\omega) \in (\Pi \cup \Phi) \times B \times \Omega$,

$$(2.5) \quad T(Q,a) = 0 \iff r[T(Q,a),\omega] = 0.$$

Hence,

$$(2.6) \quad T(P,b) = 0 \iff r[T(P,b),\omega] = 0.$$

Among the class of origin-preserving transformations of T, those with range space $T'=[0,\infty)$ are particularly useful. Such transformations translate the statement that b zeroes $T(P,*)$, which takes values in the vector space T, into a statement that b minimizes a real-valued function $r[T(P,*),\omega]$. This translation is central to the construction of analog estimates when $T(P_N,*)$ has no zero on B. See Section 2.3.

REPRESENTATIONS FORMED FROM ALTERNATIVE DEFINITIONS OF $T(*,*)$ ON $\Phi \times B$:
The above representations can be varied by defining $T(*,*)$ on $\Phi \times B$ in alternative ways. In Section 2.1, we noted that the extension of $T(*,*)$ to $\Phi \times B$ is not unique. For example, one can replace $T(*,*)$ with

$T[\pi(*),*]$, where $\pi(Q)=Q$ for $Q\in\Pi$. This construction does not restrict the behavior of $\pi(*)$ on Φ. Clearly, the behavior of $T[\pi(P_N),*]$ on B depends critically on the chosen function $\pi(*)$.

2.3. ANALOG ESTIMATES

Assume now that one has chosen an extension of $T(*,*)$ to $\Phi\times B$. The analogy principle suggests that to estimate b, one should use

$$(2.7)\quad B_N \equiv [c\in B:\ T(P_N,c) = 0].$$

If B_N is non-empty, then it is the analog estimate. Moreover, B_N remains the analog estimate under any origin-preserving transformation of T. If $P_N\in\Pi$, then B_N is the analog estimate under any representation of the estimation problem.

If $P_N\notin\Pi$ and B_N is empty, then one must select the sense in which $T(P_N,*)$ is to be brought as close as possible to zero. The following approach describes essentially all of current practice.

First, one chooses an origin-preserving transformation of T whose range space is the non-negative half line. Thus, let $r:T\times\Omega\to[0,\infty)$ be such that $T=0 \Leftrightarrow r(T,\omega)=0,\ \forall\omega\in\Omega$. Next, one sets the auxiliary variable ω equal to some function of (P,b). Thus, let $\omega:(\Pi\cup\Phi)\times B\to\Omega$. Then one expresses the knowledge that $T(P,b) = 0$ by the condition

$$(2.8)\quad r[T(P,b),\omega(P,b)] = 0.$$

To estimate b, one minimizes on B the sample analog of $r[T(P,*),\omega(P,*)]$. Provided that $r[T(P_N,*),\omega(P_N,*)]$ attains its minimum on B, the analog

estimate is

$$(2.9) \qquad B_{Nr\omega} = \underset{a \in B}{\operatorname{argmin}}\, r[T(P_N,a),\omega(P_N,a)].$$

We write B_{Nrw} rather than B_N to reflect the fact that the estimate depends on the chosen representation of the estimation problem.

CARDINALITY OF THE ANALOG ESTIMATE: In general, the analog estimate may be set-valued rather than a point estimate. The statistical literature most commonly focusses on point estimates of parameters. A point analog estimate can be obtained by applying some auxiliary rule to select within the set estimate. We shall usually not do so. The analogy principle offers no reason to select one element for special attention.

One may ask whether the presence of a set-valued estimate implies a failure of identification. The answer depends on whether the empirical measure P_N is an element of the space Π within which the population measure P is known to lie. If $P_N \in \Pi$, then it is possible that the true population measure is P_N; hence a set-valued estimate implies that b is not uniformly identified. If $P_N \notin \Pi$, then P_N cannot be the true measure. So the analog estimate may be set-valued and b uniformly identified.

2.4. CONSISTENCY OF ANALOG ESTIMATES

Subject to identification and smoothness conditions, analog estimates are generally consistent. We shall take consistency to be the sine qua non of an estimation method. If an estimator is consistent, then in a basic sense it 'works'. A heuristic explanation of the consistency of

analog estimates follows.

As the sample size N grows, the empirical measure P_N converges in various senses to the population measure P. It follows that for large N, $T(P_N,*)$ and $\omega(P_N,*)$ behave on B much like $T(P,*)$ and $\omega(P,*)$, provided that $T(*,*)$ and $\omega(*,*)$ are suitably smooth. Moreover, $r[T(P_N,*),\omega(P_N,*)]$ tends to behave like $r[T(P,*),\omega(P,*)]$, provided that r maps $T\times\Omega$ smoothly into the non-negative half line. In particular, the minima on B of $r[T(P_N,*),\omega(P_N,*)]$ tend to occur near the minima of $r[T(P,*),\omega(P,*)]$. Given identification, $r[T(P,*),\omega(P,*)]$ is minimized at b alone. So for large N, the analog estimate tends to be close to b.

Rigorous demonstration of consistency requires that one specify the desired sense of convergence of the estimate to b and give content to all the above references to 'smoothness' of $T(*,*)$, $\omega(*,*)$, and $r(*)$. It would be too much to expect one theorem to cover all the applications of interest. To the contrary, the literature contains a multitude of consistency results of varying generality.

One would like to know whether there exist identified estimation problems for which there are no consistent analog estimators but there are consistent non-analog estimators. (By a non-analog estimator, I mean one that cannot be obtained by applying the analogy principle to some representation of the estimation problem). At one time, I thought that nonparametric density and regression estimation were such problems. In the conventional statement of the density problem, $T(P_N,*)$ is not defined as P_N is not absolutely continuous with respect to Lebesgue measure. In the regression case, consistent analog estimation would seem blocked by the fact that the empirical probability measure conditioned on a given event does not converge to the corresponding population conditional measure if the conditioning event has probability zero.

It turns out that with suitable representations of the estimation problems, consistent nonparametric analog estimates of densities and regressions can be obtained. See Sections 3.2 and 5.3. As it stands, I have no example of a problem where the analogy principle does not work yet consistent estimation is possible by other means.

2.5. EFFICIENCY OF ANALOG ESTIMATES

One would like to characterize the situations in which there exists an efficient analog estimate, efficiency having been defined in some suitable sense. For example, a cornerstone of classical statistics is the fact that the maximum likelihood method is asymptotically efficient for estimation of a population density known to be a member of a smooth finite dimensional family of densities. Section 3.1 cites recent results on the asymptotic efficiency of analog estimates of parameters solving smooth finite dimensional moment problems. But a general theory of efficiency is lacking. The following comments are speculative but may be helpful.

ESTIMATION OF b VS. ESTIMATION OF (P,b): Analog estimates disregard two kinds of information that may be relevant to estimation of b. First, they use the sample data only through the empirical measure, which does not preserve information about the sample size. (Analog estimates also disregard the ordering of the observations, but this information cannot be relevant under random sampling).

Second, analog estimates use the empirical measure only to replace P in the function $T(P,*)$. To the extent that Π restricts P in ways that

are not represented by the equation $T(P,b)=0$, analog estimates ignore this information.

It seems reasonable to think that the analogy principle does make efficient use of the available information whenever the realized empirical measure is in Π. Recall that if $P_N \in \Pi$, then $T(P_N,*)$ has a zero on B and prior knowledge does not exclude the possibility that $P=P_N$. Moreover, the analog estimate B_N is invariant with respect to the representation of the estimation problem. Thus, if $P_N \in \Pi$, the sample data and prior knowledge are fully compatible with the hypothesis $(P,b)=(P_N,B_N)$. Given this, it is difficult to imagine that one can do better than use (P_N,B_N) to estimate (P,b).

The efficiency of analog estimation seems a much more complex question in those cases where $P_N \notin \Pi$. Here, the analog estimate generally depends on one's representation of the estimation problem. By definition, all representations identify b. But representations may differ in the extent to which they fully express the restriction $P \in \Pi$. Hence, as is well known, alternative analog estimates may differ in their precisions. Moreover, it may be that one can improve on any estimator that uses P_N as an estimate for P.

It would appear that a full understanding of the efficiency of analog estimation can emerge only if the problem of estimating b is embedded within the larger problem of estimating the pair (P,b). One would first consider the question of optimal estimation of this pair. Then, treating b as the parameter of interest, one would seek to determine the circumstances in which using P_N to estimate P suffices to obtain an optimal estimate of b.

3. EXAMPLES

3.1. FINITE DIMENSIONAL MOMENT PROBLEMS

FINITE DIMENSIONAL MOMENT OPTIMIZATION PROBLEMS: Much of present day econometrics is concerned with estimation of the parameter b solving a finite dimensional moment optimization problem. Here, (2.1) has the form

$$(3.1) \quad T(P,b) \equiv b - \underset{a\in B}{\operatorname{argmin}} \int h(z,a)dP = 0,$$

where B is a subset of R^K and where h(*,*) is a known function mapping Z×B into R^1. The space Π is composed of probability measures with respect to which the functions h(*,a), a∈B are integrable.

If ∫h(z,*)dP has a unique minimum on B, the parameter b is identified. Contrariwise, if ∫h(z,*)dP has a set-valued minimum, b is not identified. In the latter case, we interpret equation (3.1) to mean that b is an element of the minimizing set.

Application of the analogy principle to (3.1) yields the estimate

$$(3.2) \quad B_N = \underset{a\in B}{\operatorname{argmin}} \int h(z,a)dP_N.$$

This estimate exists provided only that $\int h(z,*)dP_N$ attains its minimum on B. Given existence, B_N is invariant under origin-preserving transformations of the estimation problem.

FINITE DIMENSIONAL MOMENT EQUATIONS: Another important part of econometric work is concerned with estimation of the parameter b solving

a finite dimensional moment equation. Here, (2.1) has the form

$$(3.3) \quad T(P,b) \equiv \int g(z,b)dP = 0,$$

where B is a subset of R^K and where $g(*,*)$ is a known function mapping ZxB into R^J. So (3.2) is a system of J equations in K unknowns. The space Π is composed of probability measures with respect to which $g(*,a)$, $a \in B$ are integrable.

