Regression Models with Infinitely Many Parameters: Consistency of Bounded Linear Functionals

by

Ker-Chau Li
Purdue University

Mimeograph Series #82-10

Department of Statistics
Purdue University

April, 1982

Regression Models with Infinitely Many Parameters: Consistency of Bounded Linear Functionals

Abbreviated Title. "Infinitely Many Parameters" Regression

by

Ker-Chau Li
Department of Statistics
Purdue University

Abstract

Consider a linear model with infinitely many parameters given by $y=\sum_{j=1}^{\infty} x_j\theta_j+\varepsilon$ where $\underset{\sim}{x}=(x_1,x_2,\ldots)'$ and $\underset{\sim}{\theta}=(\theta_1,\theta_2,\ldots)'$ are infinite dimensional vectors such that $\sum_{i=1}^{\infty} x_i^2 < \infty$ and $\sum_{i=1}^{\infty} \theta_i^2 < \infty$. Suppose independent observations $y_1,y_2,\ldots,$ are observed at levels $\underset{\sim}{x}_1,\underset{\sim}{x}_2,\ldots$. Under suitable conditions about the error distribution, the set of all bounded linear functionals $T(\underset{\sim}{\theta})$ for which there exists a sequence of estimators $\hat{T}_n$ such that $\hat{T}_n \to T(\underset{\sim}{\theta})$ in probability will be characterized. An application will be extended to the nonparametric regression problem where the response curve f is assumed to be in the Hilbert space $W_2^m[0,1]=\{f|f^{(m-1)}$ is absolutely continuous on [0,1] and $\int_0^1 f^{(m)}(t)^2dt< \infty.\}$ The connection with the results of Wu (1980) for the least squares estiamtes under the usual linear model where the number of the parameters is finite is discussed.

AMS 1980 Subject Classification: Primary 62G05
Secondary 62J05

Keywords and Phrases: Asymptotic consistency, consistent region of degree K, Fisher's information, Hilbert space, limiting point of degree K, nonparametric regression.

Regression Models with Infinitely Many Parameters: Consistency of Bounded Linear Functionals

1. Introduction.

Consistency has been considered as a minimal requirement for an estimator to behave well when the sample size is large. (Further asymptotic properties such as the convergent rates can be studied only after consistency being established). For the case that the observations are i.i.d., this property is satisfied by various types of estimators (e.g., maximal likelihood estimator, and many nonparametric estimators such as M-, L-, R-estimators etc.) under various appropriate assumptions about the structure of the paramters and the probability distribution. The list of publications addressed on this topic is too lengthy to be included here. Consistency tends to be deemed so indispensable that whenever a new estimation technique with much potential for the wide application is introduced, the first asymptotic justification has been to seek for the "weakest" assumptions for the consistency to hold; e.g., the universal consistency of the nearest neighborhood estimators (Stone 1977), the consistency of bootstrapping method (Bickel and Freedman 1981, Freedman 1981) etc. One possible theoretic sequel for such papers may be a search for some modified estimators possessing the consistency under much more general assumptions. However, such an effort would turn out fruitless, had the "weakest" assumptions been so weak that violating these assumptions would mean that no estimators can be consistent at all. In other words, it seems to be a more fundamental question to ask whether the assumed setup admits any consistent estimator or not.

The importance of this consideration becomes more apparent when the observations are not i.i.d. For instance, in the usual linear regresison setup with a finite number of parameters, it has been well-known that the least squares estimates are not consistent if the minimal eigenvalue of the information matrix does not tend to infinity. For such cases, Drygas (1976) and Wu (1980) characterized the consistent directions (i.e., the linear combinations of parameters which can be consistently estimated) for least squares estimates. Since the ridge estimators and some versions of Stein estimators have become the natural alternatives in many situations, it is reasonable to inquire whether these alternative estimates are consistent or not when we want to estimate some inconsistent directions for the least squares estimates. As a consequence of the general result of Section 2 (see Section 3 for details), we shall demonstrate that for an inconsistent direction for the least squares estimates, not only neither the least squares estimates, nor the ridge estimators (with any possible choices of ridge constants), nor any version of Stein estimators are consistent, but also the setup itself does not admit any consistent estimators, under some natural assumptions about the error distribution (in particular, it should have a finite Fisher's information).

For the more complicated illy-posed problems where the number of parameters may be infinite and no estimators are considered as standard (for example, the non-parametric regression problem where the regressors are treated as non-stochastic), it is even more important to characterize, first, the setups for which the consistency is possible (regardless of what the estimator should be). Only after such a result being established, we then can move to the discussion of the proper choice of estimators among the consistent ones. Thus in this paper, we shall mainly investigate the consistency property for a regression model (in contrast to the consistency for certain estimators as was treated in most papers). The number of parameters may be infinite and thus makes the problem illyposed. We shall focus on estimating the bounded linear functionals of the parameters only. Specifically we consider the following situation.

Suppose a linear model is given by

$$(1.1) \qquad y = \sum_{j=1}^{\infty} x_j\theta_j + \varepsilon = \langle \underset{\sim}{x}, \underset{\sim}{\theta} \rangle + \varepsilon,$$

where $\underset{\sim}{\theta}=(\theta_1,\ldots,\theta_j,\ldots)'$ and $\underset{\sim}{x}=(x_1,\ldots,x_j,\ldots)'$ are infinite dimensional vectors in the Hilbert space $\ell^2=\{(a_1,\ldots,a_i,\ldots)' \mid \sum_{i=1}^{\infty} a_i^2 < \infty\}$; $\langle\cdot,\cdot\rangle$ is the inner product of ℓ^2, ε is the random error satisfying certain conditions as will be specified later on. Suppose independent observations $y_1, y_2, \ldots$ are observed at levels $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots$. We shall characterize the set of all bounded linear functionals $T(\underset{\sim}{\theta})$ for which there exists a sequence of estimators $\hat{T}_n$, based on the first n observations $y_1,\ldots,y_n$, such that $\hat{T}_n \to T(\underset{\sim}{\theta})$ in probability as $n\to\infty$, for any $\underset{\sim}{\theta}$ in the parameter space $\Theta \subset \ell^2$. Such functionals will be called consistently-estimable bounded linear functionals (hereafter, c.e.b.l. functionals) for Θ.

Section 2 provides the main theorem of this paper. According to this theorem, $T(\cdot)$ is a c.e.b.l. functional, if it is pairwise consistency in the sense that when restricting the parameter space to $\{0, \underset{\sim}{\theta}^*\}$ for any

$\underset{\sim}{\theta}^{*'} \in \Theta$, there exist consistent estimators (possibly depend on $\underset{\sim}{\theta}^*$) for $T(\cdot)$; or, equivalently, if the probability distribution of $(y_1,\ldots,y_n)'$ under $\underset{\sim}{\theta}=\underset{\sim}{\theta}^*$ is asymptotically singular to that under $\underset{\sim}{\theta}=0$ for any $\underset{\sim}{\theta}^* \in \Theta$. Statement (iii) of Theorem 2.1 is a simple consequence of the statement (ii) if we have the normality assumption about the error distribution. Instead of assuming normality we merely require that the error distribution possess a finite Fisher's information. In Wu (1980), the equivalence between (iii) and the statement

"$T(\cdot)$ is a consistent direction for the least squares estimates"

is established when Θ is finite dimensional. It is interpreted there that a good direction $\underset{\sim}{\omega}$ with the property that

$$(1.2) \qquad \sum_{i=1}^{\infty} \langle \underset{\sim}{x}_i, \underset{\sim}{\omega} \rangle^2 = +\infty$$

is a direction where we may gain much information and $T(\cdot)$ is a consistent direction if all the bad directions $\underset{\sim}{\omega}$ with the property that $\sum_{i=1}^{\infty} \langle \underset{\sim}{x}_i, \underset{\sim}{\omega} \rangle^2 < \infty$ are orthogonal to $T(\cdot)$. While the similar interpretation holds in our case here, the equivalance between (ii) and (iii) also provides a suitable explanation for why the direction $\underset{\sim}{\omega}$ is considered as "good" if (1.2) holds. Suppose we consider the simple against simple hypotsis testing problem: $H_0:\underset{\sim}{\theta}=0$ against $H_1:\underset{\sim}{\theta}=\underset{\sim}{\omega}$. Then a version of (ii) shows that (1.2) is a necessary condition for the existence of an asymptotic power one test. Thus in a good direction $\underset{\sim}{\omega}$ we gain so much information that we can distinguish H_1 from H_0 without making any error asymptotically. Mathematically, (iii) is a relatively simple statement to verify. All the stochastic nature of the consistency problem now diminishes and what remains for our concern is only the divergence problem about the positive infinite series $\sum_{i=1}^{\infty}<\underset{\sim}{x}_i,\underset{\sim}{\omega}>^2$. The statement (iv) allows us to conduct a suitable orthogonal transformation for the construction of a consistent estimator. By the similar argument as in proving (iv) ⇒ (v) there, one may obtain many classes of consistent estimators; see for example, (2.8b) and (2.8c) of Section 3. Finally, (v) claims that the consistent estimators are also consistent in the sense of the mean squared error.

