THE LIMITING DISTRIBUTION OF THE MAXIMUM RANK CORRELATION ESTIMATOR

By Robert P. Sherman[1]

Abstract

Han's maximum rank correlation (MRC) estimator is shown to be $\sqrt{n}$-consistent and asymptotically normal. The proof rests on a general method for determining the asymptotic distribution of a maximization estimator, a simple U-statistic decomposition, and a uniform bound for degenerate U-processes. A consistent estimator of the asymptotic covariance matrix is provided, along with a result giving the explicit form of this matrix for any model within the scope of the MRC estimator. The latter result is applied to the binary choice model, and it is found that the MRC estimator does not achieve the semiparametric efficiency bound.

KEYWORDS: Generalized regression model, maximum rank correlation estimator, discontinuous criterion function, general method, U-statistic decomposition, uniform bound, empirical process, degenerate U-process, VC class, Euclidean class, numerical derivatives, semiparametric efficiency bound.

1. Introduction

Let $Z = (Y, X)$ be an observation from a distribution P on a set $S \subseteq \mathbb{R} \otimes \mathbb{R}^d$, where Y is a response variable and X is a vector of regressors. Han (1987) introduced the generalized regression model

$$Y = D \circ F(X'\beta_0, \epsilon) \tag{1}$$

where β_0 is a d-dimensional vector of unknown parameters, ϵ is a random variable independent of X, F is a strictly increasing function of each of its arguments, and D is a monotone increasing function of its argument. He showed that many interesting regression models fit into this framework. For example, take $F(u, v) = u + v$. If $D(w) = w$, the model above reduces to a standard linear regression model; for $D(w) = \{w \geq 0\}$, a binary choice model; for $D(w) = w\{w \geq 0\}$, a censored regression model. Transformation and duration models are other special cases.

[1]I wish to thank David Pollard for offering useful comments and suggestions. I also thank Chris Cavanagh and Roger Klein for useful discussions. In addition, Chris gave me the core of Theorem 4 appearing in Section 6. Finally, I am indebted to an editor and two dedicated referees for corrections and constructive suggestions for reshaping preliminary drafts. One particularly diligent referee derived a number of results appearing in Section 6.

Let $Z_1, \ldots, Z_n$ be a sample of independent observations from P. Han proposed estimating β_0 with $\beta_n = \operatorname{argmax}_{\boldsymbol{R}^d} G_n(\beta)$ where

$$G_n(\beta) = \frac{1}{n(n-1)} \sum_{i \neq j} \{Y_i > Y_j\}\{X_i'\beta > X_j'\beta\}\,. \tag{2}$$

He called β_n the maximum rank correlation (MRC) estimator of β_0.

A simple principle motivates the estimator : the monotonicity of $D \circ F$ and the independence of the u_i's and the X_i's ensure that

$$Pr(Y_i \geq Y_j \mid X_i, X_j) \geq Pr(Y_i \leq Y_j \mid X_i, X_j) \quad \text{whenever} \quad X_i'\beta_0 \geq X_j'\beta_0\,.$$

In other words, it is more likely than not that $Y_i \geq Y_j$ whenever $X_i'\beta_0 \geq X_j'\beta_0$. Ties turn out to be irrelevant, and terms for which $i = j$ make a negligible asymptotic contribution. As a result, these terms are discarded from the criterion function in (2).

Han proved strong consistency of β_n, and in Han (1988), offered an argument for $\sqrt{n}$-consistency and asymptotic normality based on a decomposition of $G_n(\beta)$ into a sum of independent, identically distributed random variables plus a remainder term satisfying a degeneracy property. He presented a more complicated, bracketing-type argument in an attempt to show that the degenerate term could be neglected. He also proposed a kernel-type estimator of the asymptotic covariance matrix.

The chief difficulty in determining the limiting distribution of β_n is the discontinuous nature of $G_n(\beta)$. Standard asymptotic methods requiring smoothness of the criterion function do not apply. The other distinctive feature of the problem is the U-process nature of $G_n(\beta)$. For each β in $\mathbb{R}^d$, $G_n(\beta)$ is a U-statistic of order two. The collection $\{G_n(\beta) : \beta \in \mathbb{R}^d\}$ is a U-process of order two, and β_n is said to maximize this U-process.

In this paper, we offer a simple proof that the MRC estimator is $\sqrt{n}$-consistent and asymptotically normal. The proof relies on a general method for establishing the limiting distribution of a maximization estimator. This method has its origins in a paper by Huber (1967), and requires neither differentiability nor continuity of the criterion function. The proof also makes use of a simple U-statistic decomposition alluded to above, and a uniform bound for degenerate U-processes established by Sherman (1991). The latter bound is used to make short work of the degenerate term that Han attacked through bracketing. A mild regularity condition on a class of functions, called a Euclidean condition, is required to apply the bound. Existing tools from the empirical process literature make it easy to verify this condition.

In the next section, the general method mentioned above is presented. Section 3 presents the U-statistic decomposition along with the uniform bound for degenerate U-processes. The limiting distribution of the MRC estimator is established in Section 4, and the Euclidean condition mentioned above is verified in Section 5. Section 6 provides a derivation of the the general form of the asymptotic covariance matrix in terms of the model primitives. This result is applied to the binary choice model, and it is found that the MRC estimator does not achieve the semiparametric efficiency bound established by Chamberlain (1986) and Cosslett (1987). Section 7 provides a consistent estimator of the asymptotic covariance matrix based on numerical derivatives, and Section 8 discusses sufficient conditions for applying the distributional result established in Section 4.

2. A General Method

Let Θ be a subset of $\mathbb{R}^m$, and θ_0 an element of Θ and a parameter of interest. Suppose θ_0 maximizes a function $\Gamma(\theta)$ defined on Θ. Suppose further that a sample analogue, $\Gamma_n(\theta)$, is maximized at a point θ_n that converges in probability to θ_0.

In this section, we present a general method for establishing that θ_n is $\sqrt{n}$-consistent for θ_0 and asymptotically normally distributed. This method has its origins in a paper by Huber (1967), and has been recast into the form presented here (apart from minor modifications) by Pollard (1989). The method is embodied in the two theorems that follow. The first of these provides conditions under which θ_n is $\sqrt{n}$-consistent for θ_0. The second theorem gives conditions under which a $\sqrt{n}$-consistent estimator is also asymptotically normally distributed. The proofs of these theorems are implicit in Pollard (1985), for example. Nonetheless, for completeness, we also provide proofs here.

