UNIVERSITY OF WISCONSIN-MADISON

SSRI

Workshop Series

ADAPTIVE ESTIMATION OF NON-LINEAR REGRESSION MODELS

Charles F. Manski

8312

SOCIAL SYSTEMS RESEARCH INSTITUTE

Social Systems Research Institute

University of Wisconsin-Madison

ADAPTIVE ESTIMATION OF NON-LINEAR REGRESSION MODELS

Charles F. Manski

8312

August 1983

ABSTRACT

This paper summarizes from an econometric perspective recent work by statisticians on adaptive estimation. It also presents new findings concerning the adaptive estimability of non-linear regression models.

The development of methods for efficient estimation of structural parameters given specified prior distributional information is an ongoing, central theme of econometric theory. At the same time, one would generally like to avoid estimators whose good properties depend on the validity of strong, unsupportable assumptions about the behavior of unobservables. The two objectives of precise estimation and unrestrictive distributional specification must ultimately be in conflict. One should not think, however, that a weakening of distributional assumptions always carries with it a loss in attainable precision.

In particular, consider a non-linear regression model with a free intercept parameter and errors known to be independent and identically distributed (i.i.d.). In this paper, it will be proved under standard regularity conditions that there exists an estimator of the model's slope parameters whose computation does not involve any knowledge of the true error density yet whose asymptotic distribution is identical to that of the most efficient estimator that could be computed were the true error density given. That is, knowledge of the error density turns out to be irrelevant, asymptotically, to estimation of the slope parameters.

The above and other important asymptotic results relating attainable precision of estimation to prior distributional information can be obtained as applications of recent work by statisticians on adaptive estimation. An estimator may be termed adaptive if its computation incorporates a data based procedure for learning unknown features of the error distribution and if such learning is asymptotically successful in the sense that the asymptotic distribution of the estimator is that of the most efficient estimator that could be computed if the distribution were known.

The literature exploring contexts in which adaptive estimation is possible and proposing specific adaptive estimators is now at least thirty years old. It was apparently Stein (1956) who first sought to characterize the situations in which a parameter is and is not adaptively estimable. The achievement

of results allowing analysis of adaptive estimation in models of econometric interest has occurred only recently, however, in seminal work of Bickel (1982), who builds on earlier important work of Stone (1975).

In Sections 1, 3, and 5, I summarize from an econometric perspective the statistical literature on adaptive estimation. In Sections 2, 4, 6, and 7, I present my own findings concerning the adaptive estimability of non-linear regression and more general non-linear models.

1. Adaptive Estimation: An Informal Historical Perspective

The statistical literature on adaptive estimation is enormously rewarding to study, both for the powerful theorems it offers and for the technical virtuosity it displays. The formal arguments in this literature tend to be rather intricate. The ideas, on the other hand, are easy to understand and intuitive. An overview of those developments most relevant to our applications will serve as a useful prelude to the more formal analysis that follows.

Stein (1956) originally posed the problem of adaptive estimability as a non-parametric generalization of the classical question concerning the asymptotic relevance of nuisance parameters. Let a sample of observations be drawn randomly from a distribution known to be a member of a family of distributions characterized by the finite parameter vector $(\theta^*,\eta^*)\varepsilon\Theta xH$, where $\Theta \subset R^M$ and $H \subset R^L$. Let θ^* be the parameters of interest and η^* be the nuisance parameters. Let $I(\theta^*)$ be the Fisher's information matrix associated with estimation of θ^* given knowledge of η^* and let $I(\theta^*,\eta^*)$ be the information matrix associated with joint estimation of (θ^*,η^*). As is well known, standard regularity conditions imply that given η^*, the best attainable asymptotic variance for an asymptotically normal estimator of θ^* is $I(\theta^*)^{-1}$. If η^* is unknown, the best attainable precision is the upper left MxM sub-matrix of $I(\theta^*,\eta^*)^{-1}$. The latter matrix exceeds the former one by a non-negative definite matrix which is null if the upper right MxL sub-matrix of $I(\theta^*,\eta^*)$ is null. That is, if and only if this last condition holds is knowledge of η^* asymptotically irrelevant to estimation of θ^*.

Now consider the more general situation in which the distribution generating the observations is known only to be a member of a family of distributions characterized by $(\theta^*,f^*)\varepsilon R^M \times \Phi$, where Φ is a function space. Then f^* is a nuisance function. For example, in a non-linear regression model with i.i.d. disturbances, f^* might be the unknown error density and Φ the set of all symmetric density functions centered on zero. Stein reasoned that in this context, the asymptotic variance of an estimate for θ^* must be at least as large as the best asymptotic variance that would be attainable were f^* known to lie in some finite parametric family $(f^\eta,\eta\varepsilon H)\subset\Phi$. This, we have noted, is the MxM upper left sub-matrix of $I(\theta^*,\eta^*)^{-1}$, where η^* indexes the true density. Stein observed that the above inequality must hold for every finite dimensional subfamily of Φ containing f^*. He then concluded that given the prior restriction of f^* to Φ, θ^* cannot be estimated adaptively if there exists any finite parametric family such that $f^*\varepsilon(f^\eta,\eta\varepsilon R^L)\subset\Phi$ and the upper right MxL sub-matrix of $I(\theta^*,\eta^*)$ is non-null. This condition, which we shall later state formally as Condition S, is conceptually simple. Unfortunately, Stein did not indicate how it might be checked in practice. A verifiable form of the condition became available only recently, in Bickel (1982). See also Begun, Hall, Huang, and Wellner (1983).

Stein's paper does contain one immediately useful result. Considering the classical problem of estimating θ^* in the presence of the nuisance parameters η^*, Stein observed that there can exist situations in which knowledge of η^* is asymptotically relevant to estimation of θ^* as a whole but irrelevant to estimation of some sub-vector of θ^*. Let $\theta^* = (\beta^*,\alpha^*)$ where $\beta^*\varepsilon R^U$, $\alpha^*\varepsilon R^V$, $U + V = M$. Given knowledge of η^*, the best attainable asymptotic variance for an estimate of β^* is the upper left UxU sub-matrix of $I(\beta^*,\alpha^*)^{-1}$. In the absence of such knowledge, it is the corresponding sub-matrix of $I(\beta^*,\alpha^*,\eta^*)^{-1}$. Stein found the necessary and sufficient condition for these two UxU matrices to be equal. This condition, which we shall state later as Lemma 3.1, is less stringent than the familiar condition for equality of

$I(\beta^*,\alpha^*)^{-1}$ and the upper left MxM sub-matrix of $I(\beta^*,\alpha^*,\eta^*)^{-1}$. Surprisingly, Stein's Lemma appears to have lay unnoticed until Bickel (1982) put it to an important application.

Following the appearance of Stein's paper, approximately fifteen years passed before a serious literature on adaptive estimation began to develop. Then, in the early 1970's, a number of authors reported positive findings for the simplest regression problem, namely the location parameter problem in which

$$y - \theta^* = u \tag{1}$$

where y, θ^*, and u are scalar, u has density f^*, and a random sample of y values are observed. The most general result was achieved by Beran (1974), who showed that if f^* is known only to be symmetric around zero, one can construct an estimator for θ^* whose asymptotic variance equals the best that would be attainable were f^* known. Thus, given symmetry of f^*, θ^* is adaptively estimable.

Beran also considered the 'two-sample' problem in which

$$y - \beta^* w - \alpha^* = u \tag{2}$$

where the arguments are all scalar, $\theta^* = (\beta^*,\alpha^*)$ and where w, which is observed, is Bernoulli distributed, independent of u. As before, u has unknown density f^* and sampling is random. For this problem, he found that the shift (or slope) parameter β^* can be adaptively estimated even when no prior restrictions on f^* beyond regularity are available.

Beran's proofs were constructive but his approach, which involved the construction of adaptive rank estimates, does not lend itself to application to more complex estimation problems. Soon after the appearance of Beran's paper, however, Stone (1975) reported an alternative constructive proof that a location parameter can be estimated adaptively, given symmetry of f^*. Stone's approach is at once generalizable, easily computable, and intuitively appealling.

Stone's construction has the following steps:

(1) Compute θ_N, any estimate for θ^* which does not use knowledge of f^* and which is $\sqrt{N}$ consistent whenever f^* is a symmetric density. For example, the sample median will do.

(2) Calculate the residuals $u_{nN} = y_n - \theta_N$, $n=1,\ldots N$.

(3) Use the residuals to form a non-parametric estimate of the density f^*. Stone chose a particular trimmed kernel estimate. Details on this will be given later.

(4) Acting as if the density estimate is the true density, take one Newton-Raphson type step from θ_N.

If the true density were used in Step (4), the generated estimate would, as is well known, be asymptotically equivalent to the maximum likelihood estimate. For this reason, Stone termed his procedure 'adaptive maximum likelihood'. Proof that the first stage residuals can be used successfully to adapt to the unknown density is decidedly non-trivial. In fact, given that the pointwise rate of convergence of a non-parametric density estimate is known to always be slower than $\sqrt{N}$, one might think that Stone's approach must fail. The Newton-Raphson step, however, requires not an estimate of the density per se but only of the information and of the sample mean score at θ_N. These functions of the density can, it turns out, be estimated well enough.

Technical differences aside, the adaptive maximum likelihood procedure should seem familiar to econometricians. The approximate generalized least squares methods ubiquitous in econometrics have similar structures. The latter are simpler in that the residuals, say from OLS, are used to estimate (adapt to) a finite set of parameters defining the second moments of a distribution, not to estimate a score. The objective in most econometric work has been to attain efficiency in the sense of the Gauss-Markov Theorem, not in the

more ambitious sense of the Cramer-Rao lower bound. The idea, however, is clearly the same.

Stone presented the adaptive maximum likelihood estimator entirely in the context of the location parameter problem. Proof that a version of the estimator is adaptive in more general regression problems is due to Bickel (1982). In a paper that must be considered a breakthrough of the first order, Bickel has done the following, all in the context of models with i.i.d. disturbances and random sampling.

First, he has shown that if the likelihood is a convex functional of the unknown error density f^*, and if that density is a priori restricted to a convex family Φ of densities, then Stein's necessary condition for adaptive estimation is equivalent to another condition far easier to verify or contradict. Essentially, Stein's necessary condition is equivalent to the requirement that the one Newton-Raphson step estimate computed using any $f \in \Phi$ be consistent and asymptotically normal whatever density $f^* \in \Phi$ actually is. The formal statement of Bickel's condition will be given later, as Condition B.

Assume now that one faces a suitably convex estimation problem for which Condition B is satisfied. Bickel's second major contribution was to prove that a modified version of Stone's procedure successfully yields an adaptive estimate if a certain verifiable condition is satisfied. This result, which will be presented here as Lemma 5.4, offers a prescription for the constructive proof of the existence of adaptive estimates.

Going further, Bickel applied his result to prove the existence of adaptive estimates in some important multi-parameter contexts. Of clear econometric interest are his applications to linear models. Consider the problem of estimating θ^* in the single equation model

$$y - x'\theta^* = u \qquad (3)$$

where y, $u \varepsilon R$ and $x, \theta^* \varepsilon R^K$. Bickel proved that if f^* is known only to be symmetric centered on zero, then, subject to standard regularity conditions on the distribution of x, θ^* can be estimated adaptively by an adaptive maximum likelihcod procedure. If the model contains an intercept, so that we may write

$$y - w' \cdot \beta^* - \alpha^* = u \qquad (4)$$

with $w, \beta^* \varepsilon R^{K-1}$; $\alpha^* \varepsilon R$; $\theta^* = (\beta^*, \alpha^*)$; $x = (w, 1)$; then the slope parameters β^* can be estimated adaptively when u is known only to be i.i.d. These findings are major generalizations of the Beran (1974) results for the models (1) and (2). The adaptive maximum likelihood estimates are, moreover, easily computable.