Section 3 is devoted to the more detailed discussion of Theorem 2.1. Assumptions under which the theorem holds are investigated and more general conclusions are obtained.

An application of Theorem 2.1 is made to solve the nonparametric regression problem of Section 4. Suppose the response curve f is $m-1^{th}$ continuously differentiable on [0,1] with $f^{(m-1)}$ absolutely continuous and $\int_0^1 f^{(m)}(t)^2dt< \infty$. Observations $y_i=f(t_i)+\varepsilon_i$ are independent and errors ε_i are with mean 0 and have a common variance and a finite Fisher's information. The goal is to characterize the set of all c.e.b.l. functionals. In particular, we would like to obtain the necessary and sufficient conditions for $f^{(k)}(t)$ to be a c.e.b.l. functional. Theorem 4.1 determines the consistent region of degree k, defined to be the set of all points at which the k-th derivative of f is consistently estimable. After introducing the notion of the limiting points of degree k for a sequence $\{t_i\}_{i=1}^{\infty}$, we show that the consistent region of degree k is equal to the set of all limiting point of degree k. A connection with Wu's results on the polynomial regression is quite interesting. In our terminology, the set of all consistent directions for the least squares in the polynomial regression model of degree m-1 is equal to the linear space generated by all $f^{(k)}(t^*)$, where $f(t)=\sum_{j=0}^{m-1}\theta_j t^i$ and t* is any limiting point of degree k, k=0,1,...,m-1. The consistent region of degree k in the polynomial case may be larger than the set of all limiting points of degree k simply because some $f^{(k)}(t)$ may be written as a linear combinations of other consistant directions. Such phenomenon does not occur in the nonparametric regression considered in Section 4. Finally, we also show that in many cases the closed linear space generated by all c.e.b.l. functionals of differential type equals the set of all c.e.b.l. functionals.

In the appendix, we derive a useful result concerning the finite Fisher's information. This is applied to connect (ii) and (iii) of Theorem 2.1.

2. Main results.

In this section, the notation $<\cdot,\cdot>$ will be used to denote either the inner product in ℓ^2 or the inner product in R^n without ambiguity.

Consider the linear regression model with infinitely many parameters given by (1.1). Since the results we shall derive here may also be applicable to the case where the parameter space Θ is not the entire ℓ^2 (e.g., $\Theta=\{\underset{\sim}{\theta}|<\underset{\sim}{\theta},\underset{\sim}{\theta}> \leq\delta^2,\underset{\sim}{\theta}\in \ell^2\}$ for some known real number δ) and to the case where the usual linear model with finitely many parameters is considered, a suitable condition on the parameter space will be given as follows. Let Θ^* be the closed linear space generated by Θ. For $\delta>0$, let $B(\delta)=\{\underset{\sim}{\theta}|<\underset{\sim}{\theta},\underset{\sim}{\theta}> \leq\delta^2, \underset{\sim}{\theta}\in\Theta^*\}$. Assume that

(2.1) Θ contains $B(\delta^*)$ for some $\delta^* > 0$.

The probability distribution of the random error ε is assumed to satisfy the following two conditions:

(2.2) $E\varepsilon = 0$ and $0 \neq \mathrm{Var}\ \varepsilon = \sigma^2 <\infty$ (σ may or may not be known),

and

(2.3) ε has a finite Fisher's information (i.e., ε has a density f which is absolutely continuous and $\int_{-\infty}^{\infty} \frac{f'(x)^2}{f(x)}\,dx < \infty$).

It is clear that (2.3) is satisfied by many important distributions including the normal distribution. In the proof of our main theorem below, we shall need a useful property about finite Fisher's information, which will be discussed in the Appendix because of its independent interest. Further discussion on these conditions will be provided in the next section. We now present the main theorem of this paper. The convention that the "inf" of an empty set is $+\infty$ will be adopted.

Theorem 2.1. Assume that (2.1)∿(2.3) hold. For the regression model (1.1), the following statements are equivalent:

(i) $T(\cdot)$ is a c.e.b.l. functional for $\underset{\sim}{\theta} \in \Theta$.

(ii) (Pairwise consistency). For any $\underset{\sim}{\theta}^* \in \Theta$, $T(\cdot)$ is a c.e.b.l. functional when the parameter space is restricted to $\{\underset{\sim}{0}, \underset{\sim}{\theta}^*\}$.

(iii) For any $\underset{\sim}{\theta} \in \Theta$ such that $T(\underset{\sim}{\theta}) \neq 0$,

$$\sum_{i=1}^{\infty} \langle \underset{\sim}{x}_i, \underset{\sim}{\theta} \rangle^2 = \infty. \tag{2.4}$$

(iv) For any $\delta > 0$,

$$\liminf_{n\to\infty} \{ \sum_{i=1}^{n} \langle \underset{\sim}{x}_i, \underset{\sim}{\theta} \rangle^2 \mid \underset{\sim}{\theta} \in B(\delta),\ T(\underset{\sim}{\theta}) = 1 \} \to \infty. \tag{2.5}$$

(v) There exists a sequence of estimators $\{\hat{T}_n\}$, where $\hat{T}_n$ is based on the first n observations, such that

$$E(\hat{T}_n - T(\underset{\sim}{\theta}))^2 \to 0, \text{ as } n\to\infty,$$

for any $\underset{\sim}{\theta} \in \Theta$.

Proof.

"(i) $\Rightarrow$ (ii)" holds obviously.

"(ii) $\Rightarrow$ (iii)". Suppose there exists a $\underset{\sim}{\theta}^* \in \Theta$ such that $T(\underset{\sim}{\theta}^*) \neq 0$ but (2.4) does not hold. Let $h_i = \langle \underset{\sim}{x}_i, \underset{\sim}{\theta}^* \rangle$. Then $\sum_{i=1}^{\infty} h_i^2 < \infty$. Let $\underset{\sim}{P}_n$ and $\underset{\sim}{Q}_n$ be the probability measure of $(y_1, \ldots, y_n)'$ when $\underset{\sim}{\theta} = \underset{\sim}{0}$ and $\underset{\sim}{\theta} = \underset{\sim}{\theta}^*$ respectively. Consider the case where the parameter space is restricted to $\{\underset{\sim}{0}, \underset{\sim}{\theta}^*\}$. It is clear that (ii) implies that $\underset{\sim}{P}_n$ and $\underset{\sim}{Q}_n$ are asymptotically mutually-singular. However, this is contradictory to the Theorem in the Appendix.

