Semiparametric Efficiency Bounds

Whitney K. Newey

Journal of Applied Econometrics, Vol. 5, No. 2. (Apr. - Jun., 1990), pp. 99-135.

Stable URL:
http://links.jstor.org/sici?sici=0883-7252%28199004%2F06%295%3A2%3C99%3ASEB%3E2.0.CO%3B2-G

Journal of Applied Econometrics is currently published by John Wiley & Sons.

JOURNAL OF APPLIED ECONOMETRICS, VOL. 5, 99–135 (1990)

SEMIPARAMETRIC EFFICIENCY BOUNDS

WHITNEY K. NEWEY
Princeton University and BellCore, Department of Economics, Princeton University, Princeton, NJ 08544, USA

SUMMARY

Semiparametric models are those where the functional form of some components is unknown. Efficiency bounds are of fundamental importance for such models. They provide a guide to estimation methods and give an asymptotic efficiency standard. The purpose of this paper is to provide an introduction to research methods and problems for semiparametric efficiency bounds. The nature of the bounds is discussed, as well as ways of calculating them. Their uses in solving estimation problems are outlined, including construction of semiparametric estimators and calculation of their limiting distribution. The paper includes new results as well as survey material.

1. INTRODUCTION

Semiparametric models, models which incorporate both parametric and nonparametric components, have received increasing attention in econometrics in recent years. This attention has been motivated primarily by the problem of misspecification of econometric models. The semiparametric approach to misspecification is to allow the functional form of some components of the model to be unrestricted. This approach is an important complement to fully nonparametric models, which may not be useful with small amounts of data or data of large dimension.

Efficiency bounds are of fundamental importance for semiparametric models. Such bounds quantify the efficiency loss that can result from a semiparametric, rather than a parametric, approach. The extent of this loss is important for the decision to use semiparametric models. These bounds also provide a guide to estimation methods. They give a standard against which the asymptotic efficiency of any particular semiparametric estimator can be measured. Their form can suggest ways of constructing estimators and calculating their limiting distribution. Also, the bounds can rule out the existence of certain types of estimators.

The purpose of this paper is to provide an introduction to research methods and problems for semiparametric efficiency bounds. To this end, the nature of the bounds will be discussed, as well as ways of calculating them. Their uses in solving estimation problems will be outlined, including construction of semiparametric estimators and calculation of their limiting distribution. A discussion of cases of interest in econometrics will be given, including open problems.

The discussion here draws heavily on several other survey papers, in particular on Wellner (1985) and Bickel Klaassen, Ritov, and Wellner (1990) (BKRW henceforth). The BKRW monograph provides a more complete treatment of semiparametric efficiency bounds. Robinson's (1988a) survey, which provides a wide-ranging account of semiparametric estimation methods, is also helpful.

Although intended primarily as a survey, this paper gives a rigorous account of some

0893–7252/90/020099–37$18.50

Received September 1988
Revised August 1989

essential components of the statistical theory of semiparametric efficiency bounds. In an attempt to keep the theory as simple as possible, several results are formulated in a way that is slightly different from the rest of the literature. In some cases conditions are imposed that may not be necessary. Also, some results that appear to be new are presented. These include the bound for the nonlinear simultaneous model, an impossibility theorem on the existence of an m-estimator in the binary choice model, small generalizations of some ideas of Bickel (1982) and BKRW on finding consistent estimators, and a generalization of Ritov (1987) and BKRW's discussion of the limiting distribution of a semiparametric m-estimator.

Section 2 includes a definition of semiparametric efficiency bounds and a discussion of their precise interpretation. Section 3 presents method of calculating the bounds, including a number of examples. Section 4 discusses some ways that semiparametric efficiency bounds can assist in solving semiparametric estimation problems. Section 5 reviews results on efficient estimation methods. Section 6 presents a survey of the literature and recent semiparametric efficiency results for econometric models.

Before proceeding it is appropriate to make two terminology notes. Throughout, 'efficiency' refers to the usual, first-order asymptotic efficiency. There are few results currently available on higher-order asymptotic efficiency in semiparametric models. Also throughout, 'nonparametric' refers to a component of a model that is infinite-dimensional. This terminology is more compact that 'infinite-dimensional parameter', although 'abstract' is an alternative used by BKRW to indicate infinite-dimensionality.

2. DEFINITION OF SEMIPARAMETRIC EFFICIENCY BOUNDS

Semiparametric efficiency bounds were introduced by Stein (1956), and developed by Koshevnik and Levit (1976), Pfanzagl and Wefelmeyer (1982), Begun *et al.* (1983), and BKRW. Such bounds are defined in an intuitive way due to Stein (1956). One could imagine that the data are generated by a parametric model that satisfies the semiparametric assumptions and contains the truth. Such a model is referred to as a *parametric submodel*, where the 'sub' prefix refers to the fact that it is a subset of the model consisting of all distributions satisfying the assumptions. One can obtain the classical Cramer–Rao bound for a parametric submodel. Any semiparametric estimator, i.e. one that is consistent and asymptotically normal under the semiparametric assumptions, has an asymptotic variance that is comparable to the Cramer–Rao bound of a semiparametric model, i.e. the semiparametric estimator is in the same class as the maximum-likelihood estimator for the submodel, and therefore has an asymptotic variance no smaller than the bound for the submodel. Since this comparison holds for each parametric submodel that one could imagine, it follows that

The asymptotic variance of any semiparametric estimator is no smaller than the supremum of the Cramer–Rao bounds for all parametric submodels, denoted V.

The matrix V is the semiparametric asymptotic variance bound.

Further discussion may help fix ideas. Let β denote a $q \times 1$ vector of parameters of interest, with true value β_0. In most of the paper we will restrict attention to models of i.i.d. data $z_1, \ldots, z_n$, although some results for time-series models will be reviewed below. For the i.i.d. case a parametric submodel corresponds to a parameter vector θ and a likelihood function $f(z|\theta)$ for a single observation. Whenever the parameter of interest is identified it will be some function $\beta(\theta)$ of θ. The definition of a parametric submodel requires that the likelihood satisfy the semiparametric restrictions and that $f(z|\theta_0)$ gives the true distribution for some θ_0.

For example, consider a model where the parameter of interest is the expectation of a single

observation z, i.e. $\beta_0 = E[z]$, where $E[\cdot]$ denotes the expectation at the truth. Let the distribution of z be unrestricted, except for regularity conditions such as $\text{var}(z) < \infty$. This model was considered by Levit (1975). A parametric submodel corresponds to any likelihood $f(z|\theta)$ (satisfying appropriate regularity conditions) such that $f(z|\theta_0)$ gives the true distribution of z_i. The parameter of interest will be $\beta(\theta) = \int z f(z|\theta) dz$. Because the distribution is unrestricted, one might expect that no estimator that is more efficient than the sample mean can be found. This conjecture can be verified by showing that $V = \text{var}(z)$, the asymptotic variance of the sample mean, Levit's (1975) result to be discussed below.

For a second example, consider the additive semiparametric regression model of Engle *et al.* (1986),

$$y_i = x_i'\beta_0 + g_0(v_i) + \varepsilon_i, \tag{1}$$

where y_i is a scalar dependent variable, x_i and v_i are vectors of exogenous variables, $g_0(v)$ is an unknown function, and ε_i is a disturbance. To focus on the semiparametric nature of the regression function, assume that the disturbance is independent of the regressors, distributed as $N(0, \sigma_0^2)$ with σ_0^2 known, and that the density $f(x, v)$ of x_i and v_i is known. This model has a parametric component β and a nonparametric component $g(v)$. A parametric submodel corresponds to a parameterization of $g(v)$, say $g(v, \eta)$, such that $g(v, \eta_0) = g_0(v)$ for some η_0. The parameters of such a parametric submodel are $\theta = (\beta', \eta')'$. Note that there $\beta(\theta)$ is simply the first q components of θ. The semiparametric variance bound for β in this model, which is given by BKRW, will be discussed in section 3.

Some of the literature on semiparametric efficiency has focused on *adaptive estimation*; e.g. Stein (1956) and Bickel (1982). Adaptive estimation refers to models where parameters of interest can be estimated equally well when the nonparametric part of the model is unknown as when it is known. Such models are those where the semiparametric bound is equal to the parametric bound that applies when the nonparametric part of the model is known.

To be precise about the bound it is necessary to impose regularity conditions on the parametric submodels and to exercise some care in the definition of asymptotic efficiency. Regularity conditions are necessary to guarantee that the Cramer–Rao bound is well-defined and gives an asymptotic efficiency bound. The regularity conditions for parametric submodels will include smoothness conditions, in the modern form of mean-square continuous differentiability of the square root of the likelihood function. Because of its inherently technical nature, the discussion of differentiable likelihoods is reserved to Appendix A. In many models it will also be helpful to impose additional 'smoothness' conditions, such as existence of $\text{var}(z)$ in the mean example. A *smooth* parametric submodel will be one that satisfies the mean-square differentiability conditions of Appendix A and possibly additional smoothness conditions. A *regular* parametric submodel will be one that is smooth with nonsingular information matrix. A precise definition of V is that it is the supremum of the Cramer–Rao bounds for all regular parametric submodels.

Ranking all consistent, asymptotically normal estimators by asymptotic variance does not lead to a useful characterization of asymptotic efficiency. There exist superefficient estimators with an asymptotic variance less than that of the maximum-likelihood estimator for some true parameter values.[1] The modern approaches in statistics to this problem are to use an

[1] The following famous example is due to Hodges: Let $\bar{X}$ be the mean from an i.i.d. $N(\mu, 1)$ sample, so $\bar{X}$ is the maximum-likelihood estimator. Consider the estimator $\hat{\mu} = \bar{X}$ if $|\bar{X}| > n^{-1/4}$ and $\hat{\mu} = 0$ if $|\bar{X}| \leq n^{-1/4}$, and note that $\hat{\mu}$ is equal to $\bar{X}$ with probability approaching one if $\mu \neq 0$ and is equal to zero with probability approaching one if $\mu = 0$. Thus, the asymptotic distribution of $\hat{\mu}$ is the same as the sample mean if $\mu \neq 0$ but has asymptotic variance zero at $\mu = 0$.

asymptotic minimax criteria for evaluating efficiency (see Hajek, 1972), and/or to restrict attention to a class of estimators satisfying uniformity conditions that rule our superefficient estimators. These approaches are justified by the fact that in finite samples superefficient estimators do worse than the maximum-likelihood estimator in neighbourhoods of the point of superefficiency; see LeCam (1953).

Uniformity conditions will be discussed here, because familiar ideas are involved and they lead to a formula that is useful in calculating the bound. The most important kind of uniformity property considered in the statistics literature is that where the convergence of an estimator to its limiting distribution is uniform in the true parameter value. Superefficient estimators do not have this property.[2] This uniformity property is desirable because without this property the answer to the often-asked question 'How big does the sample size have to be?' will depend on the true parameter value. Unfortunately, this uniformity property is too strong for many nonparametric and semiparametric models; see section 5.5 of Bickel (1982). An alternative, weaker uniformity property that is useful in semiparametric models is for the convergence to the limiting distribution to be uniform in certain shrinking neighbourhoods of the true parameter value.

To define this local uniformity property precisely, let a local data-generating process (LDGP) be one where for each sample size n the data are distributed according to θ_n, where $\sqrt{n}(\theta_n - \theta_0)$ is bounded. LDGPs are familiar from the literature on local power of test statistics. An estimator $\hat{\beta}$ is said to be *regular in a parametric submodel* if, for each θ_0, $\sqrt{n}(\hat{\beta} - \beta(\theta_n))$ has a limiting distribution that does not depend on the LDGP. An estimator $\hat{\beta}$ is said to be *regular* if it is regular in every regular parametric submodel and the limiting distribution does not depend on the parametric submodel.

The class of regular estimators has two important properties. First, superefficient estimators are excluded.[3] Second, estimators that make use of more information than is contained in the semiparametric model are excluded. Such an estimator will typically have a corresponding parametric submodel and LDGP that induces local bias, where the mean of the limiting distribution of $\sqrt{n}(\hat{\beta} - \beta(\theta_n))$ depends on $\{\theta_n\}$. As an extreme example, the estimator $\hat{\beta} = \beta_0$, where β_0 is the true value of the parameter, is not regular; for $\theta = \beta$ and $\theta_n = \beta_0 + \delta/\sqrt{n}$, the limiting distribution of $\sqrt{n}(\hat{\beta} - \beta(\theta_n))$ is degenerate at $-\delta$.

These properties suggest that the semiparametric bound V should apply to the class of regular estimators. That it does follows by the Begun *et al.* (1983) and Chamberlain (1986) semiparametric generalizations of Hajek's (1970) representation theorem. A vector version of Chamberlain's (1986) Theorem 2(i) is:

Theorem 2.1: If $\hat{\beta}$ is regular then the limiting distribution $\sqrt{n}(\hat{\beta} - \beta_0)$ is equal to the distribution of $Y + U$, where Y is distributed as $N(0, V)$ and U is some random vector that is independent of Y.

Proofs of the theorems are sketched in Appendix A. This result says that the limiting distribution of any regular estimator is equal to that of a $N(0, V)$ random variable plus noise. When U is mean zero Gaussian, this result implies the usual comparison of asymptotic covariance matrices; it follows from independence of U and Y that the asymptotic variance of

[2] When convergence in distribution is uniform in the true parameters, the limiting distribution must be continuous in the parameters; e.g. see BKRW. Typically, the limiting distribution of superefficient estimators is discontinuous in the parameters, as in the example given in the previous footnote.

[3] In the example of the previous footnote, take $\theta = \mu = 0$ and let the LDGP be $\mu_n = \delta/\sqrt{n}$. The limiting distribution of $\sqrt{n}(\hat{\mu} - \mu_n)$ is 0 if $\delta = 0$, but is δ if $\delta \neq 0$.

$\hat{\beta}$ is equal to $E[(Y+U)(Y+U)'] = V + E[UU']$, which differs from V by a positive semidefinite matrix.

This result leads to a natural definition of an efficient estimator for the parameters of interest of a semiparametric model. An estimator is said to be *efficient* if its limiting distribution is $N(0, V)$ and it is regular. It is interesting to note that the bound can be well defined without an efficient estimator existing, i.e. the bound need not be sharp. Ritov and Bickel (1987) give examples where the bound is well defined and finite but no $\sqrt{n}$-consistent estimator exists. In their examples, smoothness conditions in addition to those sufficient for the bound to be well defined are necessary for the existence of an efficient estimator.

Since the bound applies to the class of regular estimators, it is useful to have simple sufficient conditions for regularity. Also, regularity of an estimator is related to a formula that is useful for calculating the bound. For a particular class of estimators it is easy to give conditions for regularity. Define an estimator $\hat{\beta}$ to be *asymptotically linear* if it is asymptotically equivalent to a sample average, i.e. there is a function $\psi(z)$ of a single observation such that for $\psi_i = \psi(z_i)$, at the truth

$$\sqrt{n}(\hat{\beta} - \beta_0) = \sum_{i=1}^{n} \psi_i/\sqrt{n} + o_p(1), \quad E[\psi] = 0, \quad E[\psi\psi'] \text{ finite and nonsingular,}$$

where $o_p(1)$ denotes a random variable that converges in probability to zero. In what follows, it will often be convenient to suppress the z argument, as is done here. The function $\psi(z)$ is referred to as the *influence function*, motivated by the fact that to a first-order $\psi(z)$ is the influence of a single observation on $\hat{\beta}$, an observation of Hampel (1974).

Asymptotically linear estimators should be familiar. The most common argument for asymptotic normality proceeds by using a mean value expansion to show that this equation holds, and then applying a central limit theorem to the average on the right-hand side. For instance, this argument is commonly used to show asymptotic normality of an m-estimator, which is one that satisfies

$$\sum_{i=1}^{n} m(z_i, \hat{\beta})/\sqrt{n} = o_p(1), \tag{2}$$

where $m(z, \beta)$ is a $q \times 1$ vector of functions satisfying $E[m(z, \beta_0)] = 0$. The usual expansion argument, or a more general result like that of Huber (1967), shows that $\hat{\beta}$ is asymptotically linear with influence function

$$\psi(z) = -M^{-1}m(z, \beta_0), \quad M = \partial E[m(z, \beta)]/\partial\beta\,|_{\beta=\beta_0}. \tag{3}$$

A simple result on regularity of an asymptotically linear estimator is the following. Let S_θ denote the score for a parametric submodel, evaluated at the truth. The score can typically be obtained as the derivative of the log-likelihood, and is precisely defined in Appendix A. Also, let $E_\theta[\cdot]$ denote the expectation for a parametric submodel with parameters θ, and for a matrix A let $\|A\| \equiv [\text{trace}(A'A)]^{1/2}$.

Theorem 2.2: Suppose that $\hat{\beta}$ is an asymptotically linear estimator with influence function ψ, and for all regular parametric submodels $\beta(\theta)$ is differentiable and $E_\theta[\|\psi\|^2]$ exists and is continuous on a neighbourhood of θ_0. Then $\hat{\beta}$ is regular if and only if, for all regular parametric submodels,

$$\partial\beta(\theta_0)/\partial\theta = E[\psi S_\theta']. \tag{4}$$

Differentiability of $\beta(\theta)$ and continuity of $E_\theta[\|\psi\|^2]$ are examples of additional smoothness

conditions alluded to above. In models like that of the second example, where $\theta = (\beta', \eta')'$, $\beta(\theta)$ is differentiable by construction. Also, it can be shown that continuity of $E_\theta[\|\psi\|^2]$ is implied by boundedness of $E_\theta[\|\psi\|^r]$ for $r > 2$, a familiar type of condition on existence of moments. In addition, it should be noted that continuity of $E_\theta[\|\psi\|^2]$ can be replaced by local uniformity of the equation for $\hat{\beta}$ in the definition of asymptotic linearity, as in BKRW, although it apears to be easier to work with this continuity condition. In principle, such additional smoothness conditions could affect the bound, but in practice they do not. In what follows it will be convenient to include differentiability of $\beta(\theta)$ and continuity of $E_\theta[\|\psi\|^2]$ at θ_0 in the definition of a smooth parametric submodel and a regular asymptotically linear estimator.