A brief word on notation. We will be needing uniform bounds on functions of θ within shrinking neighborhoods of θ_0. A convenient notation will be "$H_n(\theta) = o_p(1/n)$ uniformly over $o_p(1)$ neighborhoods of θ_0". This means that for each sequence of random variables $\{r_n\}$ of order $o_p(1)$ there exists a sequence of random variables $\{b_n\}$ of order $o_p(1)$ such that

$$\sup_{|\theta-\theta_0|\leq r_n} |H_n(\theta)| \leq b_n/n\,.$$

Also, for simplicity, we will assume that θ_0 is the zero vector (denoted $\mathbf{0}$) in $\mathbb{R}^m$, and that $\Gamma_n(\theta_0) = \Gamma(\theta_0) = 0$. This can always be arranged by working with $\Gamma_n(\theta_0+t)-\Gamma_n(\theta_0)$ instead of $\Gamma_n(\theta)$, $\Gamma(\theta_0+t)-\Gamma(\theta_0)$ instead of $\Gamma(\theta)$, and substituting t for θ, where t satisfies $\theta_0+t\in\Theta$.

Theorem 1: *Let θ_n be a maximizer of $\Gamma_n(\theta)$, and $\mathbf{0}$ a maximizer of $\Gamma(\theta)$. Suppose θ_n converges in probability to $\mathbf{0}$, and also that*

(i) *there exists a neighborhood $\mathcal{N}$ of $\mathbf{0}$ and a constant $\kappa > 0$ for which*

$$\Gamma(\theta) \leq -\kappa|\theta|^2$$

for all θ in $\mathcal{N}$;

(ii) *uniformly over $o_p(1)$ neighborhoods of* $\mathbf{0}$*,*

$$\Gamma_n(\theta) = \Gamma(\theta) + O_p(|\theta|/\sqrt{n}) + o_p(|\theta|^2) + O_p(1/n)\,.$$

Then

$$|\theta_n| = O_p(1/\sqrt{n})\,.$$

PROOF. Since θ_n maximizes $\Gamma_n(\theta)$,

$$0 \le \Gamma_n(\theta_n)\,. \tag{3}$$

Since θ_n is consistent, it lies within at least one of the sequences of neighborhoods described in (i) and (ii) with probability tending to one as n tends to infinity. When this happens, (i) and (ii) hold for $\theta = \theta_n$. Deduce from (3), (ii), and (i) that

$$0 \le -\kappa|\theta_n|^2 + o_p(|\theta_n|^2) + O_p(|\theta_n|/\sqrt{n}) + O_p(1/n)\,.$$

With probability tending to one, the $o_p(|\theta_n|^2)$ term is bounded in absolute value by $\frac{1}{2}\kappa|\theta_n|^2$. Write $W_n|\theta_n|$ for the $O_p(|\theta_n|/\sqrt{n})$ term, where $\{W_n\}$ is a sequence of random variables of order $O_p(1/\sqrt{n})$. Then, absorbing all things that happen with probability tending to zero into the $O_p(1/n)$ term, we get

$$\frac{1}{2}\kappa|\theta_n|^2 - W_n|\theta_n| \le O_p(1/n)\,.$$

Complete the square in $|\theta_n|$, to rewrite this as

$$\frac{1}{2}\kappa\left(|\theta_n| - W_n/\kappa\right)^2 \le O_p(1/n) + \frac{1}{2}W_n^2/\kappa = O_p(1/n)\,.$$

Take square roots, then rearrange to get

$$|\theta_n| \le W_n/\kappa + O_p(1/\sqrt{n}) = O_p(1/\sqrt{n})\,.$$

□

Once $\sqrt{n}$-consistency of θ_n is established, we can prove asymptotic normality provided there exist very good quadratic approximations to $\Gamma_n(\theta)$ within $O_p(1/\sqrt{n})$ neighborhoods of $\mathbf{0}$. In the following theorem, the symbol $\Longrightarrow$ denotes convergence in distribution.

THEOREM 2: *Suppose θ_n is $\sqrt{n}$-consistent for* $\mathbf{0}$*, an interior point of Θ. Suppose also that uniformly over $O_p(1/\sqrt{n})$ neighborhoods of* $\mathbf{0}$*,*

$$\Gamma_n(\theta) = \frac{1}{2}\theta' V\theta + \frac{1}{\sqrt{n}}\theta' W_n + o_p(1/n) \tag{4}$$

where V is a negative definite matrix, and W_n converges in distribution to a $N(\mathbf{0}, \Delta)$ random vector. Then

$$\sqrt{n}\theta_n \Longrightarrow N(\mathbf{0}, V^{-1}\Delta V^{-1}) .$$

PROOF. Write t_n for $\sqrt{n}\theta_n$ and t_n^* for $-V^{-1}W_n$. Notice that $t_n^*/\sqrt{n}$ maximizes the quadratic approximation to $\Gamma_n(\theta)$ given in (4). Also, t_n^* converges in distribution to a $N(\mathbf{0}, V^{-1}\Delta V^{-1})$ random variable. We will show that $t_n = t_n^* + o_p(1)$.

Because $\mathbf{0}$ is an interior point of Θ, the point $t_n^*/\sqrt{n}$ lies in Θ with probability tending to one. When this happens, by definition of t_n,

$$\Gamma_n(t_n^*/\sqrt{n}) \leq \Gamma_n(t_n/\sqrt{n}) .$$

Apply (4) twice in the last expression, then multiply through by n, consolidate terms, and use the fact that V is negative definite to get

$$0 \leq -\frac{1}{2}(t_n - t_n^*)'V(t_n - t_n^*) \leq o_p(1) .$$

The $o_p(1)$ term can be assumed to absorb the bad cases where $t_n^*/\sqrt{n}$ does not lie in Θ. The last inequality is true without restriction, and implies that $t_n = t_n^* + o_p(1)$. □

The conditions of the theorems do not require that θ_n be a zero of the gradient of $\Gamma_n(\theta)$. Nor do they require that $\Gamma_n(\theta)$ be a continuous function of θ. Thus, this approach provides a framework within which the asymptotic distribution of the MRC estimator can be established.

