Econometrica. Vol. 61 No. 5 (September, 1993), 1199–1223

EFFICIENCY OF WEIGHTED AVERAGE DERIVATIVE ESTIMATORS AND INDEX MODELS[1]

BY WHITNEY K. NEWEY AND THOMAS M. STOKER

Weighted average derivatives are useful parameters for semiparametric index models and nonparametric demand analysis. This paper gives efficiency results for average derivative estimators, including formulating estimators that have high efficiency.

Our analysis is carried out in three steps. First, we derive the efficiency bound for weighted average derivatives of conditional location functionals, such as the conditional mean and median. Second, we derive the efficiency bound for semiparametric index models, where the location measure depends only on indices, or linear combinations of the regressors. Third, we compare the bound for index models with the asymptotic variance of weighted average derivative estimators of the index coefficients.

We characterize the form of the optimal weight function when the distribution of the regressors is elliptically symmetric. In more general cases, we discuss how to combine estimators with different weight functions to achieve efficiency. We derive a general condition for approximate efficiency of pooled (minimum chi square) estimators for index model coefficients, based on weighted average derivatives. Finally, we discuss ways of selecting the type and number of weighting functions to achieve high efficiency.

KEYWORDS: Average derivative, index model, efficiency bound, optimal weights, minimum chi-square, spanning condition.

1. INTRODUCTION

AVERAGE DERIVATIVES ARE USEFUL parameters in a number of semiparametric models. As discussed in Stoker (1986, 1992a), they can be used in estimation of index models, including limited dependent variable models and partially linear regression models. Also, they are used in nonparametric demand estimation, as in Hardle, Hildenbrand, and Jerison (1991). Efficiency of average derivative estimators is a concern, because there are several types that have been proposed. Also, the presence of a nonparametric function estimator might lead to low efficiency. The purpose of this paper is to give efficiency results for average derivative estimators, including formulating estimators that have high efficiency.

Our efficiency analysis proceeds by deriving the semiparametric efficiency bounds for average derivative and index models, and then formulating average derivative estimators with high efficiency. We first derive the bound for weighted average derivatives of conditional location functionals, such as the conditional mean and median. For the conditional mean the previously suggested estimators of Hardle and Stoker (1989) and Stoker (1991) attain this bound. This result is what one would anticipate, because this bound places no restrictions on the data distribution, and in such cases any estimator that is asymptotically equivalent to a sample average and sufficiently regular will be efficient (e.g., see Newey

[1]An earlier version of this paper was presented at the 1988 summer meeting of the Econometric Society at the University of Minnesota. Financial support was provided by the NSF and the Sloan Foundation. Helpful comments were provided by referees and G. Chamberlain, L. Hansen, R. Matzkin, and D. McFadden.

(1990a)). We also give the bound for other location functionals such as the median. We find that when the efficiency of average derivative estimators for different location functionals can be compared, the comparison is similar to that for location models, e.g. with the average derivative conditional median being more efficient than the conditional mean for "fat-tailed" distributions.

Many semiparametric limited dependent variable and regression models take the form of index models, where the location measure (e.g. conditional mean) depends only on linear combinations of the regressors, i.e. on "indices." We derive the semiparametric efficiency bound for these index models. We then compare the bound with the asymptotic variance of weighted average derivative estimators of the index coefficients.

In an index model, consistent estimators arise from the use of different weighting functions, so an important efficiency question is the choice of weights. We derive the form of an efficient weight function when the distribution of the regressors is elliptically symmetric. It is also shown that linearity of certain conditional expectations is necessary for existence of an efficient weight. We discuss the possibility of combining different weights to achieve efficiency, showing that it is possible to obtain an approximately efficient estimator by pooling using minimum chi-square. We discuss how such pooling can lead to efficient estimation of index models and how the number of estimators to combine can be chosen from the data. These results are specialized to derive conditions for achieving approximate efficiency from combining many weighted average derivative estimators. Also, we suggest ways of combining a few weighted average derivative estimators so as to achieve high efficiency.

Other papers give some efficiency results on average derivatives or index models. Chamberlain (1987) previously derived the semiparametric efficiency bound for conditional mean, single index models.[2] Following our initial work Samarov (1990) gave the efficiency bound for the (unweighted) average derivative of the conditional means. Newey (1991a) gives efficiency bounds for linear functionals of mean-square projections (that includes average derivatives of conditional means as a special case). None of these papers gives regularity conditions for the bounds.[3] Recently Hall and Ichimura (1991) derived some efficiency results for index estimators when there is a residual that is independent of the regressors.

2. WEIGHTED AVERAGE DERIVATIVES AND PARTIAL INDEX MODELS

Let y denote a dependent variable, x a $k \times 1$ vector of regressors, $\rho(\varepsilon)$ a loss function of a real-valued variable, and

$$g(x) = \operatorname{argmin}_g E[\rho(y-g)|x] \tag{2.1}$$

[2]We cite the working paper Chamberlain (1987) because the index model bound does not appear in the published version.

[3]To be precise, they do not exhibit a sequence of regular parametric submodels for which the Cramer-Rao bound for the submodel approximates the candidates for the bound they suggest.

Here $g(x)$ is a conditional location function. Examples include the conditional mean for $\rho(\varepsilon) = \varepsilon^2$, the conditional median for $\rho(\varepsilon) = |\varepsilon|$, as well as other more exotic location functionals such as quantiles or expectiles. Our interest is in properties of estimators of weighted average derivatives of $g(x)$. Partition x as $x = (x_1^T, x_2^T)^T$, suppose x_1 is continuously distributed, and for a function $a(x)$, let $a'(x) = \partial a(x)/\partial x_1$. A weighted average derivative of $g(x)$ is

$$(2.2) \qquad \delta = E[w(x)g'(x)],$$

where $w(x)$ is a scalar function. For the average derivative to be well defined, x_1 must be continuously distributed, but x_2 can be discrete.

A primary motivation for weighted average derivatives is a partial index model, with

$$(2.3) \qquad g(x) = G(x_1^T\beta, x_2)$$

Under this model, $\delta = E[w(x)\partial G(x_1^T\beta, x_2)/\partial(x_1^T\beta)] \cdot \beta$, so that the weighted average derivative is proportional to β, and hence can be used to estimate β up to scale. Rodriguez and Stoker (1992) have recently used this model in specification analysis for estimation of conditional means.

The motivation for the index model where $g(x)$ is the conditional mean and x_2 is not present is well known (e.g., see Stoker (1986, 1992a)). Other cases can be motivated by a variety of semiparametric regression models. In particular, suppose that

$$(2.4) \qquad y|x \overset{d}{=} y|(x_1^T\beta, x_2),$$

a *conditional distribution* index model that is analyzed in Newey (1990b). This model allows for y to have a conditional mean and variance (and other moments) that depend on x_2 in an arbitrary way. It is implied by many interesting more restrictive models. For instance, it is implied by $y = \tau(x_1^T\beta + \mu(x_2) + \sigma(x_2)v)$, where v is independent of x and τ is a transformation that can be either known or unknown. The transformation $\tau(r)$ could be $\tau(r) = 1\ (r > 0)$, corresponding to a binary choice model that allows for heteroskedasticity to depend on x_2 in an arbitrary way, or it could be $\tau(r) = \xi^{-1}(r)$, corresponding to a model $\xi(y) = x_1^T\beta + \eta$ that is useful for analyzing duration data. It then implies that equation (2.3) is satisfied, so that partial index models are implied by transformed, semiparametric regression models that allow heteroskedasticity to depend nonparametrically on some regressors.

For the semiparametric model in equation (2.4), the choice of loss function $\rho(\varepsilon)$ can be motivated by efficiency considerations similar to those for the linear model, namely that if the distribution of y is fat-tailed, then a more efficient estimator might be obtained by working with $\rho(\varepsilon)$ that gives less weight to large values of ε. This feature will become apparent from the semiparametric efficiency bounds derived below. Also, comparison of parameters from different choices of $\rho(\varepsilon)$ may allow one to test restrictions on the conditional distribution of y given x, similarly to Koenker and Bassett (1982).

In some cases it is possible to weaken equation (2.4) so that essentially only one choice of $\rho(\varepsilon)$ will produce a partial index model. An important such case is where

$$(2.5) \qquad y = \tau(x_1'\beta + \mu(x_2) + v, x_2),$$

$$\tau(r, x_2) \text{ is monotonic in } r,$$

$$\text{median}(v|x) = 0.$$

Because the median of a monotonic transformation is the transformation of the median, this will be a partial index model for the conditional median, where $\rho(\varepsilon) = |\varepsilon|$, but not for other conditional location measures. This model is a generalization of one considered by Powell (1991). For the case where x_2 is not present, Doksum and Samarov (1992) have suggested using average derivative estimators to estimate $\tau(v)$ and its inverse when $\tau(v)$ is monotonic.

The primary purpose of this paper is to develop the efficiency properties of weighted average derivative and partial index estimators, and discuss how and when different weighted average derivative estimators can be combined into an approximately efficient estimator of a partial index model. It is beyond the scope of this paper to discuss the properties of particular estimators, although the results here have implications for the asymptotic properties of average derivative estimators. As discussed in Newey (1991a), the semiparametric efficiency bound for an unrestricted functional, such as an average derivative, is the asymptotic variance of any sufficiently regular estimator. Thus, one would expect the asymptotic variance of any average derivative estimator to have the form described in Section 3.

Two examples of average derivative estimators are as follows. Consider a kernel estimator that is well-defined even when $\rho(\varepsilon)$ is not smooth (e.g., for $\rho(\varepsilon) = |\varepsilon|$). Let $f(x)$ be the density of x. By integration by parts

$$(2.6) \qquad \delta = E[l(x)g(x)], \quad l(x) = -w'(x) - w(x)f'(x)/f(x).$$

For a kernel $\mathcal{K}(v)$ let $K_h(x) = h^{-k}\mathcal{K}(x/h)$, let $\hat{f}(x) = \sum_{i=1}^n K_h(x - x_i)/n$ be a kernel density estimator and let $\hat{g}(x) = \text{argmin}_g \sum_{i=1}^n K_h(x - x_i)\rho(y_i - g)/n$ be the kernel estimator of Tsybakov (1982). Then an estimator of δ corresponding to equation (2.6) is

$$(2.7) \qquad \hat{\delta} = \left(\sum_{i=1}^n w(x_i)/n\right)^{-1}\left[\sum_{i=1}^n \hat{l}(x_i)x_i/n\right]^{-1}\sum_{i=1}^n \hat{l}(x_i)\hat{g}(x_i)/n,$$

$$\hat{l}(x_i) = -w'(x_i) - w(x_i)\hat{f}'(x_i)/\hat{f}(x_i).$$

For $w(x) = 1$ and $\rho(\varepsilon) = \varepsilon^2$, this estimator is similar to those analyzed in Stoker (1991).[4]

[4]The two leading terms form a nonparametric estimator of the identity matrix, and so do not contribute to the asymptotic variance, but they lead to reduction in a severe finite sample bias in the estimator, as discussed in Stoker (1992b).

Another example is a series estimator with smooth approximating functions. Let $P^K(x)$ be a $K \times 1$ vector of differentiable functions. Suppose that linear combinations of this vector can approximate functions and their derivatives. Consider the estimator

$$(2.8) \qquad \hat{\delta} = \sum_{i=1}^{n} w(x_i)\hat{g}'(x_i)/n, \quad \hat{g}(x) = \hat{\pi}^T P^K(x),$$

$$\hat{\pi} = \operatorname{argmin}_{\pi} \sum_{i=1}^{n} \rho\left(y_i - \pi^T P^K(x_i)\right).$$

This estimator is based on differentiating the series estimator $\hat{g}(x)$ of $g(x)$. For the conditional mean case this estimator is analyzed in Newey (1991a).