For m-estimators, equation (4) can be interpreted as the well-known (see e.g. Rao (1973), exercises 5·8 and 5·9) generalized information matrix equality. Note that for $\psi = -M^{-1}m(z, \beta_0)$, equation (4) can be written as

$$M\partial\beta(\theta_0)/\partial\theta + E[m(z, \beta_0)S_\theta'] = 0. \tag{5}$$

If $\beta = \theta$ and $m(z, \beta)$ were differentiable and equal to the score for β, then this equation would give equality between Hessian and outer product versions of the information matrix. It is also possible to derive equation (5) directly, which is helpful in verifying equation (4) as a sufficient condition for regularity of an m-estimator. Consider $\bar{m}(\theta, \tilde{\theta}) = E_\theta[m(z, \beta(\tilde{\theta}))]$, and assume the fundamental moment condition for an m-estimator that for each possible true value of the parameter θ, $\bar{m}(\theta, \theta) = 0$. Assuming that $\bar{m}(\theta, \tilde{\theta})$ can be differentiated with respect to θ inside the expectation (e.g. see Lemma 7.2 of Ibragimov and Hasminskii, 1981), differentiating both sides of $\bar{m}(\theta, \theta) = 0$ at θ_2 and applying the chain rule gives equation (5).

Theorem 2.2 provides one way of showing local regularity, by imposing continuity of $E_\theta[\|\psi\|^2]$ and verifying equation (4). Also, this result illustrates that equation (4) is a fundamental consequence of local regularity. Equation (4) is fundamental in the sense that it gives an important formula for the Cramer–Rao bound in a parametric submodel. The Cramer–Rao bound for θ is $(E[S_\theta S_\theta'])^{-1}$. Furthermore, if $\beta(\theta)$ is differentiable, then the Cramer–Rao bound for β is $V_\theta = (\partial\beta(\theta_0)/\partial\theta)(E[S_\theta S_\theta'])^{-1}\partial\beta(\theta_0)/\partial\theta'$. This formula follows from the invariance of maximum-likelihood and the delta-method, which imply that, for the maximum-likelihood estimator (MLE) $\tilde{\theta}$, the MLE of β is $\beta(\tilde{\theta})$ and its asymptotic variance is V_θ. It then follows by equation (4) that

$$V_\theta = (\partial\beta(\theta_0)/\partial\theta)(E[S_\theta S_\theta'])^{-1}\partial\beta(\theta_0)/\partial\theta' = E[\psi S_\theta'](E[S_\theta S_\theta'])^{-1}E[S_\theta\psi']. \tag{6}$$

This formula is important because the classifical efficiency result that the asymptotic variance $E[\psi\psi']$ of the asymptotically linear estimator $\hat{\beta}$ is no smaller than the Cramer–Rao bound follows immediately; note $E[\psi\psi'] - V_\theta = E[(\psi - AS_\theta)(\psi - AS_\theta)']$ for $A = E[\psi S_\theta'](E[S_\theta S_\theta'])^{-1}$. Also, this formula is useful in calculating the semiparametric bound, as discussed in section 3.

3. CALCULATING THE BOUND

Although the definition of the semiparametric bound V does not lead directly to a method of calculation, there is more structure to the semiparametric efficiency problem that can aid in computing the bound.[4] A hint of this structure is given by equation (6). This formula suggests that it may be useful to restrict $\beta(\theta)$ to be differentiable and satisfy an equation analogous to

[4] In fact, there is no guarantee that the supremum V exists, since it involves ranking matrices (i.e. a partial order). However, under the differentiable parameter hypothesis of this section it does exist.

(4). Following Koshevnik and Levit (1976) and Pfanzagl and Wefelmeyer (1982), define a *differentiable parameter* to be one such that $\beta(\theta)$ is differentiable for all smooth parametric submodels and there exists a $q \times 1$ random vector d such that $E[d'd]$ is finite and for all regular parametric submodels

$$\partial\beta(\theta_0)/\partial\theta = E[dS_\theta']. \tag{7}$$

This definition is actually a pathwise version of differentiability, as discussed in the literature, the term 'pathwise' referring to the fact that it is defined in terms of parametric submodels.

It is easy to show that the mean parameter is differentiable, under appropriate regularity conditions. In this example, $\beta(\theta) = E_\theta[z]$. If $f(z|\theta)^{1/2}$ is mean-square continuously differentiable (see Appendix A) and $E_\theta[z^2]$ is bounded in a neighbourhood of θ_0, then by Lemma 7.2 of Ibragimov and Hasminskii (1981), $\beta(\theta)$ is differentiable at θ_0 and the derivative can be obtained by differentiating under the integral, yielding $\partial\beta(\theta_0)/\partial\theta = E[zS_\theta']$. In this example, the mean is a differentiable parameter with $d = z$. The condition that $E_\theta[z^2]$ is bounded is another example of the additional regularity conditions alluded to earlier.

To show that a parameter is differentiable it is necessary fo find d satisfying equation (7). A general formula for d in semiparametric models with parametric and nonparametric components, as in the semiparametric regression model, will be discussed below. One way of finding d is to find the influence function ψ for any regular asymptotically linear estimator. Then by equation (4), $d = \psi$ will satisfy equation (7). Indeed, Van der Vaart (1988) shows that differentiability of $\beta(\theta)$ and existence of a regular estimator are sufficient for existence of such a d. The 'only-if' part of Theorem 2.2 specializes this sufficient condition to the existence of a regular asymptotically linear estimator.

A d satisfying equation (7) is not unique. In general, any random vector that is orthogonal to the scores for all regular parametric submodels can be added to d without affecting equation (7). For example, by the mean zero property of scores, if d satisfies equation (7) then so will $\tilde{d} = d + c$ for any constant vector c. In the mean example, $-\beta_0$ can be added to z to obtain $\tilde{d} = z - \beta_0$, the influence function for the sample mean.

For a differentiable parameter β, the formula for the Cramer–Rao bound V_θ for a parametric submodel is analogous to that of equation (6). By equation (7) and the same arguments as used in section 2, $V_\theta = (\partial\beta(\theta_0)/\partial\theta)(E[S_\theta S_\theta'])^{-1}\partial\beta(\theta_0)/\partial\theta' = E[dS_\theta'](E[S_\theta S_\theta'])^{-1}E[S_\theta d']$. This formula has a useful interpretation. Note that

$$V_\theta = E[d_\theta d_\theta'], \quad d_\theta = E[dS_\theta'](E[S_\theta S_\theta'])^{-1}S_\theta. \tag{8}$$

That is, the Cramer–Rao bound equals the variance matrix of the predicted value from the population regression of d on the score. It follows that the bound cannot decrease if more elements are added to S_θ, since adding addition elements to S_θ corresponds to adding more variables to a regression, which can only increase the variability of the prediction. Therefore, it is plausible that the Cramer–Rao bound can be made as large as possible by regressing d on a sufficient number and variety of scores for different parametric submodels. Alternatively, one can think of regressing on an infinite-dimensional vector of scores for different parametric submodels that spans the space of all possible scores.

Following Pfanzagl and Wefelmeyer (1982), a way this discussion can be formalized is as follows. Define the *tangent set* $\mathcal{S}$ to be the mean square closure of all q-dimensional linear combinations of scores S_θ for smooth parametric submodels,

$$\mathcal{S} = \{a \in \mathbb{R}^q: \ E[\|a\|^2] < \infty, \ \exists A_j S_{\theta j} \quad \text{with} \quad \lim_{j\to\infty} E[\|a - A_j S_{\theta j}\|^2] = 0\}$$

where A_j is a constant matrix with q rows. The tangent set is an infinite-dimensional set that is the nonparametric analogue of the set of linear combinations of the score vector. The predicted value from the regression of d on this set should have larger variance than the predicted value for any particular submodel, suggesting that the regression on $\mathcal{S}$ should given the semiparametric bound. The mathematical meaning of regression on the tangent set is a least-squares projection in a Hilbert space of random vectors. Define $\mathcal{S}$ to be *linear* if $a_{s1} + b_{s2} \in \mathcal{S}$ for all real scalars a and b and elements s_1 and s_2 of $\mathcal{S}$. Let an inner product on the set of $k \times 1$ random vectors with finite second moment be $E[s_1's_2]$. If $\mathcal{S}$ is linear, then the projection of d on $\mathcal{S}$ in the Hilbert space with this inner product exists, and is the unique vector δ satisfying

$$\delta \in \mathcal{S}, \quad E[(d-\delta)'s] = 0 \quad \text{for all} \quad s \in \mathcal{S}. \tag{9}$$

Theorem 3.1: Suppose that the parameter is differentiable, $\mathcal{S}$ is linear, and $\mathrm{E}[\delta\delta']$ *is nonsingular, for the projection δ of d on $\mathcal{S}$. Then $V = E[\delta\delta']$.*

The vector δ is referred to as the *efficient influence function*, motivated by the fact that the asymptotic variance of an asymptotically linear estimate with influence function δ equals V. For δ to be the efficient influence function, it is enough that equation (9) is satisfied; linearity of $\mathcal{S}$ is only used to show that such a δ exists. However, in all known examples $\mathcal{S}$ is linear, so that imposing this condition results in little lost generality.

To use this result to calculate the bound, the tangent set and the projection of d on $\mathcal{S}$ have to be calculated. Calculation of the tangent set is typically straightforward. It is often easy to conjecture a form of the tangent set from the restrictions on the scores implied by the semiparametric model. This conjecture can then be verified by showing that scores for smooth submodels lie in the set, and by exhibiting a family of smooth submodels with scores that can approximate any element of the set arbitrarily well in mean square, i.e. such that the scores are dense in the set.

In the mean example the distribution is unrestricted, so that it is natural that the tangent set should consist of all scalar random variables satisfying the zero mean restriction of scores. This conjecture can be verified by exhibiting a dense family of scores, which is done in Appendix B. Because verifying a natural conjecture concerning the form of the tangent set is a technical exercise, such verification will not be carried out for the other examples considered below.

Calculation of the projection can be difficult, although it is easy in several interesting examples. In the mean example, with $\mathcal{S} = \{s: E[s] = 0\}$, the projection is $\delta = d - E[d] = z - E[z]$; note that $\delta \in \mathcal{S}$ and $E[(d-\delta)s] = E[E[d]s] = 0$. It then follows that $V = E[d^2] = \mathrm{var}(z)$, the asymptotic variance of the sample mean, as was expected. There are also other examples where the calculation of the projection is straightforward, including the semiparametric regression example. These examples, as well as examples where the projection is difficult, will be discussed below.

The calculation for the mean example illustrates that in models where the parameter is an explicit function of the distribution (e.g. mean or median) and the distribution is unrestricted, there is only one influence function for a regular asymptotically linear estimator. To see this result, note that the only restriction on the tangent set should be the mean zero property of scores, i.e. $\mathcal{S} = \{s: E[s] = 0\}$. Also, by equation (5) any pair of influence functions ψ and $\tilde{\psi}$ will satisfy $E[(\psi - \tilde{\psi})S_\theta] = 0$, so that $\psi - \tilde{\psi}$ is orthogonal to the tangent set. But, since $\psi - \tilde{\psi} \in \mathcal{S}$, this orthogonality implies $E[(\psi - \tilde{\psi})'(\psi - \tilde{\psi})] = E[\|\psi - \tilde{\psi}\|^2] = 0$, which is equivalent to $\psi = \tilde{\psi}$ (with probability one). To interpret this result, note that a parameter of an unrestricted

distribution is exactly identified, so that it is not surprising that there can be only one asymptotically linear estimator.

Semiparametric models with a parametric and a nonparametric component, like the semiparametric regression example, have more structure that can be exploited in calculating the bound. The form of the bound for such models can be motivated by considering the bound in a parametric submodel. Let $\theta = (\beta', \eta')'$ be the parameters of a submodel, and let $S_\theta = (S_\beta', S_\eta')'$ be partitioned conformably. By the partitioned inverse formula, the Cramer–Rao bound for β is

$$\{E[(S_\beta - \tilde{B}S_\eta)(S_\beta - \tilde{B}S_\eta)']\}^{-1}, \quad \tilde{B} = E[S_\beta S_\eta'](E[S_\eta S_\eta'])^{-1}. \tag{10}$$

That is, the Cramer–Rao bound equals the inverse variance matrix of the residuals from the population regression of S_β on S_η. The bound cannot decrease as more elements are added to S_η, since adding more variables can only decrease the variability of the residuals. Therefore, it is plausible that the Cramer–Rao bound can be made as large as possible by regressing S_β on an infinite-dimensional vector that spans the set of all scores for parametric submodels.

This calculation can also be motivated by the differentiable parameter idea. Here $\beta(\theta)$ is differentiable with $\partial\beta(\theta_0)/\partial\theta = [I_k, 0]$. Specializing equation (10) gives

$$E[dS_\beta'] = I_k, \quad E[dS_\eta'] = 0. \tag{11}$$

To find such a d, one could begin with S_β, and calculate the residual S from the projection of S_β on the infinite-dimensional spanning vector. Then $E[SS_\eta'] = 0$ would follow by the orthogonality property of least-squares residuals and the definition of $\mathcal{T}$ (see Lemma A.1 of Appendix A). If $E[SS_\beta']$ is nonsingular, then $d = (E[SS_\beta'])^{-1}S$ will satisfy both conditions of equation (11). Furthermore, since the residual S is a linear combination of scores for β and η, this d will be an element of the tangent set, and will therefore be the efficient influence function.

Following Begun *et al.* (1983) and BKRW, this discussion can be formalized as follows. Let $\mathcal{T}$ be the tangent set in the nonparametric direction, defined as the mean-square closure of linear combinations of scores S_η for the nonparametric component:

$$\mathcal{T} = \{\iota \in \mathbb{R}^q: \quad E[\|\iota\|^2] < \infty, \exists B_j S_{\eta j} \quad \text{with} \quad \lim_{j\to\infty} E[\|\iota - B_j S_{\eta j}\|^2] = 0\}$$

Where it will not cause confusion, this set will be referred to simply as the tangent set. If $\mathcal{T}$ is linear, then the residual from the projection of S_β on $\mathcal{T}$ exists, and is the unique vector S satisfying

$$S_\beta - S \in \mathcal{T}, \quad E[S'\iota] = 0 \quad \text{for all} \quad \iota \in \mathcal{T}. \tag{12}$$

Theorem 3.2: If $f(z|\beta)$ is smooth with score S_β, $\mathcal{T}$ is linear, and the residual S of the projection of S_β on $\mathcal{T}$ satisfies $E[SS']$ nonsingular, then β is a differentiable parameter and has efficient influence function $(E[SS'])^{-1}S$, with $V = (E[SS'])^{-1}$.

The vector S is referred to as the *efficient* score.

This framework allows us to be more specific about conditions for adaptive estimation, as discussed in Begun *et al.* (1983) and BKRW. Recall that the condition for the Cramer–Rao bound for parameters of interest to be the same when nuisance parameters are known as when they are unknown is block diagonality of the information matrix. The geometric interpretation of block diagonality is orthogonality of the scores for parameters of interest and the nuisance parameters. The semiparametric generalization of this condition is orthogonality of S_β and the tangent set $\mathcal{T}$, a necessary condition for adaptive estimation formulated by Stein (1956). This

condition can also be modified to allow for nuisance parameters. Let α be a vector of nuisance parameters, with score S_α. In this case Stein's (1956) necessary condition for adaptive estimation is orthogonality of $\mathcal{T}$ and the partial score for β, i.e. the residuals from the projection of S_β on S_α.

Theorem 3.3: If $f(z \mid \beta, \alpha)$ is regular a necessary condition for the existence of an adaptive estimator of β is that $E[\{S_\beta - E[S_\beta S_\alpha'](E[S_\alpha S_\alpha'])^{-1} S_\alpha\}' \tau] = 0$ for all $\tau \in \mathcal{T}$.

In general, there may be no regular parametric submodel with the Cramer–Rao bound for β equal to V. All that is guaranteed by the hypotheses of Theorem 2.2, in particular by the definition of the tangent set as the *closure* of linear combinations of scores, is that V is approximated by Cramer–Rao bounds for parametric submodels. In special cases where V is attained by a parametric submodel, it may be possible to interpret the submodel. One general interpretation relates attainment of the bound to the necessary condition for adaptive estimation. If there is a parametric submodel for which S is the partial score for β, then orthogonality of S with the tangent set means that the necessary condition for adaptive estimation is satisfied, with nuisance parameters equal to those of the parametric submodel.

As noted above, finding the tangent set is typically straightforward, but calculating the projection of the tangent set can be either easy or difficult, depending on the model. Calculating the projection is easy when well-known results can be utilized. An example of such a result is the fact that the projection of a random variable U with finite second moment on the set of all functions of a random variable V that have finite second moment is $E[U|V]$. In some cases, known results give the projection immediately, or at least lead to helpful conjectures, which can then be verified by checking equation (12) (or (9)).

To illustrate, consider the semiparametric regresson model. The likelihood and scores for a parametric submodel, corresponding to a parameterization $g(v, \eta)$ of $g(v)$, are

$$\ln f(y, x, \theta) = C - (1/2)\ln \sigma_0^2 - (y - x'\beta - g(v, \eta))^2/2\sigma_0^2 + \ln f_0(x, v), S_\beta = x\varepsilon/\sigma_0^2, \quad S_\eta = g_\eta \varepsilon/\sigma_0^2, \tag{13}$$

where $f_0(x, v)$ is the density of x and v, $g_\eta = \partial g(v, \eta_0)/\partial \eta$, and $\varepsilon = y - x'\beta_0 - g_0(v)$. Note that g_η is a vector consisting only of functions of v. Therefore, a plausible conjecture for the tangent set is

$$\mathcal{T} = \{\varepsilon D(v) : E[\varepsilon^2 \|D(v)\|^2] < \infty\}. \tag{14}$$

Indeed, a parametric submodel with score for η equal to $\varepsilon D(v)$ is $g(v, \eta) = g_0(v) + \eta'[\sigma_0^2 D(v)]$, with $\eta_0 = 0$.