Viewed as a whole, Bickel's work converts what had been a set of isolated, specific results on adaptive estimation into a coherent field of study with broad application. On an aesthetic level, I find particularly appealing the relationship uncovered between the behavior of quasi-one Newton Raphson step estimates and the properties of adaptive maximum likelihood estimates. When quasi-one step estimates are $\sqrt{N}$-consistent, Stone's procedure remains $\sqrt{N}$-consistent even if a fixed, incorrect density estimate is used to perform the Newton-Raphson step. Given this, it makes sense that the use of a sequence of density estimates converging to the true density should yield an asymptotically efficient estimate. On the other hand, in an estimation problem where quasi-one step estimates are inconsistent, Stone's procedure based on a fixed density estimate is inconsistent. In this case, we might hope that use of a convergent sequence of density estimates would yield a consistent estimate but achievement of asymptotic efficiency would seem unlikely.

2. Non-Linear Regression Models: A Summary of Findings

In this paper, I shall extend Bickel's treatment of linear models to analyze the adaptive estimability of non-linear models of the general form

$$g(y,x,\theta^*) = u \tag{5}$$

where $y\varepsilon Y \subset R^J$ and $x\varepsilon X \subset R^K$ are observable, $u\varepsilon R^J$ is unobservable, the J-vector of functions g is specified up to the value of $\theta^*\varepsilon\Theta \subset R^M$ and a reduced form function $y = g^{-1}(u,x,\theta)$ exists for each $x\varepsilon X$, $\theta\varepsilon\Theta$. Maintained regularity conditions include the assumption that for each $x\varepsilon X$, the conditional distribution of u has differentiable density $f_x^*(\cdot)$ with finite, positive definite information. The conditional densities f_x^* are assumed non-informative regarding θ^*. The function $g(y,x,\theta)$ should be measureable in (y,x) for all θ and, for each x, should be jointly continuously differentiable over $(y,\theta)\ \varepsilon Y \times \Theta$. Certain integrability conditions will also be imposed on g.

The Bickel theorems apply when the sample (y_n,x_n), $n=1,\ldots,N$ is drawn by a serially independent, exogenous sampling process. That is, the sample size N and the realizations $(x_n,\ n=1,\ldots,N)$ are not per se informative regarding either θ^* or the densities $(f_x^*,\ x\varepsilon X)$. The likelihood of observing y conditional on x is the population conditional density, namely

$$\lambda(y|x,\theta^*,f_x^*) = |J(y,x,\theta^*)|\cdot f_x^*[g(y,x,\theta^*)] \tag{6}$$

where $J(y,x,\theta^*) = \det\left[\dfrac{\partial g(y,x,\theta^*)}{\partial y}\right]$. The likelihood of $(y_n,n=1,\ldots,N)$ conditional on $(x_n,\ n=1\ldots,N)$ is the product over $n=1,\ldots,N$ of the likelihoods of y_n conditional on x_n. These sampling assumptions are conventional in cross-sectional applications.

Some regularity must be imposed on the exogenous process generating realizations of x. Let Γ be the set of probability distributions on R^K having non-singular variance. We shall assume that the empirical distribution of x converges almost surely at rate $\sqrt{N}$ to some $G^* \in \Gamma$. This is satisfied if, for example, the realizations of x are randomly drawn from G^*.

An important sub-family of non-linear models are those that can be written in the non-linear regression form

$$y - h(x,\theta^*) = u \tag{7}$$

where h is a J-vector of functions. The structure (7) is considerably simpler than (5). In particular, $J(y,x,\theta) = 1$ identically and the basic likelihood expression (6) reduces to

$$\lambda(y|x,\theta^*,f_x^*) = f_x^*[y - h(x,\theta^*)]. \tag{8}$$

A further specialization of the non-linear regression family are the models with free intercept

$$y - \tilde{h}(w,\beta^*) - \alpha^* = u \tag{9}$$

where $\beta^* \in R^U$, $\alpha^* \in R^J$, $\theta^* = (\beta^*,\alpha^*)$, $x = (w,1)$, and $h(x,\theta^*) = \tilde{h}(w,\beta^*) + \alpha^*$. Here the basic likelihood expression is

$$\lambda(y|x,\beta^*, \alpha^*, f_x^*) = f_x^*[y - \tilde{h}(w,\beta^*) - \alpha^*]. \tag{10}$$

The distinctions between the likelihood expressions (6), (8) and (10) will be seen later to have important implications for the possibility of adaptive estimation.

Working in the above setting we can obtain general results on attainable precision of estimation. In Section 3, we state formally the Stein-Bickel necessary conditions for adaptive estimation. Then, in Section 4, we show that non-linear regression problems satisfy Condition B in two important settings. First, Condition B is satisfied for the entire parameter vector if the disturbances are known to be symmetrically distributed conditional on x (Proposition 4.1).

Second, Condition B is satisfied for the slope parameters of a model with free intercept if the disturbances are known to be i.i.d. (Proposition 4.2).

These results are very encouraging but it should not be thought that Condition B remains satisfied under arbitrarily weak distributional assumptions. For instance, if the disturbances are known to have densities that are close (in the sense of the weak topology) to symmetric but not necessarily symmetric, conditional on x, then Condition B is not satisfied (Proposition 4.3). If the model has a free intercept but the disturbances are not known to be i.i.d., then Condition B does not hold for the slope parameters unless the symmetry restriction holds (Proposition 4.4). In these cases then, adaptive estimation cannot be possible.

Condition B also fails when one moves from non-linear regression to models of form (5)-(6) in which the Jacobian of the transformation from y to u has a more complex form. Consider the simple location-scale parameter model $\beta^* y + \alpha^* = u$, where all the expressions are scalar and where y is known to be symmetric with mean zero and variance one. Here the Jacobian is still relatively simple yet Condition B for joint estimation of (β^*,α^*) is not satisfied (Proposition 4.5).

Following Section 4, we seek to verify the existence of adaptive estimates in those situations where Condition B is satisfied. The Stone-Bickel construction of adaptive maximum likelihood estimates is outlined in Section 5. While the adaptive maximum likelihood method should have general applicability, the central lemma on the convergence of the non-parametric estimate of the score function has thus far been proved only in the one-dimensional case. For this reason, our applications of the method, given in Section 6, are confined to single equation models.

In Section 6, we extend Bickel's proofs for linear models with i.i.d. disturbances to non-linear regression models with

x and u possibly interdependent. We prove that if the disturbances are known to be distributed symmetric around zero conditional on x, and if the space X can be partitioned into a finite system of subsets within each of which f_x^* is known to be invariant, then the Stone-Bickel estimator is adaptive for the entire parameter vector (Theorem 6.1 and Corollary). We then prove that the Stone-Bickel estimator is adaptive for the slope parameters if the model has a free intercept and the disturbances are known to be i.i.d. (Theorem 6.2).

The concluding Section 7 raises questions that are either being addressed in ongoing research or need to be addressed. Section 7.1 briefly describes the elegant new results of Begun et al. (1983) placing bounds on attainable precision when adaptation is not possible. Section 7.2 lists a number of important econometric problems which cannot be treated using the theory summarized in this paper. In Section 3, we consider the small sample behavior of adaptive maximum likelihood estimates and present some suggestive Monte Carlo findings.

3. The Stein-Bickel Necessary Conditions for the Existence of Adaptive Estimates

We present the Stein-Bickel necessary conditions in a form appropriate for our applications but not necessarily in the most general manner possible. In all that follows, we maintain the regularity and sampling assumptions imposed in Section 2 when the non-linear model (5)-(6) was introduced. Moreover, we restrict attention to problems in which the classical bound on precision of estimation would be attainable asymptotically if the set of densities $\phi^* = (f_x^*, x \in X)$ were known. That is, we assume that for all $\theta^* \in \Theta$ and $G^* \in \Gamma$,

$$I(\theta^*) = E\left(\frac{\partial \log\lambda}{\partial\theta}\,\frac{\partial \log\lambda}{\partial\theta'}\right)_{\theta=\theta^*} \tag{11}$$

$$= \int\left[\int \frac{\partial \log\lambda(y|x,\theta^*,f_x^*)}{\partial\theta}\,\frac{\partial \log\lambda(y|x,\theta^*,f_x^*)}{\partial\theta'}\,\lambda(y|x,\theta^*,f_x^*)dy\right]dG^*$$

is finite, non-singular and that, given knowledge of ϕ^*, there exists a computable estimator $\hat{\theta}_N$ such that, as $N\to\infty$,

$$\sqrt{N}\,(\hat{\theta}_N - \theta^*) \xrightarrow{L} \mathcal{N}(0, I(\theta^*)^{-1}). \tag{12}$$

Minimal regularity conditions guaranteeing the existence of $\hat{\theta}_N$ are given in LeCam (1969).

Consider next the familiar situation in which ϕ^* is not given but is known to be a member of a parametric family $[(f_x^\eta, x\varepsilon X),\ \eta\varepsilon H]$ where H is a subset of a finite dimensional real space, where each f_x^η is differentiable in u and where, for each value of u, $f_x^\eta(u)$ is differentiable in η at $\eta=\eta^*$. Here η^* indexes the true set of conditional densities. We can then write the true conditional density of y as $\lambda(y|x,\theta^*,\eta^*)$ and consider joint estimation of (θ^*,η^*). As is well known, the classical bound on the precision of an estimate for θ^* continues to equal $I(\theta^*)^{-1}$ if and only if

$$E\left(\frac{\partial \log\lambda}{\partial\theta}\,\frac{\partial \log\lambda}{\partial\eta'}\right)_{\theta=\theta^*,\ \eta=\eta^*} = 0. \tag{13}$$

Otherwise, the presence of the nuisance parameters η^* lowers the precision with which θ^* can be estimated.

It is much less well appreciated that in the presence of nuisance parameters, the bound on the precision of an estimate for a given sub-vector of θ^* can continue to equal the relevant sub-matrix of $I(\theta^*)^{-1}$ even though condition (13) is not satisfied. This was shown in Stein (1956).

Let $\theta^* = (\beta^*,\alpha^*)$, $\beta\varepsilon R^U$, $\alpha^*\varepsilon R^V$, U+V = M. For $\gamma\varepsilon(\beta,\alpha,\eta)$ and $\delta\varepsilon(\beta,\alpha,\eta)$, define

$$A_{\gamma\delta} = E\left(\frac{\partial \log\lambda}{\partial\gamma}\,\frac{\partial \log\lambda}{\partial\delta'}\right)_{\gamma=\gamma^*,\ \delta=\delta^*}. \tag{14}$$

If η^* is known, the smallest possible asymptotic variance for an asymptotically normal estimate of β^* is the UxU upper left sub-matrix of the inverted information matrix

$$I(\beta^*,\alpha^*)^{-1} = \begin{bmatrix} A_{\beta\beta} & A_{\beta\alpha} \\ A'_{\beta\alpha} & A_{\alpha\alpha} \end{bmatrix}^{-1}. \tag{15}$$

If η^* is not known, the smallest possible asymptotic variance for an estimate of β^* is the UxU upper left sub-matrix of

$$I(\beta^*,\alpha^*,\eta^*)^{-1} = \begin{bmatrix} A_{\beta\beta} & A_{\beta\alpha} & A_{\beta\eta} \\ A'_{\beta\alpha} & A_{\alpha\alpha} & A_{\alpha\eta} \\ A'_{\beta\eta} & A'_{\alpha\eta} & A_{\eta\eta} \end{bmatrix}^{-1}. \tag{16}$$

Let I_U denote the UxU identity matrix. Stein proved the following matrix algebraic lemma:

Lemma 3.1: The UxU upper left sub-matrix of $I(\beta^*,\alpha^*,\eta^*)^{-1}$ equals that of $I(\beta^*,\alpha^*)^{-1}$ if and only if

$$\left[I_U : -A_{\beta\alpha}A_{\alpha\alpha}^{-1}\right]\begin{bmatrix} A_{\beta\eta} \\ A_{\alpha\eta} \end{bmatrix} = 0. \tag{17}$$

Proof: See Stein (1956).