Regularity conditions for asymptotic normality of these estimators are beyond the scope of this paper. Nevertheless, the results of this paper do provide a formula for the asymptotic variance of these estimators. As discussed in Section 3, the semiparametric bound we derive will be the asymptotic variance of any estimator that is asymptotically equivalent to a sample average and sufficiently regular.

3. EFFICIENCY FOR WEIGHTED AVERAGE DERIVATIVE ESTIMATION

In this Section we derive the semiparametric efficiency bound for weighted average derivative estimators. The average derivative is an unrestricted parameter, in that its definition places no substantive restrictions on the data distribution. The efficiency bound for estimators of such unrestricted parameters can be calculated as the variance of the *pathwise derivative* of the parameter with respect to the distribution of the data, as shown in Pfanzagl and Wefelmeyer (1982).

To discuss the pathwise derivative it is useful to introduce more terminology. Let $z = (y, x^T)$ denote a single observation and $f(z)$ the true density of z (with respect to some dominating measure). A "parametric submodel" or "path" is a parametric family of densities $f(z|\theta)$ (with respect to some dominating measure) that pass through the truth, i.e., such that $f(z|\theta_0) = f(z)$ for some parameter value θ_0. Let $\delta(\theta)$ denote the value of the average derivative in equation (2.2) when z is distributed as $f(z|\theta)$, and let $S(z)$ denote the score of $f(z|\theta)$ at $\theta = \theta_0$, where typically $S(z) = \partial \ln f(z|\theta)/\partial\theta|_{\theta=\theta_0}$. (A more general definition of $S(z)$ is given in Newey (1990a).) The pathwise derivative is a function $\psi(z)$ with finite mean-square (i.e., $E[\psi(z)^T\psi(z)] < \infty$) such that $E[\psi(z)] = 0$ and for any sufficiently regular submodel,

$$(3.1) \qquad \partial\delta(\theta)/\partial\theta|_{\theta=\theta_0} = E\left[\psi(z)S(z)^T\right].$$

When the distribution of z is unrestricted, as it is here, it is typically possible to show that a parametric submodel can be chosen so that the score $S(z)$ approximates any function of z with mean zero and finite mean-square. In that

case, a lower bound on the asymptotic variance of semiparametric estimators of δ is

$$(3.2) \qquad V = E\left[\psi(z)\psi(z)^T\right].$$

Briefly, the idea behind this formula is that by the reasoning of Stein (1956), V should be the supremum of Cramer-Rao bounds over all parametric submodels. The Cramer-Rao bound for a submodel is

$$E\left[\psi(z)S(z)^T\right]\left(E\left[S(z)S(z)^T\right]\right)^{-1}E\left[S(z)\psi(z)^T\right]$$

by equation (3.1) and the "delta-method." This matrix is bounded above by V and it is approximately equal to V when $S(z)$ is approximately equal to $\psi(z)$.

Another interpretation of $\psi(z)$ is as the *influence function* of an efficient estimator $\hat{\delta}$, satisfying

$$(3.3) \qquad \sqrt{n}\,(\hat{\delta} - \delta) = \sum_{i=1}^{n} \psi(z_i)/\sqrt{n} + o_p(1).$$

The "influence function" terminology, which originated in the robust estimation literature, is motivated by the fact that in large samples $\psi(z)$ approximately gives the effect of a single observation on $\hat{\delta}$. It is a convenient way to think about the asymptotic properties of $\sqrt{n}$-consistent estimators, because most will satisfy equation (3.3) for some influence function (e.g., mean average derivative estimators, as shown in Stoker (1991)). By the central limit theorem the asymptotic variance of $\hat{\delta}$ will be $E[\psi(z)\psi(z)^T]$, equal to the bound when the influence function equals the pathwise derivative in equation (3.2).

In general, any influence function will be a pathwise derivative; i.e., if an estimator satisfies equation (3.3) for *some* $\psi(z)$, and certain regularity conditions hold, then its influence function satisfies equation (3.1) (e.g., see Newey (1990a)). When the model imposes no substantive restrictions on the data distribution, the set of scores is unrestricted (except for having mean zero), as it is in this Section. Therefore, there can be at most one pathwise derivative, and hence at most one influence function.[5] As discussed in Newey (1991a), this fact can be used to find the influence function of any estimator, by calculating the pathwise derivative of the parameter that is estimated under unrestricted distributions. In particular, the influence function of an average derivative estimator will be equal to the pathwise derivative derived below, because we impose no substantive restrictions on the data distribution. Thus, we are justified in interpreting V not only as the variance bound for average derivative estimators but also as the asymptotic variance of any (regular) average derivative estimator satisfying equation (3.3). For instance, the bound derived below will be the asymptotic variance of the kernel and series estimators suggested at

[5]For two different influence functions $\psi(z)$ and $\tilde{\psi}(z)$, choosing $S(z)$ (approximately) equal to $\tilde{\psi}(z) - \psi(z)$ and differencing equation (3.1) implies $0 = E[\{\tilde{\psi}(z) - \psi(z)\}^T S(z)] = E[\{\tilde{\psi}(z) - \psi(z)\}^T\{\tilde{\psi}(z) - \psi(z)\}]$.

the end of Section 2, under the plausible assumptions that they satisfy equation (3.3) and are regular.

Because the form of the pathwise derivative is the main result of this Section, we will first give this form and discuss it, postponing the regularity conditions until the end of the Section. Let $\varepsilon = y - g(x)$ and

$$(3.4) \qquad u = -\nu(x)^{-1} m(\varepsilon), \qquad m(\varepsilon) = d\rho(\varepsilon)/d\varepsilon,$$
$$\nu(x) = dE[m(y-g)|x]/dg|_{g=g(x)}.$$

Theorem 3.1 below shows that the pathwise derivative is

$$(3.5) \qquad \psi(z) = w(x)g'(x) - \delta + l(x)u$$
$$= w(x)g'(x) - \delta - l(x)\nu(x)^{-1} m(y - g(x)),$$

where $l(x)$ is given in equation (2.6). The semiparametric variance bound is then $E[\psi(z)\psi(z)^T]$.

Two interesting special cases are the conditional mean and median. For the mean, $u = \varepsilon$, so that

$$(3.6) \qquad \psi(z) = w(x)g'(x) - \delta + l(x)[y - g(x)].$$

For the median, $u = [2f(0|x)]^{-1}\,\mathrm{sgn}(\varepsilon)$, where $f(0|x)$ is the conditional density of ε given x at $\varepsilon = 0$ and $\mathrm{sgn}(\varepsilon) = 1(\varepsilon > 0) - 1(\varepsilon < 0)$, so that

$$(3.7) \qquad \psi(z) = w(x)g'(x) - \delta + l(x)[2f(0|x)]^{-1}\,\mathrm{sgn}(y - g(x)).$$

In general, the variance bound can be decomposed into two terms.[6] By $E[u|x] = 0$,

$$(3.8) \qquad E\left[\psi(z)\psi(z)^T\right] = \mathrm{Var}(w(x)g'(x)) + E\left[u^2 l(x)l(x)^T\right].$$

The first term is the asymptotic variance bound when $g(x)$ is known, being the asymptotic variance of $\sum_{i=1}^n w(x_i)g'(x_i)/n$. The second term is the bound when $f(x)$ is known, by the following argument: since $\psi(z)$ still satisfies equation (3.1) for more restrictive parametric submodels where the distribution of x is known and scores satisfy $E[S(z)|x] = 0$ (because θ parameterizes the conditional distribution of y given x), then $E[\psi(z)S(z)^T] = E[l(x)uS(z)^T]$. Also, $S(z)$ can approximate any conditional mean zero function, and hence can approximate $l(x)u$. Thus, the argument for equation (3.2) also applies here.

The magnitude of the second term $E[u^2 l(x)l(x)^T] = E[E[u^2|x]l(x)l(x)^T]$, corresponding to unknown g, depends on $\rho(\varepsilon)$ similarly to the way the asymptotic variance of location estimators depends on the loss function. In particular, $l(x)$ does not depend on the form of $\rho(\varepsilon)$, while $E[u^2|x] = \nu(x)^{-2}E[m(\varepsilon)^2|x]$ is equal to the bound for estimation of the location parameter $\mathrm{argmin}_\mu\, E[\rho(\varepsilon - \mu)]$ if ε_i were i.i.d. with density $f(\varepsilon_i|x)$, where $m(\varepsilon)$ and $\nu(x)$ are given in equation (3.4). Thus, when average derivatives for different $\rho(\varepsilon)$ functions are equal, so that the corresponding bounds can be compared, the comparisons can be

[6] We thank Gary Chamberlain for suggesting the following interpretation.

carried out in a way similar to that for estimation of location parameters. For example, when the conditional density of ε given x has "thick tails" for most values of x, the second term of the bound will tend to be smaller for the conditional median than for the conditional mean.

Turning now to the statement of regularity conditions, we first give an assumption that is essential to the result.

ASSUMPTION 3.1: *$w(x)f(x)$ is zero on the boundary of the support of x.*

This condition allows us to ignore boundary terms in the derivation of the bounds. Without this assumption, $E[w(x)g'(x)]$ may include boundary terms that depend on $g(x)$ evaluated at particular points. For continuously distributed regressors, the value of a conditional expectation at a point has an infinite variance bound, so that the average derivative will not be $\sqrt{n}$-consistently estimable. For a simple example, suppose that x is a scalar that is uniformly distributed on $[0,1]$, $g(x)$ is the conditional mean, and $w(x) = 1$. Then

$$\delta = E[g'(x)] = \int_0^1 g'(x)\,dx = g(1) - g(0). \tag{3.9}$$

It is easy to show that the semiparametric variance bound is infinite in this case, which is consistent with the well known fact that the value of a conditional expectation at a particular point is not $\sqrt{n}$-consistently estimable.

One way to guarantee that this assumption holds in index models, is to choose $w(x)$ to be zero outside the interior of the support of x. Of course, such a choice might be in conflict with the efficient choice of weights discussed in Section 5.

Additional regularity conditions are useful for deriving the result.

ASSUMPTION 3.2: *The support $\mathscr{X}$ of x is convex and compact, there is compact $\mathscr{G} \subset \mathbb{R}$ containing the closure of $g(\mathscr{X})$ in its interior such that $E[m(y-g)|x] = 0$ has a unique solution at $g(x)$ for $g \in \mathscr{G}$, and $\text{Prob}(v(x) \neq 0) = 1$. Also, $E[\|f'(x)/f(x)\|^2] < \infty$, $E[\psi(z)\psi(z)^T]$ exists and is nonsingular, $w(x)$ and $w'(x)$ are bounded on $\mathscr{X}$, the conditional distribution of y given x has conditional density $f(y|x)$ such that $f(y|x)^{1/2}$ is mean-square continuously differentiable in x_1 on $\mathscr{X}$, and for any $\zeta(y,x)$ that is bounded and continuously differentiable in x_1 with bounded derivatives, $E[\zeta(y,x)|x]$ is continuously differentiable in x_1 on $\mathscr{X}$, and $E[m(y-g)\zeta(y,x)|x]$ is continuously differentiable in x_1 and g on $\mathscr{X} \times \mathscr{G}$.*

The last smoothness condition does not seem very primitive, but it is straightforward to give sufficient conditions for particular $m(y-g)$. For example, for $m(\varepsilon) = \varepsilon$ it will follow from the other assumptions and continuity of $E[y^2|x]$, and for $m(\varepsilon) = \text{sgn}(\varepsilon)$ it will be implied by $f(y|x)$ being absolutely continuous with $f(y+\alpha|x)^{1/2}$ mean-square continuously differentiable in the scalar α and

x_1.[7] This assumption will not be satisfied if y is discrete and $m(\varepsilon)$ is not continuous, over the range of $y - g(x)$, because (for $\zeta(y, x) = 1$) $E[m(y - g)|x]$ will not be continuous in g. In particular, it does not hold if y is binary and $m(\varepsilon) = \text{sgn}(\varepsilon)$.