To see what the projection should be, note that if one ignored the presence of the term ε, the calculation would consist of projecting $\sigma_0^{-2}x$ on all possible vectors of the form $D(v)$. Such a projection would yield the conditional expectation $E[\sigma_0^{-2}x|v] = \sigma_0^{-2}E[x|v]$. It turns out that if we go back to the case where ε is present by multiplying by ε to obtain $\varepsilon \cdot \sigma_0^{-2}E[x|v]$, we do obtain the projection. This conjecture could be verified directly, by checking equation (12). There is also a general projection result that is useful in a number of models like this one, where the tangent set consists of products of a residual with all functions of some variables. Consider a random vector U, and conformable random matrices V and W.

Lemma 3.4: If WU has finite second moment and V and W are functions some T such that $E[UU'|T]$ is constant and positive definite, then the projection of WU on $\mathcal{T}_V \equiv \{D(V)U : E[\|D(V)\|^2] < \infty\}$ is $E[W|V]U$.

To apply this result to the semiparametric regression model, let $U = \varepsilon$, $W = \sigma_0^{-2}x$, and $V = v$.

Then the conclusion of Lemma 3.4 gives the efficient score

$$S = S_\beta - E[W|V]\,U = \sigma_0^{-2}\varepsilon\{x - E[x|v]\}, \tag{15}$$

a result obtained by BKRW.

Allowing for the more realistic assumption that σ_0 and the density of x and v are unknown does not change the bound. The score for σ is $S_\sigma = (\varepsilon^2 - \sigma_0^2)/\sigma_0^3$. Also, η can enter the marginal density $f(x, v, \eta)$ of x and v, so that the general form for the score for η is $S_\eta = g_\eta \cdot \varepsilon/\sigma_0^2 + S(x, v)$, where $S(x, v)$ is the score for $f(x, v, \eta)$. The tangent set becomes $\mathcal{T} = \{\varepsilon D_1(v) + D_2(x, v) : E[D_2] = 0\}$. This set is the direct sum of two orthogonal components $\{\varepsilon D_1(v)\}$ and $\{D_2(x, v)\}$, so the projection on the direct sum is the sum of the projections (e.g. Luenberger, 1969). But S_β and S_σ are orthogonal to $\{D_2(x, v)\}$, so that their projection on this set is zero. Therefore, the projection on the tangent set is just the projection on $\{\varepsilon D_1(v)\}$. By normality, S_σ is orthogonal to this set, so, that the efficient score is $(S', S_\sigma)'$, for S from equation (15). Also, $E[SS_\sigma] = 0$ by normality, so that the bound for β remains the same as before. It is interesting to note that the bound for σ is the same as the bound where $g_0(v)$ and the density of x and v are known, so that Stein's (1956) necessary condition for adaptive estimation of σ is satisfied.

Some of the calculations in the previous paragraph can be put on a general footing. A decomposition of the tangent set into orthogonal components will be available when the density for the observed data is factored into a conditional density (for y in the above example) and a marginal density (for x and v in the above example). The orthogonality between the two components corresponds to orthogonality between marginal and conditional scores. Furthermore, when the variables in the marginal density are ancillary to the parameters of interest, meaning that these parameters do not enter the marginal density, the score for the parameters of interest will be orthogonal to the component of the tangent set corresponding to the marginal density. Consequently, the efficient score can be obtained by projecting on only the component corresponding to the conditional density, implying that lack of knowledge of the marginal density does not affect the bound, as one would expect for ancillary variables.

A number of interesting semiparametric models involve unknown disturbance distributions. One example is the linear regression model with a disturbance that is independent of the regressors but otherwise unspecified. In this model there is a dependent variable y and regressors x such that

$$y = x'\beta_0 + \varepsilon, \tag{16}$$

where ε is distributed independently of x with differentiable density function $f_0(\varepsilon)$, and any constant is absorbed in ε. The nonparametric part of the model is the density of ε. A parametric submodel corresponds to a parametric family of densities $f(\varepsilon, \eta)f(x, \eta)$ such that for some η_0, $f(\varepsilon, \eta_0)f(x, \eta_0)$ is equal to the true density of ε and x. The log-likelihood and scores are

$$\ln f(y, x, \theta) = \ln f(y - x'\beta, \eta) + \ln f(x, \eta), \tag{17}$$
$$S_\beta = -xs(\varepsilon), \quad S_\eta = D(\varepsilon) + D(x),$$

where $s(\varepsilon) = f_\varepsilon(\varepsilon)/f(\varepsilon)$, $D(\varepsilon) = f_\eta(\varepsilon, \eta_0)/f(\varepsilon)$, $D(x) = f_\eta(x, \eta_0)/f(x)$ and the subscripts denote partial derivatives. Note that the distributions of ε and x are unrestricted, so that S_η should be unrestricted except for the mean zero property of scores. Therefore, a plausible conjecture for the tangent set is

$$\mathcal{T} = \{D(\varepsilon) + D(x) : E[\|D(\varepsilon)\|^2], E[\|D(x)\|^2] < \infty, E[D(x)] = E[D(\varepsilon)] = 0\}. \tag{18}$$

To see what the projection should be, note that x is ancillary for β, so that it suffices to calculate the projection on $\{D(\varepsilon): E[D(\varepsilon)] = 0\}$. It is easy to verify that if a random vector U has mean zero then the projection of U on the space of all conformable vector functions of V with mean zero is simple $E[U|V]$. Thus, setting $U = S_\beta$ and $V = \varepsilon$, the efficient score is

$$S = S_\beta - E[S_\beta|\varepsilon] = -\{x - E[x]\}s(\varepsilon), \tag{19}$$

a result due to Bickel (1982).

This example is one where the necessary conditions for adaptive estimation are satisfied in a parametric model where the only unknown piece of the distribution of ε is a location parameter, i.e. $f_0(\varepsilon) = h(\varepsilon - \alpha_0)$, where $h(u)$ is known and α_0 is unknown. The log-likelihood for this model is $\ln(h(y - x'\beta - \alpha))$, and the scores are $S_\beta = -xs(\varepsilon)$ and $S_\alpha = -s(\varepsilon)$. The partial score for β is

$$S_\beta - E[S_\beta S_\alpha](E[S_\alpha^2])^{-1}S_\alpha = -s(\varepsilon)\{x - E[s(\varepsilon)^2 x]/E[s(\varepsilon)^2]\} = -\{x - E[x]\}s(\varepsilon).$$

It follows from independence of x and ε and $E[s(\varepsilon)] = 0$ that this partial score is uncorrelated with any element of the tangent set, which is the necessary condition for adaptive estimation discussed above. Of course, in this example we can see immediately that the partial score for β is equal to the efficient score, implying equality of the Cramer–Rao bound for this parametric model with the semiparametric bound.

An example that is a little more complicated, where it is still easy to calculate the projection, is a limited or full information simultaneous-equations system with unknown distribution for the disturbances. Let y and w be vectors of endogenous variables, let x be a vector of exogenous variables, and let $z = (y, w, x)$. Assume that

$$\rho(z, \beta_0) = \varepsilon, \quad x \text{ and } \varepsilon \text{ independent}, \tag{20}$$

where $\rho(z, \beta)$ is a differentiable function of y and β that is a one-to-one function of y. The nonparametric part of the model includes the distribution of ε and x. Also, in order to allow for limited information systems, the conditional distribution of w given x and ε is included in the nonparametric component. Thus, no restriction is imposed on the reduced form for w. Full information cases are included as a special case where w is not present. The log-likelihood and score vectors for a parametric submodel are,

$$\ln \ell(z, \theta) = J(z, \beta) + \ln f(\rho(z, \beta), \eta) + \ln f(x, \eta) + \ln h(w|x, \rho(z, \beta), \eta). \tag{21}$$
$$S_\beta = J_\beta(z) + [s(\varepsilon) + s(w|x, \varepsilon)]'\rho_\beta(z), \quad S_\eta = D(\varepsilon) + D(x) + D(w|x, \varepsilon)$$

where $J(z, \beta) = \ln|\det(\rho_y(z, \beta))|$, $s(\varepsilon) = f_\varepsilon(\varepsilon)/f(\varepsilon)$, $s(w|x, \varepsilon) = f_\varepsilon(w|x, \varepsilon)/f(w|x, \varepsilon)$, $D(\varepsilon) = f_\eta(\varepsilon)/f(\varepsilon)$, $D(x) = f_\eta(x)/f(x)$, $D(w|x, \varepsilon) = f_\eta(w|x, \varepsilon)/f(w|x, \varepsilon)$, and the subscripts denote partial derivatives evaluated at true parameters. Note that the conditional score for w satisfies $E[D(w|x, \varepsilon)|x, \varepsilon] = 0$, but is otherwise unrestricted. Therefore, the tangent set should be

$$\mathcal{T} = \{D(\varepsilon) + D(x) + D(w|x, \varepsilon): E[D(\varepsilon)] = E[D(x)] = E[D(w|x, \varepsilon)|x, \varepsilon] = 0\}. \tag{22}$$

This set is the sum of three orthogonal components. By ancillarity, the projection of S_β on the $D(x)$ component is zero, while as in the regression model, the projection on the $D(\varepsilon)$ component is $E[S_\beta|\varepsilon]$. It is also easy to check that the projection on the $D(w|x, \varepsilon)$ component is $S_\beta - E[S_\beta|x, \varepsilon]$, so that the efficient score is given by

$$S = S_\beta - E[S_\beta|\varepsilon] - (S_\beta - E[S_\beta|x, \varepsilon]) = E[S_\beta|x, \varepsilon] - E[S_\beta|\varepsilon]. \tag{23}$$

A more detailed formula for the efficient score is not very useful, except in the full information

case, where w is not present. In this case,

$$S = J_\beta(z) - E[J_\beta(z)|\varepsilon] + s(\varepsilon)'(\rho_\beta(z) - E[\rho_\beta(z)|\varepsilon]). \tag{24}$$

A third example that illustrates how the projection can be calculated by inspection is the binary choice model,

$$y = 1(v(x, \beta_0) - \varepsilon > 0), \tag{25}$$

where $1(A)$ denotes the indicator function (equal to 1 if A occurs and zero otherwise), $v(x, \beta)$ a differentiable function of a parameter vector β, and ε is distributed independently of x with differentiable distribution function $F(\varepsilon)$. The nonparametric part of the model is the family of all possible distributions for ε and x. A parametric submodel corresponds to a parametric family of distribution functions $F(\varepsilon, \eta)$ for ε and density functions $f(x, \eta)$ for x. The log-likelihood and scores are

$$\ell(y, x, \theta) = y \cdot \ln F(v(x, \beta), \eta) + (1 - y) \cdot \ln[1 - F(v(x, \beta), \eta)] + f(x, \eta), \tag{26}$$
$$S_\beta = \sigma(\varepsilon)^{-2}[y - F(v)] F_\varepsilon(v) v_\beta, \quad S_\eta = \sigma(v)^{-2}[y - F(v)] F_\eta(v) + D(x),$$

where $D(x) = f_\eta(x)/f(x)$, $v = v(x, \beta_0)$, and $\sigma(v) = \{F(v)[1 - F(v)]\}^{1/2}$. Note that the distribution of x is unrestricted, so that $D(x)$ should be unrestricted except for the mean zero property of scores. Also, as long as $F_\varepsilon(v) > 0$ and $\delta(v)$ is continuously differentiable and zero outside some compact set, $F(v, \eta) = F(v) + \eta' \delta(v)$ will be a distribution function for small enough η. Chamberlain (1986) showed that such $\delta(v)$ are dense, in an appropriate sense. Thus, $F_\eta(v)$ is unrestricted and, absorbing $\sigma(v)^{-2}$ into $D(v)$, the tangent set is

$$\mathcal{T} = \{[y - F(v)] D(v) + D(x) : E[D(x)] = 0\}. \tag{27}$$

The projection follows by Lemma 3.4, with $U = \sigma(v)^{-1}[y - F(v)]$, $W = \sigma(v)^{-1} F_\varepsilon(v) v_\beta$, and $V = v$, giving

$$S = S_\beta - E[W|V]\, U = \sigma(v)^{-2}[y - F(v)] F_\varepsilon(v)\{v_\beta - E[v_\beta|v]\}, \tag{28}$$

a result due to Chamberlain (1986) and Cosslett (1987).

A class of model where calculation of the projection is not particularly easy are those with

$$y = \lambda(v(x, \beta_0) + \varepsilon), \quad x \text{ and } \varepsilon \text{ independent}, \tag{29}$$

where λ is either a known function taking on more than two values (ruling out the binary choice example) or an unknown, strictly monotonic transformation. Various limited dependent variable models are special cases with λ known. If $\lambda(q) = \max\{q, 0\}$ this is a censored regression model, the bound for which was derived by Cosslett (1987), Ritov (1984), and Ritov and Wellner (1988).

One approach to the calculation of the bounds that has proven successful in some cases, including Cosslett's (1987) calculation for censored regression, is to solve the first-order conditions for the projection. This solution involves some additional Hibert space theory. Because of its inherently technical nature, discussion of this approach is reserved to Appendix C.

Another approach that has proven useful for censored regression is to use martigale methods to calculate the efficient score (see Ritov and Wellner, 1988). Such methods might prove useful for other special cases of equation (29), such as ordered response models.

The menu of examples given above are intended to illustrate the methods available for calculation of the projection. In difficult cases the calculation involves the creative use of one of a variety of methods, as is often true with infinite dimensional optimization problems.

Chamberlain (1987a, b) has used a calculation method that sidesteps the projection and is useful for semiparametric models of moment conditions. This method makes use of the multinomial family of distributions, which has two useful properties for moment problems: (i) the semiparametric bound for the parameter of interest takes a simple form for distributions in the family; (ii) distributions in the family can approximate any distribution arbitrarily well, and the same approximation property holds for the corresponding semiparametric bounds.

To illustrate condition (i), consider the mean model, with parameter $\beta_0 = E[z]$. Suppose that the data are multinomial with known points of support z^j, $j = 1, \ldots, J$, with unknown probabilities p_0^j. Let $z = (z^1, \ldots, z^J)'$ and $p = (p^1, \ldots, p^J)'$. For this multinomial distribution $\beta_0 = z' p_0$. It is well known that the Cramer–Rao bound for estimators of p_0 is $V = \operatorname{diag}(p_0) - p_0 p_0'$, where $\operatorname{diag}(p_0)$ is a $J \times J$ diagonal matrix with the components of p_0 on the diagonal and zeros elsewhere. Then by the delta method the Cramer–Rao bound for estimators of β_0 is $z' V z = \sum_{j=1}^J p_0^j (z^j)^2 - [\sum_{j=1}^k p_0^j z^j]^2 = \operatorname{var}(z)$, the asymptotic variance of the sample mean. Indeed, a similar argument can be used to show that the maximum-likelihood estimator of β_0 is numerically equal to the sample mean. Since the multinomial family of distributions is parametric, the Cramer–Rao bound for estimators of β_0 is, trivially, the semiparametric bound.

This type of calculation and the approximation property of the multinomial distribution give the semiparametric bounds for conditional moment restrictions presented in Chamberlain (1987a). Also, Chamberlain (1987b) uses this approach to calculate the bound for a number of semiparametric regression models with conditional mean zero disturbances.

4. SEMIPARAMETRIC ESTIMATION

Semiparametric efficiency bounds and the structure used in their calculation are very helpful in understanding semiparametric estimation problems. This Section discusses impossibility results, construction of estimators, and calculation of their limiting distribution.

4.1. The Infinite Bound Case

Chamberlain (1986) has shown that if the bound is infinitely large (e.g. $E[SS']$ is singular), then no $\sqrt{n}$-consistent, regular estimator can exist. A special case of this result can be seen to hold from Theorem 2.2. By equation (6), the asymptotic variance matrix of a regular asymptotically linear estimator is an upper bound on the Cramer–Rao bounds for all parametric submodels, so that no such estimator can exist if the bound is infinite. Since, as previously discussed, regularity is quite a weak requirement, we are justified in interpreting this result as a statement that if the bound is infinite then no asymptotically linear estimator exists.

An interesting example is the binary choice model where the disturbance is restricted only to have median zero conditional on the regressors. Manski's (1975) maximum score estimator is consistent for the regression parameters of this model, under appropriate conditions. Chamberlain (1986) showed that the efficient score for this model is zero. This is an interesting result, because it implies that the maximum score estimator is not $\sqrt{n}$-consistent. Indeed, Cavanagh (1987) and Kim and Pollard (1989) have shown that for the maximum score estimator $\hat{\beta}$, $n^{1/3}(\hat{\beta} - \beta_0)$ has a non-normal limiting distribution. This example also illustrates that although the efficient score will be zero if the model is unidentified, the converse need not be true.

4.2. Other Impossibility Theorems

The same structure that is used to calculate the efficient score can also be used to show that certain types of semiparametric estimators do not exist. For example, for the binary choice model of section 3 it can be shown that no regular m-estimator based on a fixed (i.e. known) function $m(z, \beta)$ exists.

Theorem 4.1: In the binary choice model with disturbance independent of the regressors there does not exist any regular m-estimator, with fixed (known) m(z, β) as in equation 2.

The critical step in the proof of this result involves varying the conditional distribution of x given v, so that if this distribution is restricted m-estimators could conceivable exist.

It is also possible to show that an analogous result holds in the semiparametric regression model. Also, it is plausible that such a result will hold in other semiparametric models where the tangent set has a structure that is similar to that for the binary choice model, including the sample selection model considered by Chamberlain (1986).

