SOME CONVERGENCE THEORY

FOR ITERATIVE ESTIMATION PROCEDURES

WITH AN APPLICATION TO SEMIPARAMETRIC ESTIMATION

Jeff Dominitz, 412-268-5981

H. John Heinz III School of Public Policy and Management

Carnegie Mellon University

5000 Forbes Avenue

Pittsburgh, PA 15213-3890

Robert P. Sherman*, 626-395-4337

California Institute of Technology

Humanities and Social Sciences

Mail Stop 228-77

Pasadena, California 91125

* We thank a Co-Editor and two referees for comments and criticisms that led to significant improvements in this paper. We also thank Roger Klein for providing us with Gauss code to compute his estimator.

CONVERGENCE THEORY FOR ITERATIVE PROCEDURES

Robert P. Sherman

California Institute of Technology

Humanities and Social Sciences

Mail Stop 228-77

Pasadena, California 91125

Abstract

We develop general conditions for rates of convergence and convergence in distribution of iterative procedures for estimating finite-dimensional parameters. An asymptotic contraction mapping condition is the centerpiece of the theory. We illustrate some of the results by deriving the limiting distribution of a two-stage iterative estimator of regression parameters in a semiparametric binary response model. Simulation results illustrating the computational benefits of the first-stage iterative estimator are also reported.

1. Introduction

There is a substantial literature on convergence properties of various widely applied iterative estimation procedures such as the expectation-maximization (EM) algorithm and its many descendants (see, for example, McLachlan and Krishnan, 1997). Often in this literature, observed data is conditioned on, and convergence refers to numerical convergence of sample iterates to a sample fixed point. Similarly, a rate of convergence refers to how fast the sample iterates converge to the sample fixed point as the number of iterations increases. While useful computationally, such information is not sufficient for doing asymptotic inference on the parameter of interest, namely, the population parameter estimated by the sample fixed point. For example, such information neither reveals the limiting distribution of the sample sequence nor implies a rate of convergence to the parameter of interest, and gives no guidelines about the number of iterations needed to achieve these results. The literature on finite-step estimation sheds some light on these problems, but requires a starting value that converges to the parameter of interest at a known rate (see, for example, Robinson 1988, Lehmann 1983 (Chapter 6.3), or Bickel 1975). By contrast, most iterative estimation procedures do not start at consistent starting values. Hence the need for theory enabling asymptotic inference about a parameter of interest for general iterative estimation procedures. This paper provides such theory.

Specifically, this paper develops checkable conditions for consistency, rates of convergence, and convergence in distribution of iterative procedures for estimating finite-dimensional parameters. The theory covers procedures like expectation-maximization (EM), Newton-Raphson (NR), and iterative least squares (ILS). The key requirement is that the sample mapping generating the procedure and a population analogue be contraction mappings, asymptotically. We

also isolate a convenient bias condition from which sensible stopping rules can be derived.

We illustrate the theory by establishing the limiting distribution of a two-stage iterative estimator of the regression parameters in a semiparametric binary response model. The first stage is an ILS procedure. We apply the theory to show that this procedure consistently estimates the regression parameters in the model. The second stage is a NR procedure that starts at the ILS estimates and is based on the criterion function of the Klein and Spady (1993) estimator that achieves the semiparametric efficiency bound for this model established by Chamberlain (1986) and Cosslett (1987). We use the theory to show that the two-stage procedure also achieves this bound. While we use this application to illustrate the asymptotic theory, the ILS estimator is interesting in its own right because of the substantial computational advantages it can provide, particularly in applications with many observations and many regressors. To illustrate these benefits, we provide a small simulation study comparing the ILS estimator to the efficient estimator of Klein and Spady (1993). We also briefly discuss extensions of the ILS procedure to semiparametric censored regression models.

Simultaneously and independently of this work, Pastorello, Patilea, and Renault (2003, Sections 1-4) develop a similar theory of iterative estimation and consider applications to structural nonadaptive econometric models with an emphasis on financial market models. The theory developed in this paper covers all of their main applications. However, their theory requires a continuity condition (Pastorello et al., 2003, Assumption 1a p. 452) that limits the applicability of their results. For example, we show that this continuity condition does not hold for the ILS estimator developed in Section 3 for the semiparametric binary response model. Nor do these authors provide guidelines for stopping rules.

The rest of the paper is organized as follows. In Section 2, we develop conditions for

consistency, rates of convergence, and convergence in distribution of general iterative estimation procedures. We also provide guidelines for developing stopping rules for these procedures. In Section 3, we apply the theory to the iterative estimator for the semiparametric binary response model described above. Section 4 presents the simulation study. Section 5 summarizes. Proofs and other technical supporting material are provided in an appendix.

2. Theory

This section develops general conditions for rates of convergence and convergence in distribution of iterative procedures for estimating finite-dimensional parameters. We begin by developing some notation.

Let $(\Omega, \mathcal{A}, I\!P)$ denote a probability space. That is, Ω is a sample space, $\mathcal{A}$ is a σ-field of subsets of Ω, and $I\!P$ is a probability measure on $\mathcal{A}$. For $\omega \in \Omega$, let $Z_1(\omega, \beta_0), \ldots, Z_n(\omega, \beta_0)$ denote a sample of size n from $I\!P$, where β_0 denotes a fixed parameter of interest in $I\!R^k$, $k \geq 1$.

Let $M(\phi)$ denote a mapping from $I\!R^k$ to $I\!R^k$ with fixed point β_0. That is, $M(\beta_0) = \beta_0$. Population analogues of sample mappings generating common iterative estimators satisfy this fixed point condition. For example, a population analogue of the sample mapping defining an EM procedure has the form

$$M(\phi) = \underset{\beta}{\operatorname{argmax}}\, I\!E Q_n(\beta \mid \phi)$$

where $Q_n(\beta \mid \phi)$ is the conditional expectation of the complete data log-likelihood function given the observed data and ϕ. The outer expectation is over the distribution of the observed data. It follows that $I\!E Q_n(\beta \mid \beta_0)$ is the expected value of the complete data log-likelihood function, which is maximized at β_0. That is, $M(\beta_0) = \beta_0$. A population analogue of a sample

mapping defining a NR procedure has the form

$$M(\phi) = \phi - [H(\phi)]^{-1} G(\phi)$$

where $G(\phi)$ is a population analogue of the gradient of the sample objective function and $H(\phi)$ is a population analogue of the corresponding sample hessian. Under general conditions, $G(\beta_0) = 0$ and so $M(\beta_0) = \beta_0$.

We show in Section 3 that the fixed point condition just described is also satisfied by a population analogue of a semiparametric ILS mapping. However, this population mapping depends on both the sample size n and the ω that generated the sample from $\mathbb{P}$. In order to cover applications like this, from now on, we write $M_n(\phi)$ for a generic population mapping, letting the subscript n suggest the possible dependence of this mapping on both n and ω. The fixed point condition, then, becomes $M_n(\beta_0) = \beta_0$ for all n and ω. Write $\hat{M}_n(\phi)$ for a sample analogue of $M_n(\phi)$. Let $\beta_n^0 \equiv \hat{\beta}_n^0$ denote a starting point in $\mathbb{R}^k$. This starting point may depend on n and ω. For $i \geq 1$, define $\beta_n^i = M_n(\beta_n^{i-1})$ and $\hat{\beta}_n^i = \hat{M}_n(\hat{\beta}_n^{i-1})$. We call β_n^i a population iterate. We call $\hat{\beta}_n^i$ a sample iterate as well as an iterative estimator of β_0.

Contraction mappings play a central role in establishing the limiting behavior of $\hat{\beta}_n^i$. We now introduce the notion of an asymptotic contraction mapping.

DEFINITION. *For each $n \geq 1$ and $\omega \in \Omega$, let $K_n^\omega(\cdot)$ be a function defined on a set $\mathcal{X}$ where $(\mathcal{X}, d)$ is a metric space. The collection $\{K_n^\omega(\cdot) : n \geq 1, \omega \in \Omega\}$ is an asymptotic contraction mapping (denoted ACM) on $(\mathcal{X}, d)$ if there exist a constant c in $[0, 1)$ which does not depend on n or ω, and sets $\{A_n\}$ with each $A_n \subseteq \Omega$ and $\mathbb{P}A_n \to 1$ as $n \to \infty$, such that for each $\omega \in A_n$, $K_n^\omega(\cdot)$ maps $\mathcal{X}$ to itself and for all $x, y \in \mathcal{X}$,*

$$d(K_n^\omega(x), K_n^\omega(y)) \leq c\, d(x, y).$$

□

The ACM property is a property of the collection of functions $\{K_n^\omega(\cdot) : n \geq 1, \omega \in \Omega\}$. For ease of notation, we write $\{K_n^\omega(\cdot)\}$ for this collection.

If $\{K_n^\omega(\cdot)\}$ is an ACM on $(\mathcal{X}, d)$ and $\omega \in A_n$, where A_n is one of the "good" sets described in the definition, then $K_n^\omega(\cdot)$ is a contraction mapping on $(\mathcal{X}, d)$. Then, by the Banach Fixed Point Theorem (Aliprantis and Border, 1994, pp.88–89), $K_n^\omega(\cdot)$ has a unique fixed point $\hat{x}_n$ in $\mathcal{X}$, and any sequence defined by $\hat{x}_n^i = K_n^\omega(\hat{x}_n^{i-1})$ where $\hat{x}_n^0 \in \mathcal{X}$ converges to $\hat{x}_n$ as $i \to \infty$. Note that the iterates and the fixed point can depend on n and ω. Also, note that $\{K_n^\omega(\cdot)\}$ can be an ACM without $K_n^\omega(\cdot)$ being a contraction mapping for each n and ω.

To establish the limiting distribution of $\hat{\beta}_n^i$ we require that both $\{M_n(\phi)\}$ and $\{\hat{M}_n(\phi)\}$ be ACMs on (B_0, E_k). Here E_k is the Euclidean metric on $\mathbb{R}^k$ and B_0 is the closed ball centered at β_0 of radius $|\beta_n^0 - \beta_0|$. If $\{M_n(\phi)\}$ is an ACM on (B_0, E_k), then, for each $\omega \in A_n$ (see definition), β_0 is the unique fixed point of $M_n(\phi)$ on B_0. If $\{\hat{M}_n(\phi)\}$ is an ACM on (B_0, E_k), then, for each $\omega \in A_n$ (not necessarily the same A_n as for $M_n(\phi)$), $\hat{M}_n(\phi)$ has a unique fixed point on B_0, which we denote $\hat{\beta}_n$. Unlike β_0, $\hat{\beta}_n$ typically depends on both n and ω.

Theorem 1 gives conditions for rates of convergence of $\hat{\beta}_n^i$. Let Z^+ be the positive integers.