As is well known, the class of estimators must be restricted to obtain an efficiency bound result. We do this by restricting attention to estimators $\hat{\delta}$ that are *regular*, meaning that for a class of regular parametric submodels the limiting distribution of $\sqrt{n}(\hat{\delta} - \delta(\theta_n))$ does not depend on $\{\theta_n\}$ when $\sqrt{n}(\theta_n - \theta_0)$ is bounded and the data have distribution $f(z|\theta_n)$ for each sample size n.

THEOREM 3.1: *If Assumptions* 2.1 *and* 2.2 *are satisfied, and* $V = E[\psi(z)\psi(z)^T]$ *is nonsingular for* $\psi(z)$ *in equation* (3.5), *then* V *is the supremum of Cramer-Rao bounds of all regular parametric models such that* $\partial\delta(\theta)/\partial\theta|_{\theta=\theta_0} = E[\psi(z)S(z)^T]$. *Also, any estimator* $\hat{\delta}$ *of* δ *that is regular satisfies* $\sqrt{n}(\hat{\delta} - \delta_0) \xrightarrow{d} Z^* + U$ *where* Z^* *is distributed as* $N(0, V)$ *and* U *is independent of* Z^*. *Furthermore, if* $\sqrt{n}(\hat{\delta} - \delta_0) = \sum_{i=1}^n \bar{\psi}(z_i)/\sqrt{n} + o_p(1)$, *where* $E[\bar{\psi}(z_i)] = 0$ *and* $E[\bar{\psi}(z_i)^T\bar{\psi}(z_i)] < \infty$, *and* $\hat{\delta}$ *is regular, then* $\bar{\psi}(z_i) = \psi(z_i)$.

In the environment considered here, where data distribution is not restricted, the assumption that the pathwise derivative formula holds for a parametric submodel is just a convenient regularity condition. It is verified in the proof of the theorem that there exists a class of regular parametric models where the pathwise derivative formula is satisfied, with score that can be chosen to approximate any random vector in mean square.

The last conclusion of this theorem shows that any influence function must equal that given in equation (3.5). In particular, the series and kernel estimators described in equations (2.7) and (2.8) will have $\psi(z)$ as their influence function, and hence asymptotic variance $E[\psi(z)\psi(z)^T]$, as long as they satisfy equation (3.3) (for some $\psi(z)$) and are sufficiently regular.

4. EFFICIENCY BOUNDS FOR MULTIPLE INDEX MODELS

In this Section we derive the variance bound for the parameters of multiple index models, where the function $g(x)$ described earlier is restricted to depend on a function of x and parameters. Let $v(x, \beta)$ be a vector of functions of x and a $q \times 1$ parameter vector β. A multiple index model is one where there is a function $G(v)$ such that

$$g(x) = G(v(x, \beta_0)). \tag{4.1}$$

An important example is the partial index model discussed in Section 2, where $v(x, \beta) = (x_1^T\beta, x_2^T)^T$. The pathwise derivative can be used to calculate the efficiency bound for estimators of β, although this approach must be modified

[7]Assumption 3.2 follows in these cases by Lemmas C.2 and C.3 of Newey (1991b) and by $E[m(y - g)\zeta(y, x)|x] = \int m(\varepsilon)\zeta(u + g, x)f(u + g|x)\,du$ when y is continuously distributed.

to account for the restrictions imposed by equation (4.1). We now carry out this calculation, using tangent set and projection methods.

Because β is now an implicit parameter rather than an explicit functional, a more specific parameterization of the model is useful. Consider parameterizing the submodels by $\theta = (\beta^T, \eta^T)^T$, where η is a nuisance parameter vector for any feature of the distribution of z other than β. Let S_β and S_η denote the respective scores, where for notational convenience we have suppressed the z argument. Then equation (3.1) for a pathwise derivative reduces to

$$(4.2) \qquad E\left[\psi(z)S_\beta^T\right] = I, \quad E\left[\psi(z)S_\eta^T\right] = 0.$$

By the same reasoning following equation (3.2), the efficiency bound will be the variance of $\psi(z)$ such that this equation is satisfied and $\psi(z)$ can be approximated by a linear combination of $S = (S_\beta^T, S_\eta^T)^T$. This $\psi(z)$ can be found by a projection calculation. Let $\mathscr{T}$ be the mean-square closure of the union of all $q \times 1$ linear combinations of all possible nuisance scores, i.e., $\mathscr{T} = \{\varepsilon(z)$: $\exists \varepsilon > 0$, constant matrix C, S_η with $E[\|\varepsilon - CS_\eta\|^2] < \varepsilon\}$, referred to as the *tangent set*. Assuming that $\mathscr{T}$ is a linear set, let $\bar{\varepsilon}$ be the mean square projection of S_β on $\mathscr{T}$, that is characterized by the two conditions $\bar{\varepsilon} \in \mathscr{T}$ and $E[(S_\beta - \bar{\varepsilon})^T \varepsilon] = 0$ for all $\varepsilon \in \mathscr{T}$, and let $S = S_\beta - \bar{\varepsilon}$. The random vector S is referred to as the *efficient score*. Then $\psi(z) = (E[SS^T])^{-1}S$ will satisfy equation (3.1) and can be approximated by a linear combination of scores, so that the semiparametric variance bound for β will be $(E[SS^T])^{-1}$. Begun, Hall, Huang, and Wellner (1983) and Bickel, Klaasen, Ritov, and Wellner (1992) developed this projection form of the bound.

The form of the efficient score is the main result of this section, so we first present and discuss it, and then give regularity conditions. Let $v = v(x, \beta_0)$ and $v_\beta = [\partial v(x, \beta_0)/\partial\beta]^T \partial G(v)/\partial v$, where by convention each component of $\partial G(v)/\partial v$ is set equal to zero if the corresponding component of $v(x, \beta)$ does not depend on β. Theorem 4.1 shows that the efficient score is

$$(4.3) \qquad S = \sigma^{-2}(x)u\left\{v_\beta - E\left[\sigma^{-2}(x)v_\beta|v\right]/E\left[\sigma^{-2}(x)|v\right]\right\},$$

$$\sigma^2(x) = E[u^2|x] = \nu(x)^{-2}E\left[m(\varepsilon)^2|x\right].$$

Thus, the semiparametric variance bound is

$$(4.4) \qquad V = (E[SS^T])^{-1}\Big(E\left[\sigma^{-2}(x)v_\beta v_\beta^T - E\left[\sigma^{-2}(x)v_\beta|v\right]\right.$$

$$\left. \times E\left[\sigma^{-2}(x)v_\beta^T|v\right]/E\left[\sigma^{-2}(x)|v\right]\right]\Big)^{-1}.$$

Although the bound is complicated, it has a straightforward interpretation in terms of an optimally weighted m-estimator where the regression function $G(v)$ has been "concentrated out." Assuming that $\nu(x) > 0$, let $\omega(x) = 1/[\nu(x)\sigma^2(x)]$, and

$$\tilde{G}(v(x,\beta),\beta) = \operatorname{argmin}_{g(v(x,\beta))} E\left[\omega(x)\rho(y-g)\right]$$

$$= \operatorname{argmin}_{g \in \mathbb{R}} E\left[\omega(x)\rho(y-g)|v(x,\beta)\right].$$

Thus, this function minimizes the population value of a weighted m-estimation criterion. Consider the estimator of β that minimizes the sample counterpart to this criteria, with $\tilde{G}(v(x,\beta),\beta)$ substituted for g,

$$(4.5) \qquad \hat{\beta} = \operatorname{argmin}_\beta \sum_{i=1}^{n} \omega(x_i)\rho\big(y_i - \tilde{G}\big(v(x_i,\beta),\beta\big)\big)$$

This is an estimator where the unknown function $G(v)$ has been replaced by the function that minimizes the population counterpart, i.e. where $G(v)$ has been "concentrated out" in the population. By the usual formula for a parametric m-estimator, $\hat{\beta}$ will have asymptotic variance $(E[\sigma^{-2}(x)\tilde{G}_\beta(x)\tilde{G}_\beta(x)^T])^{-1}$ for $\tilde{G}_\beta(x) = \partial\tilde{G}(v(x,\beta),\beta)/\partial\beta|_{\beta=\beta_0}$. This asymptotic variance is the semiparametric bound, because[8]

$$(4.6) \qquad \tilde{G}_\beta(x) = v_\beta - E\big[\sigma^{-2}(x)v_\beta|v\big]/E\big[\sigma^{-2}(x)|v\big].$$

This interpretation suggests an approach to efficient estimation that proceeds by replacing $\omega(x)$ and $\tilde{G}(v(x,\beta),\beta)$ in equation (4.6) by nonparametric estimators. In the single index, conditional mean case, Ichimura's (1993) weighted kernel estimator that uses known $\omega(x)$ can be interpreted as an estimator of $\tilde{G}$ when $\omega(x) = \operatorname{var}(\varepsilon|x)^{-1}$ is known. In general, the estimation of G and $\omega(x)$ will not affect the asymptotic variance, so that this estimator will be efficient. In particular it follows by Proposition 2 of Newey (1991a) that the replacement of G by a nonparametric estimator will not affect the asymptotic variance, essentially because G has been "concentrated out." Also, as usual in m-estimation, estimation of $\omega(x)$ will not affect the asymptotic distribution under appropriate regularity conditions.

The first assumption gives regularity conditions for the distribution of the data as a function of β. Let $G(v(x,\beta),\beta)$ denote the location functional when β is the true parameter and $\varepsilon(\beta) = y - G(v(x,\beta),\beta)$. The way that G can depend on β directly will be left unspecified, because it does not affect the form of the bound. Let $E_\beta[\cdot]$ and $E_\eta[\cdot]$ denote expectations for a parametric submodel when $\eta = \eta_0$ and $\beta = \beta_0$, respectively, of β. Let $\mathscr{X}$ denote the support of x.

ASSUMPTION 4.1: (*i*) *The marginal distribution of x does not depend on β*; (*ii*) *$\beta \in \mathscr{B}$ for an open set β such that on $\mathscr{X}\times\mathscr{B}$, $v(x,\beta)$ is bounded and continuously differentiable in β with bounded derivative, $G(v,\beta)$ is bounded and continuously differentiable in β and v with bounded derivatives, the conditional distribution of y given x at β has density $f(y|x,\beta)$ that is regular in β with probability one with (conditional) information matrix that is nonsingular and bounded, $E_\beta[m(\varepsilon(\beta))^2|x]$ is bounded and bounded away from zero*; (*iii*) *there is a compact set $\mathscr{S}$ containing in its interior the closure of $G(v(\mathscr{X},\mathscr{B}),\mathscr{B})$ such that*

[8]The moment restriction $E[m(\varepsilon)|x] = 0$ implies that $\partial E[\omega(x)m(\varepsilon)|v(x,\beta)]/\partial\beta = \partial E[\omega(x)E[m(\varepsilon)|x]|v(x,\beta)]/\partial\beta = 0$. Then differentiation of the first order conditions $E[\omega(x)m(y - \tilde{G}(v(x,\beta),\ \beta))|v(x,\beta)] = 0$ with respect to β gives $\partial\tilde{G}(v,\beta)/\partial\beta = -E[\sigma^{-2}(x)v_\beta|v]/E[\sigma^{-2}(x)|v]$, so that differentiation separately with respect to the two β arguments in $\tilde{G}(v(x,\beta),\ \beta)$ gives equation (4.6).

on $\mathscr{X}\times\mathscr{B}\times\mathscr{I}$, $E_\beta[m(y-g)|x]$ *is continuously differentiable with bounded derivative;* (*iv*) $\partial E_\beta[m(y-g)|x]/\partial g|_{g=G(v(x,\beta),\beta)} > 0$ *and is bounded away from zero,* $G(v(x,\beta),\beta)$ *solves* $E_\beta[m(y-g)|x]=0$, *and* $\partial E_\beta[m(\varepsilon)|x]/\partial\beta = E[m(\varepsilon)S_\beta|x]$.