3. A U-statistic Decomposition and a Uniform Bound

In this section, we present a U-statistic decomposition and a uniform bound that are used in tandem with the general method of the last section to establish the asymptotic distribution of the MRC estimator.

Let $Z_1, \ldots, Z_n$ be independent, identically distributed (iid) random vectors with distribution P on a set S. Let Θ be a subset of $\mathbb{R}^m$, and for each θ in Θ suppose $f(\cdot, \cdot, \theta)$ is a real-valued function on the product space $S \otimes S$. Define

$$U_n f(\cdot, \cdot, \theta) = \frac{1}{n(n-1)} \sum_{i \neq j} f(Z_i, Z_j, \theta) .$$

For each θ in Θ, $U_n f(\cdot, \cdot, \theta)$ is a U-statistic of order two. (See Chapter 5 of Serfling (1980) for more on U-statistics.) The collection $\{U_n f(\cdot, \cdot, \theta) : \theta \in \Theta\}$ is called a U-process of order two.

By analogy with the empirical measure P_n that places mass $1/n$ on each Z_i, U_n can be viewed as a random measure putting mass $1/[n(n-1)]$ on each ordered pair (Z_i, Z_j). Let Q denote the product measure $P \otimes P$. Then

$$U_n f(\cdot,\cdot,\theta) = Qf(\cdot,\cdot,\theta) + P_n g(\cdot,\theta) + U_n h(\cdot,\cdot,\theta) \tag{5}$$

where, for each u, v in S and each θ in Θ,

$$g(u,\theta) = Pf(u,\cdot,\theta) + Pf(\cdot,u,\theta) - 2Qf(\cdot,\cdot,\theta)$$

and

$$h(u,v,\theta) = f(u,v,\theta) - Pf(u,\cdot,\theta) - Pf(\cdot,v,\theta) + Qf(\cdot,\cdot,\theta)\,.$$

Linear functional notation is used for expectations. $Qf(\cdot,\cdot,\theta)$ denotes the unconditional expectation of $f(u,v,\theta)$, while $Pf(u,\cdot,\theta)$ denotes the conditional expectation of $f(u,v,\theta)$ given its first argument, and $Pf(\cdot,v,\theta)$ the conditional expectation of $f(u,v,\theta)$ given its second argument.

Notice that for each θ in Θ, $P_n g(\cdot,\theta)$ is an average of zero-mean, iid random variables. The collection $\{P_n g(\cdot,\theta) : \theta \in \Theta\}$ is called a zero-mean empirical process. Also note that

$$Ph(u,\cdot,\theta) \equiv Ph(\cdot,v,\theta) \equiv 0\,. \tag{6}$$

Because of (6), the function $h(u,v,\theta)$ is said to be P-degenerate on $S \otimes S$ and $U_n h(\cdot,\cdot,\theta)$ is called a degenerate U-statistic of order two. The collection $\{h(\cdot,\cdot,\theta) : \theta \in \Theta\}$ is said to be a P-degenerate class of functions on $S \otimes S$ and $\{U_n h(\cdot,\cdot,\theta) : \theta \in \Theta\}$ is called a degenerate U-process of order two.

In the next section, the MRC estimator is shown to be $\sqrt{n}$-consistent and asymptotically normal. The decomposition in (5) is applied to write the criterion function as a sum of its expected value, plus a smoothly parametrized, zero-mean empirical process, plus a degenerate U-process of order two. The result is obtained by handling the first two terms using standard Taylor expansion arguments, and then showing that the degenerate term has order $o_p(1/n)$ uniformly over $o_p(1)$ neighborhoods of the parameter of interest. The following theorem is used to establish the uniformity result. We assume familiarity with the notions of a Euclidean class of functions and an envelope for a class of functions, as defined in Section 2 of Pakes and Pollard (1989).

THEOREM 3: *Let $\mathcal{F} = \{f(\cdot,\cdot,\theta) : \theta \in \Theta\}$ be a class of P-degenerate functions on $S \otimes S$. Let Q denote the product measure $P \otimes P$. Suppose there exists a point θ_0 in Θ for which $f(\cdot,\cdot,\theta_0) \equiv 0$. If*

(i) *$\mathcal{F}$ is Euclidean for a constant envelope,*

(ii) $Qf(\cdot,\cdot,\theta)^2 \to 0 \quad \text{as} \quad \theta \to \theta_0$, *then*

$$U_n f(\cdot, \cdot, \theta) = o_p(1/n)$$

uniformly over $o_p(1)$ *neighborhoods of* θ_0.

Theorem 3 is a special case of Corollary 8 in Section 6 of Sherman (1991). Pakes and Pollard provide simple criteria for determining the Euclidean property. Nolan and Pollard (1987) give complementary criteria.

4. $\sqrt{n}$-Consistency and Asymptotic Normality

In this section, the MRC estimator is shown to be $\sqrt{n}$-consistent and asymptotically normally distributed. We begin by restricting the parameter space, reparametrizing, and then introducing some convenient notation and definitions.

Notice from (2) that $G_n(\beta) = G_n(c\beta)$ for any $c > 0$. Consequently, if β_n maximizes $G_n(\beta)$, then so does $c\beta_n$, for any $c > 0$. This leads to an identifiability problem, which can be circumvented by restricting the parameter space to a $d-1$ dimensional subset of $\mathbb{R}^d$. For convenience, we take the restricted parameter space, denoted $\mathcal{B}$, to be a compact subset of $\{\beta \in \mathbb{R}^d : \beta_d = 1\}$. Rather than introduce new notation, rechristen β_n as the maximizer of $G_n(\beta)$ over $\mathcal{B}$. Also, let $G(\beta)$ denote the expected value of $G_n(\beta)$. That is,

$$G(\beta) = \mathbb{E}\{Y_1 > Y_2\}\{X_1'\beta > X_2'\beta\}\,.$$

Represent each β in $\mathcal{B}$ as $\beta(\theta) = (\theta, 1)$ where θ is an element of Θ, a compact subset of $\mathbb{R}^{d-1}$. Also, write $\beta_0 = \beta(\theta_0)$ where θ_0 consists of the first $d-1$ components of β_0. Similarly, θ_n denotes the first $d-1$ components of β_n.