THEOREM 1: *Let $i(n)$ be a function from Z^+ to Z^+. Fix $\delta > 0$. If*

(i) *$\{M_n(\phi)\}$ is an ACM on (B_0, E_k)*

(ii) *$n^\delta |\beta_n^{i(n)} - \beta_0| = O_p(1)$ as $n \to \infty$*

(iii) *$n^\delta \sup_{\phi \in B_0} |\hat{M}_n(\phi) - M_n(\phi)| = O_p(1)$ as $n \to \infty$*

then $n^\delta |\hat{\beta}_n^{i(n)} - \beta_0| = O_p(1)$ as $n \to \infty$.

REMARK 1. In order to apply Theorem 1, one must choose $i(n)$ to satisfy condition (ii). The choice depends on the order of convergence of the iterative procedure. Let β^i, $i \geq 1$, and β be points in $\mathbb{R}^k$. The sequence $\{\beta^i\}_{i=1}^{\infty}$ converges to β of order $\sigma \geq 1$ if there exists a constant $\kappa > 0$ such that for $i \geq 1$, $|e_i|/|e_{i-1}|^\sigma \leq \kappa$ where $e_i = \beta^i - \beta$. (cf. Burden et al., 1981, p.45.) The leading special cases are $\sigma = 1$ (linear convergence) and $\sigma = 2$ (quadratic convergence). For example, the population sequence for the semiparametric ILS procedure analyzed in Section 3 exhibits linear convergence to its fixed point, while the population and sample sequences for the semiparametric NR procedure analyzed in Section 3 exhibit quadratic convergence to their respective fixed points.

First, we choose $i(n)$ to satisfy condition (ii) when $\sigma = 1$. In this case, the constant κ can be taken to equal the modulus of contraction c guaranteed by condition (i). Assume $\sup_n |\beta_n^0 - \beta_0| < \infty$. A simple recursive calculation shows that $n^\delta |\beta_n^{i(n)} - \beta_0| \leq |\beta_n^0 - \beta_0|$ for each $n \geq 1$ (a stronger condition than condition (ii)) provided $i(n) \geq -\delta \ln n / \ln c$. This bound is sharp and can be used to develop a stopping rule for the iterative sequence. For instance, if $\sigma = 1$, $\delta = 1/2$, $n = 5000$, and $c \leq .9$, then the stronger condition just stated is satisfied provided $i(5000) \geq 41$. Alternatively, one can estimate c with the maximum of the ratios $|\hat{M}(\hat{\beta}_n^i) - \hat{M}(\hat{\beta}_n^{i-1})|/|\hat{\beta}_n^i - \hat{\beta}_n^{i-1}| = |\hat{\beta}_n^{i+1} - \hat{\beta}_n^i|/|\hat{\beta}_n^i - \hat{\beta}_n^{i-1}|$ over a small number of ratios and for different starting values for the sequence.

Next, we choose $i(n)$ to satisfy condition (ii) when $\sigma > 1$. Define $\alpha(\sigma) = \kappa^{\frac{1}{\sigma-1}} |\beta_n^0 - \beta_0|$ and assume $\alpha(\sigma) < 1$. A recursive calculation shows that $n^\delta |\beta_n^{i(n)} - \beta_0| \leq |\beta_n^0 - \beta_0|$ for $n \geq 1$ (again, a stronger condition than condition (ii)) provided $i(n) \geq [\ln \sigma]^{-1} \ln(-\delta \ln n / \ln \alpha(\sigma))$. To illustrate, for a smooth NR procedure, $\sigma = 2$ and the constant κ can be taken to equal $2Ck$ with $2C$ an upper bound on the absolute value of each of the second order mixed partial

derivatives of each of the k components of M_n (cf. Burden et al., 1981, p.47).[1] The constant $\alpha(2)$ can be chosen strictly less than unity by starting the procedure close enough to β_0. If $\sigma = 2$, $\delta = 1/2$, $n = 5000$, and $\alpha(2) \leq .9$, then the stronger condition stated above is satisfied provided $i(5000) \geq 6$.

Finally, we note that condition (ii) does not require that $i(n) \to \infty$ as $n \to \infty$. As a simple example, consider sampling iid observations from a normal distribution with unknown mean β_0 and known variance, and estimating β_0 using an NR procedure. Then it is trivial to show that no matter what the starting value β_n^0, $\beta_n^1 = \beta_0$. That is, condition (ii) is satisfied for all $\delta > 0$ with $i(n) = 1$ for all n.

REMARK 2. Many iterative procedures converge at rate $\sqrt{n}$, corresponding to $\delta = 1/2$ in Theorem 1. However, this need not be the case. For example, consider estimating the regression parameters in a semiparametric binary response model using a NR procedure based on the smoothed maximum score estimator of Horowitz (1992). Depending on the smoothness of the data distribution, under the conditions of Theorem 1, this NR procedure will converge at rate n^δ, for some $\delta \in [1/3, 1/2)$.

The next result gives conditions for consistency without a rate of convergence.

THEOREM 2: *Let $i(n)$ be a function from Z^+ to Z^+. If*

(i) *$\{M_n(\phi)\}$ is an ACM on (B_0, E_k)*

(ii) *$|\beta_n^{i(n)} - \beta_0| = o_p(1)$ as $n \to \infty$*

(iii) *$\sup_{\phi \in B_0} |\hat{M}_n(\phi) - M_n(\phi)| = o_p(1)$ as $n \to \infty$*

then $|\hat{\beta}_n^{i(n)} - \beta_0| = o_p(1)$ *as* $n \to \infty$.

REMARK 3. In Theorem 2, if condition (i) holds, then condition (ii) holds if $i(n) \to \infty$ as $n \to \infty$. No minimum rate of growth is required for $i(n)$ because the bias term in (ii) is not inflated by a factor of n^δ as it is in Theorem 1.

Our next objective is to develop conditions for convergence in distribution of $\hat{\beta}_n^i$. To do so, we require that $\{\hat{M}_n(\phi)\}$ be an ACM on (B_0, E_k).

Assume that the first partial derivatives of the components of $\hat{M}_n(\phi)$ and $M_n(\phi)$ exist. Let ∇_ϕ denote the differential operator $(\partial/\partial\phi_1, \ldots, \partial/\partial\phi_k)$. Write $V_n(\phi)$ for the $k \times k$ matrix $\nabla_\phi M_n(\phi)$ and $\hat{V}_n(\phi)$ for the $k \times k$ matrix $\nabla_\phi \hat{M}_n(\phi)$. For a $k \times k$ matrix $A = (a_{ij})$, let $\|A\|$ denote the matrix norm $[\sum_{ij} a_{ij}^2]^{1/2}$.

LEMMA 3. *Suppose* $V_n(\phi)$ *and* $\hat{V}_n(\phi)$ *exist. If*

(i) $\{M_n(\phi)\}$ *is an* ACM *on* (B_0, E_k)

(ii) $\sup_{\phi \in B_0} |\hat{M}_n(\phi) - M_n(\phi)| = o_p(1)$ *as* $n \to \infty$

(iii) $\sup_{\phi \in B_0} \|\hat{V}_n(\phi) - V_n(\phi)\| = o_p(1)$ *as* $n \to \infty$

then $\{\hat{M}_n(\phi)\}$ *is an* ACM *on* (B_0, E_k).

We are now in a position to establish a convergence in distribution result for $\hat{\beta}_n^i$. The limiting distribution depends on the limiting behavior of the infeasible estimator $\hat{M}_n(\beta_0)$. Recall that if $\{\hat{M}_n(\phi)\}$ is an ACM on (B_0, E_k), then, for $\omega \in A_n$, $\hat{M}_n(\phi)$ has a unique fixed point on B_0, denoted $\hat{\beta}_n$. Assume that the probability limit of $V_n(\phi)$ exists and let $V(\phi)$ denote this quantity. In what follows, we use the symbol $\Longrightarrow$ to denote convergence in distribution.

THEOREM 4: *Let $i(n)$ be a function from Z^+ to Z^+, and let $\{\epsilon_n\}$ denote an arbitrary sequence of nonnegative real numbers converging to zero as $n \to \infty$. For $\delta > 0$, if*

(i) $|\hat{\beta}_n^{i(n)} - \beta_0| = o_p(1)$ *as* $n \to \infty$

(ii) $\{\hat{M}_n(\phi)\}$ *is an ACM on* (B_0, E_k)

(iii) $n^\delta |\hat{\beta}_n^{i(n)} - \hat{\beta}_n| = o_p(1)$ *as* $n \to \infty$

(iv) *For some* $\delta > 0$, $n^\delta(\hat{M}_n(\beta_0) - \beta_0) \Longrightarrow Z$ *as* $n \to \infty$

(v) $\sup_{|\phi - \beta_0| \leq \epsilon_n} \|\hat{V}_n(\phi) - V(\phi)\| = o_p(1)$ *as* $n \to \infty$

(vi) $V(\phi)$ *is continuous in an open neighborhood of* β_0

then $n^\delta(\hat{\beta}_n^{i(n)} - \beta_0) \Longrightarrow DZ$ *as* $n \to \infty$, *where* $D = [I_k - V(\beta_0)]^{-1}$.

REMARK 4. Theorem 2 provides checkable conditions that imply condition (i). Lemma 3 provides checkable conditions that imply condition (ii). Remark 1 concerning the choice of $i(n)$ can be adapted to establish condition (iii). For example, in the application in Section 3, the sample sequence of the semiparametric NR procedure is started at a consistent estimator $\hat{\beta}^0$ and exhibits quadratic convergence to its sample fixed point. Replacing M_n with $\hat{M}_n$ and β_n^0 with $\hat{\beta}^0$ in Remark 1 and choosing $i(n) \geq [\ln 2]^{-1} \ln(-\delta \ln n / \ln \alpha(2))$ is sufficient to establish condition (iii). Finally, note that procedures exhibiting linear convergence satisfy $V(\beta_0) \neq 0$, whereas procedures exhibiting quadratic convergence (such as NR procedures) satisfy $V(\beta_0) = 0$ (Burden et. al. 1981, p.47-48). Thus, for NR procedures like the one presented in Section 3, D is the identity matrix and so $n^\delta(\hat{\beta}_n^{i(n)} - \beta_0) \Longrightarrow Z$ as $n \to \infty$.

3. AN ILLUSTRATION

In this section, we illustrate the theory developed in Section 2 by establishing the limiting behavior of a two-stage iterative estimator of regression parameters in a semiparametric binary response model. The first stage estimator is an ILS estimator. We establish consistency of this estimator by developing primitive conditions implying the conditions of Theorem 2 in Section 2.[2] We then apply Theorem 4 in Section 2 to establish the limiting distribution of an NR estimator started at the ILS estimates. The NR estimator is based on the criterion function of the Klein and Spady (1993) estimator, which achieves the semiparametric efficiency bound for this model established by Chamberlain (1986) and Cosslett (1987). We show that this two-stage estimator also achieves this bound.