This Assumption consists of more or less standard regularity conditions. The next hypothesis imposes additional smoothness conditions.

ASSUMPTION 4.2: *For any function* $\zeta(y,x,\beta)$ *that is bounded, continuously differentiable in* y *and* β, *and has bounded derivative, the integral* $E[m(y-g)\zeta(y,x,\beta)|x,\beta]$ *is continuously differentiable in* β *and* g *with bounded derivative. Also, there exists a bounded, continuously differentiable function* $\tilde{m}(\varepsilon)$ *with bounded derivatives such that* $E_\beta[\tilde{m}(\varepsilon(\beta))m(\varepsilon(\beta))|x]$ *is bounded away from zero uniformly in* x,β.

The first hypothesis is similar to the last part of Assumption 3.2, so that primitive conditions for this condition can be specified as in Section 3. Also, it is straightforward to formulate more primitive conditions for the second hypothesis with particular $m(\varepsilon)$. By $E_\beta[m(\varepsilon(\beta))^2|x]$ bounded and bounded away from zero it suffices to find $\tilde{m}(\varepsilon)$ such that $E_\beta[\{\tilde{m}(\varepsilon(\beta))-m(\varepsilon(\beta))\}^2|x]$ is small uniformly in x and β. For example, in the conditional mean case, with $m(\varepsilon)=\varepsilon$, if $E_\beta[|\varepsilon|^{2+\epsilon}|x]$ is bounded on $\mathscr{X}\times\mathscr{B}$, then choosing $\tilde{m}(\varepsilon)$ so that $|\tilde{m}(\varepsilon)|\leq|\varepsilon|$ and $\tilde{m}(\varepsilon)=\varepsilon$ except when $|\varepsilon|$ is big enough will satisfy this assumption. Also, in the conditional median case with $\varepsilon=\operatorname{sgn}(\varepsilon)$ with ε continuously distributed given x with bounded conditional density $f(\varepsilon|x)$, choosing $|\tilde{m}(\varepsilon)|\leq 1$ and $\tilde{m}(\varepsilon)=\operatorname{sgn}(\varepsilon)$ except in a small enough neighborhood of zero will satisfy this assumption.

THEOREM 4.1: *Suppose that Assumptions* 4.1 *to* 4.3 *are satisfied and* $E[SS^T]$ *is nonsingular for* S *from equation* 4.3. *Consider the class of parametric submodels such that* $G(v,\eta)$ *solves* $E_\eta[m(y-g)|x]=0$, $E_\eta[m(y-g)|x]$ *is continuously differentiable in a neighborhood of* $(G(v,\eta_0),\eta_0)$, *and* $\partial E_\eta[m(\varepsilon)|x]/\partial\eta = E[m(\varepsilon)S_\eta|x]$. *Then* $V=(E[SS^T])^{-1}$ *is the supremum of Cramer-Rao bounds and any estimator* $\hat{\beta}$ *of* β_0 *that is regular satisfies* $\sqrt{n}(\hat{\beta}-\beta_0)\xrightarrow{d} Z^*+U$ *where* Z^* *is distributed as* $N(0,V)$ *and* U *is independent of* Z^*.

Of particular interest for studying weighted average derivatives is the case where $v(x,\beta)=(x_1^T\beta,x_2^T)^T$, corresponding to a partial index model. Here β is only identified up to scale, so we normalize the first coefficient to be one, with $v(x,\beta)=(x_{11}+x_{12}^T\beta,x_2^T)^T$. In this case, for $G_1(v)=\partial G(\nu,x_2)/\partial\nu|_{\nu=x_{11}+x_{12}T_{\beta_0}}$, the efficient score is

$$S=\sigma^{-2}(x)u\cdot G_1(v)\{x_2-E[\sigma^{-2}(x)x_2|v]/E[\sigma^{-2}(x)|v]\}. \tag{4.7}$$

5. EFFICIENCY OF WEIGHTED AVERAGE DERIVATIVE ESTIMATORS OF INDEX COEFFICIENTS

As we discussed in Section 2, average derivatives estimate the parameters of partial index models up to scale. In this Section we consider the efficiency of average derivative estimators of index model coefficients. We adopt the same normalization as before, where the first coefficient equals one. For $g(x) = G(x_{11} + x_{12}^T\beta, x_2)$, let $\delta_1 = E[w(x)\partial g(x)/\partial x_{11}]$ and $\delta_2 = E[w(x)\partial g(x)/\partial x_{12}]$. Assuming that $w(x)$ obeys the identification condition $\delta_1 \neq 0$, it will be the case that $\delta_2/\delta_1 = \beta$, so that $\hat{\beta} = \hat{\delta}_2/\hat{\delta}_1$ will be consistent for β. The asymptotic efficiency of $\hat{\beta}$ is analyzed in this Section.

To evaluate the efficiency of this estimator we will assume that $\hat{\delta}$ satisfies equation (3.3), having influence function given in equation (3.5). This assumption is justified because the influence function of any (regular) average derivative estimator satisfies equation (3.5), by Theorem 3.1. We also assume that equation (4.2) is satisfied, a standard regularity condition that is known to be a consequence of regularity of the estimator (e.g., see Newey (1990a)).

5.1. *The Asymptotic Variance of Relative Average Derivative Estimators*

Given that $\hat{\delta}$ has influence function in equation (3.5), the influence function of $\hat{\beta}$ (and hence its asymptotic variance) can be derived by the delta method. Let $v = x_{11} + x_{12}^T\beta$ and note that for any function $a(x) = a(x_{11}, x_{12}, x_2)$

$$[-\beta, I]a'(x) = \partial a(v - x_{12}^T\beta, x_{12}, x_2)/\partial x_{12}, \tag{5.1}$$

i.e., $[-\beta, I]a'(x)$ is the partial derivative of $a(x)$ with respect to x_{12}, holding v constant. Recalling that x_2 is included in v, $[-\beta, I]g'(x) = 0$ since $g(x)$ depends only on v. Also, $\partial(\delta_2/\delta_1)/\partial\delta = \delta_1^{-1}[-\beta, I]$, so that by the delta method the influence function of $\hat{\beta}$ will equal

$$\begin{aligned}\psi(z) &= \delta_1^{-1}[-\beta, I]\{w(x)g'(x) - \delta - \{w'(x) + w(x)f'(x)/f(x)\}\}u \\ &= l(x)u,\end{aligned} \tag{5.2}$$

$$l(x) = -\delta_1^{-1}[-\beta, I]\{w'(x) + w(x)f'(x)/f(x)\}.$$

For notational convenience we use the same $\psi(z)$ and $l(x)$ notation here as in Sections 2 and 3, even though the expressions are not the same. The asymptotic variance of $\hat{\beta}$ is then $E[\psi(z)\psi(z)^T]$.

It is interesting to note that the term $\text{var}(w(x)g'(x))$ in the average derivative bound has dropped out of the asymptotic variance of the index estimator. This result is consistent with the interpretation of this term as the one that accounts for the density of x being unknown. Since β is a feature of the conditional distribution of y given x, the distribution of x is ancillary for estimation of β, and knowledge of this distribution should not affect the efficiency of estimators for β.

We next consider the efficiency of $\hat{\beta}$ as an estimator in the partial index model of Sections 2 and 4, where $\rho(\varepsilon)$ is given. In this model the efficiency of $\hat{\beta}$ depends on the weight function $w(x)$, and it would be useful to know whether a weight function can be chosen so that $\hat{\beta}$ is efficient, with asymptotic variance equal to the bound of Section 4. In this subsection we consider existence and the form of such an efficient weight.

Our first result is that an efficient weight function will exist when x has an elliptically symmetric distribution and $\sigma^2(x)$ depends only on v, as described in the following result.

THEOREM 5.1: *Suppose that (i) x has an elliptically symmetric distribution, with nonsingular variance and density $f((x-\mu)^T\Lambda(x-\mu))$ for a differentiable $f(\cdot)$, constant vector μ and positive definite matrix Λ; (ii) $\sigma^2(x) = E[u^2|v] = \sigma^2(v) > 0$ and $G_1(v)$ is nonzero with probability one. Then an efficient weight function is*

$$(5.5) \qquad w(x) = \sigma^{-2}(v)G_1(v)/h\big((x-\mu)^T\Lambda(x-\mu)\big),$$

$$h(q) = f(q)/\int_0^q f(r)\,dr.$$

The optimal weight depends on the density of the regressors through the inverse of the hazard $h(q)$. The regressor density does not effect the weight if and only if $h(q)$ is constant, corresponding to exponential $f(q)$ and hence normally distributed x. In cases where $f(q)$ is "thicker tailed" than normal (e.g. $f(q)$ proportional to $1/(1+q^a), a > 1$), $1/h(q)$ will tend to give more weight to larger values of $(x-\mu)^T\Lambda(x-\mu)$, while if $f(q)$ is "thinner tailed" than normal (e.g. $f(q)$ proportional to $\exp(-\varepsilon^a), a > 1$) it will tend to give less weight. The declining weight implicit in the density weighted estimator of Powell, Stock, and Stoker (1989) would tend to have high efficiency when the x distribution has thinner tails than normal, although it is difficult to find an example where $1/h(q)$ behaves exactly like $f(q)$.

Some properties of elliptically symmetric distributions are essential to existence of an efficient weight, because of implicit constraints on $l(x)$. Let $x = (x_{12}^T, v^T)^T$ and x_{-k} denote the vector of all elements of x other than the kth. Then $l(x)$ is constrained in the following way.

LEMMA 5.2: $E[l_k(x)|x_{-k}] = 0, k \leq q-1$.

This result can be interpreted in terms of the amount of information used by a single average derivative ratio. Each $l_k(x)$ is the term multiplying u in the influence function of $\hat{\beta}_k$. Also, each $\hat{\beta}_k$ uses only the restriction that the index is of the form $x_{11} + x_k\beta_k$, and allows for $g(x)$ to depend on x_{-k} in an arbitrary way. In other words, $\hat{\beta}_k$ is consistent for the coefficient in a partial index model with index $x_{11} + x_k\beta_k$. Lemma 5.2 is a consequence. It is exactly the condition

that makes the influence function uncorrelated with nuisance parameter scores in such a partial index model, as required by consistency of $\hat{\beta}_k$, as in equation (4.2). In contrast, the elements of

$$l^*(x) = G_1(v)\sigma^{-2}(x)\{x_{12} - E[\sigma^{-2}(x)x_{12}|v]/E[\sigma^{-2}(x)|v]\}$$

from the efficient score only have conditional expectation zero given v. This results from the index model imposing more restrictions than any individual average derivative ratio, leading to more information (variance) in the efficient score.

Theorem 5.2 shows that it is sometimes possible to combine the information from the individual derivative ratios to obtain efficiency. Intuitively, although the individual terms have less information than the efficient score, together they can have as much, because the index interpretation of the vector of derivative ratios comes from the index model. However, the requirement that a *single* average derivative vector contains all the information is quite restrictive, relying on linearity of certain conditional expectations as a necessary condition.