Condition (17) is less stringent than (13), which requires that $A_{\beta\eta} = A_{\alpha\eta} = 0$.

Now consider the situation in which ϕ^* is known only to belong to a class Φ of sets of densities containing the parametric family $[(f_x^{\eta},\ x\in X),\ \eta\in H]$. Clearly, a subvector β^* of θ^* can be estimated no more precisely in this case than in the case where ϕ^* is known to be in the parametric subset of Φ. This simple observation is the essence of the Stein (1956)

necessary condition for adaptive estimation of β^*. Paraphrased for our applications, Stein's condition is

Condition S: Let ϕ^* be known to lie in Φ, a specified class of sets of densities on R^J. If there exists some $(\theta^*,\phi^*,G^*) \in \Theta \times \Phi \times \Gamma$ and some finite dimensional subfamily of Φ such that $I(\beta^*,\alpha^*,\eta^*)^{-1}$ exists but (17) is not satisfied, then β^* is not adaptively estimable.

Condition S is conceptually simple but difficult to check. To obtain a practical version of the condition, Bickel (1982) restricted attention to problems in which Φ is a convex family of densities and the sampling distribution of the data is a convex functional on Φ. The former condition is satisfied if, for example, Φ is the space of all densities or of all symmetric densities centered on zero. In our applications, the structure of the basic likelihood expression (6) implies that the latter condition is always met. Simply observe that for all $(y,x,\theta) \in Y \times X \times \Theta$, all pairs of densities (f^0,f^1) and all $\eta \in [0,1]$,

$$\lambda[y|x,\theta,(\eta f^1 + (1-\eta)f^0)] \tag{18}$$

$$= |J(y,x,\theta)| \cdot [\eta f^1(g(y,x,\theta)) + (1-\eta)f^0(g(y,x,\theta))]$$

$$= \eta \cdot \lambda(y|x,\theta,f^1) + (1-\eta)\cdot\lambda(y|x,\theta,f^0).$$

Consider now any $\phi^* = (f^*_x, x \in X) \in \Phi$ and $\phi = (f_x, x \in X) \in \Phi$. Letting

$$f^\eta_x = \eta f_x + (1-\eta) f^*_x, \tag{19}$$

convexity of Φ implies that

$$[(f^\eta_x, x \in X), \eta \in [0,1]] \in \Phi. \tag{20}$$

Bickel obtained his condition by applying Condition S

informally to the parametric family defined in (20). Formally, this makes sense only if the densities f_x^η are differentiable in η at the boundary point $\eta^* = 0$. To guarantee this, we can extend the family (20) and apply Condition S only when there exists an $\eta_0 < 0$ such that

$$[(f_x^\eta, x\varepsilon X), \eta\varepsilon[\eta_0, 1]] \subset \Phi. \tag{20'}$$

This allows us to formally derive the following version of Bickel's necessary condition for adaptive estimation.

Condition B: Let ϕ^* be known to lie in Φ, a specified convex class of sets of densities on R^J. Where it exists, define the expected score function

$$S[(\theta^*,\phi^*),(\theta^*,\phi),G^*] = \int\int \frac{\partial \log\lambda(y|x,\theta^*,f_x^*)}{\partial \theta} \lambda(y|x,\theta^*,f_x) dy dG^* \tag{21}$$

A necessary condition for adaptive estimation of β^* is that whenever S exists and (20') holds for some $\eta_0 < 0$,

$$\frac{\partial \beta}{\partial \theta'} I(\theta^*)^{-1} S[(\theta^*,\phi^*),(\theta^*,\phi),G^*] = 0. \tag{22}$$

Proof: When ϕ^* and ϕ are such that (20') holds for some $\eta_0 < 0$, f_x^η is differentiable in η at $\eta^* = 0$ and Condition S can be applied.

To interpret (17) in this case, observe that by (18),

$$\frac{\partial \log\lambda(y|x,\theta^*,\eta^*)}{\partial \eta} = \frac{\lambda(y|x,\theta^*,f_x)}{\lambda(y|x,\theta^*,f_x^*)} - 1. \tag{23}$$

The adding-up condition for probabilities implies that

$$E\left(\frac{\partial \log\lambda}{\partial \theta}\right)_{\theta=\theta^*} = 0. \tag{24}$$

Together, (23) and (24) imply that

$$\begin{bmatrix} A_{\beta\eta} \\ A_{\alpha\eta} \end{bmatrix} = S[(\theta^*,\phi^*),(\theta^*,\phi),G^*]. \qquad (25)$$

Next observe that premultiplication of both sides of (17) by an arbitrary non-singular UxU matrix B yields the equivalent relationship

$$[B:-BA_{\beta\alpha}A_{\alpha\alpha}^{-1}] \begin{bmatrix} A_{\beta\eta} \\ A_{\alpha\eta} \end{bmatrix} = 0. \qquad (17')$$

Since $\frac{\partial\beta}{\partial\theta'}$ is the UxM matrix $[I_{\eta}: 0]$, a further equivalent statement is

$$\frac{\partial\beta}{\partial\theta'} \begin{bmatrix} B & -BA_{\beta\alpha}A_{\alpha\alpha}^{-1} \\ C' & D \end{bmatrix} \begin{bmatrix} A_{\beta\eta} \\ A_{\alpha\eta} \end{bmatrix} = 0 \qquad (17")$$

where C' and D are arbitrary VxU and VxV matrices. Choose $B = (A_{\beta\beta} - A_{\beta\alpha}A_{\alpha\alpha}^{-1}A_{\beta\alpha}')^{-1}$, $C = -BA_{\beta\alpha}A_{\alpha\alpha}^{-1}$, and $D = A_{\alpha\alpha}^{-1} + C'B^{-1}C$. Then (17") becomes

$$\frac{\partial\beta}{\partial\theta}, I(\beta^*,\alpha^*)^{-1} \begin{bmatrix} A_{\beta\eta} \\ A_{\alpha\eta} \end{bmatrix} = 0. \qquad (17''')$$

Combining this with (25) yields Condition B.

Q.E.D.

An important special case is that in which $\beta^* = \theta^*$. Here, $\frac{\partial\beta}{\partial\theta'} = I_M$, the presence of $I(\theta^*)^{-1}$ in (22) becomes irrelevant and (22) reduces to $S[(\theta^*,\phi^*),(\phi^*,\phi),G^*] = 0$. Observe that this is the condition that should be satisfied for consistent, quasi-maximum likelihood estimation of θ^*. See Huber (1967). More directly relevant to us is the fact that (22) is necessary and sufficient for consistent estimation of β^* by a quasi-one step procedure. That is, assume that an initial consistent estimate of θ^* has been obtained and now take one Newton-Raphson step under the assumption that the disturbances are

generated by ϕ^* when they really are by ϕ. Under standard regularity conditions, the resulting estimate remains consistent for β^*, although not necessarily for α^*, if and only if (22) is satisfied.

4. Applications

4.1 Non-Linear Regression Models

Bickel (1982) verified that Condition B is satisfied for estimation of θ^* in the single equation linear model when the disturbances are known to be i.i.d. and symmetric around zero. He also showed that Condition B is satisfied for estimation of the slope parameters β^* in the single equation linear model with free intercept when the disturbances are only known to be i.i.d. Proposition 4.1 extends the former result to the J-equation non-linear regression model when the disturbances are known to be symmetric around zero, conditional on x. Proposition 4.2 extends the latter finding to the J-equation non-linear model with i.i.d. disturbances. In what follows, the information of a density f is denoted by the JxJ matrix

$$i(f) = \int \frac{1}{f(u)} \frac{\partial f(u)}{\partial u} \frac{\partial f(u)}{\partial u'} du. \tag{26}$$

Where it exists, the expected score for a density f° computed under a possibly different density f^1 is denoted by the Jx1 vector

$$s(f^\circ, f^1) = \int \frac{1}{f^\circ(u)} \frac{\partial f^\circ(u)}{\partial u} f^1(u) du \tag{27}$$

Proposition 4.1: Let F^S be the space of all symmetric J-variate densities centered on zero, with finite non-singular information. Let $\Phi = (F^S)^X$. Then model (7) satisfies Condition B for estimation of θ^*.

Proof: As defined, Φ is convex. For a model of form (7) and $\beta^* = \theta^*$, condition (22) reduces to

$$S[(\theta^*,\phi^*),(\theta^*,\phi),\ G^*] = \int -\frac{\partial h(x,\theta^*)}{\partial \theta} s(f_x^*,f_x)dG^* = 0 \quad (28)$$

where $\frac{\partial h}{\partial \theta}$ is the MxJ matrix of terms $\frac{\partial h_j}{\partial \theta_m}$. For each $x \in X$, the restriction of f_x^* and f_x to F^s implies that $\frac{1}{f_x^*}\frac{\partial f_x^*}{\partial u} f_x$ is anti-symmetric so $s(f_x^*,f_x) = 0$. Hence, (28) is satisfied. Q.E.D.

Proposition 4.2: Let F° be the space of all J-variate densities with finite non-singular information and zero mean. Let $\Phi = [(f)^X,\ f \in F^\circ]$. Then model (9) satisfies Condition B for estimation of β^*.

Proof: Φ is convex. Rather than verify condition (22) directly, it is simpler to verify the equivalent condition

$$[I_u: -A_{\beta\alpha}A_{\alpha\alpha}^{-1}]\ S[(\theta^*,\phi^*),(\theta^*,\phi),G^*] = 0. \quad (22')$$

To do this, observe that for a model of form (9) with i.i.d. disturbances

$$A_{\beta\alpha} = E[\frac{\partial \tilde{h}(w,\beta^*)}{\partial \beta}] \cdot i(f^*) \quad (29)$$

$$A_{\alpha\alpha}^{-1} = i(f^*)^{-1} \quad (30)$$

and

$$S[(\theta^*,\phi^*),(\theta^*,\phi),G^*] = -\begin{bmatrix} E[\frac{\partial \tilde{h}(w,\beta^*)}{\partial \beta}] \\ I_J \end{bmatrix} s(f^*,f). \quad (31)$$

Inspection of (29), (30) and (31) reveals that (22') is satisfied for all densities f^* and f for which $s(f^*,f)$ exists. Q.E.D.

In Proposition 4.2, the restriction of f^* to densities with mean zero was imposed only to identify α^* and can be replaced by an alternative location parameter restriction.

This aside, it appears that the distributional assumptions of Propositions 4.1 and 4.2 are close to the minimum necessary for satisfaction of Condition B. In particular, we can prove that adaptive estimation of θ^* becomes impossible if Proposition 4.1's restriction of $(f^*_x, x\varepsilon X)$ to symmetric densities is relaxed even locally, so as to allow neighbors of such densities. We can also show that Proposition 4.2 does not generalize to cases in which u is dependent on x. These negative results are contained in Propositions 4.3 and 4.4.

Proposition 4.3: For given $\delta\varepsilon(0,1)$, let $F^{s\delta} = [(1-\eta)\cdot f + \eta\cdot z: f\varepsilon F^s, z\varepsilon(F^\circ - F^s), 0 \leq \eta < \delta]$ be the space of δ-neighbors of symmetric densities on R^J, centered on zero. Let $\Phi = (F^{s\delta})^X$. Then model (7) does not satisfy Condition B for estimation of θ^*.