Consider the binary response model $Y = \mathbf{1}\{Y^* \geq 0\}$ where the latent variable $Y^* = X'\beta_0 - u$, $X = (W_1, \ldots, W_k, W_{k+1})'$, $\beta_0 = (\beta_{01}, \ldots, \beta_{0k}, \beta_{0,k+1})'$, and u is an error term with unknown distribution. Since the distribution of u is not known in this model, restrictions are needed to identify the regression parameters. We assume that W_{k+1} is nonconstant and normalize $\beta_{0,k+1}$ to unity. Rather than introduce new notation, we reinterpret β_0 as the first k components of the true parameter vector divided by $\beta_{0,k+1}$, and u as the true error divided by $\beta_{0,k+1}$. Also, for each ϕ in $\mathbb{R}^k$, we write $X'\phi$ for $W_1\phi_1 + \cdots + W_k\phi_k + W_{k+1}$. In addition, we take $W_1 = 1$. That is, W_1 is the regressor corresponding to the intercept term in the model.

Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ denote a sample of independent observations from the model just defined, and let $\mathbf{X}_n$ denote the $n \times k$ matrix comprised of the first k components of each X_j, $j = 1, \ldots, n$. Prewhiten $\mathbf{X}_n$ so that $\mathbf{X}_n'\mathbf{X}_n = nI_k$.

Next, let F and f denote the unknown cumulative distribution function and density function of u. Fix t in $\mathbb{R}$ and ϕ in $\mathbb{R}^k$. Let $F(t, \phi) = \mathbb{P}\{Y = 1 \mid X'\phi = t\}$. Assume that $\nabla_t F(t, \phi)$

exists and write $f(t,\phi) = \nabla_t F(t,\phi)$. Note that $F(t,\beta_0) = F(t)$ and $f(t,\beta_0) = f(t)$. Define

$$\begin{aligned}
\nu(t,\phi) &= \int_{-\infty}^{t} u\, f(u,\phi)\, du / F(t,\phi) = t - \int_{-\infty}^{t} F(u,\phi)\, du / F(t,\phi) \\
\pi(t,\phi) &= \int_{t}^{\infty} u\, f(u,\phi)\, du / [1 - F(t,\phi)] = t + \int_{t}^{\infty} [1 - F(u,\phi)]\, du / [1 - F(t,\phi)]\,.
\end{aligned}$$

The second representations for $\nu(t,\phi)$ and $\pi(t,\phi)$ follow from integration by parts arguments. Note that $\nu(t,\beta_0) = I\!E[u \mid u \le t]$ and $\pi(t,\beta_0) = I\!E[u \mid u > t]$. Write $\mathbf{u}_n(\phi)$ for $(u_1(\phi), \ldots, u_n(\phi))'$ with $u_j(\phi) = F(X_j'\beta_0)\nu(X_j'\phi,\phi) + (1 - F(X_j'\beta_0))\pi(X_j'\phi,\phi)$. Write $Y_j(\phi)$ for $X_j'\phi - u_j(\phi)$. Define the population ILS mapping

$$M_n(\phi) = \operatorname*{argmin}_{\beta} \sum_{j=1}^{n} (Y_j(\phi) - X_j'\beta)^2 = \phi - n^{-1}\mathbf{X}_n'\mathbf{u}_n(\phi)\,.$$

Note that for all j, $u_j(\beta_0) = I\!Eu$. Deduce from this and prewhitening that $M_n(\beta_0) = \beta_0$. That is, β_0 is a fixed point of $M_n(\phi)$.

We now construct a sample analogue of $M_n(\phi)$ by developing numerical integral approximations of $\nu(X_j'\phi,\phi)$ and $\pi(X_j'\phi,\phi)$ using nearest neighbor estimators of the $F(X_j'\phi,\phi)$'s. Fix $\alpha > 1$ and let $\{c_n\}$ denote a sequence of nonnegative real numbers satisfying $c_n \to \infty$ as $n \to \infty$. For $t \in I\!R$, define the trimming functions $\tau_n(t) = \mathbf{1}\{|t| \le c_n\}$ and $\sigma_n(t) = \mathbf{1}\{|t| \le c_n^{\alpha}\}$. Note that $\tau_n(\cdot)$ trims more severely than $\sigma_n(\cdot)$. These functions help control tail behavior. Relabel observation numbers so that index values are ordered from smallest to largest. That is, let $X_1'\phi \le \cdots \le X_n'\phi$. Let $X_0'\phi = X_1'\phi$, $X_{n+1}'\phi = X_n'\phi$, and $\Delta(X_i'\phi) = X_i'\phi - X_{i-1}'\phi$, $i = 1, \ldots, n+1$. For $j = 1, \ldots, n$, define

$$\begin{aligned}
\hat{\nu}(X_j'\phi,\phi) &= X_j'\phi - \sum_{i=1}^{j} \sigma_n(X_i'\phi)\hat{F}(X_i'\phi,\phi)\, \Delta(X_i'\phi) / \hat{F}(X_j'\phi,\phi) \\
\hat{\pi}(X_j'\phi,\phi) &= X_j'\phi + \sum_{i=j}^{n} \sigma_n(X_i'\phi)[1 - \hat{F}(X_i'\phi,\phi)]\, \Delta(X_{i+1}'\phi) / [1 - \hat{F}(X_j'\phi,\phi)]\,.
\end{aligned}$$

In the expressions above, $\hat{F}(X_i'\phi, \phi)$ is a nearest neighbor estimator of $F(X_i'\phi, \phi)$. Let k_n be a positive integer. If $k_n < i \leq n - k_n$, then $\hat{F}(X_i'\phi, \phi)$ is the arithmetic average of the $2k_n + 1$ binary values $Y_{i-k_n}, \ldots, Y_{i+k_n}$. These are symmetric nearest neighbor estimators. If $i \leq k_n$, then $\hat{F}(X_i'\phi, \phi)$ is the average of the $k_n + 1$ binary values $Y_i, \ldots, Y_{i+k_n}$. If $i > n - k_n$, then $\hat{F}(X_i'\phi, \phi)$ is the average of the $k_n + 1$ binary values $Y_{i-k_n}, \ldots, Y_i$.

Write $\hat{\mathbf{u}}_n(\phi)$ for $(\hat{u}_1(\phi), \ldots, \hat{u}_n(\phi))'$ with $\hat{u}_j(\phi) = \left[Y_j\hat{\nu}(X_j'\phi, \phi) + (1 - Y_j)\hat{\pi}(X_j'\phi, \phi)\right] \tau_n(X_j'\phi)$. Write $\hat{Y}_j(\phi)$ for $X_j'\phi - \hat{u}_j(\phi)$.[3] Define the sample ILS mapping

$$\hat{M}_n(\phi) = \underset{\beta}{\operatorname{argmin}} \sum_{j=1}^{n} (\hat{Y}_j(\phi) - X_j'\beta)^2 = \phi - n^{-1}\mathbf{X}_n'\hat{\mathbf{u}}_n(\phi)\,.$$

Let $\hat{\beta}_n^0$ denote an arbitrary starting value and for each $i \geq 1$, define $\hat{\beta}_n^i = \hat{M}_n(\hat{\beta}_n^{i-1})$. We call $\hat{\beta}_n^i$ a semiparametric ILS estimator of β_0.

In order to prove consistency of the ILS estimator $\hat{\beta}_n^i$, we see from Theorem 2 in Section 2 that we must show that $M_n(\phi)$ is an asymptotic contraction mapping and that $\hat{M}_n(\phi)$ converges uniformly to $M_n(\phi)$. We now state and discuss primitive conditions sufficient to imply these high-level conditions.

Let $\mathcal{N}$ denote the closure of an open convex neighborhood of β_0. Write S_u for the support of u and S_ϕ for the support of $X'\phi$. Write $g(t, \phi)$ for the density of $X'\phi$ evaluated at t. Throughout, we will maintain the following assumptions for $t \in \mathbb{R}$ and $\phi \in \mathcal{N}$.

A0. $(Y_j, X_j)_{j=1}^n$ are iid observations from the model $Y = \mathbf{1}\{u < X'\beta_0\}$ described above.

A1. $S_u = \mathbb{R}$ and u is log-concave and independent of X.

A2. $S_\phi = \mathbb{R}$, $\mathbb{E}|X|^2 < \infty$, $\mathbf{X}_n$ has full rank, and $\mathbf{X}_n'\mathbf{X}_n = nI_k$.

A3. $k_n \propto n^{3/4}$.

A4. $c_n \to \infty$ as $n \to \infty$ and $c_n \ll n^\delta$ for all $\delta > 0$.

A5. For all $\delta > 0$, $\sup_{t,\phi} \sigma_n(t)/g(t,\phi) \ll n^{\delta}$.

A6. For all $\delta > 0$, $\sup_{t,\phi} \sigma_n(t)/F(t,\phi) \ll n^{\delta}$ and $\sup_{t,\phi} \sigma_n(t)/[1-F(t,\phi)] \ll n^{\delta}$.

A7. $F(t,\phi) \searrow 0$ as $t \to -\infty$ and $\sup_{\phi} \int_{-\infty}^{-c_n^{\alpha}} F(u,\phi)du/F(-c_n,\phi) \to 0$ as $n \to \infty$.

A8. $F(t,\phi) \nearrow 1$ as $t \to \infty$ and $\sup_{\phi} \int_{c_n^{\alpha}}^{\infty} [1-F(u,\phi)]du/[1-F(c_n,\phi)] \to 0$ as $n \to \infty$.

A9. $\sup_{t,\phi} g(t,\phi) < \infty$, $\sup_{t,\phi} |\nabla_t g(t,\phi)| < \infty$, and $\sup_{t,\phi} f(t,\phi) < \infty$.

A10. $g(t,\phi)$ is continuous in t and ϕ.

A11. $\nu(X'\phi,\phi)$, $\pi(X'\phi,\phi)$, and the components of $\nabla_\phi \nu(X'\phi,\phi)$ and $\nabla_\phi \pi(X'\phi,\phi)$ have square integrable envelopes and satisfy a Lipschitz condition in ϕ.

Assumption A0 describes the data and the model.

A1 and A2 are assumptions about u and X and their relationship to each other. Note that we assume that u is log-concave. This is a large class of distributions, and includes normal, logistic, extreme-value, Laplacian, Gamma, Beta, and triangular families, and many others as well. However, log-concavity is only a sufficient condition. As we will demonstrate in the next section, the ILS procedure can work well even when this assumption does not hold, as when u is log-convex. It is not necessary that $S_u = S_\phi = I\!R$. We make these assumptions for ease of exposition and because the infinite support case is a leading special case. The prewhitening in A2 is used to establish the local contraction mapping property. In general, prewhitening would not be effective in this regard if the index were not linear.