For each θ in Θ, write $\Gamma(\theta)$ for $G(\beta(\theta)) - G(\beta(\theta_0))$. Similarly, write $\Gamma_n(\theta)$ for $G_n(\beta(\theta)) - G_n(\beta(\theta_0))$. Note that θ_n maximizes $\Gamma_n(\theta)$ over Θ. As in Section 2, we shall assume that $\theta_0 = \mathbf{0}$, the zero vector in $\mathbb{R}^{d-1}$. Thus, $\Gamma_n(\mathbf{0}) = \Gamma(\mathbf{0}) = 0$.

In the course of proving consistency of β_n, Han showed that β_0 maximizes $G(\beta)$. It follows immediately that $\Gamma(\theta)$ is maximized at $\mathbf{0}$. Also, consistency of β_n immediately implies consistency of θ_n for $\mathbf{0}$.

Recall that $Z = (Y, X)$ denotes an observation from the distribution P on the set $S \subseteq \mathbb{R} \otimes \mathbb{R}^d$, and that $Z_1, \ldots, Z_n$ denotes a sample of independent observations from P. For each z in S and for each θ in Θ, define

$$\tau(z, \theta) = \mathbb{E}\{y > Y\}\{x'\beta(\theta) > X'\beta(\theta)\} + \mathbb{E}\{Y > y\}\{X'\beta(\theta) > x'\beta(\theta)\}\,.$$

The function $\tau(\cdot, \theta)$ will be the kernel of the empirical process that drives the asymptotic behavior of θ_n. Notice that $\mathbb{E}[\tau(\cdot, \theta) - \tau(\cdot, \mathbf{0})] = 2\Gamma(\theta)$.

Write ∇_m for the mth partial derivative operator with respect to θ, and

$$|\nabla_m|\sigma(\theta) \equiv \sum_{i_1,\ldots,i_m} \left| \frac{\partial^m}{\partial\theta_{i_1} \cdots \partial\theta_{i_m}} \sigma(\theta) \right| .$$

The symbol $\|\cdot\|$ denotes the matrix norm: $\|(a_{ij})\| = (\sum_{i,j} a_{ij}^2)^{1/2}$.

We now state the assumptions used in the normality proof.

A1. The element $\mathbf{0}$ is an interior point of Θ, a compact subset of $\mathbb{R}^{d-1}$.
A2. The random variables X and ϵ in (1) are independent.
A3. Let S_x denote the support of the vector of regressors X.

(i) S_x is not contained in any proper linear subspace of $\mathbb{R}^d$.

(ii) The dth component of X has an everywhere positive Lebesgue density, conditional on the other components.

A4. Let $\mathcal{N}$ denote a neighborhood of $\mathbf{0}$.

(i) For each z in S, all mixed second partial derivatives of $\tau(z,\cdot)$ exist on $\mathcal{N}$.

(ii) There is an integrable function $M(z)$ such that for all z in S and θ in $\mathcal{N}$

$$\|\nabla_2\tau(z,\theta) - \nabla_2\tau(z,\mathbf{0})\| \leq M(z)|\theta|\,.$$

(iii) $\mathbb{E}|\nabla_1\tau(\cdot,\mathbf{0})|^2 < \infty$.

(iv) $\mathbb{E}|\nabla_2|\tau(\cdot,\mathbf{0}) < \infty$.

(v) The matrix $\mathbb{E}\nabla_2\tau(\cdot,\mathbf{0})$ is negative definite.

The compactness of Θ, A2, and A3 were used by Han to establish the consistency of the MRC estimator. The conditions of A4 are standard regularity conditions sufficient to support an argument based on a Taylor expansion of $\tau(z,\cdot)$ about $\mathbf{0}$. We defer to Section 8 a discussion of sufficient conditions on the random variable Z for satisfying A4.

THEOREM 4: *If A1 through A4 hold, then*

$$\sqrt{n}\theta_n \Longrightarrow N(\mathbf{0}, V^{-1}\Delta V^{-1})$$

where $2V = \mathbb{E}\nabla_2\tau(\cdot,\mathbf{0})$ *and* $\Delta = \mathbb{E}\nabla_1\tau(\cdot,\mathbf{0})[\nabla_1\tau(\cdot,\mathbf{0})]'$.

PROOF. We will show that

$$\Gamma_n(\theta) = \frac{1}{2}\theta' V\theta + \frac{1}{\sqrt{n}}\theta' W_n + o_p(|\theta|^2) + o_p(1/n) \tag{7}$$

uniformly in $o_p(1)$ neighborhoods of $\mathbf{0}$, where W_n converges in distribution to a $N(\mathbf{0},\Delta)$ random vector. Since V is, by A4(v), a negative definite matrix, it will follow from (7) and Theorem 1 that

$$|\theta_n| = O_p(1/\sqrt{n})\,. \tag{8}$$

The result will then follow from (7), (8), and Theorem 2.

For each (z_1, z_2) in $S \otimes S$ and each θ in Θ, define

$$f(z_1, z_2, \theta) = \{y_1 > y_2\} \Big[\{x_1'\beta(\theta) > x_2'\beta(\theta)\} - \{x_1'\beta(\mathbf{0}) > x_2'\beta(\mathbf{0})\} \Big] .$$

Since $\Gamma_n(\theta)$ is a U-statistic of order two with expectation $\Gamma(\theta)$, we may apply the decomposition in (5) to write

$$\Gamma_n(\theta) = \Gamma(\theta) + P_n g(\cdot, \theta) + U_n h(\cdot, \cdot, \theta)$$

where

$$g(z, \theta) = Pf(z, \cdot, \theta) + Pf(\cdot, z, \theta) - 2\Gamma(\theta)$$

and

$$h(z_1, z_2, \theta) = f(z_1, z_2, \theta) - Pf(z_1, \cdot, \theta) - Pf(\cdot, z_2, \theta) + \Gamma(\theta) .$$

First, we show that

$$\Gamma(\theta) = \frac{1}{2}\theta' V \theta + o(|\theta|^2) \quad \text{as} \quad \theta \to \mathbf{0} . \tag{9}$$

Fix z in S and θ in $\mathcal{N}$. Invoke A4(i) and expand $\tau(z, \theta)$ about $\mathbf{0}$ to get

$$\tau(z, \theta) = \tau(z, \mathbf{0}) + \theta' \nabla_1 \tau(z, \mathbf{0}) + \frac{1}{2}\theta' \nabla_2 \tau(z, \theta^*)\theta \tag{10}$$

for θ^* between θ and $\mathbf{0}$. By A4(ii), for each z in S and each θ in $\mathcal{N}$

$$\|\theta'[\nabla_2 \tau(z, \theta) - \nabla_2 \tau(z, \mathbf{0})]\theta\| \le M(z)|\theta|^3 . \tag{11}$$

Take expectations in (10) and apply (11) and the integrability of M to get that

$$2\Gamma(\theta) = \theta' I\!E \nabla_1 \tau(\cdot, \mathbf{0}) + \theta' V \theta + o(|\theta|^2) \quad \text{as} \quad \theta \to \mathbf{0} .$$

Since $\Gamma(\theta)$ is maximized at $\mathbf{0}$, the coefficient of the linear term in the last expression must be the zero vector. Divide through by 2 to establish (9).

