Efficient Semiparametric Estimation Via Moment Restrictions*

Whitney K. Newey
Department of Economics
M.I.T, E52-262D
Cambridge, MA 02139
and
Tasneem Chipty
Charles River Associates
John Hancock Tower, T-33
200 Clarendon St.
Boston, MA 02116

August, 1999
Revised, October 2002

Abstract

Conditional moment restrictions can be combined through GMM estimation to construct more efficient semiparametric estimators. This paper is about what happens as the number of conditional moment restrictions increases. The limit of the asymptotic variance is derived and it is shown that the limit equals the semiparametric bound when the moment restrictions characterize the semiparametric model. These results are applied to transformed, censored, and truncated regression models. In each case a set of moment conditions is given that leads to approximate efficiency of the GMM estimator. Asymptotic efficiency is shown, with $J^2/n \to 0$ being sufficient for valid asymptotic inference in one important case, where J is the number of moment conditions. A sample selection application is given.

*This work was partially completed while Newey was a Fellow at the Center for Advanced Study in the Behavioral Sciences. Also, the NSF provided. financial support. This paper was presented at the 2000 World Congress of the Econometric Society. Helpful comments were provided by R. Blundell, L. Hansen, J. Horowitz, two referees, and seminar participants at USC.

1 Introduction

Generalized method of moments (GMM) provides a useful way of constructing efficient estimators, by combining moment restrictions. This approach is parsimonious and has good small sample properties in many cases (see Chamberlain, 1987 and Newey, 1988, 1993). It is particularly useful in models where the efficiency bound is complicated, so that direct construction of an efficient estimator is difficult, but there are relatively simple moment conditions that can be used for estimation. There are many important examples of such models, including several considered in this paper.

The purpose of this paper is to consider efficient estimation with an infinite sequence of conditional moment restrictions depending on nuisance parameters. We show that the limit of the GMM asymptotic variance equals the semiparametric bound when the moment conditions characterize the semiparametric model in a certain local sense, discussed below. This result enables one to check to see whether a particular sequence of moment conditions has "complete information" about parameters of interest, in the sense that they lead to full efficiency. For example, we find that, in the censored regression model with disturbance independent of regressors, the moment restrictions from Powell's (1986) quantile estimators can be combined to achieve efficiency, despite their regressor trimming. We also show that in truncated regression models, moment restrictions like those of Newey (1987) can be combined for efficiency. In both of these examples it is relatively simple to check efficiency, despite the complicated nature of the bounds. This simplicity results from the efficiency characterization being the dual of the one adopted by Chamberlain (1987).

Conditions are given for asymptotic efficiency with estimated unconditional moment restrictions. In particular we show that for Chamberlain's (1987) estimator with spline instruments, $J^2/n \to 0$ suffices for valid asymptotic inference, where J is the number of moment conditions. These results improve on those of Newey (1988, 1993), Hahn (1997), and Koenker and Machado (1999), and are the most general possible when endogeneity is present.

We also consider an application to sample selection based on the Mroz (1987) women's labor supply paper, where the residual density estimator is sharply bimodal. We find that standard errors are greatly reduced when nonlinear moments are used, being much smaller than those of the Newey, Powell, and Walker (1990). This result shows that precise results can be obtained by semiparametric estimation in these data, when nonlinear moments are used.

2 Combining Moment Restrictions

To describe the general type of estimator we consider let z denote a single data observation, β a $q \times 1$ parameter vector, and $\Gamma = (\gamma_1, \gamma_2 ...)$ a sequence of scalar parameters, and $(\rho_1(z, \beta, \gamma), \rho_2(z, \beta, \gamma), ...)$ a sequence of functions, each of which depends only on a finite number of elements of Γ. Also, let x denote a vector of conditioning variables and β_0 and γ_0 denote true values. The estimators we consider are based on the conditional moment restrictions

$$E[\rho_j(z, \beta_0, \gamma_0)|x] = 0, (j = 1, 2, ...). \tag{1}$$

The case of unconditional moment restrictions is included as a special case where $x = 1$.

A finite number of these moment conditions can be used to form a GMM estimator. Let J denote a positive integer, γ_J the $r \times 1$ subvector of Γ that enters the first J functions, $\theta = (\beta', \gamma_J')'$, and $\rho(z, \theta) = (\rho_1(z, \beta, \gamma), ..., \rho_J(z, \beta, \gamma))'$, where indexing by J of r, θ, and ρ is suppressed for notational convenience. Also, let $A(x)$ be an matrix of functions of the conditioning variables with J columns. Then equation (1) implies the unconditional moment restrictions $E[A(x)\rho(z, \theta_0)] = 0$. Let $(z_1, ..., z_n)$ denote the data and $\hat{g}_n(\theta) = \sum_{i=1}^n A(x_i)\rho(z_i, \theta)/n$. The unconditional restrictions can be combined to form an estimator $\hat{\theta}$ in the now familiar way given by Hansen (1982), as

$$\hat{\theta} = \arg\min_{\theta} \hat{g}_n(\theta)' W \hat{g}_n(\theta), \tag{2}$$

where W is a positive semi-definite matrix.

In this paper we will focus on the case where $A(x)$ is efficient, i.e. minimizes the asymptotic variance among all possible $A(x)$. We maintain this focus because the efficient $A(x)$ can generally be estimated without affecting efficiency. To describe the efficient instruments, let $\Omega(x) = E[\rho(z, \theta_0)\rho(z, \theta_0)'|x]$, $D(x) = \partial E[\rho(z, \beta, \gamma_0)|x]/\partial\beta|_{\beta=\beta_0}$, $H(x) = \partial E[\rho(z, \beta_0, \gamma)|x]/\partial\gamma_J|_{\gamma=\gamma_0}$, and $G(x) = [D(x), H(x)]$. Then, as shown by Chamberlain (1987) (and Newey, 2001 in the singular $\Omega(x)$ case), the choice of $A(x)$ that minimizes the asymptotic variance of $\hat{\theta}$ is

$$A^*(x) = G(x)'\Omega(x)^-,$$

where for a matrix B, B^- denotes any generalized inverse, satisfying $BB^-B = B$. In this paper we will give conditions for a fixed subvector of $\hat{\theta}$ to be asymptotically efficient in a semiparametric model as J grows. These efficiency results will apply to a fixed subvector of Γ as well as to the parameters β that are common to the moment conditions. Specifically, we will consider the asymptotic efficiency of $\tilde{\theta} = [I, 0]\hat{\theta}$, where the dimension of $\tilde{\theta}$ (i.e. of I) remains fixed as J grows.

In general the optimal function $A^*(x)$ will need to be estimated. It is well known from Hansen (1982) that this estimation does not affect the asymptotic variance of $\hat{\theta}$ in the unconditional case, where $x = 1$ and A^* is a matrix of constants. It has also been shown that this result also holds when x is non-trivial and $A^*(x)$ is estimated nonparametrically, in Newey (1993). This justifies us in ignoring the estimation of $A^*(x)$ in the comparison of asymptotic variances. Furthermore, these comparisons will also be valid for the case where $\rho(z, \theta)$ has components that need to be estimated, even nonparametric ones, as long as this estimation does not affect the asymptotic variance.

In the case where $\rho(z, \theta)$ is nonlinear, computation can be simplified without affecting efficiency by using a one-step method. Let $\hat{G}(x)$, $\hat{\Omega}(x)^-$, and $\hat{A}(x) = \hat{G}(x)'\hat{\Omega}(x)^-$ be estimators of the respective functions and $\bar{\theta}$ an initial root-n consistent estimator. Then the one-step estimator

$$\hat{\theta} = \bar{\theta} - \left[\sum_{i=1}^{n} \hat{A}(x_i)\partial\rho(z_i, \bar{\theta})/\partial\theta\right]^{-1} \sum_{i=1}^{n} \hat{A}(x_i)\rho(z_i, \bar{\theta}), \tag{3}$$

will be asymptotically equivalent to the optimal GMM estimator with $A(x) = A^*(x)$. Furthermore, this equivalence will continue to hold if $\rho(z, \theta)$ is replaced by an estimator in such a way that the asymptotic variance is unaffected.

Two examples are useful for illustration. The first is the semiparametric transformation model with a parametric disturbance distribution. In this model $z = (y, x)$ for a scalar dependent variable y and there is an unknown, monotonic increasing function $\tau(\cdot)$ satisfying

$$\tau(y) = x'\delta_0 + \varepsilon, \ \varepsilon \text{ and } x \text{ are independent, } \varepsilon \text{ has p.d.f. } g(\varepsilon, \lambda_0).$$

This model includes the proportional hazards model as a special case, where ε has an extreme value distribution. It also includes proportional hazards with a known distribution for the heterogeneity. In these cases $\tau(y)$ will be equal to the log of the integrated baseline hazard at y. Estimation of this model has been considered previously by Bickel et. al. (1993), where further references are given. In general the efficiency bound for this model is complicated, as are the efficient estimators that have previously been proposed (except for certain special cases), while there are simple moment conditions that can be used for approximately efficient estimation.

Parametric conditional moment restrictions can be obtained by considering the probability that y lies in intervals, as in Han and Hausman (1990). Consider a sequence $(\bar{y}_j)_{j=1}^{\infty}$ of scalars. Let $\beta = (\delta', \lambda')'$, $G(u, \lambda) = \int_{-\infty}^{u} g(\varepsilon, \lambda) d\varepsilon$ be the CDF corresponding to $g(\varepsilon, \lambda)$, and

$$\rho_j(z, \beta, \gamma_j) = 1(y \leq \bar{y}_j) - G(\gamma_j - x'\delta, \lambda).$$

These residuals will satisfy the conditional moment restrictions of equation (1) for $\gamma_{j0} = \tau(\bar{y}_j)$. Here the parameters γ_j represent values of the transformation at various points. Therefore, $\tilde{\theta}$ may include estimators of the transformation at certain points, corresponding to estimators of the integrated hazard in duration models. It turns out that these moment restrictions can be used to approximately attain the semiparametric bound for the transformation model, including for the estimators of the transformation values.

The optimal GMM estimator based on these conditions will be equivalent to the maximum likelihood estimator (MLE) for the ordered choice model based on the intervals between the cutoffs y_j. Specifically, for $\theta = (\beta', \gamma_1, ..., \gamma_J)'$, $P_j(x, \theta) = G(\gamma_j - x'\delta, \lambda) - G(\gamma_{j-1} - x'\delta, \lambda)$, $(j = 1, ..., J+1)$, with $\gamma_0 = -\infty$ and $\gamma_{J+1} = +\infty$, the ordered choice MLE will satisfy

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \sum_{j=1}^{J+1} 1(\bar{y}_{j-1} \leq y_i < \bar{y}_j) lnP_j(x_i, \theta).$$

The first-order conditions for this MLE are (see Appendix B),

$$\sum_{i=1}^{n} \sum_{j=1}^{J} \rho_j(z_i, \hat{\theta}) \partial \ln[P_j(x_i, \hat{\theta}) / P_{j+1}(x_i, \hat{\theta})] / \partial\theta = 0.$$

This has the form of a GMM estimator where $A(x)$ has j^{th} column $\partial \ln[P_j(x, \hat{\theta}) / P_{j+1}(x, \hat{\theta})] / \partial\theta$. By efficiency of MLE we know that this $A(x)$ must be efficient (where estimation of $A(x)$ does not affect the efficiency), making the MLE asymptotically equivalent to the GMM estimator with $A^*(x)$. Consequently, the efficiency results for GMM given below will apply to the ordered choice MLE. It will be shown if $(\bar{y}_j)_{j=1}^{\infty}$ is dense in $\Re$ then as $J \longrightarrow \infty$ the asymptotic variance of a fixed subvector of $\hat{\theta}$ approaches the semiparametric bound

The second example is the conditional mean index model of Ichimura (1993), where

$$E[y|w] = E[y|v(w, \beta_0)], \; E[y^2] < \infty, \tag{4}$$

for some vector of regressors w and known vector of functions $v(w, \beta)$. A simple approach to efficient estimation can be based on unconditional moment restrictions. This model implies that for any function $a(w)$ with finite second moment, $v = v(w, \beta_0)$, and $\varepsilon = y - E[y|v]$,

$$0 = E[a(w)\varepsilon] = E[\{a(w) - E[a(w)|v]\}\varepsilon]. \tag{5}$$

These moment restrictions do not have the simple parametric form of equation (1), due to the presence of conditional expectations. However, it is possible to use nonparametric estimators for the conditional expectations without affecting the asymptotic variance, so that asymptotic variance comparisons can be made as if the conditional expectations were known.

For a sequence of functions $(a_1(w), a_2(w), ...)$ let

$$\rho_j(z,\beta) = \{a_j(w) - E[a_j(w)|v(w,\beta)]\}\{y - E[y|v(w,\beta)]\}. \tag{6}$$

For these functions the moment conditions of equation (1), with $x = 1$, are equivalent to equation (5). Let $\hat{\rho}_j(z,\beta) = \{a_j(w) - \hat{E}[a_j(w)|v(w,\beta)]\}\{y - \hat{E}[y|v(w,\beta)]\}$, where $\hat{E}[\cdot|v(x,\beta)]$ denotes some nonparametric regression estimator with regressors $v(x,\beta)$, and let $\hat{\rho}(z,\beta) = (\hat{\rho}_1(z,\beta), ..., \hat{\rho}_J(z,\beta))'$. Then it is well known that $\sum_{i=1}^n \hat{\rho}(z_i,\beta_0)/\sqrt{n}$ and $\sum_{i=1}^n \rho(z_i,\beta_0)/\sqrt{n}$ have the same limiting distribution and that for $v_\beta = \partial v(w,\beta_0)/\partial\beta$ and $a(w) = (a_1(w), ..., a_J(w))'$,

$$\begin{aligned} plim(n^{-1}\sum_{i=1}^n \partial\hat{\rho}(z_i,\beta_0)/\partial\beta) &= E[\partial\rho(z_i,\beta_0)/\partial\beta] = G \\ &= -E[\{a(w) - E[a(w)|v]\}\{\partial E[y|v]/\partial v\}v_\beta] \end{aligned}$$

Consequently, a GMM estimator based on $\hat{\rho}(z,\beta)$ will have the same asymptotic variance as one based on $\rho(z,\beta)$. Also, since $x = 1$ here, each $A(x)$ just corresponds to constant linear combination coefficients, with the optimal one given by $A^*(x) = A^* = G'\Omega^{-1}$ for $\Omega = E[\rho(z,\beta_0)\rho(z,\beta_0)']$. Then an optimal GMM estimator can be constructed in the usual way, by using a preliminary estimator $\bar{\beta}$ to form $\hat{\Omega} = \sum_{i=1}^n \hat{\rho}(z,\bar{\beta})\hat{\rho}(z,\bar{\beta})'/n$, forming $\hat{\rho}_n(\beta) = \sum_{i=1}^n \hat{\rho}(z_i,\beta)/n$, and solving

$$\hat{\beta} = argmin_\beta \hat{\rho}_n(\beta)'\hat{\Omega}^{-1}\hat{\rho}_n(\beta)$$

A one-step version is given by estimating the optimal linear combination by $\hat{A} = \hat{G}'\hat{\Omega}^{-1}$, where $\hat{G} = -n^{-1}\sum_{i=1}^n\{a(w_i) - \hat{E}[a|v(w_i,\bar{\beta})]\}\partial\hat{E}[y|v(w_i,\bar{\beta})]/\partial\beta$, and forming $\hat{\beta} = \bar{\beta} - (\hat{A}\hat{G})^{-1}\hat{A}\hat{\rho}_n(\bar{\beta})$. It will be shown that this estimator is approximately efficient as J grows, if $(a_1(w), a_2(w), ...)$ is a mean square spanning set, meaning that finite linear combinations of these functions can approximate as closely as desired any function with finite mean square.

3 The Spanning Condition

A certain spanning condition is critical for the GMM estimator $\tilde{\theta} = [I, 0]\hat{\theta}$ to approximately attain the semiparametric efficiency bound. The asymptotic variance of $\tilde{\theta}$ will be

$$\Sigma_J = [I,0]\{E[G(x)'\Omega(x)^{-}G(x)]\}^{-1}[I,0]'.$$

As is usual for GMM with an increasing set of moment conditions, Σ_J will be decreasing in J, in the positive semi-definite sense. Consequently, $\Sigma_\rho = lim_{J\to\infty}\Sigma_J$ will exist (see Appendix B). The GMM estimator will be approximately efficient if Σ_ρ is equal to the semiparametric (asymptotic) variance bound. The spanning condition will be sufficient for this equality.