A3 defines the number of neighbors up to scale. From the proofs of Lemma 1A and Lemma 2A in the appendix, we see that $n^{1/2} \ll k_n \ll n$ is sufficient. However, we choose $n^{3/4}$ since this rate optimally balances convergence rates of stochastic and bias terms in the nearest neighbor estimators (see proof of Lemma 3A in the appendix). The constant of pro-

portionality could be any consistent estimate of a measure of spread in the $X_i'\phi$'s.

Assumptions A4 through A8 define trimming and tail conditions. The interval $[-c_n, c_n]$ defines the set of t values for which numerical integral estimates of $\nu(t,\phi)$ and $\pi(t,\phi)$ are computed, while the interval $[-c_n^\alpha, c_n^\alpha]$ defines the set of t values at which nearest neighbor nonparametric regression estimates of $F(t,\phi)$ are computed. These latter estimates are ratios with estimates of $g(t,\phi)$ in denominators and appear in denominators of estimates of $\nu(t,\phi)$ and $\pi(t,\phi)$. To give an example of what we have in mind for A4 through A8, suppose c_n is proportional to $\sqrt{\ln \ln n}$. Then simple calculations show that A5 is satisfied provided the tails of $X'\phi$ decrease no faster than normal tails. If, in addition, $\alpha > (1+\delta)/\delta$ for $\delta > 0$, then the conditions in A6 through A8 are satisfied at β_0 provided the tails of u decrease no faster than normal tails and no slower than $|t|^{-(2+\delta)}$ as $|t| \to \infty$.

Assumptions A9 through A11 are conditions useful for bounding remainder terms in Taylor expansions or in proving uniform convergence results.

LEMMA 5: *Suppose A0, A1, A2, and A11 hold. Then there exists an open ball centered at* β_0 *with closure* $B_0 \subseteq \mathcal{N}$ *such that* $\{M_n(\phi)\}$ *is an ACM on* (B_0, E_k).

The following result is based on Lemmas 1A, 2A, and 3A, which extend some results of Ichimura (1997) relating k_n, the number of nearest neighbors, to nearest neighbor window widths. It also uses Lemma 4A, a result on maximum sample spacings, to show that numerical integral approximations are close to their estimands.

LEMMA 6: *A0 through A11 imply* $\sup_{\phi \in \mathcal{N}} \left| \hat{M}_n(\phi) - M_n(\phi) \right| = o_p(1)$ *as* $n \to \infty$.

The next result follows from Lemma 5, Lemma 6, and Theorem 2.

THEOREM 7. *Assume A0 through A11 hold and $\hat{\beta}_n^0 \in B_0$ from Lemma 5. If $i(n) \to \infty$ as $n \to \infty$, then $|\hat{\beta}_n^{i(n)} - \beta_0| = o_p(1)$ as $n \to \infty$.*

REMARK 5. The starting value, $\hat{\beta}_n^0$, for the ILS procedure may or may not be an element of B_0. Thus, successful implementation of the procedure may require good starting values. Standard parametric estimates, such as probit or logit estimates, can be tried. For our simulations, we used OLS estimates. While these were generally poor estimates of β_0, they proved to be good enough starting values for the ILS procedure. Also, note that an informal check of the contraction mapping condition can always be made: given a sequence of ILS iterates $\hat{\beta}_n^0, \hat{\beta}_n^1, \ldots$, compute the ratios $|\hat{M}_n(\hat{\beta}_n^i) - \hat{M}_n(\hat{\beta}_n^{i-1})|/|\hat{\beta}_n^i - \hat{\beta}_n^{i-1}| = |\hat{\beta}_n^{i+1} - \hat{\beta}_n^i|/|\hat{\beta}_n^i - \hat{\beta}_n^{i-1}|$, $i \geq 1$. Ratios less than unity give informal support to the contraction mapping assumption.

We now develop the efficient estimator. We do so by starting from the consistent ILS estimates and taking Newton-Raphson steps using the criterion function for the efficient Klein and Spady (1993) estimator. Let $d = k - 1$ and reinterpret β_0 as the last d components of the true parameter vector and write $\hat{\beta}_{ILS}$ as the last d components of the ILS estimator. In short, we throw out the intercept term. We do this because the Klein and Spady estimator does not estimate the intercept. Let ϕ denote an element of $\mathbb{R}^d$. Define the sample NR mapping

$$\hat{M}_n(\phi) = \phi - \left[\hat{H}_n(\phi)\right]^{-1} \hat{G}_n(\phi)$$

where $\hat{G}_n(\phi)$ is the gradient and $\hat{H}_n(\phi)$ the Hessian or outer product gradient of the criterion function $\hat{L}_n(\phi)$ of the efficient Klein and Spady estimator as defined in Sherman (1994a). This function has the form

$$\hat{L}_n(\phi) = n^{-1} \sum_{i=1}^{n} \left[Y_i \ln \hat{F}(X_i'\phi, \phi) + (1 - Y_i) \ln \left[1 - \hat{F}(X_i'\phi, \phi)\right]\right] \tau_n(X_i)$$

where $\hat{F}(X_i'\phi, \phi)$ is a nonparametric regression estimator of $F(X_i'\phi, \phi)$ and $\tau_n(x) = \{|x| \leq c_n\}$. Define $\hat{\beta}_n^0 = \hat{\beta}_{ILS}$ and for $i \geq 1$, define $\hat{\beta}_n^i = \hat{M}_n(\hat{\beta}_n^{i-1})$. We call $\hat{\beta}_n^i$ a semiparametric NR estimator of β_0.

Define the population NR mapping

$$M(\phi) = \phi - [H(\beta_0)]^{-1} G(\phi)$$

where $G(\phi) = I\!E\tilde{G}_n(\phi)$ and $H(\phi) = I\!E\tilde{H}_n(\phi)$ with $\tilde{G}_n(\phi)$ the gradient and $\tilde{H}_n(\phi)$ the Hessian or outer product gradient of the function $\tilde{L}_n(\phi)$ where

$$\tilde{L}_n(\phi) = n^{-1} \sum_{i=1}^{n} \left[Y_i \ln F(X_i'\phi, \phi) + (1 - Y_i) \ln \left[1 - F(X_i'\phi, \phi) \right] \right] .$$

Note that $M(\phi)$ does not depend on n or ω, and $M(\beta_0) = \beta_0$.

Define $\hat{V}_n(\phi) = \nabla_\phi \hat{M}_n(\phi)$ and $V(\phi) = \nabla_\phi M(\phi)$ and note that $V(\beta_0) = 0_d$, the $d \times d$ zero matrix. Assumptions in Sherman (1994a) imply that $H(\phi)$ is continuous in a neighborhood of β_0. It follows that $V(\phi)$ is continuous in a neighborhood of β_0. Deduce that there exists an open ball centered at β_0 with closure B_0 such that $M(\phi)$ is a contraction mapping on (B_0, E_d). Arguments in Sherman (1994a) can be adapted to show that as $n \to \infty$, (i) $\sup_{\phi \in B_0} |\hat{M}_n(\phi) - M(\phi)| = o_p(1)$, (ii) $\sup_{\phi \in B_0} |\hat{V}_n(\phi) - V(\phi)| = o_p(1)$, and (iii) $\sqrt{n}(\hat{M}_n(\beta_0) - \beta_0) \Longrightarrow N(0, -[H(\beta_0)]^{-1})$. Deduce from these facts, Lemma 3, and Theorem 4 that for $i(n) \geq [\ln 2]^{-1} \ln(-.5 \ln n / \ln \kappa)$ for $\kappa \in (0, 1)$ (see Remark 1), $\sqrt{n}(\hat{\beta}_n^{i(n)} - \beta_0)$ converges in distribution to a $N(0, -[H(\beta_0)]^{-1})$ random variable as $n \to \infty$. This is the limiting distribution of the efficient Klein and Spady (1993) estimator.

Finally, we note that the ILS procedure developed in this section can be easily extended to semiparametric censored regression models. For example, if the latent variable $Y^* = X'\beta_0 - u$ and we observe (Y, X) where $Y = Y^* \mathbf{1}\{Y^* \geq 0\}$, then an ILS estimator of β_0 can be defined

exactly as in this section after setting $\hat{Y}_j(\phi) = Y_j \mathbf{1}\{Y_j > 0\} + [X_j'\phi - \hat{\pi}(X_j'\phi, \phi)]\mathbf{1}\{Y_j = 0\}$. Similar extensions can cover other censoring schemes such as those that involve top-coding.

4. Simulations

In this section, we provide a small simulation study comparing the speed and performance of the ILS procedure developed in Section 3 to the speed and performance of the efficient estimator of Klein and Spady (1993) for the semiparametric binary response model.

The model we use in the simulations has the form $Y = \mathbf{1}\{u \leq X'\beta_0\}$ where the regressor vector $X = (W_1, \ldots, W_6, W_7)'$ and

$$X'\beta_0 = \beta_{01}W_1 + \beta_{02}W_2 + \beta_{03}W_3 + \beta_{04}W_4 + \beta_{05}W_5 + \beta_{06}W_6 + \beta_{07}W_7$$

with $\beta_{01} = 0$, $\beta_{02} = -2$, $\beta_{03} = -1$, $\beta_{04} = -0.5$, $\beta_{05} = 0.5$, $\beta_{06} = 2$, and $\beta_{07} = 1$. In this model, $W_1 = 1$ and W_2 through W_7 are independent standard exponential random variables. In addition, u is independent of X and has either a $\chi^2(1)$ or a $\chi^2(3)$ distribution standardized to have mean zero and variance equal to the variance of $X'\beta_0$. Thus, the signal to noise ratio in the simulations is unity. We normalize on β_{07} and report estimates of the slope coefficients $\beta_{02}, \beta_{03}, \beta_{04}, \beta_{05}, \beta_{06}$.[4] Note that a $\chi^2(1)$ distribution is log-convex, whereas a $\chi^2(3)$ distribution is log-concave. Even though the $\chi^2(1)$ distribution violates log-concavity, we will show that the ILS procedure still works well in this setting.

For the ILS procedure, we choose the number of neighbors k_n by a "leave one out" method of least-squares cross-validation calibrated on one of the models defined above with $n = 10000$. This and A3 produce the rule $k_n \approx .14\, n^{3/4}$, which we use in all the simulations. We do no trimming. Also, we use OLS starting values and say that the procedure has converged when the

maximum relative difference between the components of successive iterates is less than 10^{-4} in absolute value. If this convergence criterion is not met after 1000 iterations, we stop and use the last five components of the 1000th iterate as an estimate of $(\beta_{02}, \ldots, \beta_{06})$.