"(iii) $\Rightarrow$ (iv)". Since the "inf" of an empty set is $+\infty$, we may consider those δ such that $\{\underset{\sim}{\theta} \mid \theta \in B(\delta),\ T(\underset{\sim}{\theta}) = 1\} \neq \phi$ only. For such a δ and any $c > 0$, define

$$A_n = \{\underset{\sim}{\theta} \mid \underset{\sim}{\theta} \in B(\delta),\ T(\underset{\sim}{\theta}) = 1, \text{ and } \sum_{i=1}^{n} \langle \underset{\sim}{x}_i, \underset{\sim}{\theta} \rangle^2 \leq c\}.$$

Since $A_n \subseteq A_{n-1}$, it suffices to show that $A_n = \phi$ for some n. Observe that (iii) implies that (2.4) holds for any $\underset{\sim}{\theta} \in B(\delta^*)$ because of the assumption (2.1). Thus it follows that (2.4) holds for any $\underset{\sim}{\theta} \in B(\delta)$. Therefore we have

$$(2.6) \qquad \bigcap_{n=1}^{\infty} A_n = \phi;$$

otherwise, for any $\underset{\sim}{\theta} \in \bigcap_{n=1}^{\infty} A_n$, we get $\sum_{i=1}^{n} \langle \underset{\sim}{x}_i, \underset{\sim}{\theta} \rangle^2 \leq c > \infty$ for any n, a contradiction to (2.4). Now, consider the weak topology on the space Θ^*

(since Θ^* is a Hilbert space, weak topology and weak* topology are identical). By Aloaglu's Theorem (c.f. Roydon 1972, page 202), $B(\delta)$ is weakly compact. Since $T(\cdot)$ and $<\underset{\sim}{x}_i,\cdot>$ are weakly continuous, it is clear that A_n is weakly closed. Moreover, since $A_n \subset B(\delta)$, A_n is also weakly compact. Therefore, by a fundamental property of compactness, from (2.6) it follows that there exists some N such that

$$\bigcap_{n=1}^{N} A_n = \phi.$$

Since $A_n \subset A_{n-1}$, we get $A_N=\phi$. Thus the proof is complete.

"(iv) $\Rightarrow$ (v)". Let $\underset{\sim}{\nu}$ be the element in Θ^* such that $<\underset{\sim}{\nu},\underset{\sim}{\theta}>=T(\underset{\sim}{\theta})$ for any $\underset{\sim}{\theta} \in \Theta^*$ (such a $\underset{\sim}{\nu}$ exists because of Rietz representation Theorem). Without loss of generality, we may assume that $<\underset{\sim}{\nu},\underset{\sim}{\nu}>=1$, because it is obvious that if $\frac{T(\cdot)}{<\underset{\sim}{\nu},\underset{\sim}{\nu}>^{1/2}}$ is a c.e.b.l. functional then $T(\cdot)$ is a c.e.b.l. functional. Before constructing the consistent estimators, we shall make a suitable orthogonal transformation on the vector of observations $(y_1,\ldots y_n)'$ and choose a convenient complete orthonormal system on Θ^* so that the design matrix takes the form

$$(2.7)\qquad \left(\begin{array}{c|ccccc|cccc} x_{10} & x_{11} & 0 & . & . & 0 & 0 & . & . & . \\ x_{20} & 0 & x_{22} & 0 & . & . & 0 & . & . & . \\ . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & 0 & . & . & . & . \\ x_{n0} & 0 & . & . & . & x_{nn} & 0 & . & . & . \end{array}\right)$$

and the first coordinate of the new parameter is what we want to estimate. To carry out this idea, let $\underset{\sim}{V}_n$ be the vector space generated by $\{\underset{\sim}{v},\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n\}$ and let $\underset{\sim}{U}_n=\{\underset{\sim}{u}|<\underset{\sim}{u},\underset{\sim}{v}>=0$ and $\underset{\sim}{u}\in\underset{\sim}{V}_n\}$. Consider the linear transformation L from $\underset{\sim}{U}_n$ to R^n defined by mapping $\underset{\sim}{u}$ to $(<\underset{\sim}{x}_1,\underset{\sim}{u}>,\ldots<\underset{\sim}{x}_n,\underset{\sim}{u}>)'$. It is a well-known fact in the linear algebra that there exists an orthonormal basis $\{\underset{\sim}{e}_1,\ldots,\underset{\sim}{e}_n\}$ in $\underset{\sim}{U}_n$ and an orthonormal basis $\{\underset{\sim}{g}_1,\ldots,\underset{\sim}{g}_n\}$ in R^n such that $L(\underset{\sim}{e}_i)=m_i\underset{\sim}{g}_i$ for some nonnegative number m_i, $i=1,\ldots,n$ (m_i may be taken as the squared roots of the eigenvalues of L'L). We extend the orthonormal basis $\{\underset{\sim}{v},\underset{\sim}{e}_1,\ldots,\underset{\sim}{e}_n\}$ in $\underset{\sim}{V}_n$ to a complete orthonormal system in Θ^* by adding an arbitrary complete orthonormal system in the orthogonal complement of $\underset{\sim}{V}_n$ in Θ^* to the set $\{\underset{\sim}{v},\underset{\sim}{e}_i,\ldots,\underset{\sim}{e}_n\}$. Write $\underset{\sim}{y}^{(n)}=(y_1,\ldots,y_n)'$, $\underset{\sim}{\varepsilon}^{(n)}=(\varepsilon_1,\ldots,\varepsilon_n)'$, $\underset{\sim}{A}^{(n)}=(<\underset{\sim}{x}_1,\underset{\sim}{v}>,\ldots,<\underset{\sim}{x}_n,\underset{\sim}{v}>)'$ and let $Z_i=<\underset{\sim}{y}^{(n)},\underset{\sim}{g}_i>$ and $\varepsilon_i'=<\underset{\sim}{\varepsilon}^{(n)},\underset{\sim}{g}_i>$. Now, it is clear that

$$(2.8)\qquad Z_i = <\underset{\sim}{A}^{(n)},\underset{\sim}{g}_i><\underset{\sim}{v},\underset{\sim}{\theta}> + m_i<\underset{\sim}{e}_i,\underset{\sim}{\theta}> + \varepsilon_i',\ i=1,\ldots,n,$$

and the random errors ε_i' are of mean 0 and uncorrelated with the common variance σ^2. Thus for $(Z_1,\ldots,Z_n)'$ and the complete orthonormal system $\{\underset{\sim}{v},\underset{\sim}{e}_1,\ldots\}$, the design matrix is of the form (2.7) with $x_{i0}=<\underset{\sim}{A}^{(n)},\underset{\sim}{g}_i>$ and $x_{ii}=m_i$. Note that to be precise we should have used the notations $x_{i0}^{(n)}$ and $x_{ii}^{(n)}$ instead of x_{i0} and x_{ii} (and $\underset{\sim}{e}_i^{(n)}$ instead of $\underset{\sim}{e}_i$) because the transformation L depends on n. However we omit the superscript (n) to avoid the complexity of notations. It should be understood that x_{i0} and x_{ii} may take different values for different sample size n and $\underset{\sim}{e}_i$ may be different too.

Now, construct the estimator $\hat{T}_n$ by setting

$$(2.8a)\qquad \hat{T}_n = \frac{\sum_{i=1}^{n} \frac{x_{i0}}{(x_{ii} v 1)^2} \quad Z_i}{\sum_{i=1}^{n} \left(\frac{x_{i0}}{x_{ii} v 1}\right)^2} \quad ,$$

where $x_{ii} v1 = \max\{x_{ii}, 1\}$.

To show that $E(\hat{T}_n - T(\underset{\sim}{\theta}))^2 \to 0$ for any $\underset{\sim}{\theta} \in \Theta$, it suffices to show that $E\hat{T}_n - T(\underset{\sim}{\theta}) \to 0$ and Var $\hat{T}_n \to 0$. Since Z_i are uncorrelated,

$$\text{Var } \hat{T}_n = \frac{\sum_{i=1}^{n} \frac{x_{i0}^2}{(x_{ii} v1)^4}}{\left(\sum_{i=1}^{n} \left(\frac{x_{i0}}{x_{ii} v1}\right)^2\right)^2} \leq \frac{1}{\sum_{i=1}^{n} \left(\frac{x_{i0}}{x_{ii} v1}\right)^2} \quad .$$

We now proceed to show that

$$(2.9)\qquad \sum_{i=1}^{n} \left(\frac{x_{i0}}{x_{ii} v1}\right)^2 \to \infty \quad \text{as } n \to \infty$$

in order to get Var $\hat{T}_n \to 0$.