Intuitively, efficiency should be closely related to whether the moment conditions characterize the semiparametric model, i.e. whether the restrictions imposed by all the moment conditions are the same as imposed by the model. Unless this condition holds, there will be information in the model that is not exploited by the GMM estimator. When this condition holds the GMM estimator based on many moments should pick up most of the information in the model, leaving only a small remainder when enough moments are used.

Because asymptotic efficiency is a local property, a local formulation of the efficiency condition, in terms of directions of departure from the truth, gives the easiest approach. The spanning condition will be that the set of directions allowed by the moment conditions is the same as allowed by the model. These direction sets are referred to as tangent spaces, so that the spanning condition is that the model tangent space is the same as the moment tangent space. Of course, if equality of tangent sets implies equality of distributions, then the local condition will coincide with the global one that the model and moments imply the same restrictions on the distribution. Tangent space equality turns out to be easy to check though, and allows us to sidestep global conditions, so the local formulation seems to be the most useful.

Before stating the spanning condition we should describe the tangent sets. The model tangent set is formulated in terms of scores, as in Bickel et. al. (1993). Partition the data observation as $z = (y, x)$, and suppose that the model specifies that the conditional density of y given x is a member of a semiparametric family

$$\{f(y|x,\beta,h) : \beta \in \mathcal{B}, h \in \mathcal{H}\}, \tag{7}$$

where $\mathcal{B}$ is an open subset of $\Re^q$ and, h denotes a function, $\mathcal{H}$ is a set of such functions. For example, in the transformation model given above the density of y given x has this form with $h = \tau$ and $f(y|x,\beta,h) = [d\tau(y)/dy]g(\tau(y) - x'\delta, \lambda)$ for the density $g(\varepsilon, \lambda) = dG(\varepsilon, \lambda)/d\varepsilon$. It will be assumed throughout that the marginal distribution of x is unrestricted, as appropriate for evaluating efficiency with conditional moment restrictions. Define a regular parametric submodel to be the family of densities $\{f(y|x, \beta_0, h(\eta))\}$, where η is a scalar parameter, with $h(\eta)$ equal to the truth at some η_0, where "regular" means that the square root of the density is mean-square differentiable with respect η, has a nonzero Fisher information, and possibly satisfies other regularity conditions (such as boundedness of conditional second moments of $\rho(z,\theta)$). Let $S_\eta = \partial \ln f(y|x, \beta_0, h(\eta))/\partial\eta|_{\eta=\eta_0}$ denote the score for the parametric submodel, where a z argument is suppressed for notational convenience and the scores are defined more precisely in terms of derivatives of the square root of the density (e.g. see Bickel, et. al. 1993). The model tangent set T is the closed linear span of the set of such scores. It represents directions of departure from the truth that are allowed by the model.

To describe the moment tangent set, consider a parametric family $\{f(y|x,\eta)\}$ of conditional densities satisfying the moment restrictions of equation (1), meaning that there exists $\gamma(\eta)$ such that

$$\int \rho_j(z, \beta_0, \gamma(\eta)) f(y|x,\eta) dy = 0, j = 1, 2, ...,$$

identically in η. Differentiating this identity with respect to η, for $j = 1, ..., J$, gives

$$E[\rho t|x] = H(x)c, c = -\partial\gamma_J(\eta_0)/\partial\eta, t = \partial lnf(y|x,\eta_0)/\partial\eta. \tag{8}$$

This suggests a tangent set for the first J moments of the form

$$T_J = \{t : E[t^2] < \infty, E[t|x] = 0, E[\rho t|x] = H(x)c \text{ for a constant vector } c\}, \tag{9}$$

where $E[t|x] = 0$ holds because of the usual zero mean property of conditional scores. Then, because T_J will be a decreasing sequence of sets (increasing J corresponds to adding moment conditions) the tangent set for all the moments will be given by

$$T_\rho = \cap_{J=1}^{\infty} T_J.$$

Here T_ρ represents the set of all directions of departure from the truth that are allowed by the moment conditions.

Assuming the moment conditions are implied by the semiparametric model, it will be the case that a score for a parametric submodel satisfies equation (8) for all J. Consequently, $T \subseteq T_\rho$. Therefore, the model and moment tangent spaces will be equal if $T_\rho \subseteq T$, meaning that any direction of departure allowed by the moment conditions is also allowed by the model. This leads to the following condition:

Spanning Condition : $T_\rho = T$.

Intuitively, using GMM will lead to approximate efficiency when imposing all the moment conditions restricts the density so as to only allow directions of departure that are given by the semiparametric model.

We use two regularity conditions to obtain a precise result. We define regularity of a parametric family of densities as in the discussion of the model tangent space above.

Assumption 1: With probability one, $f(y|x, \beta, h_0)$ is regular in β, $\partial E[\rho_j(z, \beta, \gamma_0)|x]/\partial\beta|_{\beta=\beta_0}$ exists, $\int max_{\beta\in B}\rho_j(z, \beta, \gamma_0)^2 f(y|x, \beta, h_0)dy$ is bounded, and $\rho_j(z, \beta, \gamma_0)$ is continuous at each β with probability one.

For some of the examples it will be important that this condition allows the residual to be discontinuous in β, as long as at each β this occurs with probability zero. The next condition allows for some of the residuals to be zero with positive probability, which is also important in the examples.

Assumption 2: For each J there is $R(x)$ such that $H(x) = \Omega(x)R(x)$, there is a symmetric generalized inverse $\Omega(x)^-$ such that $E[G(x)'\Omega(x)^- G(x)]$ exists and is nonsingular, *and β has a finite and nonsingular semiparametric variance bound.*

The condition $H(x) = \Omega(x)R(x)$ is easy to check in the examples we consider and should be satisfied quite generally. If there is a parametric submodel $f(y|x,\eta)$, as discussed above, with $\partial\gamma(\eta_0)/\partial\eta$ nonsingular then by equation (8), $H(x) = E[\rho t'\{-\partial\gamma(\eta_0)/\partial\eta\}^{-1}|x]$, so this condition holds by Lemma 6 of the Appendix.

Theorem 1: If Assumptions 1 and 2 and the spanning condition are satisfied then $lim_{J\to\infty}\Sigma_J = \Sigma_\rho$ is the semiparametric bound.

In the Appendix a projection formula for the GMM limit Σ_ρ is derived. This formula extends Chamberlain's (1987) bound to the case where there are a countably infinite number of moment restrictions. It is compared with a corresponding formula for the semiparametric bound to obtain the proof of Theorem 1. For ease of exposition we reserve discussion of these formulae and the proofs to the Appendix.

Consider the transformation model as an example. For a parametric submodel $\tau(y,\eta)$, the score is

$$S_\eta = \partial ln[\tau_y(y,\eta)g(\tau(y,\eta)-v)]/\partial\eta = \tau_{y\eta}(y)/\tau_y(y) + \tau_\eta(y)s(\varepsilon),$$

where subscripts denote partial derivatives, $g(\varepsilon) = g(\varepsilon,\lambda_0)$, $s(\varepsilon) = g_\varepsilon(\varepsilon)/g(\varepsilon)$, and $v = x'\delta_0$. Therefore, the tangent set T will be the closed, linear span of the set of objects of this form. To compare this set with T_ρ, note that $\rho_j(z,\beta,\gamma)$ depends only on γ_j, so that $H_{jk}(x) = 0$, $(j \neq k)$, and $H_{jj}(x) = \partial E[\rho_j(z,\beta_0,\gamma_{j0})|x]/\partial\gamma_j = -g(\tau(\bar{y}_j)-v)$. Then T_ρ will consist of those $t(y,x)$ such that $E[t|x] = 0$ and

$$E[\rho_j t|x] = \int_{-\infty}^{\bar{y}_j} t(y,x)\tau_y(y)g(\tau(y)-v)dy = -g(\tau(\bar{y}_j)-v)c(\bar{y}_j).$$

If $(\bar{y}_j)_{j=1}^{\infty}$ is dense and $g(\varepsilon)$ is differentiable and positive everywhere, then there is a $c(y)$ such that this equation holds with $\bar{y}_j$ replaced by any $y \in \Re$. Differentiating with respect to y and solving for t then gives

$$t(y,x) = -c_y(y)/\tau_y(y) - c(y)s(\varepsilon).$$

This expression has exactly the same form as the score for η for a parametric submodel, with $-c(y)$ replacing $\tau_\eta(y)$. Thus, the moment tangents satisfy the same conditions as the score for parametric submodels, and hence the spanning condition will be satisfied. Consequently, the asymptotic variance of the ordered choice MLE of the regression and distribution parameters, as well transformation values, will converge to the semiparametric bound as the intervals become finer.

Consider next the index model example. In this case a parametric submodel $f(z|\eta)$ must be such that $\int yf(y|w,\eta)dy$ is a function of only v. Therefore, its derivative will also be a function of only v, giving

$$\begin{aligned} \partial \int y \cdot f(y|w,\eta_0)dy/\partial\eta &= E[y \cdot \partial lnf(y|w,\eta_0)/\partial\eta|w] = E[\varepsilon \cdot \partial lnf(y|w,\eta_0)/\partial\eta|w] \\ &= E[\varepsilon \cdot S_\eta|w] = E[\varepsilon \cdot S_\eta|v], \end{aligned}$$

where the second equality follows by the usual mean zero property of scores $E[\partial lnf(y|w,\eta_0)/\partial\eta|w] = 0$ and the third by $S_\eta = \partial lnf(y|w,\eta_0)/\partial\eta + \partial lnf(w|\eta_0)/\partial\eta$ and $E[\varepsilon|w] = 0$. Since the score is otherwise unrestricted, it follows that $T = \{t : E[\varepsilon t|w] = E[\varepsilon t|v]\}$. Now suppose that finite linear combinations of $(a_1(w), a_2(w), ...)$ can approximate any function with finite mean-square arbitrarily well. For instance, the set of all integer power series in a bounded, one-to-one transformation of w have this property. It is well known that this property is equivalent to any function $\delta(w)$ with $E[\delta(w)^2]$ finite and $E[a_j(w)\delta(w)] = 0$ for all j being zero. Then, since $x = 1$ in this example, the moment tangent set is given by the set of t with finite mean square, such that for each j,

$$\begin{aligned} 0 &= E[\rho_j t] = E[\{a_j(w) - E[a_j(w)|v]\}E[\varepsilon t|w]] \qquad (10) \\ &= E[a_j(w)\{E[\varepsilon t|w] - E[\varepsilon t|v]\}]. \end{aligned}$$

Suppose that $Var(\varepsilon|w)$ is bounded, so that $E[\varepsilon t|w]$ has finite mean square. Then the mean-square spanning property and this equation imply that

$$E[\varepsilon t|w] = E[\varepsilon t|v].$$

Thus, the moment tangents satisfy the same conditions as the scores for parametric submodels, and hence the spanning condition will be satisfied.

The previous GMM efficiency result of Chamberlain (1987) is based on approximating by a linear combination of moment conditions. It turns out that the spanning condition is the dual of this previous approach. We have given first priority to the spanning condition because it is easiest to check in the most difficult cases.

To compare with the previous approach it needs to be generalized to allow for the nuisance parameters γ and for conditioning on x, and the efficiency of estimators of β should be considered. To do so, let V_J be the block of Σ_J corresponding to β and let V be the semiparametric variance bound for estimators of β. Also, for a set A consisting of random variables with finite mean square and conditional mean zero given x, let $A^{\perp} = \{s|E[s^2] < \infty, E[s|x] = 0, E[sa] = 0 \forall a \in A\}$ denote its orthogonal complement. As is well known, under appropriate regularity conditions there is a representation $V = (E[SS'])^{-1}$, with each component of S being the element of $T^{\perp}$ that is closest in mean-square to the corresponding component of the score for β. The random vector S is often referred to as the efficient score. As shown in Newey (1993) for the unconditional case without nuisance parameters, the optimal function $A^*(x)$ can be interpreted as the coefficients of a regression of the efficient score on the moment functions, so that the efficiency bound is approximately attained when linear combinations of the moments approximate the efficient score. The following result generalizes this previous one to conditional moment restrictions with nuisance parameters.

Theorem 2: If Assumptions 1 and 2 are satisfied then

$$V^{-1} - V_J^{-1} = min_{\pi(x):E[\pi(x)H(x)]=0} E[\{S - \pi(x)\rho\}\{S - \pi(x)\rho\}'],$$

and $V_J \to V$ *if and only if for each* J *there is* $\pi_J(x)$ *with* $E[\pi_J(x)H(x)] = 0$ *such that* $E[||S - \pi_J(x)\rho||^2] \to 0$ *as* $J \to \infty$.

This result shows that the difference of the inverses of the semiparametric bound and GMM variance is the variance of the residuals from approximating the efficient score by $\pi(x)\rho$, where $E[\pi(x)H(x)] = 0$. The presence of x in $\pi(x)$ accounts for the conditioning on x and the constraint on $\pi(x)$ of $E[\pi(x)H(x)] = 0$ accounts for the presence of γ. This result specializes to that of Newey (1993) when $x = 1$ and γ is not present. In general $V_J \to V$ when for each J there is $\pi_J(x)$ such that $\pi_J(x)\rho$ can approximate the efficient score arbitrarily well for large enough J.

One of the positive aspects of this result is that it is constructive, with efficiency following from finding $\pi_J(x)$ where $\pi_J(x)\rho$ approximates S (and $E[\pi_J(x)H(x)] = 0$). The problem is that constructing such $\pi_J(x)$ can be very hard, particularly when γ present. The root of this problem is that the structure of $T^\perp$ is often complicated, leading to a complicated form for the efficient score S. This problem leads to falling back on a more abstract sufficient condition, that any element of $T^\perp$ can be approximated by the moment conditions. Specifically, let M denote the mean-square closure of the set

$$\{\pi_J(x)\rho : E[\{\pi_J(x)\rho\}^2] < \infty,\ E[\pi_J(x)H(x)] = 0,\ J \in \{1, 2, ...\}\}.$$

That is, M is the set of random variables that can be approximated arbitrarily closely in mean-square by $\pi_J(x)\rho$, with $E[\pi_J(x)H(x)] = 0$. Then $T^\perp = M$ will be sufficient for $V_J \to V$, since the components of S are in $T^\perp$. The following result shows that the spanning condition is equivalent to this sufficient condition.

Theorem 3: If Assumptions 1 and 2 are satisfied then $T_\rho = M^\perp$ and the spanning condition is satisfied if and only if $T^\perp = M$.

Thus we see that the spanning condition is equivalent to $T^\perp = M$, and so to the previous approach of Chamberlain (1987). Furthermore, this result also shows $T_\rho = M^\perp$ while $T = T^{\perp\perp}$ is a well known result. Thus $T = T_\rho$ is the dual of $M = T^\perp$, i.e. the spanning condition is the dual of the Chamberlain (1987) type of condition for efficiency. The spanning condition has received first priority because the most difficult cases seem to correspond to $T^\perp$ having a complicated structure, but T being relatively simple. In these cases it is much easier to work with the tangent sets than their orthogonal complements. This relative simplicity is illustrated by the transformation model example, where it was straightforward to show equality of moment and model tangent sets, but Bickel et. al. (1993) shows that the orthogonal complement of the model tangent set is complicated.

Other examples are provided by censored and truncated regression with an independent disturbance, as considered in the next Section.

4 Censored and Truncated Regression with Independent Disturbance

Two important semiparametric limited dependent variable models are censored and truncated regression models with a disturbance that is independent of the regressors. There is a large literature on estimation of these models, see Powell (1994). In both of these models the efficiency bounds are complicated, but there are simple moment conditions, so that GMM may be useful for efficient estimation. In this Section we give GMM estimators that can approximately attain the semiparametric bound for each of these models.