THEOREM 5.3: *Suppose that* $\sigma^2(x) > 0$ *depends only on* v, $Prob(u = 0) = 0$, $G_1(v)$ *is nonzero with probability one, and* $E[SS']$ *is nonsingular. If an efficient weight function exists then for each* $k \leq q - 1$ *there is a vector* c_k *such that*

$$E[x_k|x_{-k}] = E[x_k|v] + c_k^T\{x_{-k} - E[x_{-k}|v]\}.$$

Thus, linearity of $E[x_k|x_{-k}]$ in x_{-k} for each k is necessary for existence of an efficient weight, when $\sigma^2(x)$ depends only on v. We could also derive a result for the case where $\sigma^2(x)$ depends on x in a more general way, but for simplicity we have not allowed for this generality here.

Although x having nonlinear conditional expectations will rule out the existence of an efficient weight, approximate efficiency can still be achieved by combining influence functions from many derivative estimators. Intuitively, combining average derivatives from different ratios imposes the index model information, that can lead to efficiency if results from different weighting functions are used. Unfortunately, the efficient combination will not generally exist in closed form, and hence is difficult to describe.

We use certain Hilbert space results to argue that average derivatives can be combined to achieve approximate efficiency, but not in closed form. The basic result that is useful here is that the closure of the direct sum of closed linear subspaces is equal to the orthogonal complement of the intersection of orthogonal complements of the subspaces. Take the Hilbert space to be the usual one of functions of x with finite mean square. Take the subspaces to be the set of functions of x satisfying Lemma 5.2 for each k. These have orthogonal complement equal to the set of functions of x_{-k}. The intersection (over $k \leq q - 1$) of this set is the set of functions of v. This intersection has orthogonal complement equal to the set of functions that have conditional mean zero given v. Then, it follows by $E[l^*(x)|v] = 0$ that each element of $l^*(x)$ is in the

closure of $\{l_1(x) + \cdots + l_{q-1}(x)\colon E[l_k(x)|x_{-k}] = 0\}$. Thus, the terms that depend on x in the efficient score, can be approximated by the terms in the average derivative that depend on x, which will lead to approximate efficiency. However, it is impossible to give an explicit form for this additive decomposition, because a direct sum of subspaces need not be closed. Also, even when it is closed the decomposition into the sum does not generally have an explicit form, except when the subspaces are orthogonal or finite dimensional. Here the subspaces $\{l_k(x)\colon E[l_k(x)|x_{-k}] = 0\}$ are never orthogonal, and many finite dimensional cases are covered by Theorem 5.1. Thus, in general an explicit form for the optimal combination of different weighted average derivative estimators cannot be found.

5.3. *Pooling Weighted Average Derivatives Via Minimum Chi-Square*

In a general setting, more efficient estimation will be possible by combining different weighted average derivative estimators. Also, when an efficient weight exists, it may have components that are unknown, and pooling estimators with known weights gives a feasible approach to efficient estimation. In this context, questions arise about how to choose weights for applications, and how to combine the resulting weighted average derivative estimators. In this section we show how minimum chi square estimation provides a sound method of combining estimators, showing the requirements for a chosen weight sequence to yield an overall efficient estimator. While we do not detail methods for choosing weights in practice, natural methods are easy to devise. For instance, one could use a sequence of weights whose first terms are appropriate when the regressors are normal, then add terms to allow for nonnormal but elliptically symmetric distributions, and finally add terms that would account for nonelliptical distributions. We give a specific formulation of this scheme following the main results and assumptions of this section.

In our framework, the minimum chi-square method of pooling different estimators is described as follows. Let $J \geqslant 2$ linearly independent weighting functions $w_j(x)$, $(j = 1, \ldots, J)$, be specified, let $\hat{\delta}_j$ denote the weighted averaged derivative estimator using $w_j(x)$, and let $\hat{\beta}_j = \hat{\delta}_{j2}/\hat{\delta}_{j1}$ be the associated ratio estimator. Stack the separate estimators into a vector $\hat{\gamma} = (\hat{\beta}_1^T, \ldots, \hat{\beta}_J^T)$ and let $H = [I, \ldots, I]^T = e_J \otimes I$, where e_J is the J vector of ones and I is the identity matrix with dimension equal to the number of elements of β. Let $\psi_j(z)$ denote the influence function of $\hat{\beta}_j$ (satisfying equation (5.2) for $w(x) = w_j(x)$) and let $\Psi(z) = (\psi_1(z)^T, \ldots, \psi_J(z)^T)^T$. The asymptotic variance of $\hat{\gamma}$ is then $\Omega = E[\Psi(z)\Psi(z)^T]$. Let $\hat{\Omega}$ be a consistent estimator of Ω, such as $\hat{\Omega} = \sum_{i=1}^n \hat{\Psi}(z_i)\hat{\Psi}(z_i)^T/n$ for $\hat{\Psi}(z) = (\hat{\psi}_1(z)^T, \ldots, \hat{\psi}_J(z)^T)^T$, with $\hat{\psi}_j(z)$ obtained by replacing all the unknown components in equation (5.2) by nonparametric estimators. The pooled estimator is

$$(5.6) \qquad \tilde{\beta} = \operatorname{argmin}_\beta (\hat{\gamma} - H\beta)^T \hat{\Omega}^{-1} (\hat{\gamma} - H\beta) = (H^T \hat{\Omega}^{-1} H)^{-1} H^T \hat{\Omega}^{-1} \hat{\gamma}.$$

It follows from standard minimum chi-square theory that $\hat{\beta}$ is $\sqrt{n}$-consistent and asymptotically normal with asymptotic variance $(H^T\Omega^{-1}H)^{-1}$ that is no larger than the asymptotic variance of any linear combination of the $\hat{\beta}_j$ $(j = 1,\ldots,J)$. Also, a consistent estimator of the asymptotic variance will be $(H^T\hat{\Omega}^{-1}H)^{-1}$.

The minimum chi-square estimator will be efficient if the efficient score is a linear combination of the influence functions. Also, it will be approximately efficient for large J if a linear combination of the influence functions can approximate the efficient score in mean-square. We state these results in the following theorem, also giving an interpretation of the difference of the efficient information matrix and the minimum chi-square precision matrix.

THEOREM 5.4: *$V^{-1} - H'\Omega^{-1}H = \min_\Pi E[\{S - \Pi\Psi\}\{S - \Pi\Psi\}^T]$, so that if there exists Π such that $S = \Pi\Psi$ then $\hat{\beta}$ is efficient. Furthermore, if for each J there exists Π_J such that $\lim_{J\to\infty} E[\|S - \Pi_J\Psi\|^2] = 0$ then $\lim_{J\to\infty}(H'\Omega^{-1}H)^{-1} = V$.*

The second condition, on existence of a mean-square approximation of the scores by the influence function, is a "spanning condition" for efficiency. It is the (minimum chi-square) analog of the generalized method of moments spanning condition given in Newey (1990b).

To achieve approximate efficiency by combining weighted average derivative estimators, the weights will have to satisfy a spanning condition corresponding to that of Theorem 5.4. This spanning condition is quite complicated, because the influence functions depend on both the weight and its derivative. For this reason, we do not try to give as general a spanning condition as possible, instead focusing on conditions that are relatively easy to verify. The first of these conditions involves the data distribution. Let $f(x)$ denote the density of x and let X_k denote the random variable with realization x_k.

ASSUMPTION 5.1: *x has compact support, $\sigma^2(x)$ is bounded and bounded away from zero, $E[\|f'(x)/f(x)\|^2] < \infty$, for $k \leq q-1$ and any positive integer r, $\partial f(x_k|x_{-k})\partial x_k$ and $f(x_k|x_{-k})^{-2}\partial f(x_k|x_{-k})/\partial x_k \operatorname{cov}(1(X_k \leq x_k), X_k^r|x_{-k})$ are continuous on the support of x.*

The last part of this condition does not seem primitive, but it is straightforward to check for particular distributions. In particular, as long as $f(x_k|x_{-k}) > 0$ on the interior of the support, the last expression will be continuous on the interior, so that it suffices to show continuity on the boundary.[9]

[9]In a previous version of this paper it was shown that this continuity condition is satisfied when $f(x_k|x_{-k})$ is a beta density, proportional to $x_k^a(1-x_k)^b$ for coefficients a, b that are continuous functions of x_{-k}.

When Assumption 5.1 is satisfied there are simple conditions on the weight functions for the spanning condition to be satisfied. Let $x_k(\epsilon, x)$ denote the vector with ϵ in the kth position and other components equal to the corresponding components of x_{-k}.

ASSUMPTION 5.2: *$w_{jJ}(x) = w_0(v)p_{jJ}(x)$ with $1/C < w_0(v) < C$ for some $C > 0$. The functions $\{p_{jJ}(x)\}$ satisfy the following condition; for each $k \leqslant q - 1$, there is a set $\mathscr{B} \subseteq \mathbb{R}$, $inf(\mathscr{B}) = -\infty$, such that for any $\varepsilon > 0$, continuous function $a(x)$, and $b \in \mathscr{B}$, such that there exists $\bar{J}$ and $\pi_{1J}, \ldots, \pi_{jJ}$ with $\sup_{\mathscr{X}} |\sum_{j=1}^{J} \pi_{jJ} \partial p_{jJ}(x)/\partial x_k - b(x)| < \varepsilon$ and $\sum_{j=1}^{J} \pi_{jJ} p_{jJ}(x) = \sum_{j=1}^{J} \pi_{jJ} \int_b^{x_k} \partial p_{jJ}(x_k(t, x))/\partial x_k \, dt$ for $J > \bar{J}$.*

This hypothesis says that the partial derivatives with respect to each x_i can approximate any continuous function, and that the linear combination $\sum_{j=1}^{J} \pi_{jJ} p_{jJ}(x)$ is a definite integral of the partial derivative. It is easy to check that this hypothesis is satisfied for particular choices of $p_{jJ}(x)$. For example, suppose $p_{jJ}(x)$ is a power series with all terms of a given integer order (sum of exponents) and below included. Then the hypotheses follow by the Weirstrass theorem and because derivatives and definite integrals of power series are also power series. Also, for similar reasons, this hypothesis will be satisfied for Gallant's (1981) flexible Fourier form.

An important assumption here is that $w_0(v)$ is a function of v and $p_{jJ}(x)$ is a function of x, both of which depend on the unknown index $x_{11} + x_{12}^T\beta$, so that these weights are not feasible. They can be made feasible by replacing β with an estimator. Because the choice of weights does not affect the consistency of the estimators, under appropriate regularity conditions the replacement of x with its corresponding estimated value will have no effect on the limiting distribution of the minimum chi-square estimator, and hence no effect on its efficiency.

The next result shows that Assumptions 5.1 and 5.2 are sufficient for the spanning condition for near efficiency of minimum chi-square.

THEOREM 5.5: *If Assumptions 5.1 are satisfied, then $\lim_{J \to \infty} (H'\Omega^{-1}H)^{-1} = V$.*

Our proposal for using a weight sequence that is initialized at normal and then elliptically symmetric regressors can be formulated as follows. Suppose $\hat{G}_1(v)$ and $\hat{\sigma}^2(v)$ are obtained from a preliminary parametric estimator of the index model.[10] Let $\hat{\mu}$ and $\hat{\Sigma}$ be the sample mean and variance of the regressors and $\omega_j(\cdot)$ be functions of a scalar argument, with $\omega_1(\cdot) = 1$, and other ω_j allowing for elliptically symmetric distributions other than normal, such as $\omega_j(\varphi) = [\varphi/(1 + \varphi)]^{j-1}$. Then a weight sequence that is initialized at normal

[10]For example, if y is binomial, $g(x)$ is the conditional mean, and $v = x_{11} + x_{12}^T\beta$, then corresponding to a preliminary probit estimator one might choose $\hat{G}_1(v) = \phi(\hat{a} + \hat{b} \cdot v)$ and $\hat{\sigma}^2(v) = \Phi(\hat{a} + \hat{b} \cdot v)[1 - \Phi(\hat{a} + \hat{b} \cdot v)]$ where $\hat{a}$ is the constant and $\hat{b}$ the coefficient of x_{11} from probit.

and elliptically symmetric regressors is as given in Assumption 5.2 for

$$w_0(v) = \hat{G}_1(v)/\hat{\sigma}^2(v), \quad p_j(x) = \omega_j\big((x-\hat{\mu})^T\hat{\Sigma}^{-1}(x-\hat{\mu})\big), \quad j \leq \bar{J},$$

where $\bar{J}$ is some (small) positive integer. Higher (than $\bar{J}$) order terms might include powers of x. As noted above, estimation of the weights will not affect the asymptotic distribution of relative average derivative estimators.