Next, we show that

$$P_n g(\cdot, \theta) = \frac{1}{\sqrt{n}}\theta' W_n + o(|\theta|^2) \tag{12}$$

uniformly over $o_p(1)$ neighborhoods of $\mathbf{0}$, where W_n converges in distribution to a $N(\mathbf{0}, \Delta)$ random vector.

Note that

$$g(z, \theta) = \tau(z, \theta) - \tau(z, \mathbf{0}) - 2\Gamma(\theta) .$$

Apply (9), (10), and (11) to see that

$$P_n g(\cdot, \theta) = \frac{1}{\sqrt{n}}\theta' W_n + \frac{1}{2}\theta' D_n \theta + o(|\theta|^2) + R_n(\theta) \tag{13}$$

uniformly over $o_p(1)$ neighborhoods of $\mathbf{0}$, where

$$\begin{aligned} W_n &= \sqrt{n} P_n \nabla_1 \tau(\cdot, \mathbf{0}) \,, \\ D_n &= P_n \nabla_2 \tau(\cdot, \mathbf{0}) - 2V \,, \end{aligned}$$

and

$$|R_n(\theta)| \leq |\theta|^3 P_n M(\cdot) \,.$$

Deduce from A4(iii) and the fact that $\mathbb{E}\nabla_1 \tau(\cdot, \mathbf{0}) = \mathbf{0}$ that W_n converges in distribution to a $N(\mathbf{0}, \Delta)$ random vector. By A4(iv) and a weak law of large numbers, D_n converges in probability to zero as n tends to infinity. Finally, deduce from the integrability of M and a weak law of large numbers that

$$R_n(\theta) = o_p(|\theta|^2)$$

uniformly over $o_p(1)$ neighborhoods of $\mathbf{0}$. This establishes (12).

In order to establish (7), it remains to show that

$$U_n h(\cdot, \cdot, \theta) = o_p(1/n) \tag{14}$$

uniformly over $o_p(1)$ neighborhoods of $\mathbf{0}$. Theorem 3 will do the job.

We show in the next section that the class of functions $\{h(\cdot, \cdot, \theta) : \theta \in \Theta\}$ is Euclidean for the constant envelope 1. Equation (14) will follow from Theorem 3 provided

$$Q h(\cdot, \cdot, \theta)^2 \to 0 \quad \text{as} \quad \theta \to \mathbf{0} \tag{15}$$

where Q is the product measure $P \otimes P$.

By A3(ii), the distribution of the random variable $X'\beta(\mathbf{0})$ is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}$. This implies that

$$Q\{x_1'\beta(\mathbf{0}) = x_2'\beta(\mathbf{0})\} = 0 \,.$$

Deduce that $f(z_1, z_2, \cdot)$ is continuous at $\mathbf{0}$ for Q almost all (z_1, z_2). The boundedness of f and a dominated convergence argument imply that the same holds true for $h(z_1, z_2, \cdot)$. Since h is bounded, another dominated convergence argument establishes (15), which, in turn, establishes (14).

Put it all together. Combine (9), (12), and (14) to get (7). This proves the theorem. □

Corollary: *If A1 through A4 hold, then*

$$\sqrt{n}(\beta_n - \beta_0) \Longrightarrow (W, 0)$$

where W has the $N(\mathbf{0}, V^{-1}\Delta V^{-1})$ distribution from Theorem 4.

5. Euclidean Properties

Consider the class of functions

$$\mathcal{F} = \{f(\cdot,\cdot,\theta) : \theta \in \Theta\}$$

where, for each (z_1, z_2) in $S \otimes S$ and each θ in Θ,

$$f(z_1, z_2, \theta) = \{y_1 > y_2\}\{x_1'\beta(\theta) > x_2'\beta(\theta)\}\,. \tag{16}$$

In this section we show that $\mathcal{F}$ is Euclidean for the constant envelope 1. The Euclidean properties of the class of P-degenerate functions $\{h(\cdot,\cdot,\theta) : \theta \in \Theta\}$ encountered in the last section can be deduced from this fact in combination with Corollary 17 and Corollary 21 in Nolan and Pollard (1987). We assume that the reader is familiar with the notions of a VC class of sets and the subgraph of a function, as defined in Section 2 of Pakes and Pollard (1989).

Let t, γ, γ_1, and γ_2 be real numbers. Let δ_1 and δ_2 be vectors in $\mathbb{R}^d$and let $\mathcal{X} = S \otimes S$. For each (z_1, z_2) in $\mathcal{X}$, define

$$g(z_1, z_2, t; \gamma, \gamma_1, \gamma_2, \delta_1, \delta_2) = \gamma t + \gamma_1 y_1 + \gamma_2 y_2 + \delta_1' x_1 + \delta_2' x_2$$

and

$$\mathcal{G} = \{g(\cdot,\cdot,\cdot; \gamma, \gamma_1, \gamma_2, \delta_1, \delta_2) : \gamma, \gamma_1, \gamma_2 \in \mathbb{R} \quad \text{and} \quad \delta_1, \delta_2 \in \mathbb{R}^d\}$$

Notice that $\mathcal{G}$ is a $(2d+3)$-dimensional vector space of real-valued functions on $\mathcal{X} \otimes \mathbb{R}$. By Lemma 2.4 in Pakes and Pollard (1989), the class of sets of the form $\{g \geq r\}$ or $\{g > r\}$ with $g \in \mathcal{G}$ and $r \in \mathbb{R}$ is a VC class. We use this fact to show that the set of subgraphs of functions belonging to $\mathcal{F}$ forms a VC class of sets. The result will then follow from Lemma 2.12 in Pakes and Pollard (1989).

