EFFICIENT ESTIMATION OF SEMIPARAMETRIC MODELS VIA MOMENT RESTRICTIONS

Whitney K. Newey
Princeton University and Bellcore

Econometric Research Program
Research Memorandum No. 352

March 1990

Econometric Research Program
Princeton University
204 Fisher Hall
Princeton, NJ 08544-1021, USA

Abstract

A semiparametric model often implies an infinite variety of restrictions that one could use to estimate parameters of interest. The purpose of this paper is to study efficient semiparametric estimation via linear combinations of moment restrictions. The motivation is the parsimonious and flexible form of these estimators. The estimators are constructed from a linear combination of moment restrictions that is chosen to approximate the efficient score. The necessary spanning condition for efficiency of these estimators is discussed, and regularity conditions for asymptotic efficiency are given. Throughout the paper a number of examples are considered, including nonlinear simultaneous equations models (which include some transformation models) and a conditional distribution regression model. The paper also gives a small Monte Carlo study concerning efficient estimation of a slope parameter in linear regression, that shows that the estimators can perform well relative to kernel type estimators and that the cross-validation suggestion is promising.

1. Introduction

A semiparametric model often implies an infinite variety of restrictions that one could use to estimate parameters of interest. The semiparametric efficiency bound gives the smallest asymptotic variance one might hope to attain by using such restrictions. The purpose of this paper is to study efficient semiparametric estimation via linear combinations of semiparametric moment restrictions. The estimators considered here are m-estimators formed by combining averages that converge to zero at the true parameters.

This study is motivated by the parsimonious, but flexible, form of these estimators. They allow the statistician to choose the type of restrictions used in estimation. For instance, one might use restrictions corresponding to those that are best (in the semiparametric efficiency bound sense) for particular parametric families of distributions. Alternatively, one might select from low order terms in an approximating family of restrictions, such as power series, with the goal of capturing most of the information. The statistician also has the freedom to choose only a few restrictions, or to allow the data to guide this choice, for example by a cross-validation method discussed below.

After some preliminary discussion of semiparametric efficiency bounds and m-estimators, the paper presents a method of combining different moment restrictions. The linear combination coefficients are chosen so as to provide a best mean-square approximation to the efficient score (which appears in the bound). In some cases these linear combination coefficients minimize the asymptotic variance of the estimator, although in general they will only guarantee that the variance is close to the efficiency bound if a sufficiently large number of a sufficiently rich set of moment restrictions is used.

The necessary *spanning condition* for efficient estimation by this method (i.e. the meaning of "sufficiently rich") is that a linear combination of a

sufficient number of influence functions corresponding to the moment restrictions can approximate the efficient score arbitrarily well in mean-square. As discussed below, this condition is related to the characterization of the model by the moment restrictions. Also, an easily checked sufficient condition is given. This spanning condition is also of interest when the goal is to find estimators with good efficiency that combine some of the restrictions, rather than construct an efficient estimator. The estimator may be ignoring important sources of information if the moment restrictions are not selected from a set that satisfies the spanning condition.

The paper formulates a set of sufficient conditions for growth of the number of restrictions with the sample size to achieve asymptotic efficiency. They include a bound on the smallest eigenvalue of a second moment matrix and convergence rates for sample moment matrices. The conditions allow for the number of moment restrictions to be chosen as a function of the data, as in the cross-validation method discussed below. Because of their generality, the conditions are not very primitive, although it is shown in an example that they can be verified in a straightforward way.

Approximating the efficient score by a linear combination of functions is really a special case of a general efficient estimation method that employs a nonparametric estimator of the efficient score; see Stone (1974), Bickel (1982), Schick (1986), Severini and Wong (1987). However, this method of estimating the score differs from that analyzed in most of the literature. Most of the literature has concentrated on kernel methods for score estimation, while the method in this paper is essentially a truncated series approximation. Indeed, the motivation for this paper given above is really just a list of characteristics of truncated series approximations, which are not shared by some other nonparametric estimation methods.

This paper includes a small Monte Carlo study consisting of a subset

of the experiments in Hsieh and Manski (1987), which concern adaptive estimation of the linear regression model with disturbance that is independent of the regressors. It is found that the estimators perform very well in comparison with kernel estimators, and that the particular cross-validation method discussed here is promising.

M-estimation based on linear combinations of moment restrictions has previously been considered by others, including Beran (1976), Hansen (1982), Chamberlain (1982), MaCurdy (1982), and Cragg (1983). The estimator considered here is most closely related to that of Beran (1976) in the linear model case he considered, although the one here applies to any semiparametric model. Also, Hayashi and Sims (1982) and Chamberlain (1987a) have previously formulated spanning conditions in the context of semiparametric models of conditional moment restrictions, although the condition here is somewhat different. For linear regression models, Newey (1988) has given sufficient conditions for the number of restrictions to grow with the sample size to achieve asymptotic efficiency.

1.1 Some Examples

To illustrate the usefulness of the results and for exposition, it is helpful to consider some examples throughout the following discussion. For simplicity, and in keeping with most of the semiparametric efficiency literature to date, all results will be limited to the i.i.d. data case, where z represents the data vector for a single observation.

The first example is a nonlinear simultaneous equations model with unknown, homoskedastic disturbance distribution. This model is useful in econometrics, e.g. for supply and demand modeling, and includes as special cases transformation models that have long been of interest in statistics. Let y be an $s \times 1$ vector of dependent variables. Also, let $\rho(y,x,\beta)$ be

a $s \times 1$ vector of functions of y, a vector of exogenous variables x, and a vector of parameters β, such that $\rho(z,\beta) \equiv \rho(y,x,\beta)$ is a one-to-one function of y. The model specifies that for the true parameter value β_0,

(1.1) $\quad \varepsilon \equiv \rho(z,\beta_0)$ is independent of x.

The nonparametric components of this model are the distributions of ε and x. No restriction is imposed on the location of ε, so that all additive constant parameters have been absorbed into ε.

Transformation models are included as a special case with y and ρ being scalars. In this case $\rho(z,\beta)$ can be interpreted as the residual from a transformation of y. For instance, $\rho(z,\beta) = (y^{\beta_1} - 1)/\beta_1 - x'\beta_2$ corresponds to the Box-Cox (1964) transformation. Note that the model (1.1) means that this is a transformation to independence, having homoskedasticity as a particular implication.

It is known that when $\rho(z,\beta)$ is jointly nonlinear in y and β, a maximum likelihood estimator based on specifying a distribution for ε may be inconsistent; e.g. Amemiya (1977). Distribution-free estimators include the nonlinear instrumental variable estimators of Sargan (1959), Kelejian (1971), Amemiya (1974,1977), Amemiya and Powell (1981), the quantile estimators of Hinkley (1977), Carroll and Ruppert (1984), and Powell (1990) for transformation models, and the moment estimators of MaCurdy (1982), Taylor (1985), and Ruppert and Aldershof (1989). Efficient estimation of this model is discussed below.

The second example is a conditional distribution regression (CDR) model. Let y be a single dependent variable and $v(x,\beta)$ a known regression function. The model specifies that the conditional distribution of y given regressors x depends only on $v(x,\beta_0)$ for an unknown parameter value β_0,

(1.3) $y|x \sim F(y|v(x,\beta_0))$,

where $F(y|v)$ is the conditional cumulative distribution function. Note that the regression function is only identified up to location and scale.

This model is similar to regression models where both the mean and variance of y depend only on the regression function: e.g. see McCullagh and Nelder (1983). Here all the conditional moments of y depend only on v. This model has been previously considered by Manski (1988) for y a binary variable, where it is less restrictive.

This model is invariant to transformations of y. That is, if $\tilde{y} = \tau(y)$, where the model for (y,x) is CDR with regression $v(x,\beta)$, then the model for $(\tilde{y},x)$ is CDR with regression $v(x,\beta)$. This statement is formalized in Appendix B in the efficiency bound context for a certain class of transformations. Consequently, when the distribution of y has a discrete component, y can often be thought of as being obtained by some fixed censoring process from a latent dependent variable obeying the CDR model.

When y is continuously distributed there is an interpretation of this model that helps to relate it to transformation models. Suppose that $F(y|v)$ is continuous and one-to-one, and define $\varepsilon = F(y|v)$. This variable ε is distributed as standard uniform (i.e. $U(0,1)$), independently of x. Thus,

(1.4) $y = T(v(x,\beta_0),\varepsilon)$, ε and x independent, $\varepsilon \sim U(0,1)$,

where the transformation $T(v,\varepsilon) \equiv F^{-1}(\varepsilon|v)$ is strictly monotonic in ε but is otherwise unrestricted. This model imposes weaker restrictions than one with $y = \tau(v(x,\beta_0)+\varepsilon)$ where ε and x are independent and τ and the distribution of ε unknown. For example, if y is time to failure, then this model is a generalization of the semiparametric proportional hazard model of Cox (1975), that allows for unobserved heterogeneity and non-proportional

hazards.

When y is discrete this model is a semiparametric generalization of well known models. For example, if y is confined to the nonnegative integers, then it is a semiparametric version of a Poisson regression model.

The third and last example is the conditional mean transformation model of Ichimura (1986), Chamberlain (1987b), and others:

$$(1.5) \qquad E[y|x] = Y(v(x,\beta_0)),$$

where $Y(\circ)$ is an unknown function. This model imposes weaker restrictions than the CDR model. It and the other models will be illustrative.

2. Preliminaries

2.1 Semiparametric Efficiency Bounds

It is helpful to briefly review semiparametric efficiency bounds, as developed by Stein (1956), Koshevnik and Levit (1976), Pfanzagl (1982), Begun, Hall, Huang, and Wellner (1983), and Bickel, Klaassen, Ritov, and Wellner (1989) (BKRW henceforth). Define a *parametric submodel* to be one that satisfies the semiparametric assumptions and contains the truth. Any semiparametric estimator must have an asymptotic variance that is no smaller than the Cramer-Rao bound for every parametric submodel, giving Stein's (1956) insight:

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

Regularity conditions are needed to make this statement precise. The

parametric submodels must be regular in that they are mean-square *smooth* (see Appendix B), have nonsingular information matrices, and satisfy other regularity conditions appropriate to the model. A precise definition of V is that it is the supremum of Cramer-Rao bounds for regular parametric submodels. The estimators must be regular in the following sense. For a parametric submodel with Euclidean parameter vector θ let $\beta(\theta)$ be the parameters of interest. A local data generating process (LDGP) is one where the for each sample size n the data is distributed according to θ_n, with $\sqrt{n}(\theta_n-\theta_0)$ bounded. An estimator $\hat{\beta}$ is said to be *regular* if for each regular parametric submodel and LDGP, $\sqrt{n}(\hat{\beta}-\beta(\theta_n))$ has a limiting distribution that does not depend on the sequence $\{\theta_n\}$ or the parametric submodel. That V is an asymptotic variance bound for regular estimators follows from semiparametric extensions of Hajek's (1970) representation theorem, e.g. Begun et. al. (1983). A vector version of Theorem 2 i) of Chamberlain (1986) is

If $\hat{\beta}$ is regular then the limiting distribution of $\sqrt{n}(\hat{\beta}-\beta_0)$ is equal to the distribution of $Y + U$, where $Y \sim N(0,V)$ and U is independent of Y.

An efficient semiparametric estimator is one that is asymptotically normal with covariance matrix V and is regular.

The projection interpretation of the bound developed by Begun et. al. (1983) and BKRW will prove useful here. Let the data consist of i.i.d. observations $z_1,\ldots,z_n$. Consider a regular parametric submodel with parameters $\theta = (\beta',\eta')'$ and likelihood function $\ell(z|\theta)$ for a single observation z_i. The $q \times 1$ vector of parameters of interest is β and the η parameters correspond to the nonparametric part of the model. Let $S_0 = (S_\beta',S_\eta')'$ be the score for θ for a single observation, evaluated at the true parameter values, where typically $S_\theta = \partial \ln \ell(z|\theta_0)/\partial\theta$ (see Appendix B). The

z argument may be suppressed for notational convenience, as here. Define the *tangent set* $\mathcal{T}$ to be the mean-square closure of $q \times 1$ linear combinations of scores S_η for the nonparametric component:

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

where each B_j is a matrix of constants. Consider S_β as an element of, and $\mathcal{T}$ as a subset of, the Hilbert space of $q \times 1$ random vectors r with inner product $E[r_1' r_2]$. 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

(2.1) $\quad S_\beta - S \in \mathcal{T}, \quad E[S't] = 0 \text{ for all } t \in \mathcal{T}.$

A version of Corollary 3.4.1 of BKRW (see Newey, 1990a, Theorem 3.2) is

If $\ell(z|\beta)$ is regular with score S_β, $\mathcal{T}$ is linear, and $E[SS']$ is nonsingular, then $V = (E[SS'])^{-1}$.

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

It will be useful for the estimation results discussed below to report the results of this calculation for the examples. In the nonlinear simultaneous equations model the parameters of interest are those of the residual $\rho(z,\beta)$. A parametric submodel corresponds to a parametric family of density functions $f_1(\varepsilon|\eta)f_2(x|\eta)$ for (ε, x) such that for some η_0 $f_1(\varepsilon|\eta_0)f_2(x|\eta_0) = f_{10}(\varepsilon)f_{20}(x)$, the true density. The likelihood and score vectors for a parametric submodel are for $J(z,\beta) = \ln|\det(\partial\rho(z,\beta)/\partial y)|$,

(2.2) $\quad \ell(z|\theta) = \exp[J(z,\beta)]f_1(\rho(z,\beta)|\eta)f_2(x|\eta),$

$$S_\beta = J_\beta(z,\beta_0) + \rho_\beta(z,\beta_0)'s(\varepsilon), \quad S_\eta = S_{\eta 1} + S_{\eta 2},$$

$$S_{\eta 1} = \partial \ln f_1(\varepsilon|\eta_0)/\partial\eta, \quad S_{\eta 2} = \partial \ln f_2(x|\eta_0)/\partial\eta$$

where $s(\varepsilon) = f_{10}(\varepsilon)^{-1}\partial f(\varepsilon)_{10}/\partial\varepsilon$ and the β subscripts denotes the partial derivative. The nuisance score S_η is unrestricted except for the additive structure implied by independence of ε and x, giving the tangent set

$$(2.3) \qquad \mathcal{T} = \{t_1(\varepsilon)+t_2(x) : E[t_1(\varepsilon)] = E[t_2(x)] = 0\}.$$

To avoid additional clutter it is here and henceforth assumed that second moments exist whenever needed. For a $q \times 1$ random vector $R(z)$, the projection $Proj(R|\mathcal{T})$ of R on $\mathcal{T}$ is

$$(2.4) \qquad Proj(R|\mathcal{T}) = E[R|\varepsilon]-E[R] + E[R|x]-E[R].$$

By ancillarity of x for β (implying $E[S_\beta|x] = 0$) the efficient score is

$$(2.5) \qquad S = S_\beta - E[S_\beta|\varepsilon] = J_\beta - E[J_\beta|\varepsilon] + \{\rho_\beta-E[\rho_\beta|\varepsilon]\}'s(\varepsilon).$$

Regularity conditions for this result are given in Newey (1989a).

For the CDR model, suppose that (y,x) is absolutely continuous with respect to the product of a dominating measure for y and the marginal distribution for x, with conditional density $f_0(y|v)$ of y given x and marginal density $f_0(x)$. For a parametric submodel $f(y|v,\eta)f(x|\eta)$, the likelihood and scores are

$$(2.6) \qquad \ell(z|\theta) = f(y|v(x,\beta),\eta)f(x|\eta), \quad S_\beta = S_v \cdot v_\beta, \quad S_\eta = S_{\eta 1}+S_{\eta 2},$$

where $S_v \equiv \partial \ln f(y|v)/\partial v$, $v_\beta \equiv \partial v(x,\beta_0)/\partial\beta$, $S_{\eta 1} = \partial \ln f(y|v,\eta_0)/\partial\eta$, and $S_{\eta 2} = \partial \ln f(x|\eta_0)/\partial\eta$. By the CDR model, $S_{\eta 1}$ is a functional only of y and v, and because it is a conditional score, $E[S_{\eta 1}|x] = 0$. Also, by the CDR model it follows that for $A(y,v)$ and $B(x)$, $E[E[A|y,v]|x] = E[A|v]$ and $E[B(x)|y,v] = E[B(x)|v]$. It follows that the tangent set, projection, and efficient score are

(2.7) $\mathcal{T} = \{t_1(y,v)+t_2(x) : E[t_1|x] = 0,\ E[t_2] = 0\}$,

$$\text{Proj}(R|\mathcal{T}) = E[R|y,v] - E[R|v] + E[R|x] - E[R],$$

$$S = S_\beta - E[S_\beta|y,v] = S_v(v_\beta - E[v_\beta|v,y]) = S_v(v_\beta - E[v_\beta|v]).$$

Details are given in Appendix B.

For the conditional mean index model, the tangent set, the projection, and the efficient score are shown by Newey and Stoker (1989) to be

(2.8) $\mathcal{T} = \{t : E[t] = 0,\ E[y(t-E[t|x])|x] = E[y(t-E[t|x])|v]\ \}$,

$$\text{Proj}(R|\mathcal{T}) = R_y - \omega(x)^{-1}(y-Y(v))\{E[yR_y|x] - E^{\omega}[yR_y|v]\} + R_x - E[R],$$

$$S = \omega(x)^{-1}[y-Y(v)](v_\beta - E^{\omega}[v_\beta|v]),$$

where $R_x = E[R|x]$, $R_y = R-R_x$, $\omega(x) = \text{Var}(y|x)$, and for $A = A(y,x)$, $E^{\omega}[A|v] = E[\omega(x)^{-1}A|v]/E[\omega(x)^{-1}|v]$ is the conditional expectation given v for the probability measure $\text{Prob}(\mathcal{E}) = E[\omega(x)^{-1}1(\mathcal{E})]/E[\omega(x)^{-1}]$. This formula for S verifies a conjecture of Chamberlain (1987b).

2.2 Semiparametric M-Estimators

The efficient estimation scheme considered below is closely related to a certain type of m-estimator, which will now be discussed. Let $m(z,\beta,\alpha)$ be $q \times 1$ vector of functions of z, β, and a function α, such that for true values β_0 and α_0,

(2.9) $E[m(z,\beta_0,\alpha_0)] = 0.$

The estimators considered solve a sample moment analog of this equation, with α replaced by an estimator. Let $\hat{\alpha}(z,\beta)$ be a consistent (in some

appropriate metric) nonparametric estimator of a function $\alpha(z,\beta)$, depending on β, with $\alpha(z,\beta_0) = \alpha_0$. A semiparametric m-estimator of the parameters of interest solves

$$(2.10) \qquad \sqrt{n}\hat{m}_n(\hat{\beta}) = o_p(1), \quad \hat{m}_n(\beta) \equiv \sum_{i=1}^n m(z_i,\beta,\hat{\alpha}(z_i,\beta))/n.$$

The general idea here is that $\hat{\beta}$ is obtained by a procedure that first "concentrates out" the function $\alpha(z,\beta)$. An early and important example is the Buckley and James (1979) estimator for censored regression. The estimator $\hat{\beta}$ should be consistent, under suitable regularity conditions, by equation (2.9) and consistency of $\hat{\alpha}(z,\beta)$.

Choices of $m(z,\beta,\alpha)$ and $\hat{\alpha}$ are required for such estimators. Finding useful $\hat{\alpha}$ can be difficult, although well known nonparametric methods often work. In contrast, it is easy to find $m(z,\beta,\alpha)$ satisfying equation (2.9). For instance, by the mean zero property of scores, $m(z,\beta,\alpha)$ such that $m(z,\beta_0,\alpha_0) = S$ would do. Indeed, as in Bickel (1982), Schick (1986), and Klaassen (1987), if $\hat{\alpha}$ is sufficiently well behaved, then the resulting estimator should be efficient. Other choices of $m(z,\beta,\alpha)$ may also be desirable. It is often possible to reduce the dimensionality of α by fixing some components of the efficient score at (possibly) false values, which could lead to improved small sample properties. As argued in Bickel (1982), BKRW, and Newey (1990a), if the semiparametric model can be nested within one that is convex in a nonparametric component (i.e. for that component the set of its values is convex and that the likelihood of a convex combination is a convex combination of likelihoods, then $m(z,\beta,\alpha)$ equal to the efficient score with that component fixed generally satisfies equation (2.9). In addition, one can draw on suggestions for particular models, e.g. Powell (1984) for regression with fixed censoring.