The analogy principle applied to (3.3) yields the estimate

$$(3.4) \quad B_N = [c \in B: \int g(z,c)dP_N = 0],$$

provided that B_N is non-empty. If B_N is empty, the analog estimate depends on the chosen origin-preserving transformation of T. A common choice is a quadratic form

$$(3.5) \quad r[\int g(z,*)dP] \equiv [\int g(z,*)dP]'\Delta[\int g(z,*)dP],$$

where $g(*,*)$ is written as a Jx1 vector and where Δ is a positive definite JxJ matrix picked by the analyst. The resulting analog estimate is

$$(3.6) \quad B_{N\Delta} = \underset{a \in B}{\operatorname{argmin}} [\int g(z,a)dP_N]'\Delta[\int g(z,a)dP_N].$$

Given regularity conditions, Hansen(1982) and Chamberlain(1986) have shown that in various first order asymptotic senses, the estimate (3.6) is the most precise possible, provided that Δ is set equal to the inverse of

$$(3.7)\quad \Sigma \equiv \int g(z,b)g(z,b)'dP.$$

The matrix Σ is not known so the ideal estimate is not computable. On the other hand, a familiar multi-step procedure yields a computable estimate that is asymptotically equivalent to the ideal(Hansen,1982). That is, one selects some positive definite JxJ matrix Δ_0, computes $B_{N\Delta_0}$ as in (3.6), and picks a point b_{N0} from $B_{N\Delta_0}$. Then one computes

$$(3.8)\quad \Sigma_N = \int g(z,b_{N0})g(z,b_{N0})'dP_N.$$

Finally, one re-estimates b by

$$(3.9)\quad B_{N1} \equiv \underset{a\in B}{\operatorname{argmin}} \left[\int g(z,a)dP_N\right]'\Sigma_N^{-1}\left[\int g(z,a)dP_N\right].$$

The derivation of B_{N1} applies the analogy principle recursively, first to obtain b_{N0}, then Σ_N, and finally B_{N1}. This recursion can be rewritten, albeit somewhat clumsily, as a single application of the analogy principle. In particular, B_{N1} minimizes on B the sample analog of the following origin-preserving transformation of T:

$$(3.10)\quad r\left[\int g(z,*)dP,\omega(P)\right] \equiv \left[\int g(z,*)dP\right]'\omega(P)^{-1}\left[\int g(z,*)dP\right],$$

where $\omega(P) \equiv \int [g(z,c(P)][g(z,c(P)]'dP$ and where

$$c(P) = \underset{a\in B}{\operatorname{argmin}} \left[\int g(z,a)dP\right]'\Delta_0\left[\int g(z,a)dP\right].$$

Simply observe that $c(P_N) = b_{N0}$ and $\omega(P_N) = \Sigma_N$.

Thus, b_{N1} is interpretable as an analog estimate. Similarly, feasible generalized least squares, one step approximations to maximum likelihood, and other multi-stage methods can be written as analog estimates.

3.2. NONPARAMETRIC DENSITY PROBLEMS

Next, we consider an estimation problem which seemingly defies treatment by the analogy principle. Let $Z=R^1$ and let Π be the space of all probability measures that are absolutely continuous with respect to Lebesgue measure, denoted μ. For $Q\in\Pi$, let $\phi_\mu(*,Q)$ denote the density of Q with respect to μ. Consider the problem of estimating the population density in the absence of restrictions on its form, that is nonparametrically. Thus, B is the space of all measurable, non-negative valued functions on the real line whose Lebesgue integral equals one. And equation (2.1) is

$$(3.11)\quad T(P,b) \equiv b - \phi_\mu(*,P) = 0.$$

As stated, this estimation problem is not amenable to application of the analogy principle. The empirical measure is not absolutely continuous with respect to Lebesgue measure. So $\phi_\mu(*,P_N)$ is not defined. Indeed, the available methods for nonparametric estimation of densities do not give the appearance of analog estimates. These methods all require that the analyst choose the value of some auxiliary parameter unrelated to P; for example, a smoothing parameter, a number of nearest neighbors, or a number of terms in an orthogonal expansion. Most of the literature supposes that the value chosen for the auxiliary parameter varies as a function of the sample size. Some analyses study the use of resampling methods, such as cross-validation, to determine this value. See Prakasa Rao(1983) for a survey of methods for density estimation.

Appearances notwithstanding, the analogy principle does apply to the problem of density estimation. In what follows, we use the analogy principle to derive the familiar kernel method.

The key is to reformulate the estimation problem so that $T(P_N,*)$ is well-defined. To do this, let Z, Π, and B remain as specified above. Let $\sigma(*)$ be any function mapping $\Pi \cup \Phi$ into $[0,\infty)$ such that $\sigma(Q)=0 \Leftrightarrow Q \in \Pi$. Let $G \in \Pi$ and let δ be a random variable distributed G, with δ independent of z. Let $P \oplus \sigma(P)G$ denote the probability measure of the random variable $z+\sigma(P)\delta$.

Now replace equation (3.11) with

$$(3.12) \quad T(P,b) \equiv b - \phi_\mu[*, P \oplus \sigma(P)G] = 0.$$

Given that $P \in \Pi$, $\sigma(P)=0$. So (3.11) and (3.12) both state that b is the density of P. On the other hand, these two statements of the estimation problem diverge with respect to the analogy principle. Whereas P_N is not absolutely continuous, $P_N \oplus \sigma(P_N)G$ is.

$\phi_\mu[*, P_N \oplus \sigma(P_N)G]$ has the form of a kernel estimate of $\phi_\mu(*,P)$. To see this, let z_N be a random variable distributed P_N, with z_N independent of δ. Let $H_N: R^1 \to [0,1]$ denote the distribution function of $z_N+\sigma(P_N)\delta$. Let g denote the density of G. Given that $P_N \notin \Pi$, $\sigma(P_N)>0$. Hence, for each $\zeta \in R^1$,

$$(3.13) \quad H_N(\zeta) \equiv \text{Prob}[z_N+\sigma(P_N)\delta \leq \zeta] = \int \int_{-\infty}^{(\zeta - z_N)/\sigma(P_N)} g(\delta)d\delta \, dP_N.$$

$H_N(*)$ is differentiable, the derivative at $\zeta' \in R^1$ being

$$(3.14) \quad h_N(\zeta') \equiv dH_N(\zeta')/d\zeta = 1/\sigma(P_N) \int g[(\zeta' - z_N)/\sigma(P_N)]dP_N.$$

Differentiability of $H_N(*)$ implies absolute continuity of $P_N \oplus \sigma(P_N)G$, with $\phi_\mu[*, P_N \oplus \sigma(P_N)G] = h_N(*)$. So $h_N(*)$ is the analog estimate of $\phi_\mu(*,P)$ obtained by applying the analogy principle to (3.12). $h_N(*)$ is also a kernel estimate of $\phi_\mu(*,P)$, the kernel being $g(*)$ and the smoothing parameter being $\sigma(P_N)$.

This derivation of the kernel method makes the smoothing parameter a functional on the space $\Pi \cup \Phi$. Thus far, we have required only that $\sigma(Q)=0 \Leftrightarrow Q \in \Pi$. How should $\sigma(*)$ behave on the space Φ of probability measures with finite support? Heuristically, we would like $\sigma(Q)$ to be closer to zero the less Q deviates from an absolutely continuous measure. Given that the elements of $\Pi \cup \Phi$ are measures with no singular continuous component, absolute continuity of an element of $\Pi \cup \Phi$ is equivalent to continuity. So we would like $\sigma(Q)$ to be closer to zero the less Q deviates from a continuous measure.

Perhaps the simplest reasonable index of a probability measure's deviation from continuity is the supremum of its point masses, denoted $M(Q)$. This suggests setting $\sigma(Q) = s[M(Q)]$, where $s:[0,1] \to [0,\infty)$ is strictly increasing and where $s(0)=0$. Continuity of P implies that $M(P_N)=1/N$ with probability one. Hence, $\sigma(P_N) = s(1/N)$ with probability one. This translates our specification of the smoothing parameter as a functional on the space of probability measures into the conventional specification as a decreasing function of sample size. So standard results on the consistency of kernel estimation(e.g. Prakasa Rao,1983, Section 2.1) imply that our analog version of kernel estimation is consistent if $s(*)$ satisfies the condition $s(M)/M \to \infty$ as $M \to 0$.

3.3. SMOOTH STATISTICAL FUNCTIONS

Recall that if b is uniformly identified with respect to (Π,B), then there exists a $t:\Pi\to B$ such that

$$(3.15)\quad b - t(P) = 0.$$

Extending the domain of $t(*)$ from Π to $\Pi\cup\Phi$ defines a statistical function $t:(\Pi\cup\Phi)\to B$ and an associated analog estimate

$$(3.16)\quad B_N = t(P_N).$$

Representation of an estimation problem in the form (3.15) and estimation of b by $t(P_N)$ is particularly appealling when $t(*)$ is a function that varies smoothly on $\Pi\cup\Phi$. Knowledge that $t(*)$ varies smoothly makes it easy to analyze the asymptotic behavior of $t(P_N)$. Moreover, smoothness brings with it the desirable property of 'robustness'.

SMOOTHNESS AND ASYMPTOTIC ANALYSIS: When $t(*)$ is appropriately smooth, characterization of the asymptotic behavior of $t(P_N)$ is almost trivial. Perhaps the most striking demonstration of this is proof of consistency by the well-known 'continuous mapping' theorem:

> Let λ be a metric on B. Let Π° denote the space of all probability measures on Z. Assume that there exists a metric ρ on Π° with respect to which (i) P_N converges to P almost surely(or in probability) and (ii) $t(*)$ is continuous at P. Then $t(P_N)$ converges to $t(P)$ with respect to λ almost surely(or in probability).

To apply the continuous mapping theorem, one draws on the literature on the convergence of empirical measures, which shows that P_N converges to P with respect to a variety of metrics. By the theorem, it suffices to find one such metric ρ such that t(*) is continuous with respect to ρ. For details and examples, see Manski(1987).

Whereas continuity of t(*) simplifies proof of consistency, differentiability eases derivation of limiting distributions. Assume that in an appropriate sense, the functional derivative dt(*)/dQ exists in a neighborhood of P. Then a Taylor's series expansion shows that for N large, $\sqrt{N}[t(P_N)-t(P)]$ behaves like $\{dt(P)/dQ\}\{\sqrt{N}(P_N-P)\}$. The limiting distribution of $\sqrt{N}(P_N-P)$ is a tied down Brownian motion process. Hence, derivation of the limiting distribution of $\sqrt{N}[t(P_N)-t(P)]$ reduces to the problem of characterizing dt(P)/dQ. See Serfling(1980), Chapter 6.

SMOOTHNESS AND ROBUSTNESS: The literature on robustness seeks to characterize the manner in which the solution to an estimation problem varies with small changes in the process generating the data. Formally, this amounts to the study of the behavior of t(*) in neighborhoods of P. A statistical function is said to be robust if it varies smoothly; the more smoothly the better. See Huber(1981).

The judgement that robustness(smoothness) is desirable is most often motivated by reference to a model of contaminated sampling. Here, one wishes to learn t(P) but P_N is not obtained by random sampling from P. Rather, it is obtained by random sampling from some measure Q. It is known that for a given metric ρ on Π^{Ω}, Q is near P. Other facts about Q may or may not be known.

In this setting, $t(P_N)$ will generally converge to t(Q), not t(P). So

$t(P_N)$ is not generally consistent for $t(P)$. Nevertheless, it is still desirable that the limit of $t(P_N)$ be close to $t(P)$. This will be the case if $t(*)$ is smooth at P, in the sense of ρ.

APPLICATIONS: Applications of the powerful theory developed for analog estimation of smooth statistical functions have focussed on a small set of problems admitting relatively simple expressions for $t(*)$. In particular, a vast literature on the estimation of location parameters has developed. Most of this has focussed on the setting in which $Z=R^1$, Π is the space of probability measures with symmetric distribution functions, and $t(P)$ is the center of symmetry. In this context, there are many interesting ways to define $t(*)$ on Φ. For example, one may select $t(*)$ to be the mean, median, or some trimmed mean. These functionals coincide on Π but not on Φ. Hence, they yield distinct analog estimates of the center of symmetry.