Proof: It is easy to show that Φ is convex. To contradict condition B, it suffices to consider the special case in which the disturbances are truly i.i.d. symmetric. Let $f^*\varepsilon F^s$ be the common true density and consider $(f_x, x\varepsilon X) = [(1-\eta)f^* + \eta z]^X$ for some $z\varepsilon(F^\circ - F^s)$, $0 < \eta < \delta$. Then the condition (22) for θ^* reduces to

$$\left[\int \frac{\partial h(x,\theta)}{\partial \theta} dG^*\right]\left[(1-\eta)s(f^*,f^*) + \eta\, s(f^*,z)\right] = 0. \qquad (32)$$

The assumed non-singularity of $I(\theta^*)$ for all $G^*\varepsilon\Gamma$ implies that $\frac{\partial h(x,\theta^*)}{\partial \theta} \neq 0$ for at least some x. Hence, $\int \frac{\partial h(x,\theta^*)}{\partial \theta} dG^* \neq 0$ for at least some $G^*\varepsilon\Gamma$. For all f^*, $s(f^*,f^*) = 0$. On the other hand, there clearly exist non-symmetric z for which $s(f^*,z) \neq 0$. Hence, (32) does not always hold as Condition B would require.

Q.E.D.

Proposition 4.4: Let $\Phi = (F^\circ)^X$. Then model (9) does not

satisfy Condition B for estimation of β^*.

Proof: It suffices to consider the two-sample model (2) where $y-\beta^*w - \alpha^* = u$ and $w=1$ with probability γ, $w=0$ otherwise. Let (f_o^*,f_1^*) be the true conditional densities for $w=0,1$ and consider any alternative pair (f_o,f_1) in Φ. In this context,

$$A_{\beta\alpha} = \gamma \cdot i(f_1^*) \tag{33}$$

$$A_{\alpha\alpha} = (1-\gamma)\cdot i(f_o^*) + \gamma \cdot i(f_1^*) \tag{34}$$

and

$$S[(\theta^*,\phi^*),(\theta^*,\phi),G^*] = \tag{35}$$

$$= - \begin{bmatrix} 0 \\ 1 \end{bmatrix} (1-\gamma)s(f_o^*,f_o) - \begin{bmatrix} 1 \\ 1 \end{bmatrix} \gamma s(f_1^*,f_1).$$

The condition (22') therefore reduces to

$$\frac{\gamma \cdot i(f_1^*)}{(1-\gamma)\cdot i(f_o^*) + \gamma \cdot i(f_1^*)} \left[(1-\gamma)s(f_o^*,f_o) + \gamma s(f_1^*,f_1)\right] \tag{36}$$

$$- \gamma s(f_1^*,f_1) = 0.$$

It is immediate that (36) is not satisfied for all choices of (f_o^*,f_1^*), (f_o,f_1) and γ. For example, if $s(f_1^*,f_1) = 0$ and $s(f_o^*,f_o) \neq 0$, condition (36) fails for all $\gamma \varepsilon (o,1)$.

Q.E.D.

4.2 Linear Systems Models

Consider the J-equation simultaneous systems model

$$b(\theta^*)y + a(\theta^*)x = u \tag{37}$$

where $b \sim (JxJ)$, $a \sim (JxK)$ and prior restrictions on the structural parameters are expressed by making those parameters functions of a lower-dimension parameter vector θ^*. Assuming $b(\theta^*)$ non-singular, we can write the equivalent reduced form model

$$y - [-b(\theta^*)^{-1} a(\theta^*)] x = b(\theta^*)^{-1} u. \qquad (37')$$

Prior knowledge that the structural disturbances u have conditional densities in a class Φ implies that the reduced form disturbances $b(\theta^*)^{-1}u$ have conditional densities in some class $\tilde{\Phi}(\theta^*)$.

When $\tilde{\Phi}(\theta^*)$ is invariant over $\theta^* \varepsilon \Theta$, the reduced form disturbances are uninformative regarding θ^* and the problem of adaptive estimation of simultaneous systems models becomes identical to that of adaptive estimation of non-linear regression models. This holds in particular if Φ is specified as in Propositions 4.1 and 4.2. In both of these cases, $\tilde{\Phi}(\theta^*) = \Phi$ for all θ^* and the simultaneous systems origin of the regression (37') introduces no new issues.

When $\tilde{\Phi}(\theta^*)$ does vary with θ^*, the reduced form model (37') cannot be analyzed using the tools of this paper. On the other hand, the structural model (37) can, in principle, be checked directly for satisfaction of Condition B. This will not be pursued here.

4.3 Joint Estimation of Location and Scale

Consider for a moment the general non-linear model defined in (5)-(6). The expected score function S has the relatively complex form

$$S[(\theta^*,\phi^*),(\theta^*,\phi),G^*] \qquad (38)$$

$$= \int\int |J(y,x,\theta^*)| \frac{\partial g(y,x,\theta^*)}{\partial \theta} \frac{\partial f_x^*[g(y,x,\theta^*)]}{\partial u} \frac{f_x[g(y,x,\theta^*)]}{f_x^*[g(y,x,\theta^*)]} dy dG^*$$

$$+ \int\int \frac{\partial |J(y,x,\theta^*)|}{\partial \theta} f_x [g(y,x,\theta^*)] dy \, dG^*$$

Note that in the special case of a non-linear regression problem, the second term disappears as $\frac{\partial |J|}{\partial \theta} = 0$ and the first term simplifies as $|J| = 1$ and $\frac{\partial g}{\partial \theta} = \frac{\partial h}{\partial \theta}$.

Given the complexity of the expression (38), a general characterization of the situations in which Condition B is and is not satisfied appears difficult to achieve. On the other hand, some insight emerges from consideration of what is, perhaps, the simplest model that is of form (5) but not (7). This is the single equation model

$$b^*y + a^* = u \tag{39}$$

where the location parameter a^* and scale parameter $b^* > 0$ are to be jointly estimated. A natural choice for Φ is the set F^{ss} of standardized symmetric densities, defined to be the space of symmetric densities centered on zero, with variance one and finite non-singular information. For this specification we can prove

Proposition 4.5: Let $\Phi = F^{ss}$. The model (39) does not satisfy Condition B for estimation of (b^*,a^*).

Proof: As defined, Φ is convex.
For model (39), $|J| = b^*$, $\frac{\partial |J|}{\partial b} = 1$, $\frac{\partial |J|}{\partial a} = 0$, $\frac{\partial g}{\partial b} = y$, $\frac{\partial g}{\partial a} = 1$.
Therefore, condition (22) for joint adaptive estimation of (b^*,a^*) has the two components

$$\frac{\partial S}{\partial b} = \int u \frac{\partial f^*(u)}{\partial u} \frac{f(u)}{f^*(u)}\, du - a^* s(f^*,f) + 1 = 0 \tag{40}$$

$$\frac{\partial S}{\partial a} = b^* s(f^*,f) = 0 \tag{41}$$

where we have, in (40), used the fact that $\beta^* y = u - a^*$. For $f \in F^{ss}$, condition (41) is always satisfied and condition (40) reduces to

$$1 = - \int u \frac{\partial f^*(u)}{\partial u} \frac{f(u)}{f^*(u)}\, du. \tag{42}$$

Clearly, (42) is not satisfied for all f^*, $f \in F^{ss}$. For example, take $f^*(u) = \frac{1}{\gamma} \exp\left(-\frac{u^4}{4\delta}\right)$ where γ and δ are such that

$\int f^*(u)du = 1$ and $\int u^2 f^*(u)\, du = 1$. Then $f^* \varepsilon F^{ss}$, $\frac{\partial f^*(u)}{\partial u} \frac{1}{f^*(u)} = -\frac{u^3}{\delta}$ and condition (42) becomes $\delta = \int u^4 f(u)du$. The restriction of f to F^{ss}, however, does not constrain the fourth moment of f to any constant value. Hence, (42) is not always satisfied as required for Condition B to hold.

Q.E.D.

Observe that the question of joint adaptive estimation of (b^*,a^*) in model (39) is logically distinct from the question of adaptive estimation of a^*/b^* in the transformed model

$$y + a^*/b^* = u' \qquad (39')$$

where $u' = u/b^*$. Restriction of the density of u to F^{ss} implies that the density of u' is in a subset of F^s, namely those symmetric densities having finite variance. It follows from Proposition 4.1 that Condition B is satisfied for a^*/b^*. Thus, we have here another instance in which the entire parameter vector θ^* cannot be adaptively estimated but an interesting function of θ^* can be.

Writing (39) as (39') points to an important re-interpretation of Proposition 4.5. This is

Corollary: Let $\Phi = (F^s)^X$. Then model (7) does not satisfy Condition B for joint estimation of θ^* and the standard error of the regression.

Proof: Model (39') is a special case of model (7) and the space of finite variance symmetric densities is a subset of F^s. In this special case, $\theta^* = a^*/b^*$ and $1/b^*$ is the standard error of the regression. By Proposition 4.5, Condition B is not satisfied for (b^*,a^*). Hence, Condition B is not satisfied for the one-to-one transformation $(a^*/b^*,1/b^*)$.

Q.E.D.

5. The Stone-Bickel Construction of Adaptive Maximum Likelihood Estimates

When Condition B is satisfied, one may attempt to confirm that adaptive estimation is possible and to construct adaptive estimates. The Stone-Bickel work on adaptive maximum likelihood (AML) estimates meets both these objectives. We shall summarize the AML approach as developed by Stone (1975) and Bickel (1982) and shall simultaneously lay the groundwork for our applications to non-linear regression models.

5.1 General Approach

Consider first the idealized situation in which (ϕ^*, G^*) is known but θ^* is not. Given the sample (y_n, x_n), $n=1,\ldots,N$, let $\theta_N \varepsilon \Theta$ be an estimate for θ^* known to satisfy the condition

$$\sqrt{N}(\theta_N - \theta^*) = O_p(1). \tag{43}$$

Recalling equation (11), let $I(\theta_N)$ be the information matrix evaluated as if θ_N were the true parameter vector. Define the sample mean score function

$$S_N(\theta_N, \phi^*) = \frac{1}{N} \sum_{n=1}^{N} \frac{\partial \log \lambda(y_n | x_n, \theta_N, f^*_{x_n})}{\partial \theta}. \tag{44}$$

Now construct the estimate

$$\hat{\theta}_N = \theta_N + I(\theta_N)^{-1} S_N(\theta_N, \phi^*), \tag{45}$$

which is a modified form of the familiar one Newton-Raphson step estimate. The modification is that $\hat{\theta}_N$ uses $I(\theta_N)$ to approximate $I(\theta^*)$ while the usual one step estimate uses minus the sample mean of $\partial^2 \log \lambda(y|x,\theta_N, f^*_x)/\partial\theta\partial\theta'$. In all that follows we assume that the estimation problem is sufficiently regular so that $\hat{\theta}_N$ satisfies

$$\sqrt{N}(\hat{\theta}_N - \theta^*) \xrightarrow{L} \mathcal{N}(0, I(\theta^*)^{-1}). \tag{46}$$

In the random sampling context in particular, the asymptotic efficiency property (46) is known to hold given minimal smoothness restrictions. See LeCam (1969) and Bickel (1982). Here, we shall simply assume (46) directly.