The asymptotic variance of these estimators plays an essential role in

the construction of efficient estimators discussed below. Well known results apply when α is not present, i.e. when $m(z,\beta,\alpha) = m(z,\beta)$. Under regularity conditions of Huber (1967) or Pakes and Pollard (1989), the asymptotic variance of $\hat{\beta}$ is $M^{-1}E[mm']M^{-1}$, where $M \equiv \partial E[m(z,\beta)]/\partial\beta|_{\beta=\beta_0}$ and $m \equiv m(z,\beta_0)$. When α is present, the asymptotic variance can be more complicated. For discussion purposes, suppose $\hat{m}_n(\beta)$ is continuously differentiable, $\sqrt{n}\hat{m}_n(\beta_0) = O_p(1)$, $\partial\hat{m}_n(\bar{\beta})/\partial\beta \xrightarrow{p} M \equiv \partial E[m(z,\beta,\alpha(z,\beta)]/\partial\beta|_{\beta=\beta_0}$ with M nonsingular for any $\bar{\beta} \xrightarrow{p} \beta_0$, and $\hat{\beta} \xrightarrow{p} \beta_0$. Then the usual mean value expansion gives

$$(2.11) \qquad \sqrt{n}(\hat{\beta}-\beta_0) = -M^{-1}\sqrt{n}\hat{m}_n(\beta_0) + o_p(1).$$

As discussed in Newey (1989b), a conjecture for the asymptotic variance of $\sqrt{n}\hat{m}_n(\beta_0)$ can often be calculated from the pathwise derivative of the functional estimated by $\hat{m}_n(\beta_0)$ under general misspecification. Let F denote a general distribution, restricted only in satisfying regularity conditions, and let $\nu(F)$ be the probability limit of $\hat{m}_n(\beta_0)$ under F. As in Pfanzagl (1982), the *pathwise derivative* of $\nu(F)$, if it exists, is a vector $u(z)$ satisfying $E[u(z)] = 0$ and

$$(2.12) \qquad \partial\nu(F_\theta)/\partial\theta|_{\theta=\theta_0} = E[uS_\theta'],$$

where F_θ is a regular parametric family of possible values for F and S_θ is the associated score for θ. If $\sqrt{n}\hat{m}_n(\beta_0)$ is asymptotically equivalent to a sample average of some function of the data and is sufficiently well-behaved,

$$(2.13) \qquad \sqrt{n}\hat{m}_n(\beta_0) = \sum_{i=1}^{n} u_i/\sqrt{n} + o_p(1).$$

When equations (2.11) and (2.13) are satisfied it follows by the central

limit theorem that the asymptotic covariance matrix of $\hat{\beta}$ is $M^{-1}E[uu']M^{-1}$. Specific formulae for u, referred to henceforth as the *influence function* (corresponding to $\hat{m}_n(\beta_0)$), when $\alpha(z,\beta)$ is the derivative of a density or a conditional expectation are derived in Newey (1989b). Primitive conditions for this pathwise derivative formula for the asymptotic variance can often be obtained directly or from general results of Andrews (1989) and Newey (1989b).

The quantity u-m is a correction term for the presence of $\hat{\alpha}$. An important special case is that where u-m is an element of the tangent set, referred to as the *efficient-*$\hat{\alpha}$ case. In this case, $u = m - Proj(m|\mathcal{T})$; see BKRW or Newey (1990a).

In the nonlinear simultaneous equations example semiparametric m-estimators can be formed by mimicking the form of the efficient score and using a residual-based (i.e. bootstrap) estimator of the conditional expectation. Let $\tilde{S}(z,\beta)$ be some function of the data and parameters. Note that since $\rho(z,\beta)$ is a one-to-one function of y there exists $\pi(x,\varepsilon,\beta)$ such that $y = \pi(x,\rho(z,\beta),\beta)$. Let $\alpha(z,\beta) \equiv \int\tilde{S}(\pi(\tilde{x},\rho(z,\beta),\beta),\tilde{x},\beta)dF_x(\tilde{x})$ be the integral of $\tilde{S}(z,\beta)$ over the marginal distribution of x, holding $\rho(z,\beta)$ fixed. By independence of x and ε, $\alpha(z,\beta_0) = E[\tilde{S}(z,\beta_0)|\varepsilon]$, so that $m(z,\beta,\alpha) \equiv \tilde{S}(z,\beta) - \alpha(z,\beta)$ satisfies equation (2.9). Furthermore, $\alpha(z,\beta)$ can be estimated by averaging over observations on x (i.e. by the bootstrap), yielding $\hat{\alpha}(z,\beta) = \sum_{j=1}^{n}\tilde{S}(\pi(x_j,\rho(z,\beta),\beta),x_j,\beta)/n$. Averaging again,

$$\hat{m}_n(\beta) = \sum_{i=1}^{n}\tilde{S}(z_i,\beta)/n - \sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{S}(\pi(x_j,\rho(z_i,\beta),\beta),x_j,\beta)/n^2 \qquad (2.14)$$

Two particular types of $\tilde{S}(z,\beta)$ are of interest. The first type is multiplicatively separable in functions of x and $\rho(z,\beta)$, say $\tilde{S}(z,\beta) = A(x)r(\rho(z,\beta))$ for A(x) a matrix and $r(\rho)$ a conformable vector. Here $\hat{m}_n(\beta)$ is a vector of sample covariances between A(x) and $r(\rho(z,\beta))$, so that the solution to equation (2.10) makes use of the uncorrelatedness

implication of independence between x and ε. An advantage of this estimator is that it does not actually use $\pi(\varepsilon,x,\beta)$, which can be hard to compute. Included as special cases are instrumental variables estimators (with a constant, additive parameter vector concentrated out of $\rho(z,\beta)$), where $r(\rho) = \rho$, as well as the higher moment estimators of MaCurdy (1982), Ruppert and Aldershof (1989), and Robinson (1989).

The second type of $\tilde{S}(z,\beta)$ more closely mimics the score for β, taking $\tilde{S}(z,\beta) = J_\beta(z,\beta) + \rho_\beta(z,\beta)'\tilde{s}(\rho(z,\beta))$, where $\tilde{s}(\varepsilon)$ is some vector of functions of ε. As shown in Newey (1989a), the resulting estimator is efficient if the true disturbance score is $\tilde{s}(\varepsilon)$, for example when $\tilde{s}(\varepsilon) = -\varepsilon$ and ε is $N(0,I)$.

These estimators can be generalized to allow for location and scale parameters for $\rho(z,\beta)$. For an s-dimensional vector of location parameters μ and a positive definite scale matrix Σ, let $\theta = (\beta',\mu',\text{hvec}(\Sigma)')'$, where $\text{hvec}(\cdot)$ denotes the usual column vectorization of a symmetric matrix. In equation (2.14) and the corresponding types of $\tilde{S}(z,\theta)$, consider replacing β by θ, $r(\rho)$ by $r(\Sigma^{-1/2}(\rho-\mu))$, and $\tilde{s}(\rho)$ by $(\Sigma^{-1/2})'\tilde{s}(\Sigma^{-1/2}(\rho-\mu))$; then the resulting $\hat{\beta}$ will be invariant to location and scale shifts of $\rho(z,\beta)$, and its asymptotic variance will be the same as if μ and Σ were equal to their true values. For example, when $\tilde{s}(\varepsilon) = -\varepsilon$ and μ and Σ are estimated as the sample mean and variance of $\rho(z,\beta)$, the second type of estimator will be efficient when ε is $N(\mu,\Sigma)$.

Because $\hat{m}_n(\beta)$ is a V-statistic, it is easy to compute directly asymptotic variance of $\hat{\beta}$ By the V-statistic projection theorem, $\hat{m}_n(\beta)$ satisfies equation (2.13) with

$$(2.15) \qquad u = \tilde{S}(z,\beta_0) - E[\tilde{S}(z,\beta_0)|x] - E[\tilde{S}(z,\beta_0)|\varepsilon] + E[\tilde{S}(z,\beta_0)].$$

The asymptotic variance of $\hat{\beta}$ will then be of the $M^{-1}E[uu']M^{-1}$, form

discussed above.

In the CDR model semiparametric m-estimators can be formed by mimicking the form of the efficient score and using a nonparametric estimator of the expectation conditional on $v(x,\beta)$. Let $\tilde{s}(y,v)$ be some function and let $\alpha(z,\beta) \equiv E[v_\beta(x,\beta)|v(x,\beta)]$. By the CDR restriction (which implies $E[\tilde{s}(y,v)|x] = E[\tilde{s}(y,v)|v]$), $v_\beta(x,\beta_0)-\alpha(z,\beta_0) = v_\beta - E[v_\beta|v]$ is uncorrelated with $\tilde{s}(y,v)$, so that $m(z,\beta,\alpha) \equiv v_\beta(x,\beta)[s(y,v(x,\beta))-\alpha(z,\beta)]$ satisfies equation (2.9). Furthermore, $\alpha(z,\beta)$ can be estimated by a nonparametric regression $\hat{\alpha}(z,\beta)$ (e.g. kernel regression) of $v_\beta(x,\beta)$ on $v(x,\beta)$. Averaging gives

$$(2.16) \qquad \hat{m}_n(\beta) = \sum_{i=1}^{n}[v_\beta(x_i,\beta) - \hat{\alpha}(x_i,\beta)]\tilde{s}(y_i,v(x_i,\beta))/n.$$

A conjectured form of the asymptotic variance of $\hat{\beta}$ follows from the results of Newey (1989b) on estimators that depend on preliminary nonparametric regressions. Noting that $\partial m(z,\beta_0,\alpha_0)/\partial\alpha = -\tilde{s}(y,v)$, it follows by $\hat{\alpha}(z,\beta_0)$ a nonparametric estimator of the conditional expectation of v_β given v that

$$(2.17) \qquad u = m + E[\partial m(z,\beta_0,\alpha)/\partial\alpha|v]|_{\alpha=E[v_\beta|v]}\{v_\beta - E[v_\beta|v]\}$$

$$= \{v_\beta - E[v_\beta|v]\}\{\tilde{s}(y,v)-E[\tilde{s}|v]\}.$$

It is interesting to note that if $\tilde{s}(y,v)$ equals the true conditional score, then $E[\tilde{s}|v] = 0$, implying that u equals the efficient score. As one might expect, it follows from this that $\hat{\beta}$ will be efficient, as will be further discussed below. Other, feasible choices of $\tilde{s}(y,v)$ will be also discussed below.

In the conditional mean transformation model it is difficult to mimic the efficient score. A more ad-hoc approach is to use an estimator similar to

that for the CDR model. The conditional mean restriction $E[y|x] = E[y|v]$ means that $A(x)-E[A(x)|v]$ will be uncorrelated with y for any vector of functions $A(x)$. Following the previous discussion, let $\alpha(z,\beta) = E[A(x)|v(x,\beta)]$ and $m(z,\beta,\alpha) = [A(x)-\alpha(z,\beta)]y$. Averaging, and applying the same asymptotic distribution argument as above, gives

$$(2.18) \qquad \hat{m}_n(\beta) = \sum_{i=1}^{n}[A(x) - \hat{\alpha}(z_i,\beta)]y_i/n.$$

$$u = \{A(x)-E[A|v]\}\{y-E[y|v]\}.$$

In contrast with CDR estimator discussed above, this estimator allows for any function of x to be interacted with y (i.e. A is not restricted to the form of v_β times some function of v). This change is important for efficiency reasons. Intuitively, the conditional variance of y can depend on x and not just v, so that the "weighting" for an efficient estimator requires this extra flexibility.

3. Combining Moment Restrictions

The efficient estimation scheme considered here is m-estimation based on a linear combination of different $m(z,\beta,\alpha)$ functions. The motivation is that if the functions are selected from a sufficiently rich set and the linear combination chosen judiciously, then the resulting estimator should be close to efficient. To describe such estimates, let there be $q \times 1$ vectors

$$(3.1) \qquad m_k(z,\beta,\alpha), \quad E[m_k(z,\beta_0,\alpha_0)] = 0, \quad (k = 0, 1, \ldots),$$

as in Section 2. A single moment vector can be formed from a linear combination of the first K vectors,

(3.2) $[(1,\gamma')\otimes I_q]m(z,\beta,\alpha)$, $m(z,\beta,\alpha) \equiv (m_0(z,\beta,\alpha)',\ldots,m_K(z,\beta,\alpha)')'$,

where $\gamma = (\gamma_1,\ldots,\gamma_K)'$ for some constants γ_K, a K subscript on m and γ is suppressed for notational convenience, and $\otimes$ denotes the Kronecker product. The first coefficient is constrained to be 1 for parsimony reasons that are further discussed below. If α is replaced by $\hat{\alpha}(z,\beta)$ as described in Section 4, and γ is replaced by some estimate $\hat{\gamma}$, then a sample moment vector can be formed as

(3.3) $\hat{m}_n(\beta) = \sum_{i=1}^n [(1,\hat{\gamma}')\otimes I_q]m(z_i,\beta,\hat{\alpha}(z_i,\beta))/n.$

An estimator might then be obtained as the solution to equation (2.10), and its asymptotic variance calculated in the way described in Section 4.

3.1 Approximating the Efficient Score

To motivate the choice of linear combination coefficients, it is useful to work with a particular expression for the asymptotic variance of $\hat{\beta}$. Under appropriate regularity conditions specified in BKRW or Newey (1990a, Theorem 2.2), if $\sqrt{n}(\hat{\beta}-\beta_0) = -M^{-1}\sum_{i=1}^n u_i/\sqrt{n} + o_p(1)$, then $E[ut'] = 0$ for all t in the tangent set $\mathcal{T}$ and $M = -E[uS_\beta']$. Furthermore, as long as the hypotheses of the projection interpretation of the bound are satisfied, $S = S_\beta - \bar{t}$ for $\bar{t}$ in $\mathcal{T}$, so that

(3.4) $M = -E[uS_\beta] + E[u\bar{t}'] = -E[uS']$,

$M^{-1}E[uu']M^{-1\prime} = (E[uS'])^{-1}E[uu'](E[Su'])^{-1}$.

The first equality in (3.4) is a semiparametric version of the well known generalized information matrix equality. It is a generalization of a result used by Beran (1976). An implication of the second equality is that when u

is close *in mean square* to S, then the asymptotic variance of $\hat{\beta}$ will be close to $(E[SS'])^{-1}E[SS'](E[SS'])^{-1} = V$, the semiparametric efficiency bound. This observation motivates an estimation scheme where the linear combination coefficients are chosen so as to approximate the efficient score.

Suppose that the entire vector $m(z,\beta,\alpha)$ satisfies an equation like (2.13),

$$\sum_{i=1}^{n} m(z_i,\beta_0,\hat{\alpha}(z_i,\beta_0))/\sqrt{n} = \sum_{i=1}^{n} u_i/\sqrt{n} + o_p(1).$$

A conjectured form for u can often be computed as in Section 2. For a positive definite matrix Q let

$$(3.5) \qquad \bar{\gamma} = \operatorname{argmin}_{\gamma} E[\{S-[(1,\gamma')\otimes I_q]u\}'Q\{S-[(1,\gamma')\otimes I_q]u\}].$$

That is, $\bar{\gamma}$ minimizes the expected squared distance between the difference of the efficient score and linear combinations of the moment functions, where Q measures the distance. By Q positive definite, $[(1,\gamma')\otimes I_q]u$ will be close to S in mean square, and an m-estimator with $m(z,\beta,\alpha) = [(1,\bar{\gamma}')\otimes I_q]m(z,\beta,\alpha)$ will be close to being efficient, when the expected distance in equation (5.4) is small.

The matrix Q is present to allow flexibility in the definition of distance between the efficient score and the moment functions. One way to choose Q is so that the distance measure is invariant to parameterization of β; e.g. Q equal to the asymptotic variance of a preliminary estimator.

Equation (3.4) can be used to estimate $\bar{\gamma}$. Let u_k correspond to m_k, so that $u = (u_0',\ldots,u_K')'$, and let $U = [u_1,\ldots,u_K]$, so that $[(1,\gamma')\otimes I_q]u = u_0+U\gamma$. Then the solution of (3.5) is

$$(3.6) \qquad \bar{\gamma} = (E[U'QU])^{-1}(E[U'QS]-E[U'Qu_0]).$$

By equation (3.4) the k^{th} element of $E[U'QS]$ is $E[u_k'QS] =$

$\text{trace}(QE[Su_k']) = -\text{trace}(QM_k')$, where $M_k \equiv \partial E[m_k(z,\beta,\alpha(z,\beta))]/\partial\beta|_{\beta=\beta_0}$. This result allows estimation of $\bar{\gamma}$ by replacing the population moments that appear in equation (3.6) with estimated sample moments. Let $\hat{Q}$ be an estimate of Q and let $\hat{M}_k$ be an estimate of M_k, $(k=0,1,\ldots,K)$. If $m_{kn}(\beta) \equiv \sum_{i=1}^n m_k(z_i,\beta,\hat{\alpha}(z_i,\beta))/n$ is continuously differentiable in β then $\hat{M}_k = \partial m_{kn}(\beta)/\partial\beta|_{\beta=\hat{\beta}}$ should do, where $\hat{\beta}$ is an initial estimate. Otherwise it may be necessary to resort to a numerical derivative (e.g. see Newey, 1990b). Also, let $\hat{u}_{ki}$, $(k=0,1,\ldots;i=1,\ldots,n)$ be estimators of u_k. When $u_k = m_k$, $\hat{u}_{ki} = m(z_i,\hat{\beta},\hat{\alpha}(z_i,\hat{\beta}))$ should do. Otherwise, the form of $\hat{u}_{ki}$ will depend on the form of u_k; e.g. see the examples discussed below. Then for $\hat{U}_i \equiv [\hat{u}_{1i},\ldots,\hat{u}_{Ki}]$, an estimator of $\bar{\gamma}$ is

$$(3.7) \qquad \hat{\gamma} = -[\textstyle\sum_{i=1}^n \hat{U}_i'\hat{Q}\hat{U}_i/n]^{-1}\{(\text{tr}[\hat{Q}\hat{M}_1'],\ldots,\text{tr}[\hat{Q}\hat{M}_K'])' + \sum_{i=1}^n \hat{U}_i'\hat{Q}\hat{u}_{0i}/n\}.$$

An estimator with this linear combination vector can be calculated as in equation (2.10) using the sample moments $\hat{m}_n(\beta)$ calculated as in equation (3.3).

It is both theoretically and computationally convenient to work with a one-step version of this estimator. The one-step estimator $\tilde{\beta}$ solves a linear approximation $\hat{m}_n(\hat{\beta}) + (\hat{M}_0 + \sum_{k=1}^K \hat{\gamma}_k\hat{M}_k)(\beta-\hat{\beta}) = 0$ of equation (2.10):

$$(3.8) \qquad \tilde{\beta} = \hat{\beta} - (\hat{M}_0 + \textstyle\sum_{k=1}^K \hat{\gamma}_k\hat{M}_k)^{-1}\hat{m}_n(\hat{\beta}).$$

Here, as in other contexts, this one-step estimator will have the same asymptotic properties as a consistent estimator solving $\sqrt{n}\hat{m}_n(\beta) = o_p(1)$.

It is important to note that the asymptotic distribution of $\tilde{\beta}$ should be unaffected by estimation of $\bar{\gamma}$, if $\hat{\gamma}$ is consistent for $\bar{\gamma}$. This feature is just the well-known fact that the estimation of linear combination coefficients does not affect the limiting distribution of m-estimators (e.g. Hansen, 1982), which results from $\sum_{i=1}^n m(z_i,\beta_0,\hat{\alpha}(z_i,\beta_0))/\sqrt{n}$ being bounded in

probability.

3.2 Minimizing the Asymptotic Variance

An interpretation of $\bar{\gamma}$ as minimizing the asymptotic variance matrix is available when $[(1,\gamma')\otimes I_q]m$ consists of an unrestricted matrix linear combination of a single vector. Suppose that $K = qr$ for a positive integer r and $m(z,\beta,\alpha) = (0,\text{vec}(g(z,\beta,\alpha)'\otimes I_q)')'$ for some $r \times 1$ vector g, so that $[(1,\gamma')\otimes I_q]m(z,\beta,\alpha) = (g(z,\beta,\alpha)'\otimes I_q)\gamma = \Gamma g(z,\beta,\alpha)$ for Γ such that $\gamma = \text{vec}(\Gamma)$. Suppose that u_g satisfies equation (2.13) for $g(z,\beta,\alpha)$ and u_g replacing $m(z,\beta,\alpha)$ and u. Then the objective function in (3.5) is $\text{trace}\{QE[(S-\Gamma u_g)(S-\Gamma u_g)']\}$, and the first-order conditions for its minimization are $E[(S-\bar{\Gamma}u_g)u_g'] = 0$. Let $G \equiv \partial E[g(z,\beta,\alpha(z,\beta))]/\partial\beta|_{\beta=\beta_0}$. Solving for $\bar{\Gamma}$ and assuming equation (3.4) applies to g,

$$(3.9) \qquad \bar{\Gamma} = E[Su_g'](E[u_g u_g'])^{-1} = -G'(E[u_g u_g'])^{-1}.$$

By Theorem 3.2 of Hansen (1982), $\bar{\Gamma}$ is a matrix of linear combination coefficients that minimizes the asymptotic variance of an m-estimator based on moment functions $\Gamma g(z,\beta,\alpha)$.