We compute the Klein and Spady (KS) criterion function using a program written in the GAUSS language by Roger Klein. This program likewise does no trimming. We compute the KS estimator using the MAXLIK optimization routine with PROBIT starting values and say that the procedure has converged when the maximum component of the gradient is less than 10^{-4} in absolute value. If this criterion is not met after 50 iterations, we stop and use the 50th iterate as an estimate of $(\beta_{02}, \ldots, \beta_{06})$. Typically, the components of the 50th iterate are not changing in the first four decimal places.

Results for both the ILS and KS estimators are based on the same simulation data sets. All simulations are performed using GAUSS 3.5 for Windows on a Pentium-III PC with 800 Megahertz of RAM.

Table 1 presents means and root-mean squared error (RMSE) statistics for the ILS and KS estimators, based on 100 simulations. Results for the PROBIT and OLS estimators are also provided as points of comparison. For both error distributions, when $n = 1000$, both ILS and KS do well in terms of bias. KS slightly outperforms ILS in terms of RMSE for $\chi^2(1)$ errors; the two estimators are close in RMSE when errors are $\chi^2(3)$. When $n = 5000$, the estimators are almost indistinguishable in terms of both bias and RMSE. Note that both absolute and relative bias in PROBIT and OLS estimates increase as the parameter values increase in magnitude, the effect being more pronounced for the more skewed $\chi^2(1)$ distribution.

Table 1: Means and RMSE of Coefficient Estimates*

	$\beta_{02} = -2$	$\beta_{03} = -1$	$\beta_{04} = -.5$	$\beta_{05} = .5$	$\beta_{06} = 2$
		$u \sim \chi^2(1)$, $n = 1000$			
ILS ($k_n = 25$)	-1.97 (.31)	-1.01 (.19)	-.52 (.17)	.52 (.15)	2.03 (.29)
KS	-1.91 (.24)	-.98 (.14)	-.51 (.11)	.51 (.10)	2.03 (.20)
PROBIT	-1.62 (.66)	-.87 (.22)	-.45 (.13)	.50 (.13)	2.06 (.28)
OLS	-1.37 (.66)	-.82 (.22)	-.45 (.13)	.52 (.13)	1.83 (.28)
		$u \sim \chi^2(3)$, $n = 1000$			
ILS ($k_n = 25$)	-2.04 (.33)	-1.01 (.23)	-.50 (.18)	.51 (.16)	2.05 (.28)
KS	-1.99 (.32)	-1.00 (.23)	-.49 (.20)	.51 (.15)	2.06 (.26)
PROBIT	-1.75 (.53)	-.90 (.21)	-.46 (.17)	.50 (.15)	2.09 (.28)
OLS	-1.53 (.53)	-.88 (.21)	-.46 (.17)	.52 (.15)	1.88 (.28)
		$u \sim \chi^2(3)$, $n = 5000$			
ILS ($k_n = 85$)	-2.01 (.14)	-.99 (.08)	-.49 (.06)	.50 (.07)	2.01 (.12)
KS	-1.97 (.14)	-.99 (.09)	-.50 (.07)	.49 (.07)	2.02 (.11)
PROBIT	-1.71 (.52)	-.88 (.16)	-.45 (.07)	.48 (.07)	2.05 (.20)
OLS	-1.49 (.52)	-.85 (.16)	-.45 (.07)	.50 (.07)	1.83 (.20)

*Results based on 100 simulations.

Table 2 presents computation statistics that allow timing and convergence comparisons of the ILS and KS estimators. These statistics correspond to the results presented in Table 1. The first column gives the median number of iterations per simulation, the second column gives the approximate time per iteration, and the third column gives the percentage of simulations where the respective convergence criteria were satisfied. We see that for both error distributions, when $n = 1000$, most of the ILS simulations do not satisfy the ILS criterion. However, we see from Table 1 that this does not imply that the ILS iterates are diverging. Rather, the iterates oscillate between vectors whose components differ in the third or fourth decimal places. We suspect that this oscillation is due to lack of smoothness in $\hat{\beta}(\phi)$ when $n = 1000$. Note

that when $n = 5000$, the median number of ILS iterations decreases to 175 while nearly all the simulations satisfy the ILS criterion. The KS estimator, on the other hand, satisfies its convergence criterion most of the time, though contrary to our expectations, the percentage decreases as skewness decreases and as sample size increases.

A notable aspect of Table 2 is the timing comparison. ILS, on average, requires only about 20 seconds to estimate 6 parameters (5 slope coefficients and an intercept) when $n = 5000$. KS, on average, requires well over 2 hours to estimate 5 parameters when $n = 5000$. The ILS procedure is fast because (i) only $O(n)$ calculations are needed to compute the nearest neighbor estimators of all the $F(X_j'\phi, \phi)$'s and (ii) computation time for a least squares calculation is nearly constant as a function of k, the number of estimated parameters. By contrast, computing the KS criterion function, which involves local smoothing, is an $O(n^2 h_n)$ calculation where h_n is the deterministic bandwidth for the kernel regression estimators of the $F(X_j'\phi, \phi)$'s. In addition, computation time can increase as a quadratic in k, since steps can require the computation of numerical gradients and hessians of the criterion function.

Table 2: Computation Statistics*

	Median # of Iterations/Simulation	Time/Iteration	Criterion Satisfied
	$u \sim \chi^2(1)$, $n = 1000$		
ILS ($k_n = 25$)	1000	.02 seconds	3%
KS	22	40 seconds	94%
	$u \sim \chi^2(3)$, $n = 1000$		
ILS ($k_n = 25$)	1000	.02 seconds	1%
KS	22	40 seconds	83%
	$u \sim \chi^2(3)$, $n = 5000$		
ILS ($k_n = 85$)	175	.11 seconds	97%
KS	27	6 minutes	77%

*Results based on 100 simulations.

5. Summary

This paper develops general conditions for rates of convergence and convergence in distribution of iterative procedures for estimating finite-dimensional parameters. The theory covers iterative estimation schemes like expectation-maximization (EM), Newton-Raphson (NR), and iterative least squares (ILS) procedures. The theory requires a combination of asymptotic contraction mapping conditions and uniform convergence conditions, with convergence in distribution requiring an additional convergence in distribution result for a certain infeasible estimator. A bias condition is isolated that can be used to derive sensible stopping rules.

We illustrate the theory by establishing the limiting distribution of a two-stage iterative estimator of regression parameters in a semiparametric binary response model. The first stage is a consistent ILS procedure. The second stage is a NR procedure that is started at the ILS

estimates and is based on the criterion function of the efficient Klein and Spady (1993) estimator. The ILS/NR procedure achieves the semiparametric efficiency bound for this model established by Chamberlain (1986) and Cosslett (1987). Simulations show that the ILS procedure is very fast to compute even for models with many observations and many estimable parameters. In addition, the ILS estimator is comparable to the Klein and Spady estimator in terms of root-mean-squared error in the given simulations. The ILS procedure developed for the semiparametric binary response model easily extends to cover various semiparametric censored regression models.

NOTES

1. Recall that $M_n(\phi) = (M_n^1(\phi), \ldots, M_n^k(\phi))'$ where each $M_n^j(\phi)$ is a real-valued funtion of $\phi = (\phi_1, \ldots, \phi_k)'$. An NR procedure is smooth if, for each j, the $k \times k$ matrix of second order mixed partial derivatives of $M_n^j(\phi)$ exists.

2. Wang and Zhou (1995) appear to have been the first to propose a semiparametric ILS estimator for binary response models. They use isotonic regression to estimate a certain conditional expectation. They do not prove consistency. The theory developed in this paper covers the ILS estimator of Wang and Zhou. However, the contraction mapping condition and uniform convergence condition of Theorem 1 are harder to prove due to the difficulty in analyzing the isotonic regression component of their estimator.

3. Note that $\hat{Y}_j(\phi)$ is a function of the nearest neighbor estimators $\hat{F}(X_i'\phi, \phi)$, which are step functions. This implies that $\hat{Y}_j(\phi)$ (and, therefore, the least squares criterion function) is a discontinuous function of ϕ, violating Assumption 1a in the theory developed in Pastorello et alia (2003, p.452).

4. While the ILS procedure directly produces an intercept estimate, the Klein and Spady (KS) estimator does not. Consequently, we do not report intercept estimates.

REFERENCES

Aliprantis, C. D., and K. C. Border (1994): *Infinite Dimensional Analysis*, New York, Springer-Verlag.

Bickel, P. (1975): "One-Step Huber Estimates in the Linear Model," *Journal of the American Statistical Association*, 70, 428–434.

Burden, R. L., Faires, J. D. and A. C. Reynolds (1981): *Numerical Analysis*, PWS Publishers, Boston, MA.

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.

Dieudonné, J. (1969): *Foundations of Modern Analysis*. New York: Academic Press.

Heckman, J. J. and B. Honoré (1990): "The Empirical Content of the Roy Model," *Econometrica*, 58, 1121–1149.

Horowitz, J. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60, 505–531.

Ichimura, H. (1997): "Asymptotic Distribution of Non-parametric and Semiparametric Estimators with Data Dependent Smoothing Parameters," Manuscript, Department of Economics, University of Pittsburgh.

Klein, R. W. and R. H. Spady (1993): "An Efficient Semiparametric Estimator for Binary

Response Models," *Econometrica* 61, 387–421.

Lehmann, E. L. (1983): *Theory of Point Estimation*. New York: Wiley.

McLachlan, G. J. and T. Krishnan (1997): *The EM Algorithm and Extensions*. New York: Wiley.

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

Pastorello, S., Patilea, V. and E. Renault (2003): "Iterative and Recursive Estimation in Structural Non-Adaptive Models," *Journal of Business and Economic Statistics*, 21, 449–509.

Robinson, P. M. (1988): "The Stochastic Difference Between Econometric Estimators," *Econometrica*, 56, 531–548.

Sherman, R. P. (1994a): "U-processes in the Analysis of a Generalized Semiparametric Regression Estimator," *Econometric Theory*, 10, 372–395.

Sherman, R. P. (1994b): "Maximal Inequalities for Degenerate U-processes with Applications to Optimization Estimators," *Annals of Statistics*, 22, 439–459.

Wang, W. and M. Zhou (1995): "Iterative Least Squares Estimator of Binary Choice Models: A Semiparametric Approach," University of Kentucky College of Business and Economics Working Paper.