The question of existence of m-estimators based on known functions is related to, but different from, the questions addressed by Gourieroux, Monfort, and Renault (1987). They take as given a moment function, and ask what minimization problems result in consistent estimates. The question that can be addressed by results like Theorem 4.1 is: For a given semiparametric model, what is the class of moment functions that exist? The complete answer to this question would give the class of all functions on which an m-estimator might be based, including those that involve estimating the nonparametric component of the model.

4.3. Convex Models and Semiparametric m-Estimation

There are a number of models where m-estimators based on a fixed function $m(z, \beta)$ exist. A leading example is Powell's (1984) least absolute deviations estimator for the censored regression model with a conditional median zero disturbance. Such estimators are quite useful, providing relatively simple estimators for parameters of interest.

Bickel (1982) and BKRW analyze a property of models that is useful for understanding when such estimators exist and how they might be constructed. In a model with parametric and nonparametric components, let g denote the nonparametric component and let $f(z|\beta, g)$ denote the likelihood as a function of g. Define the model to be *convex* in g if the set of possible values for g is convex and for $0 \leqslant \lambda \leqslant 1$,

$$f(z|\beta, \lambda g + (1 - \lambda)\tilde{g}) = \lambda f(z|\beta, g) + (1 - \lambda) f(z|\beta, \tilde{g}). \tag{30}$$

The censored regression model with a conditional median zero disturbance is an example of a convex model.

Bickel (1982) showed that if a model is convex and the necessary conditions for adaptive estimation of β are satisfied, then quasi-maximum likelihood is consistent. To be precise, in this case the score for β has expectation zero even when the nonparametric part of the model is misspecified, i.e. the true data-generating process is not equal to that corresponding to the score. BKRW generalize this result to show that, in general, in convex models, the efficient score has expectation zero under misspecification. For convenience, a misspecified efficient score function will henceforth be referred to as a *quasi-efficient score*.

Many examples of m-estimators can be viewed as being obtained from quasi-efficient scores. The moment function for Powell's (1984) estimator is equal to the efficient score for a censored

regression model with a disturbance having zero conditional median, where the conditional density at zero is equal to 1 (see Newey and Powell, 1989). Also, the moment function for Powell's (1986) symmetrically trimmed least-squares estimator for the truncated regression model with a conditionally symmetric disturbance is the efficient score for the homoskedastic normal model (see Newey, 1990). For truncated regression, some care must be exercised in the definition of g to show that the model is convex; g has to be defined as the density of the truncated data rather than the density of the latent data.

The importance of the convexity property for quasi-efficient scores to have zero expectation can be seen from the following heuristic argument, which is taken from BKRW. By convexity, a parametric submodel can be formed as $(1-\eta)\ell_0(z)+\eta\ell(z)$, where $\ell_0(z)$ is the true likelihood, and $\ell(z)$ is some other likelihood satisfying the restrictions of the model. If this parametric submodel is regular (see Bickel, 1982, for regularity conditions for convex models), then the score should be

$$S_\eta = \partial \ln[(1-\eta)\ell_0(z)+\eta\ell(z)]/\partial\eta|_{\eta=\eta_0} = \ell(z)/\ell_0(z) - 1. \tag{31}$$

Also, because the efficient score is a linear combination of scores, it will have expectation zero at the truth, i.e. $E[S]=0$. Then by orthogonality of the efficient score S and the tangent set,

$$0 = E[SS_\eta] = E[S\{\ell(z)/\ell_0(z)\}] - E[S] = \int S\ell(z)\,dz. \tag{32}$$

That is, the efficient score S corresponding to the distribution $\ell_0(z)$ has expectation zero when the data are distributed according to some other distribution $\ell(z)$.

An m-estimator based on a quasi-efficient score $\tilde{S}$ has the interesting property of being efficient when the truth is equal to the distribution corresponding to $\tilde{S}$: By equation (5), the influence function for such an estimator should be $(E[\tilde{S}S_\beta'])^{-1}\tilde{S}$, and when $\tilde{S}$ is the true efficient score, it follows by orthogonality of $\tilde{S}$ with the tangent set that $(E[\tilde{S}S_\beta'])^{-1}\tilde{S} = (E[\tilde{S}\tilde{S}'])^{-1}\tilde{S}$, the efficient influence function. BKRW refer to estimators which are efficient for particular values of the nonparametric component, as *locally efficient*.

The discussion of convex models in Bickel (1982) and BKRW can be generalized in ways that are useful in constructing semiparametric estimators in nonconvex models. One useful generalization involves nested convexity. If there is a model $\ell(z|\beta,\alpha,h)$ depending on an additional parametric component α and a nonparametric component h such that for each g, $\ell(z|\beta,g)=\ell(z|\beta,\alpha,h)$ for some α and h, and $\ell(z|\beta,\alpha,h)$ is convex in h, then an m-estimator for β (and α) can be based on the quasi-efficient score obtained by fixing the value of h. For example, consider the linear model $y=x'\beta_0+\varepsilon$, with ε and x independent, which is not convex. Also consider some function $\rho(u)$, and suppose that for each distribution of ε there exists α_0 with $E[\rho(\varepsilon-\alpha_0)]=0$. Then the linear model is nested within a model where ε and x need not be independent, but there is some α_0 such that $\int\rho(\varepsilon-\alpha_0)h(\varepsilon,x)\,d\varepsilon=0$ for the density $h(\varepsilon,x)$ of ε and x. This is a convex model. The efficient score for $\gamma=(\beta',\alpha')'$ is $E[\rho s|x](E[\rho^2|x])^{-1}\cdot X\rho(y-X'\gamma_0)$, where $X=(1,x')'$ and $s=g_\varepsilon(\varepsilon,x)/g(\varepsilon,x)$; e.g. see Newey (1989). Specializing to a case where the leading terms are 1 gives a quasi-efficient score of $X\rho(y-X'\gamma)$, on which an m-estimator of γ (and hence β) can be based. Of course, it can be verified directly that $E[X\rho(y-X'\gamma_0)]=0$. The point here is to show that this result could be derived by nesting the nonconvex regression model where x and ε are independent within a larger convex model. A less transparent example is given in Newey (1989), where a wide class of m-estimators for censored and truncated regression models with an independent disturbance are derived from the efficient scores for a conditional moment restriction model.

It appears that all known examples of m-estimators based on fixed functions and a finite-dimensional nuisance parameters can be interpreted as being based on a quasi-efficient score

for a larger, convex model. It would be interesting to know if this condition is necessary for the existence of such an m-estimator.

Another useful generalization involves component-wise convexity. If a semiparametric model has multiple nonparametric components, and the model is convex in one of these components, then fixing that component should give a quasi-efficient score on which an m-estimator could be based. For example, the linear model with disturbance independent of regressors is convex in either the distribution of the regressors or the distribution of the disturbances. Replacing the piece of the efficient score corresponding to the disturbance distribution, which is $s(\varepsilon)$, by a fixed value $\tilde{s}(\varepsilon)$, such as $\tilde{s}(\varepsilon) = \varepsilon$, gives $\tilde{S} = (x - E[x])\tilde{s}(\varepsilon)$. By independence, $E[\tilde{S}] = 0$, irrespective of the distribution of ε and x. An estimator $\hat{\beta}$ can then be formed by solving

$$\sum_{i=1}^{n} (x_i - \bar{x})\tilde{s}(y - x_i'\hat{\beta})/n = 0, \tag{33}$$

where $\bar{x}$ is the sample mean of x. When $\tilde{s}(\varepsilon) = \varepsilon$, the estimator of β is the least-squares coefficient of x in a regression of y on a constant and x. It is important to note that this estimator involves estimation of part of the nonparametric component, namely $E[x]$. In general, fixing only some pieces of the efficient score means that other parts may have to be estimated. Thus, for estimation of component-wise convex models, it is useful to allow m-estimators to depend on the nonparametric component of the model.

One general way to formulate such estimators, as in Ritov (1987) and Andrews (1989), is the following. Let α denote some function of the nonparametric component, which is possibly infinite-dimensional. Let $m(z, \beta, \alpha)$ be a fixed $q \times 1$ vector such that for some α_0,

$$E[m(z, \beta_0, \alpha_0)] = 0, \tag{34}$$

Because α_0 is unknown, an m-estimator based on $m(z, \beta, \alpha_0)$ is not feasible. A feasible version might be based on a preliminary estimate $\hat{\alpha}(\beta)$ of some $\alpha(\beta)$ such that $\alpha(\beta_0) = \alpha_0$. For example, if α is finite-dimensional, $\hat{\alpha}(\beta)$ might itself be an m-estimator, obtained as the solution to $\sum_{i=1}^{n} h(z_i, \beta, \alpha)/n = 0$. Here $\alpha(\beta)$ would be the solution to $E[h(z, \beta, \alpha)] = 0$, and $\alpha_0 = \alpha(\beta_0)$ would follow from $E[h(z, \beta_0, \alpha_0)] = 0$. The estimate $\hat{\alpha}(b)$ can then be used to estimate β by substituting it for α_0 in the estimation equation corresponding to equation (34), i.e. by choosing $\hat{\beta}$ to solve,

$$\hat{m}_n(\hat{\beta}) = o_p(1/\sqrt{n}), \quad \hat{m}_n(\beta) = \sum_{i=1}^{n} m(z_i, \beta, \hat{\alpha}(\beta))/n. \tag{35}$$

If $\hat{\alpha}(\beta)$ is consistent for $\alpha(\beta)$ in an appropriate metric then one might expect from equation (34) that $\hat{\beta}$ is consistent. The general idea here is that β is obtained by a procedure that first 'concentrates out' α.

A simple example of this type of estimator is that of equation (33), where $\alpha_0 = E[x]$, $\hat{\alpha} = \bar{x}$, and $m(z, \beta, \alpha) = (x - \alpha)\tilde{s}(y - x'\beta)$. An important example is the Buckley and James (1979) estimator for the censored regression model, where $\hat{\alpha}(\beta)$ is the product limit estimator of the distribution of the residuals. Another important example is Robinson's (1988b) estimator for the semiparametric regression model, where $\alpha_0 = (E[y|v], E[x|v]') \equiv (\alpha_{10}(v), \alpha_{20}(v)')'$, $m(z, \beta, \alpha) = [x - \alpha_2(v)]\,[y - \alpha_1(v) - \{x - \alpha_2(v)\}'\beta]$, and $\hat{\alpha}(\beta) = \hat{\alpha}$ is a vector of (trimmed) kernel estimators of the conditional expectations.

To construct such an estimator in a particular model, one needs to find $m(z, \beta, \alpha)$ satisfying equation (34) and suitable $\hat{\alpha}(\beta)$. There is no totally general method of constructing $\hat{\alpha}(b)$, although there is a large body of useful literature on nonparametric estimation that can

provide estimates in particular models. When the efficient score exists in closed form it is easy to find $m(z,\beta,\alpha)$ satisfying equation (34). As noted above, if a model is convex in a nonparametric component, then fixing the value of this component in the efficient score yields such a function. Indeed, since S is a linear combination of scores, a moment function such that $m(z,\beta_0,\alpha_0) = S$ will do. As discussed below, one would even expect that the resulting m-estimator of β is efficient. However, if the efficient score is complicated, it may be computationally convenient to simplify by fixing some of the nonparametric components. Also, the estimate may behave better in finite samples if fewer nuisance functions have to be estimated.

A final useful generalization of convexity involves combining nested and component-wise convexity. If a model can be nested within a model that is convex in a nonparametric component, then a quasi-efficient score can be formed by fixing that nonparametric component. An example is the semiparametric regression model. This model is not convex in g because mean mixtures of normal distributions are not normal. A model which nests this one, and is convex in g, is one where $E[y|x] - x'\beta_0$ is some function of v, with efficient score given by Chamberlain (1987b). Fixing $\tilde{g}(v) = 0$ and $\text{var}(y|x) = 1$ gives

$$m(z,\beta,\alpha) = -[x - \alpha(v)](y - x'\beta), \tag{36}$$

where $\alpha(v)$ indexes $E[x|v]$. It can be verified directly that equation (34) is satisfied for $\alpha_0 = E[x|v]$. A semiparametric estimator of β can be obtained by replacing $E[x|v]$ by nonparametric regression estimate $\hat{E}[x|v]$. The result can be interpreted as an instrumental variables estimate of β for instruments $x - \hat{E}[x|v]$. This estimator is somewhat simpler than and, as argued below, asymptotically equivalent to Robinson's (1988b) estimator.

It is interesting to note that the consistency of this estimator is sensitive to the estimate of $E[x|v]$, a function of the marginal distribution of the regressors, despite the fact that this distribution is ancillary for β. This phenomenon occurs because the component of the efficient score corresponding to the conditional distribution of y given x has been fixed at a value that may be false. Intuitively, misspecification of the conditional distribution can destroy ancillarity. In the parametric quasi-maximum-likelihood context, scores for the conditional and marginal likelihoods need not be orthogonal if the conditional likelihood is misspecified.

A consequence of this phenomenon is that restrictions on the marginal distribution can affect the consistency of estimators of conditional distribution parameters, an observation made by Ruud (1983) for discrete choice models. For example, in the semiparametric regression model, suppose that the conditional expectation of x given v is linear in v, say $E[x|v] = \alpha_{10} + \alpha_{20}v$, such as would be the case if x and v are multivariate normal. Then $E[x|v]$ can be estimated by $\hat{\alpha}_1 + \hat{\alpha}_2 v$, where $\hat{\alpha}_1, \hat{\alpha}_2$ are the coefficients of the linear regression of x_i on a constant and v_i. It is easily verified that the above estimator of β is just the least-squares estimator from a regression of y_i on x_i, v_i, and a constant. It follows that this least-squares estimator is consistent whenever $E[x|v]$ is linear in v, *irrespective of the form of* $g_0(v)$. Thus, the restriction that $E[x|v]$ is linear results in consistency of an estimator where $g(v)$ is misspecified as linear in v.

An important identification condition for consistency of a semiparametric m-estimator is uniqueness of the solution to

$$E[m(z,\beta,\alpha(\beta))] = 0, \tag{37}$$

for all $\beta \in B$, where B is some known set containing β_0 in which $\hat{\beta}$ is restricted to lie. This condition can be difficult to verify and/or impose strong conditions on B. BKRW give some useful primitive conditions for the case where $\alpha(\beta)$ is not present. Also, Gourieroux, Monfort,

and Renault (1987) give a characterization of when $E[m(z,\beta)] = 0$ corresponds to the first-order conditions of some maximization problem, $\max_{\beta \in B} E[l(z,\beta)]$. Uniqueness of the maximum can hold under weaker conditions than uniqueness of the solution to the first-order conditions, and can be easier to verify; e.g. see Powell (1984).

4.4. The Asymptotic Distribution of Semiparametric *m*-estimators

Calculation of the limiting distribution of a semiparametric m-estimator is an important step for asymptotic inference procedures for parameters of interest. This calculation is straightforward when the moment function depends only on parameters of interest. Under the regularity conditions of Huber (1967) or Pollar (1985), the estimator will be asymptotically linear with influence function $-M^{-1}m(z,\beta_0)$, where $M = \partial E[m(z,\beta_0)]/\partial\beta$, so that its asymptotic covariance matrix is $M^{-1}E[mm']M^{-1\prime}$. The limiting distribution can be more complicated when estimated nonparametric components $\hat{\alpha}(\beta)$ are present.

In order to study the limiting distribution with $\hat{\alpha}(\beta)$ present, impose the following assumption.

Assumption 4.1: (i) $\hat{\beta} \xrightarrow{p} \beta_0$; (ii) $\hat{m}_n(\beta)$ is continuously differentiable with probability one; (iii) for any $\bar{\beta} \xrightarrow{p} \beta_0$, $\partial\hat{m}_n(\bar{\beta})/\partial\beta \xrightarrow{p} M \equiv \partial E[m(z,\beta,\alpha(\beta))]/\partial\beta\,|_{\beta=\beta_0}$, and M is nonsingular.

Conditions (i) and (iii) are convergence conditions that might require much work to check in a particular model, and condition (ii) is imposed to simplify the following discussion. This assumption is not meant to be primitive, only to allow a simple but rigorous analysis.

If, in addition to Assumption 4.1, $\sqrt{n}\hat{m}_n(\beta_0)$ is bounded in probability, then the usual mean value expansion gives

$$\sqrt{n}(\hat{\beta} - \beta_0) = -M^{-1}\sqrt{n}\hat{m}_n(\beta_0) + o_p(1); \tag{38}$$

e.g. see Lemma A.3 of Appendix A. Here calculating the asymptotic distribution reduces to finding a formula for M and for the distribution of $\sqrt{n}\hat{m}_n(\beta_0)$.