Of course, given functions as specified in equation (3.1), it is always possible to construct an unrestricted moment vector, leading to an estimator with no larger asymptotic variance. The reason that other cases are considered here is that it is often possible to reduce the dimension of γ by choosing a restricted form for $m(z,\beta,\alpha)$, without affecting the ability to approximate the efficient score. As discussed below, this is possible in the first two examples. It is plausible that this increased parsimony could result in improved finite sample properties. The first coefficient in the linear combination of equation (3.2) is restricted to be one for similar

reasons. In the nonlinear simultaneous equations model this restriction gives a useful way of accounting for the Jacobian term in the efficient score.

3.3 Cross-Validated Number of Moments

The estimator $\tilde{\beta}$ depends on the choice of $\{m_k(z,\beta,\alpha)\}$ and the number of terms K. The nature of $\{m_k\}$ is specific to particular models, and will be discussed below for the examples. Choosing K is a generic problem, that is important for implementation. Here a cross-validated choice of K will be discussed. The asymptotic theory will allow for this (or other data-based) choice of K. Its properties will be evaluated in the Monte Carlo example.

A cross-validated choice of K can be based on the distance measure used in calculation of the coefficients $\hat{\gamma}$. The mean-square approximation depends on the linear combination $u_0+U\gamma$ through the terms

$$(3.10)\qquad -2E[(u_0+U\gamma)'QS] + E[(u_0+U\gamma)'Q(u_0+U\gamma)]$$

$$= 2[\mathrm{trace}(QM_0'+\textstyle\sum_{k=1}^{K} QM_k'\gamma_k)] + E[(u_0+U\gamma)'Q(u_0+U\gamma)],$$

where the constant term $E[S'QS]$ has been omitted and the equality uses equation (3.4). Note that it would not be useful to choose K to minimize this function by replacing unknown quantities with the estimates discussed above; for each K, $\hat{\gamma}$ minimizes the result, while increasing K corresponds to relaxing a zero restriction on the next higher-order coefficient, leading to a decrease in this object. Intuitively, the choice of $\hat{\gamma}$ as the minimizing coefficient biases the estimate downward.

One way past this difficulty, which has proven fruitful in other contexts, is delete-one cross-validation. To describe how this idea might apply here, let $\hat{Q}$, $\hat{M}_k$, and $\hat{u}_{ki}$ be the estimates considered above, and suppose that $\hat{M}_k^{\cdot} = \sum_{i=1}^{n}\hat{M}_{ki}/n$ for some $\hat{M}_{ki}$ (for example, $\hat{M}_{ki}$ =

$\partial m(z_i,\beta,\hat{\alpha}(z_i,\beta))/\partial\beta|_{\beta=\hat{\beta}}$ in the differentiable case). Let $\hat{\gamma}_{-i}$ denote $\hat{\gamma}$ with the i^{th} observation excluded from the averages, i.e., for $\hat{M}_{k,-i} = \sum_{j\neq i}\hat{M}_{ki}/(n-1)$

$$(3.11)\qquad \hat{\gamma}_{-i} = -[\sum_{j\neq i}\hat{U}_j'\hat{Q}\hat{U}_j/(n-1)]^{-1}\cdot$$

$$\{(\mathrm{tr}[\hat{Q}\hat{M}_{1,-i}'],\ldots,\mathrm{tr}[\hat{Q}\hat{M}_{K,-i}'])' + \sum_{j\neq i}\hat{U}_j'\hat{Q}\hat{u}_{0j}/(n-1)\}.$$

Replacing in equation (3.10) the expectations with the i^{th} observation, Q by $\hat{Q}$, γ by $\hat{\gamma}_{-i}$, and averaging gives,

$$(3.12)\qquad CV(K) =$$

$$= \sum_{i=1}^{n}\{2[\mathrm{trace}(\hat{Q}\hat{M}_{0i}'+\sum_{k=1}^{K}\hat{Q}\hat{M}_{ki}'\hat{\gamma}_{k,-i})] + (\hat{u}_{0i}+\hat{U}_i\hat{\gamma}_{-i})'\hat{Q}(\hat{u}_{0i}+\hat{U}_i\hat{\gamma}_{-i})\}/n.$$

One cross-validation method chooses K to minimize $CV(K)$.

The formal motivation for this procedure is similar to that of cross-validation in other nonparametric regression contexts. By dropping the i^{th} observation in computation of the linear combination coefficients one source of bias in the distance estimator is removed. It would be interesting to give a precise justification for this procedure, say, by showing that it can guarantee that the distribution of the resulting estimator of parameters of interest converges to its limit as fast as possible. Such a result is beyond the scope of this paper, although the following general reasoning suggests that it is plausible: Under regularity conditions, the rate of convergence of the nonparametric estimator $[(1,\hat{\gamma}')\otimes I_q]\hat{m}(z,\hat{\alpha}(z,\beta))$ to the efficient score S, which should be less than $\sqrt{n}$, should be the determining factor, since the bound V is a parameter that should be $\sqrt{n}$-consistently estimated by $(\hat{M}_0+\sum_{k=1}^{K}\hat{\gamma}_k\hat{M}_k)^{-1}$. Also, it is plausible that cross-validation could give the best rate.

One could also use the bootstrap to choose K, in a way analogous to the

window width choice suggested by Hsieh and Manski (1987), by minimizing some measure of the magnitude of a bootstrap estimate of the variance matrix of $\hat{\beta}$. This procedures appears to be computationally more expensive than the one suggested above. Also, the conjectures in the previous paragraph suggest that cross-validation on the efficient score estimate might be closely related to the properties of $\hat{\beta}$, as is the bootstrap variance estimate, and as turns out to be the case in the Monte-Carlo example.

3.4 Examples

In each of the examples, $m(z,\beta,\alpha)$ as discussed in Section 2 can be combined to form estimators. In the nonlinear simultaneous equations example, quantities required for the one-step estimator can be calculated as

$$(3.13)\qquad m_k(z,\beta,\hat{\alpha}(z,\beta)) = \tilde{S}_k(z,\beta) - \sum_{j=1}^{n}\tilde{S}_k(\pi(x_j,\rho(z,\beta),\beta),x_j,\beta)/n,$$

$$\hat{M}_{ki} = \partial m_k(z_i,\beta,\hat{\alpha}(z_i,\beta))/\partial\beta|_{\hat{\beta}},\quad \hat{M}_k = \sum_{i=1}^{n}\hat{M}_{ki}/n,$$

$$\hat{u}_{ki} = \tilde{S}^k_{ii} - \sum_{j=1}^{n}\tilde{S}^k_{ij}/n - \sum_{j=1}^{n}\tilde{S}^k_{ji}/n + \sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{S}^k_{ij}/n^2,$$

where $\hat{\beta}$ is a preliminary estimator and $\tilde{S}^k_{ij} = \tilde{S}_k(\pi(x_j,\rho(z_i,\hat{\beta}),\hat{\beta}),x_j,\hat{\beta})$.

The functions $\tilde{S}_k(z,\beta)$ can be chosen to be either of the two types discussed in Section 2. The first type, where $\tilde{S}_k(z,\beta) = A_k(x)r_k(\hat{\Sigma}^{-1/2}[\rho(z,\beta)-\hat{\mu}])$ for $\{A_k(x),r_k(\rho))$ and preliminary location and scale estimates $\hat{\Sigma}$ and $\hat{\mu}$, has the advantage that the reduced form $\pi(\varepsilon,x,\beta)$ is not required for its computation, but the disadvantage that it does not use much of the model structure. This disadvantage will be reflected in the fact that approximation of the efficient score requires approximation in both the ε and x dimension. The second type, with

$$(3.14)\qquad \tilde{S}_0(z,\beta) = J_\beta(z,\beta),$$

$$\tilde{S}_k(z,\beta) = \rho_\beta(z,\beta)'\hat{\Sigma}^{-1/2}\tilde{s}_k(\hat{\Sigma}^{-1/2}[\rho(z,\beta)-\hat{\mu}]),$$

requires $\pi(\varepsilon,x,\beta)$, but has the advantage of imposing enough structure that only approximation in the ε dimension will be required for efficiency.

Each of these types of estimators depend on certain functions, A_k and r_k in the first case and $\tilde{s}_k$ in the second. One possible choice of these functions are as basis sequences, such as polynomials. To avoid conditions on the existence of higher order moments, one could use polynomials in some bounded, one-to-one functions of the components. To be specific, let $\tau(\cdot)$ denote a one-to-one, smooth univariate function, such as $\tau(\cdot) = \exp(\cdot)/[1+\exp(\cdot)]$. For a vector ζ with $\dim(\zeta)$ components, let $p_0(\zeta) = (\tau(\zeta_1),\ldots,\tau(\zeta_{\dim(\zeta)}))'$, $\lambda(\ell) = (\lambda_1(\ell),\ldots,\lambda_{\dim(\zeta)}(\ell))$, $(\ell=1,2,\ldots)$, be vectors of nonnegative numbers, and $p_0(\zeta)^{\lambda(\ell)} = \prod_{j=1}^{\dim(\zeta)}[\tau(\zeta_j)]^{\lambda_j(\ell)}$. For the second type of estimator one could choose

$$(3.15)\qquad \tilde{s}_k(\rho) = e_{k-(s[k/s])}p_0(\rho)^{\lambda([k/s+1])},\quad (k = 1, 2, \ldots),$$

where e_i is the s-dimensional unit vector with 1 in the i^{th} position and zeros elsewhere, and [a] denotes the largest integer less than or equal to a. The unit vector e_i is present in order to make possible the approximation of each element of the vector $s(\varepsilon)$. One could choose $A_k(x)r_k(\rho)$ similarly, with ρ replaced by $(x',\rho')'$ and s by q. In either case, if $\{\lambda(\ell)\}$ is chosen to so that $\{p_0(\zeta)^{\lambda(\ell)}\}$ forms a complete sequence, in metrics discussed in the next Section, the resulting estimator could be approximately efficient.

Such a power series works well in the Monte Carlo example below, where ε is a scalar. However, when ε has several components one could get a large

number of terms, even with small polynomial orders, which should adversely affect finite sample performance. An alternative procedure, that is more parsimonious, is to choose $\tilde{s}_k(\rho)$ from among a small number of location scores for different possible distributions for ε. For example, one might choose $\tilde{s}_k(\rho) = \xi_k(\rho'\hat{\Sigma}^{-1}\rho)\hat{\Sigma}^{-1}\rho$, corresponding to densities depending only on $\rho'\hat{\Sigma}^{-1}\rho$, with $\xi_1 = 1$ corresponding to the normal distribution and higher orders corresponding to thick-tailed and asymmetric alternatives to the normal. With such an approach one could only hope to be approximately efficient if the true density depended only on $\rho'\Sigma^{-1}\rho$, but with typical data set sizes and multivariate systems, efficiency for such a reduced dimension class of densities may be all one can hope for.

In the CDR example, the quantities required for the one-step estimator can be calculated as

$$(3.16)\qquad m_k(z,\beta,\hat{\alpha}(z,\beta)) = [v_\beta(x,\beta) - \hat{\alpha}(x,\beta)]\tilde{s}_k(y,v(x,\beta)),$$

$$\hat{M}_{ki} = \partial m_k(z_i,\beta,\hat{\alpha}(z_i,\beta))/\partial\beta|_{\hat{\beta}}, \qquad \hat{M}_k = \textstyle\sum_{i=1}^{n}\hat{M}_{ki}/n,$$

$$\hat{u}_{ki} = [v_\beta(x_i,\hat{\beta}) - \hat{\alpha}_i]\{\tilde{s}_k(y_i,\hat{v}_i) - \hat{E}[\tilde{s}_k|\hat{v}_i]\},$$

where $\hat{\alpha}_i \equiv \hat{\alpha}(x_i,\hat{\beta})$, $\hat{v}_i \equiv v(x_i,\hat{\beta})$, $m_k(z_i,\beta,\hat{\alpha}(z_i,\beta))$ is assumed to be differentiable, and $\hat{E}[\tilde{s}_k|v]$ is a nonparametric regression of $\tilde{s}_k(y_i,\hat{v}_i))$ on $\hat{v}_i$ evaluated at v.

One could choose $\tilde{s}_0 = 0$ and $\tilde{s}_k$ a power series for $k \geq 1$. For example, if y is confined to the nonnegative integers, with small values of y most likely, then one might choose to use nonnegative integer powers of $\tau(v)$ and fractional integer powers of $\tau(y)$ in such a power series. In any such power series, terms that depend only on v should be excluded, because the population residual $\tilde{s}_k(y,v)-E[\tilde{s}_k|v]$ is identically zero, causing a

potential singularity in the calculation of $\hat{\gamma}$.

It is also possible to use more of the structure of the model in the specification of $\tilde{s}_k$. An example is $\tilde{s}_k(y,v) = \partial \ln \tilde{f}(y|v)/\partial v$ for fixed $\tilde{f}(y|v)$ among possible conditional densities for y. Also, if y has separate discrete and continuous components, one might treat them differently by specifying $\tilde{s}_k(y,v)$ to be nonzero for only one of the components.

In the mean transformation example, the quantities required for the one-step estimator can be calculated as

$$(3.17) \qquad m_k(z,\beta,\hat{\alpha}(z,\beta)) = [A_k(x) - \hat{\alpha}_k(x,\beta)]y,$$

$$\hat{M}_{ki} = \partial m_k(z_i,\beta,\hat{\alpha}(z_i,\beta))/\partial\beta|_{\hat{\beta}}, \qquad \hat{M}_k = \textstyle\sum_{i=1}^{n} \hat{M}_{ki}/n,$$

$$\hat{u}_{ki} = [A_k(x_i) - \hat{\alpha}_{ki}]\{y_i - \hat{E}[y|\hat{v}_i]\},$$

where $\hat{\alpha}_k(x,\beta)$ is the nonparametric regression estimate of $A(x)$ on $v(x,\beta)$, assumed to be differentiable, $\hat{\alpha}_{ki} = \hat{\alpha}_k(x_i,\hat{\beta})$, $\hat{v}_i \equiv v(x_i,\hat{\beta})$, and $\hat{E}[y|v]$ is a nonparametric regression of y_i on $\hat{v}_i$ evaluated at v.

One could choose $A_k(x)$ as in equation (3.15), with ρ replaced by x and s by q. Any basis sequence that is complete for $q \times 1$ vector functions of x in a metric discussed below will yield an approximately efficient estimator. One could be more parsimonious by taking $A_k(x) = v_\beta(x,\beta)\cdot p_\ell(v(x,\beta))$, which would lead to efficiency if the conditional variance of y given x depended only on v.

4. The Spanning Condition

The fundamental necessary condition for combining moment functions to lead to efficiency is that finite linear combinations of corresponding $\{u_k\}$ can approximate the efficient score S arbitrarily well in mean square. A precise statement is

Spanning Condition: $S \in \mathcal{U}$,

$$\mathcal{U} \equiv \{u : \exists \{\gamma_{kK}\}_{k=1,K=1}^{K\ \ \infty} \text{ with } \lim_{K\to\infty} E[\|u-u_0-\textstyle\sum_{k=1}^{K}\gamma_{kK}u_k\|^2] = 0\}$$

There are useful geometric, dual sufficient conditions for spanning when $u_0 = 0$. Let $\mathcal{S} = \{a: E[a] = 0\}$ and let $\mathcal{T}^\perp$ and $\mathcal{U}^\perp$ be the orthogonal complements of $\mathcal{T}$ and $\mathcal{U}$ in $\mathcal{S}$ for the inner product $E[a_1'a_2]$, respectively, e.g. $\mathcal{T}^\perp = \{a \in \mathcal{S}: E[a't] = 0, \forall\ t \in \mathcal{T}\}$.

Theorem 4.1: Suppose that $u_0 = 0$. *If* $\mathcal{U} = \mathcal{T}^\perp$ or $\mathcal{U}^\perp = \mathcal{T}$ *then the spanning condition is satisfied.*

Proof: Note that since $\mathcal{U}$ and $\mathcal{T}$ are mean square closed by definition, $\mathcal{U} = \mathcal{T}^\perp$ if and only if $\mathcal{U}^\perp = \mathcal{T}$. Then recall from Section 2.1 that $S \in \mathcal{T}^\perp$, so that the conclusion that $\mathcal{U} = \mathcal{T}^\perp$ implies spanning is trivial. ∎

The second condition $\mathcal{U}^\perp = \mathcal{T}$ can be interpreted as a condition that the moment restrictions used in estimation characterize the semiparametric model. There is a precise interpretation for the case where α is not present, where $m_k = u_k$ for all k. Consider a smooth likelihood $\ell(z|\eta)$, passing through the truth at η_0, with score S_η at η_0, but not necessarily obeying the restrictions of the semiparametric model, and let $E_\eta[\cdot]$ denote the expectation with respect to this density. Here "moment conditions characterize the semiparametric model" will be taken to mean that $E_\eta[m_k] =$

0 in a neighborhood of η_0 for each k implies $S_\eta \in \mathcal{T}$.

Theorem 4.2: Suppose that $u_0 = 0$, $m_k = u_k$ $(k \geq 1)$, *bounded functions are dense in* $\mathcal{U}^\perp$, *and limit attention to parametric families where* $E_\eta[m_k' m_k]$ *is bounded in a neighborhood of* η_0. *The moment conditions characterize the semiparametric model if and only if* $\mathcal{U}^\perp = \mathcal{T}$.

Proof: Note that $\mathcal{U}^\perp \supseteq \mathcal{T}$ follows by $\mathcal{U} \subseteq \mathcal{T}^\perp$ (see Section 2), so that here $\mathcal{U}^\perp = \mathcal{T}$ can be taken to be equivalent to $\mathcal{U}^\perp \subseteq \mathcal{T}$. For the "only if" part, let B be the dense set in the hypotheses. For $b(z) \in B$, let $f(z|\eta) = f_0(z)[1+\eta' b(z)]$. It is straightforward to check that $f(z|\eta)$ is smooth with $S_\eta = b(z)$ (e.g. see BKRW). Then if $b(z) \in \mathcal{U}^\perp$, $E_\eta[m_k] = E[m_k] + E[m_k b(z)']\eta = 0$, so that $b(z) \in \mathcal{T}$ follows by the moment characterization hypothesis. Thus $B \subseteq \mathcal{T}$, implying $\mathcal{U}^\perp \subseteq \mathcal{T}$ by B dense in $\mathcal{U}^\perp$ and $\mathcal{T}$ closed. For the "if" part, suppose that the parametric family satisfies $E_\eta[m_k] = 0$. By differentiation and Ibragimov and Hasminski, 1981, Lemma 7.2, $E[m_k S_\eta'] = 0$ so $S_\eta \in \mathcal{U}^\perp \subseteq \mathcal{T}$, giving moment characterization. ∎

This result may not be useful for verifying spanning when the denseness hypothesis is difficult to check.

In general, where α may be present, $\mathcal{U}^\perp = \mathcal{T}$ is simply a local version of this characterization statement. If one thinks of $a \in \mathcal{A}$ as indexing a direction of departure from the truth, and $a \in \mathcal{U}^\perp$ as the directional analog of the moment condition then $\mathcal{U}^\perp = \mathcal{T}$ means that the only directions of departure allowed by the moment conditions are those allowed by model.

When restrictions are imposed on the form of the moment functions, these results may not be useful for checking the spanning condition. There is another sufficient condition that is often useful.

Theorem 4.3: If there is a set of random vectors $\{a_\ell\}_{\ell=0}^{\infty}$ *such that* $\{a_\ell - E[a_\ell] - \text{Proj}(a_\ell|\mathcal{T})\} \subset \{0\} \cup \{u_k\}$ *and there exists* $\{\gamma_{\ell L}\}$ *with* $\lim_{L\to\infty} E[\|S_\beta - a_0 - \sum_{\ell=1}^{L} \gamma_{\ell L} a_\ell\|^2] = 0$ *then the spanning condition holds.*

Proof: Consider γ_{kK} such that for $\epsilon_K = S_\beta - a_0 - \sum_{k=1}^{K} \gamma_{kK} a_k$, $E[\|\epsilon_K\|^2] \longrightarrow 0$. Let $\bar{a}_\ell = a_\ell - E[a_\ell] - \text{Proj}(a_\ell|\mathcal{T})$. By linearity of projections and $E[S_\beta] = 0$, $S - \bar{a}_K - \sum_{k=1}^{K} \gamma_{kK} \bar{a}_k = \epsilon_K - E[\epsilon_k] - \text{Proj}(\epsilon_K|\mathcal{T})$, so that by $E[\|\text{Proj}(\circ|\mathcal{T})\|^2] \leq E[\|\circ\|^2]$ and $\|E[\epsilon_k]\|^2 \leq E[\|\epsilon_k\|^2]$,

$$(4.1) \qquad E[\|S - \bar{a}_0 - \sum_{k=1}^{K} \gamma_{kK} \bar{a}_k\|^2] = E[\|\epsilon_K - E[\epsilon_k] - \text{Proj}(\epsilon_K|\mathcal{T})\|^2]$$

$$\leq 3(E[\|\epsilon_K\|^2] + \|E[\epsilon_k]\|^2 + E[\|\text{Proj}(\epsilon_K|\mathcal{T})\|^2]) \leq 9E[\|\epsilon_K\|^2] \longrightarrow 0.$$

Thus, S is in the closed linear span of $\{\bar{a}_\ell\}$. The conclusion then follows by the inclusion hypothesis $\{\bar{a}_\ell\} \subset \{0\} \cup \{u_k\}$, implying that the closed linear span of $\{\bar{a}_\ell\}$ is a subset of $\mathcal{U}$. ∎

This result is useful because S and the influence functions u_k are often more complicated objects than S_β and a_ℓ, when α is present, making it easier to check the approximation hypothesis for S_β. The first hypothesis allows $\{a_\ell\}$ to include some values where $a_\ell - E[a_\ell] - \text{Proj}(a_\ell|\mathcal{T}) = 0$, e.g. a_ℓ a constant vector. Such values can ease the task of verifying the approximation of S_β, but are important to exclude from the estimation procedure to avoid singularity in the calculation of the linear combination coefficients $\hat{\gamma}$.