These models can be formulated as missing data models for the latent regression

$$y^* = x'\beta_0 + \varepsilon, \varepsilon \text{ and } x \text{ are independent, } \varepsilon \text{ has p.d.f. } g(\varepsilon). \tag{11}$$

The censored regression model is one where x is always observed, but only $y = max\{0, y^*\}$ is observed. The truncated regression model is one where (y, x) is only observed if $y^* > 0$.

To construct moment conditions in each model we consider functions $m_j(\varepsilon), (j = 1, ..., n)$, and suppose that there is γ_{j0} such that $E^*[m_j(\varepsilon - \gamma_{j0})] = 0$, where $E^*[\cdot]$ represents the expectation for the latent data. For censored regression we require that $m_j(\varepsilon)$ is constant below some value, and let $\tau_j = \sup\{\bar{\varepsilon} : m_j(\varepsilon) = m_j(\bar{\varepsilon}), \varepsilon \leq \bar{\varepsilon}\}$ For truncated regression we require that $m_j(\varepsilon)$ is zero below some value, and let $\tau_j = \sup\{\bar{\varepsilon} : m_j(\varepsilon) = 0, \varepsilon \leq \bar{\varepsilon}\}$. Then for $\theta = (\beta', \gamma_1, ..., \gamma_J)'$ and

$$\rho_j(z, \theta) = 1(\gamma_j + x'\beta > -\tau_j)m_j(y - \gamma_j - x'\beta), (j = 1, ..., J), \tag{12}$$

the conditional moment restriction of equation (1) is satisfied, as shown by Newey(2001), where references and examples are given.

The optimal matrix $A^*(x)$ has the same form for both censored and truncated regression. Let Λ be the $J \times J$ matrix with $\Lambda_{jk} = E^*[m_j(\varepsilon - \gamma_{j0})m_k(\varepsilon - \gamma_{k0})], (j, k =$

$1,...,J)$. Also, let d be the $J \times 1$ vector with $d_j = \partial E^*[m_j(\varepsilon - \gamma_{j0} + \alpha)]/\partial\alpha|_{\alpha=0}$, and $D = diag(d_1, ..., d_J)$ be the diagonal matrix with j^{th} diagonal element d_j. Also, let $I(x, \theta)$ be the selection matrix that selects those $\rho_j(z, \theta)$ with $\gamma_j + x'\beta > -\tau_j$, and $I(x) = I(x, \theta_0)$. Then, as shown in Newey (2001), $G(x) = I(x)'I(x)[dx', D]$ and $\Omega(x)^- = I(x)'[I(x)\Lambda I(x)']^{-1}I(x)$, so that

$$A^*(x) = [dx', D]'I(x)'[I(x)\Lambda I(x)']^{-1}I(x). \tag{13}$$

4.1 Censored Regression

For censored regression we consider quantile estimation, where $m_j(\varepsilon) = 1(\varepsilon < 0) - \alpha_j$, $0 < \alpha_j < 1$, as in Powell (1986). Here $\tau_j = 0$ and γ_{j0} is the α_j^{th} quantile of the distribution of ε. Also, it is straightforward to estimate the unknown components $I(x), d$, and Λ of $A^*(x)$. Here $\Lambda_{jk} = \min\{\alpha_j, \alpha_k\} - \alpha_j\alpha_k$ is known and $d_j = g(\gamma_{j0})$. Let $\bar{\beta}$ and $\bar{\gamma}_j$ be preliminary estimators of the parameters and $\bar{v}_i = x_i'\bar{\beta}$. For example, $\bar{\beta}$ could be obtained from some censored regression quantile estimator and each $\bar{\gamma}_j$ from minimizing the censored regression quantile objective function $\sum_{i=1}^n q_j(y_i - \max\{0, \bar{v}_i + \gamma_j\})$, where $q_j(u) = [\alpha_j - 1(u > 0)]u$. For $\bar{\varepsilon}_i = y_i - \bar{v}_i$, let $K(u)$ denote a kernel function, satisfying $\int K(u)du = 1$ and other regularity conditions, h_j a bandwidth parameter, $K_{ji} = K((\bar{\varepsilon}_i - \bar{\gamma}_j)/h_j)1(y_i > 0)$, and $\bar{K}_{ji} = \int_{-(\bar{v}_i+\bar{\gamma}_j)/h_j}^{\infty} K(u)du$. The kernel density estimator of d_j from Hall and Horowitz (1990) is $\hat{d}_j = \sum_{i=1}^n K_{ji}/(h_j \sum_{i=1}^n \bar{K}_{ji})$. Let $\hat{d} = (\hat{d}_1, ..., \hat{d}_J)$ and $\hat{D} = diag(\hat{d}_1, ..., \hat{d}_J)$. Then $A^*(x)$ can be estimated by

$$\hat{A}(x) = [\hat{d}x', \hat{D}]'I(x, \hat{\theta})'[I(x, \hat{\theta})\Lambda I(x, \hat{\theta})']^{-1}I(x, \hat{\theta})$$

A one step estimator can be formed as in equation (3).

By comparing the model and moment tangent sets we can see why the asymptotic variance will approach the bound as the quantiles become dense on the real line. By independence of ε and x, a parametric submodel for the conditional density of y given x will have the form $f(y|x, \eta) = 1(y > 0)g(\varepsilon, \eta) + 1(y = 0)\int_{-\infty}^{-v} g(u, \eta)du$, where $g(\varepsilon, \eta)$ is a parametric submodel for the density of ε. Then for $s(\varepsilon) = \partial \ln g(\varepsilon, \eta)/\partial\eta|_{\eta=\eta_0}$, the score will be

$$\begin{aligned} S_\eta &= 1(y > 0)s(\varepsilon) + 1(y = 0)E[s(\varepsilon)|y = 0, x] \\ &= 1(y > 0)s(\varepsilon) - 1(y = 0)\Pr(y > 0|x)\Pr(y = 0|x)^{-1}E[s(\varepsilon)|y > 0, x], \end{aligned}$$

where the second equality follows by $E[s(\varepsilon)|x] = 0$. Thus, the model tangent set consists of functions that depend only on ε for $y > 0$, and that are determined by their values for $y > 0$. Thus, to show that the spanning condition holds, it suffices to show that the moment tangents depend only on ε when y is positive. Intuitively, conditional on $v > -\gamma_{j0}$, quantile independence of ε from x holds at all quantiles with $\gamma_{j'0} \geq \gamma_{j0}$, implying independence of ε and x (and hence $t(\varepsilon, x)$ depending only on ε) on the set where $\varepsilon > \gamma_{j0}$ and $v > -\gamma_{j0}$. The spanning condition then holds because the set where $y = \varepsilon + v > 0$ is the union of the sets where $\varepsilon > \gamma_{j0}$ and $v > -\gamma_{j0}$ over the countable, dense set of quantiles.

The following result gives precise conditions for efficiency.

Theorem 4: If $g(\varepsilon)$ is positive, v is continuously distributed, and $(\alpha_j)_{j=1}^{\infty}$ is dense in $(0, 1)$, then the asymptotic variance of the GMM estimator for censored regression quantiles converges to the semiparametric bound as $J \to \infty$.

Because the spanning condition is satisfied, as J grows the asymptotic variance of the slope estimator will approach the bound derived by Cosslett (1987) and Ritov (1990), and the quantile estimators will also approach efficiency. Thus, combining moment restrictions from censored regression quantiles leads to efficiency of regression slope and quantile estimators. This approach provides a simple alternative to the efficient estimator of Ritov (1990). It should also be noted that using quantiles amounts to a step function approximation of the efficient estimator, which approximation might be improved by using $\rho_j(z, \beta, \gamma)$ that are smooth in ε.

4.2 Truncated Regression

For truncated regression we consider $m_j(\varepsilon) = \alpha_j 1(\varepsilon > 0) - 1(\varepsilon > \tau_j)$, $0 < \alpha_j < 1$, $\tau_j > 0$, similarly to Newey (1987). Here, for $\Pr^*$ denoting the latent probability distribution of ε, γ_{j0} is the solution to $\Pr^*(\varepsilon > \gamma + \tau_j)/\Pr^*(\varepsilon > \gamma) = \alpha_j$, which will exist when the density of $g(\varepsilon)$ is strictly log-concave and a boundary condition holds, as specified below. Estimating $A^*(x)$ is more difficult for truncated regression because it is not possible to form direct estimators of the constants d and Λ. One can use a GMM estimator as in Newey (2001). Order j so that $\gamma_{j+1,0} < \gamma_{j,0}$, and assume that there are no ties. Let $\rho^j(z,\theta)$ denote the vector of the first j elements of $\rho(z,\theta)$, $X = (1, x')'$, and

$$\begin{aligned} g^j(z,\theta) &= 1(-\gamma_{j+1} > x'\beta \geq -\gamma_j)\rho^j(z,\theta) \otimes X, (j = 1, ..., J-1), \qquad (14) \\ g^J(z,\theta) &= 1(x'\beta \geq -\gamma_J)\rho(z,\theta) \otimes X, g(z,\theta) = (g^1(z,\theta)', ..., g^J(z,\theta)')'. \end{aligned}$$

Evidently, $g(z,\theta_0) = A(x)\rho(z,\theta_0)$ for some matrix $A(x)$ and $\rho(z,\theta)$ from equation (12). Also, as shown in Newey (2001), $Bg(z,\theta_0) = A^*(x)\rho(z,\theta_0)$ for a matrix B. It follows that the one-step optimal GMM estimator using the moment functions from equation (14) is as efficient as the estimator with the best instruments. This estimator can be formed using some $\hat{G}$ and $\hat{\Omega}$ as

$$\hat{\theta} = \bar{\theta} - (\hat{A}\hat{G})^{-1}\hat{A}\sum_{i=1}^{n}\hat{\rho}(z_i,\bar{\theta})/n, \hat{A} = \hat{G}'\hat{\Omega}^{-1}. \qquad (15)$$

We refer the interested reader to Newey (2001) for a fuller description of this estimator, including construction of $\hat{G}$ and $\hat{\Omega}$.

To check equality of the moment and model tangent sets, as needed for the spanning condition, we first derive the model tangent set. By independence of ε and x, a parametric submodel for the conditional density of y given x will have the form $f(y|x,\eta) = g(\varepsilon,\eta)/\int_{-v}^{\infty} g(u,\eta)du$, where $g(\varepsilon,\eta)$ is a parametric submodel for the density of ε. Then for $s(\varepsilon) = \partial \ln g(\varepsilon,\eta)/\partial\eta|_{\eta=\eta_0}$, the score will be

$$S_\eta = s(\varepsilon) - E^*[s(\varepsilon)|y^* > 0, x]. \qquad (16)$$

Here the score is an additively separable function of ε and x that has conditional mean zero given x. Thus, to show that the spanning condition holds, it suffices to show that the moment tangents must be additively separable functions of ε and x. Intuitively, conditional on $v > -\gamma_{j0}$, the moment restrictions hold at all j' with with $\gamma_{j'0} \geq \gamma_{j0}$, implying that $\Pr(\varepsilon > \gamma + \tau|x)/\Pr(\varepsilon > \gamma|x)$ does not depend on x for all $\gamma \geq \gamma_{j0}$ and $\tau > 0$, i.e. $\Pr(\varepsilon > \gamma|x) = c(\gamma)\Pr(\varepsilon > \gamma_{j0}|x)$. This form implies scores that are additive in ε and x for $v > -\gamma_{j0}$ and $\varepsilon > \gamma_{j0}$. The spanning condition then holds because the set where $y > 0$ is the union of these sets, similarly to the censored case. The following result makes this intuition precise.

Theorem 5: If $g(\varepsilon)$ is positive, differentiable, and strictly log concave, $\lim_{\gamma\to\infty}[1 - G(\gamma + \tau)]/[1 - G(\gamma)] = 0$ *for every $\tau > 0$, v is continuously distributed, and $(\alpha_j, \tau_j)_{j=1}^{\infty}$ is dense in* $(0,1) \times (0,\infty)$, *then the asymptotic variance of the truncated moment GMM estimator converges to the semiparametric bound as $J \to \infty$.*

Similarly to quantiles, it should be noted that these moment functions correspond to a step function approximation of the efficient estimator, which might be improved by using $\rho_j(z, \beta, \gamma)$ that are smooth in ε.

5 Asymptotic Efficiency

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

In the formulation of regularity conditions there is always a trade-off between generality and ease of verification. Here we give one general result and one example showing how the regularity conditions can be checked. Although substantial work is involved in applying the general result, we believe it to be useful, because it allows one to side-step algebraic and probabilistic arguments that will be required for showing asymptotic efficiency in many cases.

The general result covers cases with estimated unconditional moment restrictions, no nuisance parameter estimates, and smooth moment functions. Specifically, it applies to a one-step estimator as given in equation (3), with $x = 1$, and $\theta = \beta$. Many of the estimators can be thought of as having this form, with the nuisance parameter estimates subsumed in $\hat{\rho}$. Much work may be required to verify these conditions when nuisance parameters are present.

Let $S_i = S(z_i)$ be the efficient score for the i^{th} observation $\hat{\rho}_i = \hat{\rho}(z_i, \beta_0), \rho_i = \rho(z_i, \beta_0)$, and $A^* = G'\Omega^{-1}$.

Theorem 6: Suppose that $J \to \infty$ and $n \to \infty$ and i) for each J, $G = -E[\rho_i S_i']$ and there exists π_J such that $E[\|S_i - \pi_J \rho_i\|^2] \to 0$; for each J there is a nonsingular constant matrix B such that when $\hat{\rho}(z, \beta)$ is replaced by the nonsingular linear transformation $B\hat{\rho}(z, \beta)$, the following conditions are satisfied: ii) the smallest eigenvalue of Ω is bounded away from zero; iii) $\|\hat{\Omega} - \Omega\| \xrightarrow{p} 0$; iv) $\sqrt{J}\|A^*(\hat{\Omega} - \Omega)\| \xrightarrow{p} 0$; v) for any $\dot{\beta}$ on the line joining $\bar{\beta}$ and β_0, $\sqrt{J}\|\sum_{i=1}^n \hat{\rho}_\beta(z_i, \dot{\beta})/n - G\| \xrightarrow{p} 0$; and vi) $\|\sum_{i=1}^n (\hat{\rho}_i - \rho_i)/\sqrt{n}\| \xrightarrow{p} 0$. Then for $V^* = (E[S_i S_i'])^{-1}$,

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V^*), (\hat{G}'\hat{\Omega}^{-1}\hat{G})^{-1} \xrightarrow{p} V^*.$$

The first part of condition i) is a semiparametric version of the generalized information matrix equality, that is often straightforward to show. The second part of condition i) will follow from the spanning condition via Theorem 3 (because the components of S_i are in $T^\perp$). The other convergence conditions lead to asymptotic efficiency.

An example is the Chamberlain (1987) estimator that uses many unconditional moment restrictions to estimate a model with a fixed number of conditional moment restrictions. Specifically, suppose there is a vector $u(z, \beta)$ satisfying $E[u(z, \beta_0)|w] = 0$ and let $p^J(w)$ be a vector of approximating functions (e.g. powers or splines). Consider the one-step estimator in equation (3) with $\hat{\rho}(z, \beta) = \rho(z, \beta) = u(z, \beta) \otimes p^J(w)$. Let $D(w) = E[u_\beta(z, \beta_0)|w]$, $u = u(z, \beta_0)$, $\Sigma(w) = E[uu'|w]$, $\mathcal{N}$ be some neighborhood of β_0, $\delta_1(z) = \sup_{\beta \in \mathcal{N}} \|u_\beta(z, \beta)\|$, and $\delta_2(z) = \max_{j \le q} \sup_{\beta \in \mathcal{N}} \|\partial u_\beta(z, \beta)/\partial \beta_j\|$.