It appears rather difficult to apply smooth statistical function theory to typical econometric problems. To use this theory, one must be able to determine the smoothness characteristics of the functional $t(*)$. In econometric work, however, $t(*)$ is usually not a simple function on $\Pi \cup \Phi$. For example, in moment optimization problems, $t(*)$ is an argmin operator. In nonlinear moment equation problems, $t(*)$ can often only be defined implicitly. So determination of the senses in which $t(*)$ is and is not smooth is problematic.

3.4. INDEX PROBLEMS

The parameter space B indexes the family Π of probability measures if

(3.17) $T(P,b) \equiv P - \tau(b) = 0,$

where $\tau(*)$ is an invertible function mapping B onto Π. Thus, an index problem has the special characteristic that B and Π are one-to-one. In particular, $\Pi = [\tau(a), a \in B]$ and $B = [\tau^{-1}(Q), Q \in \Pi]$.

Two prominent approaches to the estimation of indices are the maximum likelihood method and minimum distance estimation. These are obtained by application of the analogy principle to alternative representations of the estimation problem.

THE MAXIMUM LIKELIHOOD METHOD: Assume that all the probability measures in Π are absolutely continuous with respect to a common measure ν on Z. For $Q \in \Pi$, let $\phi_\nu(*,Q)$ denote the density of Q with respect to ν. For $a \in B$, let $\phi_\nu(*,a)$ denote the density of $\tau(a)$. Then we may replace (3.17) with

(3.18) $\phi_\nu(*,P) - \phi_\nu(*,b) = 0.$

As is well-known, b solves (3.18) if and only if

(3.19) $b - \underset{a \in B}{\operatorname{argmax}} \int \log[\phi_\nu(z,a)]\,dP = 0.$

See, for example, Rao(1973), p.58. So the moment optimization problem (3.19) is an alternative representation of the index problem (3.17). Application of the analogy principle to (3.19) yields the maximum likelihood estimate

$$(3.20)\quad B_N = \underset{a\in B}{\operatorname{argmax}} \int \log[\phi_\nu(z,a)]\,dP_N.$$

MINIMUM DISTANCE ESTIMATION: Let $\rho(*,*)$ be a metric on the space Π^{Ω} of all probability measures on Z. Then an alternative representation of (3.17) is

$$(3.21)\quad b - \underset{a\in B}{\operatorname{argmin}}\ \rho[P,\tau(a)] = 0.$$

Applying the analogy principle to (3.21) yields

$$(3.22)\quad B_N = \underset{a\in B}{\operatorname{argmin}}\ \rho[P_N,\tau(a)],$$

the minimum distance estimate introduced by Wolfowitz(1953,1957).

Minimum distance estimation is a class of methods whose members are distinguished by the chosen metric ρ. Following the original work of Wolfowitz, it has been observed that the theme of minimum distance estimation does not require that ρ be a metric. In particular, equation (3.21) remains a valid representation of the index problem if ρ is any mapping from $(\Pi\cup\Phi)\times\Pi$ to $[0,\infty]$ such that $\rho(Q_1,Q_2)=0$ if and only if $Q_1=Q_2$. Analog estimates obtained under such general ρ are termed 'minimum discrepancy' estimates. See Sahler(1970).

One may also generalize the definition of ρ to let it depend on an auxiliary variable ω in the manner of the origin-preserving transformations discussed in Section 2. Let Ω be an arbitrary space. Let $\rho:(\Pi\cup\Phi)\times\Pi\times\Omega \to [0,\infty]$ be such that for all $\omega\in\Omega$, $\rho(Q_1,Q_2,\omega)=0 \Leftrightarrow Q_1=Q_2$. Now set $\omega=\omega(P,a)$, where $\omega(*,*)$ maps $(\Pi\cup\Phi)\times B$ into Ω. Then the index problem is representable as

$$(3.23)\quad b - \underset{a\in B}{\operatorname{argmin}}\ \rho[P,\tau(a),\omega(P,a)] = 0.$$

Applying the analogy principle to (3.23) yields

$$(3.24)\quad B_N = \underset{a\in B}{\operatorname{argmin}}\ \rho[P_N,\tau(a),\omega(P_N,a)].$$

Minimum chi-square estimation is a familiar example of (3.24). Let $Z=(1,\ldots,I)$ for I finite, let $B\subset R^K$, and let $[\tau(a),a\in B]$ be a family of multinomial distributions on Z. Let $\Omega\subset R^I$ and let $\omega_i(P,a)>0$ be the i^{th} component of $\omega(P,a)$. Let $\tau(a)(z=i)$ denote the probability under $\tau(a)$ that $z=i$. Let

$$(3.25)\quad \rho[P,\tau(a),\omega(P,a)] = \sum_{i=1}^{I} [P(z=i) - \tau(a)(z=i)]^2/\omega_i(P,a).$$

With this choice of ρ, (3.24) is a minimum chi-square estimate for b. Different versions of the minimum chi-square method are obtained by defining the weighting function $\omega(*,*)$ in alternative ways. See, for example, Rao(1973), p.352.

RELATIVE MERITS OF THE TWO APPROACHES: It is of interest to compare the maximum likelihood and minimum distance approaches to estimation. The former is favored for its asymptotic efficiency properties and for the relative simplicity of its moment optimization form. The latter has a broader domain of application; it does not require that the measures in Π be absolutely continuous with respect to any common measure.

Some recent literature has emphasized the superior robustness of certain minimum distance estimates. Maximum likelihood and minimum distance estimators apply the analogy principle to different statistical functions. In general, the functional $t(*) \equiv \operatorname{argmax}_{a\in B} \int \log[\phi_\nu(z,a)]d*$ is not continuous with respect to the usual(weak) topology on Π^∞. On the other hand, one can often select ρ so that $t(*) \equiv \operatorname{argmin}_{a\in B} \rho[*,\tau(a)]$ is continuous. See, for example, Parr and Schucany(1980).

3.5. SEPARABLE ECONOMETRIC MODELS

The reader will have observed that we have, thus far, made no mention of unobservable random variables. The basic equation $T(P,b)=0$ defining an estimation problem relates the probability measure $P \in \Pi$ of an observable random variable z to a parameter $b \in B$.

Econometric models, on the other hand, posit restrictions on an assumed probabilistic process generating realizations of a random pair (z,u), where realizations of u are not observable by the researcher. Suppose that one wishes to estimate an unknown feature of this process. Then one must derive from the available information a relationship connecting observables and the parameter of interest.

STATEMENT OF THE PROBLEM: Many econometric models can be formulated as follows. It is assumed that realizations of (z,u) are drawn by independent sampling from some probability measure P_{zu} on $Z \times U$, where U is the domain of u. It is known that P_{zu} is a member of a given space Ψ of probability measures on $Z \times U$.

Let $C \subset R^K$. Let $f(*,*,*)$ be a given function mapping $Z \times U \times C$ into R^J. It is known that for some $c \in C$, the random pair (z,u) satisfies

$$f(z,u,c) = 0. \tag{3.26}$$

That is, almost every realization (ζ,η) of (z,u) satisfies the equation $f(\zeta,\eta,c)=0$. The estimation problem is to combine sample data on z with the knowledge that $P_{zu} \in \Psi$, $c \in C$, and $f(z,u,c)=0$ so as to learn c.

To apply the analogy principle, we need to express this knowledge in a form that relates P, the probability measure of z, to the parameter c.

In what follows, we describe approaches that are applicable if f is separable in either u or z.

MODELS WITH f SEPARABLE IN u: Consider the class of econometric models in which

(3.27) $f(z,u,c) \equiv u_\circ(z,c) - u,$

where $u_\circ: Z \times C \to U$ is a given measurable function. Then (3.26) implies that

(3.28) $(z,u) = [z, u_\circ(z,c)].$

This expresses the unobservable random variable u as a function of the observable z and of the parameter c.

Recall that Π° denotes the space of all probability measures on Z. For $Q \in \Pi^\circ$ and $a \in C$, let $\psi(Q,a)$ denote the probability measure of $[z, u_\circ(z,a)]$ when z is distributed Q. Then $P_{zu} = \psi(P,c)$. Moreover, knowing that $P_{zu} \in \Psi$ is the same as knowing that c solves the equation

(3.29) $T(P,c) \equiv c - \underset{a \in C}{\operatorname{argmin}}\, r[\psi(P,a)] = 0,$

where r(*) is a function mapping the space $[\psi(Q,a), Q \in \Pi^\circ, a \in C]$ of measures on $Z \times U$ into $[0,\infty]$ and satisfying the condition $r[\psi(Q,a)] = 0 \Leftrightarrow \psi(Q,a) \in \Psi$. Application of the analogy principle to (3.29) yields

(3.30) $C_N = \underset{a \in C}{\operatorname{argmin}}\, r[\psi(P_N,a)],$

the 'closest empirical distribution' estimate studied in Manski(1983).

In the most familiar application of (3.30), Ψ is the space of measures on $Z \times U$ that satisfy a given finite dimensional moment equation.

Let it be known that for a given measurable function $g:Z\times U\to R^J$,

$$(3.31)\quad \int g(z,u)dP_{zu} = \int g(z,u)d\psi(P,c) = \int g[z,u_\circ(z,c)]dP = 0.$$

Select $r(*)$ to be a quadratic form function

$$(3.32)\quad r[\psi(Q,a)] = [\int g(z,u)d\psi(Q,a)]'\Delta[\int g(z,u)d\psi(Q,a)]$$
$$= [\int g\{z,u_\circ(z,a)\}dQ]'\Delta[\int g\{z,u_\circ(z,a)\}dQ],$$

where Δ is a chosen $J\times J$ positive definite matrix. Then

$$(3.34)\quad C_N = \underset{a\in C}{\operatorname{argmin}}\ [\int g[z,u_\circ(z,a)]dP_N]'\Delta[\int g[z,u_\circ(z,a)]dP_N]$$

is a moment equation estimate of the form seen in Section 3.1.

MODELS WITH f SEPARABLE IN z: Now consider the class of models in which

$$(3.35)\quad f(z,u,c) \equiv z - z_\circ[x(z),u,c],$$

where $x:Z\to X$ and $z_\circ:X\times U\times C\to Z$ are given measurable functions. Let P_{xu} be the probability measure of (x,u). Assume that the space Ψ within which P_{zu} is known to lie restricts P_{zu} only through P_{xu}. That is, for some given space Ψ_{xu} of probability measures on $X\times U$, $Q_{zu}\in\Psi \Leftrightarrow Q_{xu}\in\Psi_{xu}$. For $(Q_{xu},a)\in\Psi_{xu}\times C$, let $\tau(Q_{xu},a)$ denote the probability measure of $z_\circ[x,u,a]$ when (x,u) is distributed Q_{xu}. Then (3.26) implies that (P_{xu},c) solves

$$(3.36)\quad T[P,(P_{xu},c)] \equiv P - \tau(P_{xu},c) = 0.$$

Equation (3.36) defines an index problem of the kind discussed in

Section 3.4, with the parameter (P_{xu},c) indexing P.