The hypotheses state that included in $\{u_k\}$ are objects of the form $a_\ell - E[a_\ell] - \text{Proj}(a_\ell|\mathcal{T})$, which has an interesting interpretation. For such u_k, if $m_k = a_\ell - E[a_\ell]$, then one has the efficient-$\hat{\alpha}$ case of Section 2, where $u_k - m_k = -\text{Proj}(m_k|\mathcal{T})$. Theorem 4.3 therefore says that in efficient-$\hat{\alpha}$ cases, spanning holds if $\{a_\ell\}$ can approximate S_β. It is also interesting to note

that the efficient-$\hat{\alpha}$ case is not necessary for the spanning condition, as shown for the conditional mean transformation model below.

Theorem 4.3 is useful for checking the spanning condition in the first two examples. Let a sequence $\{b_\ell(z)\}$ be complete with respect to a distribution $P(z)$ if for any $q(z)$ with $\int\|q(z)\|^2dP(z) < \infty$ there exists $\{\gamma_{\ell L}\}$ with $\lim_{L\to\infty}\int\|q(z)-\sum_{\ell=1}^{L}\gamma_{\ell L}b_\ell(z)\|^2dP(z) = 0$. Also, let $\{e_i\}_{i=1}^{d}$ denote the set of all unit vectors with dimension d.

Corollary 4.4: For the nonlinear simultaneous equations model, the spanning condition is satisfied if $m_k(z,\beta,\hat{\alpha}(z,\beta))$ *is specified as in equation (3.4.1), with either i)* $\{\tilde{S}_k(z,\beta)\} = \{A_k(x)r_k(\rho(z,\beta))\}$ *and* $\{A_k(x)r_k(\varepsilon)\} \cup \{r_k(\varepsilon)\} \cup \{A_k(x)\} \cup \{e_i\}_{i=1}^{s}$ *is complete with respect to the distribution of* (ε,x)*; ii)* $\{\tilde{S}_k(z,\beta)\} = \{J_\beta(z,\beta)\} \cup \{\rho_\beta(z,\beta)'\tilde{s}_k(\rho(z,\beta))\}$, $0 < E[\|\rho_\beta\|^2] < \infty$, *and either* a) $\{\tilde{s}_k(\varepsilon)\}$ is complete *with respect to the distribution for* ε *with* $\text{Prob}(\varepsilon\in A) = E[\|\rho_\beta\|^2 1(A)]/E[\|\rho_\beta\|^2]$ *or* b) $\rho_\beta-E[\rho_\beta|\varepsilon]-E[\rho_\beta|x]+E[\rho_\beta] = 0$ *and* $\{\tilde{s}_k(\varepsilon)\} \cup \{e_i\}_{i=1}^{s}$ *is complete with respect to this distribution.*

Proof: By equation (2.4), $u_k = \tilde{S}_k-E[\tilde{S}_k|\varepsilon]-E[\tilde{S}_k|x]+E[\tilde{S}_k] = \tilde{S}_k-E[\tilde{S}_k]-\text{Proj}(\tilde{S}_k|\mathcal{T})$. In case i), let $\{a_\ell\} = \{A_k(x)r_k(\varepsilon)\} \cup \{r_k(\varepsilon)\} \cup \{A_k(x)\} \cup \{e_i\}$, and note that the first hypothesis of Theorem 4.3 is satisfied by $r_k(\varepsilon)-E[r_k|\varepsilon]-E[r_k|x]+E[r_k] = A_k(x)-E[A_k|\varepsilon]-E[A_k|x]+E[A_k] = e_i-E[e_i|\varepsilon]-E[e_i|x]+E[e_i] = 0$, while the second hypothesis holds by completeness. In case ii), let $\{a_\ell\} = \{\tilde{S}_k(z,\beta_0)\}$ in subcase a), and $\{a_\ell\} = \{\tilde{S}_k(z,\beta_0)\} \cup \{\rho_\beta'e_i\}$ in subcase b). The first hypothesis of Theorem 4.3 holds in both subcases, by $\rho_\beta-E[\rho_\beta|\varepsilon]-E[\rho_\beta|x]+E[\rho_\beta] = 0$ in the subcase b). The second hypothesis follows by $E[\|S_\beta-a_0-\sum_{\ell=1}^{L}\gamma_{\ell L}a_\ell\|^2] = E[\|\rho_\beta'[s(\varepsilon)-\sum_{\ell=1}^{L}\gamma_{\ell L}\tilde{s}_\ell(\varepsilon)\|^2] \le E[\|\rho_\beta\|^2\|s(\varepsilon)-\sum_{\ell=1}^{L}\gamma_{\ell L}\tilde{s}_\ell(\varepsilon)\|^2]$ and completeness. ∎

Note the dimensionality reduction for the second type of $\tilde{S}_k(z,\beta)$ is evident

in this result. The second type only requires approximation of functions of ε, while the first requires approximation of functions of both ε and x.

Corollary 4.5: For the CDR model, the spanning condition is satisfied if $m_k(z,\beta,\hat{\alpha}(z,\beta))$ *is specified as in equation (3.4.4),* $0 < E[\|v_\beta\|^2] < \infty$, *and there is* $\{b_k(v)\}$ *such that* $\{\tilde{s}_k(y,v)\} \cup \{b_k(v)\}$ *is complete with respect to the distribution for* (y,v) *with* $\text{Prob}((y,v)' \in A) = E[\|v_\beta\|^2 1(A)]/E[\|v_\beta\|^2]$.

Proof: Let $\{a_\ell\} = \{v_\beta \tilde{s}_k(y,v)\} \cup \{v_\beta b_k(v)\}$. For $a_\ell = v_\beta \tilde{s}_k(y,v)$, $u_k = v_\beta \tilde{s}_k - E[v_\beta \tilde{s}_k] - \{E[v_\beta|v]\tilde{s}_k - E[v_\beta|v]E[\tilde{s}_k|v] + v_\beta E[\tilde{s}_k|v] - E[v_\beta \tilde{s}_k]\} = a_\ell - E[a_\ell] - \text{Proj}(a_\ell|\mathcal{T})$, while for $a_\ell = v_\beta b_k(v)$, $a_\ell - E[a_\ell] - \text{Proj}(a_\ell|\mathcal{T}) = v_\beta b_k(v) - E[v_\beta|y,v]b_k(v) + E[v_\beta|v]b_k(v) - v_\beta b_k(v) = 0$, so that the first hypothesis of Theorem 4.3 is satisfied. To see that the second hypothesis holds, note that for $\{d_\ell(y,v)\} = \{\tilde{s}_k(y,v)\} \cup \{b_k(v)\}$, $E[\|S_\beta - a_0 - \sum_{\ell=1}^{L} \gamma_{\ell L} a_\ell\|^2] = E[\|v_\beta\{s(y,v) - \sum_{\ell=1}^{L} \gamma_{\ell L} d_\ell\}\|^2] \leq E[\|v_\beta\|^2 \|\{s(y,v) - \sum_{\ell=1}^{L} \gamma_{\ell L} d_\ell\}\|^2] \longrightarrow 0$ by completeness. ∎

More primitive conditions for completeness are readily available for the power series choices for $\{\tilde{s}_k(\varepsilon)\}$, $\{A_k(x)r_k(\varepsilon)\}$, and $\{\tilde{s}_k(y,v)\}$ discussed in Section 3; see equation (3.15). Suppose that in each case, the coefficient vector sequence $(\lambda(\ell))_{\ell=1}^{\infty}$ is selected from the nonnegative integers and ordered according to $\sum_{j=1}^{\dim(\lambda)} \lambda_j(\ell)$. Then from Gallant (1980, Theorem 3) and Newey (1988a, Theorem 3.1) it follows that in each case a sufficient condition for completeness is the existence of the moment generating function of $p_0(\zeta)$, where ζ equals $\Sigma^{-1/2}(\varepsilon - \mu)$, $(x, \Sigma^{-1/2}(\varepsilon - \mu))$, or (y,v) respectively, for the distributions that appear in the respective completeness conditions. This condition is automatically satisfied if p_0 is bounded. For example, a sufficient condition for ii) a) of Corollary 4.3 is existence of some $C > 0$ such that

(4.2) $\quad E[\exp\{C\|p_0(\Sigma^{-1/2}(\varepsilon-\mu))\|\}] < \infty.$

In the mean transformation model the first hypothesis of Theorem 4.3 is not satisfied, i.e. $u_k \neq m_k - E[m_k] - \text{Proj}(m_k|\mathcal{T})$, except in the special case where $\text{Var}(y|x)$ depends only on v. An intuitive explanation for this feature is that $\hat{E}[y|v]$ is pointwise inefficient as an estimator of $E[y|x]$ under the conditional mean transformation model. It is easy to show that $\hat{E}[w(x)y|v]/\hat{E}[w(x)|v]$ with $w(x) = 1/\text{Var}(y|x)$ is more efficient at a particular value of v.

Despite the fact that this is not an efficient-$\hat{\alpha}$ case, the spanning condition can be satisfied.

Theorem 4.6: For the conditional mean transformation model, the spanning condition is satisfied if $m_k(z,\beta,\hat{\alpha}(z,\beta))$ *is specified as in equation (3.4.5),* $E[y^2] < \infty$, $\text{Var}(y|x) > 0$, *and there is* $\{b_k(v)\}$ *such that* $\{A_k(x)\} \cup \{b_k(v)\}$ *is complete with respect to the distribution for* x *given by* $\text{Prob}(x\in A) = E[\omega(x)1(A)]/E[\omega(x)]$.

Proof: The conclusion will follow from Theorem 4.1 by showing $\mathcal{U}^\perp \subseteq \mathcal{T}$ (as noted above $\mathcal{U}^\perp \supseteq \mathcal{T}$ holds automatically). Consider $t \in \mathcal{U}^\perp$, implying that for $\eta = y - E[y|x] = y - E[y|v]$,

(4.3) $\quad 0 = E[\{A_k - E[A_k|v]\}'\eta t] = E[\{A_k - E[A_k|v]\}E[\eta\{t - E[t|x]\}|x]]$

$$= E[\omega(x)^{1/2}\{A_k - E[A_k|v]\}'r(x)], \quad r(x) \equiv E[\omega(x)^{-1/2}\eta\{t - E[t|x]\}|x].$$

By the Cauchy-Schwarz inequality, $\|r(x)\|^2 \leq \omega(x)^{-1}E[\eta^2|x]\cdot E[\|t - E[t|x]\|^2|x] \leq E[\|t\|^2|x]$, implying $E[\|r(x)\|^2] < \infty$. Then for $\{a_\ell(x)\} = \{A_{k_\ell}(x)\} \cup \{b_{k_\ell}(v)\}$, by $E[\omega(x)\{\omega(x)^{-1/2}r(x)\}^2] < \infty$ and the completeness condition, there exists $\gamma_{\ell L}$ such that $E[\omega(x)\|\omega(x)^{-1/2}r(x) - \sum_{\ell=1}^{L}\gamma_{\ell L}a_\ell(x)\|^2] =$

$E[\|r(x)-\sum_{\ell=1}^{L}\gamma_{\ell L}\omega(x)^{1/2}a_\ell(x)\|^2] \longrightarrow 0$, so that by the Cauchy-Schwartz inequality, equation (6.3), and a_ℓ a function only of v if $a_\ell \neq A_k$ for some A,

$$(4.4)\qquad 0 = \sum_{\ell=1}^{L}\gamma_{\ell L}E[\omega(x)^{1/2}\{a_\ell-E[a_\ell|v]\}'r(x)]$$

$$= E[\{\sum_{\ell=1}^{L}\gamma_{\ell L}\omega(x)^{1/2}a_\ell-E[\sum_{\ell=1}^{L}\gamma_{\ell L}\omega(x)^{1/2}a_\ell|v]\}r(x)]$$

$$\longrightarrow E[\|r(x)-E[r(x)|v]\|^2],$$

Thus, $r(x)$ is a function of only v, which implies $t \in \mathcal{T}$, by the form of the tangent set in equation (2.8). ∎

As for the other examples, a primitive completeness condition when $A_k(x)$ consists of a power series is that $E[\exp\{C\|p_0(x)\|\}] < \infty$ for some $C > 0$.

5. Asymptotic Efficiency

When the spanning condition is satisfied, the moment estimator will be close to being efficient for K large enough, an approximate efficiency result. Full asymptotic efficiency requires a specification of a rate of growth of K with the sample size so that the bound is achieved asymptotically. This section provides such conditions.

The following result is a useful intermediate one. Let $\lambda(B)$ denote the smallest eigenvalue for a symmetric matrix B and

$$\mathcal{M} \equiv [M_0', \ldots, M_K']', \quad \hat{\mathcal{M}} \equiv [\hat{M}_0', \ldots, \hat{M}_K']', \quad \hat{L} \equiv (1, \hat{\gamma})' \otimes I_q, \quad \hat{u}_i \equiv [\hat{u}_{0i}', \ldots, \hat{u}_{Ki}']'.$$

Lemma 5.1: *Suppose that* $\sqrt{n}\|\hat{\beta}-\beta\| = O_p(1)$, *and there exists* ϵ, Δ, *and* $\nu = \nu(K)$ *such that for nonrandom* $K = K(n) \longrightarrow \infty$: i) Q *is positive definite,* $K = O(\nu^{\Delta\nu})$, $\|\mathcal{M}\| = O(\nu^{\Delta\nu})$, $E[\|u\|^{2+\epsilon}] = O(\nu^{\Delta\nu})$, $1/\lambda(E[U'U]) = O(\nu^{\Delta\nu})$; ii) $\|\hat{Q}-Q\| = O_p(n^{-\epsilon})$, $\|\hat{\mathcal{M}}-\mathcal{M}\| = O_p(n^{-\epsilon}\nu^{\Delta\nu})$, $\sum_{i=1}^n \|\hat{u}_i-u_i\|^2/n = O_p(n^{-\epsilon}\nu^{\Delta\nu})$, $\sqrt{n}\|\hat{m}_n(\hat{\beta})-\hat{m}_n(\beta_0)-\mathcal{M}(\hat{\beta}-\beta_0)\| = O_p(n^{-\epsilon}\nu^{\Delta\nu})$, $\sqrt{n}\|\hat{m}_n(\beta_0)-\sum_{i=1}^n u_i/n\| = O_p(n^{-\epsilon}\nu^{\Delta\nu})$; iii) $\mathcal{M} = -E[mS']$ *for the efficient score* S; iv) *There exists* $\tilde{\gamma}^K$ *such that* $E[\|S-((1,\tilde{\gamma}^{K\prime})\otimes I_q)u\|^2] = o(1)$; v) K *is chosen to be* $\hat{K}$, *a (possibly) random function of sample size, such that* $\hat{K} \xrightarrow{p} \infty$ *and there exists nonrandom* K_n *such that* $\hat{K} \leq K_n$ *with probability approaching one, and* $\nu(K_n)\ln[\nu(K_n)]/\ln(n) \longrightarrow 0$; vi) *Either* $\hat{K} \in \mathcal{K}_n$ *with probability approaching one and the number of elements of* $\mathcal{K}_n$ *is bounded, or* $\sum_{K=1}^{\infty}\{E[\|S-((1,\tilde{\gamma}^{K\prime})\otimes I_q)u\|^2]\}^{1/2}$ *is finite. Then* $\tilde{\beta}$ *is well defined with probability approaching one and*

$$(5.1) \qquad \sqrt{n}(\tilde{\beta}-\beta_0) \xrightarrow{d} N(0,V),\ (\hat{L}\hat{\mathcal{M}})^{-1}\hat{L}(\sum_{i=1}^n \hat{u}_i\hat{u}_i'/n)\hat{L}'(\hat{L}\hat{\mathcal{M}})^{-1\prime} \xrightarrow{p} V.$$

Furthermore, if for every regular parametric submodel $E_\theta[\|S\|^2]$ *is continucus in a neighborhood of* θ *then* $\tilde{\beta}$ *is regular.*

This result specifies an upper bound on the growth rate for K that gives efficiency under certain boundedness and rate of convergence conditions. These conditions are not very primitive, which is to be expected, since they will depend on specifics of particular nonparametric estimation methods for the nuisance functions α. When the nuisance function is finite dimensional, it is possible to give more primitive conditions; see Newey (1989c). In any case, this Lemma helps to simplify the task of verifying that a particular estimator is asymptotically efficient, as in the nonlinear simultaneous equations example below.

The ν term in this result is an index for the order of the moment functions, such as the maximum included power for power series. Indeed,

the conditions of Lemma 5.1 are designed so as to apply as easily as possible to power series. In particular, one of the more stringent and difficult to check conditions is the eigenvalue bound in i). One result that is useful for power series is the eigenvalue bound of Newey (1988a), which requires that the distribution have a continuous component. A lower bound on the eigenvalue of a second moment matrix of such a multivariate power series that includes powers of individual components up to ν is $c\nu^{-\Delta\nu}$ for constants c and Δ, motivating the form of the bounds and rate conditions given in Lemma 5.1.

An important feature of this result is that K can be data based. This feature allows for K to be chosen, say, by the cross-validation method discussed earlier. However, it should be noted that an approximation rate hypothesis is imposed in the more interesting case where K is allowed to vary freely between upper and lower bounds, requiring that the remainder in the spanning condition go to zero fast enough that the infinite sum in condition v) is finite. A literature search has not yet revealed useful primitive conditions for this approximation rate, except under restrictive conditions such as boundedness of the random variables of power series and continuous differentiability of the efficient score in these variables (e.g. see Powell, 1981).

Lemma 5.1 is useful for showing efficiency of the estimators discussed in the examples. For brevity, only the simultaneous equations example will be discussed here. In particular, the second type of estimator will be considered, where the moment functions are formed as in equations (3.13) and (3.14). To give primitive regularity conditions, it is necessary to specify the sequence $\{\tilde{s}_K(\rho)\}$ of approximating functions. Here power series in bounded, monotonic functions will be considered, as in:

Assumption 5.1: $\{\tilde{s}_k(\rho)\}$ satisfies equation (3.15), where $p_0(\rho) = (\tau(\rho_1),\ldots,\tau(\rho_s))$, $\tau(\rho)$ is one-to-one with positive, bounded derivative and bounded second derivative, and $\lambda(\ell)$ are distinct vectors. Also, for $\nu(K) = 1 + \max_{m\leq[K/s]+1}\sum_{\ell=1}^{s}\lambda_\ell(m)$ and $\underline{\nu}(K)$ equal to the maximum integer $\underline{\nu}$ such that $\{p_0(\rho)^\lambda:\sum_{\ell=1}^{s}\lambda_\ell \leq \underline{\nu}\} \subseteq \{p_0(\rho)^{\lambda(m)}\}_{m=1}^{[K/s]+1}$, $\underline{\nu}(K) \longrightarrow \infty$ and there is $C > 0$ such that $K = O(\nu(K)^{C\nu(K)})$.

This assumption states that $\{s_k(\rho)\}$ is a power series in a function $\tau(\cdot)$, such as $\tau(\cdot) = \exp(\cdot)/[1+\exp(\cdot)]$, such that all possible nonnegative integer powers are included, and that the highest order included term does not grow too fast relative to the number of terms. If the power series terms are ordered in the natural way, with all terms of a given order included before the next highest order, then $K \leq s\nu(K)^s$, which satisfies the assumption.

The next assumption imposes sufficient conditions for the eigenvalue hypothesis in Lemma 5.1 i).