The parametric case provides insight concerning $\sqrt{n}\hat{m}_n(\beta_0)$. Suppose that $\hat{\alpha}(\beta)$ is a finite-dimensional vector m-estimator solving $\sum_{i=1}^n h(z_i,\beta,\hat{\alpha}(\beta))/n = 0$, where $E[h(z,\beta_0,\alpha_0)] = 0$. Ignoring the presence of $\hat{\alpha} \equiv \hat{\alpha}(\beta_0)$ in $\sqrt{n}\hat{m}_n(\beta_0) = \sum_{i=1}^n m(z_i,\beta_0,\hat{\alpha})/\sqrt{n}$ could lead to the wrong formula for the asymptotic covariance matrix of $\sqrt{n}\hat{m}_n(\beta_0)$. This phenomenon is familiar from the literature on two-step estimators, e.g. Pierce (1982), Newey (1984), and Pagan (1986). The presence of $\hat{\alpha}$ can be accounted for by a mean-value expansion in $\hat{\alpha}$. Assuming that $m(z,\beta,\alpha)$ is differentiable in α, the sample average of the derivative converges uniformly in a neighbourhood of α_0, and $\hat{\alpha}$ obeys the usual asymptotic linearity condition for m-estimators (see section 2), then for $m_i = m(z_i,\beta_0,\alpha_0)$, $h_i = h(z,\beta_0,\alpha_0)$, $m_n = \sum_{i=1}^n m_i/n$,

$$\sqrt{n}\hat{m}_n(\beta_0) = \sqrt{n}m_n + \left[\sum_{i=1}^n \partial m(z_i,\beta_0,\bar{\alpha})/\partial\alpha/n\right]\sqrt{n}(\hat{\alpha} - \alpha_0) \tag{39}$$

$$= \sum_{i=1}^n \{m_i - E[\partial m(z,\beta_0,\alpha_0)/\partial\alpha](E[\partial h(z,\beta_0,\alpha_0)/\partial\alpha])^{-1}h_i\}/\sqrt{n}.$$

The term following the minus sign corrects for the presence of $\hat{\alpha}$ in the sample average $\hat{m}_n(\beta_0)$. In a model with parameters $\theta = (\beta', \alpha')'$, a more illuminating expression for this term can be obtained from the generalized information matrix equality. Differentiating $E_\theta[m(z,\theta)] = 0$ and $E_\theta[h(z,\theta)] = 0$ with respect to α under the integral and substituting the resulting formulas in

equation (39) gives

$$\sqrt{n}\hat{m}_n(\beta_0) = \sum_{i=1}^{n} \{m_i - E[mS_\alpha'](E[hS_\alpha'])^{-1}h_i\}/\sqrt{n} + o_p(1), \tag{40}$$

where S_α is the score for α.

There are two features of equation (40) that are useful for understanding the semiparametric case. First, note that if $E[mS_\alpha'] = 0$, then equation (40) reduces to $\sqrt{n}\hat{m}_n(\beta_0) = \sqrt{n}m_n + o_p(1)$, and the presence of $\hat{\alpha}$ can be ignored in calculating the limiting distribution of $\sqrt{n}m_n(\hat{\beta})$. The nonparametric analogue of $E[mS_\alpha'] = 0$ is orthogonality of m and the tangent set, since α represents the components of the model other than β, and $E[mS_\alpha'] = 0$ is equivalent to orthogonality of m and combinations of the score (see Lemma A.1 in Appendix A). Thus, one would expect that for semiparametric models, orthogonality of m and the tangent set is an important condition for $\sqrt{n}\hat{m}(\beta_0)$ and $\sqrt{n}m_n$ to have the same limiting distribution. Secondly, note that if $\hat{\alpha}$ is efficient, so that $h = S_\alpha$, then $E[mS_\alpha'](E[hS_\alpha'])^{-1}h = E[mS_\alpha'](E[S_\alpha S_\alpha'])^{-1}S_\alpha$, which is the projection of m on S_α. The nonparametric analogue of this term is the projection of m on the tangent set. Thus, in semiparametric models, one would expect to find that when $\hat{\alpha}$ is efficient in an appropriate sense, then the presence of $\hat{\alpha}$ can be corrected for by replacing m with the residual from the projection of m on the tangent set. Indeed, BKRW give quite primitive conditions under which this calculation is appropriate when $\hat{\alpha}(\beta)$ is a function of the nonparametric maximum-likelihood estimator of a distribution.

It is possible to make these observations rigorous for semiparametric models in the following way. Define $\bar{m}(\alpha) = E[m(z, \beta_0, \alpha)]$, and note that $\bar{m}(\alpha_0) = 0$. Suppose

Assumption 4.2: (i) $\sqrt{n}[\hat{m}_n(\beta_0) - \bar{m}(\hat{\alpha}) - m_n] = o_p(1)$; (ii) $\sqrt{n}\bar{m}(\hat{\alpha}) = \sum_{i=1}^{n}\tau_i/\sqrt{n} + o_p(1)$, where $E[\tau'\tau] < \infty$ and $E[\tau] = 0$.

Condition (i) is a type of stocastic equicontinuity condition, that has been used by Andrews (1989). If condition (i) is strengthened, then it is possible to weaken the assumption that $\hat{m}_n(\beta)$ is continuously differentiable (see Andrews, 1989). It follows from Assumption 4.2 that $\sqrt{n}\hat{m}_n(\beta_0) = \sum_{i=1}^{n}(m_i + \tau_i)/\sqrt{n} + o_p(1)$, so that τ_i here is a correction term for the presence of $\hat{\alpha}$ analogous to that of equation (40).

The previous discussion leads one to expect that $\tau = 0$ is related to orthogonality of m and the tangent set. To be rigorous, it is helpful to impose regularity of $\hat{\beta}_0$. The following result is a simple consequence of Theorem 2.2.

Theorem 4.2: Suppose that Assumptions 4.1 and 4.2 are satisfied and $\hat{\beta}$ is regular. Then if $\tau = 0$, m is orthogonal to the tangent set. Also, if m is orthogonal to the tangent set and $\tau \in \mathscr{T}$, then $\tau = 0$, and $\hat{\beta}$ has influence function $(E[mS'])^{-1}m$.

The condition that $\tau \in \mathscr{T}$ can be interpreted as an efficiency condition for $\bar{m}(\hat{\alpha})$. If each possible α corresponds to a particular value g for the nonparametric part of the model, then $\bar{m}(\alpha)$ is a parameter of the nonparametric component. Also, Assumption 4.2 specifies that $\sqrt{n}\bar{m}(\hat{\alpha})$ is an asymptotically linear estimator (of zero) with influence function τ. By Theorem 3.1, efficiency of a regular asymptotically linear estimator at a particular distribution is equivalent to the influence function being in the tangent set.

When m is orthogonal to the tangent set, one might expect that $\bar{m}(\hat{\alpha})$ is efficient for a wide variety of choices of α. In the parametric analogue, if an additional expansion were applied to the correction term following the first equality in equation (39), then it would be possible to show that $\text{plim}(n^{1/4}(\hat{\alpha} - \alpha_0)) = 0$ would be sufficient for the correction term to be zero. The nonparametric version of this is that sufficient convergence rates for $\hat{\alpha}$ should suffice for $\tau = 0$.

Robinson's (1988b) semiparametric regression estimator is an example with $m = (x - E[x|v])\varepsilon$ orthogonal to the tangent set. Here it is easily checked that $\|\bar{m}(\alpha)\| \leq E[\|\alpha(v) - \alpha_0(v)\|^2]$ so that, e.g. $n^{1/4} sup_v|\hat{\alpha}(v)^- \alpha_0(V)| = o_p(1)$ would suffice for $\sqrt{n}\nu(\hat{\alpha}) = o_p(1)$, i.e. for $\tau = 0$.

One circumstance where $\tau = 0$ is guaranteed is if $m(z, \beta_0, \alpha_0)$ is the quasi-efficient score for a convex model. In this case $\bar{m}(\alpha) = 0$, since α is a function of the nonparametric component, implying $\sqrt{n}\bar{m}(\hat{\alpha}) = 0$.

In general, efficiency of $\bar{m}(\hat{\alpha})$ leads to the projection formula for the correction term. Let u be the residual from the projection of m on τ.

Theorem 4.3: Suppose that Assumptions 4.1 and 4.2 are satisfied and $\hat{\beta}$ is regular. If $\tau \in \mathcal{T}$, $\hat{\beta}$ has an influence function $(E[us'])^{-1}u$.

One circumstance where $\tau \in \mathcal{T}$ must hold is if $\bar{m}(\alpha)$ depends only one functions of z with an unrestricted distribution. Then, as discussed in section 3, all regular asymptotically linear estimators, such as $\bar{m}(\hat{\alpha})$, are efficient.

Proposition 4.4: Suppose that Assumptions 4.1 and 4.2 are satisfied and $\hat{\beta}$ is regular. If for $w = r(z)$, $\hat{\alpha}(\beta_0)$ depends only on $(w_1, \ldots, w_n)$, and the semiparametric model does not restrict the distribution of w, then $\hat{\beta}$ is locally regular with influence function $(E[uS'])^{-1}u$.

For example, consider the m-estimator for the semiparametric regression model, with moment function in equation (36). The model does not restrict the distribution of x and v, so that as long as $\hat{E}[x|v]$ depends only on (x_i, v_i), the hypotheses of this result are satisfied. Recall that $\mathcal{T} = \{\varepsilon D_1(v) + D_2(x, v)\}$, where D_1 and D_2 are unrestricted except for $E[D_2] = 0$. Then by $m = (x - E[x|v])(\varepsilon + g_0(v))$ is orthogonal to $\varepsilon D_1(v)$,

$$u = m - E[m|x, v] = m - (x - (E[x|v])g_0(v) = (x - E[x|v])\varepsilon$$

Then by Proposition 4.4, the influence function is $\{E[\text{var}(x|v)]\}^{-1}(x - E[x|v])\varepsilon$. This influence function is the same as that of Robinson's (1988b) estimator, suggesting that the two are asymptotically equivalent.

These results suggest that it is possible to modify a particular m-estimator so that estimation of the nonparametric components does not affect its limiting distribution. Consider a moment function equal to the residual $u(z, \beta, \tilde{\alpha})$ from the projection of $m(z, \beta, \alpha)$ on the tangent set for given β and α, where $\tilde{\alpha}$ includes any additional components that enter in the projection. This moment function is orthogonal to the tangent set by construction, suggesting that estimation of the nonparametric components should not affect the limiting distribution of the corresponding m-estimator. Note, though, that this estimator will not be asymptotically equivalent to the original m-estimator unless the correction term for the original estimator is an element of the tangent set.

When the correction term is in the tangent set, this modification of the moment function is useful for estimation of the asymptotic convariance matrix of $\hat{\beta}$, which by the conclusion of Theorem 4.3 is $M^{-1}E[uu']M^{-1\prime}$, where $u = m - \text{Proj}(m|\mathcal{T})$. If Assumption 4.1 is satisfied then M can be estimated by $\partial\hat{m}_n(\hat{\beta})/\partial\beta$. If $\hat{m}_n(\beta)$ is not differentiable, then one may have to resort to a numerical derivative, e.g. Newey (1990). Estimation of $E[uu']$ is more difficult, but if $u(z, \beta, \alpha_u)$ is calculated as discussed in the previous paragraph and estimated by $\hat{u}_i = u(z_i, \hat{\beta}, \hat{\alpha}_u)$, where α_u are nuisance parameters of the efficient score with corresponding estimates $\hat{\alpha}_u$, then $\sum_{i=1}^n \hat{u}_i\hat{u}_i/n$ should be a consistent estimator of $E[uu']$ under appropriate regularity conditions. A slightly more primitive condition for consistency of $\sum_{i=1}^n \hat{u}_i\hat{u}_i/n$ is $\sum_{i=1}^n \|\hat{u}_i - u_i\|^2/n = o_p(1)$; see Powell, Stock, and Stoker (1990).

5. EFFICIENT ESTIMATION

One efficiency criterion is local efficiency, which is efficiency for particular values of the nonparametric component of the model, e.g. for normal disturbances. Such estimators are best estimators for particular distributions, subject to the constraint that the estimator be semiparametric. Important examples are Powell's (1986) symmetrically trimmed least-squares estimator for truncated regression models and Robinson's (1988b) estimator for the additive semiparametric regression model, both of which are efficient for normal, homoskedastic disturbances.

A more ambitious efficiency criterion is global efficiency, which is efficiency for all values of the nonparametric component of the model. An example of such an estimator is Bickel's (1982) adaptive estimator for the slope parameters (coefficients of nonconstant regressors) of a regression model with independence of disturbance and regressors. This estimator uses a nonparametric estimate of the score function of the disturbance density to obtain an estimate of the parameters that is efficient for all disturbance distributions. The nature of other globally efficient estimators is similar; to achieve efficiency they must be able to adjust their form, depending on the value of the nonparametric part of the model. The remainder of this section discusses results on globally efficient estimation.

Bickel (1982), Schick (1986), and Klaasen (1987) have presented fundamental results on the construction of (globally) efficient estimates. Their constructions are based on a one-step estimator, starting at an initial $\sqrt{n}$-consistent estimator. Let $\hat{d}(z, \beta)$ denote an estimate of the efficient influence function, for a given value of β. Their estimates take the form

$$\tilde{\beta} = \hat{\beta} + \sum_{i=1}^{n} \hat{d}(z_i, \hat{\beta})/n, \tag{41}$$

where $\sqrt{n}(\hat{\beta} - \beta_0)$ is bounded in probability. For example, if $\hat{S}(z, \beta)$ is an estimate of the efficient score function, then $[\Sigma_{i=1}^{n} \hat{S}(z_i, \hat{\beta})\hat{S}(z_i, \hat{\beta})/n]^{-1}\hat{S}(z, \beta)$ is an estimate of the influence function, and $\tilde{\beta}$ becomes

$$\tilde{\beta} = \hat{\beta} + \left[\sum_{i=1}^{n} \hat{S}(z_i, \hat{\beta})\hat{S}(z_i, \hat{\beta})/n\right]^{-1} \sum_{i=1}^{n} \hat{S}(z_i, \hat{\beta})/n. \tag{42}$$

This is a one-step version of a semiparametric m-estimator, with $\hat{m}_n(\beta) = \Sigma_{i-1}^{n} \hat{S}(z_i, \beta)/n$. By the reasoning of section 4, one would expect that by orthogonality of S with the tangent set, the estimation of the score should not affect the limiting distribution of $\tilde{\beta}$ as long as the score function is estimated at a sufficiently fast rate. Consequently, $\tilde{\beta}$ should be efficient.

Bickel (1982) and Schick (1986) give weak conditions on the influence function estimate $\hat{d}(z, \beta)$ that are sufficient for efficiency of $\tilde{\beta}$. Klaasen (1987) shows that under a uniform integrability condition for the efficient influence function, a certain set of sufficient conditions is also necessary. Their specific constructions make use of discretization of the parameter space and sample splitting. Discretization of the parameter space involves restricting $\check{\beta}$ to a finite set that is becoming finer at rate $1/\sqrt{n}$. This allows one to treat the estimator as a nonstochastic sequence; e.g. see Manski (1984). Sample splitting involves using a part of the sample that does not include the ith observations to estimate the influence function $\hat{d}(z_i, \beta)$ for the ith observation. With sample splitting, the convergence rate for $\hat{d}(z_i, \beta)$ is minimal. In Schick's (1986) construction, essentially half of the sample is used in the estimation of $\hat{d}(z_i, \beta)$. See BKRW for a more complete account.

Methods of estimating the efficient influence function or the efficient score are inherently specific to the model. Bickel (1982) and BKRW have presented kernel-based estimates for

models where the nonparametric piece of the efficient score is the (location parameter) score for a density. The estimate of the density score equals the ratio of the derivative of a kernel density estimate, trimmed to be zero for some data configurations. Efficient estimators in a variety of other semiparametric models are given in the references in the next section. Newey (1988b) has presented some general results on series estimation of the efficient score by means of a sufficient number and variety of moment functions.

One general approach to efficient estimation is nonparametric maximum-likelihood. Econometric examples are Cosslett (1983) for binary choice, and Heckman and Singer (1984) for duration models with unobserved heterogeneity. One might hope that, as in parametric models, nonparametric maximum-likelihood often results in efficiency. Unfortunately, nonparametric maximum-likelihood estimators do not always exist, and when they do they can be hard to analyze. These problems have led to the introduction of a number of approximate nonparametric maximum-likelihood methods.

One such method is sieve estimation, terminology due to Grenander (1981). Sieve estimation involves a finite-dimensional approximation of the nonparametric component of the model, where the dimension of the approximation is allowed to grow with the sample size. Econometric examples are Duncan (1986) for censored regression, and Gallant and Nychka (1987) for sample selection. As discussed by Severini and Wong (1987), one would expect that sieve estimation results in an efficient estimator of the parameters of interest. Intuitively, in terms of the projection interpretation of the bound, as the parametric approximation becomes richer, the Cramer–Rao bound should approach the semiparametric bound, because the scores for the parametric approximation approximately span the tangent set. The precise argument is somewhat delicate, because of the bias introduced by the parametric approximation, but this intuition should be correct if the bias disappears sufficiently fast as the degree of approximation increases and other appropriate regularity conditions are satisfied.

The finite sample behaviour of the estimators, including comparative performance of different efficient estimators for the same model, remains largely an open question. The most seems to be known about adaptive estimation of the slope parameters in a regression model with disturbance independent of regressors, where several Monte Carlo experiments have been carried out; see Hsieh and Manski (1987), Newey (1988a), and Portnoy and Koenker (1989). In those examples the series estimator of Newey (1988a) and the L-estimator of Portnoy and Koenker (1988) perform substantially better than the kernel-based estimator.

6. A REVIEW OF THE LITERATURE AND RECENT RESULTS

This section contains a review of semiparametric efficiency results for econometric models. To organize the material it is helpful to classify semiparametric econometric models. In econometrics it is typically the case that the parameters of interest, such as a demand or labour supply elasticity, quantify the response of an economic variable to change in conditions. A way to formalize this is that there is some response

$$y = R(w, \varepsilon, \beta_0), \tag{43}$$

where the vector of functions R quantifies the effect of observed variables w and unobserved variables ε on variables y of interest, and depends on unknown parameters β_0. In this context semiparametric models are those where the distribution of w and ε is not restricted to a parametric family and/or the functional form of $R(w, \varepsilon, \beta)$ is partly unknown. Both types of models are important.

6.1. Parametric Response Function

Identification of the parameter β of the response function requires that there be some conditions placed on the statistical relationship between observables and unobservables. These conditions typically take the form of a restriction on the conditional distribution of ε given a subset of variables x of w. Three important kinds of restrictions are:

1. conditional moment restrictions of the form $E[m(\varepsilon, \alpha_0)|x] = 0$ where $m(\varepsilon, \alpha)$ is some known function and α_0 is unknown;
2. independence of ε and x;
3. conditional symmetry of ε given x.