In general, minimum distance estimation is applicable to the problem of estimating (P_{xu},c). If the measures $\tau(Q_{xu},a)$, $(Q_{xu},a)\in\Psi_{xu}\times C$ are absolutely continuous with respect to a common measure on Z, the maximum likelihood method is applicable. Observe, that these analog estimation procedures call for estimation of P_{xu} along with the parameter of interest c. In special cases, the estimation problem may decompose in a manner that makes it possible to estimate c without explicit consideration of P_{xu}.

4. REGRESSION PROBLEMS

The estimation of regressions is a central theme of econometrics. In common usage, the regression of y on x refers to the expected value of some measurable function $y:Z\to Y$ conditional on the realization of some other measurable function $x:Z\to X$, considered as a function on the space X where x lives. More generally, a regression of z on x is some property of the probability measure of z conditional on the realization of x(z), again considered as a function on X.

In this section and the next, we apply the analogy principle to the estimation of regression functions. A recurring theme is that alternative representations of a regression problem generate distinct analog estimation methods. The relative usefulness of these methods varies with the regression function under study, with the characteristics of the population measure P, and with the nature of the parameter space B.

Section 4.1 formally defines regression problems and discusses the identification of regression functions. Then Section 4.2 exposits a familiar and widely used analog estimation method, the method of moments. This approach to estimation has the very attractive feature that it avoids reference to conditional probability measures. It is applicable if the regression is defined by a collection of moment problems and the parameter space is sufficiently small. Section 4.3 describes the application of the method of moments to a leading moment regression problem, that of conditional prediction. Section 4.4 gives an example of a regression problem that is not a collection of moment problems.

Section 5 considers analog methods that explicitly cope with the fact

that regressions are properties of conditional measures. Section 5.1 briefly discusses 'naive' analog estimation of regression problems in which X is discrete. Section 5.2 applies the kernel density estimation method of Section 3.2 to regression problems in which P is absolutely continuous. Our main contribution is presented in Sections 5.3 and 5.4. There we introduce a novel representation of regression problems. Applying the analogy principle to this representation yields an analog estimation method that is applicable quite generally. This method is termed 'smallest neighborhood' estimation.

4.1. THE ESTIMATION PROBLEM

MAINTAINED ASSUMPTIONS: Henceforth, X is a measurable subset of a finite dimensional real space, $x:Z\to X$ is a measurable function, and P_x is the probability measure on X of the random variable x. The parameter space B is a metric space of functions mapping X into some space Θ. Thus, for each $a\in B$ and $\xi\in X$, $a(\xi)\in\Theta$.

For measurable $A\subset X$ and $Q\in\Pi\cup\Phi$, let $Q|A$ be the probability measure Q conditioned on the event $[x\in A]$. For $\xi\in X$ and $Q\in\Pi\cup\Phi$, let $Q|\xi$ be Q conditioned on the event $[x=\xi]$. Let $\Pi|X$ denote the collection of measures $[Q|\xi, \xi\in X, Q\in\Pi]$. Let $S(*,*)$ be a given function mapping $(\Pi|X\cup\Phi)\times\Theta$ into some vector space Γ. In regression problems, (2.1) has the form

$$(4.1)\quad T(P,b) \equiv [S(P|\xi,b(\xi)),\ \xi\in X] = 0.$$

That is, for each $\xi\in X$, $b(\xi)$ solves the equation $S[P|\xi,b(\xi)] = 0$.

IDENTIFICATION: One would like to say that the regression function b is identified relative to (P,B) if T(P,*) defined in (4.1) has a unique zero on B. This statement is unexceptional if the space X is discrete with $P_x(\xi)>0$ for all $\xi\in X$. On the other hand, if there exist an $X_\circ\subset X$ such that $P_x(X_\circ)=0$, then one must contend with the fact that knowledge of P does not distinguish the collection $[P|\xi,\xi\in X]$ from any other collection $[Q|\xi,\xi\in X]$ such that $Q|\xi=P|\xi$, $\xi\in X-X_\circ$. It follows that if b solves (4.1), then any other $c\in B$ such that $b(x)=c(x)$, a.e. P_x must also be said to solve (4.1). Hence, b cannot be identified relative to such c.

The literature copes in two ways with the inherent indeterminacy of a regression function whose domain contain sets of probability zero. One approach relies on the specification of the parameter space to exclude functions that differ only on sets of P_x-probability zero. Assume that for all distinct $a\in B$, $c\in B$ and for all $Q\in\Pi$, there exists some $X_1\subset X$, which may depend on (a,c,Q), such that $Q_x(X_1)>0$ and $a(\xi)\neq c(\xi)$, $\xi\in X_1$. Then the parameter space contains only functions that differ on some set of positive probability. Hence, identification of b may be defined in the traditional manner.

The other approach is used in problems where one does not have sufficient prior information to restrict the parameter space as above. Then it is conventional to weaken the definition of identification. In particular, b is said to be identified relative to (P,B) if, for $c\in B$, $[S(P|x,c(x))=0$, a.e. $P_x] \Rightarrow c(x)=b(x)$, a.e. P_x. See, for example, the Stone(1977) treatment of nonparametric regression.

AN ALTERNATIVE FORMULATION: It is worth noting that the indeterminacy of regression functions can be rephrased as a sampling problem rather than

as a failure of identification.

To do this, let us not take P as the primitive probability concept for the analysis of regression. Rather, let us begin by positing the existence of the collection of measures $(P|\xi, \xi \in X)$ and of the marginal measure P_x. Then it is unambiguous to say that b is identified relative to $[(P|\xi, \xi \in X), B]$ if the solution to (4.1) is unique.

In this formulation of the regression problem, P is defined as the mixture of $(P|\xi, \xi \in X)$ with respect to the mixing measure P_x. That is, for measurable $A \subset Z$, $P(A) \equiv \int P(A|x)dP_x$, where $P(A|*)$ is assumed a measurable function on X. The indeterminacy of regressions is now a consequence of the fact that we can sample only from the mixture P, not from each of the measures $(P|\xi, \xi \in X)$.

4.2. METHOD OF MOMENT ESTIMATION OF MOMENT REGRESSIONS

With few exceptions, the regression problems that have been studied to date are members of the subclass of moment regressions. These are problems in which b solves either a collection of moment optimization problems or a collection of moment equations. In the former case, $S(*,*)$ has the form

$$(4.2) \quad S[P|\xi, b(\xi)] \equiv b(\xi) - \underset{\theta \in \Theta}{\operatorname{argmin}} \int h(z,\theta)dP|\xi,$$

where $h: Z \times \Theta \to R^1$. In the latter,

$$(4.3) \quad S[P|\xi, b(\xi)] \equiv \int g[z, b(\xi)]dP|\xi,$$

where $g: Z \times \Theta \to R^J$. In both cases, $\Theta \subset R^K$ and B is some subset of the

Cartesian product space $(\times\Theta,\xi\in X)$.

REPRESENTATION OF OPTIMIZATION REGRESSIONS BY OPTIMIZATIONS IN (b,P): Regression problems of the moment optimization type have representations which avoid reference to the collection of conditional measures $P|\xi, \xi\in X$. Assume that $S(*,*)$ has the form (4.2). Let $w:X\to[0,\infty)$ be any measurable function such that $w(x)>0$ a.e. P_x. Then b solves (4.1) a.e. P_x if and only if b also solves the problem

$$(4.4)\quad b - \underset{a\in B}{\operatorname{argmin}} \int w(x)\left[\int h(z,a(x))dP|x\right]dP_x = 0$$

$$b - \underset{a\in B}{\operatorname{argmin}} \int w(x(z))h[z,a(x(z))]dP = 0.$$

Thus, a collection of moment optimization problems, each relating $b(\xi)$ to $P|\xi$ for a given $\xi\in X$, can be represented by a single such problem relating b to P. Note that if B is a finite dimensional space of functions, this representation is a finite dimensional moment optimization problem of the type described in Section 3.1.

Application of the analogy principle to (4.4) yields the estimate

$$(4.5)\quad B_N = \underset{a\in B}{\operatorname{argmin}} \int w(x(z))h[z,a(x(z))]dP_N.$$

The attractivenss of this method of moments estimate depends critically on the parameter space. If B is finite dimensional, the discussion of Section 3.1 implies that given regularity conditions, estimates of the form B_N are very appealling. Indeed, they dominate practice.

On the other hand, method of moments estimation of moment regressions breaks down if the parameter space is too large. This is most easily seen in the extreme case in which B is unrestricted; that is $B=(\times\Theta,\xi\in X)$. Let P_{Nx} be the empirical measure of x. Let X_N denote the support of

P_{Nx}. For each $\xi \in X_N$, let $B_N(\xi)$ be the subset of Θ on which $\int h(z,*)dP_N|\xi$ is minimized. For each $\xi \in X - X_N$, let $B_N(\xi)=\Theta$. Then the analog estimate B_N is the set of functions $(\times B_N(\xi), \xi \in X)$. Except in the special case where $P_x(\xi)>0$, $B_N(\xi)$ does not generally converge to $b(\xi)$. Hence, B_N does not generally converge to b.

It should be understood that when B is unrestricted, the failure of the method of moments to be consistent does not derive from an absence of identification; the unconditional moment problem (4.4) inherits the identification properties of the original regression problem (4.1). It is rather that as $N \to \infty$, $\int w\{x(z)\}h[z,*\{x(z)\}dP_N$ does not converge as a function on B to $\int w\{x(z)\}h[z,*\{x(z)\}]dP$.

At present, we can offer no general characterization of the behavior of method of moments estimation in regression problems where B is not finite dimensional but is a proper subset of $(\times\Theta, \xi \in X)$. Results have, however, been obtained in specific settings. We shall cite some findings in Section 4.3, where we consider the class of conditional prediction problems.

REPRESENTATION OF EQUATION REGRESSIONS BY EQUATIONS IN (b,P): The foregoing discussion of moment optimization regressions applies to some, but not all, moment equation regressions. Assume that $S(*,*)$ has the form (4.3) and let $v: X \to R^{J \times J}$ be any measurable function such that $v(x)$, written as a $J \times J$ matrix, is non-singular a.e. P_x. If b solves (4.1) a.e. P_x, then b also solves the problem

$$(4.6) \quad \int v(x)[\int g\{z,b(x)\}dP|x]dP_x = \int v\{x(z)\}g[z,b\{x(z)\}]dP = 0.$$

So a collection of moment equations, each relating $b(\xi)$ to $P|\xi$ for a

given $\xi \in X$, implies a moment equation relating b to P.

Solutions to (4.6), however, do not necessarily solve the regression problem defined by (4.3). That is, there may exist $a \in B$, $a \neq b$ such that $\int v\{x(z)\}g[z,a\{x(z)\}]dP = 0$ even though $\int g[z,a(x)]dP|x \neq 0$ on a set of positive P_x-measure. If so, then solution of (4.6) does not identify b. Application of the analogy principle to (4.6) can be an attractive estimation method, but only if, for the given specification of (g,B,Π) and chosen $v(*)$, one can verify that (4.6) identifies b.

4.3. CONDITIONAL PREDICTION PROBLEMS

Perhaps the most familiar class of moment regression problems are the conditional prediction problems. Here, one observes a realization of $x(z)$ and wishes to make an optimal point prediction of the realization of some other random variable $y(z)$ conditional on the realization of $x(z)$. An optimal prediction is one that minimizes expected loss with respect to a specified loss function. In general, the best predictor of y given the event $x=\xi$ is some function of the probability measure of y conditional on $x=\xi$. The transformation from this measure to the best predictor depends on the loss function.