Assumption 5.2: The density of ε is bounded away from zero on an open set and either 1) $\rho_\beta - E[\rho_\beta|\varepsilon]$ is a function only of x with $E[\{\rho_\beta - E[\rho_\beta|\varepsilon]\}\{\rho_\beta - E[\rho_\beta|\varepsilon]\}']$ nonsingular, and $\{e_i\}_{i=1}^{s}$ excluded from $\{\tilde{s}_k(\rho)\}$; or 2) there are vectors $\rho_{\ell\beta}$, $(\ell=1,\ldots,s)$, such that $\rho_\beta = \text{diag}[\rho'_{1\beta},\ldots,\rho'_{s\beta}]$, for each ℓ, some element of $\rho_{\ell\beta}$ is not multiplicatively separable in a function of x and ε, is bounded, and has conditional variance given ε that is bounded away from zero.

This assumption is somewhat restrictive. The leading examples where 1) is satisfied are models where y enters linearly and has constant coefficients,

$$(5.2) \qquad \rho(z,\beta) = B(\beta)y - f(x,\beta), \quad B_0 = B(\beta_0) \text{ nonsingular},$$

for which $\rho_\beta(z,\beta_0)$ is linear in ε with constant coefficients. Part 2)

requires that each element of β enter only one residual, as well as a restrictive boundedness and conditional variance condition. The first of these conditions can be circumvented if $\rho(z,\beta) = \tilde{\rho}(z,\zeta(\beta))$, and $\tilde{\rho}(z,\zeta)$ satisfies the assumptions. Here the above procedure can be used to form an efficient estimator $\tilde{\zeta}$ of ζ_0, and then an efficient estimator of β_0 constructed by minimum chi-square (i.e. $\tilde{\beta} = \text{argmin}_\beta[\tilde{\zeta}-\zeta(\beta)]'\hat{V}_\zeta^{-1}[\tilde{\zeta}-\zeta(\beta)]$ where $\hat{V}_\zeta$ is an estimator of the asymptotic variance of $\tilde{\zeta}$), or a linearized version. The boundedness condition in 2) may require the strong restriction that x and ε are bounded, and the conditional variance condition may also restrict the range of ε. It would be possible to relax these conditions if it was possible to relax the conditions of Lemma A.2 in Appendix A.

The following pair of assumptions imposes dominance conditions that are used in verifying conditions i) and ii) of Lemma 5.1:

Assumption 5.3: $\rho(z_i,\beta)$ is twice continuously differentiable on a neighborhood $\mathcal{N}$ of β_0. Also, there exists $B_i = B(z_i)$ such that $E[B_i] < \infty$ and for all $\beta \in \mathcal{N}$; $\|\rho(z_i,\beta)\|^2$, $\|\rho_\beta(z_i,\beta)\|^2$, $\|\partial\rho_\beta(z_i,\beta)/\partial\beta\| \leq B_i$.

Assumption 5.4: There is $\epsilon > 0$ such that $E[\|J_\beta\|^{2+\epsilon}] < \infty$ and $E[\|\rho_\beta\|^{2+\epsilon}] < \infty$. Also, $b^J_{ij}(\beta) = J_\beta(z_i,\beta) - J_\beta(\pi(x_j,\rho(z_i,\beta),\beta),x_j,\beta)$ and $b^\rho_{ij}(\beta) = \rho_\beta(z_i,\beta) - \rho_\beta(\pi(x_j,\rho(z_i,\beta),\beta),x_j,\beta)$ are twice continuously differentiable on a neighborhood $\mathcal{N}$ of β_0. Also, for $i \neq j$ there exists $B_{ij} = B(z_i,z_j)$ such that $E[B_{ij}] < \infty$ and for all $\beta \in \mathcal{N}$; $\|b^J_{ij}(\beta)\|^2$, $\|\partial b^J_{ij}(\beta)/\partial\beta\|^2$, $\|\partial^2 b^J_{ij}(\beta)/\partial\beta^2\|$, $\|b^\rho_{ij}(\beta)\|^2$, $\|\partial b^\rho_{ij}(\beta)/\partial\beta\|^2$, $\|\partial^2 b^\rho_{ij}(\beta)/\partial\beta^2\|$, $\|b^\rho_{ij}(\beta)\|^2 B_i \leq B_{ij}$ for B_i from Assumption 5.3.

For an example, consider the linear regression model $y = x'\beta + \varepsilon$, where $J(z,\beta)$ is nonexistent and $\rho(z,\beta) = y-x'\beta$, $\rho_\beta(z,\beta) = -x$, $b^\rho_{ij}(\beta) = x_i-x_j$. Here assumptions 5.3 and 5.5 require $E[\|x\|^4] < \infty$, $E[\varepsilon^2] < \infty$. It is possible

to relax these assumptions by using the special structure of this model; see Newey (1988a). In general, it is possible to relax the assumption about the existence of second moments of $\rho(z,\beta)$ if preliminary estimates of location and scale are not used.

The next Assumption imposes smoothness conditions that are useful for verifying Lemma 5.1 iii). Let $f(\varepsilon)$ denote the density of ε:

Assumption 5.5: $\ell(z|\beta) = \exp(J(z,\beta))\cdot f(\rho(z,\beta))$ is mean-square in a neighborhood $\mathcal{N}$ of β_0 and $E[\{\sup_{\mathcal{N}}\|b_{12}^{J}(\beta)\|^2\}|\beta]$ and $E[\{\sup_{\mathcal{N}}\|b_{12}^{\rho}(\beta)\|^2\}|\beta]$ are continuous on $\mathcal{N}$.

The final condition imposes convergence rates for initial estimators.

Assumption 5.6. $\sqrt{n}(\hat{\beta}-\beta) = O_p(1)$, Q and Σ_0 are positive definite, and for some $\epsilon > 0$, $\|\hat{Q}-Q\| = O_p(n^{-\epsilon})$, $\|\hat{\Sigma}-\Sigma_0\| = O_p(n^{-1/4-\epsilon})$, $\|\hat{\mu}-\mu_0\| = O_p(n^{-1/4-\epsilon})$.

Theorem 5.2: Suppose that Assumptions 5.1-5.6, and Lemma 5.1 v) and vi) are satisfied. Then equation (5.1) holds.

Lemma 5.1 is also useful for showing asymptotic efficiency of semiparametric estimators for other models. Examples are given in Newey (1988b, 1989c, 1990b).

6. A Sampling Experiment

To obtain information concerning the small sample performance of the estimator, a sampling experiment was carried out. The experiment concerned the simplest special case of the nonlinear simultaneous equations model, which is the linear regression model. The same regression design, sample size, and

a subset of the distributions in Hsieh and Manski (1987) were considered. The model was $y_i = \beta_0 x_i + \varepsilon_i$, with x_i a binomial random variable and $Prob(x_i = 0) = Prob(x_i = 1) = 1/2$. The distributions for ε_i were standard normal, variance contaminated mixture of normals with relative scale of nine, being $.1N(0,9) + .9N(0,1/9)$, bimodal symmetric mixture of normals, being $.5N(-3,1) + .5N(3,1)$, lognormal, being $\exp(u)$ where u is distributed as standard normal. The sample size was 50.

The estimator of β_0 considered was essentially the linearized semiparametric m-estimator for the nonlinear simultaneous equations model of Section 3, applied to this regression model, with moment functions chosen as in equations (3.13) and (3.14). The preliminary estimator $\hat{\beta}$ of β was chosen to be the least squares estimator form a regression of y_i on $(1,x_i')'$, and the location and scale estimators were chosen to be the constant $\hat{\mu}$ and residual standard deviation $\hat{\sigma}$ from this regression. The matrix $\hat{Q}$ was chosen to the the sample variance of x_i. The approximating functions were

$$(6.1) \qquad \tilde{s}_k(\rho) = \tau(\rho)^k, \ (k = 1, 2, \ldots), \quad \tau(\rho) = \rho/(1+|\rho|).$$

The function $\tau(\rho)$ is bounded and continuously differentiable with bounded Lipschitz derivative. Although this function does not satisfy the twice differentiability hypothesis of Assumption 5.1, it is possible to weaken this assumption to allow the first derivative to just be Lipschitz. This function was used to make the results comparable with Newey (1988a).

Because the estimator differed slightly from that described in Section 3, a brief description is appropriate. Also, this description may help to illustrate the previously described calculations for a particular example. For $p(\rho) = (\tau(\rho),\ldots,\tau(\rho)^K)'$, $\hat{\xi}_i = (y_i - x_i'\hat{\beta} - \hat{\mu})/\hat{\sigma}$, let

(6.2) $$m_k(z,\beta,\hat{\alpha}(z,\beta)) = -(x-\bar{x})\hat{\sigma}^{-1}\tau((y-x'\beta-\hat{\mu})/\hat{\sigma})^k,$$

$$\hat{M}_{ki} = (x_i-\bar{x})(x_i-\bar{x})'k\tau(\hat{\xi}_i)^{k-1}\tau'(\hat{\xi}_i)/\hat{\sigma}^2, \quad \hat{M}_k = \sum_{i=1}^n \hat{M}_{ki}/n,$$

$$\hat{u}_{ki} = -(x_i-\bar{x})\{\tau(\hat{\xi}_i)^k - \sum_{j=1}^n \tau(\hat{\xi}_j)^k/n\}/\hat{\sigma}.$$

These objects are exactly those described in eq. (3.13), except that the additional term $-(x_i-\bar{x})\bar{x}'k\tau(\hat{\xi}_i)^{k-1}\tau'(\hat{\xi}_i)/\hat{\sigma}^2$ has been added to $\hat{M}_{ki}$. This change makes the calculations somewhat easier, without affecting the efficiency result; note that $\sum_{i=1}^n\{-(x_i-\bar{x})\bar{x}'k\tau(\hat{\xi}_i)^{k-1}\tau'(\hat{\xi}_i)/\hat{\sigma}^2\}/n$ should converge to zero by independence of x and ε. The estimator was constructed from equation (6.2), as in equations (3.13) and (3.14). For $\hat{p}_i = (\tau(\hat{\xi}_i),\ldots,\tau(\hat{\xi}_i)^K)'$, $d\hat{p}_i = \tau'(\hat{\xi}_i)(1,\ldots,K\tau(\hat{\xi}_i)^{K-1})'/\hat{\sigma}$, $\bar{p} = \sum_{i=1}^n \hat{p}_i/n$, $\hat{\psi}_i = (x_i-\bar{x})'\hat{Q}(x_i-\bar{x}) = (x_i-\bar{x})'[\sum_{j=1}^n (x_j-\bar{x})(x_j-\bar{x})'/n]^{-1}(x_i-\bar{x})$, the linear combination parameters $\hat{\gamma}$ and the estimator take the form

(6.3) $$\hat{\gamma} = -(\sum_{i=1}^n \hat{\psi}_i(\hat{p}_i-\bar{p})(\hat{p}_i-\bar{p})']^{-1}\sum_{i=1}^n \hat{\psi}_i d\hat{p}_i,$$

$$\tilde{\beta} = \hat{\beta} + [\sum_{i=1}^n (x_i-\bar{x})(x_i-\bar{x})'(d\hat{p}_i'\hat{\gamma})]^{-1}\sum_{i=1}^n (x_i-\bar{x})\{(\hat{p}_i-\bar{p})'\hat{\gamma}\}$$

This estimator differs from that of Newey (1988a), which does not include $\hat{\psi}_i$ in the calculation of $\hat{\gamma}$, and replaces $\sum_{i=1}^n (x_i-\bar{x})(x_i-\bar{x})'(d\hat{p}_i'\hat{\gamma})$ by $-\{\sum_{i=1}^n [(\hat{p}_i-\bar{p})'\hat{\gamma}]^2/n\}\sum_{i=1}^n (x_i-\bar{x})(x_i-\bar{x})'$ in the calculation of $\tilde{\beta}$.

Before describing the results, it is useful to note some facts about this estimator. Because of the use of preliminary estimators of location and scale, the ratio of its mean-square error relative to ordinary least squares in the simulations is invariant to location and scale of ε. Also, because of the use of powers of an odd function, $\tilde{\beta}-\beta_0$ is an odd function of the $(\varepsilon_1-\mu_0,\ldots,\varepsilon_n-\mu_0)'$. Therefore, when the disturbance is symmetrically distributed around μ_0, each estimator will be symmetrically distributed

around β_0, which occurs for each of the first three distributions. The estimators also appear to be unbiased for all the distributions. In calculations not reported here it was found that the average deviation of the estimators from β_0 was much smaller than the mean-square error.

Computations were performed using GAUSS on a microcomputer, with 500 replications. Table One reports the root mean square error (RMSE), relative to ordinary least squares, of the estimator for $K \in \{1,\ldots,5\}$, as well as that for the cross-validated choice $\hat{K}$ described in Section 3. The distribution of $\hat{K}$ in the simulations is also given.

The RMSE results here are much like those for the similar moment estimator considered in Newey (1988a), and, except in the normal case, are much better than those for the kernel-based adaptive estimator given in Hsieh and Manski (1987). The cross-validated choice of K performs quite well. In each case there is only a few percentage points loss in efficiency from using the cross-validated choice rather than the (nonfeasible) value of K that minimizes the mean-square error. Also, cross-validation seems to be quite good at avoiding disasters; for the location mixture, cross-validation never chooses K to be 1 or 2, for which the performance of the estimator is very poor. These results seem quite promising.

Appendix A: Proofs for Section 5

Throughout, c and C will denote generic, positive constants that can be different in different uses. First, a simple but useful Lemma on convergence rates of regression coefficients will be given.

Lemma A.1: Suppose $\hat{\Sigma}$, $\hat{G}$, Σ, G *are random submatrices of* $\hat{\Sigma}_n$, $\hat{G}_n$, Σ_n, G_n, *respectively, with* Σ_n, G_n *nonrandom, such that i)* $\|\hat{\Sigma}_n-\Sigma_n\| = O_p(\epsilon_{\Sigma n})$, $\|\hat{G}_n-G_n\| = O_p(\epsilon_{Gn})$; *ii)* $\|G_n\| = O(\Delta_{Gn})$, $\|\Sigma_n\| = O(\Delta_{\Sigma n})$, Σ_n *positive semi-definite,* $\lambda(\Sigma_n)^{-1} = O(\Delta_{\lambda n})$, $dim(\Sigma_n) = O(\Delta_{dn})$. *Then if* $\epsilon_{\Sigma n}\Delta_{\lambda n} = o(1)$,

$$\|\hat{\Sigma}^{-1}\hat{G}-\Sigma^{-1}G\| = O_p(\Delta_{dn}^{1/2}\Delta_{\lambda n}\epsilon_{Gn}) + O_p(\Delta_{dn}^{1/2}\Delta_{\lambda n}\epsilon_{\Sigma n}\|\Sigma_n^{-1}G_n\|)$$

$$= O_p(\Delta_{dn}^{1/2}\Delta_{\lambda n}\epsilon_{Gn}) + O_p(\Delta_{dn}\Delta_{\lambda n}^{2}\Delta_{Gn}\epsilon_{\Sigma n})$$

Proof: Let $\hat{\lambda}$, λ, $\hat{\lambda}_n$, λ_n denote the smallest eigenvalues of $\hat{\Sigma}$, Σ, $\hat{\Sigma}_n$, and Σ_n respectively. Note that by $|\lambda(A)-\lambda(B)| \leq \|A-B\|$,

$$|1-\hat{\lambda}_n/\lambda_n| \leq |\lambda_n-\hat{\lambda}_n|/\lambda_n \leq \|\hat{\Sigma}_n-\Sigma_n\|/\lambda_n = O_p(\epsilon_{\Sigma n}\Delta_{\lambda n}) = o_p(1),$$

so that with probability approaching one $\hat{\lambda}_n \geq C\lambda_n$. Also, by the submatrix assumption $\lambda \geq \lambda_n$ and $\hat{\lambda} \geq \hat{\lambda}_n$, and in particular $\hat{\Sigma}$ is positive definite with probability approaching one. Then by $\|A\| \leq \dim(A)^{1/2}\bar{\lambda}(A)$, where $\bar{\lambda}(A)$ is the maximum eigenvalue of a symmetric matrix A, by $\bar{\lambda}(A) = 1/\lambda(A^{-1})$ for a positive definite matrix A, and the Cauchy-Schwartz and triangle inequalities,

$$\|\hat{\Sigma}^{-1}\hat{G} - \Sigma^{-1}G\| \leq \|\hat{\Sigma}^{-1}(\hat{G}-G)\| + \|(\hat{\Sigma}^{-1}-\Sigma^{-1})G\| \leq \|\hat{\Sigma}^{-1}\|\|\hat{G}-G\| + \|\hat{\Sigma}^{-1}(\Sigma-\hat{\Sigma})\Sigma^{-1}G\|$$

$$\leq \|\hat{\Sigma}^{-1}\|(\|\hat{G}-G\| + \|\Sigma-\hat{\Sigma}\|\|\Sigma^{-1}G\|) \leq \dim(\hat{\Sigma})^{1/2}\hat{\lambda}^{-1}(\|\hat{G}_n-G_n\| + \|\Sigma_n-\hat{\Sigma}_n\|\|\Sigma^{-1}G\|)$$

$$\leq \dim(\Sigma_n)^{1/2}\hat{\lambda}_n^{-1}[O_p(\epsilon_{Gn}) + O_p(\epsilon_{\Sigma n}\|\Sigma^{-1}G\|)]$$

$$= O_p(\Delta_{dn}^{1/2}\lambda_n^{-1})[O_p(\epsilon_{Gn}) + O_p(\epsilon_{\Sigma n}\|\Sigma^{-1}G\|)]$$

$$= O_p(\Delta_{dn}^{1/2}\Delta_{\lambda n}\epsilon_{Gn}) + O_p(\Delta_{dn}^{1/2}\Delta_{\lambda n}\epsilon_{\Sigma n}\|\Sigma_n^{-1}G_n\|),$$

giving the first equality of the conclusion. The second equality follows by $\|\Sigma_n^{-1}G_n\| \leq \|\Sigma_n^{-1}\|\|G_n\| \leq \lambda(\Sigma_n)^{-1}\dim(\Sigma_n)^{1/2}\|G_n\| = O(\Delta_{dn}^{1/2}\Delta_{\lambda n}\Delta_{Gn})$. ∎

Proof of Lemma 5.1: The n argument will be suppressed where convenient. By $\hat{K} \xrightarrow{p} \infty$ there exists nonrandom $\tilde{K}$ such that $\tilde{K} \leq \hat{K}$ with probability approaching one and $\tilde{K} \longrightarrow \infty$. Let $\bar{K} \equiv K_n$ denote the upper bound specified in iv). In the case that the number of elements of $\mathcal{K}$ is bounded it can be assumed without loss of generality that the smallest element of $\mathcal{K}$ goes to infinity and that the largest element is less than or equal to $\bar{K}$. For the other case specified in v), let $\mathcal{K} = \{\tilde{K}, \tilde{K}+1, \ldots, \bar{K}\}$. Note that in both cases $\hat{K} \in \mathcal{K}$ with probability approaching one. Therefore, it suffices to show that the result holds when K is always an element of $\mathcal{K}$. (Define a new estimator equal to $\tilde{\beta}$ if $\hat{K} \in \mathcal{K}$, and equal to the estimator for some choice of $K \in \mathcal{K}$ if $\hat{K} \notin \mathcal{K}$; $\tilde{\beta}$ and this new estimator are equal with probability approaching one, and therefore have the same asymptotic properties, so that it suffices to consider just the new estimator. Similar logic applies to the asymptotic variance estimator.)

Some additional notation is useful in what follows. For any given K let $\bar{\gamma}$ be obtained from eq. (3.5) and $L \equiv (1, \bar{\gamma}')\otimes I_q$. Also, let

$$u_n = \sum_{i=1}^n u_i/n \quad \Omega \equiv E[uu'], \quad \tilde{\Omega} \equiv \sum_{i=1}^n u_i u_i'/n, \quad \hat{\Omega} \equiv \sum_{i=1}^n \hat{u}_i \hat{u}_i'/n,$$

$$\Sigma \equiv E[U'QU], \quad \hat{\Sigma} \equiv \sum_{i=1}^n \hat{U}_i'\hat{Q}\hat{U}_i/n,$$

and let a subscript denote the same matrices calculated at the subscript value of K. For $\tilde{\gamma}^K$ as specified in v), let $\bar{S}_K \equiv L_K u$, and $\tilde{S}_K \equiv [(1, \tilde{\gamma}^{K\prime})\otimes I_q]u$.

Let $\bar{\lambda}(\cdot)$ denote the largest eigenvalue of a symmetric matrix.