Theorem 7: Suppose that i) $E[\|u\|^4|w], E[\delta_1(z)^2|w]$, and $E[\delta_2(z)|w]$ are bounded; ii) for any scalar function $b(w)$ with $E[b(w)^2] < \infty$ there exists π_J such that $E[\{b(w) -$

$\pi_J' p^J(w)\}^2] \to 0$ as $J \to \infty$; iii) $\Sigma(w)$ has smallest eigenvalue that is bounded away from zero; iv) for each J there is a nonsingular constant matrix B such that $\tilde{p}^J(w) = Bp^J(w)$ satisfies $\sup_w \|\tilde{p}^J(w)\| \le \zeta(J)$ and $E[\tilde{p}^J(w)\tilde{p}^J(w)']$ has smallest eigenvalue that is bounded away from zero; and v) $J\zeta(J)^2/n \to 0$. Then for $V^* = (E[D(w)'\Sigma(w)^{-1}D(w)])^{-1}$,

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V^*), (\hat{G}'\hat{\Omega}^{-1}\hat{G})^{-1} \xrightarrow{p} V^*.$$

Condition v) restricts the rate of growth of J in a way that depends on condition iv), which is a normalization like that adopted by Newey (1997). The allowed rate of growth for J depends on $p^J(w)$ and the distribution of w. Under the conditions of Newey (1993) and Hahn (1997), where w is bounded, the components of $p^J(w)$ are powers of w, and the density of w is bounded away from zero on some interval, $\zeta(J)$ can be taken equal to CJ^{CJ} for some constant C. In that case the allowed growth rate for J is $J\ln(J)/\ln(n) \to 0$. If stronger restrictions are placed on the distribution of w then J can be allowed to grow at a faster rate. For power series and the density of w bounded away from zero on a rectangular support, $\zeta(J) = CJ$ (Andrews, 1991), so that the rate condition is $J^3/n \to 0$. Under the same condition on w, for regression splines (Newey, 1997) or Fourier series over the whole support (Andrews, 1991), $\zeta(J) = CJ^{1/2}$, so that $J^2/n \to 0$ is the rate condition. For efficiency reasons regression splines or power series would be preferred to Fourier series, because the periodicity of Fourier series would lead to poor approximation of the optimal instruments, that need not be periodic.[1]

Koenker and Machado (1999) obtained a nice result showing that asymptotic efficiency is possible with $J^3/n \to 0$ for Fourier series. This result shows that the less restrictive condition $J^2/n \to 0$ gives efficiency, for Fourier series and regression splines. The source of this improvement is condition iv) of Theorem 6. Previous work has used $\sqrt{J}\|\hat{\Omega} - \Omega\| \xrightarrow{p} 0$, which is sufficient for iv) because $\|A^*\|$ is bounded under the other conditions. Condition iv) exploits the fact that all that is really needed for asymptotic efficiency is that a linear combination of all the moment conditions be well behaved in

[1] This efficiency problem can be remedied by constraining the support to lie strictly inside the full domain of a Fourier series, but then $\zeta(J)$ will no longer be $CJ^{1/2}$ for some constant C.

large samples.

For other estimators it may be possible to weaken further the condition $J^2/n \longrightarrow 0$. For example, Donald and Newey (2001) show that in the linear simultaneous equations model the limited information maximum likelihood estimator is efficient when $J/n \longrightarrow 0$. In general, though, the estimator considered here has a bias term in the expansion of $\sqrt{n}(\hat{\beta} - \beta_0)$ that is of order $J/\sqrt{n}$, so that $J^2/n \longrightarrow 0$ will be required for the asymptotic efficiency.

For applications it is important to know how to choose J as a function of the data. Donald and Newey (2001) give some results for conditional moment restrictions with homoskedasticity. Extending those results to the general case considered here is beyond the scope of this paper, but has been done in Donald, Imbens, and Newey (2002).

6 A Sample Selection Empirical Example

Sample selection bias is important in many econometric applications, but correcting for it can lead to imprecise estimators. In some applications, using moment conditions beyond the usual least squares ones may lead to large efficiency gains. As an example, we consider Mroz's (1987) data on female labor supply. The data consists of measurements on the characteristics of 753 married women, drawn from the 1975 University of Michigan Panel Study of Income Dynamics. Of the 753 women in the analysis sample, 428 were working at the time of the study. The equation of interest has a left-hand side variable that is the annual hours of work. The right-hand side variables include the logarithm of the wage rate, family income less wife's labor income, indicators for young and older children in the family, and the wife's number of years of age and education. The (binary choice) selection equation has as regressors all of the exogenous right-hand side variables from the hours equation, labor force experience, and other background variables, and various interaction terms. The model will be identified by the exclusion from the labor supply equation of several variables that are included in the selection equation. These variables are labor force experience, mother's education, father's education, and a regional unemployment variable. Exclusion of each of these variables from labor supply seems reasonable.

Figure 1 graphs the density of the residual from a Heckman (1979) two-step least squares estimator like that given in Table X of Mroz (1987). Evidently, it is multimodal with sharp peaks, as might be expected if individuals cluster around specific amounts of full and part time work. Gaussian disturbances generally only lead to some asymmetry in the selected data, so that this shape suggests the disturbances are not Gaussian. Of course this means that parametric methods based on Gaussian distributions may be inconsistent. It also means that, by analogy with linear regression, we might expect nonlinear moment conditions to add greatly to the efficiency of least squares in this setting. For example, Newey (1988) shows that in regression large efficiency gains are possible with bimodal error distributions by using multiple, nonlinear moment conditions.

6.1 The Model and Estimator

To describe the model we consider let y^* be the left-hand side variable, w be a vector of right-hand side variables, and β_0 a vector of regression parameters. Some components of w are allowed to be endogenous. Also, let x be a vector of exogenous variables that includes the exogenous components of w. The model is then

$$\begin{aligned} y^* &= w'\beta_0 + \varepsilon, \ y^* \text{ only observed if } d = 1, \\ \Pr(d &= 1|x) = p; \ \varepsilon \text{ and } x \text{ independent given } p \text{ and } d = 1 \end{aligned}$$

The selection probability p is often referred to as the propensity score. This model is implied by a latent variable model for d, where $d = 1(\tau(x) + v \geq 0)$ for an unknown function $\tau(x)$ and a disturbance v, p is a one-to-one function of $\tau(x)$, and the conditional distribution of (ε, v) given x depends only on $\tau(x)$. A conditional mean version of this model was considered by Ahn and Powell (1993). The semiparametric efficiency bound for this model with exogenous w was derived by Newey and Powell (1993).

This model implies that any function of ε should be uncorrelated with any function of x, conditional on p. We consider a particular class of moment restrictions that exploit the conditional independence of ε and x. Let $m(\varepsilon, p)$ be a vector of functions, D denote the event $d = 1$, and subscripts denote partial derivatives. Let $\hat{p}$ denote some nonparametric estimator of the propensity score. Also, let $\hat{E}[\bullet|\hat{p} = p, D]$ denote the predicted value from a nonparametric regression on $\hat{p}$ in the selected sample, evaluated at p, $\hat{w}$ a linear combination of x that is used to instrument for w, and

$$\hat{\lambda}(\beta, p) = \hat{E}[m(y - w'\beta, \hat{p})|\hat{p} = p, D], \hat{\mu}(p) = \hat{E}[\hat{w}|\hat{p} = p, D].$$

Then moment conditions can be formed as follows. For $\hat{\lambda}_p(\beta, p) = \partial\hat{\lambda}(\beta, p)/\partial p$,

$$\begin{aligned} \hat{\eta}_i(\beta) &= d_i[m(y_i - w_i'\beta, \hat{p}_i) - \hat{\lambda}(\beta, \hat{p}_i)] - \hat{p}_i\hat{\lambda}_p(\beta, \hat{p}_i)(d_i - \hat{p}_i), \\ \hat{\rho}_i(\beta) &= \hat{\eta}_i(\beta) \otimes [\hat{w}_i - \hat{\mu}(\hat{p}_i)]. \end{aligned}$$

The term in $\hat{\eta}_i(\beta)$ involving λ_p corrects for the estimation of the unknown propensity score p_i, so that the GMM estimator can be formed as if $\hat{\rho}_i(\beta)$ was not estimated. For an initial estimator $\bar{\beta}$, $\bar{\varepsilon}_i = y_i - w_i'\bar{\beta}$, and $m_\varepsilon(\varepsilon, p) = \partial m(\varepsilon, p)/\partial\varepsilon$ let

$$\begin{aligned} \hat{\Omega} &= \sum_{i=1}^{n} \hat{\rho}_i(\bar{\beta})\hat{\rho}_i(\bar{\beta})'/n, \\ \hat{G} &= \sum_{i=1}^{n} d_i m_\varepsilon(\bar{\varepsilon}_i, \hat{p}_i) \otimes \{[\hat{w}_i - \hat{\mu}(\hat{p}_i)]w_i'\}/n. \end{aligned}$$

The estimator we consider is a one-step GMM estimator formed as

$$\hat{\beta} = \bar{\beta} - (\hat{A}\hat{G})^{-1}\hat{A}\sum_{i=1}^{n} \hat{\rho}_i(\bar{\beta})/n, \hat{A} = \hat{G}'\hat{\Omega}^{-1}. \tag{17}$$

In the case where w is exogenous these moment conditions contain all the information available from the model in equation (17). Indeed, Newey and Powell (1999) showed that the spanning condition is satisfied as the dimension of $m(\varepsilon, p)$ grows, as long as linear combinations of $m(\varepsilon, p)$ can approximate any function of (ε, p) in mean-square. When some components of w are endogenous, as in our application, the optimal instruments $\hat{w}$ may be nonlinear in x. We leave the issue of the optimal instruments to future work, choosing here to focus on the shape of the distribution of ε, as motivated by Figure 1 and the spanning condition for the exogenous case, by using nonlinear functions of ε in the moment conditions.

For $m(\varepsilon, p)$ we consider using subvectors of $(\varepsilon, \varepsilon^2, \varepsilon^3, \varepsilon^4)'$ or $(\Phi(\varepsilon), \Phi(\varepsilon)^2, \Phi(\varepsilon)^3, \Phi(\varepsilon)^4)'$, where $\Phi(\varepsilon)$ is the standard normal CDF. Powers in $\Phi(\varepsilon)$ were considered because they should be more robust, i.e. less sensitive to outliers in ε, than raw powers of ε. We only considered up to fourth order powers because the shape of Figure 1 indicates that much of the efficiency gain should come from these. The bimodal shape of the density corresponds to a log derivative of the density that changes sign three times, and has different slopes at the ends. Such a function is well approximated by a fourth order polynomial. Also, as moment restrictions are added, standard errors become less reliable. Monte Carlo results in Newey (1988) for regression showed high accuracy of standard errors up until about the fourth order, but rapid deterioration beyond that point. Although those results are only suggestive, being for a different model, we should be cautious about going beyond the number suggested by Figure 1.

For the estimated propensity score $\hat{p}_i$ we use the estimated probit probabilities given in Newey, Powell, and Walker (1990). The probit specification includes many nonlinear and interaction terms in background variables, and is thus based on flexible functional forms. In several likelihood ratio tests we found no evidence of additional nonlinear or interaction terms beyond those previously specified. On this basis we view the probit probabilities as being nonparametric. This view is consistent with the results of Newey, Powell, and Walker (1990), where the kernel estimator of Ahn and Powell (1993) produced nearly identical coefficient estimates to the estimators based on $\hat{p}_i$.

For the conditional expectation estimators $\hat{\lambda}(\beta, p)$ and $\hat{\mu}(p)$ we used the predicted values from regression on $(1, \tau(p), \tau(p)^2, \tau(p)^3)$ in the selected data, where $\tau(p) = \varphi(\Phi^{-1}(p))/p$ and φ denotes the standard normal density. This choice corresponds to a series estimator of the conditional expectation, with approximating function given by powers of the inverse Mills ratio. The corresponding sample selection correction is exactly right in the Gaussian case for the Heckman (1979) two step estimator, while including higher order terms allows for any functional form. A cubic specification was chosen because it was slightly more flexible than the quadratic one found by cross-validation in Newey, Powell, and Walker (1990).

6.2 Estimation Results

The initial estimator $\bar{\beta}$ is a two step semiparametric instrumental variables estimator like that of Newey, Powell, and Walker (1990). Our results differ in allowing nonlabor income to be endogenous, which may be important due to family choices (e.g. see the discussion in Mroz, 1987). When we did not instrument for this variable we found that it was significantly positive when higher moments are used, which is contrary to usual presumption that leisure is a normal good.

The first set of estimation results are presented below in Table 1. The first Column contains the results of our initial estimator $\bar{\beta}$. These results are comparable to those of Newey, Powell, and Walker (1990), the main difference being the negative sign of nonlabor income. The remaining three Columns contain the results for more efficient estimators, using information from nonlinear moments. The Column $L, (L = 2, ..., 4)$, uses powers of ε up to the L^{th}.

Table 1 - Estimates of Hours Equation Using Powers of Residual

Variable	(1)	(2)	(3)	(4)
Log Wage	183	296	216	126
	(259)	(229)	(141)	(90)
Nonwife Income	-5.5	-12.3	-1.2	0.8
	(16.7)	(14.4)	(11.5)	(8.0)
Young Children	57	-66	40	7
	(188)	(140)	(90)	(64)
Older Children	-65	-37	-50	-71
	(40)	(33)	(30)	(25)
Age	5.2	7.5	4.5	-2.3
	(7.1)	(6.2)	(4.5)	(3.0)
Education	-75	-62	-88	-65
	(40)	(37)	(20)	(14)

Notes: Col. (1) Series; Col. (2) Two Powers, Col. (3) Three Powers, Col. (4) Four Powers.

The results indicate clearly that use of information in the higher moments significantly reduces the standard errors. Comparing Column 1 to 4, t-statistics increase greatly. The regression changes from one where no variables are significant at a .05 level to one where two are significant. In addition, the wage coefficient is much more precisely estimated, with a standard error that is only about 1/3 the size when four powers are used, and a t-statistic that is significant at the .10 level.

The next set of results, presented in Table 2 below, are generated using powers of $\Phi(\varepsilon)$ instead of simple powers. The Column $L, (L = 2, ..., 4)$, uses powers of $\Phi(\varepsilon)$ up to the L^{th}.

Table 2 - Estimates of Hours Equation Using Powers of Normal CDF of Residual

Variable	(2)	(3)	(4)
Log Wage	159	118	220
	(292)	(194)	(130)
Nonwife Income	-9.4	-5.9	-3.2
	(10.7)	(8.3)	(5.6)
Young Children	-52	-45	99
	(230)	(166)	(86)
Older Children	-74	-93	-26
	(43)	(32)	(22)
Age	6.5	4.3	5.2
	(5.3)	(3.6)	(3.2)
Education	-56	-64	-90
	(32)	(28)	(17)

Notes: Col. (1) Series; Col. (2) Two Powers, Col. (3) Three Powers, Col. (4) Four Powers.

Here also the standard errors decrease, although not as much. The p-value for the estimated effect of the wage variable is .091, which is very small relative to that for the estimator which just uses the linear restriction. These results are quite good, considering that the sample size here is very small relative to many cross-section and panel data applications with selection. Overall, we find statistically significant results, unlike the findings in Newey, Powell, and Walker (1990) for the two step least squares estimator.

COMMENT ON DIFFERENCES BETWEEN 1 AND 2

One concern about this approach is that use of multiple moment conditions requires stronger assumptions about distributions, namely that all the conditions are satisfied rather than just those for least squares. These assumptions could well be violated in empirical applications, where some misspecification is bound to be present. We test for this violation using a sequence of Hausman tests, testing each column of Tables 1 and 2 versus the previous one. We use Hausman tests because the effect of adding additional moment restrictions can be clearly seen from the estimates. For Table 1 this procedure tests the full set of moment restrictions, because the additional number used by each column is equal to the number of parameters, which is equal to the degrees of freedom of each Hausman test. Under the null hypothesis that the moment restrictions are true, these tests will be independent, so that correct p-values can easily be computed.[2]

For the Hausman tests corresponding to Table 1, where we tested column 2 versus 1, column 3 versus 2, and column 4 versus 3, we take the test statistic to be the larges of the three, which was 7.4, for column 3 versus 4. This is not a very large value for a chi-squared distribution with 6 degrees of freedom, with a p-value .29. The p-value test statistic given by the maximum of the three, which is computed as described in the above footnote, is .64.

For the Hausman tests corresponding to Table 2, where we tested column 2 versus 3 and column 3 versus 4, the largest Hausman test statistic was 17, for column 3 versus 4. This is a large value, with corresponding p-value .01 for the individual test and p-value .02 for the joint test. The test statistic for column 2 versus 3 is only 2.3, suggesting the misspecification only is a problem for column 4 of Table 2.