Practical applications also involve the choice of the number of weighting functions to use. One proposal for doing this is to use the minimum chi-square analogue of the *GMM* cross validation criteria developed in Newey (1990b). For brevity, we omit the explicit formulation of this method. However, there are reasons for optimism regarding the practical performance of such methods. For instance, Newey (1990b) gives an example where the choice from cross-validation delivered virtually all the efficiency gains available.

5.4. *Efficiency For Conditional Distribution Index Models*

When the conditional index distribution model of equation (2.4) holds, relative average derivatives will estimate the same coefficients for different $\rho(\varepsilon)$ functions, so that their asymptotic efficiencies can be compared. The asymptotic variance is $E[\sigma^2(x)l(x)l(x)^T]$ for $\sigma^2(x) = \nu(x)^{-2}E[m(\varepsilon)^2|x]$. As discussed in Section 3, $\sigma^2(x)$ is the asymptotic variance for estimation of the location parameter $\mu(x) = \text{argmin}_\mu E[\rho(y-\mu)|x]$, so that the asymptotic variances of relative average derivative estimators depend on $\rho(\varepsilon)$ in a way that is analogous to the location parameter. Here this dependence is even more direct, since the first term in the average derivative has dropped out. For example, if the conditional distribution of y given x is "fat-tailed" enough at each x so that the median has lower asymptotic variance than the mean, then the asymptotic variance of a weighted average derivative estimator based on the median will be smaller than that based on the mean.

It is possible to approximately attain the semiparametric bound for the conditional index model, that is given in Newey (1990b), by combining relative average derivative estimators for different weights and $\rho(\varepsilon)$ functions. A sufficient spanning condition is that the weighted average derivative estimators are calculated from all combinations of a sequence of weights satisfying Assumptions 5.1 and 5.2 and a sequence of $m(\varepsilon)$ functions for which linear combinations can, for the conditional distribution of ε given any x, approximate any function with finite mean square. For brevity, we have omitted a full description and the proof of this result.

Dept. of Economics, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

and

Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

Manuscript received June, 1988; final revision received March, 1993.

APPENDIX

PROOFS OF THEOREMS

Throughout the Appendix let C denote a generic matrix of positive constants that may be different in different appearances.

PROOF OF THEOREM 3.1: Let $\tilde{\zeta}(y,x)$ and $\tilde{\gamma}(x)$ be bounded and continuously differentiable, let $\zeta(y,x) = \tilde{\zeta}(y,x) - E[\tilde{\zeta}|x]$, $\gamma(x) = \tilde{\gamma}(x) - E[\tilde{\gamma}(x)]$, and for the density $f(z)$ of z consider the parametric submodel

$$(A.1) \qquad f(z|\theta) = f(z)[1 + \theta^T\zeta(y,x)][1 + \theta^T\gamma(x)].$$

Both $\zeta(y,x)$ and $\gamma(x)$ are bounded, so this is a density function for θ close enough to $\theta_0 = 0$. Also,

$$(A.2) \qquad E_\theta[a|x] = E[a|x] + \theta^T E[a\zeta|x].$$

In addition, mean-square continuous differentiability of $f(z|\theta)^{1/2}$ in a neighborhood Θ of $\theta_0 = 0$ follows by Lemma C.4 of Newey (1991b), with score

$$(A.3) \qquad S_\theta(z) = \zeta(y,x) + \gamma(x).$$

By Assumption 3.2 and Lemma C.2 of Newey (1991b), $E[\zeta(y,x)|x]$ is continuously differentiable in x_1 with bounded derivatives, and hence so is $\zeta(y,x)$, so that $E[m(y-g)|x]$ and $E[m(y-g)\zeta(y,x)|x]$ are continuously differentiable in (x_1, g) on $\mathscr{X}\times\mathscr{G}$, and hence by equation (A.2), $E_\theta[m(y-g)|x]$ is continuously differentiable on $\mathscr{X}\times\mathscr{G}\times\Theta$. In particular, by continuity Θ can be chosen small enough that for all $\theta \in \mathscr{G}$ there exists a unique solution $g(x,\theta)$ to $E_\theta[m(y-g)|x] = 0$ for $g \in \mathscr{G}$. By the implicit function theorem, for each x, θ in $\mathscr{X}\times\Theta$ there is a neighborhood such that $\partial g(x,\theta)/\partial\theta$ and $g'(x,\theta)$ exist and are continuous on that neighborhood, so that by compactness of $\mathscr{X}, \Theta$ can be chosen small enough that $\partial g(x,\theta)/\partial\theta$ and $g'(x,\theta)$ exist, are continuous and bounded on $\mathscr{X}\times\Theta$, and

$$(A.4) \qquad \partial g(x,\theta)/\partial\theta|_{\theta=0} = -\nu(x)^{-1}E[m(\varepsilon)\zeta|x] = E[u\zeta|x].$$

Next, the marginal density of x in the parametric submodel $f(z|\theta)$ is $f(x|\theta) = f(x)[1 + \theta\gamma(x)]$ so that by integration by parts,

$$(A.5) \qquad \delta(\theta) = E_\theta[w(x)g'(x,\theta)] = E[w(x)g'(x,\theta)] + \theta^T E[w(x)\gamma(x)g'(x,\theta)]$$
$$= E[l(x)g(x,\theta)] - \theta^T E\left[f(x)^{-1}[w(x)\gamma(x)f(x)]'(x,\theta)\right].$$

By $g(x,\theta)$ differentiable in θ with bounded derivative, $\delta(\theta)$ is differentiable in θ, and by equation (A.4) and another integration by parts,

$$(A.6) \qquad \partial\delta(\theta)/\partial\theta|_{\theta=0} = E\left[l(x)\partial g(x,\theta)/\partial\theta^T|_{\theta=0}\right] - E\left[f(x)^{-1}[w(x)\gamma(x)f(x)]'g(x)\right]$$
$$= E\left[l(x)u\zeta^T\right] + E[w(x)g'(x)\gamma(x)] = E\left[\psi(z)S_0(z)^T\right].$$

Thus, $\psi(z)$ is the pathwise derivative for all parametric submodels as specified above.

Next, it follows, e.g. by Lemma C.7 of Newey (1991b), that for any $s(z)$ with finite mean-square and $E[s(z)] = 0$, and for any $\varepsilon > 0$, there are $\zeta(y,x)$ and $\tilde{\gamma}(x)$ satisfying the above boundedness and smoothness hypotheses with $E[\|s(z) - \tilde{\zeta}(y,x)\|^2] < \varepsilon$ and $E[\|E[s|x] - \tilde{\gamma}(x)\|^2] < \varepsilon$, so that

$$(A.7) \qquad E\left[\|s(z) - S_\theta(z)\|^2\right] = E\left[\|s(z) - E[s|x] + E[s|x] - \zeta - \gamma\|^2\right]$$
$$\leq 2E\left[\|s(z) - E[s|x] - \zeta\|^2\right] + 2E\left[\|E[s|x] - E[s] - \gamma\|^2\right]$$
$$\leq 4\varepsilon,$$

where the last inequality follows by the Cauchy-Schwartz inequality. The first conclusion then follows by Bickel, Klaasen, Ritov, and Wellner (1992, Chapter 3, Theorem 2).

To obtain the second conclusion, note that $E_\theta[\|\psi(z)\|^2]$ is continuous in θ. Then by regularity and Theorem 2.2 of Newey (1990a), $\partial\delta(\theta)/\partial\theta|_{\theta=0} = E[\psi(z)D_\theta(z)^T]$, so by equation (A.6),

$e[(\psi(z)-\psi(z))S_\theta(z)^T]=0$. The second conclusion then follows because $S_\theta(z)^T$ can approximate any mean zero vector function in mean square, and hence can approximate $\psi(z)-\psi(z)$, so that $E[\|\psi(z)-\psi(z)\|^2]=0$. *Q.E.D.*

We prove Theorem 4.1 using two Lemmas. The first gives the form of the tangent set and the second the projection on the tangent set.

LEMMA A.1: *If Assumptions* 4.1 *are satisfied then the tangent set, for all parametric submodels satisfying the hypotheses of Theorem* 4.1, *is*

$$\mathscr{T}=\left\{\varepsilon(z): E\left[\varepsilon(z)^2\right]<0,\ E[\varepsilon(z)]=0,\ E[\varepsilon u|x]=E[\varepsilon u|v]\right\}.$$

PROOF: We prove this result by showing that any nuisance score must lie in $\mathscr{T}$ and exhibiting a class of parametric submodels that can approximate anything in $\mathscr{T}$ in mean square. Consider first any nuisance score for a submodel satisfying the hypotheses of Theorem 4.1. By the implicit function theorem $G(v,\eta)$ is differentiable at η_0 and $\partial G(v,\eta)/\partial\eta=-\nu(x)^{-1}\partial E_\eta[m(\varepsilon)|x]=E[uS_\eta|x]$. Thus, $E[uS_\eta|x]$ is a function of v and so S_η lies in $\mathscr{T}$.

To construct a regular parametric submodel with score that can approximate anything in $\mathscr{T}$, let $\mathscr{D}$ denote the statement "the function is bounded, continuously differentiable in y and β, and has bounded derivatives." Let $\tilde\zeta(y,x)$ satisfy $\mathscr{D}$. Then by Assumption 4.1 and Lemma C.2 of Newey (1991b), $E_\beta[\tilde\zeta(y,x)|x]$ is continuously differentiable in β, with derivative $E_\beta[\tilde\zeta(y,x)S_\beta(y|x)|x]$, where $S_\beta(y|x)$ is the (conditional) score for $f(y|x,\beta)$. This matrix is bounded by $\tilde\zeta$ bounded and boundedness of the conditional information $E_\beta[S_\beta(y|x)S_\beta(y|x)^T|x]$. Thus, $\bar\zeta(y,x,\beta)=\tilde\zeta(y,x)-E_\beta[\tilde\zeta(y,x)|x]$ satisfies $\mathscr{D}$. Also, let $\bar m(\varepsilon)$ be as specified in Assumption 4.2. Then $\overline{\bar m(\varepsilon(\beta))}=\bar m(\varepsilon(\beta))-E_\beta[\bar m(\varepsilon(\beta))|x]$ satisfies $\mathscr{D}$ by Assumption 4.1 and Lemma C.2 of Newey (1991b), so that by Assumptions 4.1 and 4.2 $E_\beta[\overline{\bar m(\varepsilon(\beta))}m(\varepsilon(\beta))|x]$ and $E_\beta[\bar\zeta(y,x,\beta)m(\varepsilon(\beta))|x]$ satisfy $\mathscr{D}$. For a function $\alpha(v)$ that is continuously differentiable with bounded derivative, let

$$\begin{aligned}\bar\zeta(y,x,\beta)=\tilde\zeta(y,x,\beta)-\overline{\bar m(\varepsilon(\beta))}\left(E_\beta[\bar m(\varepsilon(\beta))m(\varepsilon(\beta))|x]\right)^-\\ \times\left\{E_\beta[\tilde\zeta(y,x,\beta)m(\varepsilon(\beta))|x]-\alpha(v(x,\beta))\right\}.\end{aligned}$$