Next, by iii), $\bar{\gamma}^K$ minimizes $E[\{S-((1,\gamma')\otimes I_q)u\}'Q\{S-((1,\gamma')\otimes I)u\}]$ over γ, so that $E[\|S-\bar{S}_K\|^2] \leq (1/\lambda(Q))E[(S-\bar{S}_K)'Q(S-\bar{S}_K)] \leq (1/\lambda(Q))E[(S-\tilde{S}_K)'Q(S-\tilde{S}_K)] \leq (\bar{\lambda}(Q)/\lambda(Q))E[\|S-\tilde{S}_K\|^2]$. Then by v) and $\tilde{K} \longrightarrow \infty$,

$$(A.1) \qquad \sum_{\mathcal{K}}\{E[\|S-\bar{S}_K\|^2]\}^{1/2} \leq C\sum_{\mathcal{K}}\{E[\|S-\tilde{S}_K\|^2]\}^{1/2} = o(1),$$

and for n large enough,

$$(A.2) \qquad \max_{\mathcal{K}}E[\|S-\bar{S}_K\|^2] \leq \max_{\mathcal{K}}\{E[\|S-\bar{S}_K\|^2]\}^{1/2} \leq \sum_{\mathcal{K}}\{E[\|S-\bar{S}_K\|^2]\}^{1/2} = o(1).$$

By iv) and Jensen's and Cauchy-Schwartz inequalities,

$$(A.3) \qquad \|L_{\hat{K}}M + V^{-1}\| \leq \max_{\mathcal{K}}\|-E[S_K S'] + E[SS']\| \leq C\max_{\mathcal{K}}E[\|S_K - S\|^2] = o(1),$$

$$\|L_{\hat{K}}\Omega_{\hat{K}}L_{\hat{K}} - V^{-1}\| \leq \max_{\mathcal{K}}\|E[S_K S_K'] - E[SS']\| = o(1),$$

Also, by the Markov and Cauchy-Schwartz inequalities, $E[S_K]$ and $E[S] = 0$, and independence of the observations it follows that

$$(A.4) \qquad \|\sum_{i=1}^n S_{\hat{K}i}/\sqrt{n} - \sum_{t=1}^n S_i/\sqrt{n}\| \leq \sum_{\mathcal{K}}\|\sum_{i=1}^n (S_{Ki}-S_i)/\sqrt{n}\|$$

$$= O_p(\sum_{\mathcal{K}}\{E[\|\sum_{i=1}^n (S_{Ki}-S_i)/\sqrt{n}\|^2]\}^{1/2}) = O_p(\sum_{\mathcal{K}}\{E[\|S_K-S\|^2]\}^{1/2}) = o_p(1).$$

Next note that by v), for any $c, C > 0$, for $\nu = \nu(\bar{K})$,

$$(A.5) \qquad n^{-c}\nu^{C\nu} = \exp\{-C\ln(n)[c/C - (\nu\ln(\nu)/\ln(n))]\} = o(1).$$

By Lemma A.2 of Newey (1988a), i), and eq. (A.5),

$$(A.6) \qquad \|\tilde{\Omega}_{\bar{K}}-\Omega_{\bar{K}}\| \leq qK\max_{j\ell}|(\tilde{\Omega}_{\bar{K}})_{j\ell}-(\Omega_{\bar{K}})_{j\ell}| = O_p(n^{-c\epsilon}(\nu^{\Delta\nu})^C) = O_p(n^{-c}).$$

By i), ii), (A.5), and the triangle, Cauchy-Schwartz, and Markov inequalities,

(A.7) $$\|\hat{\Omega}_{\bar{K}}-\tilde{\Omega}_{\bar{K}}\| \le \sum_{i=1}^{n}\|\hat{u}_{i\bar{K}}\hat{u}'_{i\bar{K}} - u_{i\bar{K}}u'_{i\bar{K}}\|/n$$

$$\le \sum_{i=1}^{n}\|\hat{u}_{i\bar{K}} - u_{i\bar{K}}\|^2/n + \sum_{i=1}^{n}\|\hat{u}_{i\bar{K}}-u'_{i\bar{K}}\|\|u_{i\bar{K}}\|/n$$

$$\le O_p(n^{-\epsilon}\nu^{\Delta\nu}) + (\sum_{i=1}^{n}\|\hat{u}_{i\bar{K}}-u_{i\bar{K}}\|^2/n)^{1/2}(\sum_{i=1}^{n}\|u_{i\bar{K}}\|^2/n)$$

$$= O_p(n^{-c}) + O_p([n^{-\epsilon}\nu^{\Delta\nu}]^{1/2})O_p(\mathrm{tr}(\Omega_{\bar{K}})) = O_p(n^{-c}).$$

where tr denotes the trace. It follows that $\|\hat{\Omega}_{\bar{K}}-\Omega_{\bar{K}}\| = O_p(n^{-c})$. Each element of $\hat{\Sigma}_{\bar{K}}$ and $\Sigma_{\bar{K}}$ consists of a sum of q elements of $(I_{\bar{K}}\otimes\hat{Q})\hat{\Omega}_{\bar{K}}$ and $(I_{\bar{K}}\otimes Q)\Omega_{\bar{K}}$. Then by $\|\Omega_K\| \le \dim(\Omega_K)\bar{\lambda}(\Omega_K) \le qK\mathrm{tr}(\Omega_K) = O(\nu^{C\nu})$,

(A.8) $$\|\hat{\Sigma}_{\bar{K}}-\Sigma_{\bar{K}}\| \le C\|(I_{\bar{K}}\otimes\hat{Q})\hat{\Omega}_{\bar{K}} - (I\otimes Q_{\bar{K}})\Omega_{\bar{K}}\|$$

$$\le C\|I_{\bar{K}}\otimes\hat{Q}\|\|\hat{\Omega}_{\bar{K}}-\Omega_{\bar{K}}\| + \|(I_{\bar{K}}\otimes\hat{Q})-(I_{\bar{K}}\otimes Q)\|\|\Omega_{\bar{K}}\| = CK(\|\hat{Q}\|\|\hat{\Omega}_{\bar{K}}-\Omega_{\bar{K}}\|+\|\hat{Q}-Q\|\|\Omega_{\bar{K}}\|)$$

$$= O_p(\nu^{\Delta\nu})[O_p(n^{-c}) + O_p(n^{-\epsilon})O(\nu^{C\nu}) = O_p(n^{-c}).$$

Similarly,

(A.9) $$\|[\mathrm{tr}(\hat{Q}\hat{M}'_1),\ldots,\mathrm{tr}(\hat{Q}\hat{M}_{\bar{K}}')]-[\mathrm{tr}(QM'_1),\ldots,\mathrm{tr}(QM_{\bar{K}}')]\| = O_p(n^{-c})$$

$$\|\sum_{i=1}^{n}\hat{U}'_{i\bar{K}}\hat{Q}\hat{u}_{0i}/n - E[U_{\bar{K}}'Qu_0]\| = O_p(n^{-c})\sum_{i=1}^{n}\| = O_p(n^{-\epsilon}\nu^{C\nu}).$$

Then by $\lambda(\Sigma_{\bar{K}})^{-1} = O(\nu^{C\nu})$, Lemma A.1 gives,

(A.10) $$\|\hat{L}-L_{\hat{K}}\| = O_p(n^{-c}\nu^{C\nu}) = O_p(n^{-c}), \quad \|L_{\hat{K}}\| \le C\|\gamma\| = O_p(\nu^{C\nu}).$$

Then by $\hat{M}$ a submatrix of $\hat{M}_{\bar{K}}$,

(A.11) $$\|\hat{L}\hat{M} - L_{\hat{K}}M_{\hat{K}}\| \le \|\hat{L}-L_{\hat{K}}\|\|\hat{M}-M_{\hat{K}}\| + \|\hat{L}-L_{\hat{K}}\|\|M_{\hat{K}}\| + \|L_{\hat{K}}\|\|\hat{M}-M_{\hat{K}}\|$$

$$\le O_p(n^{-c})(\|\hat{M}_{\bar{K}}-M_{\bar{K}}\| + \|M_{\bar{K}}\|) + O_p(\nu^{C\nu})\|\hat{M}_{\bar{K}}-M_{\bar{K}}\| = o_p(1).$$

It follows similarly that $\|\hat{L}\hat{\Omega}\hat{L}' - L_{\hat{K}}\Omega_{\hat{K}}L_{\hat{K}}'\| = o_p(1)$. Then by eq. (A.3), the

triangle inequality, and continuity of the matrix inverse at nonsingularity,

$$\text{(A.12)} \qquad \|(\hat{L}\hat{M})^{-1} - V\| = o_p(1), \quad \|\hat{L}\hat{\Omega}\hat{L}' - V\| = o_p(1).$$

Next, note that by the Markov and Cauchy-Schwartz inequalities,

$$\text{(A.13)} \qquad \sqrt{n}\|u_{n\bar{K}}\| = O_p(\{E[n\|u_{n\bar{K}}\|^2]\}^{1/2}) = O_p(\{tr(\Omega_{\bar{K}})\}^{1/2}) = O_p(\nu^{C\nu}).$$

Then by the triangle inequality, and by, e.g. $\hat{m}_n(\beta_0)-u_{n\hat{K}}$ a subvector of $\hat{m}_{n\bar{K}}(\beta_0)-u_{n\bar{K}}$,

$$\begin{aligned}
\text{(A.14)} \qquad & \sqrt{n}\|\hat{L}\hat{m}_n(\hat{\beta})-\textstyle\sum_{i=1}^n S_i/n+V^{-1}(\hat{\beta}-\beta_0)\| \\
& \le \|\hat{L}\|\sqrt{n}(\|\hat{m}_n(\hat{\beta})-\hat{m}_n(\beta_0)-M_{\hat{K}}(\hat{\beta}-\beta_0)\|+\|\hat{m}_n(\beta_0)-u_{n\hat{K}}\|) \\
& \quad + \|\hat{L}-L_{\hat{K}}\|\sqrt{n}\|u_{n\hat{K}}\| + \|\hat{L}M_{\hat{K}}+V^{-1}\|\sqrt{n}\|\hat{\beta}-\beta_0\| + \sqrt{n}\|L_{\hat{K}}u_{n\hat{K}}-\textstyle\sum_{i=1}^n S_i/n\| \\
& \le O_p(\nu^{C\nu})\sqrt{n}(\|\hat{m}_{n\bar{K}}(\hat{\beta})-\hat{m}_{n\bar{K}}(\beta_0)-M_{\bar{K}}(\hat{\beta}-\beta_0)\|+\|\hat{m}_{n\bar{K}}(\beta_0)-u_{n\bar{K}}\|) \\
& \quad + O_p(n^{-c})\sqrt{n}\|u_{n\bar{K}}\| + \|\hat{L}M_{\hat{K}}+V^{-1}\|\sqrt{n}\|\hat{\beta}-\beta_0\| + \sqrt{n}\|L_{\bar{K}}u_{n\bar{K}}-\textstyle\sum_{i=1}^n S_i/n\| \\
& \le (\|L_{\hat{K}}\|+\|\hat{L}-L_{\hat{K}}\|)\sqrt{n}\|\hat{m}_{n\bar{K}}(\hat{\beta})-m_{n\bar{K}}(\beta_0)-M_{\bar{K}}(\hat{\beta}-\beta_0)\}\| \\
& \quad + \|\hat{L}-L_{\hat{K}}\|(\sqrt{n}\|m_{n\bar{K}}(\beta_0)\| + \|M_{\bar{K}}\|\sqrt{n}\|\hat{\beta}-\beta_0\|) + o_p(1) = o_p(1).
\end{aligned}$$

It then follows

$$\begin{aligned}
\text{(A.15)} \qquad & \sqrt{n}\|(\hat{L}\hat{M})^{-1}\hat{L}\hat{m}_n(\hat{\beta}) - [-V\textstyle\sum_{i=1}^n S_i + (\hat{\beta}-\beta_0)]\| \\
& \le \|(\hat{L}\hat{M})^{-1}+V\|\sqrt{n}\|\hat{L}\hat{m}_n(\hat{\beta})\| + \|V\|\|\hat{L}\hat{m}_n(\hat{\beta})-\textstyle\sum_{i=1}^n S_i/n+V^{-1}(\hat{\beta}-\beta_0)\| \\
& \le o_p(1)\sqrt{n}\|\hat{L}\hat{m}_n(\hat{\beta})-\textstyle\sum_{i=1}^n S_i+V^{-1}(\hat{\beta}-\beta_0)\|+\|\textstyle\sum_{i=1}^n S_i/\sqrt{n}\| +\sqrt{n}\|\hat{\beta}-\beta_0\|]+o_p(1) = o_p(1).
\end{aligned}$$

Then by equation (A.14) and the definition of $\tilde{\beta}$ it follows that

(A.16) $$\sqrt{n}(\tilde{\beta}-\beta_0) = \sqrt{n}(\hat{\beta}-\beta_0) - \sqrt{n}(\hat{L}\hat{M})^{-1}\hat{L}\hat{m}_n(\hat{\beta}) = V\sum_{i=1}^{n} S_i/\sqrt{n} + o_p(1).$$

The first conclusion now follows by the Lindbergh-Levy central limit theorem. The second conclusion follows by eq. (A.12). The final conclusion follows by Theorem 2.2 of Newey (1990a)

The following Lemma is useful for proving the eigenvalue hypothesis for the nonlinear simultaneous equations example.

Lemma A.2: Suppose ε *and* x *are independent,* $g(\varepsilon,x)$ *is bounded,* $Prob(g(\varepsilon,x) = 0) = 0$, $E[g(\varepsilon,x)^2|\varepsilon] \geq c > 0$, *and for all* $a(x)b(\varepsilon)$, $Prob(g(\varepsilon,x)=a(x)b(\varepsilon)) < 1$. *Then* $\inf_{\{\delta(\varepsilon):E[\delta^2]=1\}} E[Var\{\delta(\varepsilon)g(\varepsilon,x)|x\}] > 0$.

Proof: Proceed by showing the contrapositive: Suppose the infimum is zero and let $\{\delta_\ell(\varepsilon)\}$ be a sequence such that $E[Var\{\delta_\ell(\varepsilon)g(\varepsilon,x)|x\}] \longrightarrow 0$, i.e. $\delta_\ell(\varepsilon)g(\varepsilon,x) - E[\delta_\ell(\varepsilon)g(\varepsilon,x)|x] \xrightarrow{m.s.} 0$. Consider a subsequence where this convergence is almost sure. Apply Alagolu's Theorem, on the usual Hilbert space of functions of ε with finite mean square, to find a further subsequence where $\delta_\ell(\varepsilon)$ converges weak-* (in the function space sense) to some $\delta(\varepsilon)$. This means that $E[\delta_\ell(\varepsilon)r(\varepsilon)] \longrightarrow E[\delta(\varepsilon)r(\varepsilon)]$ for any $r(\varepsilon)$ with $E[r(\varepsilon)^2] < \infty$, so that for almost all x, $a_\ell(x) \equiv E[\delta_\ell(\varepsilon)g(\varepsilon,x)|x] = \int\delta_\ell(\varepsilon)g(\varepsilon,x)f(\varepsilon)d\varepsilon \longrightarrow \int\delta(\varepsilon)g(\varepsilon,x)f(\varepsilon)d\varepsilon = E[\delta(\varepsilon)g(\varepsilon,x)|x] \equiv a(x)$. It follows that $\delta_\ell(\varepsilon) = [\delta_\ell(\varepsilon)g(\varepsilon,x)]/g(\varepsilon,x) \xrightarrow{a.s.} a(x)/g(\varepsilon,x)$. Since $\delta_\ell(\varepsilon)$ is a function only of ε, so is its almost sure limit, i.e. $a(x)/g(\varepsilon,x) = \tilde{\delta}(\varepsilon)$ for some $\tilde{\delta}(\varepsilon)$. Consider the events $A = \{a(x) \neq 0\}$, $B = \{b(\varepsilon,x) \neq 0\}$, $\Delta = \{\tilde{\delta}(\varepsilon) = 0\}$. Then by independence, $0 = Prob(A\cap\Delta\cap B) = Prob(A\cap\Delta) = Prob(A)Prob(\Delta)$, implying $Prob(A) = 0$ or $Prob(\Delta) = 0$. If $prob(\Delta) = 0$, then $g(\varepsilon,x) = a(x)/\tilde{\delta}(\varepsilon)$. Thus, it suffices to show $Prob(A) > 0$. Suppose $Prob(A) = 0$, i.e. $a(x) = 0$. Then $a_\ell(x) \xrightarrow{a.s.} 0$, while by boundedness of

$g(\varepsilon,x)$ and independence, $|a_\ell(x)|^2 \leq CE[\delta_\ell(\varepsilon)^2|x] = CE[\delta_\ell(\varepsilon)^2] = C$, so that by the dominated convergence theorem, $E[a_\ell(x)^2] \longrightarrow 0$. Therefore, $\{E[\{\delta_\ell(\varepsilon)g(\varepsilon,x)-a_\ell(x)\}^2]\}^{1/2} \geq \{E[\{\delta_\ell(\varepsilon)g(\varepsilon,x)\}^2]\}^{1/2} - \{E[a_\ell(x)^2]\}^{1/2} \geq \{E[\delta_\ell(\varepsilon)^2E[g(\varepsilon,x)^2|\varepsilon]]\}^{1/2} - \{E[a_\ell(x)^2]\}^{1/2} \geq c\{E[\delta_\ell(\varepsilon)^2]\}^{1/2} - o(1) \geq c-o(1)$, contradicting the starting hypothesis of the proof. Therefore, $Prob(A) > 0$ under this hypothesis, implying $Prob(\Delta) = 0$, implying $g(\varepsilon,x) = a(x)/\tilde{\delta}(\varepsilon)$. ■

The following Lemmas are useful for verifying the convergence rate hypotheses. Let γ be a vector of parameters with corresponding true value γ_0 and estimator $\hat{\gamma}$, and $b_{ij}(\gamma) = b(z_i,z_j,\gamma)$ a matrix function of a pair of observations, with dimension that can depend on the sample size. Also, let $\hat{b}_{ij} = b_{ij}(\hat{\gamma})$, $b_{ij} = b_{ij}(\gamma_0)$, $\bar{b} = E[b_{12}]$, and for $i \neq j$, $\bar{b}_{i.} = E[b_{ij}|z_i]$, $\bar{b}_{.i} = E[b_{ji}|z_i]$.

Lemma A.3: $\|\sum_{ij}b_{ij}/n^2 - \sum_i(\bar{b}_{i.}+\bar{b}_{.i})/n + \bar{b}\| = O_p(E[\|b_{11}\|]/n + (E[\|b_{12}\|^2])^{1/2}/n)$.