APPENDIX

Proof of Theorem 1. By (i), there exist a constant c in $[0,1)$ which does not depend on n or ω, and sets $\{A_n\}$ with each $A_n \subseteq \Omega$ and $I\!P A_n \to 1$ as $n \to \infty$, such that for each $\omega \in A_n$, $|M_n(\phi) - M_n(\gamma)| \leq c|\phi - \gamma|$ for each ϕ, γ in B_0. Let $\rho_0 = |\beta_n^0 - \beta_0|$. By (iii), there exist sets $\{B_n\}$

with each $B_n \subseteq \Omega$ and $\mathbb{P}B_n \to 1$ as $n \to \infty$, such that for each $\omega \in B_n$ for all n large enough, $n^\delta \sup_{\phi \in B_0} |\hat{M}_n(\phi) - M_n(\phi)|$ is bounded and $\sup_{\phi \in B_0} |\hat{M}_n(\phi) - M_n(\phi)|$ is arbitrarily small, in particular, smaller than $1 - c\rho_0$. Let $C_n = A_n \bigcap B_n$ and note that $\mathbb{P}C_n \to 1$ as $n \to \infty$. It follows that for each $\omega \in C_n$ and for each γ in B_0,

$$\begin{aligned} |\hat{M}_n(\gamma) - \beta_0| &\leq |M_n(\gamma) - M_n(\beta_0)| + |\hat{M}_n(\gamma) - M_n(\gamma)| \\ &\leq c\rho_0 + \sup_{\phi \in B_0} |\hat{M}_n(\phi) - M_n(\phi)| < 1 \,. \end{aligned}$$

Deduce that for each $\omega \in C_n$, $\hat{M}_n$ maps B_0 to itself.

Next, note that $|\hat{\beta}_n^{i(n)} - \beta_0|$ can be bounded by a bias term plus a stochastic term. That is,

$$|\hat{\beta}_n^{i(n)} - \beta_0| \leq |\beta_n^{i(n)} - \beta_0| + |\hat{\beta}_n^{i(n)} - \beta_n^{i(n)}| \,.$$

By (ii), the bias term has order $O_p(n^{-\delta})$ as $n \to \infty$. Consider the stochastic term. Recall that $\hat{\beta}_n^0 = \beta^0$ and that for each $\omega \in C_n$, $\hat{M}_n$ maps B_0 to itself. It follows that for each $\omega \in C_n$, $\hat{\beta}_n^{i(n)} \in B_0$. Thus, for each $\omega \in C_n$,

$$\begin{aligned} |\hat{\beta}_n^{i(n)} - \beta_n^{i(n)}| &= |\hat{M}_n(\hat{\beta}_n^{i(n)-1}) - M_n(\beta_n^{i(n)-1})| \\ &\leq |\hat{M}_n(\hat{\beta}_n^{i(n)-1}) - M_n(\hat{\beta}_n^{i(n)-1})| + |M_n(\hat{\beta}_n^{i(n)-1}) - M_n(\beta_n^{i(n)-1})| \\ &\leq \sup_{\phi \in B_0} |\hat{M}_n(\phi) - M_n(\phi)| + c\,|\hat{\beta}_n^{i(n)-1} - \beta_n^{i(n)-1}| \,. \end{aligned}$$

Apply the last inequality recursively to see that for each $\omega \in C_n$,

$$\begin{aligned} |\hat{\beta}_n^{i(n)} - \beta_n^{i(n)}| &\leq \sup_{\phi \in B_0} |\hat{M}_n(\phi) - M_n(\phi)|[1 + c + c^2 + \cdots + c^{i(n)-1}] \\ &\leq \sup_{\phi \in B_0} |\hat{M}_n(\phi) - M_n(\phi)|[1/(1-c)] \,. \end{aligned}$$

Condition (iii) implies that the stochastic term has order $O_p(n^{-\delta})$, which proves the result. □

Proof of Theorem 2. This involves a trivial modification of the proof of Theorem 1.□

PROOF OF LEMMA 3. Fix ϕ and γ in B_0. Since B_0 is convex, we may apply a multivariate Taylor expansion (e.g., Dieudonné, 1969, p.190) to write $M_n(\phi)-M_n(\gamma) = \Lambda_n(\phi,\gamma)[\phi-\gamma]$ where $\Lambda_n(\phi,\gamma) = \int_0^1 V_n(\gamma+\xi(\phi-\gamma))d\xi$. Similarly, we may write $\hat{M}_n(\phi) - \hat{M}_n(\gamma) = \hat{\Lambda}_n(\phi,\gamma)[\phi-\gamma]$ where $\hat{\Lambda}_n(\phi,\gamma) = \int_0^1 \hat{V}_n(\gamma+\xi(\phi-\gamma))d\xi$. Use (i) and (ii) and argue as in the proof of Theorem 1 that there exist sets $\{C_n\}$ with each $C_n \subseteq \Omega$ and $I\!PC_n \to 1$ as $n \to \infty$, such that for each $\omega \in C_n$, $\hat{M}$ maps B_0 to itself, and $M_n(\phi)$ is a contraction mapping on (B_0, E_k) with fixed point β_0 and modulus of contraction $c \in [0,1)$ independent of n and ω. Then for each $\omega \in C_n$ and for each ϕ, γ in B_0,

$$\begin{aligned} |\hat{M}_n(\phi) - \hat{M}_n(\gamma)| &\leq |M_n(\phi) - M_n(\gamma)| + |[\hat{M}_n(\phi) - \hat{M}_n(\gamma)] - [M_n(\phi) - M_n(\gamma)]| \\ &\leq c|\phi-\gamma| + |[\hat{\Lambda}_n(\phi,\gamma) - \Lambda_n(\phi,\gamma)][\phi-\gamma]|\,. \end{aligned}$$

By (iii), the last term has order $o_p(|\phi-\gamma|)$ as $n \to \infty$, from which the result follows. □

PROOF OF THEOREM 4. By (ii), there exist sets $\{A_n\}$ with each $A_n \subseteq \Omega$ and $I\!PA_n \to 1$ as $n \to \infty$ such that for each $\omega \in A_n$, $\hat{\beta}_n$ is a fixed point of $\hat{M}_n(\phi)$. For $\omega \in A_n$, write $\hat{\beta}_n^{i(n)} - \beta_0$ as $\hat{\beta}_n^{i(n)} - \hat{\beta}_n + \hat{\beta}_n - \beta_0$. By (iii), $|\hat{\beta}_n^{i(n)} - \hat{\beta}_n|$ has order $o_p(n^{-\delta})$ as $n \to \infty$. The result will follow if, as $n \to \infty$,

$$n^{\delta}(\hat{\beta}_n - \beta_0) \Longrightarrow DZ\,.$$

Since $\hat{\beta}_n$ is a fixed point of $\hat{M}_n(\phi)$, $\hat{M}_n(\hat{\beta}_n) = \hat{\beta}_n$. Thus, for each $\omega \in A_n$,

$$\hat{\beta}_n - \beta_0 = \hat{M}_n(\hat{\beta}_n) - \hat{M}_n(\beta_0) + \hat{M}_n(\beta_0) - \beta_0\,.$$

Expand $\hat{M}_n(\hat{\beta}_n)$ about β_0 to get

$$\hat{M}_n(\hat{\beta}_n) - \hat{M}_n(\beta_0) = \hat{V}_n(\beta_n^*)[\hat{\beta}_n - \beta_0]$$

where β_n^* is on the line segment connecting $\hat{\beta}_n$ and β_0. Combine the last two expressions to get

$$n^\delta(\hat{\beta}_n - \beta_0) = [I_k - \hat{V}_n(\beta_n^*)]^{-1} n^\delta(\hat{M}_n(\beta_0) - \beta_0)\,.$$

Note that

$$\hat{V}_n(\beta_n^*) = V(\beta_0) + [\hat{V}_n(\beta_n^*) - V(\beta_n^*)] + [V(\beta_n^*) - V(\beta_0)]\,.$$

Conditions (i) and (ii) imply that $|\hat{\beta}_n - \beta_0| = o_p(1)$ as $n \to \infty$. This and the definition of β_n^* imply that $|\beta_n^* - \beta_0| = o_p(1)$ as $n \to \infty$. This and condition (v) imply that the first term in brackets has order $o_p(1)$ as $n \to \infty$. Condition (vi) and the fact that $|\beta_n^* - \beta_0| = o_p(1)$ as $n \to \infty$ imply that the second term in brackets has order $o_p(1)$ as $n \to \infty$. This and condition (iv) imply the result. □

Next, we prove several lemmas used in the proof of consistency of the semiparametric ILS estimator. The first is an extension of a result of Ichimura (1997, Lemma 4.1, p.29) relating a nearest neighbor window width to the number of neighbors.

Recall that $\mathcal{N}$ is a neighborhood of β_0, and for $\phi \in \mathcal{N}$, $g(t,\phi)$ is the density of $X'\phi$ evaluated at $t \in I\!R$. Define $c(t,\phi) = [\int_t^\infty g(v,\phi)dv]^{-1}$ and $d(t,\phi) = [\int_{-\infty}^t g(v,\phi)dv]^{-1}$. Note that for all $t \in I\!R$, $c(t,\phi) \geq 1$ and $d(t,\phi) \geq 1$. In addition, by A10, both $c(t,\phi)$ and $d(t,\phi)$ are continuous in t and ϕ. Define $a_n(t,\phi)$ to be the distance from t to its k_nth nearest neighbor to the right and $b_n(t,\phi)$ the distance from t to its k_nth nearest neighbor to the left. Recall that the trimming function $\sigma_n(t) = \{|t| \leq c_n^\alpha\}$. Fix $\mu > 0$. Define S_n to be the interval $[-c_n^\alpha, c_n^\alpha]$, L_n the lower interval $[-c_n^\alpha, -\mu]$, M the middle interval $[-\mu, \mu]$, and U_n the upper interval $[\mu, c_n^\alpha]$.

LEMMA 1A: *Suppose A2, A3, A5, A9, and A10 hold. Then, as $n \to \infty$,*

$$\sup_{M \times \mathcal{N}} |r_n(t,\phi) - 1| = o_p(1)$$

where either $r_n(t,\phi) = [nc(t,\phi)g(t,\phi)a_n(t,\phi)]/k_n$ *or* $r_n(t,\phi) = [nd(t,\phi)g(t,\phi)b_n(t,\phi)]/k_n$.

PROOF. We shall prove the result for the case $r_n(t,\phi) = [nc(t,\phi)g(t,\phi)a_n(t,\phi)]/k_n$. The proof for the other case is similar.