These types of restriction are related, and give an interesting menu of semiparametric models. Typically the conditional moment restrictions are weaker than independence or symmetry. For example, if ε is a scalar and $m(\varepsilon, \alpha)$ equals $\varepsilon - \alpha$ $(\text{sgn}(\varepsilon - \alpha))$, then (1) means that the conditional mean (median) or ε does not depend on x. Both conditional symmetry and independence would imply this restriction. Indeed, a conditional moment restriction can be the weakest type of restriction that gives identification of β, although it has the disadvantage that the statistical interpretation of the parameters of interest can depend on the function $m(\varepsilon - \alpha)$, e.g. mean versus median.

Independence and conditional symmetry are complementary restrictions. Heteroskedasticity is allowed for under conditional symmetry, but not under independence. Independence does not impose the strong shape restriction imposed by symmetry. Although symmetry has received little attention in econometrics, Powell's (1986) estimators for Tobit models show its potential. Models involving some symmetrizing transformation of the dependent variable or residual, such as the Box–Cox model, may prove particularly useful.

To give a coherent review the references will be organized into tables. The columns of the tables will correspond to three different types of restrictions on the conditional distribution of ε given x, and the rows to different models. The entries in the tables will be references where bounds are derived, or efficient estimators presented, or both. The efficient estimators referred to in the tables are globally efficient, rather than locally efficient. The references with no asteriks are bounds results without construction of efficient estimators.

Table I summarizes the literature and recent work on regression and nonlinear simultaneous equations models. The multiple references under the conditional moment heading result from consideration of different types of conditional moment restrictions and from work on different

Table I. Regression and nonlinear simultaneous equation

	Conditional moment	Independence	Conditional symmetry
Regression	Chamberlain (1987a) Carroll (1982) Robinson (1987)* Newey (1987)* Newey and Powell (1989)** Newey (1989)**	Bickel (1982)** Manski (1984)* Newey (1988a)* Portnoy and Koenker (1989)*	Manski (1984) Newey (1988a)*
Nonlinear simultaneous equation	Chamberlain (1987a) Newey (1987)*	Newey (1988b)**	—

*Efficient Estimator.
**Efficiency Bound and Estimator.

types of estimators. Chamberlain (1987a) gives the efficiency bound for a conditional mean zero disturbance, showing that it corresponds to heteroskedasticity corrected generalized least-squares (GLS). Carroll (1982), Robinson (1987) and Newey (1987) give different efficient estimators, involving a nonparametric regression estimate of the conditional variance in a feasible GLS procedure and many moment restrictions. Newey and Powell (1989) carry through with analogous exercises for the conditional median zero case. The most complete set of results is contained in Newey (1989), where a formula for the efficiency bound and construction of an efficient estimator is given that applies to any single conditional moment restriction.

The multiple references in the independence case are a consequence of the amount of work that has gone into the efficient estimation problem. Bickel's (1982) work was followed by that of Manski (1984), extending Bickel's (1982) results to nonlinear models, and Newey (1988a) and Portnoy and Koenker (1989), given different estimators with improved small sample performance.

The second set of results that will be reviewed concerns univariate limited dependent variable models consisting of the binary choice, truncated, and censored regression models. Each of these models can be thought of as arising from a latent variable model of the form

$$y_i^* = x_i\beta_0 + \varepsilon_i,$$

where y_i^* and/or x_i are not always observed. In the binary choice model x_i is always observed but only the sign of y_i^* is observed, in the truncated model y_i^* and x_i are only observed if y_i^* is positive, and in the censored model y_i^* is only observed when it is positive, but x_i is always observed.

Interest in semiparametric versions of these models arose because misspecification of the distribution of ε_i generally makes the maximum-likelihood estimator of β inconsistent; see Goldberger (1983), Hurd (1979), Abrazamar and Schmidt (1981, 1982). This inconsistency led to work on semiparametric estimation of these models, notably Manski (1975), Powell (1984), and Ruud (1986); see Robinson (1988a) for a more complete set of references. Table II summarizes the literature on semiparametric efficiency of these models.

There has been little work on multivariate limited dependent variable models, despite the fact that several of these models, such as ordered and multinomial choice, and disequilibrium models, are important for applications. An exception is the sample selection model. This model has two dependent variables, one of which is binary, and the other only observed if the binary variable is positive. Chamberlain (1986) has derived the bound for the case with disturbances independent of regressors. Estimation of this model has been discussed in Cosslett

Table II. Univariate limited dependent variable

	Conditional moment	Independence	Conditional symmetry
Binary choice	Chamberlain (1986)	Chamberlain (1986) Cosslett (1987) Klein and Spady (1987)*	—
Truncated	Newey (1989)**	Cosslett (1987) BKRW	Newey (1990)**
Censored	Newey and Powell (1989)** Newey (1989)**	Cosslett (1987) Ritov (1984)**	Newey (1990)**

*Efficient estimator.
**Efficiency bound and efficient estimator.

(1988), Gallant and Nychka (1987), Powell (1987), and Newey (1988c), with an efficient estimator given in Newey (1988c).

Another important type of semiparametric model involving unknown distributions of unobservables is an endogenous sampling model. In such a model the sample has been selected on the basis of one or more dependent variables, rather than randomly. A fairly complete set of semiparametric efficiency bounds is given in independent work by Cosslett (1985) and BKRW. BKRW give elegant characterization results for the efficiency bounds, without computing closed form expressions for many of them. Cosslett (1985) computes closed form expressions for several different cases.

There are gaps in the efficient estimation literature for these models. Cosslett (1981a, b) and Manski and McFadden (1981) present results for choice-based sampling in discrete choice models. Other cases remain to be considered, although there has been some work on semiparametric estimation in such models; see Hausman and Wise (1981) and Jewell (1985).

6.2. Semiparametric Response Functions

Models with parametric response function and nonparametric distributions corresponds loosely to the notion that, by virtue of economic theory, we tend to know more about response functions than distributions. Of course, the restrictions on the response function implied by economic theory are typically nonparametric, so that it is useful to be able to weaken functional form assumptions on the response function. Specifying the response function to have a parametric and a nonparametric component is one way to do this. Such statistical models make up a useful intermediate ground between parametric and fully nonparametric response functions.

Four type of semiparametric response models will be discussed here. They are weighted average derivative (WAD), semiparametric regression (SR), single index—referred to as projection pursuit (PP), and monotonically transformed regresson (MR). Each of these models can be thought of as special cases of the response function of equation (43), involving restrictions on the form of $R(w, \varepsilon, \beta_0)$.

WAD parameters, due to Stoker (1986) and Powell, Stock, and Stoker (1990), are a weighted expectation of the derivative of the conditional expectation of y given w. That is, for a weight function $\omega(w)$

$$\beta_0 = E[\omega(w)\partial E[y|w]/\partial w].$$

Estimators of these parameters have been developed by Powell, Stock, and Stoker (1990). Conditions for differentiability of these parameters, as discussed in section 3, and a corresponding efficiency result for the estimators, are given in Newey and Stoker (1989).

The other three models involve explicit functional form assumptions about the response function. The SR model is that of Engle *et al.* (1986):

$$y = x'\beta_0 + g_0(v) + \varepsilon,$$

where $w = (x', v')'$ and the function $g_0(\cdot)$ is unknown. The PP model, terminology motivated by Friedman and Stueltze (1981), is

$$y = h_0(w'\beta_0) + \varepsilon,$$

where the function $h_0(\cdot)$ is unknown. The MR model is

$$\lambda_0(y) = w'\beta_0 + \varepsilon,$$

where $\lambda_0(\cdot)$ is an unknown monotonic function. The response function for this model is $R(w, \varepsilon, \beta_0) = \lambda_0^{-1}(w'\beta_0 + \varepsilon)$, which is qualitatively different from the response functions of the other models in that it is nonlinear in ε. It should be noted that, in both this model and the PP model, the parameters are only identified up to scale.

Efficiency for semiparametric response functions has been considered under several different assumptions on the conditional distribution of ε. Here results for three different assumptions will be discussed: conditional mean zero, independence of ε and w, and normality of ε. The normality assumption serves to focus attention on the semiparametric nature of the response function, and gives efficiency interpretations to some estimators that have been proposed for semiparametric response models. Table III summarizes the results. It is plausible that the MR model is not identified in the conditional mean case, because nonlinear transformations do not preserve mean restrictions. On the other hand, if ε were restricted to have median zero, then the bound should be finite. Since the median of a monotonic transformation is a monotonic transformation of the median, the conditional median of y would be $\lambda_0^{-1}(w'\beta_0)$, a median version of PP.

The efficient estimators under normality for the PP and SR models can also be interpreted as locally efficient at the homoskedastic, Gaussian model, in a model where the disturbance has conditional mean zero; see Chamberlain (1987b) and Newey and Stoker (1989). In particular, the consistency of these estimators does not depend on the normality assumption.

There are some relationships among the semiparametric response models and the semiparametric distribution models of section 6.1. First, the MR model with ε is independent of w is a special case of PP model with zero conditional mean, since $E[y|w] = E[\lambda_0^{-1}(w'\beta_0 + \varepsilon)|w] = h_0(w'\beta_0)$. Also, the univeriate limited dependent variable models with an independent latent disturbance are special cases of the PP model. For example, the binary choice and censored models are monotonic regression models with known transformations, so the same calculation gives a PP model. A useful consequence of this relationship is that estimators that work for the PP model, such as Ichimura's (1986), also work for the specal cases. However, one typically loses information by using a PP estimator for a special case. One loses the identification of the scale for β_0 in the censored and truncated regression models. There is also loss of information in estimating the parameters up to scale. The models imply that the entire conditional distribution of the dependent variable, and not just its mean, depends on the single index $w'\beta_0$. In general the efficiency bound for such a model is different from the efficiency bound for the PP model; see Newey (1988b) for the bound and discussion. Furthermore, it appears that independence gives still more information than just the conditional distribution depending on the index $x'\beta_0$. Similar remarks about loss in information apply to many other special case results.

Table III. Semiparametric response functions

	Conditional mean	Independence	Normality
Semiparametric regression	Chamberlain (1987b)	BKRW	BKRW Robinson (1988b)*
Projection pursuit	Chamberlain (1987b) Newey (1988b)*	BKRW	BKRW Ichimura (1986)*
Monotonically transformed regression	—	BKRW	BKRW

*Efficient estimator.

6.3. Duration Models

An important model that involves nonparametric assumptions on both distributions and response functions is the duration model. To save space, a definition of the model will be omitted. Instead, some results will be related and important open problems mentioned that may be of interest to readers familiar with this model.

The semiparametric versions of this model that have been considered can be classified according to whether the baseline hazard, or the distribution of unobserved heterogeneity, is allowed to be nonparametric. For the case with no unobserved heterogeneity Cox (1975) developed a partial-likelihood estimator of the covariate cofficients that did not utilize functional form restrictions on the baseline hazard. Begun *et al.* (1983) derived the semiparametric bound for this model and showed that the partial-likelihood estimator attains this bound. Ritov and Wellner (1988) use martingale theory to give an easier treatment of the bounds.

There appear to be no known efficiency results for the case where there is unobserved heterogeneity with unknown distribution. Results on identification and estimation of such models are recent, e.g. Heckman and Singer (1984), and asymptotic distribution theory for the proposed estimators has not yet been developed. Derivation of semiparametric efficiency bounds is also an important open problem.

6.4. Time-series Models

There are a few results on semiparametric efficiency in time-series models. Here two papers will be discussed, that of Hannan (1963) and that of Hansen (1988).

Some of the earliest work on efficient semiparametric estimation was carried out by Hannan (1963). He developed a feasible generalized least-squares (GLS) estimator that is efficient with autocorrelation of unknown form. This estimator is an efficient semiparametric estimator for the following model. Suppose that a dependent variable y_t satisfies

$$y_t = x_t'\beta_0 + \varepsilon_t,$$

where x_t and ε_t are independent, stationary processes, and ε_t is Gaussian. The nonparametric part of this model is the autocorrelation function of ε_t. In terms of the classification of subsections 6.1 and 6.2, it seems appropriate to view this model as one with a semiparametric response function, since the autocorrelation function helps to determine how a change in the observed variable x_t will affect future y_t.

It is easy to see that the semiparametric efficiency bound for this model is the asymptotic variance of the GLS estimator. The scores for the parameters of the autocorrelation function are orthogonal to the score for β, and the score for β at the true parameters is the first-order condition for the GLS estimator. In this context Hannan's (1963) contribution can be thought of as the construction of an adaptive estimator. Of course, Hannan's efficient estimator has efficiency properties that do not depend on the normality of the disturbance, although these efficiency properties do not pertain to a semiparametric model. They pertain to the Gauss–Markov class of estimators that are linear functions of the observations on the dependent variable.

Hansen (1988) derives the efficiency bound for a restricted, linear, Gaussian model. The model is:

$$\sum_{j=0}^{J} a_j(\beta_0) y_{t-j} + c(\beta_0) = \sum_{j=0}^{\tau-1} b_j \varepsilon_{t-j}, \tag{44}$$

where $\{y_t\}$ is an n-dimensional stationary, Gaussian process; $\{\varepsilon_t\}$ is an n-dimensional Gaussian white noise with an identity covariance matrix; $a_j(\beta)$ and b_j are $l \times n$ matrices; and $c(\beta)$ is an $l \times 1$ vector, with $l < n$. The parametric feature of this model is that this equation does not give the full dynamic specification of the y_t process; there are $n - l$ other linear equations, with unrestricted coefficients and possibly infinite lag-lengths, that make up the system.

Hansen (1988) shows that the bound for the β parameters is equal to the asymptotic covariance matrix of the optimal instrumental variables estimator of β. To discuss this result, define $e_t(\beta) = \sum_{j=0}^{f} a_j(\beta) y_{t-j} + c(\beta)$ and let J_t be the information set generated by $y_t, y_{t-1}, \ldots$. The model implies the conditional moment restriction:

$$E[e_t(\beta_0)|J_{t-\tau}] = 0.$$

This restriction means that past values of y_t, dated $t - \tau$ and before, qualify as instruments for the residual $e_t(\beta)$. Hansen (1985) and Hansen, Heaton, and Ogaki (1988) characterize the optimal form of these instruments. Hansen (1988) shows that the semiparametric bound for estimators of β is the covariance matrix of an optimal instrumental variables estimator.

ACKNOWLEDGEMENT

This paper was prepared by invitation for presentation at the 1988 European meeting of the Econometric Society. Helpful comments were provided by D. W. K. Andrews, R. Klein, R. Koenker, B. Meyer, A. Pagan, J. Powell, J. Robins, P. Robinson, and a referee. Financial support was provided by the NSF and the Sloan Foundation.

APPENDIXES

A: Proofs of Theorems

The conditions for smoothness and regularity of likelihoods are like those of Ibragimov and Hasminskii (1981, Ch. 7).

Definition A.1: $f(z|\theta)$ is *smooth* if (i) $\theta \in \Theta$, Θ open; (ii) there is a measure μ dominating $f(z|\theta)$ for $\theta \in \Theta$ such that $f(z|\theta)$ is continuous on Θ a.s. μ; (iii) $f(z|\theta)^{1/2}$ is μ-mean square continuously differentiable with respect to θ on Θ with derivative $D(z, \theta)$, i.e. for each $\theta \in \Theta$, $\int \|D(z, \theta)\|^2 d\mu$ is finite, and for every $\theta_i \to \theta$, $\int \|D(z, \theta_i) - D(z, \theta)\|^2 d\mu \to 0$, $\int [f(z|\theta_i)^{1/2} - f(z|\theta)^{1/2} - D(z, \theta)'(\theta_i - \theta)_2]^2 d\mu / \|\theta_i - \theta\|^2 \to 0$. For smooth $f(z|\theta)$ the score for θ is $S_\theta \equiv 21(f(z|\theta) > 0) D(z, \theta) / f(z|\theta)^{1/2}$ and the information matrix is $\int S_\theta S_\theta' f(z|\theta) d\mu$. The likelihood is *regular* if it is smooth and the information matrix is nonsingular on Θ.