Let $I_n = \{i | i \leq n, x_{ii} < 1\}$. Write

$$\sum_{i=1}^{n} \left(\frac{x_{i0}}{x_{ii} v1}\right)^2 = \sum_{i \in I_n} x_{i0}^2 + \sum_{i \notin I_n} \left(\frac{x_{i0}}{x_{ii}}\right)^2 \quad .$$

Suppose (2.9) does not hold. Then there exists some positive number M such that

(2.10) $$\sum_{i \in I_n} x_{i0}^2 \leq M,$$

and

(2.11) $$\sum_{i \notin I_n} \left(\frac{x_{i0}}{x_{ii}}\right)^2 \leq M,$$

for any n. Let $\underset{\sim}{\omega}^{(n)}=\underset{\sim}{\nu}-\sum_{i \notin I_n} \frac{x_{i0}}{x_{ii}} \underset{\sim}{e}_i$. Because of (2.11) it is clear that

$\underset{\sim}{\omega}^{(n)} \in \{\underset{\sim}{\theta} | T(\underset{\sim}{\theta})=1$ and $\underset{\sim}{\theta} \in B(1+M)\}$. It follows that

$$\inf \{\sum_{i=1}^{n} <\underset{\sim}{x}_i,\underset{\sim}{\theta}>^2 | \underset{\sim}{\theta} \in B(1+M),\ T(\underset{\sim}{\theta}) = 1\} \leq \sum_{i=1}^{n} <\underset{\sim}{x}_i,\underset{\sim}{\omega}^{(n)}>^2$$

$= ||L(\underset{\sim}{\omega}^{(n)} - \underset{\sim}{\nu}) + \underset{\sim}{A}^{(n)}||^2$ (by the definition of L and $\underset{\sim}{A}^{(n)}$, here $||\cdot||$ is the Euclidean norm in R^n)

$= \sum_{i=1}^{n} <\underset{\sim}{g}_i,\ L(\underset{\sim}{\omega}^{(n)} - \underset{\sim}{\nu}) + \underset{\sim}{A}^{(n)}>^2$ (since $\{\underset{\sim}{g}_i\}$ is an orthonormal basis)

$= \sum_{i=1}^{n} <\underset{\sim}{g}_i,\ -\sum_{j \notin I_n} \frac{x_{j0}}{x_{jj}} x_{jj}\underset{\sim}{g}_j + \underset{\sim}{A}^{(n)}>^2$ (Since $L(e_j)=x_{jj}\underset{\sim}{g}_j$)

$= \sum_{i \in I_n} <\underset{\sim}{g}_i,\ \underset{\sim}{A}^{(n)}>^2 + \sum_{i \notin I_n} (<\underset{\sim}{g}_i,\underset{\sim}{A}^{(n)}> - x_{i0})^2$

$= \sum_{i \in I_n} x_{i0}^2$ (recall the definition of x_{i0})

$\leq M < \infty$ (by (2.10))

This is contradictory to (2.5) for δ=M+1. Hence (2.9) holds and thus Var $\hat{T}_n \to 0$. It remains to show that $E\hat{T}_n - T(\underset{\sim}{\theta}) \to 0$ for any $\underset{\sim}{\theta} \in \Theta$.

For this purpose, it suffices to verify that

$$(2.12)\qquad \frac{\sum_{i=1}^{n} \frac{x_{i0}}{(x_{ii} \vee 1)^2}\; x_{ii}\; <\underset{\sim}{e}_i, \underset{\sim}{\theta}>}{\sum_{i=1}^{n} \left(\frac{x_{i0}}{x_{ii} \vee 1}\right)^2} \rightarrow 0, \text{ as } n \rightarrow \infty.$$

Now, by Cauchy-Schwartz inequality,

$$\sum_{i=1}^{n} \frac{x_{i0}}{(x_{ii} \vee 1)^2}\; x_{ii}\; <\underset{\sim}{e}_i, \underset{\sim}{\theta}>$$

$$\leq \left[\sum_{i=1}^{n} \left(\frac{x_{i0}}{x_{ii} \vee 1}\right)^2\right]^{1/2} \cdot \left[\sum_{i=1}^{n} \left(\frac{x_{ii}}{x_{ii} \vee 1}\right)^2 <\underset{\sim}{e}_i, \underset{\sim}{\theta}>^2\right]^{1/2}$$

$$\leq \left[\sum_{i=1}^{n} \left(\frac{x_{i0}}{x_{ii} \vee 1}\right)^2\right]^{1/2} \cdot ||\underset{\sim}{\theta}||.$$

Therefore by (2.9), we see that (2.12) holds and consequently $\hat{T}_n$ is consistent. The proof for "(iv) ⇒ (v)" is now complete. Finally, "(v) ⇒ (i)" holds obviously.

□

3. Discussion.

Several important features about Theorem 2.1 are now in order. First, "(ii) ⇒ (i)" means "pairwise consistency implies consistency", which certainly may not be a true statement in other context. In fact, the following example demonstrates what may happen without the structure given by (1.1).

Example 1. Suppose $\Theta=\{(\theta_1,\ldots,\theta_n,\ldots)|\sum_{i=1}^{\infty}\theta_i^2=\infty\}\cup\{(0,0,\ldots)\}$.

The observations y_i satisfy the model $y_i=\theta_i+\varepsilon_i, i=1,2,\ldots$, where ε_i are i.i.d. normal with mean 0 and variance σ^2. Suppose we want to estimate θ_1. Without much difficulty, it can be verified that when the parameter space is restricted to two points $\{(0,0,\ldots), (\theta_1^*,\ldots,\theta_n^*,\ldots)\}$, then consistent estimates exist. But, it is also clear that when the parameter space is the whole Θ, θ_1 is not consistently estimable.

One may also find that the following example illustrates the similar phenomenon but in a simpler situation where only one observation is made.

Example 2. Take $\Theta=R\cup\{\infty\}$. Assume that the observation Y is a normal random variable with mean 0 and variance 1 if $\theta=\infty$; otherwise, $Y=\theta$. Obviously, when restricting the parameter space to any pair of points (say θ_1, and $\theta_2\neq\infty$), a perfect estimator (without any error) exists; i.e., $\hat{\theta}=Y$ if $Y=\theta_2$ and $\hat{\theta}=\theta_1$ if $Y\neq\theta_2$. But when the parameter space is the whole Θ then no estimator can estimate θ without error.

The next important feature about Theorem 2.1 concerns the statement (iii). By the equivalent between (iii) and (i), the consistency problem (which is stochastic in nature and therefore is relatively complicated) can now be reduced to verifying the deterministic equation(2.4), which retains some intuitive interpretation as was already given in Wu (1980). Roughly speaking, when y_i is observed, not only the "information" along the direction $\underset{\sim}{x}_i$ is obtained, but also partial information can be gained along directions not orthogonal to $\underset{\sim}{x}_i$. A direction $\underset{\sim}{\theta}$ is called a good direction if (2.4) holds; otherwise, it is called a bad direction. A bad direction is a direction where the total "information" is finite. Therefore, we conclude that T is a c.e.b.l. functional if all the bad directions are orthogonal to T. A useful consequence is given in the following corollary.

Corollary 3.1. The set of all c.e.b.l. functional is a closed linear space.

Proof. This follows from the remark above and the fact that the orthogonal complement of any subset in a Hilbert space is closed. □

However, unlikely in the finite dimensional case, the set of all bad directions may not be a closed space. Consequently, the space of all $\underset{\sim}{\theta}$ such that $\sum_{i=1}^{\infty} <\underset{\sim}{x}_i,\underset{\sim}{\theta}>^2 < \infty$ may be only a dense subset of the orthogonal complement of the space of all c.e.b.l. functionals. Thus caution should always be taken when one wants to characterize the set of all c.e.b.l. functionals; see, for instance, Example 3 of Section 3.

Strictly speaking, in view of the equivalence between (ii) and (iii) (and the proof of (ii) $\Rightarrow$ (iii)) the reason why a $\underset{\sim}{\theta}$ satisfying (2.4) can be deemed as a "good"direction strongly depends on the condition of the finite Fisher's information of the error distribution. If the error distribution has an infinite Fisher's information, then (ii) may not imply (iii); a bad direction $\underset{\sim}{\theta}^*$ in the sense that $\sum_{i=1}^{\infty} <x_i, \underset{\sim}{\theta}^*>^2 < \infty$, may also be "good" enough so that we can asymptotically discriminate the distribution of $(y_1,\ldots,y_n)'$ under $\underset{\sim}{\theta}=0$ from that under $\underset{\sim}{\theta}=\underset{\sim}{\theta}^*$ perfectly. This is illustrated in the following example.