Now return to the situation of interest, in which it is known only that $(\theta^*, \phi^*, G^*) \in \Theta \times \Phi \times \Gamma$. Assume that there exists a computable initial estimate θ_N that satisfies (43). Let I_N and S_N be computable estimates of $I(\theta_N)$ and $S_N(\theta_N, \phi^*)$ respectively. Then construct the estimate for θ^*

$$\tilde{\theta}_N = \theta_N + I_N^- S_N \tag{47}$$

where I_N^- is a generalized inverse of I_N. Given (46), the subestimate $\tilde{\beta}_N = \frac{\partial \beta}{\partial \theta'} \tilde{\theta}_N$ will be adaptive for β^* if and only if

$$\frac{\partial \beta}{\partial \theta'} \sqrt{N}[\, I_N^- S_N - I(\theta_N)^{-1} S_N(\theta_N, \phi^*)] = o_p(1). \tag{48}$$

The question is whether computable estimates θ_N, I_N and S_N satisfying (43) and (48) exist.

5.2 Estimation of the Density

Stone (1975) answered this question in the affirmative in the case of the location parameter model (1) with symmetric disturbances. In this setting, suitable initial estimates θ_N are readily available and the problem of estimating $I(\theta_N)$ and $S_N(\theta_N, \phi^*)$ reduces to one of appropriate estimation of the unknown f^* and its score function $\frac{1}{f^*} \frac{\partial f^*}{\partial u}$. To accomplish this, Stone used residuals to form a kernel estimate for f^* and, after some trimming, formed a score function estimate from this. The procedure is as follows.

For $n = 1, \ldots N$, compute the residuals

$$u_{nN} = y_n - \theta_N \tag{49}$$

Define u^N to be a random variable having the empirical distribution of the residuals. That is, $\text{Prob}(u^N = u_{nN}) = \frac{1}{N}$. Let ν^N be a random variable distributed normal with mean zero and variance σ_N^2, where σ_N^2 is a positive parameter selected by the analyst. Now define the random variable

$$\zeta^N = u^N + \nu^N. \tag{50}$$

Then the distribution of ζ^N is a smoothed version of that of u^N, with σ_N^2 controlling the degree of smoothing. In particular, ζ^N has the infinitely differentiable density

$$f_N(u) = \sum_{n=1}^{N} \frac{1}{N} \left[\frac{1}{\sigma_N} \phi\left(\frac{u-u_{nN}}{\sigma_N}\right)\right] \tag{51}$$

where ϕ is the standard normal density. Recall that f^* is known to be symmetric. The estimate f_N is not generally symmetric but the derived estimate

$$f_N^s(u) = \frac{1}{2}\left[f_N(u) + f_N(-u)\right] \tag{52}$$

is.

Stone used f_N^s as his estimate for f^*. He proposed a family of trimmed score function estimates, the simplest of which is

$$r_N^s(u) = \frac{1}{f_N^s(u)} \frac{df_N^s(u)}{du} \quad \text{if } |u| < b_N \tag{53a}$$

$$r_N^s(u) = 0 \quad \text{otherwise.} \tag{53b}$$

Here b_N is a positive parameter selected by the analyst to control the degree of trimming. Using f_N^s and r_N^s, Stone proposed estimates I_N and S_N. His estimates are considerably more burdensome to compute than the ones later introduced by

Bickel, in a more general setting. For this reason, details will not be given here. The important point is that Stone proved his estimate $\tilde{\theta}_N$ to be adaptive for the location parameter θ^*, provided that the degree of smoothing in f_N^s and trimming in r_N^s are reduced at appropriate rates as the sample size increases. Neither Stone nor Bickel discusses how the smoothing and trimming parameters should be set in the context of a given data sample. We shall return to this problem in Section 7.

While Stone's approach would seem generalizable well beyond the location parameter problem, his dense, specific presentation makes it difficult to see beyond his paper's confines. Happily, Bickel (1982) has succeeded in doing so. To appreciate Bickel's contribution, observe that in any AML problem, the question of appropriate estimation of the unknown density can be decomposed into three components. These are

(1) What properties should an estimate of the density have in order that it be possible to construct estimates I_N and S_N that satisfy (48)?

(2) In the idealized situation where the disturbances $u_1,\ldots,u_N$ are observed, can a density estimate with the appropriate properties be formed?

(3) If the answer to question 2 is positive but only residuals $u_{1N},\ldots,u_{NN}$ associated with a θ_N satisfying (43) are observed, can a density estimate with the appropriate properties still be formed?

Working in the context of single equation linear models with i.i.d. disturbances, Bickel found the following.

First, appropriate estimation of $I(\theta_N)$ and $S_N(\theta_N,\phi^*)$ requires suitably convergent estimation of the score function $\frac{1}{f^*}\frac{\partial f^*}{\partial u}$ but not of the density f^* per se. Let q_N be an estimate for the score function. Then q_N should satisfy the mean square

convergence condition

$$\int [q_N(u) - \frac{1}{f^*(u)} \frac{df^*(u)}{du}]^2 \; f^*(u)du \xrightarrow{p} 0 \qquad (54)$$

as $N \to \infty$. Rather than derive (54) here, we shall show in Section 6 that it remains the relevant condition in the more general context of single equation non-linear regression models.

Second, if $u_1, \ldots u_N$ are observed, a modified version of Stone's score function estimate satisfies (54). In particular, define

$$f_N^*(u) = \sum_{n=1}^{N} \frac{1}{N} [\frac{1}{\sigma_N} \phi(\frac{u-u_n}{\sigma_N})] \qquad (55)$$

and

$$q_N^*(u) = \frac{1}{f_N^*(u)} \frac{df_N^*(u)}{du} \text{ if } |u| \leq b_N, \; f_N(u) \geq c_N \qquad (56a)$$

$$\text{and } | \frac{df_N^*(u)}{du} | \leq d_N f_N^*(u)$$

$$q_N^*(u) = 0 \qquad \text{otherwise.} \qquad (56b)$$

Here b_N, c_N, and d_N are positive parameters chosen by the analyst to control trimming. Bickel proved

Lemma 5.1: Let $u_1, \ldots, u_N$ be a random sample from a univariate density f^* having finite information. Then the score function estimate q_N^* defined in (56) converges in mean square as specified in (54) provided that $b_N \to \infty$, $c_N \to 0$, $d_N \to \infty$, $\sigma_N \to 0$, $N^{-1} b_N \sigma_N^{-3} \to 0$ and $\sigma_N d_N \to 0$ as $N \to \infty$.

Proof: See Bickel (1982), Section 6.1 for the lengthy and delicate proof. Related results are given in Stone (1975), Section 3.

When f^* is known to be symmetric, it is desirable that the score function estimate be anti-symmetric. The estimate q_N^* is not generally anti-symmetric but the derived estimate

$$q_N^{*s}(u) = \frac{1}{2}\left[q_N^*(u) - q_N^*(-u)\right] \tag{57}$$

is. The following Corollary to Lemma 5.1 is a simple application of the Cauchy-Schwarz inequality.

Corollary: If $f^* \in F^s$, Lemma 5.1 continues to be satisfied when q_N^{*s} replaces q_N^*.

Now turn to the third question. Observe that Lemma 5.1 and its Corollary are general results, not dependent on the model in which the disturbances appear. On the other hand, extension of the Lemma to score function estimation based on residuals requires consideration of the model generating the residuals. For $n=1,\ldots,N$ let

$$u_{nN} = y_n - x_n' \theta_N \tag{58}$$

be the single equation linear model residuals. Define f_N as in (51) and define q_N as in (56), but with f_N replacing f_N^*. In this context, Bickel proved that Lemma 5.1 continues to be satisfied when q_N replaces q_N^*. The argument that linear model residuals can replace disturbances in the estimation of the score function relies on two very useful ideas of LeCam and on a theorem of Hajek and Sidak. We shall need to extend the argument to non-linear models and so shall present it in some detail.

LeCam (1960) introduced the concept of 'contiguity' of two sequences of probability measures. In the present setting, the sequence P_N', $N=1,\ldots\infty$ of N dimensional densities of the residuals $(u_{nN}, n=1,\ldots,N)$ is said to be contiguous to the sequence P_N, $N=1,\ldots\infty$ of N dimensional densities of the disturbances $(u_n, n=1,\ldots,N)$ if, as $N \to \infty$,

$$\int_{A_N} P'_N(u_{1N},\ldots,u_{NN})du_{1N},\ldots,du_{NN} \to 0 \qquad (59a)$$

for every measureable sequence of events $A_N \subset R^N$, $N=1,\ldots\infty$ such that

$$\int_{A_N} P_N(u_1,\ldots,u_N)\, du_1,\ldots du_N \to 0. \qquad (59b)$$

To see the relevance of contiguity to the present problem, choose any $\varepsilon > 0$ and let

$$A_N^{\varepsilon} = \left[(u_1,\ldots,u_N)\varepsilon R^N : \int\left[q_N^*(u) - \frac{1}{f^*(u)}\frac{df^*(u)}{du}\right]^2 f^*(u)du > \varepsilon\right] \qquad (60)$$

By Lemma 5.1, $\{A_N^{\varepsilon}\}$ is a sequence satisfying (59b). If $\{P'_N\}$ is contiguous to $\{P_N\}$, then the sequence $\{A_N^{\varepsilon}\}$ also satisfies (59a), implying that q_N does converge in mean square. Thus, the problem of extending Lemma 5.1 to estimation by residuals is solved if it can be shown that $\{P'_N\}$ is contiguous to $\{P_N\}$.

The result enabling demonstration of contiguity is due to Hajek and Sidak (1967). Let z_n, $n=1,\ldots\infty$ be a sequence of univariate random variables such that the joint density of z_n, $n=1,\ldots,N$ is

$$Q'_N = \prod_{n=1}^{N} f(z_n - \mu_{nN}). \qquad (61)$$

Here f is a fixed density with $0 < i(f) < \infty$ and μ_{nN}, $n=1,\ldots,N$ are a set of location parameters. Let $\bar{\mu}_N = N^{-1}\sum\mu_{nN}$ and define

$$Q_N = \prod_{n=1}^{N} f(z_n - \bar{\mu}_N). \qquad (62)$$

Hajek and Sidak proved

Lemma 5.2: Assume that as $N \to \infty$,

$$\operatorname*{Max}_{n=1,\ldots,N} (\mu_{nN} - \bar{\mu}_N)^2 \to 0 \qquad (63)$$

and

$$\sum_{n=1}^{N} (\mu_{nN} - \bar{\mu}_N)^2 \to \rho \tag{64}$$

where $0 < \rho < \infty$. Then $\{Q'_N\}$ is contiguous to $\{Q_N\}$.

Proof: See Hajek and Sidak (1967), Section VI.2.1.

To apply Lemma 5.2 to the problem at hand, we select the initial estimate θ_N in a manner suggested by LeCam. Let $\bar{\theta}_N$ be any computable estimate satisfying (43). Recalling that $\theta^* \varepsilon R^M$, define the lattice of coordinates in R^M

$$R_N^M = [N^{-1/2}(i_1,\ldots,i_M): i_k = -\infty,\ldots -1,0,1,\ldots\infty, k=1,\ldots M]. \tag{65}$$

Now choose θ_N to be a point in R_N^M closest in Euclidean distance to $\bar{\theta}_N$. Clearly θ_N satisfies (43), hence is a legitimate initial estimate. The estimate θ_N has a technical advantage over $\bar{\theta}_N$, as follows.