Proof of Theorem 2.1: Follows from Hajek's (1970) representation theorem by the multivariate version of the characteristic function argument of the proof of Theorem 2(i) of Chamberlain (1986). ■

Proof of Theorem 2.2: Consider an LDGP with parameter θ_n and let $E_n[\cdot]$ denote the expectation taken at θ_n. Since LDGPs for regular parametric submodels are contiguous to the process with $\theta_n = \theta_0$, $\sqrt{n}(\hat{\beta} - \beta_0) = \sum_{i=1}^{n} \psi_i / \sqrt{n} + o_p(1)$, also holds under the LDGP. Then by addition of appropriate terms,

$$\sqrt{n}(\hat{\beta} - \beta_n) = \sum_{i=1}^{n} (\psi_i - E_n[\psi]) / \sqrt{n} + \sqrt{n}(\beta_0 - \beta_n) + \sqrt{n} E_n[\psi] + o_p(1), \tag{45}$$

where $\beta_n = \beta(\theta_n)$. Let $\ell(\theta)$ denote the likelihood for a single observation, where the z argument is suppressed for notational convenience, and let $\ell_n \equiv \ell(\theta_n)$ and $\ell_0 \equiv \ell(\theta_0)$. By regularity $\ell_n \xrightarrow{a.s.} \ell_0$, so that for $K_n \to \infty$, note that $1(\|\psi\| \geq K_n)\|\psi\|^2\ell_n \xrightarrow{a.s.} 0$. Also, $1(\|\psi\| \geq K_n)\|\psi\|^2\ell_n \leq \|^2\ell_n \xrightarrow{a.s.} \|\psi\|^2\ell_0$ and by the continuity hypothesis, $\int\|\psi\|^2\ell_n d\mu$ converges to $\int\|\psi\|^2\ell_0 d\mu$. Then by the dominated convergence theorem of Pitman (1979),

$$\int 1(\|\psi\| \geq K_n)\|\psi\|^2\ell_n d\mu \to 0. \tag{46}$$

It follows similarly that $\text{var}_n(\psi) = E_n[\psi\psi'] - E_n[\psi]E_n[\psi'] \to E[\psi\psi']$, so that by equation (45) the Lindbergh–Feller conditions are satisfied and

$$\sum_{i=1}^{n} (\psi_i - E_n[\psi])/\sqrt{n} \xrightarrow{d} N(0, E[\psi\psi']). \tag{47}$$

By continuity, $\int\|\psi\|^2\ell(\theta)d\mu$ is bounded on a neighbourhood of θ_0, so that by Lemma 7.2 of Ibragimov and Hasminskii (1981), $\int\psi\ell(\theta)d\mu$ is differentiable with derivative at θ_0 equal to $E[\psi S_\theta']$. Thus,

$$\sqrt{n}E_n[\psi] = \sqrt{n}\{E[\psi] + E[\psi S_\theta'](\theta_n - \theta_0) + o(\|\theta_n - \theta_0\|)\} = E[\psi S_\theta']\sqrt{n}(\theta_n - \theta_0) + o(1).$$

Also, by $\beta(\theta)$ differentiable,

$$\sqrt{n}(\beta - \beta_n) = \sqrt{n}\{[-\partial\beta(\theta_0)/\partial\theta](\theta_n - \theta_0) + o(\|\theta_n - \theta_0\|)\} = [-\partial\beta(\theta_0)/\partial\theta]\sqrt{n})\theta_n - \theta_0) + o(1). \tag{48}$$

Since $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, E[\psi\psi'])$ for $\theta_n = \theta_0$, it follows from (45)–(47) that the limiting distribution of $\sqrt{n}(\hat{\beta} - \beta(\theta_n))$ exists and does not depend on the sequence $\{\theta_n\}$ if and only if

$$\{\partial\beta(\theta_0)/\partial\theta - E[\psi S_\theta']\}\sqrt{n}(\theta_n - \theta_0) = o(1). \tag{49}$$

This equation holds for all sequences such that $\sqrt{n}(\theta_n - \theta_0)$ is bounded if and only if $\partial\beta(\theta_0)/\partial\beta - E[\psi S_\theta'] = 0$. ■

The following elementary, tedious Lemmas are useful in the following proofs.

Lemma A.1: $E[q'\delta] = 0\ \forall\delta \in \mathcal{S}$ *if and only if* $E[q\delta'] = 0 \forall\delta \in \mathcal{S}$ *if and only if* $E[qS_\theta'] = 0$ *for all scores* S_θ *for smooth parametric submodels.*

Proof: Note that $\{F\delta : \delta \in \mathcal{S},\ F \text{ is } q \times q \text{ constant}\} \subseteq \mathcal{S}$, since $E[\|\delta - A_jS_{\theta j}\|^2] \to 0$ implies $E[\|F\delta - (FA_j)S_{\theta j}\|^2] \leq \|F\|^2E[\|\delta - A_jS_{\theta j}\|^2] \to 0$. In particular if $\delta \in \mathcal{S}$, then $e_ie_j'\delta \in \mathcal{S}$, $(i, j = 1, \ldots, q)$, where e_i denotes the ith unit vector. Then since $E[q'(e_ie_j'\delta)] = E[q_i\delta_j] = 0$, where q_i and δ_j denote the ith and jth components respectively, it follows by i and j arbitrary that $E[q\delta'] = 0$. Also, note that $FS_\theta \in \mathcal{S}$ for any constant matrix F with q rows, so that if $E[q\delta'] = 0$ for all $\delta \in \mathcal{S}$, $0 = E[q(FS_\theta)'] = E[qS_\theta']F'$ for all constant matrices F, implying $E[qS_\theta'] = 0$. Finally, note that if $E[qS_\theta'] = 0$ for all scores, then for any $\delta \in \mathcal{S}$ and $A_jS_{\theta j}$ converging in mean square to δ, $E[q'\delta] = \lim_{j\to\infty}E[q'A_jS_{\theta j}] = \lim_{j\to\infty}\text{trace}(E[qS_{\theta' j}]A_j) = 0$. Thus, the first statement implies the second, the second the third, and the third the first, and all the statements are equivalent, as claimed. ■

Lemma A.2: The scores in the definition of the tangent set can be restricted to be those for regular parametric submodels (i.e. smooth submodels with a nonsingular information matrix) without shrinking the tangent set.

Proof: Consider $\delta \in \mathcal{S}$ and a sequence of scores $\{S_{\theta j}\}$ for smooth parametric submodels such that $E[\|\delta - A_jS_{\theta j}\|^2] \to 0$ for a conformable sequence $\{A_j\}$ of constant matrices. For each j

there exists a subvector $S_{\theta j}^1$ of $S_{\theta j}$ such that $S_{\theta j} = FS_{\theta j}^1$ for a constant matrix F, and $E[S_{\theta j}^1 S_{\theta' j}^1]$ is nonsingular. To see this, let H be a basis for the null space of $E[S_{\theta j} S_{\theta' j}]$, and order the elements of $S_{\theta j}$ and H so that $H = [H_1', H_2']'$ with H_2 nonsingular. Then for $S_{\theta j} = (S_{\theta' j}^1, S_{\theta' j}^2)'$ partitioned conformably, pre and post multiplying $E[S_{\theta j} S_{\theta' j}]$ by H gives $E[\|H_1' S_{\theta j}^1 + H_2' S_{\theta j}^2\|^2] = \text{trace}\ E[(H_1' S_{\theta j}^1 + H_2' S_{\theta J}^2)(H_1' S_{\theta j}^1 + H_2' S_{\theta j}^2)'] = 0$, implying $H_1' S_{\theta j}^1 + H_2' S_{\theta j}^2 = 0$ (with probability one). Solving for $S_{\theta j}^2$ then gives $S_{\theta j}^2 = (H_2')^{-1} H_1' S_{\theta j}^1$, so that $S_{\theta j} = [I, -H_1 (H_2)^{-1}]' S_{\theta j}^1$. Also, $E[S_{\theta j}^1 S_{\theta' j}^1]$ is nonsingular, as otherwise there would exist a vector $(a, 0)'$ with $a \neq 0$ in the nullspace of $E[S_{\theta j} S_{\theta' j}]$, which is not spanned by H because of the nonsingularity H_2. Consider a parametric submodel corresponding to $f(a|\theta_1, \theta_{20})$, for θ_1 in $\Theta_j^1 \equiv \{\theta_1 : (\theta_1', \theta_{20}')' \in \Theta_j\}$, which is open by openness of the parameter set Θ_j for the jth smooth parametric submodel. This submodel is smooth. By smoothness and $E[S_{\theta j}^1 S_{\theta' j}^1]$ nonsingular, it has a nonsingular information matrix for all θ in a small enough open subset of Θ_j^1. Choosing this smaller subset to be the parameter set makes $f(z|\theta_1, \theta_{20})$ regular. Finally, note that $A_j S_{\theta j} = (A_j F) S_{\theta j}^1 = A_j^1 S_{\theta j}^1$ for $A_j^1 = A_j F$, so that $E[\|a - A_j^1 S_{\theta j}^1\|^2] = E[\|a - A_j S_{\theta j}\|^2] \to 0$. The conclusion then follows by a arbitrary. ■

Proof of Theorem 3.1: Consider any regular parametric submodel with score S_θ. By Lemma A.1 and $d - \delta$ orthogonal to the tangent set, $E[\delta S_\theta'] - E[dS_\theta'] = E[(\delta - d)S_\theta'] = 0$, so that δ satisfies equation (3.1). Then, as discussed in the text, the Cramer–Rao bound V_θ for the parametric submodel satisfies

$$V_\theta = E[\delta S_\theta'](E[S_\theta S_\theta'])^{-1} E[S_\theta \delta'] \leq E[\delta\delta'], \tag{50}$$

in the positive definite sense, implying $E[\delta\delta']$ is an upper bound. By $\delta \in \mathcal{S}$ and Lemma A.2 it is possible to find a sequence of scores $S_{\theta j}$ for regular parametric submodels and constant matrices such that $E[\|\delta - A_j S_{\theta j}\|^2] \to 0$. Let $V_{\theta j} = E[d_{\theta j} d'_{\theta j}]$ be the Cramer–Rao bound for the jth parametric submodel, where $d_{\theta j} = E[\delta S'_{\theta j}](E[S_{\theta j} S'_{\theta j}])^{-1} S_{\theta j}$. Since $d_{\theta j}$ are least-squares coefficients, it follows that:

$$E[\|\delta - d_{\theta j}\|^2] \leq E[\|\delta - A_j S_{\theta j}\|^2] \to 0. \tag{51}$$

Since each element of $d_{\theta j}$ converges in mean-square to the corresponding element of δ, the second moment matrix $V_{\theta j} = E[d_{\theta j} d_{\theta' j}]$ converges to the second moment matrix $E[\delta\delta']$ of δ. ■

Proof of Theorem 3.2: By $f(z|\beta)$ smooth, S_β and S are well defined. By nonsingularity of $E[SS']$, $\delta \equiv (E[SS'])^{-1} S$ is well defined. Note that $\beta(\theta) = (\beta', \eta')'$, so that by construction $\beta(\theta)$ is differentiable with derivative $\partial\beta(\theta)/\partial\theta = [I_q, 0]$. Furthermore, by Lemma A.1 and orthogonality of S and $\mathcal{T}$,

$$E[\delta S_\beta'] = (E[SS'])^{-1} E[S(S_\beta - \bar{\imath})'] = I_q, \quad E[\delta S_\eta'] = (E[SS'])^{-1} E[SS_\eta'] = 0, \tag{52}$$

where $\bar{\imath}$ denotes the projection of S_β on $\mathcal{T}$, so that δ satisfies equation (3.1). Now, by Theorem 3.1, to show that δ is the efficient influence function, it suffices to show that $\delta \in \mathcal{S}$, since then the projection of δ on $\mathcal{S}$ is just δ. By definition, $\mathcal{S}$ is the set of a such that there exists a sequence of smooth parametric submodels with scores $S_{\theta j} = (S_\beta', S_{\eta' j})'$ and of constant matrices $A_j = [A_j^1, B_j]$ such that:

$$0 = \lim_{j \to \infty} E[\|a - A_j S_{\theta j}\|^2] = \lim_{j \to \infty} E[\|a - (A_j^1 S_\beta + B_j S_{\eta j})\|^2]. \tag{53}$$

Note that $FS = FS_\beta - F\bar{\imath} \in \mathcal{S}$ for any square constant matrix F, since for $B_j S_{\eta j}$ converging in mean square to $\bar{\imath}$, $FS_\beta + (-FB_j) S_{\eta j}$ converges in mean square to FS. In particular, $\delta = (E[SS'])^{-1} S \in \mathcal{S}$. The final conclusion then follows by $E[\delta\delta'] = (E[SS'])^{-1}$. ■

Proof of Theorem 3.3: Adaptive estimability of β means that there is a regular estimator such that $\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0,(E[UU'])^{-1}$, for $U \equiv S_\beta - E[S_\beta S_\alpha'](E[S_\alpha S_\alpha'])^{-1}S_\alpha$. Then by Theorem 2.1, there cannot exist any regular parametric submodel $f(z|\beta,\alpha,\eta)$ such that $V_\theta^\beta - (E[UU'])^{-1}$ is nonzero and positive semidefinite, where V_θ^β is the Cramer–Rao bound for β in the submodel. Equivalently, by inversion of positive semidefinite ranking for inverse and the partitioned inverse formula, there cannot exist a regular parametric submodel such that for $S_\xi = (S_\alpha', S_\eta')'$ and $U_\xi = S_\beta - E[S_\beta S_\xi](E[S_\xi S_\xi'])^{-}S_\xi$,

$$E[UU'] - E[U_\xi U_\xi'] = E[(U-U_\xi)(U-U_\xi)']$$

nonzero and positive semi-definite, where the equality follows by orthogonality of least-squares residuals (e.g. $E[(U_\xi - U)U_\xi']$). But this implies that $U = U_\xi$, so that $E[US_\eta'] = 0$ follows by orthogonality of least-squares residuals. Since this must hold for all regular submodels, it follows by an argument analogous to that for Lemma A.2 that $E[US_\eta'] = 0$ for all smooth parametric submodels, so the conclusion follows by Lemma A.1. ∎

Proof of Lemma 3.4: First, note that $E[U'U|T]$ is constant, so that $E[\|D(V)U\|^2] \leqslant E[U'U\|D(V)\|^2] = E[E[U'U|T]\|D(V)\|^2] = CE[\|D(V)\|^2]$ and $E[\|D(V)\|^2] = E[\mathrm{tr}\{D(V)D(V)'\}] \leqslant CE[\mathrm{tr}\{D(V)E[UU'|T]D(V)'\}] =$ $CE[\mathrm{tr}\{E[D(V)UU'D(V)'|T]\}] = CE[E[\mathrm{tr}\{D(V)UU'D(V)'\}|T]] =$ $CE[E[\|D(V)U\|^2|T]] = CE[\|D(V)U\|^2]$, so that $D(V)U$ is an element of $\mathscr{T}_V$ if and only if $E[\|D(V)U\|^2] < \infty$. Next, $E[W|V]$ has finite second moment and $E[W|V]U = D(V)U$ for $D(V) = E[W|V]$, so that $E[W|V]U \in \mathscr{T}_V$. Also $E[(WU - E[W|V]U)'(D(V)U)]$ $= trE[D(V)UU'(W - E[W|V])']$ $= \mathrm{tr}E[D(V)UU'(W-E[W|V])'] = \mathrm{tr}E[D(V)E[UU'|T](W-E[W|V])']$ $= \mathrm{tr}E[\tilde{D}(V)(W-E[W|V])'] = 0$, for $\tilde{D}(V) = D(V)E[UU'|T]$. ∎

Proof of Theorem 4.1: Consider any $S_\eta = \sigma(v)^{-2}[y-F(v)]D(v)$. Suppose that $E[mS_\eta'] = 0$. Then:

$$\begin{aligned} 0 &= E[mS_\eta'] = E[E[mS_\eta'|v]] \\ &= E[E[m(z)\{y-F(v)\}|v]\sigma^{-4}(v)D(v)] \\ &= E[E[m(z)\{y-F(v)\}|v]A(v)], \end{aligned}$$

where $A(v) = \sigma(v)^{-4}D(v)$. Since by varying the parametric submodel, $A(v)$ can be essentially any function of v, it follows that $E[m(z)\{y-F(v)\}|v] = 0$. Then:

$$\begin{aligned} 0 &= E[m(z)\{y-F(v)\}|v] = E[E[m(z)\{y-F(v)\}|x]|v] \\ &= E[m(1,x)\{1-F(v)\}F(v) + m(0,x)\{-F(v)\}\{1-F(v)\}|v] \\ &= E[m(1,x)-m(0,x)|v]\{1-F(v)\}F(v). \end{aligned}$$

Then by choosing $F(v)$ to be such that $0 < F(v) < 1$, we obtain

$$E[m(1,x)-m(0,x)|v] = 0.$$

Now, if $m(1,x)-m(0,x) \neq 0$, then one can choose the conditional distribution of x given v in such a way that this equality is violated, e.g. by putting positive probability weight on a value of x where it is nonzero. Therefore, this equality implies $m(1,x) = m(0,x)$, i.e. that $m(z)$ depends only on x. But then $E[mS_\beta'] = E[mE[S_\beta'|x]] = 0$, violating nonsingularity of $D = -E[mS_\beta']$. It follows that there exists no $m(z)$ that depends only on z and satisfies equation (4.5), implying that one of the stated hypotheses must be false. ∎

Lemma A.3: If Assumptions 4.1 and 4.2 are satisfied then $\hat{\beta}$ is asymptotically linear with influence function $-M^{-1}(m+\tau)$.

Proof: By equation (4.6), Assumption 4.1(ii), and the mean value theorem,

$$\sqrt{n}\hat{m}_n(\beta_0) + \bar{M}\sqrt{n}(\hat{\beta} - \beta_0) = o_p(1), \tag{54}$$

where $\bar{M} = \partial \hat{m}_n(\bar{\beta})/\partial\beta$ and $\bar{\beta}$ is the mean value, which actually differs from row to row of $\bar{M}$. By $\|\bar{\beta} - \beta_0\| \leq \|\hat{\beta} - \beta_0\|$ and Assumption 4.1(i), $\text{plim}(\bar{\beta}) = \beta_0$, so that by 4.1(ii), $\text{plim}\bar{M} = M$. Then by M nonsingular, $\bar{M}$ is nonsingular with probability approaching one, and $\text{plim}\bar{M}^{-1} = M^{-1}$. Also, by 4.2,

$$\sqrt{n}\hat{m}_n(\beta_0) = \sqrt{n}[\hat{m}_n(\beta_0) - m_n - \bar{m}(\hat{\alpha})] + \sqrt{n}[m_n + \bar{m}(\hat{\alpha})] = \sum_{i=1}^{n} (m_i + \tau_i)/\sqrt{n} + o_p(1) \tag{55}$$

so by the central limit theorem, $\sqrt{n}\hat{m}_n(\beta_0)$ is bounded in probability. Therefore, by equation (55), solving equation (54) for $\sqrt{n}(\hat{\beta} - \beta_0)$ gives

$$\sqrt{n}(\hat{\beta} - \beta_0) = -\bar{M}^{-1}\sqrt{n}\hat{m}_n(\beta_0) + o_p(1) = -M^{-1}\sqrt{n}\hat{m}_n(\beta_0) + o_p(1) \tag{56}$$

$$= \sum_{i=1}^{n} [-M^{-1}(m_i + \tau_i)/\sqrt{n} + o_p(1). \quad \blacksquare$$

Proof of Theorem 4.2: By Lemma A.3 and Theorem 4.2, $E[(m+\tau)S_\eta'] = 0$ for all scores for regular parametric submodels, so by Lemmas A.1 and A.2, $m + \tau$ is orthgonal to $\mathcal{T}$. Therefore if $\tau = 0$, m is orthogonal to $\mathcal{T}$. The second conclusion follows by Theorem 4.3, since if m is orthogonal to τ, $\text{Proj}(m|\mathcal{T}) = 0$.