Example 3. Take $\Theta=R$ and $y_i=x_i\theta+\varepsilon_i$, where x_i is a real nonnegative number such that $\sum_{i=1}^{\infty} x_i=\infty$ and $\sum_{i=1}^{\infty} x_i^2<\infty$; ε_i, $i=1,2,\ldots$, are i.i.d. with the common distribution uniform on $[-1/2,1/2]$. It is clear that for any $\theta^*\neq 0$, the likelihood ratio of $(y_1,\ldots,y_n)'$ between θ^* and 0 is either 0, or $+\infty$, or 1; the probability measure (either when $\theta=0$ or when $\theta=\theta^*$) of the set of points for which the likelihood ratio equals 1 is at most $\prod_{i=1}^{n} (1-x_i\theta^*)_+$ where $(1-x_i\theta^*)_+=\text{Max}\{1-x_i\theta^*, 0\}$. From the condition that $\sum_{i=1}^{\infty} x_i=\infty$, it follows that $\prod_{i=1}^{n} (1-x_i\theta^*)_+\to 0$. Thus for $\theta=\theta^*\neq 0$, the distribution of $(y_1,\ldots,y_n)'$ is asymptotically singular to that for $\theta=0$. Hence (ii) of Theorem 2.1 holds but not (iii).

Note that in the proof of Theorem 2.1, the only part involving the assumption (2.3) is (ii) $\Rightarrow$ (iii). It follows that even if ε has an infinite Fisher's information (iii) still implies (i) (and of course (ii)). In particular, while a bad direction may become "good" because of infinite Fisher's information error, a good direction is always good no matter the error distribution has finite or infinite Fisher's information.

We now discuss the cases where (2.2) is violated and the observations may be dependent or correlated and may have unequal variances. First, we observe that the inpenendence assumption about observations is needed only when verifying "(ii) ⇒ (iii)". Thus, if the observations are dependent but uncorrelated with common error distribution and satisfy (2.2), then it still holds that (iii) ⇒ (iv) ⇒ (v) ⇒ (i) ⇒ (ii). Thus (iii) is a sufficient condition for T to be a consistent estimator in most situations. Next, suppose the covariances of the observations are known up to a constant. Denote the covariance matrix of the first n observations by V_n. Let A_n be the lower triangular matrix such that $A_n V_n A_n' = I_{n\times n}$. Now, transform the original data $(y_1,\ldots,y_n)'$ to $(z_1,\ldots,z_n)' = A_n(y_1,\ldots,y_n)'$. For the new data, the observations are now homosedastic and uncorrelated. Let $\underset{\sim}{A}_i' = (A_{i1},\ldots,A_{ii},0,\ldots)$ be the i-th row of A_n. The regression model for z_i becomes

$$z_i = <\sum_{i=1}^{i} A_{ij}\underset{\sim}{x}_j, \underset{\sim}{\theta}> + \varepsilon_i'.$$

(Note that since A_{n-1} is the left-upper submatrix of A_n, z_i should be independent of n.) Thus writing $\underset{\sim}{x}_i^* = \sum_{j=1}^{i} A_{ij}\underset{\sim}{x}_j$, we can establish "(iii) ⇒ (iv) ⇒ (v) ⇒ (i) ⇒ (iii)" after substituting $\underset{\sim}{x}_i$ by $\underset{\sim}{x}_i^*$ in the Theorem 2.1. Moreover, if $\{\varepsilon_i\}$ is Gaussian, then the ε_i, i=1,2,... are independent; hence "(ii) ⇒ (iii)" holds and the analogue of Theorem 2.1 is now established. However, to what extent the condition about the existence of the second moment of the error distribution can be released is not clear to the present author yet.

A comment about the consistent estimators constructed by the method used in the proof of "(iv) ⇒ (v)" is given below. Examining the proof carefully it is not hard to see that instead of using the estimator of (2.8a), the following types of estimators also work:

(2.8b)

$$\hat{T}_n = \frac{\sum_{i=1}^{n} \frac{x_{i0}}{(x_{ii} \vee \lambda)^2} Z_i}{\sum_{i=1}^{n} \left(\frac{x_{i0}}{x_{ii} \vee \lambda}\right)^2} \quad , \text{ or}$$

(2.8c)

$$\hat{T}_n = \frac{\sum_{i=1}^{n} \frac{x_{i0}}{\lambda^2 + x_{ii}^2} Z_i}{\sum_{i=1}^{n} \frac{x_{i0}^2}{\lambda^2 + x_{ii}^2}} \quad ,$$

where λ is any fixed positive number. The role of λ here is similar to the role of the ridge constant in the ridge regression or the role of a smoothing parameter in any illy-posed problem; it controls the tradeoff between the variance and the bias. Therefore, one might expect that an adaptive choice of λ will be more useful in practice. This should be investigated in the future. Also, for other commonly-used estimation procedures such as the smoothing spline method in the non-parametric regression setting of Section 4, their consistency property should also be examined under the general framework discussed here.

The equivalence between (v) and (i) is also interesting. Without the specific setup, particularly the conditions (2.2) and (2.3), (i) generally does not imply (v).

Finally, let's consider the finite dimensional case (i.e., $\Theta = R^p$), and discuss the connection with Wu's work. By the natural confinement to the least squares estimates as usual, Wu showed that T(·) is a consistent direction if and only (iii) holds (although the nonsingularity of the information matrix is assumed there, it is removable if one works with the generalized inverse instead). Thus combining Wu's result with Theorem 2.1 here, we obtain the following corollary.

Corollary 3.2. Suppose Θ is finite dimensional. Then T(·) is a c.e.b.l. functional if and only if T(·) is a consistent direction for the least squares estimates.

The important consequence of this corollary is that when the least squares method fails to provide a consistent estimator, no other types of estimators (for instance, ridge estimators with adaptive or nonadaptive choices of ridge constants or Stein estimator or alike) can be consistent. While this fact does not discredit these alternative estimators, it does point out one of the difficulties for such illy-posed problems and thus more attention should be given before drawing the conclusion; even if the sample size is very large, the inference error may still be sizable.

Before closing this section, let's comment on the estimability of the linear combinations of parameters as defined in Scheffé (1958); i.e., a linear combination of parameters is estimable if there exists an unbiased estimator. In the finite dimensional case, it is true that if T(·) is consistently estimable then T(·) is estimable. But this is not necessarily the case in the infinite dimensional situation. T(·), when represented as an element in the Hilbert space concerned, can be outside of the linear space generated by $\{\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n\}$ for any n but still retain the property of consistent estimability. This will become clearer when we consider the nonparametric regression setting of the next section.

4. Nonparametric regression.

In this section, the following nonparametric regression problem will be considered. For $m \geq 1$ (m is considered as a fixed integer hereafter), let $W_2^m [0,1]=\{f | f^{(m-1)}$ is absolutely continuous on [0,1] and $\int_0^1 f^{(m)}(t)^2 dt < \infty$. $W_2^m[0,1]$ is a seperable Hilbert space when equipped with the inner product $<f,g>= \sum_{i=0}^{m} \int_0^1 f_{(t)}^{(i)} g_{(t)}^{(i)} dt$. (It should be clear from the context whether $<\cdot,\cdot>$ is the inner product of $W_2^m[0,1]$ or the inner product of ℓ^2). Suppose a sequence of points in [0,1], $\{t_1,t_2,\ldots\}$, is given. We observe the y_i, $i=1,2,\ldots$, which follow the nonparametric regression model:

$$(4.1) \qquad y_i = f(t_i) + \varepsilon_i,$$

where $f \in W_2^m[0,1]$ and ε_i are i.i.d. with a common distribution satisfying the conditions (2.2) and (2.3). Our goal is then to characterize the set of all c.e.b.l. functionals on $W_2^m[0,1]$.

Let $\{f_1,f_2,\ldots\}$ be a complete orthonormal system of $W_2^m[0,1]$. Any f in $W_2^m[0,1]$ can be represented as $\sum_{j=1}^{\infty} <f,f_j>f_j$. In particular, for any $t \in [0,1]$, the bounded linear functional $D_t^{(0)}$, defined by $<D_t^{(0)}, f>=f(t)$, can be written as $\sum_{j=1}^{\infty} <D_t^{(0)},f_j>f_j$. Take $\theta_j=<f,f_j>$ and $x_j=<D_t^{(0)},f_j>$. Denote $\underset{\sim}{\theta}=(\theta_1,\theta_2,\ldots)'$ and $\underset{\sim}{x}=(x_1,x_2,\ldots)'$. When any observation y is made at the point t we can rewrite (4.1) as

$$\begin{aligned} y = f(t) + \varepsilon &= <D_t^{(0)},f> + \varepsilon \\ &= \sum_{j=1}^{\infty} x_j\theta_j + \varepsilon \\ &= <\underset{\sim}{x},\underset{\sim}{\theta}> + \varepsilon \quad . \end{aligned}$$

Hence our setup is indeed a special case of (1.1) with $\Theta = \ell^2$). We may apply Theorem 2.1 to derive the desired results as follows.