For each θ in Θ,

$$\begin{aligned} \text{subgraph}(f(\cdot,\cdot,\theta)) &= \{(z_1, z_2, t) \in \mathcal{X} \otimes \mathbb{R} : 0 < t < f(z_1, z_2, \theta)\} \\ &= \{y_1 - y_2 > 0\}\{x_1'\beta(\theta) - x_2'\beta(\theta) > 0\}\{t \geq 1\}^c\{t > 0\} \\ &= \{g_1 > 0\}\{g_2 > 0\}\{g_3 \geq 1\}^c\{g_4 > 0\} \end{aligned}$$

for $g_i \in \mathcal{G}$, $i = 1, 2, 3, 4$. The subgraph of $f(\cdot,\cdot,\theta)$ is the intersection of four sets, three of which belong to a polynomial class, and the fourth is the complement of a set belonging to a polynomial class. Deduce from Lemma 2.5 in Pakes and Pollard (1989) that $\{\text{subgraph}(f) : f \in \mathcal{F}\}$ forms a VC class of sets.

6. The Asymptotic Covariance Martrix

In this section, we provide general expressions for V and Δ in Theorem 4 in terms of the models primitives. For ease of notation, we return to the parametrization in terms of β, with the understanding that partial derivatives are taken with respect to the first $d-1$ components of this vector. Consequently, we shall write $\tau(\cdot,\beta)$ for $\tau(\cdot,\theta)$, and so on.

Notice that

$$\tau(Z,\beta) = \int_{x'\beta < X'\beta} S(Y, x'\beta_0)\, G(dx) + \int \rho(Y, x'\beta_0)\, G(dx) \tag{17}$$

where $G(\cdot)$ denotes the probability distribution of X,

$$S(y,t) = \mathbb{E}\left[\{y > Y\} - \{y < Y\} \mid X'\beta_0 = t\right] ,$$

and

$$\rho(y,t) = \mathbb{E}[\{y < Y\} \mid X'\beta_0 = t] .$$

The second integral does not depend on β. For each t, $S(Y,t)$ is bounded, and has a symmetric distribution conditional on $X'\beta_0 = t$. Consequently,

$$\mathbb{E}[S(Y,t) \mid X'\beta_0 = t] = 0 . \tag{18}$$

Let $\mathcal{X}$ denote the first $d-1$ components of the vector of regressors. Let $g_0(\cdot \mid r)$ denote the conditional density of $X'\beta_0$ given $\mathcal{X} = r$, and $g_0(\cdot)$ the marginal density of $X'\beta_0$. Write $\tilde{\mathcal{X}}_0$ for $\mathbb{E}(\mathcal{X} \mid X'\beta_0)$ and $S_2(y,t)$ for $\frac{\partial}{\partial t} S(y,t)$.

THEOREM 4: *If $S_2(y,t)$ and $\frac{\partial}{\partial t} g_0(t)$ exist, and if $\mathbb{E}|X|^2 < \infty$, then*

$$\Delta = \mathbb{E}(\mathcal{X} - \tilde{\mathcal{X}}_0)(\mathcal{X} - \tilde{\mathcal{X}}_0)'\, S(Y, X'\beta_0)^2\, g_0(X'\beta_0)^2 \tag{19}$$

and

$$2V = \mathbb{E}(\mathcal{X} - \tilde{\mathcal{X}}_0)(\mathcal{X} - \tilde{\mathcal{X}}_0)'\, S_2(Y, X'\beta_0)\, g_0(X'\beta_0) . \tag{20}$$

PROOF: Let u_i denote the unit vector in $\mathbb{R}^{d-1}$ with ith component equal to unity. Write ∇^i for the ith component of ∇_1. By definition,

$$\nabla^i \tau(Z,\beta_0) = \lim_{\epsilon \to 0} \epsilon^{-1} [\tau(Z, \beta_0 + \epsilon u_i) - \tau(Z,\beta_0)] .$$

Apply (17) to see that the term in brackets equals

$$\int_{x'\beta_0 < X'\beta_0 + \epsilon(X_i - x_i)} S(Y, x'\beta_0)\, G(dx) - \int_{x'\beta_0 < X'\beta_0} S(Y, x'\beta_0)\, G(dx) .$$

Change variables from $x \equiv (r, x_d)$ to $(r, x'\beta_0)$. The resulting integral equals

$$\int \left[\int_{X'\beta_0}^{X'\beta_0 + \epsilon(X_i - x_i)} S(Y,t)\, g_0(t \mid r)\, dt \right] G_{\mathcal{X}}(dr)$$

where $G_{\mathcal{X}}(\cdot)$ denotes the distribution of $\mathcal{X}$. The inner integral equals

$$\epsilon(X_i - x_i)\, S(Y, X'\beta_0)\, g_0(X'\beta_0 \mid r) + |X_i - x_i| o(|\epsilon|) \quad \text{as} \quad \epsilon \to 0 .$$

Integrate, then apply the moment condition to see that

$$\nabla_1 \tau(Z, \beta_0) = (\mathcal{X} - \tilde{\mathcal{X}}_0)\, S(Y, X'\beta_0)\, g_0(X'\beta_0)\,,$$

from which (19) follows immediately.

Write $\lambda(y,t)$ for $S(y,t)g_0(t)$ and $\lambda_2(y,t)$ for $\frac{\partial}{\partial t}\lambda(y,t)$. Calculations similar to those above show that

$$2V = I\!\!E(\mathcal{X} - \tilde{\mathcal{X}}_0)(\mathcal{X} - \tilde{\mathcal{X}}_0)'\lambda_2(Y, X'\beta_0)\,.$$

Apply (18) to establish (20). □

If Y is continuously distributed conditional on $X'\beta_0 = t$, then $S(Y,t)$ is uniformly distributed on $[-1,1]$ conditional on $X'\beta_0 = t$. In this case,

$$I\!\!E[S(Y,t)^2 \mid X'\beta_0 = t] = 1/3$$

and (19) reduces to

$$\Delta = I\!\!E(\mathcal{X} - \tilde{\mathcal{X}}_0)(\mathcal{X} - \tilde{\mathcal{X}}_0)'\, g_0(X'\beta_0)^2/3\,.$$

Provided S_2 exists, the monotonicity of $D \circ F$ in (1) implies

$$S_2(y,t) \leq 0\,.$$

For simplicity, suppose $S_2 < 0$ and $g_0 > 0$. Since $-2V$ has an interpretation as a covariance matric for W where

$$W = (\mathcal{X} - \tilde{\mathcal{X}}_0)\sqrt{-S_2(Y, X'\beta_0)\, g_0(X'\beta_0)}\,,$$

nonsingularity of V follows from assumption A3.