Proof of Theorem 4.3: Note that $-\tau \in \mathcal{T}$ by $\tau \in \mathcal{T}$ and the definition of $\mathcal{T}$. Then, as in the proof of Theorem 4.2, $m - (-\tau)$ and $\mathcal{T}$ are orthogonal, so that $-\tau = \text{Proj}(m|\mathcal{T})$ follows by equation (3.6). The remainder of the conclusion follows by Theorem 2.2, which implies $M = -E[(m+\tau)S_\beta'] = -E[(m - \text{Proj}(m|\mathcal{T}))S_\beta'] = -E[(m - \text{Proj}(m|\mathcal{T}))S']$.

Proof of Proposition 4.4: Note that $\bar{m}(\hat{\alpha})$ is a function of only $(w_1, \ldots, w_n)$ by the hypothesis on $\hat{\alpha}$ and the definition of $\bar{m}(\alpha)$. Therefore, $\tau = \tau(w)$ is a function of only w. By Assumption 4.2, $\tau \in \mathcal{T}_w = \{a(w) : E[a'a] < \infty, E[a] = 0\}$. Next, consider parametric submodels for the marginal distribution of w. The scores for such parametric submodels are elements of $\mathcal{S}$. Furthermore, since the distribution of w is unrestricted, except possibly for regularity conditions, such scores are dense in $\mathcal{S}_w$; e.g. see section 3. It follows that $\mathcal{S}_w \subset \mathcal{T}$, so that $\tau \in \mathcal{T}$. Theorem 4.3, then gives the conclusion. $\blacksquare$

B. Tangent Set Verification for Mean Example

Consider parametric submodels of the form $f(z|\theta) = f_0(z)\{1 + \theta[q(z) - \bar{q}]\}$, where $f_0(z)$ is the true density, $q(z)$ is bounded, and $\bar{q} = E[q(z)]$. For all θ close enough to $\theta_0 = 0$, $f(z|\theta) \geq 0$, so that this will be a density function by the definition of $\bar{q}$. Also, it is easy to check that the likelihood is smooth with score $q(z) - \bar{q}$; e.g. see Newey (1990, Lemma C.4). Furthermore, by $q(z)$ bounded, there is a constant C such that $E_\theta[z^2] \leq CE[z^2]$ for all θ small enough, so that $f(z|\theta)$ satisfies the conditions for differentiability of the mean. Therefore, the family of scores $\{q(z) - \bar{q} : q(z) \text{ bounded}\}$ must be a subset of $\mathcal{S}$. Denseness of this set in $\{a(z) : E[a] = 0\}$ follows from the fact that bounded functions can approximate any function with finite mean square arbitrarily close (choose $q(z)$ close to $a(z)$; then $\bar{q}$ will also be close to $E[a] = 0$).

C. The Normal Equations Approach to Calculating the Bound

Following Begun *et al.* (1983) and BKRW, suppose that

$$\mathcal{T} = \{T(q) : q \in Q\}, \tag{57}$$

where Q is a set of random vectors, with inner product $E_Q[q_1 q_2]$, where $E_Q[\cdot]$ denotes the expectation for some probability measure, and $T(\cdot)$ is a continuous linear mapping. Let T^a denote the adjoint mapping from the set $\{r(z) : E[\|r(z)\|^2] < \infty\}$ to Q, satisfying $E[r(z)'T(q)] = E_Q[T^a(r)'q]$, and consider the composed function $T^aT(q) = T^a(T(q))$. The normal equations for the projection $\bar{\imath} = T(\bar{q})$ of S_β on $\mathcal{T}$ are:

$$T^aT(\bar{q}) = T^a(S_\beta). \tag{58}$$

See Luenberger (1969, p. 160).

If an inverse to T^aT can be found, then this equation can be solved to find $\bar{q}$, and the efficient score calculated as $S = S_\beta - \bar{\imath} = S_\beta - T(\bar{q})$. Let $(T^aT)^-$ denote a generalized inverse defined as an operator such that $T^aT(T^aT)^-T^aT(q) = T^aT(q)$, where the product notation denotes composition. Such an operator exists here because equation (57) and the definition of $\mathcal{T}$ imply that T, and hence T^aT has closed range: see Luenberger (1969, section 6.11). Then by $T^aT(T^aT)^-T^a(r) = T^a(r)$ (again, see Luenberger), a solution to equation (58) is $\bar{q} = (T^aT)^-T^a(S_\beta)$, and the efficient score is:

$$S = S_\beta - \bar{\imath} = S_\beta - T(T^aT)^-T^a(S_\beta). \tag{59}$$

The condition that $T(q)$ has closed range, which is implicit in equation (57) and the definition of the tangent set, is restrictive. However, even when this condition does not hold, or is difficult to check, a heuristic calculation using equation (59) can lead to a candidate for the efficient score. The candidate can then be verified by showing that equation (12) holds. For example, Cosslett (1987) essentially follows this procedure.

Calculating the efficient score via equation (59) involves choosing Q in such a way that both the adjoint and inverse can be computed. Some choices of Q, such as $Q = \mathcal{T}$, may not be useful. A choice of Q that is fruitful in the models of equation (29) is to choose Q to be the scores for the distrubance ε. The adjoint is easy to compute in this case, but the inverse can be quite difficult. For censored regression, Cosslett's (1987) calculation involves the general solution of a second-order linear differential equation with noncostant coefficients. It also appears to be difficult to calculate the inverse for other special cases of equation (29). For the case where λ is an unknown strictly monotonic transformation, and ε is restricted to be normal, BKRW find that the inverse corresponds to the solution of a differential equation that is not known to have a closed form.

REFERENCES

Abrazamar, A., and P. Schmidt (1981). 'Further evidence of the robustness of the Tobit estimator to heteroscedasticity', *Journal of Econometrics*, **17**, 253–258.

Abrazamar, A., and P. Schmidt (1982). 'An investigation of the robustness of the Tobit estimator to non-normality', *Econometrica*, **50**, 1055–1063.

Andrews, D. W. K. (1989), 'Asymptotics for semiparametric econometric models: I. Estimation', Working Paper, Cowles Foundation, Yale University.

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. (1982). 'On adaptive estimation', *Annals of Statistics*, **10**, 647–671.
Bickel, P., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1990). 'Efficient and adaptive interence in semiparametric models', monograph, in preparation.
Buckle, J., and I. James (1979). 'Linear regression with censored data', *Biometrika*, **66**, 429–436.
Carroll, R. J. (1982). 'Adapting for heteroskeadasticity in linear models', *Annals of Statistics*, **10**, 1224–1233.
Cavanagh, C. (1987). 'Asymptotic distribution theory for the maximum score estimator', presented at the North American summer meeting of the econometric Society.
Chamberlain, G. (1986). 'Asymptotic efficiency in semiparametric models with censoring', *Journal of Econometrics*, **32**, 189–218.
Chamberlain, G. (1987a), 'Asymptotic efficiency in estimation with conditional moment restrictions', *Journal of Economics*, **34**, 305–334.
Chamberlain, G. (1987b). 'Efficiency bounds for semiparametric regression', manuscript, Department of Economics, University of Wisconsin.
Cosslett, S. (1981a). 'Maximum likelihood estimation for choice-based samples', *Econometrica*, **49**, 1289–1316.
Cosslett, S. (1981b). 'Efficient estimation of discrete choice models', *in* C. Manksi and D. McFadden (eds) *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge, Mass.
Cosslett, S. R. (1983). 'Distribution-free maximum likelihood estimator of the binary choice model', *Econometrica*, **51**, 765–782.
Cosslett, S. R. (1985). 'Efficiency bounds for distribution-free estimators from endogenously stratified samples', mimeo, presented at the 1985 World Congress of the Econometric Society.
Cosslett, S. R. (1987). 'Efficiency bounds for distribution-free estimators of the binary choice and censored regression models', *Econometrica*, **55**, 559–585.
Cosslett, S. R. (1988). 'Distribution-free estimator for a regression model with sample selectivity', mimeo, Ohio State University.
Cox, D. R. (1975). 'Partial likelihood', *Biometrika*, **62**, 269–276.
Duncan, G. M. (1986). 'A semiparametric censored regression estimator', *Journal of Econometrics*, **32**, 5–34.
Engle, R. F., C. W. J. Granger, J. Rice, and A. Weiss (1986). 'Semiparametric estimates of the relation between weather and electricity sales'. *Journal of the American Statistical Association*, **81**, 310–320.
Friedman, J. H., and W. Stuetzle (1981). 'Projection pursuit regression', *Journal of the American Statistical Association*, **76**, 817–823.
Gallant, A. R., and D. W. Nychka (1987). 'Semi-nonparametric maximum likelihood estimation', *Econometrica*, **55**, 363–390.
Goldberger, A. S. (1983). 'Abnormal selection bias', in S. Darlin, *et al.* (eds), *Studies in Econometrics, Time Series, and Multivariate Statistics*, Academic Press, New York.
Gourieroux, C., A. Monfort, and E. Renault (1987). 'Consistent *m*-estimators in a semiparametric model', Working paper 8706, INSEE.
Grenander, U. (1981). *Abstract Inference*, Wiley, New York.
Hajek, J. (1970). 'A characterization of limiting distributions of regular estimates', *Z. Wahrscheinlichkeitstheorie verw. Geb.*, **14**, 323–330.
Hajek, J. (1972). 'Local asymptotic minimax and admissibility in estimation', in *proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, Vol. 1, pp. 175–194.
Hampel, F. R. (1974). 'The influence function and its role in robust estimation', *Journal of the American Statistical Association*, **62**, 1179–1186.
Hannan, E. J. (1963), 'Regression for time series', in M. Rosenblatt, (ed.), *Time Series Analysis*, John Wiley, New York.
Hansen, L. P. (1985). 'A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators', *Journal of Econometrics*, **30**, 203–238.
Hansen, L. P. (1988). 'Semiparametric efficiency bounds for linear time series models', paper presented at the Econometric Society North American 1988 Summer Meeting.
Hansen, L. P., J. C. Heaton, and M. Ogaki (1988). 'Efficiency bounds implied by multiperiod conditional moment restrictions', *Journal of the American Statistical Association*, **83**, 863–871.

Hausman, J. A., and D. A. Wise (1981). 'Stratification on endogenous variables and estimation, the Gary income maintenance experiment', in C. Masnki and D. McFadden (eds), *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge, Mass.

Heckman, J. J., and B. Singer (1984). 'A method for minimizing the impact of distributional assumptions in econometric models for duration data', *Econometrica*, **52**, 271–320.

Hsieh, D., and C. Manski (1987). 'Monte-Carlo evidence on adaptive maximum likelihood estimation of a regression', *Annals of Statistics*, **15**, 541–551.

Huber, P. (1967), 'The behavior of maximum likelihood estimates under nonstandard conditions', in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, vol. 1.

Hurd, M. (1979). 'Estimation in truncated samples when there is heteroskedasticity', *Journal of Econometrics*, **11**, 247–258.

Ibragimov, I. A., and R. Z. Hasminskii (1981). *Statistical Estimation: Asymptotic Theory*, Springer-Verlag, New York.

Ichimura, H. (1986). 'Estimation of index model coefficients', Ph.D. thesis, Department of Economics, MIT.

Jewell, N. P. (1985). 'Least squares regression with data arising from stratified samples of the dependent variable', *Biometrika*, **72**, 11–21.

Kim, J., and D. Pollard (1989). 'Cube root asymptotics', manuscript, Statistics Department, Yale University.

Klaasen (1987). 'Consistent estimation of the influence function of locally asymptotically linear estimators', *Annals of Statistics*, **15**, 1548–1562.

Klein, R. W., and R. S. Spady (1987). 'An efficient semiparametric estimator of the binary response model', manuscript, Bell Communications Research.

Koshevnik, Y. A., and Levit, B. Y. (1976). 'On a non-parametric analogue of the information matrix', *Theory of Probability and Applications*, **21**, 738–753.

LeCam, L. (1953). 'on some asymptotic properties of maximum likelihood estimates and related bayes' estimates', *University of California Publications in Statistics*, **1**, 277–329.

Levit, B. Y. (1975). 'On the efficiency of a class of nonparametric estimates', *Theory of Probability and its applications*, **20**, 723–740.

Luenberger, D. G. (1969). *Optimization by Vector Space Methods*, John Wiley and Sons, New York.

Manski, C. (1975). 'Maximum score estimation of the stochastic utility model of choice', *Journal of Econometrics*, **3**, 205–228.

Manski, C. (1984). 'Adaptive estimation of nonlinear regression models', *Econometric Reviews*, **3**, 145–194.

Manski, C., and D. McFadden (1981). 'Alternative estimators and sample designs for discrete choice analysis', in C. Manski and D. McFadden (eds), *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge, Mass.

Newey, W. K. (1984). 'A method of moments interpretation of sequential estimators', *Economics Letters*, **14**, 201–206.

Newey, W. K. (1987). 'Efficient estimation of models with conditional moment restrictions', manuscript, Princeton University.

Newey, W. K. (1988a). 'Adaptive estimation of regression models via moment restrictions', *Journal of Econometrics*, **38**, 301–339.

Newey, W. K. (1988b). 'Efficient estimation of semiparametric models via moment restrictions', manuscript, Department of Economics, Princeton University.

Newey, W. K. (1988c). 'Two-step series estimation of sample selection models', manuscript, Department of Economics, Princeton University.

Newey, W. K. (1989). 'Efficiency in univariate limited dependent variable models under conditional moment restrictions', manuscript, department of economics, Princeton University.

Newey, W. K. (1990). 'Efficient estimation of Tobit models under symmetry', forthcoming in W. Barnett, J. Powell, and G. Tauchen, (eds), *Semiparametric and Nonparametric Methods in Econometrics and Statistics*, Cambridge University Press, Cambridge.

Newey, W. K., and J. L. Powell (1989). 'Efficient estimation of type I censored regression models under conditional quantile restrictions', manuscript, Department of Economics, Princeton University.

Newey, W. K., and T. Stoker (1989). 'Efficiency of weighted average derivatives', working paper, Sloan School of Management, MIT.

Pagan, A. (1986). 'Two-stage and related estimators and their applications', *Review of Economic Studies*, **53**, 517–538.
Pfanzagl, J. and W. Wefelmeyer (1982). *Contributions to a General Asymptotic Statistical Theory*, Springer-Verlag, New York.
Pierce, D. A. (1982). 'The asymptotic effect of substituting estimators for parameters in certain types of statistics', *Annals of Statistics*, **10**, 475–478.
Pitman, E. J. G. (1979). *Some Basic Theory for Statistical Inference*, Chapman and Hall, London.
Pollard, D. (1985). 'New ways to prove central limit theorems', *Econometric Theory*, **1**, 295–314.
Portnoy, S., and R. Koenker (1989)). 'Adaptive L-estimation of linear models', *Annals of Statistics*, **17**, 362–381.
Powell, J. L. (1984). 'Least absolute deviations estimation for the censored regression model', *Journal of Econometrics*, **25**, 303–325.
Powell, J. L. (1986). 'Symmetrically trimmed least squares estimation for Tobit models', *Econometrica*, **54**, 1435–1460.
Powell, J. L. (1987). 'Semiparametric estimation of bivariate limited dependent variable models', manuscript, University of Wisconsin.
Powell, J. L., J. H. Stock, and T. M. Stoker (1990). 'Semiparametric estimation of index coefficients', *Econometrica*, forthcoming.
Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, John Wiley and Sons, New York.
Ritov, Y. (1984). 'Efficient and unbiased estimation in nonparametric regression with censored data', manuscript, University of California, Berkeley.
Ritov, Y. (1977). 'Estimation in a linear regression model with censored data', Technical Report no. 114, Department of Statistics, University of California, Berkley.
Ritov, Y., and P. J. Bickel (1987). 'Achieving information bounds in non and semiparametric models', Technical Report no. 116, Department of Statistics, University of California, Berkeley.
Ritov, Y., and J. A. Wellner (1988). 'Censoring, martingales, and the Cox model', *Contemporary Mathematics*, **80**, 191–219.
Robinson, P. (1987). 'Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form', *Econometrica*, **55**, 875–891.
Robinson, P. (1988a). 'Semiparametric econometrics: a survey', *Journal of Applied Econometrics*, **3**, 35–51.
Robinson, P. (1988b). 'Root-N-consistent semiparametric regression', *Econometrica*, **56**, 931–954.
Ruud, P. A. (1983). 'Sufficient conditions for the consistency of maximum likelihood despite misspecification of distribution', *Econometrica*, **51**, 225–228.
Ruud, P. A. (1986). 'Consistent estimation of limited dependent variable models despite misspecification of distribution', *Journal of Econometrics*, **32**, 157–187.
Schick, A. (1986), 'On asymptotically efficient estimation in semiparametric models', *Annals of Statistics*, **14**, 1139–1151.
Severini, T. A., and W. H. Wong (1987). 'Profile likelihood and semiparametric models', manuscript, University of Chicago.
Stein, C. (1956). 'Efficient nonparametric testing and estimation', in *Proceedings of the third Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, vol. 1.
Stoker, T. M. (1986). 'Consistent estimation of scaled coefficients', *Econometrica*, **54**, 1461–1481.
Van der Vaart, A. (1988). 'On differentiable functionals', Working Paper, Department of Statistics, University of Washington.
Wellner, J. A. (1985). 'Semiparametric models: progress and problems', *Bulletin of the I.S.I., Proceedings of the 45th Session: Invited Papers*, Book 4, 23.1.