Let $y:Z \to Y \subset R^1$ be a given measurable function and let $\Theta \subset R^1$ be a given space of feasible predictor values. Let $L:R^1 \to [0,\infty]$ be a specified loss function, that is some measurable function such that

(4.7) $\quad 0 \leq v < w \Rightarrow 0 = L(0) \leq L(v) < L(w)$ and $L(0) \leq L(-v) < L(-w)$.

A conditional prediction problem is a collection of moment optimizations

in which h(*,*) of equation (4.2) has the form

$$(4.8)\quad h(z,\theta) = L[y(z)-\theta].$$

A best predictor of y conditional on x=ξ solves the problem

$$(4.9)\quad b(\xi) - \underset{\theta\in\Theta}{\operatorname{argmin}} \int L[y(z)-\theta]dP|\xi = 0.$$

METHOD OF MOMENT ESTIMATION: In problems where the parameter space is finite dimensional, the dominant approach to estimation of the best predictor function b is the method of moments. Applying (4.4) with w(*)=1, one observes that b(ξ) solves the conditional prediction problem (4.9) a.e. P_x if and only if b solves the unconditional prediction problem

$$(4.10)\quad b - \underset{a\in B}{\operatorname{argmin}} \int L[y(z)-a\{x(z)\}]dP = 0.$$

Then one estimates b by

$$(4.11)\quad B_N = \underset{a\in B}{\operatorname{argmin}} \int L[y(z)-a\{x(z)\}]dP_N.$$

For example, under the absolute loss function $L(y-\theta)=|y-\theta|$, B_N is the least absolute deviations estimate of b. Under the square loss function $L(y-\theta)=(y-\theta)^2$, B_N is the least squares estimate.

It is important to understand the substantive distinction between the prediction problems (4.9) and (4.10). In (4.9), the event [x=ξ] has been observed. The problem is to minimize over $\theta\in\Theta$ the expectation of L[y(z)-θ] with respect to the measure P|ξ. In (4.10), a realization of x will be drawn, following which a prediction of y will be made. The problem is to minimize over $a\in B$ the expectation of L[y(z)-a{x(z)}] with respect to the measure P. The fact that the same function b solves both

problems is a simple but remarkable consequence of the linearity of the expectation operator.

ESTIMATION WHEN THE PARAMETER SPACE IS LARGE: To what extent does the estimate (4.11) remain attractive when the parameter space is not finite dimensional? One class of problems admitting positive results is isotonic regression. Here, X is a space endowed with a known partial ordering and B is the space of functions that are monotone in this partial ordering. The simplest case is that in which $X=R^1$; then B is the space of increasing functions on the real line. See, for example, Barlow et al.(1972) and Sager and Thisted(1982).

A second class that has received attention are the binary response problems. Here, y is distributed Bernoulli conditional on x. Manski and Thompson(1986) consider various specifications for the loss function and for the parameter space.

They find that (4.11) is consistent for b if L is the log loss function $L(y-\theta)=-\log[1-|y-\theta|]$ and B is the space of functions that are increasing in an index $x\beta$, where β is a parameter vector. This specification of B generalizes that of isotonic regression; with β unknown, the ordering on X is not known. It turns out that B_N is the maximum likelihood estimate studied by Cosslett(1983).

They also study estimation when B is the space of functions $c:X\rightarrow R^1$ satisfying the 'single-crossing' condition $c(x)\geq\gamma \Leftrightarrow x\beta\geq 0$, where $\gamma\in[0,1]$ is known and β is again an unknown parameter. Here, B_N is consistent for b if an absolute loss function is imposed but not otherwise. Whatever loss function is imposed, (4.11) reduces to some version of maximum score estimation(Manski,1985).

4.4. PREDICTION BY MINIMIZATION OF QUANTILE LOSS

The literature on best prediction has focussed exclusively on optimality defined by minimization of the expectation of the loss function. There is, however, no compelling reason why one might not wish to minimize some other location parameter of the loss distribution, say some quantile, a trimmed mean, or the mode. For a decision theoretic analysis of some alternative decision rules, see Manski(1986).

In general, conditional prediction problems minimizing location parameters other than the expectation are not moment problems. As an example, we shall consider predictors minimizing quantile loss.

Let the loss function have the form $L(y-\theta)=|y-\theta|$. Let $\alpha\in(0,1)$. Given a realization of $x(z)$, suppose that one predicts $y(z)$ by minimizing the α-quantile of the conditional distribution of $|y(z)-\theta|$. For simplicity, assume that the probability measure of y conditional on the event $[x=\xi]$ has no mass points. Then the best predictor of y conditional on ξ solves the problem

$$(4.12)\quad b(\xi) \;-\; \underset{\theta\in\Theta}{\operatorname{argmin}}\,[\eta\colon P\{|y(z)-\theta|\leq\eta \mid \xi\} = \alpha] \;=$$

$$b(\xi) \;-\; \underset{\theta\in\Theta}{\operatorname{argmin}}\,[\eta\colon P\{\theta-\eta\leq y(z)\leq\theta+\eta \mid \xi\} = \alpha] \;=\; 0.$$

The first equality in (4.12) shows that the best predictor $b(\xi)$ has a pleasing interpretation in terms of conditional confidence intervals for y. That is, the best predictor of y given ξ is the center of the smallest confidence interval for y that has coverage probability α.

Equation (4.12) is not a moment problem. Let $\eta(\alpha,\xi,\theta)$ denote the

α-quantile of $|y(z)-\theta|$ conditional on ξ. For a given predictor value $\theta \in R^1$, $\eta(\alpha,\xi,\theta)$ solves the moment equation

$$(4.13) \quad \int 1[\theta-\eta(\alpha,\xi,\theta) \leq y(z) \leq \theta+\eta(\alpha,\xi,\theta)]dP|\xi = \alpha.$$

But $b(\xi)$ minimizes $\eta(\alpha,\xi,*)$ on Θ and does not itself solve a moment problem.

ANALOG ESTIMATION: Analog estimation of $b(\xi)$ has been studied in the setting where one can sample directly from $P|\xi$. In particular, see Andrews et al.(1972), who discuss analog estimation of the closely related 'shorth', or mean of the shortest interval containing a fraction α of the probability mass of y.

We are concerned with estimation in the setting where one samples from P, not from $P|\xi$. In particular, it is of interest to learn whether the regression problem (4.12) can be represented in a manner that avoids reference to the conditional measures $P|\xi, \xi \in X$. This question is being investigated by the author in work in progress.

One finding obtained thus far is that such a representation does exist if a homoskedasticity condition holds. Specifically, it suffices that $\eta[\alpha,\xi,b(\xi)]$ be the same for all $\xi \in X$. In this case, application of the analogy principle to a suitable representation of (4.12) yields an estimator that has recently drawn attention for its robustness properties. This is the 'least median of squares' method proposed by Rousseuw(1984).

5. ANALOG ESTIMATION OF GENERAL REGRESSIONS

We now apply the analogy principle to representations of regression problems that refer explicitly to the conditional measures $P|\xi, \xi \in X$. The approaches to be developed here are less convenient than is method of moment estimation. On the other hand, they apply much more generally. We have seen that method of moments estimation of a moment regression is not consistent if the parameter space is too large or, in the case of moment equation regressions, if the derived moment equation does not identify b. Regression problems which are not collections of moment problems may have no representations that avoid reference to $P|\xi, \xi \in X$. The methods presented below are directed toward such problems.

5.1. 'NAIVE' ANALOG ESTIMATION OF REGRESSIONS WITH DISCRETE X

We begin with so-called 'naive' analog estimation. Naive estimation applies the analogy principle to our basic representation of a regression problem, equation (4.1). The sample analog of $P|\xi$ is the empirical conditional measure $P_N|\xi$. Hence, the analogy principle suggests that to estimate b, one might use

$$(5.1) \quad B_N \equiv [c \in B: [S(P_N|\xi, c(\xi)) = 0, \ \xi \in X],$$

provided that B_N is non-empty. Otherwise, one might seek to minimize the distance of $[S(P_N|\xi, *(\xi)), \ \xi \in X]$ from zero, in some sense.

This application of the analogy principle works if X is a discrete set and $P_x(\xi)>0$ for all $\xi\in X$. Here, $N\to\infty$ implies that for each $\xi\in X$, $P_N|\xi$ converges to $P|\xi$ almost surely. So for each $\xi\in X$, $S[P_N|\xi,*(\xi)]$ behaves well as an approximation to $S[P|\xi,*(\xi)]$, provided only that $S(*,*)$ is smooth. If X is finite, the convergence of $P_N|\xi$ to $P|\xi$ is uniform on X. In this case, we can make the stronger statement that $[S\{P_N|\xi,*(\xi)\},\xi\in X]$ behaves well as an approximation to $[S\{P|\xi,*(\xi)\},\xi\in X]$.

More generally, however, naive analog estimation does not work. The empirical measure of x, P_{Nx}, puts all its mass on its finite support $X_N\subset X$. For $\xi\notin X_N$, $P_N|\xi$ is arbitrary. For $\xi\in X_N$, $P_N|\xi$ is well-defined but does not converge to $P|\xi$ unless $P_x(\xi)>0$. Hence, wherever $P_x(\xi)=0$, $S[P_N|\xi,*(\xi)]$ behaves poorly as an approximation to $S[P|\xi,*(\xi)]$.

5.2. KERNEL ESTIMATION OF REGRESSIONS WITH ABSOLUTELY CONTINUOUS P

The failure of naive analog estimation when X is not discrete has fostered a widespread presumption that the analogy principle cannot be applied to general regression problems. The recent literature on nonparametric estimation of regressions has contributed to this view. Available nonparametric regression methods are in large part outgrowths of earlier work on nonparametric estimation of density functions. As indicated earlier, nonparametric density methods give the appearance of being divorced from the analogy principle.

In Section 3.2, we demonstrated that the density estimation problem can be represented in a manner that allows the analogy principle to work. Here, we use this representation to obtain an analog estimate of regressions that works when P has a Lebesgue density.

Assume that P is absolutely continuous with respect to Lebesgue

measure μ. Then $P|\xi$ is completely characterized by the conditional density $\phi_\mu(*,P|\xi)$. So an alternative representation of (4.1) is

$$(5.2) \quad [S\{\phi_\mu(*,P|\xi),b(\xi)\},\ \xi\in X] = 0,$$

where the domain of $S\{*,b(\xi)\}$ is now a space of density functions rather than a space of probability measures.