Finally, consider the binary choice model

$$Y = \{X'\beta_0 + \epsilon \geq 0\}\,.$$

Write $F(\cdot)$ for the cumulative distribution function of $-\epsilon$ and $f(\cdot)$ for the derivative of $F(\cdot)$. Then

$$S(Y,t) = Y - F(t)\,.$$

Apply Theorem 4 to get

$$\Delta = I\!\!E(\mathcal{X} - \tilde{\mathcal{X}}_0)(\mathcal{X} - \tilde{\mathcal{X}}_0)'\, g_0^2(X'\beta_0)\, F(X'\beta_0)[1 - F(X'\beta_0)]$$

and

$$2V = -I\!\!E(\mathcal{X} - \tilde{\mathcal{X}}_0)(\mathcal{X} - \tilde{\mathcal{X}}_0)'\, f(X'\beta_0)\, g_0(X'\beta_0)\,.$$

The semiparametric efficiency bound (Chamberlain (1986), Cosslett (1987)) for this model is the inverse of the matrix

$$I\!\!E(\mathcal{X} - \tilde{\mathcal{X}}_0)(\mathcal{X} - \tilde{\mathcal{X}}_0)'\, f^2(X'\beta_0)/F(X'\beta_0)[1 - F(X'\beta_0)]\,.$$

It is evident that the MRC estimator does not achieve this bound, in general.

7. A Consistent Estimator of the Asymptotic Covariance Matrix

In this section, we construct consistent estimators of the components Δ and V from Theorem 4. We use numerical derivatives as in Pakes and Pollard (1989), and the following uniformity result:

LEMMA: *Let $\mathcal{G}$ be a class of zero-mean functions on a set S. If $\mathcal{G}$ is Euclidean for a constant envelope, then*

$$\sup_{\mathcal{G}} |P_n g| = O_p(1/\sqrt{n})\,.$$

This lemma is a special case of Corollary 7 in Section 6 of Sherman (1991).

Recall the definition of $f(\cdot,\cdot,\theta)$ given in (16). For each z in S and each θ in Θ, define

$$\tau_n(z,\theta) = P_n[f(z,\cdot,\theta) + f(\cdot,z,\theta)]\,.$$

Notice that $I\!E\tau_n(z,\theta) = \tau(z,\theta)$.

A trivial modification of the argument in Section 5 shows that the class of functions $\{f(z,\cdot,\theta) + f(\cdot,z,\theta) : z \in S, \theta \in \Theta)\}$ is Euclidean for the constant envelope 2. Deduce from the previous lemma that

$$\sup_{S\otimes\Theta} |\tau_n(z,\theta) - \tau(z,\theta)| = O_p(1/\sqrt{n})\,. \tag{21}$$

Let $\{\epsilon_n\}$ denote a sequence of real numbers converging to zero, and let u_i denote the unit vector in $I\!R^{d-1}$ with ith component equal to unity. Estimate Δ with $\hat{\Delta} = (\hat{\delta}_{ij})$, where

$$\hat{\delta}_{ij} = P_n[\hat{p}_i(\cdot,\theta_n)\hat{p}_j(\cdot,\theta_n)]$$

and, for each z in S and each θ in Θ,

$$\hat{p}_i(z,\theta) = \epsilon_n^{-1}[\tau_n(z,\theta+\epsilon_n u_i) - \tau_n(z,\theta)]\,.$$

Deduce from (21) that

$$\hat{p}_i(z,\theta) = \epsilon_n^{-1}[\tau(z,\theta+\epsilon_n u_i) - \tau(z,\theta)] + \epsilon_n^{-1}O_p(1/\sqrt{n})\,.$$

Provided $n^{1/2}\epsilon_n \to \infty$, the consistency of θ_n and the smoothness of $\tau(z,\cdot)$ near $\mathbf{0}$ imply that $\hat{p}_i(z,\theta_n)$ converges in probability to $\nabla^i\tau(z,\mathbf{0})$. Then $\hat{\delta}_{ij}$ converges in probability to $I\!E\nabla^i\tau(\cdot,\mathbf{0})\nabla^j\tau(\cdot,\mathbf{0})$ by a law of large numbers.

Write ∇^{ij} for the ijth element of ∇_2. Estimate V with $\hat{V} = (\hat{v}_{ij})$ where

$$2\hat{v}_{ij} = P_n\hat{p}_{ij}(\cdot,\theta_n)$$

and, for each z in S and each θ in Θ,

$$\hat{p}_{ij}(z,\theta) = \epsilon_n^{-2}[\tau_n(z,\theta+\epsilon_n(u_i+u_j)) - \tau_n(z,\theta+\epsilon_n u_j) \\ -\tau_n(z,\theta+\epsilon_n u_i) + \tau_n(z,\theta)]\,.$$

Argue as before. Provided $n^{1/4}\epsilon_n \to \infty$, $\hat{p}_{ij}(z,\theta_n)$ converges in probability to $\nabla^{ij}\tau(z,\mathbf{0})$, and $2\hat{v}_{ij}$ converges in probability to $I\!E\nabla^{ij}\tau(\cdot,\mathbf{0})$.

For notational simplicity, we have defined $\hat{p}_i(z,\theta)$ and $\hat{p}_{ij}(z,\theta)$ above with difference quotients not centered at θ. Since $\tau_n(z,\cdot)$ is a step function, centered difference quotients perform better in practice, especially for small sample sizes.

Proper choices of ϵ_n are crucial for obtaining reliable estimates of Δ and V, and merit further study. We offer some general guidelines below.

Notice that

$$\hat{p}_i(z,\theta) = \nabla^i\tau(z,\theta) + \epsilon_n\nabla^{ii}\tau(z,\theta^*) + \epsilon_n^{-1}O_p(1/\sqrt{n})$$

for θ^* between θ and $\theta+\epsilon_n$. Minimize the right-hand side with respect to ϵ_n, ignoring the dependence of θ^* and the $O_p(1/\sqrt{n})$ term on ϵ_n. Provided $\nabla^{ii}\tau(z,\theta^*) \neq 0$, $n^{-1/4}$ is a reasonable rate for ϵ_n. A similar calculation and proviso lead to $n^{-1/6}$ as reasonable for estimating $\nabla^{ij}\tau(z,\theta)$.