Then this function satisfies $\mathscr{D}$ and $E_\beta[\zeta(y,x,\beta)|x]=0$ by construction. Therefore, for any function $\zeta(x)$ with mean zero, $\Delta(z,\theta)=(1+\eta^T\zeta(y,x,\beta)(1+\eta^T\zeta(x))$ is continuously differentiable in $\theta=(\beta,\eta)$ with bounded derivative and is bounded away from zero and one for η in a small enough neighborhood of zero, and $E_\beta[\Delta(z,\theta)]=0$, so that by Lemma C.4 of Newey (1991b), $f(z|\theta)=f(z|\beta)\Delta(z,\theta)$ is a density with mean-square continuously differentiable square root, with

$$\text{(A.8)}\qquad S_\eta=\bar\zeta(z)-\overline{\bar m(\varepsilon)}\left(E[\bar m(\varepsilon)m(\varepsilon)|x]\right)^{-1}\left\{E[\bar\zeta(z)m(\varepsilon)|x]-\alpha(v)\right\}.$$

Thus, $f(z|\theta)$ is smooth. To show that it is a parametric submodel, note that

$$\text{(A.9)}\qquad E_\theta[m(\varepsilon)|x]=\eta^T\alpha(v(x,\beta)).$$

By Assumption 4.1, $E_\theta[m(y-g)|x]=E_\beta[m(y-g)|x]+\eta^TE_\beta[m(y-g)\zeta(y,x,\beta)|x]$ is continuously differentiable in g and by Assumption 4.1, there is $\epsilon>0$ such that for all η small enough $\partial E_\theta[m(y-g)|x]/\partial g$ is bounded away from zero on $[G(v(x,\beta),\beta)-\epsilon, G(v(x,\beta),\beta)+\epsilon]$, uniformly in x and β. Therefore, by equation (A.9), for all η small enough there is $\nu(v(x,\beta),\eta)$ in a neighborhood of zero such that

$$\begin{aligned}&\text{(A.10)}\qquad 0=E_\theta[m(\varepsilon+\nu(v(x,\beta),\eta)|x]=E_\theta[m(y-G(v(x,\beta),\theta))|x],\\ &G(v(x,\beta),\theta)=G(v(x,\beta),\beta)+\nu(v(x,\beta),\eta).\end{aligned}$$

This is a local minimum of $E_\theta[\rho(y-g)|x]$ by continuous differentiability of $E_\theta[m(y-g)|x]$ and the derivative bounded away from zero, and a global minimum for small enough η by the theorem of the maximum and compactness of $\mathscr{A}$. Thus, $f(z|\theta)$ satisfies the index restriction, and hence is a smooth parametric submodel. Furthermore, the other hypotheses in Theorem 4.1 for the parametric submodels are satisfied by construction.

To show that S_η can approximate anything in $\mathscr{T}$, note that for $\varepsilon\in\mathscr{T}$, by boundedness of $E[m(\varepsilon)^2|x]$, $E[\|E[m(\varepsilon)\varepsilon|v]\|^2]=E[\|E[m(\varepsilon)\varepsilon|x]\|^2]=E[E[m(\varepsilon)^2|x]E[\|\varepsilon\|^2|x]]\leq CE[\|\varepsilon\|^2]$. Then

for any $\epsilon > 0$ it follows by Lemma C.7 of Newey (1991b) that there exists $\bar{\zeta}(y,x)$, $\alpha(v)$, and $\gamma(x)$ that are bounded and continuously differentiable with bounded derivative such that $E[\|\varepsilon - \bar{\zeta}\|^2] < \epsilon$, $E[\|E[m(\varepsilon)\varepsilon|v] - \alpha(v)\|^2] < \epsilon$, and $E[\|E[\varepsilon|x] - \gamma(x)\|^2] < \epsilon$. Also, $E[\|\varepsilon - E[\varepsilon|x] - \bar{\zeta}\|^2] \leq CE[\|\varepsilon - \bar{\zeta}\|^2] + CE[\|E[\varepsilon|x] - E[\bar{\zeta}|x]\|^2] \leq C\epsilon$. Therefore,

$$
\begin{aligned}
(A.11)\qquad E\left[\|\varepsilon - S_\eta\|^2\right] &\leq C\Big\{E\left[\|\varepsilon - E[\varepsilon|x] - \bar{\zeta}\|^2\right] + E\left[\|E[\varepsilon|x] - \zeta(x)\|^2\right] \\
&\qquad + E\Big[\mathrm{Var}(\overline{m}(\varepsilon)|x)\left(E[\overline{m}(\varepsilon)m(\varepsilon)|x]\right)^{-2}\|E\left[m(\varepsilon)\bar{\zeta}|x\right] - \alpha(v)\|^2\Big]\Big\} \\
&\leq C\epsilon + CE\left[\|E\left[m(\varepsilon)\bar{\zeta}|x\right] - E[m(\varepsilon)\varepsilon|x] + E[m(\varepsilon)\varepsilon|v] - \alpha(v)\|^2\right] \\
&\leq C\epsilon + CE\left[\|E\left[m(\varepsilon)(\bar{\zeta} - \varepsilon)|x\right]\|^2\right] + CE\left[\|E[m(\varepsilon)\varepsilon|v] - \alpha(v)\|^2\right] \\
&\leq C\epsilon + CE\left[E\left[m(\varepsilon)^2|x\right]E\left[\|\bar{\zeta} - \varepsilon\|^2|x\right]\right] \leq C\epsilon. \qquad \textit{Q.E.D.}
\end{aligned}
$$

LEMMA A.2: *If $E[u^2] < \infty$ and $E[\sigma^{-2}(x)] < \infty$ then the projection of a $q \times 1$ random vector s with finite mean-square on $\mathscr{T}$ is*

$$
s - E[s] - uR(x), \quad R(x) = \sigma^{-2}(x)\left\{E[u \cdot s|x] - E\left[\alpha^{-2}(x)u \cdot s|v\right]/E\left[\sigma^{-2}(x)|v\right]\right\}.
$$

PROOF: For notational simplicity let $q = 1$. By the Cauchy-Schwartz inequality,

$$
E\left[\left\{\sigma^{-2}(x)uE[u \cdot s|x]\right\}^2\right] \leq E\left[E\left[\left\{u/\sigma^2(x)\right\}^2|x\right]E\left[u^2|x\right]E\left[s^2|x\right]\right] = E[s^2] < \infty,
$$

and

$$
\begin{aligned}
&E\left[\left\{\sigma^{-2}(x)uE[\sigma^{-2}(x)u \cdot s|v]/E\left[\sigma^{-2}(x)|v\right]\right\}^2\right] \\
&\quad \leq E\left[\sigma^{-2}(x)E\left[\sigma^{-2}(x)^2u^2|v\right]E\left[s^2|v\right]/E\left[\sigma^{-2}(x)|v\right]^2\right] \\
&\quad = E\left[\sigma^{-2}(x)E\left[\sigma^{-2}(x)^2u^2|v\right]E\left[s^2|v\right]/E\left[\sigma^{-2}(x)|v\right]^2\right] \\
&\quad = E\left[E\left[\sigma^{-2}(x)^2E[u^2|x]|v\right]E\left[s^2|v\right]/E\left[\sigma^{-2}(x)|v\right]\right] = E[s^2],
\end{aligned}
$$

where all the expressions are finite by $E[\sigma^{-2}(x)] < \infty$ (implying $E[\sigma^{-2}(x)|v]$ exists and $\mathrm{Prob}(\sigma^2(x) = 0) = 0$). Thus, for the expression $\bar{\varepsilon}$ given in the statement of the Lemma, $E[\bar{\varepsilon}^2] < \infty$. Also, note that

$$
\begin{aligned}
E[R(x)|v] &= E\left[E\left[\sigma^{-2}(x)us|x\right]|v\right] \\
&\quad - E\left[\sigma^{-2}(x)|v\right]E\left[\sigma^{-2}(x)us|v\right]/E\left[\sigma^{-2}(x)|v\right] = 0.
\end{aligned}
$$

Then

$$
E\left[\bar{\varepsilon}u|x\right] = E[su|x] - E[u^2|x]R(x) = E\left[\sigma^{-2}(x)u \cdot s|v\right]/E\left[\sigma^{-2}(x)|v\right]
$$

is a function of v, so that $\bar{\varepsilon} \in \mathscr{T}$. Also, for any $\varepsilon \in \mathscr{T}$,

$$
\begin{aligned}
E\left[(s - \bar{\varepsilon})\varepsilon\right] &= E\left[E[\varepsilon u|x]R(x)\right] \\
&= E\left[E[\varepsilon u|v]R(x)\right] = E\left[E[\varepsilon u|v]E[R(x)|v]\right] = 0. \qquad \textit{Q.E.D.}
\end{aligned}
$$

PROOF OF THEOREM 4.1: By Assumption 4.1 and the implicit function theorem, $f(z|\beta)$ is smooth with score S_β satisfying

$$
(A.12)\qquad E\left[uS_\beta|x\right] = v_\beta + \partial G(v,\beta)/\partial\beta.
$$

This equation implies that

$$(A.13)\qquad E\left[\sigma^{-2}(x)uS_\beta|v\right] = E\left[\sigma^{-2}(x)E\left[uS_\beta|x\right]|v\right] = E\left[\sigma^{-2}(x)v_\beta|v\right] + \partial G(v,\beta)/\partial\beta E\left[\sigma^{-2}(x)|v\right]$$

By Lemma 4.1 the tangent set is $\mathscr{T}$, so by Lemma 4.2 and equation (A.13) the residual from the projection of S_β on $\mathscr{T}$ is S. The conclusion now follows by Bickel, Klaasen, Ritov, and Wellner (1992, Chapter 3, Theorem 2). *Q.E.D.*

There is a convenient characterization of efficiency that is useful for proving Theorem 5.1.