For each $\xi\in X$, the conditional density $\phi_\mu(*,P|\xi)$ can be written as the ratio of the density of P evaluated at realizations of z satisfying $[x(z)=\xi]$ to the density of P_x evaluated at ξ, that is

$$(5.3) \quad \phi_\mu(*,P|\xi) = \frac{1[x(z)=\xi]\ \phi_\mu(*,P)}{\phi_\mu(\xi,P_x)}.$$

We showed in Section 3.2 that $\phi_\mu(*,P)$ and $\phi_\mu(*,P_x)$ are the same as $\phi_\mu(*,P\oplus\sigma(P)G]$ and $\phi_\mu[*,P_x\oplus\sigma(P_x)G]$, where $\sigma(*)$ and G were defined in equation (3.12). Hence, the regression problem (5.2) is equivalent to the problem,

$$(5.4) \quad \left[S\left\{ \frac{1[x(z)=\xi]\ \phi_\mu(*,P\oplus\sigma(P)G]}{\phi_\mu[\xi,P_x\oplus\sigma(P_x)G]}, b(\xi)\right\}, \xi\in X \right] = 0.$$

The analogy principle may be applied to the representation (5.4). For example, let (4.1) and (5.2) have the form

$$(5.5) \quad [b(\xi)-\int y(z)dP|\xi,\ \xi\in X] = [b(\xi)-\int y(z)\{\phi_\mu(z,P|\xi)\}dz,\ \xi\in X] = 0$$

and let $B = (x\Theta,\xi\in X)$. Then b is the mean regression of y on x. For each $\xi\in X$, the analog estimate of $b(\xi)$ is, by (3.14),

$$(5.6)\quad B_N(\xi) = \frac{1/\sigma(P_N)\ \int y(z)\ 1[x(z)=\xi]\ [\int g\{(z-z_N)/\sigma(P_N)\}dP_N]\ dz}{1/\sigma(P_{Nx})\ \int g[(\xi-x_N)/\sigma(P_{Nx})]dP_{Nx}},$$

where z_N and x_N are distributed P_N and P_{Nx} respectively. Equation (5.6) is a kernel regression estimate. See Prakasa Rao(1983), p.239-240.

If P is not absolutely continuous, one should not expect analog estimates derived from (5.4) to be well-behaved. In particular, if P_x has mass points and $\sigma(P_N)$ measures the distance of P_N from a continuous measure, then $\sigma(P_{Nx})$ does not converge to zero as $N\to\infty$. So the local averaging on which kernel estimation is predicated goes askew.

5.3. SMALLEST NEIGHBORHOOD ESTIMATION OF GENERAL REGRESSIONS

So far, we have given a variety of analog estimates for regression functions, each appropriate to a different class of problems. In each case, the key to successful application of the analogy principle was selection of a suitable representation of the estimation problem. Thus, we transformed moment regressions into moment problems. We maintained the original regression form (4.1) for problems with discrete X. And we represented problems with absolutely continuous P as ones with smoothed densities.

Here, we introduce a new representation of regression problems that yields an appealing, generally applicable analog estimate. In short, we replace probability measures conditioning on events of probability zero by ones that condition on neighborhoods having vanishingly small positive probability. This done, application of the analogy principle

works whether or not X is discrete and whether or not P is absolutely continuous. We term the resulting analog method 'smallest neighborhood' estimation.

REPRESENTATION OF THE REGRESSION PROBLEM: Let $\rho(*,*)$ be a metric generating the usual topology on the space X. Let $m(*)$ be a strictly increasing function mapping $[0,\infty)$ into $[0,\infty)$ with $m(0)=0$ and $m(d)\geq d$ for $d>0$. For $\xi \in X$ and $d\geq 0$, define

$$(5.7) \quad X(\xi,d) \equiv [\xi' \in X: \rho(\xi,\xi') \leq d],$$

$$(5.8) \quad d(\xi,P_x) \equiv \inf d: P_x[X(\xi,d)]>0,$$

and

$$(5.9) \quad A(\xi,P_x) \equiv X[\xi, m\{d(\xi,P_x)\}].$$

Thus, $X(\xi,d)$ is the closed ball of radius d centered at ξ and $d(\xi,P_x)$ is the infimum of d such that $X(\xi,d)$ has positive probability under P_x. The set $A(\xi,P_x)$ is the closed ball of radius $m\{d(\xi,P_x)\}$ centered at ξ.

Now consider the regression problem (4.1) with $[P|A(\xi,P_x), \xi \in X]$ replacing $(P|\xi, \xi \in X)$. That is, let b solve

$$(5.10) \quad [S\{P|A(\xi,P_x), b(\xi)\}, \xi \in X] = 0.$$

We shall show that the estimation problems defined by (4.1) and (5.10) are equivalent.

To see this, let $X_s \subset X$ be the support of P_x. That is,

(5.11) $X_s \equiv [\xi \in X: P_x[X(\xi,d)]>0, \ \forall d>0]$.

It follows from (5.7) through (5.9) that

(5.12) $\xi \in X_s \Rightarrow d(\xi,P_x)=0 \Rightarrow m\{d(\xi,P_x)\}=0 \Rightarrow A(\xi,P_x)=\{\xi\} \Rightarrow P|A(\xi,P_x)=P|\xi$.

Hence, $[P|A(\xi,P_x),\xi \in X_s] = [P|\xi,\xi \in X_s]$.

It remains to consider $\xi \in X-X_s$. In general, $P|A(\xi,P_x)$ need not equal $P|\xi$ for such ξ. But $X-X_s$ has probability zero under P_x. See Chung, (1974), p.31. So the behavior of $P|A(\xi,P_x),\xi \in X-X_s$ is immaterial.

SMALLEST NEIGHBORHOOD ESTIMATES: A smallest neighborhood estimate of b is obtained by applying the analogy principle to (5.10). Thus, the estimate is

(5.13) $B_N = [c \in B: S\{P_N|A(\xi,P_{Nx}),c(\xi)\} = 0, \ \xi \in X]$,

provided that B_N is non-empty. Otherwise, one minimizes the distance of $[S\{P_N|A(\xi,P_{Nx}),*(\xi)\}, \ \xi \in X]$ from zero, in some sense.

The expression $P_N|A(\xi,P_{Nx})$ appears forbidding but actually has a simple interpretation. By (5.8),

(5.14) $d(\xi,P_{Nx}) = \min_{\xi' \in X_N} \rho(\xi,\xi')$,

where $X_N \subset X$ is the finite support of P_{Nx}. Thus, $d(\xi,P_{Nx})$ is the distance from ξ to its nearest neighbor among the sample observations of x. And $X[\xi,d(\xi,P_{Nx})]$ is the smallest neighborhood of ξ having positive

empirical probability.

If $\xi \in X_N$, then $d(\xi, P_{Nx})=0$. Hence, by (5.12), $P_N|A(\xi, P_{Nx}) = P_N|\xi$, as in naive analog estimation. If $\xi \notin X_N$, then $m\{d(\xi, P_{Nx})\} \geq d(\xi, P_{Nx}) > 0$. So $A(\xi, P_{Nx})$ is the neighborhood $X[\xi, d(\xi, P_{Nx})]$ 'blown up' to radius $m\{d(\xi, P_{Nx})\}$. In this case, $P_N|A(\xi, P_{Nx})$ is the empirical probability measure of z conditioned on the event that x is within distance $m\{d(\xi, P_{Nx})\}$ of ξ.

COMPARISON WITH THE NEAREST NEIGHBOR AND HISTOGRAM METHODS: Smallest neighborhood estimation is reminiscent of but distinct from the nearest neighbor and histogram methods. All three methods impose a metric on X and estimate $P|\xi$ by the empirical measure of z conditioned on the event that x is within some neighborhood of ξ. They differ in the way this neighborhood is determined.

In nearest neighbor estimation, a positive integer k, dependent on the sample size N, is chosen by the analyst. Let $d_{Nk}(\xi)$ be the distance from ξ to its k^{th} nearest neighbor among the N observations of x. Then $P|\xi$ is estimated by $P_N|X[\xi, d_{Nk}(\xi)]$. Thus, the number of observations used to estimate $P|\xi$ is predetermined and the neighborhood of ξ that contains these observations is random.

In histogram estimation, a neighborhood radius $\delta(N)>0$, dependent on N, is chosen by the analyst. Then $P|\xi$ is estimated by $P_N|X[\xi, \delta(N)]$. Here, the number of observations within the selected neighborhood of ξ is random.

In smallest neighborhood estimation, $m:[0,\infty) \rightarrow m[0,\infty)$ is chosen by the analyst. When $m(*)$ is evaluated at the random distance $d(\xi, P_{Nx})$ from ξ to its nearest neighbor, a random neighborhood $A(\xi, P_{Nx})$ results. The

number of observations within $A(\xi,P_{Nx})$ is random but always positive.

It would be of interest to know whether the nearest neighbor and histogram methods can be derived as analog estimates. I have not yet found representations of the regression problem that yield these methods.

k^{th}-SMALLEST NEIGHBORHOOD ESTIMATION: Smallest neighborhood estimation has one irritating feature not shared by the nearest neighbor and histogram methods. Fix $\xi \in X$. We pointed out earlier that if one or more sample observations of x have the value ξ, then the smallest neighborhood estimate of $P|\xi$ is the naive estimate $P_N|\xi$. This is desirable if $P_x(\xi)>0$ but not if $P_x(\xi)=0$.

The equivalence of smallest neighborhood and naive estimates on $\xi \in X_N$ is immaterial if one is concerned only with pointwise consistency. If $P_x(\xi)=0$, then with probability one, no observation of x equals ξ. Hence, $B_N(\xi)$ can still be a consistent estimate of $b(\xi)$. See Section 5.4. On the other hand, this property of smallest neighborhood estimation implies that unless P_x has finite support X_s, $B_N(*)$ cannot, in general, be uniformly consistent for $b(*)$ on X_s.

The problem noted here can be fixed by generalizing smallest neighborhood estimation to 'k^{th}-smallest neighborhood estimation', as follows. Fix an integer $k \geq 1$. As earlier, let $d_{Nk}(\xi)$ be the distance from ξ to its k^{th}-nearest neighbor among the N observations of x. Let $A_{Nk}(\xi) \equiv X[\xi, m\{d_{Nk}(\xi)\}]$. Now let $P|A_{Nk}(\xi)$ define the k^{th}-smallest neighborhood estimate of $P|\xi$.

A k^{th}-smallest neighborhood estimate of $P|\xi$ reduces to the naive estimate only if k or more sample observations of x have the value ξ. With probability one, there exists no $\xi \in X_s$ with $P_x(\xi)=0$ such that more

than one observation of x has the value ξ. Hence, with probability one, k^{th}-smallest neighborhood estimation with $k \geq 2$ does not misbehave anywhere on X_s.

I have not been able to find a representation of $P|\xi$ for which $P|A_{Nk}(\xi)$ is the sample analog. The integer k refers to a number of sample observations. For $k \geq 2$, the distance to the k^{th}-nearest neighbor of ξ is not determined by the empirical measure P_{Nx} alone. Hence, this distance has no obvious counterpart in the population, which is characterized only by P.

The case k=1, which yields smallest neighborhood estimation, is special. The distance to the nearest neighbor is the same as the distance to the smallest neighborhood having positive empirical probability. The latter distance is determined fully by P_{Nx}.

5.4. CONSISTENCY OF SMALLEST NEIGHBORHOOD ESTIMATES OF MEAN REGRESSIONS

To investigate the properties of smallest neighborhood estimation in a comprehensive way would require us to disgress too much from the theme of this paper. We shall therefore restrict attention to a central asymptotic question, the pointwise weak consistency of the smallest neighborhood estimate of a mean regression.

In what follows, a theorem gives conditions that are sufficient for consistency. Then a set of lemmas show that these conditions can be satisfied if m(*) is selected appropriately, provided only that P be minimally regular. Proofs are in an Appendix.