Overall, the results of Table 2 show some evidence against the independence hypothesis on which these estimators are based, but do not reject the use of higher moments. None of the columns in Table 1 are rejected, and only column 4 in Table 2. For the columns that are not rejected we find large reductions in standard errors that are not accompanied by misspecification of the information from the additional moment conditions used in estimation.

[2] To see why, consider several estimators based on increasing numbers of moment restrictions $\hat{\beta}^k$, $(k = 1, ..., K)$. They will be joint asymptotically normal such that for $\bar{k} \geq k$, asymptotic variance V, and covariance C we have $C(\hat{\beta}^{\bar{k}}, \hat{\beta}^k) = V(\hat{\beta}^{\bar{k}})$. Then for $k_1 \leq k_2 \leq k_3 \leq k_4$

$$C(\hat{\beta}^{k_1} - \hat{\beta}^{k_2}, \hat{\beta}^{k_3} - \hat{\beta}^{k_4}) = V(\hat{\beta}^{k_3}) - V(\hat{\beta}^{k_4}) - V(\hat{\beta}^{k_3}) + V(\hat{\beta}^{k_4}) = 0.$$

Since the differences are joint asymptotically normal, zero covariance implies independence of the differences, and hence of the Hausman tests. For two independent tests with the same critical values, and minimum p-value $\hat{p}$ over the two tests, the correct p-value will be $2\hat{p} - \hat{p}^2$. For three tests, with minimum p-value $\hat{p}$, the correct p-value will be $3\hat{p}(1 - \hat{p}) + \hat{p}^3$.

7 Appendix A: Proofs

Throughout the Appendix C will denote a generic constant that may be different in different uses and I will denote the same identity matrix that gives $\tilde{\theta} = [I, 0]\hat{\theta}$. To prove Theorem 1, we derive a projection formula for the moment limit Σ_ρ and compare it with a well known formula for the semiparametric variance bound Σ. Let $proj(Y|A)$ denote the vector of orthogonal projections of the elements of a random vector Y on a closed linear set A, in the Hilbert space of random variables with inner product $\langle Y_1|Y_2\rangle = E[Y_1 \cdot Y_2]$. Also, let $S_\beta = \partial \ln f(y|x, \beta, h_0)/\partial\beta|_{\beta=\beta_0}$. The semiparametric variance bound for estimators of β is $V = Var(S_\beta - proj(S_\beta|T))^{-1}$. Consider any fixed $J = \bar{J}$ big enough that $(\gamma_1, ..., \gamma_{\bar{J}})$ includes the nuisance parameters that are present in $\tilde{\theta} = [I, 0]\hat{\theta}$. Let $\bar{\rho}$ and $\bar{G}(x)$ be the corresponding residual vector and derivative expectation. Consider any $\bar{A}(x)$ such that $E[\bar{A}(x)G(x)]$ is nonsingular and $Var(\bar{A}(x)\bar{\rho})$ exists, and let

$$\psi(z) = \bar{B}(x)\bar{\rho}, \bar{B}(x) = -[I, 0](E[\bar{A}(x)\bar{G}(x)])^{-1}\bar{A}(x).$$

Then $\psi(z)$ is the influence function of a GMM estimator with $J = \bar{J}$ and $A(x) = \bar{A}(x)$, meaning that $\sqrt{n}(\tilde{\theta} - \theta_0) = \sum_{i=1}^{n} \psi(z_i)/\sqrt{n} + o_p(1)$. Also let the full tangent space of the semiparametric model be $\Psi = \{a'S_\beta + t : a \in \Re^q, t \in T\}$. Then, as shown in Bickel et. al. (1993, Proposition 3.3.1), the semiparametric bound for the asymptotic variance of $\tilde{\theta}$ is

$$\Sigma = Var(proj(\psi|\Psi)). \tag{18}$$

Let $\Psi_\rho = \{a'S_\beta + t : a \in \Re^q, t \in T_\rho\}$ be the corresponding space for the moment functions. Then it turns out that

$$lim_{J\to\infty}\Sigma_J = \Sigma_\rho = Var(proj(\psi|\Psi_\rho)), \tag{19}$$

as will be shown below. The proof of Theorem 1 will follow from this result.

To show eq. (19), we need several intermediate results and some additional notation. For a vector $h = (h_1, ..., h_q)$ of elements of a Hilbert space P let $[h] = \{\gamma' h : \gamma \in \Re^q\}$ denote the linear span of h. Also, let $\oplus$ denote the direct sum of two linear subspaces, i.e. $M \oplus N = \{m+n : m \in M, n \in N\}$. Also, for a closed linear subspace L let $proj(h|L)$ denote the vector of orthogonal projections of the elements of h on L, satisfying $proj(h_j|L) \in L$ and $\langle h_j - proj(h_j|L)|t\rangle = 0$ for all $t \in L$.

Lemma A1: *If $L_1, L_2, ...$ is a sequence of closed linear subsets of a Hilbert space P and h is a vector of elements of P such that $m = h - proj(h|\cap_{j=1}^{\infty} L_j)$ has a corresponding nonsingular matrix $Q = [\langle m_k|m_l\rangle_{k,l=1}^{q}]$ of inner products, then for any $a \in P$, we have $proj(a|[h] \oplus \cap_{j=1}^{J} L_j) \to proj(a|[h] \oplus \cap_{j=1}^{\infty} L_j)$ as $J \to \infty$.*

Proof: Denote $L^J = \cap_{j=1}^{J} L_j$, $L^\infty = \cap_{j=1}^{\infty} L_j$, and $m_J = h - proj(h|L^J)$. By Lemma 4.5 of Hansen and Sargent (1991), $m_J \to m = h - proj(h|L^\infty)$. Then $Q_J = [\langle m_{Jk}|m_{Jl}\rangle_{k,l=1}^{q}] \to Q$, so that Q_J is nonsingular for large enough J. Therefore,

$$proj(a|[m_J]) = m_J' Q_J^{-1}(\langle m_{Jk}|a\rangle)_{k=1}^{q} \to m'Q^{-1}(\langle m_k|a\rangle)_{k=1}^{q} = proj(a|[m])$$

Then by orthogonality of $[m_J]$ and L^J and orthogonality of $[m]$ and L^∞, standard Hilbert space theory and Lemma 4.5 of Hansen and Sargent (1991) gives,

$$\begin{aligned} proj(a|[h] \oplus L^J) &= proj(a|[m_J] \oplus L^J) = proj(a|[m_J]) + proj(a|L^J) \\ &\to proj(a|[m]) + proj(a|L^\infty) = proj(a|[m] \oplus L^\infty) = proj(a|[h] \oplus L^\infty). \end{aligned}$$

Q.E.D.

We now consider the Hilbert space of random variables with inner product $\langle X|Y\rangle = E[XY]$.

Lemma A2: *For $T_{J\rho} = \{t : E[\rho t|x] = 0\}$ and $T_{JH} = \{c'H(x)'\Omega(x)^{-}\rho : c \in \Re^r\}$, $T_J = T_{J\rho} \oplus T_{JH}$ and $T_{J\rho}$ and T_{JH} are orthogonal.*

Proof: Consider $t \in T_J$. Then $E[\rho t|x] = H(x)c$. Let $t_H = c'H(x)'\Omega(x)^{-}\rho$ and $t_\rho = t - t_H$. Then $t_H \in T_{JH}$ by construction, while $t_\rho \in T_{J\rho}$ by $E[\rho t_\rho|x] = E[\rho t|x] - E[\rho\rho'|x]\Omega(x)^{-}H(x)c = H(x)c - \Omega(x)\Omega(x)^{-}\Omega(x)R(x)c = H(x)c - \Omega(x)R(x)c = 0$. Then $t = t_\rho + t_H \in T_{J\rho} \oplus T_{JH}$. Furthermore, for $t = t_\rho + t_H \in T_{J\rho} \oplus T_{JH}$, we have $E[\rho t|x] = E[\rho t_\rho|x] + E[\rho t_H|x] = E[\rho\rho'|x]\Omega(x)^{-}H(x)c = H(x)c$, so $t \in T_J$. Furthermore, for any $t_\rho \in T_{J\rho}, t_H \in T_{JH}$, $E[t_\rho t_H] = E[E[t_\rho t_H|x]] = E[c'H(x)'\Omega(x)^{-}E[\rho t_\rho|x]] = 0$.

The following result will hold with $F(x) = \Omega(x)^{-1}E[a\rho|x]$ when $\Omega(x)$ is nonsingular, but requires a proof for $\Omega(x)$ singular.

Lemma A3: *For any a with $E[a^2]$ finite, there exists $F(x)$ such that $E[a\rho|x] = \Omega(x)F(x)$.*

Proof: Consider $J \times 1$ random vector $\delta(x)$ such that $\|\delta(x)\| \leq 1$ and $\Omega(x)\delta(x) = 0$ with probability one. Then $E[\{\rho'\delta(x)\}^2]$ exists, so that so does $E[\{\rho'\delta(x)\}^2|x] = \delta(x)'\Omega(x)\delta(x) = 0$. Then by iterated expectations, $E[\{\rho'\delta(x)\}^2] = 0$, and hence $\rho'\delta(x) = 0$. It follows that $E[a\rho'|x]\delta(x) = E[a\rho'\delta(x)|x] = 0$. Since this equality holds for any such $\delta(x)$, it follows that with probability one $E[a\rho'|x]$ is orthogonal to the null space for $\Omega(x)$. By symmetry of $\Omega(x)$, its range and the null space are orthogonal subspaces of $\Re^J$, so that $\Re^J$ is the direct sum of the range and null space. Consequently, $E[a\rho|x]$ must be in the range of $\Omega(x)$. Q.E.D.

Lemma A4: *For any generalized inverse $\Omega(x)^-$ and any a with $E[a^2]$ finite, $E[\{E[a\rho'|x]\Omega(x)^-\rho\}^2]$ and $E[H(x)'\Omega(x)^-E[\rho a|x]]$ are finite, and $E[E[a\rho'|x]\Omega(x)^-E[\rho a|x]] \leq E[a^2]$.*

Proof: By Lemma A3 there is $F(x)$ such that $E[a\rho'|x] = \Omega(x)F(x)$, so that for $\bar{a} = E[a\rho'|x]\Omega(x)^-\rho$,

$$\begin{aligned} E[\bar{a}^2|x] &= E[a\rho'|x]\Omega(x)^-\Omega(x)\Omega(x)^{-\prime}E[a\rho|x] \\ &= F(x)'\Omega(x)\Omega(x)^-\Omega(x)\Omega(x)^{-\prime}\Omega(x)F(x) \\ &= F(x)'\Omega(x)F(x) = E[a\rho'|x]\Omega(x)^-E[a\rho|x] \end{aligned}$$

is invariant to the generalized inverse. Let $\Lambda(x)$ denote a diagonal matrix of eigenvalues of $\Omega(x)$ and $B(x)$ an orthonormal matrix with $\Omega(x) = B(x)'\Lambda(x)B(x)$. Let $\Lambda(x)^{-1/2}$ denote the matrix with diagonal elements equal to the inverse square root of corresponding nonzero elements of $\Lambda(x)$ and zeros where $\Lambda(x)$ is zero, and $L(x) = B(x)'\Lambda(x)^{-1/2}B(x)$. Then $L(x)^2$ is a generalized inverse, and by the Cauchy-Schwartz inequality,

$$\begin{aligned} E[a\rho'|x]\Omega(x)^-E[a\rho|x] &\leq E[a^2|x]E[\rho'L(x)^2\rho|x] = E[a^2|x]tr(L(x)\Omega(x)L(x)) \\ &= E[a^2|x]tr(L(x)\Omega(x)L(x)) = E[a^2|x]rank(\Omega(x)). \end{aligned}$$

Taking expectations of both sides give the first conclusion. To show the second conclusion, let $\bar{b} = H(x)'\Omega(x)^-\rho$ and note that $H(x)'\Omega(x)^-E[\rho a|x] = E[\bar{b}\bar{a}|x]$. By Assumption 2,

$E[\bar{b}\bar{b}']$ is finite, so that by the Cauchy Schwartz inequality,

$$E[\|E[\bar{b}\bar{a}|x]\|] \leq E[E[\|\bar{b}a\||x]] = E[\|\bar{b}a\|] < \infty.$$

To show the last conclusion, note that $E[a\rho'|x]\Omega(x)^{-}E[\rho a|x] = E[\bar{a}^2|x]$ and that $E[a\bar{a}|x] = E[\bar{a}^2|x]$. Q.E.D.

Lemma A5: *$T_{J\rho}$ is closed in mean-square.*

Proof: Consider $t_k \longrightarrow t, t_k \in T_{J\rho}$. Then by Lemma A4, for any symmetric, p.s.d. generalized inverse

$$\begin{aligned} 0 &\leq E[E[\rho t|x]'\Omega(x)^{-}E[\rho t|x]] = E[E[\rho(t-t_k)|x]'\Omega(x)^{-}E[\rho(t-t_k)|x]] \\ &\leq E[(t-t_k)^2] \longrightarrow 0. \end{aligned}$$

It follows that $E[\rho t|x]'\Omega(x)^{-}E[\rho t|x] = 0$. By Lemma A3, $0 = E[\rho t|x]'\Omega(x)^{-}E[\rho t|x] = F(x)'\Omega(x)F(x)$. Then for any square root matrix $B(x)$ with $B(x)'B(x) = \Omega(x)$, we have $\|B(x)F(x)\|^2 = F(x)'\Omega(x)F(x) = 0$. It follows that $B(x)F(x) = 0$, and hence that $E[\rho t|x] = \Omega(x)F(x) = B(x)'[B(x)F(x)] = 0$. Therefore, $t \in T_{J\rho}$. Q.E.D.

Lemma A6: *For $m_\gamma = H(x)'\Omega(x)^{-}\rho$*

$$proj(a|T_J) = a - E[a\rho'|x]\Omega(x)^{-}\rho + E[am_\gamma'](E[m_\gamma m_\gamma'])^{-1}m_\gamma.$$

Proof: By Lemma A4 $a_\rho = a - E[a\rho'|x]\Omega(x)^{-}\rho$ has finite mean square. Also, $E[\rho a_\rho|x] = E[\rho a|x] - \Omega(x)\Omega(x)^{-}E[\rho a|x] = 0$ by Lemma A3, so that $a_\rho \in T_{J\rho}$, and for any $t \in T_{J\rho}$, $E[(a-a_\rho)t] = E[E[a\rho'|x]\Omega(x)^{-}\rho t] = E[E[a\rho'|x]\Omega(x)^{-}E[\rho t|x]] = 0$. Therefore, $a_\rho = proj(a|T_{J\rho})$. Also,

$$\begin{aligned} E[m_\gamma m_\gamma'] &= E[H(x)'\Omega(x)^{-}\Omega(x)\Omega(x)^{-}H(x)] \qquad (20) \\ &= E[H(x)'\Omega(x)^{-}\Omega(x)\Omega(x)^{-}\Omega(x)R(x)] \\ &= E[H(x)'\Omega(x)^{-}\Omega(x)R(x)] = E[H(x)'\Omega(x)^{-}H(x)] \end{aligned}$$

is nonsingular by Assumption 2, so that $E[am_\gamma'](E[m_\gamma m_\gamma'])^{-1}m_\gamma = proj(a|T_{JH})$. The conclusion then follows Lemmas A2 and A5 and from the standard Hilbert space result that the projection on a sum of orthogonal subspaces is the sum of the projections. Q.E.D.

Lemma A7: *For all J large enough and $\psi_J = proj(\psi|[S_\beta] \oplus T_J)$, it is the case that $E[\psi_J \psi_J'] = \Sigma_J$.*

Proof: Let $m_\beta = D(x)'\Omega(x)^- \rho$. It follows similarly to eq. (20) that $E[m_\beta m_\gamma'] = E[D(x)'\Omega(x)^- H(x)]$. Also, it follows from Assumption 1 and Lemma 5.4 of Newey and McFadden (1994) that $D(x) = -E[\rho S_\beta'|x]$. Therefore, by Lemma A6,

$$\begin{aligned} S_\beta - proj(S_\beta|T_J) &= E[S_\beta \rho'|x]\Omega(x)^- \rho - E[E[S_\beta \rho'|x]\Omega(x)^- H(x)](E[m_\gamma m_\gamma'])^{-1} m_\gamma \\ &= -m_\beta + E[m_\beta m_\gamma'](E[m_\gamma m_\gamma'])^{-1} m_\gamma = U, \end{aligned}$$

where the last equality defines U.