<u>Lemma 4.1</u>. $T(\cdot)$ is a c.e.b.l. functional if and only if

(4.2) $$\sum_{i=1}^{\infty} f(t_i)^2 = \infty, \text{ for any } f \text{ in } W_2^m[0,1] \text{ such that } T(f) \neq 0.$$

Proof.

Represent f and $D_{t_i}^{(0)}$ as $\underset{\sim}{\theta}$ and $\underset{\sim}{x}_i$. It follows that $<\underset{\sim}{x}_i, \underset{\sim}{\theta}>=f(t_i)$. Therefore Theorem 2.1 applies. (4.2) follows from (2.4). □

Now let us consider an important class of bounded linear functionals on $W_2^m[0,1]$, namely the differential functional $D_t^{(k)}$, which maps any f in $W_2^m[0,1]$ to its k-th derivative at the point t, $f^{(k)}(t)$. Note that $D_t^{(k)}$ is a bounded linear functional only when $0 \leq k \leq m-1$. To characterize the set of all $D_t^{(k)}$ which are consistently estimable we shall, equivalently, determine the <u>consistent region of degree k</u>, defined to be the set

(4.3) $$C_k = \{t \mid t \in [0,1] \text{ and } D_t^{(k)} \text{ is consistently estimable}\}.$$

(This notion is obviously an extension of the definition of the consistent region given in Wu (1980).) Obviously, the limiting behavior of $\{t_i\}$ is crucial here. For any nonnegative integer k, we call a point t* in

[0,1] a <u>limiting point of degree k</u> for the sequence $\{t_i\}$, if

(4.4) there exists a subsequence $\{t_i'\}$ of $\{t_i\}$ such that $t_i' \to t^*$ as $i \to \infty$ and $\sum_{i=1}^{\infty} (t_i' - t^*)^{2k} = +\infty$.

When k=0, this notion is exactly the usual definition of the limiting points of a sequence of real numbers. Note that by our definition here if a point is a limiting point of degree k, then it is also a limiting point of degree less than k. Other useful properties about limiting points are described in the following lemma. Note that the topology considered here is restricted to [0,1]; e.g., (1/2,1] is an open set, etc.

<u>Lemma 4.2.</u> A point t* is a limiting point of degree k if and only if

(4.5) $$\sum_{t_i' \in N}^{\infty} (t_i - t^*)^{2k} = \infty, \text{ for any open neighborhood } N \text{ of } t^*.$$

In particular, the set of all limiting points of degree k is a closed set, and the set of all limiting points which is of degree 0 but is not of degree k is a discrete set.

Proof.

Obviously, (4.4) implies (4.5). We now show (4.5) implies (4.4). There are two cases.

Case (i), let us assume that there exists an infinite sequence of distinct limiting points of degree 0,$\{t^*_i\}$,such that $t^*_i \to t^*$. Then, without much difficulty, one can select a subsequence $\{t'_i\}$ of $\{t_i\}$ such that $t'_i \to t^*$ and the first n_1 items of $\{t'_i\}$ are in a small neighborhood of t^*_1, the next n_2 items of $\{t'_i\}$ are in another small neighborhood of t^*_2,..., etc. The ranges of the neighborhoods should tend to 0 to insure that $t'_i \to t^*$. The numbers $n_1, n_2, \ldots,$ should be large enough to guarantee that $\sum_{i=1}^{n_1}(t'_i-t^*)^{2k}\geq 1$, $\sum_{i=n_1+1}^{n_1+n_2}(t'_i-t^*)^{2k}\geq 1,\ldots$ etc. Thus $\sum_{i=1}^{\infty}(t'_i-t^*)^{2k}=\infty$ and (4.4) holds consequently.

Case (ii), there exists a neighborhood N of t^* such that there is no limiting point of degree 0 in $N-\{t^*\}$. Now, let $\{t'_i\}=N\cap\{t_1,t_2,\ldots\}$. By (4.5), it is obvious that $t'_i \to t^*$ and $\sum_{i=1}^{\infty}(t'_i-t^*)^{2k}=+\infty$. Thus (4.4) holds.

The other statements of the lemma are easy to establish; the proofs are omitted. □

The consistent region of degree k is now characterized below.

Theorem 4.1. For any integer k such that $0\leq k\leq m-1$, the consistent region of degree k, C_k, consists of all limiting points of degree k.

By Lemma 4.2 and this theorem we obtain some useful properties of C_k.

Corollary 4.1. C_k is compact, $C_k \supset C_{k+1}$, and C_0-C_{m-1} is discrete (and is therefore countable).

Proof of Theorem 4.1.

First, we show that for any limiting point t of degree k, $D_t^{(k)}$ is a c.e.b.l. functional. By Lemma 4.1, it suffices to show that for any f such that $f^{(k)}(t^*)\neq 0$, we have $\sum_{i=1}^{\infty} f(t_i)^2=\infty$. Let γ be the smallest nonnegative integer such that $f^{(\gamma)}(t^*)\neq 0$. Because of the continuity of $f^{(\gamma)}$, it is easy to see that $|f(t)|\geq\frac{1}{2}|f^{(\gamma)}(t^*)(t-t^*)|^{\gamma}$ for any t in an open neighborhood N of t*. Since $\gamma\leq k$ and t* is a limiting point of degree k, it follows that t is also a limiting point of degree γ. Apply Lemma 4.2 (taking k to be γ) and we conclude that

$$\sum_{i=1}^{\infty} f(t_i)^2 \geq \sum_{t_i\in N} f(t_i)^2 \geq \frac{1}{2} f^{(\gamma)}(t^*)^2 \sum_{t_i\in N} (t_i - t^*)^{2\gamma} = \infty.$$

(4.2) is now established and $D_{t^*}^{(k)}$ is therefore a c.e.b.l. functional.

Next, for any t* which is not a limiting point of degree k, we shall demonstrate that there exists a f in $W_2^m[0,1]$ such that $f^{(k)}(t^*)\neq 0$ and $\sum_{i=1}^{\infty} f(t_i)^2<\infty$; this then implies that $D_{t^*}^{(k)}$ is not a c.e.b.l. functional by Lemma 4.1. By Lemma 4.2, let N be an open neighborhood of t* such that $\sum_{t_i\in N}(t_i-t^*)^{2k}<\infty$. Construct a function f in $W_2^m[0,1]$ such that $f(t)=0$ for any $t\notin N$, $f^{(\gamma)}(t^*)=0$ for any $\gamma<k$, and $f^{(k)}(t^*)\neq 0$. It follows that $\sum_{i=1}^{\infty} f(t_i)^2=\sum_{t_i\in N} f(t_i)^2\leq\sum_{t_i\in N} M\cdot(t_i-t^*)^{2k}<\infty$, where $M\geq\sup\{f^{(k)}(t)^2 \mid t\in[0,1]\}$. Therefore $D_{t^*}^{(k)}$ is not a c.e.b.l. functional, and the proof is complete.□

By Theorem 4.1, we may easily see whether $D_t^{(k)}$ is a c.e.b.l. functional or not. We call $D_t^{(k)}$ a <u>c.e.b.l. functional of differential type</u> if $t\in C_k$. An application of Corollary 3.1 then shows that any element in the closed linear space generated by all c.e.b.l. functionals of differential type is also a c.e.b.l. functional. Naturally, we would like to know if there are any other c.e.b.l. functionals or not. The following example gives some clues to the answer. It also demonstrates that the space of all $\underset{\sim}{\theta}$ such that (2.4) does not hold may not be closed as was already pointed out in Section 3.