Consistency Theorem: Let $y:Z\to Y\subset R^1$ be a given measurable function such that $b(\xi')\equiv\int y(z)dP|\xi'$ exists for all $\xi'\in X$. Fix $\xi\in X$. Let $N(\xi,P_{Nx})$ be the number of sample observations of z for which $x(z)\in A(\xi,P_{Nx})$. Assume that the following conditions hold:

[1a] $\xi'\to\xi \Rightarrow b(\xi')\to b(\xi)$.

[1b] $\exists d_\circ>0$ and $\lambda>0$ s.t. $Var(y|x=\xi') \leq \lambda$ for $\xi'\in X(\xi,d_\circ)$.

[1c] As $N\to\infty$, $d(\xi,P_{Nx})\to 0$ in probability.

[1d] As $N\to\infty$, $N(\xi,P_{Nx})\to\infty$ in probability.

Then as $N\to\infty$, $\int y(z)dP_N|A(\xi,P_{Nx}) \to b(\xi)$ in probability.■

The four conditions of this theorem are unsurprising. Smallest neighborhood estimates, like histogram and nearest neighbor estimates, approximate the conditional mean $b(\xi)$ by local averages. For such local averages to be consistent, the population must be sufficiently regular. Conditions [1a] and [1b] suffice. That is, it is enough that $b(*)$ be continuous at ξ and that, for ξ' near ξ, the variances of the measures $P|\xi'$ be bounded.

Given these regularity conditions on P, the local average converges if Conditions [1c] and [1d] hold. That is, as $N\to\infty$, the neighborhood of ξ on which the average is taken should shrink toward ξ and, at the same time, the average should be computed on increasingly many observations.

SELECTION OF $m(*)$: There is, of course, a tension between Conditions [1c] and [1d]. To make the consistency theorem operational, we need to show that it is possible to select $m(*)$ so that both [1c] and [1d] hold.

To do this, we work with the distribution $G_\xi(*)$ of the distance of the random variable x to the point $\xi\in X$. That is, for $d\geq 0$, define

(5.15) $G_\xi(d) \equiv P_x[X(\xi,d)]$,

where $X(\xi,d)$ was defined in (5.7). By the Lebesgue decomposition theorem, the probability measure on $[0,\infty)$ generated by G_ξ can be decomposed uniquely into the sum of a discrete measure, a singular continuous measure, and a measure that is absolutely continuous with respect to Lebesgue measure. See Chung(1974), p.12. Let $g_\xi(*)$ denote the density of the absolutely continuous component.

With this as background, we have the following:

Lemma 1: Assume that $\xi \epsilon X_s$. Let $N\to\infty$. Then $d(\xi,P_{Nx})\to 0$ in probability. ■

Lemma 2: Assume that $P_x(\xi)>0$. Let $N\to\infty$. Then $N(\xi,P_{Nx})\to\infty$ almost surely. ■

Lemma 3: Assume that for some $d_1>0$, $G_\xi(d_1)=\int_0^{d_1} g_\xi(\delta)d\delta$. Moreover, $g_1 \le g_\xi(\delta) \le g_2$ for $\delta \le d_1$, where $0<g_1<g_2<\infty$. Let $m(*)$ be differentiable with with derivative $m_1(*)$ satisfying the condition $m_1(\delta)\to\infty$ as $\delta\to 0$. Then $N(\xi,P_{Nx})\to\infty$ in probability. ■

Lemma 1 states that Condition [1c] holds if ξ is in the support of P_x. Lemma 2 says that Condition [1d] holds if P_x places positive mass at ξ. These simple results require no conditions on $m(*)$ beyond the maintained assumptions that $m(0)=0$ and that $m(d)\ge d$.

Lemma 3 addresses a much more subtle question. Can $m(*)$ be chosen so that Condition [1d] is satisfied when P_x places zero mass at ξ? We obtain a positive answer provided only that $G_\xi(*)$ is well-behaved in a neighborhood of zero. In particular, it suffices that in a neighborhood

of zero, the density $g_\xi(*)$ of the absolutely continuous component of $G_\xi(*)$ be bounded away from zero and infinity. The Lemma also assumes that $G_\xi(*)$ has no singular continuous component in a neighborhood of zero but this condition is inessential.

We find that Condition [1d] is satisfied if m(*) is a function whose derivative $m_1(\delta)\to\infty$ as $\delta\to 0$. This property is essential. It can be shown that if $m_1(\delta)$ stays bounded as $\delta\to 0$, then $N(\xi,P_{Nx})$ stays bounded with positive probability.

One class of functions m(*) that work are the power functions

$$(5.16)\quad m(d) = d + \alpha_1 d^{\alpha_2},$$

for $0<\alpha_1<\infty$ and $0<\alpha_2<1$. Here, $m_1(\delta) = 1 + \alpha_1\alpha_2 d^{\alpha_2-1}$.

Given that all m(*) of the form (5.16) yield consistent estimates, one would like guidance on the selection of the constants (α_1,α_2). More generally, one would like a criterion for selection of m(*) from the space of all functions that satisfy the assumptions of Lemma 3. This question, which will not be pursued here, resembles questions that arise in nearest neighbor and histogram estimation. There, the analyst must decide how to increase the number of neighbors or shrink the window width as $N\to\infty$. Here, the problem of selecting a function whose argument is the sample size is replaced by one of selecting a function whose argument is the distance to the nearest neighbor.

APPENDIX: PROOFS OF RESULTS ON SMALLEST NEIGHBORHOOD ESTIMATION

Proof of Theorem: For each N, let $I(\xi,N)$ index the $N(\xi,P_{Nx})$ observations of z for which $x(z)\in A(\xi,P_{Nx})$. In general,

$$(A1)\quad b_N(\xi) \equiv \int y(z)\,dP_N|A(\xi,P_{Nx}) = \frac{1}{N(\xi,P_{Nx})} \sum_{i\in I(\xi,N)} y(z_i).$$

Hence, conditional on the sample size and on the realizations of x, the mean and variance of $b_N(\xi)$ are, provided that the relevant terms exist,

$$(A2)\quad E[b_N(\xi)|P_{Nx}] = \frac{1}{N(\xi,P_{Nx})} \sum_{i\in I(\xi,N)} b(x_i)$$

and

$$(A3)\quad Var[b_N(\xi)|P_{Nx}] = \frac{1}{N(\xi,P_{Nx})^2} \sum_{i\in I(\xi,N)} Var(y|x=x_i),$$

where $x_i=x(z_i)$.

Condition [1a] implies that given any $\eta>0$, there exists $d_\eta>0$ such that

$$(A4)\quad x_i\in X(\xi,d_\eta) \Rightarrow |b(x_i)-b(\xi)| < \eta.$$

Recall that $A(\xi,P_{Nx}) \equiv X[\xi,m\{d(\xi,P_{Nx})\}]$. Hence,

$$(A5)\quad E[b_N(\xi)|P_{Nx},d(\xi,P_{Nx})<m^{-1}(d_\eta)] \in [b(\xi)-\eta,b(\xi)+\eta].$$

Condition [1b] implies that the variances $Var(y|x=x_i)$, $i\in I(\xi,N)$ are bounded by λ, provided that $A(\xi,P_{Nx}) \subset X(\xi,d_\circ)$. Hence,

$$(A6)\quad Var[b_N(\xi)|P_{Nx},d(\xi,P_{Nx})<m^{-1}(d_\circ)] \leq \frac{\lambda}{N(\xi,P_{Nx})}.$$

Now let $0<\delta<\min[m^{-1}(d_\eta), m^{-1}(d_\infty)]$ and let $0<J<\infty$. Consider the mean and variance of $b_N(\xi)$ conditional on the sample size and on the event that the empirical measure of x is a member of the set of measures

$$(A7) \quad C(\delta,J) \equiv [Q_x : d(\xi,Q_x)<\delta \cap N(\xi,Q_x)>J].$$

It follows from (A5) that for all feasible δ and J,

$$(A8) \quad E_{N\delta J} \equiv E[b_N(\xi) \mid P_{Nx} \in C(\delta,J)] \in [b(\xi)-\eta, b(\xi)+\eta].$$

It follows from (A5) and (A6) that

$$\begin{aligned}(A9) \quad V_{N\delta J} &\equiv \mathrm{Var}[b_N(\xi) \mid P_{Nx} \in C(\delta,J)] \\ &= \mathrm{Var}\left[E[b_N(\xi) \mid P_{Nx}, d(\xi,P_{Nx})<m^{-1}(\delta)] \,\middle|\, P_{Nx} \in C(\delta,J)\right] \\ &\quad + \\ &\quad E\left[\mathrm{Var}[b_N(\xi) \mid P_{Nx}, d(\xi,P_{Nx})<m^{-1}(\delta)] \,\middle|\, P_{Nx} \in C(\delta,J)\right] \\ &\leq 4\eta^2 + \lambda/J.\end{aligned}$$

Chebychev's inequality and (A9) imply that for any $\nu>0$,

$$(A10) \quad \mathrm{Prob}[|b_N(\xi)-E_{N\delta J}|<\nu \mid P_{Nx} \in C(\delta,J)] > 1-V_{N\delta J}/\nu^2 \geq 1-(4\eta^2+\lambda/J)/\nu^2.$$

Hence, by (A8),

$$(A11) \quad \mathrm{Prob}[|b_N(\xi)-b(\xi)|<\eta+\nu \mid P_{Nx} \in C(\delta,J)] > 1 - (4\eta^2+\lambda/J)/\nu^2.$$

Finally, remove the conditioning on $C(\delta,J)$. In general,

(A12) $\text{Prob}[|b_N(\xi)-b(\xi)|<\eta+\nu] =$

$$\text{Prob}[|b_N(\xi)-b(\xi)|<\eta+\nu \mid P_{Nx}\in C(\delta,J)] \times \text{Prob}[P_{Nx}\in C(\delta,J)]$$

$$+ \text{Prob}[|b_N(\xi)-b(\xi)|<\eta+\nu \mid P_{Nx}\notin C(\delta,J)] \times \text{Prob}[P_{Nx}\notin C(\delta,J)].$$

Conditions [1c] and [1d] imply that as $N\to\infty$,

(A13) $\text{Prob}[P_{Nx}\in C(\delta,J)] \to 1.$

This and (A12) imply that

(A14) $\liminf_{N\to\infty} \text{Prob}[|b_N(\xi)-b(\xi)|<\eta+\nu] \geq 1 - (4\eta^2+\lambda/J)/\nu^2.$

Now let $\eta\to 0$ and $J\to\infty$. By (A14), $\text{Prob}[|b_N(\xi)-b(\xi)|<\nu] \to 1$ for every $\nu>0$.

Q.E.D.

Proof of Lemma 1: $d(\xi,P_{Nx})$ is the distance from ξ to its nearest neighbor among the N observations of x. By (5.15) and the assumption of random sampling,

(A15) $\text{Prob}[d(\xi,P_{Nx})\leq\delta] = 1 - [1-G_\xi(\delta)]^N$

for $\delta\geq 0$. By (5.11), $\xi\in X_s \Rightarrow G_\xi(\delta)>0$ for all $\delta>0$. Hence, for all $\delta>0$, $\text{Prob}[d(\xi,P_{Nx})\leq\delta] \to 1$ as $N\to\infty$.