LEMMA A.3: *A regular estimator with influence function $\psi(z)$ is efficient if and only if there is a constant matrix B such that $\psi(z) = BS$.*

PROOF: By equation (4.2) it follows that $E[\psi(z)S^T] = I$, so that $E[\psi(z)\psi(z)^T] - (E[SS^T])^{-1} = E[\psi(z)\psi(z)^T] - E[\psi(z)S^T](E[SS^T])^{-1}E[S\psi(z)^T] = \min_B E[\{\psi(z) - BS\}\{\psi(z) - BS\}^T] = 0$ if and only if there is a B such that $\psi(z) = BS$. *Q.E.D.*

PROOF OF THEOREM 5.1: Let $q(x) = (x - \mu)'\Lambda(x - \mu)$. Note that $d\lambda(q)^{-1}/dq + \lambda(q)^{-1}f(q)^{-1}\partial\lambda(q)/\partial q = 1$. Then by equation (5.1),

$$(A.14)\qquad l(x) = \sigma^{-2}(v)G_1(v)F(x), \qquad F(x) = -[-\beta, I][I_1, 0]\Lambda(x - \mu),$$

where I_1 is an identity matrix with dimension equal to the number of elements of x_1. Also, by equation (4.2) $E[l(x)^T u\,\varepsilon] = 0$ for all $\varepsilon \in \mathscr{T}$. Note that $\sigma^{-2}(x)u\alpha(v) \in \mathscr{T}$ for any bounded function $\alpha(v)$, since it has finite mean-square (by $E[\sigma^{-2}(x)]$ finite) and

$$E\left[u\{\sigma^{-2}(x)\}\alpha(v)u|x\right] = \alpha(v)\sigma^{-2}(x)E[u^2|x] = \alpha(v)$$

Thus,

$$0 = E\left[l(x)^T u\{\sigma^{-2}(x)\}\alpha(v)u\right]$$

$$= E\left[l(x)^T\alpha(v)\right] = E\left[E[l(x)|v]^T\alpha(v)\right]$$

for any bounded $\alpha(v)$, implying $E[l(x)|v] = 0$ (by choosing $\alpha(v)$ to approximate $E[l(x)|v]$ in mean-square). Hence, $0 = E[l(x)|v] = -\sigma^{-2}(v)G_1(v)E[F(x)|v]$, so that $E[F(x)|v] = 0$ by $-\sigma^{-2}(v)G_1(v) \neq 0$. In particular, $E[F(x)v^T] = 0$. Furthermore, $[-\beta, I]$, $[I_1, 0]$, and Λ all having full row rank and nonsingularity of $\mathrm{var}(x)$ imply $\mathrm{var}(F(x))$ is nonsingular. Then since $z = (x_{12}^T, v^T)^T$ is a nonsingular linear combination of x, it follows that $F(x)$ is a linear combination of $z - E[z]$ that is orthogonal to v with nonsingular variance. Standard linear regression arguments then imply that there is a nonsingular matrix D such that $DF(x)$ is the residual vector from the population linear regression of x_{12} on v. Linearity of conditional expectations for spherically symmetric distributions then gives $DF(x) = x_{12} - E[x_{12}|v]$, so that by equation (A.14), $\sigma^2(x) = \sigma^2(v)$, Lemma A.3 gives the conclusion as a result of

$$D\psi(z) = D \cdot Bl(x)u = \sigma^{-2}(v)G_1(v)DF(x)u = \sigma^{-2}(v)G_1(v)\{x_{12} - E[x_{12}|v]\}u$$

$$= \sigma^{-2}(v)G_1(v)\{x_{12} - E\left[\sigma^{-2}(v)x_{12}|v\right]/E\left[\sigma^{-2}(v)|v\right]\}u = S. \qquad \textit{Q.E.D}$$

PROOF OF LEMMA 5.2: As noted in the text, $\hat{\delta}_k/\hat{\delta}_1$ is an estimator of β_k in the partial index model $g(x) = G(x_{11} + x_{1k}\beta_k, x_{12}, \ldots, x_{1,k-1}, x_{1,k+1}, \ldots, x_{1q}, x_2)$, with influence function $l_k(x)u$. Then by the argument following equation (A.14), the conditional expectation of $l_k(x)$ given the arguments of G is zero. Furthermore, the arguments of G are a nonsingular linear combination of x_{-k}, giving the conclusion. *Q.E.D.*

PROOF OF THEOREM 5.3: Let $w(x)$ equal the efficient weight, and without changing notation let $w(x) = \sigma^2(v)G_1(v)^{-1}w(x)$. Then by equation (5.1) and Lemma A.3 there is a matrix B such that

$S = Bl(x)u$, so by $\text{Prob}(u \neq 0) = 1$,

$$l(x) = B\{x_{12} - E[x_{12}|v]\}. \tag{A.15}$$

Let $x_j = (x_{12})_j$ and $x_{-j} = (x_{12})_{-j}$. Equation (A.15) implies $l_j(x) = b_j(x_j - E[x_j|v]) + b_{-j}^T(x_{-j} - E[x_{-j}|v])$ for constant scalar b_j and vector b_{-j}. Also, $b_j \neq 0$, because either b_j or b_{-j} is not zero (because nonsingularity of B follows from nonsingularity of $E[SS^T]$) and if $b_j = 0$, nonzero $c_{-j}^T(x_{-j} - E[x_{-j}|v])$ (again implied by finiteness of the variance bound for the index model) contradicts Lemma 5.1. Then, dividing by b_j, we obtain

$$E[x_j|v, x_{-j}] = E[x_j|v] + \left(-b_{-j}^T/b_j\right)\{x_{-j} - E[x_{-j}|v]\}. \qquad \textit{Q.E.D}$$

PROOF OF THEOREM 5.4: By equation (4.6),

$$\begin{aligned} V^{-1} - H'\Omega^{-1}H &= E[SS^T] - E\left[S\Psi(z)^T\right]\left(E\left[\Psi(z)\Psi(z)^T\right]\right)^{-1}E\left[\Psi(z)S^T\right] \\ &= \min_{\Pi} E\left[(S - \Pi\Psi)(S - \Pi\Psi)^T\right], \end{aligned}$$

and the other statements follow as immediate consequences.

PROOF OF THEOREM 5.5: It suffices to prove the result with $w_0(v) = 1$, since $w_0(v)$ factors out of each $l_j(x)$. By $\sigma^2(x)$ bounded away from zero, $E[\|l^*(x)\|^2]$ is finite. Note first that for any vector $b(x)$, $E[\|b(x)u\|^2] \leq CE[\|b(x)\|^2]$, so by Theorem 5.5 it suffices to show existence of square matrices Π_{jJ} such that $E[\|l^*(x) - \sum_{j=1}^J \Pi_{jJ} l_j(x)\|^2] \to 0$ as $J \to \infty$. Also, by the Hilbert space fact cited in the text, each element of $l^*(x)$ is in the closure of $\mathcal{L}_{21} \oplus \cdots \oplus \mathcal{L}_{2,q-1}$, for $\mathcal{L}_{2k} = \{l: E[l|x_{-k}] = 0, E[l^2] < \infty\}$. Thus, it suffices to show that for each k and $a(x) \in \mathcal{L}_{2k}$ there are π_{jJ} such that

$$E\left[\left|a(x) - \sum_{j=1}^J \pi_{jJ} l_{jk}(x)\right|^2\right] \to 0. \tag{A.16}$$

Also, by x bounded, polynomials in x are dense in $\mathcal{L}_2$, while

$$\begin{aligned} &E\left[\left\{a(x) - \left\{p(x) - E\left[p(x)|x_{-j}\right]\right\}\right\}^2\right] \\ &\quad = E\left[\text{Var}\left(a(x) - p(x)|x_{-j}\right)\right] \\ &\quad \leq \text{Var}(a(x) - p(x)) \leq E\left[\{a(x) - p(x)\}^2\right] \end{aligned}$$

so that the set $\{p(x) - E[p(x)|x_{-j}]: p(x)$ is a polynomial$\}$ is dense in $\mathcal{L}_{2k}$. Therefore, it suffices to show that equation (A.16) is satisfied where $a(x) = p(x) - E[p(x)|x_{-j}]$ for a polynomial $p(x)$. It follows by Assumption 5.1 that

$$b(x) = a(x) - f(x)^{-2}\partial f(x)/\partial x_k \int_b^{x_k} a(x_k(t, x)) f(x_k(t, x))\, dt$$

is continuous, for $b \in \mathcal{B}$ such that $x_k > b$ on the support of x. Note that

$$a(x) = b(x) + f(x)^{-1}\partial f(x)/\partial x_k \int_b^{x_k} b(x_k(t, x))\, dt,$$

so that for the π_{jJ} of Assumption 5.2,

$$\begin{aligned} &\left|a(x) - \sum_{j=1}^J \pi_{jJ} l_{jk}(x)\right| \\ &\quad \leq \left|b(x) - \sum_{j=1}^J \pi_{jJ}\partial p_{jk}(x)/\partial x_k\right| + \left|f(x)^{-1}\partial f(x)/\partial x_k\right| \\ &\qquad \times \left|\int_b^{x_k}\left[b(x_k(t, x)) - \sum_{j=1}^J \pi_{jJ}\partial p_{jk}(x_k(t, x))/\partial x_k\right] dt\right| \\ &\quad \leq \left(1 + |f(x)^{-1}\partial f(x)/\partial x_k| \max_{\mathcal{B}} |x_k - b|\right)\varepsilon \leq C\left(1 + |f(x)^{-1}\partial f(x)/\partial x_k|\right)\varepsilon. \end{aligned}$$

Therefore, by ε arbitrarily small, it follows that equation (A.16) is satisfied. *Q.E.D.*

REFERENCES

BEGUN, J., W. HALL, W. HUANG, AND J. WELLNER (1983): "Information and Asymptotic Efficiency in Parametric-Nonparametric Models," *Annals of Statistics*, 11, 432–452.

BICKEL, P., C. A. J. KLAASEN, Y. RITOV, AND J. WELLNER (1992): "Efficient and Adaptive Inference in Semiparametric Models," forthcoming, Johns Hopkins University Press.

CHAMBERLAIN, G. (1987): "Efficiency Bounds for Semiparametric Regression," Working Paper, University of Wisconsin.

DOKSUM, K., AND A. SAMAROV (1992): "Average Derivative Estimators of Transformations," work in progress.

GALLANT, A. R. (1981): "On the Bias in Flexible Functional Forms and an Essentially Unbiased Form: The Fourier Flexible Form," *Journal of Econometrics*, 15, 211–245.

HALL, P., AND H. ICHIMURA (1991): "Optimal Semiparametric Estimation in Single Index Models," preprint, Australian National University.

HARDLE, W., W. HILDENBRAND, AND M. JERISON (1991): "Empirical Evidence on the Law of Demand," *Econometrica*, 59, 1525–1549.

HARDLE, W., AND T. STOKER (1989): "Investigation of Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association*, 84, 986–995.

ICHIMURA, H. (1993): "Semiparametric Least Squares (SLS) and Weighted Least Squares Estimation of Single Index Models," forthcoming, *Journal of Econometrics*.

KOENKER, R., AND G. BASSETT (1982): "Robust Tests for Heteroskedasticity Based on Regression Quantiles," *Econometrica*, 50, 43–61.

NEWEY, W. K. (1990a): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

——— (1990b): "Efficient Estimation of Semiparametric Models Via Moment Restrictions," preprint, MIT Department of Economics.

——— (1991a): "The Asymptotic Variance of Semiparametric Estimators," MIT Department of Economics Working Paper No. 583.

——— (1991b): "Efficient Estimation of Tobit Models Under Symmetry," in *Nonparametric and Semiparametric Methods*, ed. by W. A. Barnett, J. L. Powell, and George Tauchen. Cambridge: Cambridge University Press.

PFANZAGL, J., AND WEFELMEYER (1982): *Contributions to a General Asymptotic Statistical Theory*. New York: Springer-Verlag.

POWELL, J. L. (1991): "Estimation of Monotonic Regression Models Under Quantile Restrictions," in *Nonparametric and Semiparametric Methods*, ed. by W. A. Barnett, J. L. Powell, and George Tauchen. Cambridge: Cambridge University Press.

POWELL, J. L., J. H. STOCK, AND T. M. STOKER (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403–1430.

RODRIGUEZ, D., AND T. STOKER (1992): "A Regression Test of Semiparametric Index Model Specification," Working Paper, Sloan School of Management, MIT.

SAMAROV, A. (1990): "On Asymptotic Efficiency of Average Derivative Estimates," preprint, Massachusetts Institute of Technology.

STEIN, C. (1956): "Efficient Nonparametric Testing and Estimation," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. Berkeley: University of California Press.

STOKER, T. M. (1986): "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54, 1461–1481.

——— (1991): "Equivalence of Direct, Indirect and Slope Estimators of Average Derivatives," in *Nonparametric and Semiparametric Methods*, ed. by W. A. Barnett, J. L. Powell, and George Tauchen. Cambrige: Cambridge University Press.

——— (1992a): *Lectures on Semiparametric Econometrics*. Louvain-La-Neuve: CORE Foundation.

——— (1992b): "Smoothing Bias in Density Derivative Estimation," forthcoming, *Journal of the American Statistical Association*.

TSYBAKOV, A. B. (1982): "Robust Estimates of a Function," *Problems of Information Transmission*, 18, 190–201.