Let $H(\cdot,t,\phi)$ denote the cdf of $(X'\phi - t)^+$. That is, for $s \geq 0$,

$$H(s,t,\phi) = I\!P\{(X'\phi - t)^+ \leq s\} = c(t,\phi)\int_t^{t+s} g(v,\phi)dv\,.$$

Let $H_n(s,t,\phi)$ denote the corresponding empirical cdf. That is, for $s \geq 0$,

$$H_n(s,t,\phi) = n^{-1}\sum_{i=1}^n \mathbf{1}\{(X_i'\phi - t)^+ \leq s\}\,.$$

Note that $H_n(a_n(t,\phi),t,\phi) = k_n/n$ and so

$$H(a_n(t,\phi),t,\phi) = k_n/n + [H(a_n(t,\phi),t,\phi) - H_n(a_n(t,\phi),t,\phi)]\,.$$

Standard empirical process results (e.g., Pakes and Pollard, 1989) imply that as $n \to \infty$,

$$\sup_{I\!R^+ \times I\!R \times \mathcal{N}} |H(s,t,\phi) - H_n(s,t,\phi)| = O_p(n^{-1/2})\,.$$

Thus, uniformly over $I\!R \times \mathcal{N}$,

$$H(a_n(t,\phi),t,\phi) = k_n/n + O_p(n^{-1/2})\,.$$

This, and a Taylor expansion of H about $s = 0$ imply that uniformly over $I\!R \times \mathcal{N}$,

$$a_n(t,\phi) = k_n/[nc(t,\phi)g(t+s^*,\phi)] + O_p(1/[n^{1/2}c(t,\phi)g(t+s^*,\phi)]) \tag{1}$$

where s^* is between zero and $a_n(t,\phi)$. Since $k_n \gg n^{1/2}$, we get from (1) that uniformly over $I\!R \times \mathcal{N}$,

$$[nc(t,\phi)g(t+s^*,\phi)a_n(t,\phi)]/k_n = 1 + o_p(1)\,. \tag{2}$$

The continuity of $g(t,\phi)$ implies that $\mathbb{P}\{t \leq X'\phi \leq t+1\}$ is a continuous function of t and ϕ on the compact set $M \times \mathcal{N}$. Thus, it achieves its minimum on this set. Moreover, this minimum must be positive, since the support of $X'\phi$ is $\mathbb{R}$ for each ϕ in $\mathcal{N}$. Deduce from a standard uniform law of large numbers that $wp \to 1$ as $n \to \infty$, there exists a positive number p independent of $t \in M$ and $\phi \in \mathcal{N}$, such that there are at least pn points in $[t, t+1]$ for each $t \in M$ and $\phi \in \mathcal{N}$. Since $k_n \ll n$, $wp \to 1$ as $n \to \infty$, $a_n(t,\phi)$ (and therefore, s^*) must be in the interval $[0,1]$ for each t in M and ϕ in $\mathcal{N}$. Since $n^{1/2} \ll k_n \ll n$ and $1/[c(t,\phi)g(t,\phi)]$ is bounded over $M \times \mathcal{N}$, it follows from (1) that $a_n(t,\phi) = O_p(k_n/n) = o_p(1)$ uniformly over $M \times \mathcal{N}$. The left-hand side of (2) equals

$$\frac{na_n(t,\phi)c(t,\phi)}{k_n}\left[g(t,\phi) + [g(t+s^*,\phi) - g(t,\phi)]\right] . \tag{3}$$

The result follows from (3), $a_n(t,\phi) = O_p(k_n/n)$, and a Taylor expansion of $g(t+s^*,\phi)$ about $s=0$ together with the uniform convergence of s^* to zero and $\sup_{t,\phi}|\nabla_t g(t,\phi)| < \infty$. □

LEMMA 2A: *Suppose A2, A3, A5, A9, and A10 hold. Then*

(i) $\sup_{L_n \times \mathcal{N}} |r_n(t,\phi) - 1| = o_p(1)$ *as* $n \to \infty$ *where* $r_n(t,\phi) = [nc(t,\phi)g(t,\phi)a_n(t,\phi)]/k_n$.

(ii) $\sup_{U_n \times \mathcal{N}} |r_n(t,\phi) - 1| = o_p(1)$ *as* $n \to \infty$ *where* $r_n(t,\phi) = [nd(t,\phi)g(t,\phi)b_n(t,\phi)]/k_n$.

PROOF. We shall prove (i). The proof of (ii) is similar.

To establish (i), note that an argument similar to the one given in the proof of Lemma 1A shows that $wp \to 1$ as $n \to \infty$, $a_n(t,\phi)$ (and therefore, s^*) must be in the interval $[0, c_n^{\alpha} - \mu + 1]$ for each t in L_n and ϕ in $\mathcal{N}$. Fix $\delta \in (0, 1/8)$. Since $n^{1/2} \ll k_n \ll n^{1-\delta}$ and $1/[c(t,\phi)g(t,\phi)] \ll n^{\delta}$ uniformly over $L_n \times \mathcal{N}$, it follows from (1) that $a_n(t,\phi) = o_p(k_n/n^{1-2\delta})$ uniformly over $L_n \times \mathcal{N}$.

Result (i) now follows from (2) and (3), a Taylor expansion of $g(t+s^*,\phi)$ about $s=0$, the fact that $s^* = o_p(k_n/n^{1-\delta})$ uniformly over $L_n \times \mathcal{N}$, and $\sup_{t,\phi} |\nabla_t g(t,\phi)| < \infty$. □

Recall that we use three types of nearest neighbor estimators of $F(t,\phi)$: symmetric, asymmetric from the right, and asymmetric from the left. The nearest neighbor estimator that is asymmetric from the right has the form

$$\hat{F}_r(t,\phi) = \frac{\sum_{j=1}^n Y_j \mathbf{1}\{0 \le X_j'\phi - t \le a_n(t,\phi)\}}{\sum_{j=1}^n \mathbf{1}\{0 \le X_j'\phi - t \le a_n(t,\phi)\}}$$

whereas the the nearest neighbor estimator that is asymmetric from the left has the form

$$\hat{F}_l(t,\phi) = \frac{\sum_{j=1}^n Y_j \mathbf{1}\{-b_n(t,\phi) \le X_j'\phi - t \le 0\}}{\sum_{j=1}^n \mathbf{1}\{-b_n(t,\phi) \le X_j'\phi - t \le 0\}}\,.$$

The symmetric nearest neighbor estimator has the form

$$\hat{F}_s(t,\phi) \;=\; \frac{\sum_{j=1}^n Y_j \mathbf{1}\{-b_n(t,\phi) \le X_j'\phi - t \le a_n(t,\phi)\}}{\sum_{j=1}^n \mathbf{1}\{-b_n(t,\phi) \le X_j'\phi - t \le a_n(t,\phi)\}}\,.$$

We see that if $t \neq X_j'\phi$ for any j, $\hat{F}_s(t,\phi)$ is an arithmetic average of $\hat{F}_r(t,\phi)$ and $\hat{F}_l(t,\phi)$. We shall establish rates of uniform consistency for $\hat{F}_r(t,\phi)$. The same results with similar proofs can be obtained for the other types. Note that we can write

$$\hat{F}_r(t,\phi) = (na_n(t,\phi))^{-1} \sum_{j=1}^n Y_j \mathbf{1}\{0 \le X_j'\phi - t \le a_n(t,\phi)\}/\hat{g}(t,\phi)$$

where

$$\hat{g}(t,\phi) = (na_n(t,\phi))^{-1} \sum_{j=1}^n \mathbf{1}\{0 \le X_j'\phi - t \le a_n(t,\phi)\}\,.$$

Our next result gives a rate of uniform convergence of $\hat{g}(t,\phi)$ to $g(t,\phi)$ and $\hat{F}_r(t,\phi)$ to $F(t,\phi)$.

LEMMA 3A: *Suppose A2, A3, A5, A9, and A10 hold. Then, as $n \to \infty$, for all $\delta > 0$,*

(i) $n^{1/4-\delta} \sup_{S_n \times \mathcal{N}} |\hat{g}(t,\phi) - g(t,\phi)| = o_p(1)$.

(ii) $n^{1/4-\delta} \sup_{S_n \times \mathcal{N}} \left| \hat{F}_r(t,\phi) - F(t,\phi) \right| = o_p(1)$.

PROOF. We first prove (i). Decompose $\hat{g}(t,\phi) - g(t,\phi)$ into a sum of a stochastic term and a bias term:

$$[\hat{g}(t,\phi) - I\!E\hat{g}(t,\phi)] + [I\!E\hat{g}(t,\phi) - g(t,\phi)] \; . \tag{4}$$

Let $\{\epsilon_n\}$ denote an arbitrary sequence of nonnegative real numbers satisfying $\epsilon_n \to 0$ as $n \to \infty$. By Lemma 1A and Lemma 2A, $wp \to 1$ as $n \to \infty$, for $t \in S_n$, $\phi \in \mathcal{N}$, and a satisfying $|nc(t,\phi)g(t,\phi)a/k_n - 1| \leq \epsilon_n$, the first term in (4) is bounded by

$$\sup_{t,\phi,a} a^{-1} \left| n^{-1} \sum_{j=1}^{n} \mathbf{1}\{0 \leq X_j'\phi - t \leq a\} - I\!P\{0 \leq X'\phi - t \leq a\} \right| . \tag{5}$$

Standard empirical process results (e.g., Pakes and Pollard, 1989) show that uniformly over $t \in I\!R$, $\phi \in I\!R^k$, and $a \geq 0$, the term in absolute value signs in (5) has order $O_p(n^{-1/2})$ as $n \to \infty$. Deduce from this, A3, A5, and the condition on a, that for all $\delta > 0$, the term in (5) has order $o_p(n^{-1/4+\delta})$ as $n \to \infty$.

Next, consider the bias term in (4). This term, $wp \to 1$ as $n \to \infty$, for $t \in S_n$, $\phi \in \mathcal{N}$, and a satisfying $|nc(t,\phi)g(t,\phi)a/k_n - 1| \leq \epsilon_n$, is bounded by

$$\sup_{t,\phi,a} \left| a^{-1} I\!P\{0 \leq X'\phi - t \leq a\} - g(t,\phi) \right| . \tag{6}$$

After the change of variable $z = (v-t)/a$,

$$\begin{aligned} a^{-1} I\!P\{0 \leq X'\phi - t \leq a\} &= a^{-1} \int_{-\infty}^{\infty} \mathbf{1}\{0 \leq v - t \leq a\} g(v,\phi) dv \\ &= \int_0^1 g(t + az, \phi) dz \, . \end{aligned}$$

A Taylor expansion of $g(t+az,\phi)$ about t, together with the bounded derivative in A9, A3, A5, and the condition on a, imply that for each $\delta > 0$, the term in (6) has order $o_p(n^{-1/4+\delta})$ as $n \to \infty$. This proves (i).