Example 3. Suppose $t_i \to t^*$ as $i \to \infty$ and $\sum_{i=1}^{\infty} (t_i - t^*)^{2m} = \infty$. By Theorem 4.1, $D_{t^*}^{(k)}$, $k=0,1,\ldots,m-1$, are the only c.e.b.l. functionals of differential type. We now show that the linear space generated by $D_{t^*}^{(k)}$, $k=0,1,\ldots,m-1$, is exactly the set of all c.e.b.l. functionals.

Consider an f in $W_2^m[0,1]$ which is orthogonal to $D_{t^*}^{(k)}$, $k=0,1,\ldots,m-1$; i.e., $f^{(k)}(t^*)=0$, for $0 \le k \le m-1$. To show that $\langle f,\cdot\rangle$ is not a c.e.b.l. functional we have to find a g such that $\langle f,g\rangle \neq 0$ and $\sum_{i=1}^{\infty} g(t_i)^2 < \infty$ (by Lemma 4.1). By some trivial argument it can be shown that for any $\varepsilon > 0$, there exists a g which equals 0 in a small open interval I containing t^* such that $||f-g|| < \varepsilon$. Take ε small enough to insure that $\langle f,g\rangle \neq 0$. It follows that $\sum_{i=1}^{\infty} g(t_i)^2 = \sum_{i=1}^{n} g(t_i)^2 < \infty$, where n is an integer such that $t_i \in I$ for $i > n$. Thus the desired result is established.

Note that in this example the set of all g such that $\sum_{i=1}^{\infty} g(t_i)^2 < \infty$ is not closed. To see this, consider the function $g_0(t) = (t-t^*)^m$. Obviously, $g_0^{(0)}(t^*) = \ldots = g_0^{(m-1)}(t^*) = 0$ and $\sum_{i=1}^{\infty} g_0(t_i)^2 = \infty$. However, if the set of all g such that $\sum_{i=1}^{\infty} g(t_i)^2 < \infty$ were closed, then this set would equal the orthogonal complement of the space of all c.e.b.l. functionals (i.e., $\{g|g^{(0)}(t^*)=\ldots=g^{(m-1)}(t^*)=0\} = \{g|\sum_{i=1}^{\infty} g(t_i)^2<\infty\}$). Thus a contradiction is obtained because g_0 can't be in both sets.

The argument used in the Example 3 can be generalized to show that the set of all c.e.b.l. functionals equals the closed space generated by all c.e.b.l. functionals of differential type when the consistent region of degree less than m-1 is a finite set (but the cardinality of the consistent region of degree m-1 may be infinite). Consider f in $W_2^m[0,1]$ such that $f^{(0)}(t)=\ldots=f^{(m-1)}(t)=0$ for all $t\in C_{m-1}$ and $f^{(k)}(t)=0$ for all $t\in C_k$, $k=0,1,\ldots,m-2$. Since C_{m-1} is compact and $\bigcup_{k=0}^{m-2}$ is a finite set, without much difficulty we can construct a function g in $W_2^m[0,1]$ such that

(4.6) $g(t) = 0$ for t in a union of finitely many open intervals covering C_{m-1}; $g^{(k)}(t) = 0$ for $t\in C_k$, $k=0,1,\ldots,m-2$; and $||f-g||<\varepsilon$ for $\varepsilon>0$.

It follows that for ε small enough, we have $<f,g>\neq0$ and $\sum_{i=1}^{\infty} g(t_i)^2<\infty$. Thus f is not a c.e.b.l. functional and the desired result is obtained.

However, for the case that $\bigcup_{k=0}^{m-2} C_k$ is not a finite set, it is still not clear to the present author whether a g satisfying (4.6) exists or not. Thus it remains unknown whether the set of all c.e.b.l. functionals equals the closed linear space generated by all c.e.b.l. functionals of differential type or not.

Finally, we draw the connection between our results here and those obtained by Wu (1980) in the consideration of the polynomial regression model. Suppose the model is $y_j=f(t_j)+\varepsilon_j=\sum_{i=0}^{m-1}\theta_i t_j^i+\varepsilon_j$, where ε_j are independent random errors satisfying conditions (2.2) and (2.3) and $t_j\in[0,1]$. Wu's characterization on the set of all consistent directions for the least squares estimates can be summarized as below.

(i) The consistent region (of degree 0) contains (but may not equal) all the limiting points (of degree 0).

(ii) The consistent region of degree k ($1\le k\le m-1$) contains (but may not equal) all the limiting points of degree k.

(iii) The set of all consistent directions equals the linear space generated by the consistent directions obtained by (i) and (ii).

(iv) From (iii), it is easy to identify the consistent region (of degree k): it equals [0,1] if the dimension of the set of all consistent directions is m; otherwise it equals the set of all limiting points of degree k.

(Note that if the sequence $\{t_i\}$ is unbounded, then it was shown in Proposition 3 of Wu (1980) that the characterization problem is easy to solve. We thus omit the discussion for this case.)

We leave the verification of the above statements to the readers (who have read Wu's paper) while simply reminding them of the following (and other similar) fact:

$$\text{"}\sum_{i=1}^{\infty}(t_i-t_1^*)^{2k}(t_i-t_2^*)^{2n}=\infty\text{" if and only if "}\sum_{i=1}^{\infty}(t_i-t_1^*)^{2k}=\infty$$

$$\text{or }\sum_{i=1}^{\infty}(t_i-t_2^*)^{2n}=\infty\text{",}$$

where t_1^* and t_2^* are two distinct limiting points of $\{t_i\}_{i=1}^{\infty}$.

Now, the similarity between the polynomial model and the nonparametric regression setup considered here becomes transparent. The only difference lies in the dimensions of the parameter space. In polynomial regression, as a consequence of the finite dimension property, some $f^{(k)}(t)$, where t is not a limiting point of degree k, may become a consistent direction simply because it is a linear combination of other consistent directions. But this is not the case for the nonparametric setting. Thus the consistent region of degree k in the polynomial case may be larger than C_k.

Appendix. A property about finite Fisher's information.

Suppose ε, $\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n,\ldots$, are i.i.d. random variables satisfying (2.2) and (2.3). Let $\{h_1,h_2,\ldots,\}$ be an infinite sequence of real numbers such that $\sum_{i=1}^{\infty} h_i^2<\infty$. Let $\underset{\sim}{P}_n$ be the probability measure of $(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n)'$ and $\underset{\sim}{Q}_n$ be the probability measure of $(\varepsilon_1+h_1,\varepsilon_2+h_2,\ldots,\varepsilon_n+h_n)'$. Consider the simple against simple hypothesis testing problem $H_n:\underset{\sim}{P}_n$ against $\underset{\sim}{Q}_n$, based on the observation $(y_1,\ldots,y_n)'$, where $y_i=\varepsilon_i$ under $\underset{\sim}{P}_n$ and $y_i=\varepsilon_i+h_i$ under $\underset{\sim}{Q}_n$. Let A_n be any measurable acceptance region in R^n. The following theorem claims that there does not exist a sequence of regions with an asymptotic power equal to 1. In other words, $\underset{\sim}{P}_n$ and $\underset{\sim}{Q}_n$ are not asymptotically mutually singular.

Theorem. For any measurable set $A_n\subset R^n$, we have

$$"\underset{\sim}{P}_n(A_n) \to 1 \text{ as } n\to\infty" \text{ implies that } "\underset{\sim}{Q}_n(A_n) \not\to 0 \text{ as } n\to\infty".$$

To prove the theorem, we first observe that by Neyman-Pearson's Lemma, we may consider the following special type of acceptance region

$$A_n=\{(y_1,\ldots,y_n)' \,|\, \frac{q_n(y_1,\ldots,y_n)}{p_n(y_1,\ldots,y_n)} \le C_n\}$$ where p_n and q_n are the density (with respect to Lebesque measure) of $\underset{\sim}{P}_n$ and $\underset{\sim}{Q}_n$, and C_n is a nonnegative real number. Suppose we can demonstrate that

(A.1) $$\underset{n\to\infty}{\underline{\lim}}\ \underset{\sim}{P}_n \left\{ \frac{q_n(y_1,\ldots,y_n)}{p_n(y_1,\ldots,y_n)} \leq \delta \right\} \leq 1-\alpha, \text{ for some } \alpha,\ \delta>0.$$

Then it follows that to have "$\underset{\sim}{P}\ (A_n)\to 1$" it is necessary that $\underset{n\to\infty}{\overline{\lim}}\ C_n>\delta$. Now,

$$\underset{\sim}{Q}_n(A_n) = \int_{A_n} q_n(y_1,\ldots,y_n)dy_1,\ldots,dy_n$$

$$\geq \int_{c_n \geq \frac{q_n(y_1,\ldots,y_n)}{p_n(y_1,\ldots,y_n)} \geq \delta} \frac{q_n(y_1,\ldots,y_n)}{p_n(y_1,\ldots,y_n)}\ p_n(y_1,\ldots,y_n)dy_1,\ldots,dy_n$$

$$\geq \left(\underset{\sim}{P}(A_n) - \underset{\sim}{P}_n \left\{ \frac{q_n(y_1,\ldots,y_n)}{p_n(y_1,\ldots,y_n)} \leq \delta \right\}\right) \cdot\ \delta.$$

Hence $\underset{n\to\infty}{\overline{\lim}}\ \underset{\sim}{Q}_n(A_n)\geq\delta(1-(1-\alpha))=\delta\alpha>0$ and the theorem holds.