Proof: This Lemma is a V-statistic projection result, proved as follows: First, consider the case with $\bar{b} = 0$. Note that

$$\|\sum_{ij}b_{ij}/n^2 - \sum_i(\bar{b}_{i.}+\bar{b}_{.i})/n\| = \|\sum_{ij}(b_{ij}-\bar{b}_{i.}-\bar{b}_{.j})/n^2\|$$

$$\leq \|\sum_{i\neq j}(b_{ij}-\bar{b}_{i.}-\bar{b}_{.j})/n^2\| + \|\sum_i(b_{ii}-\bar{b}_{i.}-\bar{b}_{.i})/n^2\| \equiv T_1+T_2$$

Note $E[T_2] \leq (E[\|b_{11}\| + 2E[\|b_{12}\|])/n$. Also, for $i\neq j$, $k\neq\ell$ let $\nu_{ijk\ell} \equiv E[(b_{ij}-\bar{b}_{i.}-\bar{b}_{.j})'(b_{k\ell}-\bar{b}_{k.}-\bar{b}_{.\ell})]$. By i.i.d. observations, if neither k nor ℓ is equal to i or j, then $\nu_{ijk\ell} = 0$. Also for ℓ not equal to i or j,

$$\nu_{iji\ell} = E[(b_{ij}-\bar{b}_{i.})'(b_{i\ell}-\bar{b}_{i.})] = E[E[(b_{ij}-\bar{b}_{i})'(b_{i\ell}-\bar{b}_{i.})|z_i,z_j]]$$

$$= E[(b_{ij}-\bar{b}_{i.})'(E[b_{i\ell}|z_i,z_j]-\bar{b}_{i.})] = 0 = \nu_{ijj\ell}.$$

Similarly, $\nu_{ijk\ell} = 0$ if k equals neither i nor j. Thus,

$$E[T_1^2] = \sum_{i\neq j}\sum_{k\neq\ell}\nu_{ijk\ell}/n^4 = \sum_{i\neq j}(\nu_{ijij}+\nu_{ijji})/n^4$$

$$= 2(n^2-n)E[\|b_{12}-\bar{b}_{1.}-\bar{b}_{.2}\|^2]/n^4 = E[\|b_{12}-\bar{b}_{1.}-\bar{b}_{.2}\|^2]O(n^{-2}),$$

and $T_1 = O_p(\{E[\|b_{12}-\bar{b}_{1.}-\bar{b}_{.2}\|^2]\}^{1/2}n^{-1}) = O_p(\{E[\|b_{12}\|^2]\}^{1/2}n^{-1})$. The conclusion then follows by the Markov and triangle inequalities. Finally, the conclusion when $\bar{b} \neq 0$ follows by replacing b_{ij}, $\bar{b}_{i.}$, $\bar{b}_{.i}$ by $b_{ij}-\bar{b}$, $\bar{b}_{i.}-\bar{b}$, $\bar{b}_{.i}-\bar{b}$, respectively in the above argument, and noting that $\|\bar{b}\| \leq E[\|b_{12}\|] \leq \{E[\|b_{12}\|^2]\}^{1/2}$, which implies $E[\|b_{11}-\bar{b}\|] \leq E[\|b_{11}\|] + \{E[\|b_{12}\|^2]\}^{1/2}$ and $\{E[\|\bar{b}_{.i}-\bar{b}\|^2]\}^{1/2} \leq 2\{E[\|b_{12}\|^2]\}^{1/2}$. ■

Lemmma A.4: Suppose $\hat{\gamma} \xrightarrow{p} \gamma_0$ *and there is a neighborhood* N *of* γ_0 *and* $B(z,\tilde{z})$ *such that for* $B_{ij} = B(z_i,z_j)$ *and* $\gamma \in N$, $b_{ij}(\gamma)$ *is continuously differentiable and* $\sup_N \|\partial b_{ij}(\gamma)/\partial\gamma\| \leq B_{ij}$. *Then*

$$\|\sum_{ij}\hat{b}_{ij}/n^2 - \bar{b}\|$$

$$\leq O_p(\{E[B_{11}]/n\}+E[B_{12}])\|\hat{\gamma}-\gamma_0\| + O_p(E[\|b_{11}\|]/n + (E[\|b_{12}\|^2])^{1/2}/\sqrt{n}).$$

Proof: By a mean-value expansion,

$$\|\sum_{ij}\hat{b}_{ij}/n^2 - \sum_{ij}b_{ij}/n^2\| \leq \|\sum_{ij}\partial b_{ij}(\bar{\gamma})/\partial\gamma/n^2\|\cdot\|\hat{\gamma}-\gamma_0\|$$

$$\leq (\sum_{ij}B_{ij}/n^2)\|\hat{\gamma}-\gamma_0\| = O_p(E[B_{11}]/n + E[B_{12}])\|\hat{\gamma}-\gamma_0\|,$$

where $\bar{\gamma}$ is the mean value, the inequalities hold with probability approaching one, and the last equality follows by the Markov inequality. Also, $\|\sum_{ij}b_{ij}/n^2 - \bar{b}\| \leq \|\sum_{ij}b_{ij}/n^2 - \sum_i(\bar{b}_{i.}+\bar{b}_{.i})/n + \bar{b}\| + \|\sum_i(\bar{b}_{i.}+\bar{b}_{.i})/n - 2\bar{b}\|$ by the triangle inequality. Thus, the conclusion follows by Lemma A.3 and $E[\|\sum_i(\bar{b}_{i.}+\bar{b}_{.i})/n - 2\bar{b}\|] \leq \{E[\|\bar{b}_{1.}+\bar{b}_{.1}\|^2]\}^{1/2}/\sqrt{n} = O(\{E[\|b_{12}\|^2]\}^{1/2}/\sqrt{n})$. ■

Lemmma A.5: Under the hypotheses of Lemma A.4,

$$\sum_i \|\sum_j [b_{ij}(\hat{\gamma})+b_{ji}(\hat{\gamma})]/n - (\bar{b}_{i.}+\bar{b}_{.i})\|^2/n$$

$$\leq O_p(\{E[B_{11}^2]/n\}+E[B_{12}^2])\|\hat{\gamma}-\gamma_0\|^2 + O_p(E[\|b_{11}\|^2+\|b_{12}\|^2]/n).$$

Proof: First, note that

$$\text{(A.17)} \quad \sum_i \|\sum_j [(b_{ij}(\hat{\gamma})+b_{ji}(\hat{\gamma}))-(b_{ij}+b_{ji})]/n\|^2/n \leq 4\sum_{ij}\|b_{ij}(\hat{\gamma})-b_{ij}\|^2/n^2$$

$$\leq \sum_{ij}\|\partial b_{ij}(\bar{\gamma})/\partial\gamma\|^2\|\hat{\gamma}-\gamma_0\|^2/n^2$$

$$\leq (\sum_{ij}B_{ij}^2/n^2)\|\hat{\gamma}-\gamma_0\|^2 \leq O_p(\{E[B_{11}]/n\} + E[B_{12}^2])\|\hat{\gamma}-\gamma_0\|,$$

where $\bar{\gamma}$ is the mean value and the last two inequalities follow as in the proof of Lemma A.4. Also note that $E[\|\sum_j b_{1j}/n - \bar{b}_{1.}\|^2] \leq CE[\|\sum_{j\neq 1}(b_{1j}-\bar{b}_{1.})/n\|^2] + CE[\|b_{11}-\bar{b}_{1.}\|^2]/n \leq CE[E[\|\sum_{j\neq 1}(b_{1j}-\bar{b}_{1.})/n\|^2|z_1]] + CE[\|b_{11}\|^2+\|\bar{b}_{1.}\|^2]/n \leq E[E[\|b_{12}-\bar{b}_{1.}\|^2|z_1]](n-1)/n^2 + O(E[\|b_{11}\|^2+\|b_{12}\|^2]/n) = O(E[\|b_{11}\|^2+\|b_{12}\|^2]/n)$. Similarly $E[\|\sum_j b_{j1}/n - \bar{b}_{.1}\|^2] = O(E[\|b_{11}\|^2+\|b_{12}\|^2]/n)$. It then follows by i.i.d. observations that

$$\text{(A.18)} \quad E[\sum_i \|\sum_j [b_{ij}+b_{ji}]/n - (\bar{b}_{i.}+\bar{b}_{.i})\|^2]/n$$

$$\leq CE[\|\sum_j b_{1j}/n - \bar{b}_{1.}\|^2] + E[\|\sum_j b_{j1}/n - \bar{b}_{.1}\|^2] = O(E[\|b_{11}\|^2+\|b_{12}\|^2]/n).$$

Eqs. (A.17)-(A.18) and the triangle inequality give the conclusion. ∎

Proof of Theorem 5.2: The proof proceeds by verifying the hypotheses of Lemma 5.1. Let $\gamma = (\beta', \tilde{\mu}', \text{vec}(\Sigma^{-1/2})')'$, $\gamma_0 = (\beta_0', (\Sigma_0^{-1/2}\mu_0)', \text{vec}(\Sigma_0^{-1/2})')'$, $\rho(z,\gamma) = \Sigma^{-1/2}\rho(z,\beta) - \tilde{\mu}$, $\xi = \rho(z,\gamma_0) = \Sigma_0^{-1/2}(\varepsilon-\mu_0)$. Let $u_k = v_k - E[v_k|x]$ for $v_k = (\rho_\beta - E[\rho_\beta|\xi])'\Sigma^{-1/2}\tilde{s}_k(\xi)$. Let $\nu = \bar{\nu}(K)$ from Assumption 5.1. First, the smallest eigenvalue hypothesis of Lemma 5.1 will be shown. Let $\bar{p}(\xi) = (1,\ldots,\tau(\xi_1)^\nu)'\otimes\cdots\otimes(1,\ldots,\tau(\xi_s)^\nu)'$ and $\tilde{p}(\xi)$ equal the same

vector except that the first element (i.e. 1) is excluded. By Assumption 5.1, U is a submatrix of $\bar{U} = (\bar{V}-E[\bar{V}|x])(I\otimes\Sigma^{-1/2},)$, where $\bar{V} = p(\xi)'\otimes(\rho_\beta-E[\rho_\beta|\xi])'$, with $p(\xi) = \tilde{p}(\xi)$ under Assumption 5.2 i) and $p(\xi) = \bar{p}(\xi)$ under Assumption 5.2 ii). Then, by $I\otimes\Sigma^{-1/2}$ nonsingular,

$$\text{(A.19)} \qquad \lambda(E[U'U]) \geq \lambda(E[\bar{U}'\bar{U}]) \geq c\lambda(E[\{\bar{V}-E[\bar{V}|x]\}'\{\bar{V}-E[\bar{V}|x]\}]),$$

Now, in case i), $\bar{V}-E[\bar{V}|x] = \{p(\xi)-E[p(\xi)]\}'\otimes(\rho_\beta-E[\rho_\beta|\xi])'$, so that by eq. (A.19) and Lemma A.1 of Newey (1988a),

$$\lambda(E[U'U]) \geq \lambda(\ Var(p(\xi))\otimes E[\{\rho_\beta-E[\rho_\beta|\xi]\}\{\rho_\beta-E[\rho_\beta|\xi]\}']\ )$$

$$\geq c\cdot\lambda(Var(\tilde{p}(\xi))) = c\lambda((-E[\tilde{p}(\xi)],I)E[\bar{p}(\xi)\bar{p}(\xi)'](-E[\tilde{p}(\xi)],I)')$$

$$\geq c\lambda(E[\bar{p}(\xi)\bar{p}(\xi)'])\lambda(I + E[\tilde{p}(\xi)]E[\tilde{p}(\xi)]') \geq c\nu^{-C\nu}.$$

Next, under Assumption 5.2 ii), note that the smallest eigenvalue is invariant with respect to permutations of corresponding rows and columns, so that by eq. (A.19), it suffices to find a bound for the smallest eigenvalue of $E[\{\tilde{V}-E[\tilde{V}|x]\}'\{\tilde{V}-E[\tilde{V}|x]\}]$, where $\tilde{V} = (\rho_\beta-E[\rho_\beta|\xi])'\otimes p(\xi)'$. By ρ_β block diagonal, this is a block diagonal matrix, with s blocks, so that it suffices to find a bound on an arbitrary, say the first, diagonal block. This block is $B = \sum_{j=1}^{m}E[Var\{p(\xi)(\rho_{1\beta j}-E[\rho_{1\beta j}|\xi])|x\}]$, where m is the dimension of $\rho_{1\beta}$. Suppose, say, that $\rho_{1\beta 1}$ is the element of $\rho_{1\beta}$ satisfying the hypotheses of Assumption 5.2 ii). Then for $\|\lambda\| = 1$ and $g(\varepsilon,x) = \rho_{1\beta j}-E[\rho_{1\beta j}|\xi]$, Lemma A.2 gives

$$\lambda'B\lambda \geq E[Var\{\lambda'p(\xi)g(\varepsilon,x)|x\}]$$

$$= E[\{\lambda'p(\xi)\}^2]E[Var\{\{\lambda'p(\xi)/E[\{\lambda'p(\xi)\}^2]^{1/2}\}g(\varepsilon,x)|x\}]$$

$$\geq \lambda(E[p(\xi)p(\xi)'])c \geq c\nu^{-C\nu}.$$

Next, the rest of hypotheses i) and ii) of Lemma 5.1 will be verified. Note, by Assumption 5.1, that $\tilde{s}_k(\rho)$ is twice continuously differentiable in ρ, and that for a matrix R and vector μ,

$$\text{(A.20)} \qquad \|\tilde{s}_k(R\rho-\mu)\| = |p_0(R\rho-\mu)^{\lambda([k/s]+1)}| \leq C^{\nu},$$

$$\|\partial\tilde{s}_k(\rho)/\partial\rho\| = \|\partial[p_0(\rho)^{\lambda([k/s]+1)}/\partial\rho\| \leq \nu C^{\nu}, \quad \|\partial\tilde{s}_k(\rho)/\partial\rho\|^2 \leq \nu^2 C^{\nu}.$$

Let

$$b^0_{ij}(\gamma) = J_\beta(z_i,\beta)-J_\beta(\pi(x_j,\rho(z_i,\beta),\beta),x_j,\beta),$$

$$b^k_{ij}(\gamma) = [\rho_\beta(z_i,\beta)-\rho_\beta(\pi(x_j,\rho(z_i,\beta),\beta),x_j,\beta)]'\Sigma^{-1/2}\tilde{s}_k(\rho(z_i,\gamma)),$$

$$b_{ij}(\gamma) = (b^0_{ij}(\gamma),\ldots,b^K_{ij}(\gamma))'.$$

Note that $b_{ii}(\gamma) = 0$. It is easy, but tedious, to check that it follows from Assumptions 5.3 and 5.4 that $b_{ij}(\gamma)$ is twice continuously differentiable in γ and that for a bounded, convex neighborhood $\mathcal{N}$ of γ_0, such that the coordinate projection for β is contained in the neighborhoods of Assumptions 5.3 and 5.4,

$$\text{(A.21)} \qquad \|b_{ij}(\gamma)\|^2 \leq C(K+1)B_{ij}C^{\nu} \leq C\cdot\nu^{C\nu}B_{ij}$$

$$\|\partial b_{ij}(\gamma)/\partial\gamma\|^2 \leq C\cdot\nu^{C\nu}(B_{ij}+\tilde{B}_{ij}B_i), \quad \|\partial^2 b_{ij}(\gamma)/\partial\gamma^2\| \leq C\cdot\nu^{C\nu}(B_{ij}+\tilde{B}_{ij}B_i).$$

For the partitioning $\gamma = (\gamma_1',\gamma_2')'$, with $\gamma_1 = \beta$, note that

$$\hat{m}_n(\beta) = \textstyle\sum_{ij} b_{ij}(\beta,\hat{\gamma}_2)/n^2, \quad \hat{u}_i = \textstyle\sum_j [b_{ij}(\hat{\gamma})+b_{ji}(\hat{\gamma})]/n,$$

$$\hat{\mathcal{M}} = \partial\hat{m}_n(\hat{\beta})/\partial\beta = \textstyle\sum_{ij}\partial b_{ij}(\hat{\gamma})/\partial\beta/n^2, \quad \mathcal{M} = E[\partial b_{12}(\gamma_0)/\partial\beta].$$

By eq. (A.21) $\|\mathcal{M}\| \leq E[\|\partial b_{12}(\gamma_0)/\partial\beta\|] \leq 1 + C\cdot\nu^{C\nu}E[B_{12}+\tilde{B}_{12}B_1] = O(\nu^{C\nu})$, while by eq. (A.20), for ϵ from Assumption 5.4 and $|\circ|_\epsilon \equiv (E[\|\circ\|^{2+\epsilon}])^{1/(2+\epsilon)}$,

$|u|_\epsilon \le |u_0|_\epsilon + \max_{1\le k\le K}|u_k|_\epsilon \le 4(|J_\beta|_\epsilon + K\max_{1\le k\le K}|\rho_\beta'\Sigma^{-1/2},\tilde{s}_k(\xi)|_\epsilon \le C(|J_\beta|_\epsilon + \nu^{C\nu}|\rho_\beta|_\epsilon) = O(\nu^{C\nu})$, giving i). Next, by Assumption 5.6, Lemma A.4 (with the object in the lemma equal to $\partial b_{ij}(\gamma)/\partial\beta$ here), and eq. (A.21),

$$\text{(A.22)}\qquad \|\hat{M}-M\| = O_p(\nu^{C\nu}E[B_{12}+\tilde{B}_{12}B_1])\|\hat{\gamma}-\gamma_0\| + O_p(\nu^{C\nu}\{E[B_{12}+\tilde{B}_{12}B_1]\}^{1/2}/\sqrt{n})$$

$$= O_p(\nu^{C\nu}n^{-c}).$$

Also, by Assumption 5.6, Lemma A.5, eq. (A.21), and reasoning similar to that for eq. (A.22), $\sum_{i=1}^n\|\hat{u}_i-u_i\|^2/n = O_p(\nu^{C\nu}n^{-c})$. Also, by a mean value expansion with mean value $\bar{\beta}$, $\sqrt{n}\|\hat{m}_n(\hat{\beta})-\hat{m}_n(\beta_0)-M(\hat{\beta}-\beta_0)\| = \sqrt{n}\|[\partial\hat{m}_n(\bar{\beta})/\partial\beta-M](\hat{\beta}-\beta_0)\| \le \|\partial\hat{m}_n(\bar{\beta})/\partial\beta-M]\sqrt{n}\|\hat{\beta}-\beta_0\| = \|\partial\hat{m}_n(\bar{\beta})/\partial\beta-M\|O_p(1) = O_p(\nu^{C\nu}n^{-c})$, where the last equality holds by eq. (A.22) with $\bar{\beta}$ substituted for $\hat{\beta}$. Also,

$$\text{(A.23)}\qquad \sqrt{n}\|\hat{m}_n(\beta_0) - \sum_{i=1}^n u_i/n\| \le \sqrt{n}\|\sum_{ij}b_{ij}(\gamma_0)/n^2 - \sum_{i=1}^n u_i/n\|$$

$$+ \sqrt{n}\|\hat{m}_n(\beta_0) - \sum_{ij}b_{ij}(\gamma_0)/n^2\| \equiv T_1 + T_2.$$

and $T_1 = O_p(\nu^{C\nu}n^{-c})$ by Lemma A.3, Assumption 5.6, and eq. (A.21), and $u_i = E[b_{ij}(\gamma_0)+b_{ji}(\gamma_0)|z_i]$, for $j \ne i$. Also, note that $\partial b^0_{ij}(\gamma_0)/\partial\gamma_2 = 0$ and for $k \ge 1$, $\partial b^k_{ij}(\gamma_0)/\partial\gamma_2$ is a linear combination of $b^\rho_{ij}(\beta_0)'\otimes s_k(\xi_i)$ and $b^\rho_{ij}(\beta_0)\otimes\partial s_k(\xi_i)/\partial\xi\otimes(\xi_i',1)'$, each of which have expectation zero independence of ε and x; e.g. for $j \ne i$, $E[b^\rho_{12}(\beta_0)'\otimes s_k(\xi_1)] = E[E[b^\rho_{12}(\beta_0)'\otimes s_k(\xi_1)|z_1]] = E[\{\rho_\beta-E[\rho_\beta|\varepsilon]\}\otimes s_k(\xi)] = 0$. Thus, for $\tilde{b}_{i.} = E[\partial b^k_{ij}(\gamma_0)/\partial\gamma_2|z_i]$ and $\tilde{b}_{.i} = E[\partial b^k_{ji}(\gamma_0)/\partial\gamma_2|z_i]$, $(j \ne i)$, $E[\tilde{b}_{i.}] = E[\tilde{b}_{.i}] = 0$. Furthermore, by eq. (A.21), $E[\|\partial b^k_{ij}(\gamma_0)/\partial\gamma_2\|^2] = O(\nu^{C\nu})$, implying $E[\|\tilde{b}_{i.}+\tilde{b}_{.i}\|^2] = O(\nu^{C\nu})$. Then by a second order mean-value expansion,

$$(A.24)\quad T_2 \le \sqrt{n}\|\textstyle\sum_{ij}\partial b_{ij}(\gamma_0)/\partial\gamma_2/n^2\|\cdot\|\hat{\gamma}-\gamma_0\| + \sqrt{n}\|\textstyle\sum_{ij}\partial b_{ij}(\bar{\gamma})/\partial\gamma_2/n^2\|\cdot\|\hat{\gamma}-\gamma_0\|^2$$

$$\le \|\textstyle\sum_{ij}\partial b_{ij}(\gamma_0)/\partial\gamma_2/n^2\|O_p(n^{1/4-\epsilon}) + \textstyle\sum_{ij}\|\partial b_{ij}(\bar{\gamma})/\partial\gamma\|/n^2 O_p(n^{-2\epsilon})$$

$$\le \|\textstyle\sum_{ij}\partial b_{ij}(\gamma_0)/\partial\gamma_2/n^2 - \sum_i(\tilde{b}_{i.}+\tilde{b}_{.i})/n\|O_p(n^{1/4-\epsilon})$$

$$+ \|\textstyle\sum_i(\tilde{b}_{i.}+\tilde{b}_{.i})/n\| + O_p(n^{-c_\nu C\nu})$$

$$= O_p(\{E[\|\partial b_{ij}(\gamma_0)/\partial\gamma_2\|^2]\}^{1/2}n^{-c}) + O_p(n^{-c_\nu C\nu}) = O_p(n^{-c_\nu C\nu}).$$

Lemma 5.1 ii) now follows by eqs. (A.23) and (A.24).

Turning to Lemma 5.1 iii), let $b_{12}(\beta) = b_{12}(\beta,\gamma_{20})$. It follows by Assumptions 5.3 and 5.4 and Bartle (1966, Corollary 5.9) that $E[b_{12}(\beta,\gamma_{20})]$ is differentiable in β and $\partial E[b_{12}(\beta,\gamma_{20})]/\partial\beta|_{\beta_0} = \mathcal{M}$. Also, $b_{12}(\beta)$ is continuous in β,

and by Assumption 5.5 and $\tilde{s}_k(\rho)$ bounded, for $\beta \in \mathcal{N}$, $\|b_{12}(\beta)\|^2 \le C(\sup_{\mathcal{N}}\|b^J_{12}(\beta)\|^2 + \sup_{\mathcal{N}}\|b^\rho_{12}(\beta)\|^2) \equiv \bar{B}_{12}$, and $E[\bar{B}_{12}|\beta]$ is continuous in β. Furthermore, by $f(z|\beta)$ smooth, so is $f(z_1|\beta)f(z_2|\beta)$, with score $S_\beta(z_1)+S_\beta(z_2)$. Then by Lemma C.3 of Newey (1990b),

$$\mathcal{M} = -E[b_{12}(\beta_0)\{S_\beta(z_1)+S_\beta(z_2)\}']$$

$$= -E[\{E[b_{12}(\beta_0)|z_1]+E[b_{21}(\beta_0)|z_1]\}S_\beta(z_1)'] = -E[\omega S'_\beta] = -E[\omega S'],$$

where the last equality follows by eq. (2.3).