$\sqrt{N}$-consistency implies that for every $\lambda > 0$ there exists a $\delta(\lambda) > 0$ such that

$$\lim_{N \to \infty} \text{Prob}[|\bar{\theta}_N - \theta^*| < N^{-1/2}\delta(\lambda)] > 1-\lambda \tag{66a}$$

$$\lim_{N \to \infty} \text{Prob}[|\theta_N - \theta^*| < N^{-1/2}\delta(\lambda)] > 1-\lambda. \tag{66b}$$

Here $|\cdot|$ is the Euclidean norm. Consider the sets

$$\bar{\Theta}_{\lambda N} = [\theta: |\theta-\theta^*| < N^{-1/2}\delta(\lambda)] \tag{67a}$$

$$\Theta_{\lambda N} = \bar{\Theta}_{\lambda N} \cap R_N^M. \tag{67b}$$

By (66), any proposition that holds uniformly for all θ in $\bar{\Theta}_{\lambda N}$ (respectively $\Theta_{\lambda N}$) must hold asymptotically with probability at

least $1-\lambda$ for $\bar{\theta}_N$ (respectively θ_N). Now $\bar{\Theta}_{\lambda N}$ is uncountable but $\Theta_{\lambda N}$ is a finite subset whose cardinality depends on λ but does not vary with N. It is thus often easy to prove propositions uniformly over $\Theta_{\lambda N}$. This opens a convenient approach to proving propositions concerning the random variable θ_N. That is, prove that for each N and $\lambda > 0$, the proposition holds for all θ in $\Theta_{\lambda N}$. Then let $\lambda \to 0$ to prove that the proposition holds in probability, asymptotically, for θ_N.

With the above as preliminaries, Bickel proved that the answer to Question 3 is affirmative for single equation linear models. Our Lemma 5.3 extends this result to single equation non-linear regression models.

Lemma 5.3: For a single equation model of form (7) with i.i.d. disturbances, Lemma 5.1 continues to be satisfied when q_N replaces q_N^*.

Proof: For $\theta\varepsilon\Theta$ and $n=1,\ldots,\infty$ define the residuals

$$u_n(\theta) = y_n - h(x_n,\theta). \tag{68}$$

Let $\Delta = \theta-\theta^*$. Then the relationship between $u_n(\theta)$ and the disturbance $u_n = u_n(\theta^*)$ is

$$u_n(\theta) = u_n - \dot{h}_n' \Delta \tag{69}$$

where $\dot{h}_n = \dfrac{\partial h(x_n,\theta_n^1)}{\partial\theta}$ and θ_n^1 is intermediate between θ^* and θ. Observe that in the notation of equation (61), $f=f^*$, $z_n = u_n(\theta)$, $\mu_{nN} = \dot{h}_n'\Delta$ and $\bar{\mu}_N = \bar{\dot{h}}_N'\Delta$, where $\bar{\dot{h}}_N = N^{-1}\sum \dot{h}_n$. This sets the stage for application of Lemma 5.2.

Fix $\lambda > 0$. By the definition of $\Theta_{\lambda N}$ in (67), $\Delta'\Delta < N^{-1}\delta(\lambda)^2$ for all $\theta\varepsilon\Theta_{\lambda N}$. By this and the Cauchy-Schwarz inequality,

$$\underset{\theta\varepsilon\Theta_{\lambda N}}{\text{Max}}\Big[\underset{n=1,\ldots,N}{\text{Max}}(\mu_{nN} - \bar{\mu}_N)^2\Big]$$

$$< \delta(\lambda)^2 \underset{\theta\varepsilon\Theta_{\lambda N}}{\text{Max}}\Big[\underset{n=1,\ldots,N}{\text{Max}} N^{-1}(\dot{h}_n - \bar{\dot{h}}_N)'(\dot{h}_n - \bar{\dot{h}}_N)\Big] \quad (70)$$

$$\underset{\theta\varepsilon\Theta_{\lambda N}}{\text{Max}}\Big[\sum_{n=1}^{N}(\mu_{nN} - \bar{\mu}_N)^2$$

$$< \delta(\lambda)^2 \underset{\theta\varepsilon\Theta_{\lambda N}}{\text{Max}}\Big[N^{-1}\sum_{n=1}^{N}(\dot{h}_n - \bar{\dot{h}}_N)'(\dot{h}_n - \bar{\dot{h}}_N)\Big]. \quad (71)$$

Now let $N \to \infty$. Recall that by assumption, the empirical distribution of x has almost sure limit G^* and that $\frac{\partial h(x,\theta^*)}{\partial\theta}$ has finite, positive definite variance under G^*. It follows from this and from the fixed cardinality, convergent construction of the sequence of sets $\Theta_{\lambda N}$, $N=1,\ldots\infty$ that the r.h.s. of (70) has limit zero, almost surely in x and the r.h.s. of (71) has finite, positive limit, almost surely in x. Also note that $\bar{\mu}_N = \bar{\dot{h}}_N'\Delta \xrightarrow{a.s.} 0$. Lemma 5.2 then implies that given any sequence of values $\theta_{\lambda N} \varepsilon \Theta_{\lambda N}$, $N=1,\ldots,\infty$, the sequence of densities of the residuals $[u_n(\theta_{\lambda N}), n=1,\ldots,N]$ is contiguous to the sequence of densities of the disturbances $(u_n, n=1,\ldots,N)$, almost surely in x. Therefore, Lemma 5.1 extends to the sequence of score function estimates constructed using the sequence of sets of residuals $[u_n(\theta_{\lambda N}), n=1,\ldots,N]$, $N=1, \infty$.

Since $\theta_N \varepsilon \Theta_{\lambda N}$ with asymptotic probability at least $1-\lambda$, we can conclude that for all $\gamma > 0$,

$$\lim_{N\to\infty} \text{Prob}\Big[\int\Big(q_N(u) - \frac{1}{f^*(u)}\frac{df^*(u)}{du}\Big)^2 f^*(u)du < \gamma\Big] > 1-\lambda. \quad (72)$$

Letting $\lambda \to 0$ proves that (54) is satisfied.

Q.E.D.

The following Corollary is an immediate consequence of Lemma 5.3 and of the Corollary to Lemma 5.1.

Corollary 1: If $f^* \varepsilon F^s$, Lemma 5.3 continues to be satisfied when q_N^s replaces q_N.

A simple extension of Lemma 5.3 is to models involving a finite number of unknown conditional densities.

Corollary 2: Let there exist $A < \infty$ unknown densities f_a^*, $a=1,\ldots,A$. Assume that X partitions into A known, mutually exclusive subsets X_a, $a=1,\ldots,A$ such that $G^*(X_a) > 0$, $\sum_{a=1}^{A} G^*(X_a) = 1$ and $x \varepsilon X_a \rightarrow f_x^* = f_a^*$. Let N(a) be that subset of observations $n=1,\ldots,N$ for which $x_n \varepsilon X_a$ and let $q_{N(a)}$ be the score function estimate constructed using the residuals in N(a). For a single equation model of form (7), $q_{N(a)}$ converges in mean square to $\frac{1}{f_a^*} \frac{df_a^*}{du}$. If $f_a^* \varepsilon F^s$, $\bar{q}_{N(a)}^s$ converges as well.

Proof: Since $G^*(X_a) > 0$, $N \rightarrow \infty \Rightarrow N(a) \xrightarrow{a.s.} \infty$. Lemma 5.3 and Corollary 1 can therefore be applied to $q_{N(a)}$ and $q_{N(a)}^s$ respectively.

Q.E.D.

Corollary 2 will be used in Section 6 to demonstrate that adaptive estimation remains possible in the presence of at least some forms of interdependence between u and x. On the other hand, it is clear that we cannot adapt if the set of

conditional densities is too rich. For example, consider the case in which x has a limiting continuous distribution and there is no prior information relating the conditional densities f_x^*, $x \in X$ to one another. In this setting, the only observations whose residuals can yield information on f_x^* are those for which $x_n = x$ but $N(x)$ does not go to infinity with N. It would be of considerable interest to determine how rich the set of conditional densities can be and convergent score function estimation in the sense of (54) still remain possible.

A second important open question concerns the generalizability of the Stone-Bickel approach to problems of multivariate score function estimation. A natural idea would be to convolute the residuals of a multivariate regression with a multivariate normal random variable, leading to multivariate versions of f_N and q_N. To prove that this works in the manner of Lemma 5.3, however, requires appropriate multivariate generalizations of both Lemma 5.1 and Lemma 5.2. These non-trivial tasks are not attempted here.

5.3 A Sufficient Condition for Successful Adaptation

Return to the general AML problem of finding estimates I_N and S_N that satisfy (48). Bickel (1982) shows that if $I_N - S_N$ is constructed in a certain manner, this problem reduces to one of verifying a simpler condition. Bickel presents his sufficient condition in great generality. For our applications, a less abstract presentation is adequate and may have advantages in clarity.

For given $T < N$, let q_T be the score function estimate constructed using θ_T and the residuals $u_{nT}(\theta_T)$, $n=1,\ldots,T$. Let $\ell(y|x,\theta_N,q_T)$ denote a computable estimate for $\partial \log \lambda(y|x,\theta_N,f_x^*)/\partial\theta$. Now define

$$S_{NT} = \frac{1}{N-T} \sum_{n=T+1}^{N} \ell(y_n|x_n,\theta_N,q_T) \tag{73}$$

The unusual feature in this construction of S_{NT} is the splitting of the sample into two parts. All the observations are used to form the initial estimate θ_N. Then the first T observations are used to estimate the score functions and the last N-T to determine the step taken from θ_N. Splitting the sample in this way is very convenient technically because it allows ones to condition on q_T as a predetermined function when examining the behavior of S_{NT}. Of course, maintenance of desirable asymptotic properties requires that T grow with N in an appropriate manner. It is easy to see that q_T remains a convergent score function estimate in the sense of (54) as long as $T \to \infty$ as $N \to \infty$. Comparison of (73) with (48) indicates that S_{NT} can serve as a successful estimate for $S_N(\theta_N,\phi^*)$ only if $\frac{N-T}{N} \to 1$ as $N \to \infty$. Together these two requirements imply that we should select T so that

$$T \to \infty,\ T/N \to 0 \text{ as } N \to \infty. \tag{74}$$

Note that asymptotic theory gives no guidance on the choice of T for a given data sample. In fact, one should not infer that sample splitting is necessary to successful AML estimation. Stone's estimator for a location parameter does not split the sample nor does it discretize the initial estimate, at the cost of a more difficult proof of adaptation.

Bickel proved

Lemma 5.4: Let $\theta_N \in R_N^M$ be an estimate satisfying (43). Let $\tilde{\theta}_N = \theta_N + I_N^- S_{NT}$ where S_{NT} is constructed in the manner of (73) and (74). Then a sufficient condition for (48) to hold is that as $N \to \infty$,

$$\sqrt{N}\,\frac{\partial\beta}{\partial\theta'}\,\bar{I}_N^{-}\int\left[\int \ell(y|x,\theta_N,q_T)\lambda(y|x,\theta_N,f_x^*)dy\right]dG^* = o_p(1) \quad (75)$$

and

$$\frac{\partial\beta}{\partial\theta'}\int\left[\int\left|\bar{I}_N^{-}\ell(y|x,\theta_N,q_T) - I(\theta_N)^{-1}\,\frac{\partial\log\lambda(y|x,\theta_N,f_x^*)}{\partial\theta}\right|^2 \quad (76)$$

$$\lambda(y|x,\theta_N,f_x^*)dy\right]dG^* = o_p(1).$$

Proof: Bickel (1982) Theorem 3.1 proves the Lemma for the case $\beta^* = \theta^*$ while his Theorem 3.2 extends it to the case where β^* is a sub-vector. Both of these Theorems impose the condition that the l.h.s. expression of (75) actually equal zero. However, it is sufficient that the expression be $o_p(1)$. Bickel implicitly uses this weaker condition in his proof of adaptation for his Example 3.