Next, let $\bar{\Omega}(x) = E[\bar{\rho}\bar{\rho}'|x]$ and consider any $J \geq \bar{J}$. Then for any $t \in T_{J\rho}$, $E[\psi t] = E[\bar{B}(x)E[\bar{\rho}t|x]] = 0$. Hence, the components of ψ are in the orthogonal complement of $T_{J\rho}$, so that $proj(\psi|T_{J\rho}) = 0$. It follows by Lemma A2 that $proj(\psi|T_J) = proj(\psi|T_{JH})$. Then, since $(U', m_\gamma')'$ is a nonsingular linear combination of $m = (m_\beta', m_\gamma')' = G(x)'\Omega(x)^- \rho$, it follows by standard Hilbert space theory that

$$\begin{aligned} \psi_J &= proj(\psi|[S_\beta] \oplus T_J) = proj(\psi|[U] \oplus T_J) = proj(\psi|[U]) + proj(\psi|T_J) \\ &= proj(\psi|[U]) + proj(\psi|T_{JH}) = proj(\psi|[(U', m_\gamma')']) = proj(\psi|[m]) \\ &= E[\psi m'](E[mm'])^{-1} m. \end{aligned}$$

Let K be the selection matrix so that $\bar{\rho} = K\rho$. Note that by construction, $\bar{\rho}(z, \beta, \gamma_{\bar{J}})$ does not depend on the parameters in $\hat{\theta}$ that are not in $\tilde{\theta}$, so that $KG(x) = \partial E[\bar{\rho}(z, \beta, \gamma_{\bar{J}})|x]/\partial \theta = [\bar{G}(x), 0]$. Also, by $D(x) = -E[\rho S_\beta'|x]$, Lemma A3, and Assumption 2 there is an $F(x)$ such that $G(x) = \Omega(x)F(x)$. Therefore, for $\bar{\Sigma} = (E[\bar{A}(x)\bar{G}(x)])^{-1}$,

$$\begin{aligned} -E[\psi m'] &= [I, 0]\bar{\Sigma}E[\bar{A}(x)\bar{\rho}\rho'\Omega(x)^- G(x)] = [I, 0]\bar{\Sigma}E[\bar{A}(x)K\Omega(x)\Omega(x)^- G(x)] \quad (21) \\ &= [I, 0]\bar{\Sigma}E[\bar{A}(x)KG(x)] = [I, 0]\bar{\Sigma}[E[\bar{A}(x)\bar{G}(x)], 0] = [I, 0], \end{aligned}$$

where the identity matrix I has the same number of rows as $\tilde{\theta}$ throughout. Finally, noting that $(E[mm'])^{-1}$ is the bound for $\hat{\theta}$, it follows from the last equation that $E[\psi_J \psi_J'] = E[\psi m'](E[mm'])^{-1}E[m\psi'] = [I, 0](E[mm'])^{-1}[I, 0]' = \Sigma_J$. Q.E.D.

Proof of Theorem 1: By Lemma A1, $\psi_J \to \psi^* = proj(\psi|[S_\beta] \oplus \cap_{J=1}^{\infty} T_J)$, so that by Lemma A7, $\Sigma_J = E[\psi_J \psi_J'] \to E[\psi^* \psi^{*\prime}]$. Q.E.D.

Proof of Theorem 2: For simplicity suppress the x argument in $\pi(x), D(x), H(x)$, and $\Omega(x)$. Let $\pi^* = -D'\Omega^- + E[D'\Omega^- H](E[H'\Omega^- H])^{-1}H'\Omega^-$. Note that $E[\pi^*\rho(\pi^*\rho)'] = V_J$. Also, let $K = E[\rho S'|x]$. Then by Assumption 1 and Lemma 5.4 of Newey and McFadden (1994), $D = -E[\rho S_\beta|x]$, and by $T \subseteq T_\rho$, $E[\rho \cdot proj(S_\beta|T)'|x] = HC$ for some matrix C, so that $K = E[\rho S_\beta'|x] - E[\rho \cdot proj(S_\beta|T)'|x] = -D - HC$. Therefore,

$$\begin{aligned} & K'\Omega^- - E[K'\Omega^- H](E[H'\Omega^- H])^{-1}H'\Omega^- \\ = \; & \pi^* - C'H'\Omega^- + C'E[H'\Omega^- H](E[H'\Omega^- H])^{-1}H'\Omega^- = \pi^* \end{aligned}$$

Note that $E[\pi^* H] = 0$. Also, for any π with $E[\pi H] = 0$, since Lemma A3 implies that there is F with $K = \Omega F$ and by Assumption 2, $H = \Omega R$, for $B = E[K'\Omega^- H](E[H'\Omega^- H])^{-1}$ we have

$$\begin{aligned} E[\pi^*\rho\rho'\pi'] &= E[\pi^*\Omega\pi'] = E[F'\Omega\Omega^-\Omega\pi'] - BE[R'\Omega\Omega^-\Omega\pi'] \\ &= E[K'\pi'] - BE[H'\pi'] = E[E[S\rho'|x]\pi'] = E[S\rho'\pi']. \end{aligned}$$

Then it follows that for any π with $E[\pi H] = 0$,

$$E[\{S - \pi^*\rho\}\{\pi\rho\}'] = E[S\rho'\pi'] - E[\pi^*\rho\rho'\pi'] = 0.$$

Therefore,

$$E[(S - \pi^*\rho)(S - \pi^*\rho)'] = E[SS'] - E[\pi^*\rho\rho'\pi^{*\prime}] = V^{-1} - V_J^{-1}, \tag{22}$$

so that for any $\bar{\pi}$ with $E[\bar{\pi}H] = 0$, since $\pi = \pi^* - \bar{\pi}$ also satisfies $E[\pi H] = 0$,

$$\begin{aligned} E[(S - \bar{\pi}\rho)(S - \bar{\pi}\rho)'] &= E[(S - \pi^*\rho + \pi\rho)(S - \pi^*\rho + \pi\rho)'] \\ &= E[(S - \pi^*\rho)(S - \pi^*\rho)'] + E[\pi\rho(\pi\rho)'] \end{aligned}$$

The conclusion then follows from the last two eqs. and $E[\pi\rho(\pi\rho)']$ p.s.d.. Q.E.D.

Proof of Theorem 3: Consider any $t \in T_\rho$. Then t will satisfy $E[\rho t|x] = Hc$ for every J. Then for every $m = \pi(x)\rho \in M$, $E[mt] = E[\pi(x)E[\rho t|x]] = E[\pi(x)H]c = 0$.

Thus we have $T_\rho \subseteq M^\perp$. Now consider $t \in M^\perp$. For a symmetric, p.s.d. Ω^- let $K = E[\rho t|x]$, $c = (E[H'\Omega^- H])^{-1}E[H'\Omega^- K]$, and $\pi(x) = (K - Hc)'\Omega^-$, where $E[H'\Omega^- K]$ exists by Lemma A4. Note that $E[\|\pi(x)\rho\|^2] = E[(K - Hc)'\Omega^-(K - Hc)]$ also exists by Lemma A4 and that $E[\pi H] = E[K'\Omega^- H\ ] - c'E[H'\Omega^- H] = 0$. Therefore $\pi(x)\rho \in M$, implying

$$0 = E[\pi(x)\rho t] = E[\pi(x)K] = E[(K - Hc)'\Omega^- K] = E[(K - Hc)'\Omega^-(K - Hc)],$$

where the last equality follows by $E[\pi H] = 0$. It follows that $(K - Hc)'\Omega^-(K - Hc) = 0$. Since $H = \Omega R$ by Assumption 2 and $K = \Omega F$ by Lemma A3, it follows that $(F - Rc)'\Omega(F - Rc) = 0$, implying $\Omega^{1/2}(F - Rc) = 0$ for any square root matrix, implying

$$0 = \Omega(F - Rc) = K - Hc = E[\rho t|x] - H(x)c.$$

Therefore, $t \in T_\rho$. Since this inclusion holds for any $t \in M^\perp$ it follows that $M^\perp \subseteq T_\rho$, and hence that $T_\rho = M^\perp$. To prove the second conclusion, note that M is linear and closed, so that by standard Hilbert space theory $(T_\rho)^\perp = M$. Q.E.D.

Proof of Theorem 4: Let $1^\varepsilon_j = 1(\varepsilon > \gamma_{j0})$ and (a^x_k) be a countable basis of bounded functions of x. Then $(1^\varepsilon_j a^x_k)_{j,k=1}^\infty$ is a basis (the proof is available upon request), meaning that if $E[r(\varepsilon, x)^2] < \infty$ and $E[1^\varepsilon_j a^x_k r(\varepsilon, x)] = 0$ for all j and k then $r(\varepsilon, x) = 0$. Note that for $v > -\gamma$ and $y = 0$ we have $\varepsilon = y^* - v \leq -v = y - v$. Hence, $1(v > -\gamma)1(y - v < \gamma) = 1(v > -\gamma)1(\varepsilon < \gamma)$. Therefore,

$$\rho_j(z, \beta_0, \gamma_j) = 1(v > -\gamma_j)[1(\varepsilon < \gamma_j) - \alpha_j].$$

By v continuously distributed, it follows that $E[\rho_j(z, \beta_0, \gamma)|x]$ is differentiable at γ with probability one (w.p.1), with derivative $1(v > -\gamma)g(\gamma)$. Hence, $H_{jj}(x) = -1^v_j g(\gamma_{j0})$ for $1^v_j = 1(v > -\gamma_{j0})$. Consider $t(y, x)$ in the moment tangent set. Since y is a function of x and ε we can regard t as a function of ε and x. Then for $1^\varepsilon_j = 1(\varepsilon > \gamma_{j0})$

$$\begin{aligned} E[\rho_j t|x] &= 1^v_j \int_{-\infty}^{\gamma_{j0}} t(\varepsilon, x)g(\varepsilon)d\varepsilon = -1^v_j E[1^\varepsilon_j t(\varepsilon, x)|x] \\ &= -1^v_j g_j(\gamma_{j0})c_j = -1^v_j d_j = H_{jj}(x)c_j, d_j = g_j(\gamma_{j0})c_j. \end{aligned}$$

Next, consider j with $P_j = E[1^v_j] > 0$, and let $s_j(\varepsilon) = E[1^v_j t|\varepsilon]/P_j = E[t|\varepsilon, 1^v_j = 1]$, and $J(j) = \{j' : \gamma_{j'0} \geq \gamma_{j0}\}$. Note that for any $j' \in J(j)$ we have $1^v_j 1^v_{j'} = 1^v_j$. Then replacing j by j' in the previous equation and multiplying through by 1^v_j gives

$$1^v_j E[1^\varepsilon_{j'} t(\varepsilon, x)|x] = 1^v_j d_{j'} \text{ for all } j' \in J(j).$$

Taking expectations of both sides and dividing by P_j gives $E[1^\varepsilon_{j'} s_j] = E[1^v_j 1^\varepsilon_{j'} t]/P_j = d_{j'}$. Let $1_j = 1^\varepsilon_j 1^v_j$. Note that for any $\tilde{j}$, $1^\varepsilon_{\tilde{j}} 1_j = 1^\varepsilon_{j'} 1^v_j$ for some $j' \in J(j)$. Then

$$\begin{aligned} E[1^\varepsilon_{\tilde{j}} a^x_k 1_j(t - s_j)] &= E[a^x_k 1^v_j 1^\varepsilon_{j'} t] - E[a^x_k 1^v_j 1^\varepsilon_{j'} s_j] \\ &= E[a^x_k 1^v_j E[1^\varepsilon_{j'} t|x]] - E[a^x_k 1^v_j] E[1^\varepsilon_{j'} s_j] \\ &= E[a^x_k 1^v_j d_{j'}] - E[a^x_k 1^v_j] d_{j'} = 0. \end{aligned}$$

Since this equality holds for all $\tilde{j}$ and k it follows that $1_j(t - s_j) = 0$ w.p.1. Since this equality holds for all j, a countable number of these, w.p.1 we have $1_j(t - s_j) = 0$ for *all* j with $P_j > 0$.

Next, consider any j and j' with $\gamma_{j'} \geq \gamma_j$ and $P_j > 0$. Then

$$0 = 1_j 1_{j'}(s_j - s_{j'}) = 1^v_j 1^\varepsilon_{j'}(s_j - s_{j'}).$$

Taking conditional expectations given ε, by independence of x and ε, $P_j 1^\varepsilon_{j'}(s_j - s_{j'}) = 0$, implying $1^\varepsilon_{j'}(s_j - s_{j'}) = 0$, so that $s_j(\varepsilon) = s_{j'}(\varepsilon)$ w.p.1 for $\varepsilon > \gamma_{j'}$. It then follows in a straightforward way that there is an $s(\varepsilon)$ such that $1_j(t - s) = 0$ for all j (details available upon request). Then, noting that by denseness of the quantiles, $\cup_{j=1}^\infty \{(\varepsilon, x) : \varepsilon > \gamma_{j0}, v > -\gamma_{j0}\} = \{(\varepsilon, x) : \varepsilon + v > 0\}$, it follows that $t(\varepsilon, x) = s(\varepsilon)$ for $y > 0$. Then, as noted in the text, it follows that $t(\varepsilon, x)$ is an element of the model tangent set. Thus the spanning condition is satisfied. Q.E.D.

Details for the proof of Theorem 4: Let $\bar{v}$ be the least upper bound for the support of v, that may be ∞. Define $s(\varepsilon) = 0$ for $\varepsilon \leq -\bar{v}$. Let $(\gamma_\ell)_{\ell=1}^\infty$ denote a monotonic decreasing subsequence of $(\gamma_{0j})_{j=1}^\infty$, with $\gamma_\ell > -\bar{v}$, $\gamma_\ell \to -\bar{v}$ as $\ell \to \infty$. Let $j(\ell)$ solve $\gamma_\ell = \gamma_{j(\ell)0}$ and define $s(\varepsilon) = s_{j(\ell+1)}(\varepsilon)$ for $\gamma_{\ell+1} \leq \varepsilon < \gamma_\ell$ and $s(\varepsilon) = s_{j(1)}(\varepsilon)$ for $\varepsilon \geq \gamma_{j(1)}$. Then for any j with $\gamma_j \geq \gamma_{j(1)}$ we have $P_{j(1)} > 0$, implying $s_j(\varepsilon) = s(\varepsilon)$ for $\varepsilon > \gamma_j$, and hence $1_j(t-s)=0$. For any j with $\gamma_{\ell+1} \le \gamma_j < \gamma_\ell$ we have $P_{j(\ell+1)} > 0$, implying $s_j(\varepsilon) = s_{j(\ell+1)}(\varepsilon)$ for $\gamma_j < \varepsilon$, and hence $s_j(\varepsilon) = s(\varepsilon)$ for $\gamma_j < \varepsilon$, so that $1_j(t-s)=0$. Summarizing, we find that $1_j(t-s)=0$ for all j with $P_j > 0$. Furthermore, for all j with $P_j = 0$ it is automatically true that $1_j(t-s)=0$. Thus for each j, $1_j(t-s)=0$ w.p.1. Since there are a countable of these equalities, it follows that w.p.1, $1_j(t-s)=0$ for *all* j.

.