Finally, Lemma 5.1 iv) holds by Corollary 4.4 and the remarks that follow it, while the remainder of the conditions of Lemma 5.1 hold by hypothesis. ■

Appendix B: Efficiency Bound Theorem for CDR Model

First, mean-square (m.s.) smoothness will be defined. M.s. continuity and (Frechet) differentiability of functions of θ are these properties for the mean-square norm on the space of measurable functions. The following condition for smoothness and regularity of parametric submodels is like Ibragimov and Hasminskii (1981, Ch. 7), referred to as IH henceforth. Suppose that $\mathcal{P}_\theta = \{\ell(z|\theta) : \theta \in \Theta\}$ is a family of densities $\ell(z|\theta)$ with respect to some measure, and let dz denote integration with respect to that measure.

Definition B.1: $\mathcal{P}_\theta$ is *smooth* if Θ is open and i) $\ell(z|\theta)$ is continuous on Θ a.s.; ii) $\ell(z|\theta)^{1/2}$ is m.s. differentiable with respect to θ on Θ with derivative $\psi(z,\theta)$, i.e. $\int\|\psi(z,\theta)\|^2 dz$ is finite on Θ and for each θ and $\theta_i \longrightarrow \theta$, $\int[\ell(z|\theta_i)^{1/2}-\ell(z|\theta)^{1/2}-\psi(z,\theta)'(\theta_i-\theta)]^2 dz/\|\theta_i-\theta\|^2 \longrightarrow 0$; iii) $\psi(z,\theta)$ is m.s. continuous. Also, for smooth $\mathcal{P}_\theta$ the score is defined by $S_\theta \equiv 2\cdot 1(\ell(z|\theta)>0)\psi(z,\theta)/\ell(z|\theta)^{1/2}$ and the information matrix by $\int S_\theta S_\theta' \ell(z|\theta)dz$. $\mathcal{P}_\theta$ is *regular* if it is smooth and the information matrix is nonsingular on Θ.

See IH for further details.

To emphasize that the CDR model and the form of its efficiency bound are invariant to transformations of the dependent variable, such a transformation will be explicitly considered here. Let $\tau(y)$ equal the identity map on a set $\mathcal{Y}$ and $\tau(\mathcal{Y}^c)$ be discrete, and suppose that

$$(B.1) \qquad y = \tau(y^*), \qquad y^*|z \sim F(y^*|v(x,\beta_0)).$$

This specification includes a number of models of interest, such as ordered choice and regression with fixed censoring. Suppose that (y^*,x) is absolutely continuous with respect to the product of some measure for y^* (e.g. Lebesgue measure) and the marginal distribution for x. Let $f_0(y^*|v)$ be the conditional density of y^* given x, and $f_0(x)$ the marginal density

of x. Suppose that for $v \in \mathbb{R}$, $f_0(y^*|v)$ gives a smooth parametric model for densities for y with parameter v and score $s_v(y^*,v)$. Let $S_v = E[s_v(y^*,v)|y,v]$ be the score with respect to v for y $(= \tau(y^*))$, a well known conditional expectation formula for the score for the observed data in terms of the latent density (e.g. IH, Theorem 7.2). The following result proves the validity of the ef..cient score formula in equation (2.7).

Theorem B.1: Suppose i) $v(x,\beta)$ *is continuously differentiable in* β *in a neighborhood* N *of* β_0 a.s.-x, $\sup_N \|v_\beta(x,\beta)\| = M(x)$ *satisfies* $E[M(x)^2] < +\infty$; ii) *The parametric submodels correspond to a family of latent densities* $f(y^*|v(x,\beta),\eta)f(x|\eta)$ *where* $f(y^*|v,\eta)$ *is smooth with bounded information for* $v \in \mathbb{R}$ *and* $f(y^*|v(x,\beta_0),\eta_1)f(x|\eta_2)$ *is smooth in the separate* η_1, η_2 *arguments; iii)* $E[SS']$ *is nonsingular for* S *given in equation (3.10). Then the* S *is the efficient score. Furthermore, if* $E_\beta[\|S\|^2]$ *is continuous at* β_0 *then this conclusion remains true for the class of parametric submodels such that* $E_\theta[\|S\|^2]$ *is continuous at* θ_0.

Proof: Proceed by verifying the hypotheses of the projection result cited following eq. (2.1). First, to show smoothness of $\ell(z|\beta)$, let $z^* = (y^*,x)$. By ii), $f_0(y^*|v)f_0(x)$ is smooth with bounded information, so by Newey (1990b, Lemma C.5), $f_0(z^*|\beta) = f_0(y^*|v(x,\beta))f_0(x)$ is smooth with score $s_v v_\beta$. It then follows as in the proof of Theorem 7.2 of IH that $\ell(z|\beta)$ is smooth with score $E[s_v v_\beta|y,x] = S_v v_\beta$.

Next, consider a parametric submodel as spe..fied in the statement of the Theorem. By ii) and Newey (1990b, Lemma C.5), $f(y^*|v(x,\beta),\eta)$ is smooth in β and η, with score $s_{\theta 1} = (s_v v_\beta', s_{\eta 1}')'$. By Lemma 7.2 of IH and $1 = \int f(y^*|v,\eta)dy^*$, differentiation of this identity with respect to v and η gives $E[s_v|x] = 0$ and $E[s_{\eta 1}|x] = 0$. It also follows by taking an almost sure convergent subsequence of the mean-square convergent sequence in the

definition of the score that s_v and $s_{\eta 1}$ depend only on y^* and v. Then since the conditional distribution of y^* given x depends only on v, and s_v depends only on y^* and v, $S_v = E[s_v|z]$ and $S_{\eta 1} = E[s_{\eta 1}|z]$ will depend only on y and v, and by iterated expectations $E[S_v|x] = E[s_v|x] = 0$ and $E[S_{\eta 1}|x] = 0$. Now, for the latent model $f(y^*|v(x,\beta_0),\eta_1)f(x|\eta_2)$, it can be shown that by these results and an almost-sure subsequence argument that the score for η_1 is $s_{\eta 1}$, that the score $s_{\eta 2}$ for η_2 is a function only of x, and $E[s_{\eta 2}] = 0$. Then by the chain rule $\ell(z^*|\theta)$ is smooth with score s_η for η satisfying $s_\eta = s_{\eta 1}+s_{\eta 2}$. Furthermore, it follows as in the proof of Theorem 7.2 of IH that $\ell(z|\theta)$ is smooth with score for η given by $S_\eta = E[s_\eta|z] = E[s_{\eta 1}|z] + s_{\eta 2}$. Then it follows from the above noted properties of the two terms in S_η that any conformable linear combination of the score for η will be an element of $\mathcal{T}$.

Next, to show that any element of $\mathcal{T}$ can be approximated arbitrarily closely in mean square by the score for a regular parametric submodel, consider a submodel with $\ell(z^*|\beta,\eta_1,\eta_2) = f(z^*|\beta)\Delta(z^*,\beta,\eta_1,\eta_2)$ and $\Delta(z^*,\theta) = [1+\eta_1'h_1(y^*,v(x,\beta))][1+\eta_2'h_2(x)]$ where $h_2(x)$ is bounded, $E[h_2(x)] = 0$, $h_1(y^*,v) = a(y^*,v) - \int a(t,v)f(t|v)dt$, and $a(t,v)$ is bounded and continuously differentiable in t and v with bounded derivatives. By smoothness of $f(y|v)$ and Lemma 7.2 of IH, $h_1(y^*,v)$ is continuously differentiable in v, with derivative $a_v - E[a_v|v] + E[as_v|v]$, where $a_v = \partial a(y^*,v)/\partial v$ and s_v is the score for $f_0(y^*|v)$. This derivative is dominated by $C(1+E[\|s_v\||v] \le C(1+\{E[s_v's_v|v]\}^{1/2})$, which is bounded by boundedness of the information matrix for $f(y^*|v)$. Let $f(y^*|v,\eta) = f_0(y^*|v)[1+\eta'h_1(y^*,v)]$. By Newey (1990b, Lemma C.4), $f(y^*|v,\eta)$ is smooth on $v \in \mathbb{R}$ and a bounded set for η containing $\eta_0 = 0$. Also, by boundedness of h_1 and its derivative with respect to v, the information matrix for $f(y^*|v,\eta)$ will be bounded, so that this conditional density family satisfies

the hypotheses of the theorem. Also, by Newey (1990b, Lemma C.4), $f(z^*|\beta,\eta_1,\eta_2)$ is smooth with score $h_1(y^*,v)$ for η_1 and $h_2(x)$ for η_2. It then follows as in the previous paragraph that $f(z^*|\theta)$ is smooth with $S_\eta = E[h_1(y^*,v)|y,v] + h_2(x)$. Let $t = t_1(y,v) + t_2(x)$ be an element of $\mathcal{T}$ and consider $\epsilon > 0$. By Newey (1990b, Lemma C.6), there exists $a = a(y^*,v)$ such that for $t_1^* \equiv t_1(\tau(y^*),v))$, $E[\|t_1^*-a\|^2] < \epsilon$, implying

$$(B.2) \qquad E[\|t_1-E[a|y,v]\|^2] \le E[E[\|t_1^*-a\|^2|y,v]] = E[\|t_1^*-a\|^2] < \epsilon.$$

Then by $h_1(y^*,v) = a(y^*,v)-E[a|v]$ and $E[t_1^*|x] = E[t_1|x] = 0$

$$(B.3) \qquad E[\|t_1-E[h_1|y,v]\|^2] \le 2E[\|t_1-E[a|y,v]\|^2] + 2E[\|E[a|v]\|^2]$$

$$< 2\epsilon + 2E[\|E[E[t_1^*-a|x]|v]\|^2] \le 2\epsilon + E[\|t_1^*-a\|^2] \le 4\epsilon,$$

Since by Newey (1990b, Lemma C.6) there exists h_2 such that $E[\|t_2-h_2\|^2] \le \epsilon$, the triangle inequality and ϵ arbitrary imply t is an element of the mean-square closure of the set of scores, and hence is in the tangent set.

Next, to verify the projection formula in eq. (2.7), let $\bar{R} = E[R|y,v]-E[R|v]$ $E[R|x]-E[R]$. Note that for any $A(y,v)$, by the distribution of y given x depending only on v,

$$(B.4) \qquad E[A(y,v)|x] = E[A(y,v)|v].$$

Thus (for $A = E[R|y,v]$), it follows that $\bar{R} \in \mathcal{T}$. Also, note that for any vectors $t = t(y,v)$ and $A = A(x)$, by eq. (B.4), $E[\{E[A|y,v]-E[A|v]\}'t] = E[A't] - E[E[A'|v]E[t|v]] = E[A'E[t|x]] - E[A'E[t|v]] = 0$, so that

$$(B.5) \qquad E[A(x)|y,v] = E[A(x)|v].$$

For $U = R-E[R|x]$, it follows by eq. (B.5) that $R-\bar{R} = U-E[U|y,v] + E[R]$. Then by equation (B.4) and $E[U|v] = 0$, for any $t \in \mathcal{T}$,

(B.6) $$E[(R-\bar{R})'t] = E[(U-E[U|y,v])'t] = E[E[U|y,v]'t_2(x)]$$

$$= E[E[E[U|y,v]|x]'t_2(x)] = E[E[U|v]'t_2(x)] = 0,$$

Thus, $R-\bar{R}$ is orthogonal to $\mathcal{T}$, proving $\bar{R} = \text{Proj}(R|\mathcal{T})$.

To verify the formula for S, recall from above that $E[S_v|x] = 0$, implying $E[S_\beta|x] = 0$, while $E[v_\beta|y,v] = E[v_\beta|v]$ follows by eq. (B.5).

To verify regularity of $\ell(z|\beta)$, recall from above that $\ell(z|\beta)$ is smooth, to that it remains to show that the information matrix is nonsingular on an open set containing β_0. It follows as in eq. (B.6) that $E[St'] = 0$ for all $t \in \mathcal{T}$. Therefore $E[S_\beta S_\beta] - E[SS']$ is positive semidefinite, so a nonsingular information matrix on some neighborhood of β_0 follows by nonsingularity of $E[SS']$ and continuity of the information matrix for smooth models.

For the second conclusion, it suffices to show that $E_\theta[\|S\|^2]$ is continuous at θ_0 for the class of parametric submodels that were used above to approximate the elements of the tangent set. For this class of parametric submodels it follows by h_1 and h_2 bounded that

(B.7) $$|f(z^*|\theta)-f(z^*|\beta)| \le f(z^*|\beta)|1-\Delta(z^*,\theta)| \le Cf(z^*|\beta)\|\eta\|$$

for $\|\eta\|$ small enough, implying $|f(z|\theta)-f(z|\beta)| \le Cf(z^*|\beta)\|\eta\|$, implying

$$E_\theta[\|S\|^2] - E_\beta[\|S\|^2]| \le C\|\eta\|E_\beta[\|S\|^2].$$

Thus, continuity of $E_\theta[\|S\|^2]$ at $\theta_0 = (\beta_0', 0)'$ follows by continuity of $E_\beta[\|S\|^2]$ at β_0. ■

References

Amemiya, T. (1974): "The Nonlinear Two-Stage Least-Squares Estimator," *Journal of Econometrics*, 2, 105-110.

Amemiya, T. (1977): "The Maximum Likelihood and Nonlinear Three-Stage Least Squares Estimator in the General Nonlinear Simultaneous Equations Model," *Econometrica*, 45, 955-968.

Amemiya, T. and J.L. Powell (1981): "A Comparison of the Box-Cox Maximum Likelihood Estimator and the Non-Linear Two-Stage Least Squares Estimator," *Journal of Econometrics*, 17, 351-381.

Andrews, D.W.K. (1989): "Asymptotics for Semiparametric Econometric Models: I Estimation," mimeo, Cowles Foundation, Yale University.

Bartle, R. G. (1966): *The Elements of Integration*, New York: John Wiley and Sons.

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

Beran, R. (1976): "Adaptive Estimates for Autoregressive Processes," *Annals of the Institute of Statistical Mathematics*, 26, 77-89.

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

Bickel P., C.A.J. Klaasen, Y. Ritov, and J.A. Wellner (1988): "Efficient and Adaptive Inference in Semiparametric Models" Forthcoming monograph, Johns Hopkins University Press.

Box, G.E.P. and D.R. Cox (1964): "An Analysis of Transformations," *Journal of the Royal Statistical Society,* Series B, 26, 211-252.

Buckley, J. and I. James (1979): "Linear Regression with Censored Data," *Biometrika*, 66, 429-436.

Carroll, R.J. and D. Ruppert (1984): "Power Transformations When Fitting Theoretical Models to Data," *Journal of the American Statistical Assocation*, 79, 321-328.

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 Econometrics*, 34, 305-334.

Chamberlain, G. (1987b): "Efficiency Bounds for Semiparametric Regression," manuscript, Department of Economics, University of Wisconsin

Cox D.R. (1975): "Partial Likelihood" *Biometrika*, 62, 269-276

Cragg, J.G. (1983): "More Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form," *Econometrica*, 751-764.

Gallant, A.R. (1980): "Explicit Estimators of Parametric Functions in Nonlinear Regression," *Journal of the American Statistical Association*, 75, 182-193.

Hajek, J. (1970): "A Characterization of Limiting Distributions of Regular Estimates," Z. *Wahrscheinlichkeitstheorie verw. Geb.*, 14, 323-330.

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

Hardle, W. and J.S. Marron (1985): "Optimal Bandwidth Selection in Nonparametric Regression Function Estimation," *Annals of Statistics*, 13, 1465-1485.

Hayashi, F., and C. Sims (1983): "Nearly Efficient Estimation of Time Series Models with Predetermined, But Not Exogenous, Instruments," *Econometrica*, 51, 783-798.

Hinkley, D.V. (1975): "On Power Transformations to Symmetry," *Biometrika*, 62, 101-111.

Hsieh, D., and C. Manski (1987): "Monte-Carlo Evidence on Adaptive Maximum Likelihood Estimation of a Regression," *Annals of Statistics*, forthcoming.

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

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

Ichimura, H. (1986): "Estimation of Index Model Coefficients," manuscript, Department of Economics, MIT.

Kelejian, H.H. (1971): "Two-Stage Least Squares and Econometric Systems Linear in the Parameters but Nonlinear in the Endogenous Variables," *Journal of the American Statistical Association*, 66, 373-374.

Klaassen, C.J. (1987), "Consistent Estimation of the Influence Function of Locally Asymptotically Linear Estimators," *Annals of Statistics*, 15, 1548-1562.

Koshevnik, Yu. A. and B. Ya. Levit (1976): "On a Non-parametric Analouge of the Information Matrix," *Theory of Probability and Applications*, 21, 738-753.

MaCurdy, T. E. (1982): "Using Information on the Moments of the Disturbance to Increase the Efficiency of Estimation," Stanford University, manuscript.

Manski, C. (1988): "Identification of Binary Response Models," *Journal of the American Statistical Association*, forthcoming.

McCullagh, P. and J.A. Nelder (1983): *Generalized Linear Models*, New York: Chapman and Hall.

Newey, W. K. (1988a): "Adaptive Estimation of Regression Models Via Moment Restrictions," *Journal of Econometrics*, 38, 301-339.

Newey, W.K. (1988b): "Two-Step Series Estimation of Sample Selection Models," Princeton University, manuscript.

Newey, W.K. (1989a): "Locally Efficient Residual-Based Estimation of Nonlinear Simultaneous Equations Models," working paper, Bell Communications Research.

Newey, W.K. (1989b): "The Asymptotic Variance of Semiparametric Estimators," Princeton University, Econometric Research Program Memo. No. 346.

Newey, W.K. (1989c): "Efficiency in Limited Dependent Variable Models Under Conditional Location Restrictions," Princeton University, manuscript.

Newey, W.K. (1990a): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics,* forthcoming.

Newey, W.K. (1990b): "Efficient Estimation of Tobit Models Under Conditional Symmetry," in W. Barnett, J. Powell, G. Tauchen eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, New York: Cambridge University Press, forthcoming.

Newey, W.K. and T. Stoker (1989): "Efficiency Properties of Average Derivative Estimators," working paper, Sloan School of Management, MIT.

Pakes, A. and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027-1057.

Pfanzagl, J. and W. Wefelmeyer (1982): *Contributions to a General Asymptotic Statistical Theory*, New York: Springer-Verlag.

Pitman, E.J.G. (1979): *Some Basic Theory for Statistical Inference*, London: Chapman and Hall.

Powell, M.J.D. (1981): *Approximation Theory and Methods*, Cambridge, England: Cambridge University Press.

Powell, J.L. (1984): "Least Absolute Deviations Estimation for the Censored Regression Model," *Journal of Econometrics*, 25, 303-325.

Powell, J.L. (1990): "Estimation of Monotonic Regression Models Under Conditional Quantile Restrictions," in W. Barnett, J. Powell, G. Tauchen eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, New York: Cambridge University Press, forthcoming.

Robinson, P.M. (1989): "Best Nonlinear Three-Stage Least Squares Estimation of Certain Econometric Models," preprint, London School of Economics

Ruppert, D. and Aldershof, B. (1989), "Transformations to Symmetry and Homoscedasticity," *Journal of the American Statistical Association,* 84, 437-446.

Sargan, J.D. (1959): "The Estimation of Relationships with Autocorrelated Residuals by the Use of Instrumental Variables," *Journal of the Royal Statistical Society, Series B*, 21, 91-105.

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," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, Berkeley: University of California Press.

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

Taylor, J.M.G. (1985): "Power Transformations to Symmetry," *Biometrika*, 72, 145-152.

Table 1: Root-Mean Square Errors Relative to OLS and the Distribution of Cross-Validated Choice of K.

Gaussian

	CV	K = 1	K = 2	K = 3	K = 4	K = 5
RMSE	1.05	1.03	1.06	1.08	1.11	1.12
Freq $\hat{K}$		.73	.11	.09	.04	.03

Gaussian Scale Mixture

	CV	K = 1	K = 2	K = 3	K = 4	K = 5
RMSE	.43	.51	.46	.41	.42	.45
Freq $\hat{K}$		.26	.22	.35	.14	.03

Lognormal

	CV	K = 1	K = 2	K = 3	K = 4	K = 5
RMSE	.35	.69	.41	.34	.34	.42
Freq $\hat{K}$		.00	.18	.29	.35	.18

Gaussian Location Mixture

	CV	K = 1	K = 2	K = 3	K = 4	K = 5
RMSE	.40	1.53	1.32	.36	.38	.41
Freq $\hat{K}$		.00	.00	.74	.18	.08