To show that (A.1) holds, we prove the following proposition which obviously is equivalent to (A.1).

Proposition. If $\sum_{i=1}^{\infty} h_i^2<\infty$, then

$$\sum_{i=1}^{n} \log f(\varepsilon_i-h_i) - \log f(\varepsilon_i) \not\to -\infty \text{ in probability, as } n\to\infty$$

where f is the density of ε.

Note that because the ε_i's are independent, the convergence in probability is equivalent to the almost sure convergence.

To establish this proposition, we shall apply a powerful tool devised by LeCam (1960) in the proof of his second lemma on the contiguity theory (see also Hajek (1967)); i.e., we shall approximate the random variable $\log \frac{f(\varepsilon_i - h_i)}{f(\varepsilon_i)}$ by a quadratic function of the random variable $\sqrt{\frac{f(\varepsilon_i - h_i)}{f(\varepsilon_i)}} - 1$, which always has a finite variance. The reader may refer to Hajek (1967) for the idea about the proof of some lemmas below.

Write $s(y) = \sqrt{f(y)}$ and $I(f) = \int_{-\infty}^{\infty} \frac{f'(y)^2}{f(y)} dy = 4\int_{-\infty}^{\infty} s'(y)^2 dy$

Lemma 1. $\int_{-\infty}^{\infty} \{[s(y-h) - s(y)]/h\}^2 dy \leq \frac{1}{4} I(f)$.

Proof.

See Hajek (1967), page 212, (12). □

Lemma 2. $E\left(\frac{s(\varepsilon - h)}{s(\varepsilon)} - 1\right)^2 \leq \frac{h^2}{4} I(f)$.

Proof.

$$E\left(\frac{s(\varepsilon-h)}{s(\varepsilon)} - 1\right)^2 = h^2 \int_{-\infty}^{\infty} \left\{\frac{s(y-h) - s(y)}{h}\right\}^2 dy \leq \frac{h^2}{4} I(f) \quad \text{(by Lemma 1).}\square$$

Lemma 3. For any $\delta > 0$,

$$\sum_{i=1}^{\infty} P\left(\left|\frac{s(\varepsilon_i - h_i)}{s(\varepsilon_i)} - 1\right| \geq \delta\right) < \infty$$

Proof.

By Chebychev's inequality, we have

$$\sum_{i=1}^{\infty} P\left(\left|\frac{s(\varepsilon_i - h_i)}{s(\varepsilon_i)} - 1\right| \geq \delta\right)$$

$$\leq \frac{1}{\delta^2} \sum_{i=1}^{\infty} E\left(\frac{s(\varepsilon_i - h_i)}{s(\varepsilon_i)} - 1\right)^2$$

$$\leq \frac{1}{4\delta^2} \sum_{i=1}^{\infty} h_i^2 \cdot I(f) \quad \text{(by Lemma 2)}$$

$$< \infty \quad . \qquad \square$$

Lemma 4. $\sum_{i=1}^{\infty}\left(\frac{s(\varepsilon_i-h_i)}{s(\varepsilon_i)} - 1\right) \not\to -\infty$ in probability.

Proof.

By Chebychev's inequality, it suffices to show that the first moment is bounded below and the variance is bounded above.

Now, $$\sum_{i=1}^{\infty} E\left(\frac{s(\varepsilon_i-h_i)}{s(\varepsilon_i)} - 1\right)$$

$$= \sum_{i=1}^{\infty}\left\{\int_{-\infty}^{\infty} s(y-h_i)\,s(y)\,dy - 1\right\}$$

$$= -\tfrac{1}{2}\sum_{i=1}^{\infty} E\left(\frac{s(\varepsilon_i-h_i)}{s(\varepsilon_i)} - 1\right)^2$$

$$\geq -\frac{1}{8}\sum_{i=1}^{\infty} h_i^2 \cdot I(f) > -\infty.$$

Next, $$\sum_{i=1}^{\infty} \mathrm{Var}\left(\frac{s(\varepsilon_i-h_i)}{s(\varepsilon_i)} - 1\right)$$

$$\leq \sum_{i=1}^{\infty} E\left(\frac{s(\varepsilon_i-h_i)}{s(\varepsilon_i)} - 1\right)^2$$

$$\leq \tfrac{1}{4}\sum_{i=1}^{\infty} h_i^2 \cdot I(f) \quad \text{(by Lemma 2).}$$

$$< \infty \,.$$

Thus, Lemma 4 is proved. □

Proof of Proposition.

By (7), (8), (9) on page 206, Hajek (1967), we obtain

$$\sum_{i=1}^{n} \{\log f(\varepsilon_i - h_i) - \log f(\varepsilon_i)\}$$

$$= 2\sum_{i=1}^{n} \left\{\frac{s(\varepsilon_i - h_i)}{s(\varepsilon_i)} - 1\right\} - \sum_{i=1}^{n} \left(\frac{s(\varepsilon_i - h_i)}{s(\varepsilon_i)} - 1\right)^2 \int_0^1 \left\{2(1-\lambda)/ [1+\lambda(s(\varepsilon_i - h_i)/s(\varepsilon_i) - 1)]\right\} d\lambda.$$

$$\equiv (I) - (II)$$

By Lemma 4, the first term (I) does not tend to $-\infty$. Therefore, it suffices to show that the second term (II) does not tend to ∞.

Let $T_i^\delta = \text{Min}\left\{\left|\frac{s(\varepsilon_i - h_i)}{s(\varepsilon_i)} - 1\right|, \delta\right\}$, for sufficiently small $\delta > 0$.

Then, by Lemma 3 and Borel-Cantelli's lemma, we need only to verify that

$$\sum_{i=1}^{n} (T_i^\delta)^2 \cdot \int_0^1 \{2(1-\lambda)/(1-\lambda\delta)\} d\lambda \not\to \infty,$$

or, equivalently,

(A.2)
$$\sum_{i=1}^{n} (T_i^\delta)^2 \not\to \infty.$$

Now, since $\sum_{i=1}^{n} E(T_i^\delta)^2 \leq \sum_{i=1}^{n} E\left(\frac{s(\varepsilon_i - h_i)}{s(\varepsilon_i)} - 1\right)^2 \leq ¼ \sum_{i=1}^{n} h_i^2 \cdot I(f)$, (A.2) follows easily. Hence the proof is complete. □

References

Bickel, P.J. and Freedman, D.A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196-1217.

Drygas, H. (1976). Weak and strong consistency of the least squares estimates in regression models. Z. Wahrsheinlichkeitstheorie und Verw. Gebiete. **34** 119-127.

Freedman, D.A. (1981). Bootstrapping regression models. *Ann. Statist.* **9** 1218-1228.

Hajek, J. and Sidak, Z. (1967). *Theory of Rank Tests*. Academic Press, New York.

LeCam, L. (1960). Locally asymptotically normal families of distributions. Univ. of Calif. Publ. in Stat. 3, 37-98.

Roydon, H.L. (1972). *Real Analysis*. The Macmillan Company, New York.

Scheffé, H. (1958). *The Analysis of Variance*. Wiley, New York.

Stone, C.J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595-645.

Wu, C.F. (1980). Characterizing the consistent directions of least squares estimates. *Ann. Statist.* **8** 789-801.