Proof of Theorem 5: Let $S(\gamma) = \int_\gamma^\infty g(\varepsilon)d\varepsilon$ be the survivor function for $g(\varepsilon)$. By $\ln g(\varepsilon)$ strictly concave and Pratt (1981) it follows that $\ln S(\gamma)$ is strictly concave. Then $d\ln S(\gamma)/d\gamma = -g(\gamma)/S(\gamma)$ is strictly decreasing, so that for any $\tau > 0$,

$$\frac{\partial}{\partial\gamma}\frac{S(\gamma+\tau)}{S(\gamma)} = \frac{S(\gamma+\tau)}{S(\gamma)}\left[-\frac{g(\gamma+\tau)}{S(\gamma+\tau)} + \frac{g(\gamma)}{S(\gamma)}\right] < 0$$

Then by $\lim_{\gamma\to\infty}[S(\gamma+\tau)/S(\gamma)] = 0$, it follows that for any $\tau > 0$ and $0 < \alpha < 1$ there is a unique solution $\gamma(\alpha,\tau)$ to $S(\gamma+\tau)/S(\gamma) = \alpha$, i.e. to $E[\alpha 1(\varepsilon > \gamma) - 1(\varepsilon > \gamma+\tau)] = 0$. By the implicit function theorem, $r(\alpha,\tau) = (\gamma(\alpha,\tau), \tau + \gamma(\alpha,\tau))$ is continuous in (α,τ) and has range $\Gamma = \{(\gamma,\zeta) : \zeta > \gamma\}$. It follows that $\bar{\Gamma} = \{(\gamma_{j0}, \gamma_{j0} + \tau_j)\}$ is dense in Γ.

Now let $m_j^\varepsilon = \alpha_j 1(\varepsilon > \gamma_{j0}) - 1(\varepsilon > \gamma_{j0} + \tau_j)$. Consider any $r(\varepsilon)$ with $E^*[r(\varepsilon)^2] < \infty$, and suppose that $0 = E^*[m_j^\varepsilon r]$ for all j. Then by continuity of the integral it follows that for all $\gamma < \zeta$,

$$S(\zeta)\int_\gamma^\infty r(\varepsilon)g(\varepsilon)d\varepsilon/S(\gamma) - \int_\zeta^\infty r(\varepsilon)g(\varepsilon)d\varepsilon = 0.$$

Differentiating with respect to ζ holding γ fixed we see that $r(\zeta) = \int_\gamma^\infty r(\varepsilon)g(\varepsilon)d\varepsilon/S(\gamma)$ almost everywhere for all $\zeta > \gamma$. By repeated application of this equality for different γ, we find that $r(\varepsilon)$ is constant almost everywhere. Then, if $r = r(\varepsilon,x)$ and a_k^x is as in the proof of Theorem 4, by the Fubini Theorem, $E^*[m_j^\varepsilon a_k^x r] = 0$ implies $\int r(\varepsilon,x)a_k^x F(dx) = c_k$ for some constant c_k. Taking expectations of each of these equalities and subtracting, it follows that for $\bar{r}(x) = \int r(\varepsilon,x)g(\varepsilon)d\varepsilon$ that $\int[r(\varepsilon,x) - \bar{r}(x)]a_k^x F(dx) = 0$ for each k, which implies that $r(\varepsilon,x) = r(x)$. Thus, $E^*[m_j^\varepsilon a_k^x r] = 0$ for all j and k implies $r(\varepsilon,x) = \bar{r}(x)$.

Next we proceed as in the proof of Theorem 4 with notation as given there. It follows similarly to the last paragraph that if $E^*[1_j m_{j'} a_k^x r] = 0$ for all $j' \in J(j)$ and all k then $r(\varepsilon, x) = \bar{r}(x) = \int_{\gamma_{j0}}^{\infty} r(\varepsilon, x)g(\varepsilon)d\varepsilon/P_j$ for all $\varepsilon > \gamma_{j0}$ and $v > -\gamma_{j0}$. Also, it follows similary to the proof of Theorem 4 that $E[\rho_j(z, \beta_0, \gamma)|x]$ is differentiable at γ with probability one (w.p.1), with derivative $H_{jj}(x) = -1_j^v \tilde{d}_j/S(v)$ where $\tilde{d}_j = \alpha_j g(\gamma_{j0}) - g(\gamma_{j0} + \tau_j)$. Then, multiplying through by $S(v)$, for the moment tangents must satisfy

$$S(v)E[\rho_j t|x] = -1_j^v E^*[m_j^\varepsilon t|x] = -1_j^v d_j = H_{jj}(x)c_j S(v), d_j = \tilde{d}_j c_j.$$

It then follows anaogously to Theorem 4 that $E^*[1_j m_{j'}^\varepsilon a_k^x(t - s_j)] = 0$ for all $j' \in J(j)$. Therefore, for $r(\varepsilon, x) = t(\varepsilon, x) - s_j(\varepsilon)$ and the $r_j(x) = \int_{\gamma_{j0}}^{\infty} t(\varepsilon, x)g(\varepsilon)d\varepsilon/S(\gamma_{j0})$ we find that

$$1_j(t - s_j - r_j) = 0,$$

for all j with $P_j > 0$. Then, because t is additive in a function of ε and x for $1_j = 1$ it follows in a straightforward way, similarly to the proof of Theorem 4, that there is $s(\varepsilon)$ and $r(x)$ with $1_j(t - s - r) = 0$ (details available upon request). Then $t(\varepsilon, x) = s(\varepsilon) + t(x)$ for all $y > 0$, so that t is an element of the model tangent set, and the spanning condition is satisfied. Q.E.D.

Details for the proof of Theorem 5: Consider any j and j' with $\gamma_{j'} \geq \gamma_j$ and $P_j > 0$.

$$0 = 1_j 1_{j'}(s_j + r_j - s_{j'} - r_{j'}) = 1_j^v 1_{j'}^\varepsilon(s_j - s_{j'} + r_j - r_{j'}).$$

Define $c_{jj'} = E[r_j - r_{j'}|v > -\gamma_j]$. Taking conditional expectations given ε, and using the definition of $c_{jj'}$ we find that

$$1_{j'}^\varepsilon(s_j - s_{j'} + c_{jj'}) = 0, 1_j^v(r_j - r_{j'} - c_{jj'}) = 0.$$

Also, for $\gamma_{j''} > \gamma_{j'}$, we have $1_{j''}^\varepsilon(s_{j'} - s_{j''} + c_{j'j''}) = 0$, so that summing gives $1_{j''}^\varepsilon(s_j - s_{j''} + c_{jj'} + c_{j'j''}) = 0$. Then, by the previous equality with $j' = j''$ it follows by $\Pr(\varepsilon > \gamma_{j''}) > 0$ that $c_{jj'} + c_{j'j''} = c_{jj''}$.

Next, let $(\gamma^\ell)_{\ell=-\infty}^{\infty} \subseteq \{\gamma_{j0}\}_{j=1}^{\infty}$ be a monotonic increasing sequence with $\lim_{\ell \to -\infty} \gamma^\ell = -\bar{v}$, where $\bar{v}$ is the least upper bound for the support of v, and $\lim_{\ell \to \infty} \gamma^\ell = \infty$. Define $s^\ell(\varepsilon) = s_j(\varepsilon)$ and $r^\ell(x) = r_j(x)$ for j with $\gamma^\ell = \gamma_{j0}$. Define $c^{\ell\ell'} = c_{jj'}$ for $\gamma^\ell = \gamma_{j0} <$

$\gamma_{j'0} = \gamma^{\ell'}$. Let

$$
\begin{aligned}
s(\varepsilon) &= \begin{cases} s^0(\varepsilon), \varepsilon > \gamma^0 \\ s^\ell(\varepsilon) + c^{\ell 0}, \gamma^{\ell-1} < \varepsilon \leq \gamma^\ell, \ell \leq 0 \end{cases}, \\
r(x) &= \begin{cases} r^0(x), v > -\gamma^0, \\ r^\ell(x) + c^{0\ell}, -\gamma^\ell < v \leq -\gamma^{\ell-1}, \ell \geq 1, \end{cases}.
\end{aligned}
$$

Consider any γ_{j0}, and let $c_j^\ell = c_{jj'}$ for $\gamma^\ell = \gamma_{j'} > \gamma_j$ and $c_j^\ell = c_{j'j}$ for $\gamma^\ell = \gamma_{j'} < \gamma_j$. Consider first the case where $\gamma_{j0} > \gamma^0$. Then for $\varepsilon > \gamma_{j0}$ and $v > -\gamma_{j0}$,

$$t = s_j + r_j = s + c_j^0 + r_j.$$

Now let $\gamma^{\ell-1} < \gamma_{j0} < \gamma^\ell$, i.e. $-\gamma^\ell < -\gamma_{j0} < -\gamma^{\ell-1}$. For x with $-\gamma_{j0} < v \leq -\gamma^{\ell-1}$, note that $c_j^0 + c_j^\ell = c^{0\ell}$, so that

$$r(x) = r^\ell(x) + c^{0\ell} = r_j(x) - c_j^\ell + c^{0\ell} = r_j(x) + c_j^0.$$

so that $t = s + r$ for $\varepsilon > \gamma_{j0}$ and x with $-\gamma_{j0} < v \leq -\gamma^{\ell-1}$. It also follows analogously that $t = s + r$ for all x with $v > -\gamma_{j0}$, so that $1_j(t - s - r) = 0$. The case with $\gamma_{j0} < \gamma^0$. Follows analogously.

Proof of Theorem 6: Without changing notation let $B\hat{\rho}(z, \beta)$ replace $\hat{\rho}(z, \beta)$, noting that this replacement does not affect the estimator. Consider a further transformation, where $B\hat{\rho}(z, \beta)$ replaces $\hat{\rho}(z, \beta)$, for a symmetric square root $B = \Omega^{-1/2}$. By the smallest eigenvalue of Ω bounded away from zero, the largest eigenvalue of Ω^{-1} is bounded. It follows that

$$
\begin{aligned}
\|B(\hat{\Omega} - \Omega)B\|^2 &= tr(B(\hat{\Omega} - \Omega)B^2(\hat{\Omega} - \Omega)B) \leq Ctr(B(\hat{\Omega} - \Omega)(\hat{\Omega} - \Omega)B) \\
&= Ctr((\hat{\Omega} - \Omega)BB(\hat{\Omega} - \Omega)) \leq C\|\hat{\Omega} - \Omega\|^2,
\end{aligned}
$$

Similarly, $\|A^*(\hat{\Omega} - \Omega)B\| \leq C\|A^*(\hat{\Omega} - \Omega)\|$ and $\|B(\hat{G} - G)\| \leq C\|\hat{G} - G\|$. It follows that all the hypotheses of the theorem hold for this replacement, where $\Omega = I$.

Next, note that $-G$ are the coefficients of the mean-square projection of S_i by hypothesis i) (and $\Omega = I$). Then, also by i), it follows that $E[\|S_i - (-G'\rho_i)\|^2] \longrightarrow 0$, and hence that

$$G'G = E[G'\rho_i\rho_i'G] \longrightarrow (V^*)^{-1}. \tag{23}$$

Next, let $\lambda_{\min}(F)$ and $\lambda_{\max}(F)$ denote the smallest and largest eigenvalues of a symmetric matrix F. Note that $\|\hat{\Omega} - I\| \xrightarrow{p} 0$ implies $\lambda_{\max}(\hat{\Omega}^{-1}) = \lambda_{\min}(\hat{\Omega})^{-1} \xrightarrow{p} 1$. It follows that for any conformable sequence Z_n of random matrices, $\|\hat{\Omega}^{-1}Z_n\| \leq O_p(1)\|Z_n\|$. Therefore,

$$\begin{aligned}\|\hat{\Omega}^{-1}\hat{G} - G\| &\leq \|\hat{\Omega}^{-1}(\hat{G} - G)\| + \|\hat{\Omega}^{-1}(I - \hat{\Omega})G\| \\ &\leq O_p(1)(\|\hat{G} - G\| + \|\hat{\Omega} - I\|) \xrightarrow{p} 0.\end{aligned}$$

It then follows that $\|\hat{\Omega}^{-1}\hat{G}\| = O_p(1)$ and hence

$$\|\hat{G}'\hat{\Omega}^{-1}\hat{G} - G'G\| \leq \|(\hat{G} - G)'\hat{\Omega}^{-1}\hat{G}\| + \|G'(\hat{\Omega}^{-1}\hat{G} - G)\| \xrightarrow{p} 0. \tag{24}$$

The second conclusion then follows from eq. (23) and the triangle inequality.

To show the first conclusion, note that by i.i.d. observations and $\Omega = I$, $\bar{\rho} = \sum_{i=1}^{n} \rho_i/\sqrt{n}$ satisfies $E[\|\bar{\rho}\|^2] = E[\|\rho_i\|^2] = J$, so that $\bar{\rho} = O_p(J^{1/2})$, and hence

$$\|\hat{\Omega}^{-1}\bar{\rho}\| = O_p(J^{1/2}). \tag{25}$$

It then follows by conditions iv) and v) that

$$\begin{aligned}\|(\hat{G}'\hat{\Omega}^{-1} - G')\bar{\rho}\| &\leq \|(\hat{G} - G)'\hat{\Omega}^{-1}\bar{\rho}\| + \|G'(I - \hat{\Omega})\hat{\Omega}^{-1}\bar{\rho}\| \\ &\leq (\|\hat{G} - G\| + \|G'(I - \hat{\Omega})\|)\|\hat{\Omega}^{-1}\bar{\rho}\| = o_p(J^{-1/2})O_p(J^{1/2}) \xrightarrow{p} 0.\end{aligned}$$

Also, by condition i) and the Markov inequality,

$$G'\bar{\rho} - \sum_{i=1}^{n} S_i/\sqrt{n} = O_p(\{E[\|S_i - G'\rho_i\|^2]\}^{1/2}) \xrightarrow{p} 0.$$

It then follows by the triangle inequality that $\hat{G}'\hat{\Omega}^{-1}\bar{\rho} = \sum_{i=1}^{n} S_i/\sqrt{n} + o_p(1)$. Also, by vi),

$$\begin{aligned}&\|\hat{G}'\hat{\Omega}^{-1}(\sum_{i=1}^{n} \hat{\rho}_i/\sqrt{n} - \bar{\rho})\| \\ \leq\ &\|\hat{G}'\hat{\Omega}^{-1}\|\|\sum_{i=1}^{n} \hat{\rho}_i/\sqrt{n} - \bar{\rho}\| = O_p(1)o_p(1) \xrightarrow{p} 0.\end{aligned} \tag{26}$$

It follows that $\hat{G}'\hat{\Omega}^{-1}\sum_{i=1}^{n} \hat{\rho}_i/\sqrt{n} \xrightarrow{d} N(0, V^*)$. Then by a mean-value expansion,

$$\begin{aligned}\sqrt{n}(\hat{\beta} - \beta_0) &= [I - (\hat{G}'\hat{\Omega}^{-1}\hat{G})^{-1}\hat{G}'\hat{\Omega}^{-1}\bar{G}]\sqrt{n}(\bar{\beta} - \beta_0) \\ &\quad -(\hat{G}'\hat{\Omega}^{-1}\hat{G})^{-1}\hat{G}'\hat{\Omega}^{-1}\sum_{i=1}^{n} \hat{\rho}_i/\sqrt{n}.\end{aligned} \tag{27}$$

where $\bar{G} = \sum_{i=1}^{n} \hat{\rho}_\beta(z_i, \dot{\beta})/n$ and $\dot{\beta}$ is a mean value. It follows similarly to eq. (24) that $\hat{G}'\hat{\Omega}^{-1}\bar{G} \xrightarrow{p} (V^*)^{-1}$, so by the Slutzky theorem, the first term following the equality in the last eq. converges to zero in probability. The conclusion then follows by applying the Slutzky theorem to the second term. Q.E.D.

Proof of Theorem 7: We proceed by verifying the hypotheses of Theorem 6 for $\rho(z,\beta) = u(z,\beta) \otimes p^J(w)$. Note first that the estimator is invariant to nonsingular linear transformations of $p^J(w)$, so that by iv) we can assume that $p^J(w) = \tilde{p}^J(w)$. Here let $p_i = p^J(w_i)$ and $Q = E[p_i p_i']$. By iv), $\lambda_{\min}(Q)$ is bounded away from zero, so that $\|Q^{-1/2}p^J(w)\|^2 \leq C\zeta(J)^2$. Therefore, all of the hypotheses are satisfied with $Q^{-1/2}p^J(w)$ replacing $p^J(w)$, and hence we may assume throughout that $E[p_i p_i'] = I$.