Q.E.D.

Proof of Lemma 2: By the strong law of large numbers, $P_{Nx}(\xi) \to P_x(\xi) > 0$ almost surely. Let $0 < \eta < P_x(\xi)$. Then with probability one, there exists a finite $N_\circ$ such that $N > N_\circ \Rightarrow P_{Nx}(\xi) > \eta$. But $P_{Nx}(\xi) > \eta \Rightarrow d(\xi, P_{Nx}) = 0 \Rightarrow A(\xi, P_{Nx}) = \{\xi\} \Rightarrow N(\xi, P_{Nx}) = NP_{Nx}(\xi) > N\eta$.

Q.E.D.

Proof of Lemma 3: Let J be any positive integer. The Lemma states that as $N \to \infty$, $\text{Prob}[N(\xi, P_{Nx}) < J] \to 0$. But

$$\text{(A16)} \quad \text{Prob}[N(\xi, P_{Nx}) < J] = \sum_{j=0}^{J-1} \text{Prob}[N(\xi, P_{Nx}) = j].$$

By construction, $N(\xi, P_{Nx}) \geq 1$ always. Hence, it suffices to show that as $N \to \infty$, $\text{Prob}[N(\xi, P_{Nx}) = j] \to 0$ for each positive integer j.

For any $d > 0$,

$$\begin{aligned} \text{(A17)} \quad & \text{Prob}[N(\xi, P_{Nx}) = j] \\ & = \text{Prob}[N(\xi, P_{Nx}) = j \cap d(\xi, P_{Nx}) \leq d] + \text{Prob}[N(\xi, P_{Nx}) = j \cap d(\xi, P_{Nx}) > d] \\ & \leq \text{Prob}[N(\xi, P_{Nx}) = j \cap d(\xi, P_{Nx}) \leq d] + \text{Prob}[d(\xi, P_{Nx}) > d]. \end{aligned}$$

By assumption, $g_\xi(*) > 0$ in a neighborhood of zero. So $\xi \epsilon X_s$. Hence, by Lemma 1, $\text{Prob}[d(\xi, P_{Nx}) > d] \to 0$. Therefore, we need only to show that $\text{Prob}[N(\xi, P_{Nx}) = j \cap d(\xi, P_{Nx}) \leq d] \to 0$. In particular, it suffices to choose some $d_2 < m^{-1}(d_1)$ and show that $\text{Prob}[N(\xi, P_{Nx}) = j \cap d(\xi, P_{Nx}) \leq d_2] \to 0$.

By (5.14), $d(\xi, P_{Nx})$ is the first order statistic in a random sample of size N from G_ξ. Also, $N(\xi, P_{Nx}) = j$ if and only if the j^{th} order statistic is less than or equal to $m[d(\xi, P_{Nx})]$ and the $(j+1)^{st}$ is

greater than $m[d(\xi,P_{Nx})]$. By assumption, the mass of G_ξ in the interval $[0,\delta_0]$ derives entirely from the absolutely continuous component of G_ξ. It follows that

$$(A18)\quad \text{Prob}[N(\xi,P_{Nx})=j \cap d(\xi,P_{Nx})\leq d_{\cong}] =$$

$$= \int_0^{d_{\cong}} N\, g_\xi(\delta)\frac{(N-1)!}{(j-1)!(N-j)!}\left[G_\xi[m(\delta)]-G_\xi(\delta)\right]^{j-1}\left[1-G_\xi[m(\delta)]\right]^{N-j}\,d\delta$$

$$\leq \int_0^{d_{\cong}} g_\xi(\delta)\frac{N!}{(j-1)!(N-j)!}\left[G_\xi[m(\delta)]\right]^{j-1}\left[1-G_\xi[m(\delta)]\right]^{N-j}\,d\delta.$$

Thus, it suffices to show that as $N\to\infty$,

$$(A19)\quad \int_0^{d_{\cong}} g_\xi(\delta)\frac{N!}{(j-1)!(N-j)!}\left[G_\xi[m(\delta)]\right]^{j-1}\left[1-G_\xi[m(\delta)]\right]^{(N-j)}d\delta \to 0.$$

The integrand in (A19) is closely related to the density of the j^{th} order statistic of a random sample of N observations drawn from the distribution function $G_\xi[m(*)]$. To see this, first observe that strict monotonicity of $m(*)$ implies that $G_\xi[m(*)]$ is a legitimate distribution function. Differentiability of $m(*)$ implies that the absolutely continuous component of $G_\xi[m(*)]$ has density $g_\xi[m(*)]m_1(*)$. Hence, the density of the absolutely continuous component of the distribution of the j^{th} order statistic from $G_\xi[m(*)]$ is (see Lehmann, 1983, p.353)

$$(A20)\quad \phi_{\xi jN}(*) \equiv g_\xi[m(*)]\, m_1(*)\,\frac{N!}{(j-1)!(N-j)!}\left[G_\xi[m(*)]\right]^{j-1}\left[1-G_\xi[m(*)]\right]^{(N-j)}.$$

It follows that (A19) is equivalent to the condition

$$(A21)\quad \int_0^{d_\infty} \phi_{\xi jN}(\delta)\, \frac{g_\xi(\delta)}{g_\xi[m(\delta)]\; m_1(\delta)}\, d\delta \;\to\; 0.$$

By assumption, $0<g_1\le g_\xi(\delta)\le g_\infty$ for $\delta\le d_\infty$. Hence, (A21) is equivalent to

$$(A22)\quad \int_0^{d_\infty} \phi_{\xi jN}(\delta)\, \frac{1}{m_1(\delta)}\, d\delta \;\to\; 0.$$

Also by assumption, $m_1(\delta)\to\infty$ as $\delta\to 0$. Hence, given any $\zeta>0$, there exists a $\delta_\zeta>0$ such that $\delta\le\delta_\zeta \Rightarrow m_1(\delta)>\zeta$. Let $d_\infty=\delta_1$ and let $\zeta>1$. Then

$$(A23)\quad \int_0^{d_\infty} \phi_{\xi jN}(\delta)\, \frac{1}{m_1(\delta)}\, d\delta$$

$$= \int_0^{\delta_\zeta} \phi_{\xi jN}(\delta)\, \frac{1}{m_1(\delta)}\, d\delta \;+\; \int_{\delta_\zeta}^{d_\infty} \phi_{\xi jN}(\delta)\, \frac{1}{m_1(\delta)}\, d\delta$$

$$\le \frac{1}{\zeta} \int_0^{\delta_\zeta} \phi_{\xi jN}(\delta)\, d\delta \;+\; \int_{\delta_\zeta}^{d_\infty} \phi_{\xi jN}(\delta)\, d\delta.$$

As $N\to\infty$, the j^{th} order statistic of $G_\xi[m(*)]$ approaches zero with probability one. Hence, for all $\zeta>0$,

$$(A24)\quad \int_0^{\delta_\zeta} \phi_{\xi jN}(\delta)\, d\delta \;\to\; 1$$

and

$$(A25)\quad \int_{\delta_\zeta}^{d_\infty} \phi_{\xi jN}(\delta)\, d\delta \;\to\; 0.$$

It follows that the left hand side of (A23) is asymptotically bounded above by $1/\zeta$. Letting $\zeta\to\infty$ completes the proof.

Q.E.D.

REFERENCES

Andrews, D., Bickel, P., Hampel, F., Huber, P., Rogers, W., and Tukey, J.(1972), Robust Estimates of Location: Survey and Advances, Princeton: Princeton University Press.

Barlow, R., Bartholomew, D., Bremner, J., and Brunk, H.(1972), Statistical Inference Under Order Restrictions, New York: Wiley.

Burguete, J., Gallant, R., and Souza, G.(1982), "On Unification of the Asymptotic Theory of Nonlinear Econometric Models", Econometric Reviews, 1, 151-190.

Chamberlain, G.(1986), "Asymptotic Efficiency in Estimation With Conditional Moment Restrictions", Journal of Econometrics, forthcoming.

Chung, K.(1974), A Course in Probability Theory, Orlando:Academic Press.

Cosslett, S.(1983), Distribution-Free Maximum Likelihood Estimation of the Binary Choice Model", Econometrica 51, 765-782.

Goldberger, A.(1968), Topics in Regression Analysis, New York: McMillan.

Hansen, L.(1982), "Large Sample Properties of Generalized Method of Moment Estimators", Econometrica, 50, 1029-1054.

Huber, P.(1967), "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions", Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 221-233, Berkeley: University of California Press.

Huber, P.(1981), Robust Statistics, New York: Wiley.

Lehmann, E.(1983), Theory of Point Estimation, New York: Wiley.

Manski, C.(1983), "Closest Empirical Distribution Estimation", Econometrica, 51, 305-319.

Manski, C.(1985), "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator", Journal of Econometrics, 27, 303-333.

Manski, C.(1986), "Ordinal Utility Models of Decision Making Under Uncertainty", Social Systems Research Institute paper no. 8621, University of Wisconsin-Madison.

Manski, C.(1987), Analog Estimation Methods in Econometrics, in preparation.

Manski, C. and T. Thompson(1986), "Estimation of Best Predictors of Binary Response", unpublished.

Neyman, J.(1949), "Contributions to the Theory of the χ^2 Test", in Berkeley Symposium on Mathematical Statistics and Probability, Berkeley: University of California Press.

Parr, W. and Schucany, W.(1980), "Minimum Distance and Robust Estimation", Journal of the American Statistical Association, 75, 616-624.

Prakasa Rao, B.L.S.(1983), Nonparametric Functional Estimation, Orlando: Academic Press.

Rao, C.R.(1973), Linear Statistical Inference and Its Applications, New York: Wiley.

Reiersol, O.(1941), "Confluence Analysis by Means of Lag Moments and Other Methods of Confluence Analysis", Econometrica, 9, 1-23.

Reiersol, O.(1945), "Confluence Analysis by Means of Instrumental Sets of Variables", Arkiv Fur Matematik, Astronomi Och Fysik, 32A, no.4, 1-119.

Rousseeuw, P.(1984), "Least Median of Squares Regression", Journal of the American Statistical Association, 79, 871-880.

Sager, T. and Thisted, R.(1982), "Maximum Likelihood Estimation of Isotonic Modal Regression", Annals of Statistics, 10, 690-707.

Sahler, W.(1970), "Estimation by Minimum Discrepancy Methods", Metrika, 16, 85-106.

Serfling, R.(1980), Approximation Theorems of Mathematical Statistics, New York: Wiley.

Stone, R.(1977), "Consistent Nonparametric Regression", Annals of Statistics, 5, 595-645.

Von Mises, R.(1947), "On the Asymptotic Distribution of Differentiable Statistical Functions", Annals of Mathematical Statistics, 18, 309-348.

Wright, S.(1928), Appendix B to Wright, P. The Tariff on Animal and Vegetable Oils, New York: McMillan.

Wolfowitz, J.(1953), "Estimation by the Minimum Distance Method", Annals of the Institute of Statistics and Mathematics, 5, 9-23.

Wolfowitz, J.(1957), "The Minimum Distance Method", Annals of Mathematical Statistics, 28, 75-88.