The perturbation ϵ_n need not be deterministic. For example, $\hat{p}_i(z,\theta_n)$ will converge in probability to $\nabla^i\tau(z,\mathbf{0})$ if ϵ_n has the form $\hat{\sigma}_n\gamma_n$ where $n^{1/2}\gamma_n \to 0$ and $\hat{\sigma}_n$ consistently estimates some appropriate measure of scale in the data. Such an estimate could comprise both global and local measures of spread, as is done in density estimation (See Silverman (1986).).

An appropriate value of ϵ_n could also be obtained by minimizing a measure of discrepancy between estimator and estimand. For example, write $\hat{p}_i(z,\theta,\epsilon)$ for $\hat{p}_i(z,\theta)$ with ϵ replacing ϵ_n, and consider the mean square error criterion

$$I\!E[\hat{p}_i(\cdot,\mathbf{0},\epsilon) - \nabla^i\tau(\cdot,\mathbf{0})]^2\,.$$

Minimizing this last quantity with respect to ϵ is equivalent to minimizing

$$L_n(\epsilon) = I\!E[\hat{p}_i(\cdot,\mathbf{0},\epsilon)]^2 - 2I\!E\hat{p}_i(\cdot,\mathbf{0},\epsilon)\nabla^i\tau(\cdot,\mathbf{0})\,.$$

One could estimate $L_n(\epsilon)$ with

$$\hat{L}_n(\epsilon) = P_n[\hat{p}_i(\cdot,\theta_n,\epsilon)]^2 - 2P_n\hat{p}_{i1}(\cdot,\theta_n,\epsilon)\hat{p}_{i2}(\cdot,\theta_n,\epsilon)$$

where p_{i1} and p_{i2} have the same form as p_i, but are based on observations from the first and second half of the sample, respectively. An appropriate value for ϵ_n would be the minimizer of $\hat{L}_n(\epsilon)$.

One might also consider bootstrap estimates of both Δ and V. While such estimates do not require the choice of a smoothing parameter, they will be computationally expensive for moderate to large sample sizes, since in general, each evaluation of the maximand in (2) requires $O(n^2)$ computations.

8. Sufficient Conditions

In this section, we provide a brief discussion of sufficient conditions on $Z = (Y, X)$ for satisfying A4 of Theorem 4. We make no attempt to find the most general conditions. Rather, our intent is to demonstrate that A4 is satisfied in some interesting, nontrivial circumstances, and to provide the reader with the tools needed to discover the full range of applicability of Theorem 4.

Notice that

$$\tau(z,\theta) = \int_{-\infty}^{x'\beta(\theta)} \int_{-\infty}^{y} g(t \mid s,\theta)\, G_Y(ds)\, dt + \int_{x'\beta(\theta)}^{\infty} \int_{y}^{\infty} g(t \mid s,\theta)\, G_Y(ds)\, dt$$

where $G_Y(\cdot)$ denotes the marginal distribution of Y, and $g(\cdot \mid s,\theta)$ denotes the conditional density of $X'\beta(\theta)$ given $Y = s$.

Write $f(\cdot \mid r, s)$ for the conditional density of X_d given $\mathcal{X} = r$ and $Y = s$. For simplicity, assume that each component of $\mathcal{X}$ has bounded support. We now show that if $f(\cdot \mid r, s)$ has bounded derivatives up to order three for each (r, s) in the support of $\mathcal{X} \otimes Y$, then the first four conditions of A4 are met.

Write $g(\cdot \mid r, s, \theta)$ for the conditional density of $X'\beta(\theta)$ given $\mathcal{X} = r$ and $Y = s$. Notice that for each t in $\mathbb{R}$, $g(t \mid r, s, \theta) = f(t - r'\theta \mid r, s)$. It follows from our assumptions that $g(t \mid r, s, \cdot\,)$ has bounded mixed partial derivatives up to order three. Observe that

$$g(t \mid s,\theta) = \int g(t \mid r, s, \theta)\, G_{\mathcal{X}|s}(dr)$$

where $G_{\mathcal{X}|s}(\cdot)$ denotes the conditional distribution of $\mathcal{X}$ given $Y = s$. Deduce that $g(t \mid s, \cdot\,)$ and therefore $\tau(z, \cdot\,)$ have bounded mixed partial derivatives up to order three. This is enough to satisfy the first four conditions of A4. Nonsingularity of V essentially follows from A3, as discussed in Section 6.

Bell Communications Research, Morristown, NJ 07962, U.S.A.

REFERENCES

Chamberlain, G. (1986): "Asymptotic Efficiency in Semi-parametric Models with Censoring," *Journal of Econometrics*, 32, 189–218.

Cosslett, S. R. (1987): "Efficiency Bounds for Distribution-free Estimators of the Binary Choice and Censored Regression Models," *Econometrica*, 55, 559–586.

Han, A. K. (1987): "Non-parametric Analysis of a Generalized Regression Model," *Journal of Econometrics*, 35, 303–316.

—— (1988): "Large Sample Properties of the Maximum Rank Correlation Estimator in Generalized Regression Models," Preprint, Department of Economics, Harvard University, Cambridge, MA.

Huber, P. J. (1967): "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," in *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 221–233. Berkeley, CA: University of California.

Nolan, D., and D. Pollard (1987): "U-Processes: Rates of Convergence," *Annals of Statistics*, 15, 780–799.

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

Pollard D. (1984): *Convergence of Stochastic Processes.* New York: Springer-Verlag.

—— (1985): "New Ways to Prove Central Limit Theorems," *Econometric Theory*, 1, 295–314.

—— (1989): Unpublished lecture notes for an asymptotic methods course taught at Yale University.

Serfling, R. J. (1980): *Approximation Theorems of Mathematical Statistics.* New York: Wiley.

Sherman, R. P. (1992): "Maximal Inequalities for Degenerate U-processes with Applications to Optimization Estimators," forthcoming in *Annals of Statistics.*

Silverman, B. W. (1986): *Density Estimation for Statistics and Data Analysis.* New York: Chapman and Hall.