Let $\Sigma_i = E[u_i u_i' | w_i], D_i = E[u_\beta(z_i, \beta_0) | w_i]$, s denote the dimension of $u(z,\beta)$, and $S_i = -D_i{}'\Sigma_i^{-1}u_i$. Note that

$$G = E[u_\beta(z_i,\beta_0) \otimes p_i] = E[D_i \otimes p_i]. = E[(u_i \otimes p_i)u_i'\Sigma_i^{-1}D_i] = -E[\rho_i S_i']. \qquad (28)$$

Also, by i) and iii), $A_i = -D_i'\Sigma_i^{-1}$ is bounded. By ii) there is a $q \times sJ$ matrix π_J such that $E[\|A_i - \pi_J(I_s \otimes p_i)\|^2] \to 0$ as $J \to \infty$. Therefore, for $\rho_i = u_i \otimes p_i$ and $S_i = A_i\varepsilon_i$,

$$\begin{aligned} E[\|S_i - \pi_J\rho_i\|^2] &= tr(E[\{A_i - \pi_J(I_s \otimes p_i)\}\Sigma_i\{A_i - \pi_J(I_s \otimes p_i)\}']) \\ &\leq CE[\|A_i - \pi_J(I_s \otimes p_i)\|^2] \to 0. \end{aligned}$$

Thus, condition i) of Theorem 6 are satisfied.

Next, note that by iii)

$$E[\rho_i\rho_i'] = E[\Sigma_i \otimes p_i p_i'] \geq E[CI \otimes p_i p_i'] \geq CI,$$

so that condition ii) of Theorem 6 is satisfied.

Next, let $\hat{u}_i = u(z_i, \bar{\beta})$, so that $\hat{\Omega} = \sum_i \hat{u}_i\hat{u}_i' \otimes p_i p_i'/n$, and $\delta_{1i} = \delta_1(z_i)$. By mean value expansions, $\|\hat{u}_i - u_i\| \leq \delta_{1i}\|\bar{\beta} - \beta_0\|$ for each $i \leq n$ with probability approaching one (w.p.a.1). Let $\tilde{\Omega} = \sum_i u_i u_i' \otimes p_i p_i'/n$. Then by $\|\bar{\beta} - \beta_0\| \leq 1$ and $E[\|p_i\|^2] = J$, by the

Markov inequality,

$$\begin{aligned}
\|\hat{\Omega} - \tilde{\Omega}\| &\leq \sum_i \|\hat{u}_i \hat{u}_i' - u_i u_i'\| \|p_i p_i'\|/n \\
&\leq \sum_i (\|\hat{u}_i - u_i\|^2 + 2\|\hat{u}_i - u_i\| \|u_i\|) \|p_i\|^2/n \\
&\leq \|\bar{\beta} - \beta_0\| \sum_i (\delta_{1i}^2 + 2\delta_{1i}\|u_i\|) \|p_i\|^2/n \\
&= O_p(E[(\delta_{1i}^2 + 2\delta_{1i}\|u_i\|)\|p_i\|^2]/\sqrt{n}) = O_p(J/\sqrt{n}) \xrightarrow{p} 0.
\end{aligned}$$

Also, we have

$$\begin{aligned}
E[\|\tilde{\Omega} - \Omega\|^2] &= tr\{E[(u_i u_i' \otimes p_i p_i')^2] - \Omega^2\}/n \leq E[\|u_i\|^4 \|p_i\|^4]/n \\
&\leq CE[\|p_i\|^4]/n \leq C\zeta(J)^2 E[\|p_i\|^2]/n \leq C\zeta(J)^2 J/n \to 0,
\end{aligned}$$

so that by the Markov inequality, $\|\tilde{\Omega} - \Omega\| \xrightarrow{p} 0$. Then by the triangle inequality condition iii) of Theorem 4 is satisfied.

Next, for any $\dot{\beta}$ with $\dot{\beta} = \beta_0 + O_p(n^{-1/2})$, $\dot{u}_{\beta i} = u_\beta(z_i, \dot{\beta})$, $u_{\beta i} = u_\beta(z_i, \beta_0)$, and $\delta_{2i} = \delta_2(z_i)$, let $\dot{G} = \sum_i \rho_\beta(z_i, \dot{\beta})/n = \sum_i \dot{u}_{\beta i} \otimes p_i/n$ and $\tilde{G} = \sum_i u_{\beta i} \otimes p_i/n$. We have

$$\begin{aligned}
\|\dot{G} - \tilde{G}\| &\leq \sum_i \|\dot{u}_{\beta i} - u_{\beta i}\| \|p_i\|/n \leq \|\tilde{\beta} - \beta_0\| \sum_i \delta_{2i} \|p_i\|/n \\
&= O_p(n^{-1/2}\zeta(J)) = o_p(J^{-1/2}).
\end{aligned}$$

Also,

$$\begin{aligned}
E[\|\tilde{G} - G\|^2] &= tr(E[u_{\beta i}' u_{\beta i} \|p_i\|^2] - G'G)/n \leq E[\|u_{\beta i}\|^2 \|p_i\|^2]/n \\
&\leq E[\delta_{1i}^2 \|p_i\|^2]/n \leq C\zeta(J)^2/n.
\end{aligned}$$

By the Markov inequality it follows that $\|\tilde{G} - G\| = O_p(n^{-1/2}\zeta(J)) = o_p(J^{-1/2})$, so that condition v) of Theorem 6 is satisfied.

Next, assume for the moment that β is a scalar. By boundedness of $E[\delta_{1i}^2 | w_i]$ and Σ_i having smallest eigenvalue bounded away from zero (implying $I \leq C\Sigma_i$), for $\rho_{\beta i} = u_{\beta i} \otimes p_i$,

$$\begin{aligned}
E[\|A^* \rho_{\beta i}\|^2] &\leq Ctr\{E[A^*(I \otimes p_i p_i')A^{*\prime}]\} \leq Ctr\{E[A^*(\Sigma_i \otimes p_i p_i')A^{*\prime}]\} \\
&= CE[\|A^* \rho_i\|^2] \leq C.
\end{aligned}$$

Let $\hat{\rho}_i = \hat{u}_i \otimes p_i$, and note that by expansions, w.p.a.1

$$\begin{aligned} \|\hat{\rho}_i - \rho_i\| &\leq C\delta_{1i}\zeta(J)\|\bar{\beta} - \beta_0\|, \\ \|A^*\{\hat{\rho}_i - \rho_i - \rho_{\beta i}(\bar{\beta} - \beta_0)\}\| &\leq \|A^*\|\delta_{2i}\|p_i\|\|\bar{\beta} - \beta_0\|^2 \leq C\delta_{2i}\zeta(J)\|\bar{\beta} - \beta_0\|^2. \end{aligned}$$

Therefore, it follows that

$$\begin{aligned} \|A^*(\hat{\Omega} - \tilde{\Omega})\| &= \|\sum_i A^*(\hat{\rho}_i\hat{\rho}_i' - \rho_i\rho_i')/n\| \\ &\leq \sum_i\{\|A^*\|\|\hat{\rho}_i - \rho_i\|^2 + \|A^*\rho_i\|\|\hat{\rho}_i - \rho_i\| \\ &\quad +\|A^*(\hat{\rho}_i - \rho_i)\|\|\rho_i\|\}/n \\ &\leq C\sum_i\{\delta_{1i}^2\zeta(J)^2\|\bar{\beta} - \beta_0\|^2 + \|A^*\rho_i\|\delta_{1i}\zeta(J)\|\bar{\beta} - \beta_0\| \\ &\quad +\delta_{2i}\zeta(J)^2\|u_i\|\|\bar{\beta} - \beta_0\|^2 + \|A^*\rho_{\beta i}\|\|u_i\|\zeta(J)\|\bar{\beta} - \beta_0\|\}/n \\ &= O_p(\zeta(J)^2/n + E[\|A^*\rho_i\|\delta_{1i} + \|A^*\rho_{\beta i}\|\|u_i\|]n^{-1/2}\zeta(J)^{1/2}) = o_p(J^{-1/2}). \end{aligned}$$

Next, note that by $E[\|u_i\|^4|w_i]$ bounded, for

$$\begin{aligned} E[\|A^*(\tilde{\Omega} - \Omega)\|^2] &= tr\{E[\|\rho_i\|^2A^*\rho_i\rho_i'A^{*\prime}] - A^*\Omega^2A^*\}/n \\ &\leq C\zeta(J)^2tr\{A^*E[\|u_i\|^2\rho_i\rho_i']A^{*\prime}\}/n \\ &\leq C\zeta(J)^2tr\{A^*E[I \otimes p_ip_i']A^{*\prime}\}/n = O(\zeta(J)^2/n), \end{aligned}$$

so by the Markov inequality, $\|A^*(\tilde{\Omega} - \Omega)\| = O_p(n^{-1/2}\zeta(J)^{1/2}) = o_p(J^{-1/2})$. Then by the triangle inequality, condition iv) of Theorem 6 is satisfied. Furthermore, the last condition of Theorem 6 holds by since $\rho(z,\beta) = \hat{\rho}(z,\beta)$, so the conclusion of Theorem 7 follows from Theorem 6. Q.E.D.

8 Appendix B:Additional Derivations

The first result shows the formula for the first-order conditions of the ordered choice MLE in the transformation model. Let $Y_{ij} = 1(\bar{y}_{j-1} \leq y_i < \bar{y}_j)$ and $P_{ij}(\theta) = P_j(x_i, \theta)$.

Differentiating the log-likelihood gives first-order conditions

$$\begin{aligned} 0 &= \sum_{i=1}^{n}\sum_{j=1}^{J+1} Y_{ij}\partial lnP_{ij}(\hat{\theta})/\partial\theta = \sum_{i=1}^{n}\sum_{j=1}^{J+1}[Y_{ij} - P_{ij}(\hat{\theta})]\partial lnP_{ij}(\hat{\theta})/\partial\theta \\ &= \sum_{i=1}^{n}\sum_{j=1}^{J}[Y_{ij} - P_{ij}(\hat{\theta})]\partial ln[P_{ij}(\hat{\theta})/P_{i,J+1}(\hat{\theta})]/\partial\theta = \sum_{i=1}^{n}\sum_{j=1}^{J}\rho_j(z_i,\hat{\theta})\partial ln[P_{ij}(\hat{\theta})/P_{i,j+1}(\hat{\theta})]/\partial\theta, \end{aligned}$$

where the second and third equalities follow by $\sum_{j=1}^{J+1} P_{ij}(\theta) = 1$ identically in θ and the fourth equality by $Y_{ij} - P_{ij}(\theta) = \rho_j(z,\theta) - \rho_{j-1}(z,\theta)$, $j = 2, ..., J$, and $Y_{i1} - P_{i1}(\theta) = \rho_1(z,\theta)$.

The next result shows that a sequence of positive semi-definite (p.s.d.) matrices that is monotonically decreasing in the p.s.d. semi-order has a limit.

Lemma A8: *If Σ_J is positive semi-definite and $\Sigma_J \geq \Sigma_{J+1}$ for each J then $lim_{J\to\infty}\Sigma_J$ exists.*

Proof: Let $tr(M)$ denote the trace of a square matrix M. By $\Sigma_J - \Sigma_{J+1}$ p.s.d, $tr(\Sigma_J) - tr(\Sigma_{J+1}) = tr(\Sigma_J - \Sigma_{J+1}) \geq 0$, so $tr(\Sigma_J)$ is a nonnegative, monotonic decreasing sequence that converges. Therefore, $tr(\Sigma_J)$ is a Cauchy sequence, implying $tr(\Sigma_J - \Sigma_K) \to 0$ as $J \to \infty$ and $K \to \infty$, $K \geq J$. Let $\|M\| = \sqrt{tr(M'M)}$. For M p.s.d., $M = B'\Lambda B$ for an orthonormal matrix B and a diagonal matrix of nonnegative eigenvalues Λ, so that by $\Lambda_{jj} \geq 0$

$$\|M\|^2 = tr(B'\Lambda BB'\Lambda B) = tr(B'\Lambda^2 B) = tr(\Lambda^2) = \sum_j \Lambda_{jj}^2 \leq tr(\Lambda)^2 = tr(M)^2.$$

Applying this equation with $M = \Sigma_J - \Sigma_K$, we obtain $\|\Sigma_J - \Sigma_K\|^2 \leq tr(\Sigma_J - \Sigma_K)^2$. Because $\|\Sigma_J - \Sigma_K\|$ is just the usual Euclidean norm, it follows that each element of Σ_J is a Cauchy sequence, and hence converges. Q.E.D.

References

[1] Ahn, H. and J.L. Powell (1993): "Semiparametric Estimation of Censored Selection Models," *Journal of Econometrics* 58, 3-29.

[2] Andrews, D.W.K. (1991): "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Models," *Econometrica* 59, 307-345.

[3] Bickel, P.J., C.A.J. Klaassen, Y. Ritov, and J.A. Wellner (1993): *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore: Johns Hopkins University Press.

[4] Chamberlain, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics* 34, 305-334.

[5] Cosslett, S.R. (1987): "Efficiency Bounds for Distribution Free Estimators of the Binary Choice and Censored Regression Models," *Econometrica* 55, 559-585.

[6] Donald, S.G., G. Imbens, and W.K. Newey (2002): "Choosing the Number of Moment Conditions with Conditional Moment Models," working paper, University of Texas.

[7] Donald, S.G. and W.K. Newey (2001): "Choosing the Number of Instruments," *Econometrica 69,* 1161-1191.

[8] Hall, P. and J.L. Horowitz (1990): "Bandwidth Selection in Semiparametric Estimation of Censored Linear Regression," *Econometric Theory* 6, 123-150.

[9] Hahn, J. (1997): "Efficient Estimation of Panel Data Models with Sequential Moment Restrictions, J*ournal of Econometrics* 79, 1-21.

[10] Han, A. and J.A. Hausman (1990): "Flexible Parametric Estimation of Duration and Competing Risks Models," *Journal of Applied Econometrics* 5, 1-28.

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

[12] Hansen, L.P. and T.J. Sargent (1991): *Rational Expectations Econometrics*, Westview Press, San Francisco.

[13] Heckman, J.J (1979): "Sample Selection Bias as a Specification Error," *Econometrica* 47, 153-161.

[14] Ichimura, H. (1993): "Estimation of Single Index Models," *Journal of Econometrics* 58, 71-120.

[15] Koenker, R. and J.A.F. Machado (1999): "GMM Inference When the Number of Moment Conditions is Large," *Journal of Econometrics 93, 327-344.*

[16] Mroz, T. (1987), "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica, 55, 765-799.*

[17] Newey, W.K. (1987): "Interval Moment Estimation of the Truncated Regression Model," mimeo, Department of Economics, MIT, presented at the 1987 Summer Meeting of the Econometric Society.

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

[19] Newey, W.K. (1993): "Efficient Estimation of Models with Conditional Moment Restrictions," in G.S. Maddala, C.R. Rao, and H.D. Vinod, eds., *Handbook of Statistics, Volume 11: Econometrics.* Amsterdam: North-Holland.

[20] Newey, W.K. (1997):"Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics* 79, 147-168.

[21] Newey, W.K. (2001): "Conditional Moment Restrictions in Censored and Truncated Regression Models," *Econometric Theory* 17, 863-888..

[22] Newey, W.K., J.L. Powell, and J.R. Walker (1990): "Semiparametric Estimation of Selection Models: Some Empirical Results," *American Economic Review, Papers and Proceedings* 80, 324-328.

[23] Newey, W.K. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in Engle, R.F. and D.L. McFadden, eds., *Handbook of Econometrics, Volume IV*, Chapter 36.

[24] Newey, W.K. and J.L. Powell (1993): "Efficiency Bounds for Semiparametric Selection Models," *Journal of Econometrics* 58, 169-184.

[25] Newey, W.K. and J.L. Powell (1999): "Two-step Estimation, Optimal Moment Conditions, and Sample Selection Models," MIT Working Paper 99-06.

[26] Powell, J.L. (1986): "Censored Regression Quantiles," *Journal of Econometrics* 32, 143-155.

[27] Powell, J.L. (1994): "Estimation of Semiparametric Models," in Engle, R.F. and D.L. McFadden, eds., *Handbook of Econometrics, Volume IV*, Chapter 41.

[28] Pratt, J.W. (1981): "Concavity of the Log-Likelihood," *Journal of the American Statistical Association 76, 103-106.*

[29] Ritov, Y. (1990): "Estimation in a Linear Regression Model with Censored Data," *Annals of Statistics* 18, 303-328.