In the next section, we shall verify Lemma 5.4 in two non-linear regression settings. Before doing this, we should clarify the content of conditions (75) and (76). Condition (75) is essentially equivalent to Condition B, the necessary condition for successful adaptation introduced in Section 3. Condition (76) more or less requires consistent estimation of the information matrix $I(\theta^*)$ and mean square convergent estimation of the score function $\partial\log\lambda/\partial\theta$. This explains our focus in Section 5.2 on the problem of mean square convergent estimation of the score function associated with the density f^*.

6. Applications to Non-Linear Regression Models

We are now in position to extend Bickel's findings for linear models with i.i.d. disturbances to non-linear models and to models allowing some interdependence between u and x. Theorem 6.1 and its Corollary prove that AML estimates do

successfully adapt in the context studied in Proposition 4.1. Theorem 6.2 proves adaptation in the context of Proposition 4.2.

Theorem 6.1: Let $y-h(x,\theta^*) = u$ with $y \varepsilon R^1$. Let $\Phi = [(f)^X, f \varepsilon F^S]$. Let $\theta_N \varepsilon R_N^M$ be an estimate satisfying (43). Let I_N be a consistent estimate for $I(\theta^*)$ and S_{NT} be constructed in the manner of (73) and (74) using

$$\ell(y|x,\theta_N,q_T^s) = -\frac{\partial h(x,\theta_N)}{\partial \theta} q_T^s [y-h(x,\theta_N)]. \tag{77}$$

Then $\tilde{\theta}_N = \theta_N + I_N^{-} S_{NT}$ is adaptive for θ^*.

Proof: Both here and in Theorem 6.2, a menu of suitable initial estimates θ_N exists. For example, discretized versions of the least squares estimates of Jennrich (1969) and White (1980) will do. There also exist a number of satisfactory ways to define I_N. In the present setting,

$$I(\theta^*) = \tag{78}$$

$$\left[\int \frac{\partial h(x,\theta^*)}{\partial \theta} \frac{\partial h(x,\theta^*)}{\partial \theta'} dG^*\right] \cdot \left[\int \left(\frac{1}{f^*(u)} \frac{df^*(u)}{du}\right)^2 f^*(u) du\right].$$

One consistent estimate for $I(\theta^*)$ is

$$I_N = Q_N i_N \tag{79}$$

where

$$Q_N = \frac{1}{N-T} \sum_{n=T+1}^{N} \frac{\partial h(x_n,\theta_N)}{\partial \theta} \frac{\partial h(x_n,\theta_N)}{\partial \theta'} \tag{80}$$

$$i_N = \frac{1}{N-T} \sum_{n=T+1}^{N} q_T^s (u_{nN})^2. \tag{81}$$

By consistency of θ_N and by the weak law of large numbers, Q_N converges in probability to the first bracketted integral in (78). By contiguity of $\{P_N'\}$ to $\{P_N\}$, by Lemma 5.3 and by the

weak law of large numbers, i_N converges in probability to the second integral, that is to $i(f^*)$.

Now consider conditions (75) and (76). With ℓ defined in (77) and with $\beta^* = \theta^*$, the l.h.s. of (75) becomes

$$\sqrt{N}\, I_N^{-} \left[\int - \frac{\partial h(x,\theta_N)}{\partial \theta} dG^*\right] \left[\int q_T^s(u) f^*(u) du\right].$$

By symmetry of f^* and by anti-symmetry of q_T^s, the second bracketted integral is identically zero. By consistency of θ_N and I_N and by non-singularity of $I(\theta^*)$, the leading term is $o_p(\sqrt{N})$. Hence, condition (75) is satisfied.

Given that I_N and $I(\theta_N)$ both converge to $I(\theta^*)$, the l.h.s. of (76) can be written as

$$\left[\int |I(\theta^*)^{-1} \frac{\partial h(x,\theta_N)}{\partial \theta}|^2 dG^*\right] \left[\int \left(q_T^s(u) - \frac{1}{f^*(u)} \frac{df^*(u)}{du}\right)^2 f^*(u) du\right] + o_p(1).$$

The first bracketted integral has a finite probability limit. By Lemma 5.3, the second integral is $o_p(1)$. Hence condition (76) is satisfied.

By Lemma 5.4, $\tilde{\theta}_N$ is adaptive for θ^*.

Q.E.D.

Corollary: Let $y-h(x,\theta^*) = u$ with $y \varepsilon R^1$. Assume that X partitions into $A < \infty$ known, mutually exclusive subsets X_a, $a=1,\ldots,A$ such that $G^*(X_a) > 0$ and $\sum_{a=1}^{A} G^*(X_a) = 1$. Let $\Phi = \prod_{a=1}^{A} \left[(f)^{X_a}, f \varepsilon F^s\right]$. Let $\theta_N \varepsilon R_N^M$ be an estimate satisfying (43). Let I_N be a consistent estimate for $I(\theta^*)$ and let S_{NT} be constructed in the manner of (73) and (74) using

$$\lambda(y|x,\theta_N,q^s_{Ta(x)}) = -\frac{\partial h(x,\theta_N)}{\partial\theta}\, q^s_{Ta(x)}[y-h(x,\theta_N)] \qquad (82)$$

Here $a(x)$ denotes the subset X_a containing x and q^s_{Ta} is the anti-symmetric score function estimate introduced in Lemma 5.3, Corollary 2. Then $\tilde{\theta}_N = \theta_N + I_N^- S_{NT}$ is adaptive for θ^*.

Proof: For each $a=1,\ldots,A$, define Q_{Na} and i_{Na} as in (80) and (81) but using only the sub-sample $N(a)$. The argument in the proof of Theorem 6.1 that $Q_N i_N$ is consistent for $I(\theta^*)$ implies in the setting of this Corollary that $Q_{Na} i_{Na}$ is consistent for $I(\theta^*|X_a)$. Let $N_a = |N(a)|$ and $T_a = |T(a)|$. Then

$$I_N = \sum_{a=1}^{A} \frac{N_a - T_a}{N-T}\, Q_{Na} i_{Na} \qquad (83)$$

is consistent for $I(\theta^*)$.

With ℓ defined in (82) and with $\beta^* = \theta^*$, the l.h.s. of (75) becomes

$$\sqrt{N}\, I_N^- \sum_{a=1}^{A} \left[\int_{X_a} - \frac{\partial h(x,\theta_N)}{\partial\theta}\, dG^*\right]\left[\int q^s_{Ta}(u) f^*_a(u)\, du\right].$$

By the argument of the proof to Theorem 6.1, this expression is identically zero. Likewise, the l.h.s. of (76) can be written as the sum of A terms, each of which is $o_p(1)$. Hence, (75) and (76) are satisfied. By Lemma 5.4, $\tilde{\theta}_N$ is adaptive for θ^*.

Q.E.D.

Theorem 6.2: Let $y - h(x,\theta^*) = y - \tilde{h}(w,\beta^*) - \alpha^* = u$ with $y \in R^1$. Let $\Phi = [(f)^X,\ f \in F^\circ]$. Let $\theta_N \in R_N^M$ be an estimate satisfying (43). Let I_N be the estimate defined in (79), (80), and (81). Let S_{NT} be constructed in the manner of (73) and (74) using

$$\ell(y|x,\theta_N,q_T) = -\frac{\partial h(x,\theta_N)}{\partial\theta}\, q_T[y - h(x,\theta_N)]. \qquad (84)$$

Then $\tilde{\beta}_N = \beta_N + \frac{\partial \beta}{\partial \theta'} I_N^- S_{NT}$ is adaptive for β^*.

Proof: Condition (76) is satisfied by the same argument as used in the proof of Theorem 6.1. With ℓ defined in (84), the l.h.s. of (75) becomes

$$\sqrt{N} \frac{\partial \beta}{\partial \theta'} i_N^- Q_N^- \left[\int - \frac{\partial h(x,\theta_N)}{\partial \theta} dG^*\right] \left[\int q_T(u) f^*(u) du\right].$$

In contrast to the situation in Theorem 6.1, the second bracketted integral is not identically zero. It is, however, $o_p(1)$. This follows from Lemma 5.3 and from the fact that $\int \frac{df^*(u)}{du} du = 0$.

To determine the limiting behavior of the leading expression, recall first that $i_N^- \xrightarrow{p} i(f^*)^{-1}$. Next define

$$\dot{h}_N(x) = \frac{\partial h(x,\theta_N)}{\partial \theta} = \begin{bmatrix} \frac{\partial \tilde{h}(w,\beta_N)}{\partial \beta} \\ 1 \end{bmatrix} = \begin{bmatrix} \dot{\tilde{h}}_N(w) \\ 1 \end{bmatrix}. \tag{85}$$

Using this notation, define

$$\tilde{Q}_N = \frac{1}{N-T} \sum_{n=T+1}^{N} \dot{\tilde{h}}_N(w_n) \dot{\tilde{h}}_N(w_n)', \tag{86}$$

$$\tilde{E}_N = \frac{1}{N-T} \sum_{n=T+1}^{N} \dot{\tilde{h}}_N(w_n) \tag{87}$$

$$\tilde{V}_N = \tilde{Q}_N - \tilde{E}_N \tilde{E}_N'. \tag{88}$$

We can now write

$$\sqrt{N}\,\frac{\partial \beta}{\partial \theta}\,,\ Q_N^{-}\left[\int -\frac{\partial h(x,\theta_N)}{\partial \theta}\,dG^*\right]$$

$$= \sqrt{N}\,[I_{K-1}:0]\begin{bmatrix} \tilde{V}_N^{-} & -\tilde{V}_N^{-}\tilde{E}_N \\ -\tilde{E}_N'\tilde{V}_N^{-} & 1+\tilde{E}_N'\tilde{V}_N^{-}\tilde{E}_N \end{bmatrix}\begin{bmatrix} -\int \dot{\tilde{h}}_N(w)dG^* \\ -1 \end{bmatrix}$$

$$= \tilde{V}_N^{-}\ \sqrt{N}\left[\tilde{E}_N - \int \dot{\tilde{h}}_N(w)dG\right]$$

The variable $\tilde{V}_N^{-}$ has a non-singular probability limit. The expression $\sqrt{N}\left[\tilde{E}_N - \int \dot{\tilde{h}}_N(w)dG^*\right]$ is $O_p(1)$ by the Central Limit Theorem. Hence, the l.h.s. of (75) is $o_p(1)$ as required. By Lemma 5.4, $\tilde{\beta}_N$ is adaptive.

Q.E.D.

7. Some Questions

We have earlier called attention to a number of specific unresolved issues that deserve attention. Do Theorems 6.1 and 6.2 extend to multivariate regression models? Does the Corollary to Theorem 6.1 extend to other forms of interdependence between u and x? In a given sample, how should one select the parameters T, σ_T, b_T, c_T, and d_T in constructing the score function estimate q_T? Can we characterize the situations in which a non-linear systems model of form (5) satisfies Condition B? In this concluding section, we attempt to organize in a coherent manner some more general open questions.

7.1 Attainable Precision When Adaptive Estimates Do Not Exist

Many researchers are initially surprised to learn that adaptive estimates exist in settings as general as those

considered in Theorems 6.1 and 6.2. It remains the case, nevertheless, that adaptation is not possible in "most" estimation problems. It is then natural to ask how well one can do.

An important part of an answer has recently been achieved in a paper by Begun, Hall, Huang, and Wellner (1983). Working in the random sampling context, these authors consider the infinite dimensional problem of joint estimation of (θ^*, f^*). Using projection arguments on Hilbert spaces, they derive the appropriate infinite dimensional generalization of the classical bound on precision of estimation. A special case is that in which $\partial \log \lambda / \partial \theta$ is orthogonal to the score for f^*, a functional derivative $\partial \log \lambda / \partial f$. This is a necessary condition for adaptation. When the scores are not orthogonal, the authors' bound for estimation of θ^* differs from the classical bound given knowledge of f^*. Thus, lack of knowledge of f^* causes a quantifiable loss in attainable precision of estimation of θ^*.

An appealling feature of the Begun et al. work is that its treatment of the finite dimensional parameter θ^* and the functional parameter f^* is entirely symmetric. Consider any estimate θ_N such that $\sqrt{N}(\theta_N - \theta^*)$ has a limiting distribution. The authors show that asymptotically,

$$\sqrt{N}(\theta_N - \theta^*) \stackrel{d}{=} Z_* + W \qquad (89)$$

where $Z_* \sim \mathcal{N}(0, I_*^{-1})$, I_*^{-1} is the bound on precision, and W is an independent random variable. Analogously, consider any estimate F_N of the distribution function $F^* = \int_{-\infty}^{\bullet} f^*(u)du$ such that $\sqrt{N}(F_N - F^*)$ has a limiting stochastic process. The authors show that asymptotically,

$$\sqrt{N}\,(F_N - F^*) \overset{d}{\cong} \zeta_* + \omega \qquad (90)$$

where ζ_* is a Gaussian stochastic process and ω is an independent process. The results (89) and (90) are nonparametric generalizations of the Hajek (1972) convolution theorem characterizing limiting distributions in parametric models.

Begun et al. do not attempt to construct an estimator that achieves the best asymptotic distributions Z_* and ζ_*. Nor do they verify that their bounds are sharp. They do, however, offer an intriguing conjecture. They speculate, that when adaptation is not possible, the nonparametric maximum likelihood estimator is asymptotically efficient under weak regularity conditions.

7.2 Attainable Precision in Other Settings

The analytical framework assumed in this paper is general enough to treat many important econometric problems but certainly not all. For one reason or another, various classes of problems do not satisfy the assumptions imposed in Sections 1 and 2. The following is a partial list.

(i) Models with serially dependent observations - The sample likelihood does not decompose into the product of the likelihoods of y_n conditional on x_n.

(ii) Discrete Dependent Variable Models - For most such models, it is not known whether $\sqrt{N}$ consistent initial estimates exist. For quantal response models there are consistent, distribution free estimators but rates of convergence have not been established.

(iii) Endogenous Sampling Problems - The likelihood under stratified endogenous sampling processes (e.g. truncated sampling, choice-based sampling) is not a convex functional of

the density f^*. Moreover, the sample distribution of x is informative regarding θ^* and f^*.

(iv) Models in which f^* is informative for θ^* - In such problems as Poisson regression, the linear probability model, and exponential family models, the distribution of disturbances is functionally dependent on θ^*.

Among these four classes of problems, I see no fundamental reason why the arguments of Stein-Stone-Bickel should not extend to class (i). Applications to problems of class (ii) hinge on the resolution of the initial estimate question. Problems of classes (iii) and (iv) are not treatable in Bickel's setup but may be using the Begun et al. approach.

7.3 Beyond First Order Efficiency

To close this paper, it is appropriate that we recall the sense in which adaptive estimation is a desirable objective. In the presence of a nuisance density function, an adaptive estimate achieves the first order asymptotic efficiency of the best estimate that would be computable were the density known. In its present state, the literature on adaptive estimation makes no claims beyond first order efficiency. Indeed, the AML estimator is based on a first-order approximation to the likelihood function of a correctly specified model. To the best of my knowledge, there are not now available any theoretical results on the exact distributions, higher order asymptotic properties, or behavior in misspecified models of AML estimates.

In discussions of adaptive estimation, I have occasionally heard more than the usual concern expressed about the limitations of first order asymptotic theory. Satisfactory density estimation, it is feared, requires inordinately large data samples. In small samples, the reasoning goes, adaptive

estimates are likely to be inferior to conventional ones such as ordinary least squares.

I believe that this concern is unfounded. While it is true that nonparametric density estimates converge slowly pointwise, the AML method requires only estimation of the information and of the sample mean score associated with the unknown density. These estimation problems have more in common with the problem of nonparametric estimation of a mean than with that of pointwise nonparametric estimation of a density. Moreover, the smoothing and trimming performed in constructing the score function estimate q_T prevents outlying residuals from being overly influential and constrains the size of the step taken from the initial $\sqrt{N}$-consistent estimate.

A potentially troublesome aspect of AML estimation is the need for the analyst to select the parameters T, σ_T, b_T, c_T, and d_T. Successful adaptation imposes restrictions only on the rates at which these parameters change with the sample size. By its nature first order asymptotic theory can provide no guidance on the parameter settings appropriate for a given sample. The development of a second order theory of adaptive estimation might conceivably yield implications but I see no early prospect of a breakthrough in second order theory.

I have recently begun a series of Monte Carlo experiments designed to reveal the exact distributions of some AML estimates and of certain alternative non-parametric estimates. When these experiments are completed, I plan to report them in a separate paper. For now, I have a few early, suggestive findings to report, as follows.

Consider the model $y = \alpha + \beta x + u$ with with $\alpha=1$, $\beta=-1$, x distributed uniform on $[-1,1]$ and u i.i.d. with density f^* having mean zero and variance one. In the experiments, five alternative densities were used to draw the realizations of u. These include 1) normal, 2) contaminated normal, being the

convolution .9 $\mathcal{N}(0, \frac{1}{9})$ + .1 $\mathcal{N}(0,9)$; 3) log-normal; 4) Type I extreme value, and 5) exponential.

Given each density, a random sample of observations (y,x) was drawn. I shall report findings for N=25 and for N=100. Ordinary least squares provided the initial estimate θ_N. The restriction of θ_N to the lattice R_N^M is ignored here. I chose the parameter values T, σ, b, c, and d in a manner that seemed subjectively reasonable. In particular, for N=25, I set T=10, σ=.08, b=4.0, c=.004 and d=30.0. For N=100, I set T=30, σ=.06, b=5.0, c=.002 and d=36.0. These settings are consistent with the requirements of Lemma 5.3 and Lemma 5.4. In a second set of experiments, I did not split the sample as Bickel calls for. Instead, I set T=N and re-used all the observations to compute the step from θ_N. The values of σ, b, c, and d were not altered. In all the experiments, the score function estimate q_T was used to compute the AML estimate. By Theorem 6.2, the estimate for β is adaptive but that for α need not be.

Each experiment consisted of 400 independent replications in which a sample was drawn and the AML estimates were computed. Table 7.1 presents the results on precision of estimation, as measured by the root mean square errors of the estimates over the 400 replications. The columns labelled "Scale" refer to estimates of the standard error of u. The estimate used in each case is the square root of the sample variance of the residuals.

Inspection of the Table reveals some clear patterns, as follows.

1. The OLS and AML estimates of α do not differ at all in precision.

2. In the case of normal disturbances, the OLS estimates of β slightly outperform the AML ones. The difference in precision is always less than 6 percent for N=25 and less than 4.5 percent for N=100. Since OLS is the maximum likelihood

estimate here, these results are consistent with the theoretical prediction that the AML estimates should approach the MLE in precision as $N \to \infty$.

3. When the true distribution is contaminated normal, log normal, or exponential, the AML estimates outperform the OLS ones. The difference in precision between the OLS and split sample (SAML) adaptive estimates is marginal. On the other hand, the re-used sample (RAML) estimates perform strikingly better than OLS. For N=25, the RAML estimates have root mean square errors, 11, 17, and 4 percent lower than the corresponding OLS ones. For N=100, the RAML root mean square errors are 28, 35, and 18 percent lower than those for OLS. These findings strongly suggest that sample splitting is unnecessary and moreover, not to be recommended in small samples.

4. In the case of extreme value disturbances, the OLS estimates of β slightly outperform the AML ones. The pattern is very similar to that observed under normality. This suggests that OLS may be close to efficient when the distribution of u is extreme value. In fact, examination reveals that for small disturbances, the extreme value likelihood equations are approximated to first order by the OLS normal equations.

5. The OLS and RAML estimates of the scale parameter do not differ at all in their precisions. The SAML estimates are noticeably less precise. This is presumably due to the fact that the SAML estimates are based on sub-samples of size N-T=15, 70 rather than on the full samples.

Overall, the experiments indicate that RAML estimates of β can range from slightly less precise to substantially more precise than OLS ones. This conclusion holds for samples as small as N=25 but is more pronounced in samples of size N=100. For estimation of α and the scale parameter, RAML and OLS have more or less identical precisions. These preliminary results

are encouraging, particularly, in light of the fact that we have not attempted to optimize the settings for σ, b, c and d.

TABLE 7.1 Monte Carlo Experiments

Distribution	Estimator	Root Mean Square Errors of Estimates					
		Alpha		Beta		Scale	
		N=25	N=100	N=25	N=100	N=25	N=100
Normal	OLS	.1919	.0972	.3579	.1842	.1432	.0703
	SAML	.1915	.0973	.3624	.1864	.1843	.0835
	RAML	.1931	.0977	.3793	.1922	.1455	.0707
Contaminated Normal	OLS	.1901	.0998	.3239	.1731	.4519	.2402
	SAML	.1900	.1022	.3216	.1611	.5323	.2791
	RAML	.1896	.0999	.2898	.1242	.4472	.2405
Log Normal	OLS	.1822	.0925	.3361	.1753	.4193	.2715
	SAML	.1896	.0929	.3299	.1669	.4830	.3065
	RAML	.1797	.0910	.2790	.1143	.4157	.2716
Extreme Value	OLS	.2098	.0947	.3599	.1762	.1957	.1040
	SAML	.2120	.0964	.3623	.1762	.2379	.1241
	RAML	.2124	.0950	.3757	.1792	.1961	.1036
Exponential	OLS	.2109	.1003	.3590	.1740	.2676	.1327
	SAML	.2136	.1008	.3679	.1707	.3048	.1590
	RAML	.2127	.0997	.3451	.1419	.2632	.1318

Abbreviations: SAML= split sample AML
RAML= reused sample AML

REFERENCES

Begun, J., Hall, W., Huang, W., and Wellner, J., (1983) "Information and Asymptotic Efficiency in Parametric-Nonparametric Models", Annals of Statistics, 11, 432-452.

Beran, R., (1974) "Asymptotically Efficient Adaptive Rank Estimates in Location Models", Annals of Statistics, 2, 63-74.

Bickel, P., (1982) "On Adaptive Estimation", Annals of Statistics, 10, 647-671.

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

Hajek, J., (1972) "Local Asymptotic Minimax and Admissibility in Estimation", Proc. Sixth Berkeley Symp. Math. Statist. Prob. 1, 175-194, University of California Press, Berkeley.

Huber, P., (1967) "The Behavior of Maximum Likelihcod Estimates Under Nonstandard Conditions", Proc. Fifth Berkeley Symp. Math. Statist. Prob., 1, 221-233, University of California Press, Berkeley.

Jennrich, R., (1969) "Asymptotic Properties of Non-Linear Least Squares Estimation", Annals of Mathematical Statistics, 40, 633-643.

LeCam, L., (1960) "Locally Asymptotically Normal Families of Distributions", University of California Publications in Statistics, 3, 37-98.

LeCam, L., (1969) Theorie Asymptotique de la Decision Statistique, Seminaire de Mathematiques Superieures ete 1968, Les Presses de l'Universite de Montreal, Montreal.

Stein, C., (1956) "Efficient Nonparametric Testing and Estimation", Proc. Third Berkeley Symp. Math. Statist. Prob., 1, 187-196, University of California Press, Berkeley.

Stone, C., (1975) "Adaptive Maximum Likelihood Estimators of a Location Parameter", Annals of Statistics, 3, 267-284.

White, H., (1980) "Non-Linear Regression on Cross-Section Data", Econometrica, 48, 721-746.