Mimic the proof of (i) to prove that

$$n^{1/4-\delta} \sup_{S_n \times \mathcal{N}} \left| \hat{F}_r(t,\phi)\hat{g}(t,\phi) - F(t,\phi)g(t,\phi) \right| = o_p(1)\,. \tag{7}$$

To prove (ii), note that

$$\hat{F}_r(t,\phi) = \frac{\hat{F}_r(t,\phi)\hat{g}(t,\phi)}{g(t,\phi)} \left[1 - \frac{\hat{g}(t,\phi) - g(t,\phi)}{g(t,\phi) + [\hat{g}(t,\phi) - g(t,\phi)]} \right] .$$

Apply (i), (7), and A5 to get the result. □

LEMMA 4A: *Let P be a probability distribution and $V_1, \ldots, V_n$ a sample of independent observations from P. Let λ_n be the infimum of the density of P over $[-\kappa_n, \kappa_n]$, $0 < \kappa_n < \infty$, and suppose $\lambda_n > 0$. For $\alpha_n > 0$, partition $[-\kappa_n, \kappa_n]$ into $N = (2\kappa_n n)/\alpha_n$ intervals of length α_n/n. For $i = 1, \ldots, N$, let A_i be the event that the ith interval contains at least one sample point. As $n \to \infty$,*

(i) *If $\kappa_n \to C < \infty$, then $\lambda_n \to c > 0$ and $I\!P\{\cap_{i=1}^N A_i\} \to 1$ provided $c\alpha_n > \log n$.*

(ii) *If $\kappa_n \to \infty$, then $\lambda_n \to 0$ and $I\!P\{\cap_{i=1}^N A_i\} \to 1$ provided $\kappa_n \ll \alpha_n$ and $\lambda_n \alpha_n > \log n$.*

PROOF. By Bonferroni's inequality,

$$\begin{aligned} I\!P\{\cap_{i=1}^N A_i\} &\geq 1 - \sum_{i=1}^N I\!P A_i^c \\ &\geq 1 - \sum_{i=1}^N (1 - (\lambda_n \alpha_n)/n)^n \\ &= 1 - N(1 - (\lambda_n \alpha_n)/n)^n \,. \end{aligned}$$

Both (i) and (ii) now follow from simple calculus. □

Lemma 5A is used in the proof of Lemma 5. It is proved in a more general form in Klein and Spady (1993). For ease of reference, we restate the result in the form in which we use it in the proof of Lemma 5.

LEMMA 5A: *Suppose $f(\cdot)$ is bounded and $\mathbb{E}|X| < \infty$. Then*

$$\left[\nabla_\phi F(X'\phi, \phi)\right]_{\beta_0} = f(X'\beta_0)\left[X - \mathbb{E}[X \mid X'\beta_0]\right] .$$

PROOF OF LEMMA 5. Note that $V_n(\phi) = \nabla_\phi M_n(\phi)$ is an average of iid random variables. By A11 and Lemma 2.13 in Pakes and Pollard (1989), $V_n(\phi)$ converges uniformly to $\mathbb{E}V_n(\phi)$ on $\mathcal{N}$. Moreover, A11 implies that $\mathbb{E}V_n(\phi)$ is continuous on $\mathcal{N}$.

Fix ϕ and γ in $\mathcal{N}$. By a multivariate Taylor expansion, there exists a ϕ^* on the line segment between ϕ and γ such that as $n \to \infty$,

$$\begin{aligned} |M_n(\phi) - M_n(\gamma)| &= |V_n(\phi^*)[\phi - \gamma]| \\ &\leq |V_n(\beta_0)[\phi - \gamma]| + |[\mathbb{E}V_n(\phi^*) - \mathbb{E}V_n(\beta_0)][\phi - \gamma]| + o_p(|\phi - \gamma|) . \end{aligned}$$

By the continuity of $\mathbb{E}V_n(\phi)$ on $\mathcal{N}$ and a standard linear algebra result, it is enough to show that $wp \to 1$ as $n \to \infty$, the largest eigenvalue of $V_n(\beta_0)$ in absolute value is strictly less than unity. To this end, recall that

$$M_n(\phi) = (\phi_1 - n^{-1}\sum_{j=1}^n X_{j1}u_j(\phi), \ldots, \phi_k - n^{-1}\sum_{j=1}^n X_{jk}u_j(\phi))'$$

where

$$u_j(\phi) = \left[F(X_j'\beta_0)\nu(X_j'\phi, \phi) + [1 - F(X_j'\beta_0)]\pi(X_j'\phi, \phi)\right] .$$

Write ∇_c for $\frac{\partial}{\partial \phi_c}$, $c = 1, 2, \ldots, k$. The rcth element of $V_n(\beta_0)$ equals

$$\mathbf{1}\{r = c\} - n^{-1}\sum_{j=1}^n X_{jr}\left[F(X_j'\beta_0)\left[\nabla_c \nu(X_j'\phi, \phi)\right]_{\beta_0} + [1 - F(X_j'\beta_0)]\left[\nabla_c \pi(X_j'\phi, \phi)\right]_{\beta_0}\right] . \quad (8)$$

Recall that $X = (W_1, \ldots, W_k, W_{k+1})$ and $W_1 = 1$. Integrate by parts, then differentiate and apply Lemma 5A to get that $\left[\nabla_c \nu(X_j'\phi, \phi)\right]_{\beta_0}$ equals

$$\frac{\int_{-\infty}^{X_j'\beta_0} F(u)\,du\, f(X_j'\beta_0)\left[W_{jc} - I\!E[W_{jc} \mid X_j'\beta_0]\right] + F(X_j'\beta_0)\int_{-\infty}^{X_j'\beta_0} f(u) I\!E[W_c \mid X'\beta_0 = u]\,du}{F^2(X_j'\beta_0)}.$$

Similarly, $\left[\nabla_c \pi(X_j'\phi, \phi)\right]_{\beta_0}$ equals

$$\frac{\int_{X_j'\beta_0}^{\infty}[1 - F(u)]\,du\, f(X_j'\beta_0)\left[W_{jc} - I\!E[W_{jc} \mid X_j'\beta_0]\right] + [1 - F(X_j'\beta_0)]\int_{X_j'\beta_0}^{\infty} f(u) I\!E[W_c \mid X'\beta_0 = u]\,du}{[1 - F(X_j'\beta_0)]^2}.$$

Integration by parts arguments show that

$$\begin{aligned}
\nu_1(X_j'\beta_0) &= \frac{\int_{-\infty}^{X_j'\beta_0} F(u)\,du\, f(X_j'\beta_0)}{F^2(X_j'\beta_0)} \\
\pi_1(X_j'\beta_0) &= \frac{\int_{X_j'\beta_0}^{\infty}[1 - F(u)]\,du\, f(X_j'\beta_0)}{[1 - F(X_j'\beta_0)]^2}.
\end{aligned}$$

Write $d(X_j'\beta_0)$ for $F(X_j'\beta_0)\nu_1(X_j'\beta_0) + [1 - F(X_j'\beta_0)]\pi_1(X_j'\beta_0)$. The term in outer brackets in (8) is equal to

$$(W_{jc} - I\!E[W_{jc} \mid X_j'\beta_0])d(X_j'\beta_0) + \kappa_c$$

where

$$\kappa_c = \int_{-\infty}^{\infty} f(u) I\!E[W_c \mid X'\beta_0 = u]\,du. \tag{9}$$

This and prewhitening let us write the rcth element of $V_n(\beta_0)$ as

$$n^{-1}\sum_{j=1}^{n} W_{jr}\left[W_{jc} - (W_{jc} - I\!E[W_{jc} \mid X_j'\beta_0])d(X_j'\beta_0) - \kappa_c\right] \tag{10}$$

Since $W_1 = 1$, $\kappa_1 = 1$. Prewhitening implies that $W_{j1} = 1$ for all j. Deduce that for $r = 1, \ldots, k$ and $c = 1$, the rcth element of $V_n(\beta_0)$ equals zero. That is, the first column of $V_n(\beta_0)$ is a vector of zeros. This implies that one solution of the characteristic equation of $V_n(\beta_0)$ must be zero. Thus, to prove that $wp \to 1$ as $n \to \infty$, the maximum eigenvalue of $V_n(\beta_0)$ in absolute value

is strictly less that unity, it is enough to prove this for the $(k-1)\times(k-1)$ lower right-hand submatrix of $V_n(\beta_0)$. For convenience, we call this submatrix A_n.

Prewhitening implies that $\sum_{j=1}^n W_{jr} = 0$ for $r > 1$. Deduce from this and (10) that the rcth element of A_n equals

$$n^{-1}\sum_{j=1}^{n} W_{jr}\left[W_{jc} - (W_{jc} - I\!E[W_{jc} \mid X_j'\beta_0])d(X_j'\beta_0)\right] .$$

Let $\overline{\mathbf{X}}_n$ denote the $n\times(k-1)$ matrix comprised of the second through kth components of each regressor vector, and let $\mathbf{X}\beta_0$ denote the $n\times 1$ vector with jth component $X_j'\beta_0$. Define $T_n = n^{-1/2}\overline{\mathbf{X}}_n$ and $S_n = n^{-1/2}\left[\overline{\mathbf{X}}_n - I\!E[\overline{\mathbf{X}}_n \mid \mathbf{X}\beta_0]\right]$. Also, let D_n denote the $n\times n$ diagonal matrix with jjth element $d(X_j'\beta_0)$. We see that

$$A_n = T_n'T_n - T_n'D_nS_n\,.$$

By prewhitening, $T_n'T_n = I_{k-1} = T_n'I_nT_n$. The fact that $T_n'D_nS_n = S_n'D_nS_n + o_p(1)$ as $n\to\infty$ implies that $wp\to 1$ as $n\to\infty$,

$$\begin{aligned} A_n &= T_n'I_nT_n - S_n'D_nS_n \\ &= T_n'[I_n - D_n]T_n + T_n'D_nT_n - S_n'D_nS_n \\ &= T_n'[I_n - D_n]T_n + [T_n - S_n]'D_n[T_n - S_n] \\ &= W_n'W_n + Z_n'Z_n \end{aligned}$$

where $W_n = [I_n - D_n]^{1/2}T_n$ and $Z_n = D_n^{1/2}[T_n - S_n]$. (Note that log-concavity of u and Proposition 1 in Heckman and Honoré (1990) imply that the diagonal elements of D_n are in $[0,1]$, making it possible to form $D_n^{1/2}$ and $[I_n - D_n]^{1/2}$.) Thus, $wp\to 1$ as $n\to\infty$, A_n is a nonnegative definite matrix and so must have all nonnegative eigenvalues. Recall from above

that $wp \to 1$ as $n \to \infty$,

$$\begin{aligned} A_n &= I_{k-1} - S_n' D_n S_n \\ &= I_{k-1} - \Sigma_n' \Sigma_n \end{aligned}$$

where $\Sigma_n = D_n^{1/2} S_n$. It follows that $wp \to 1$ as $n \to \infty$, the maximum eigenvalue of A_n is equal to

$$\max_{|x|=1} x' A_n x = \max_{|x|=1} \left[1 - x' \Sigma_n' \Sigma_n x\right] .$$

A1 and A2 imply that $I\!P\{u \in S_u \cap S_\phi\} = 1$. Thus, $x' \Sigma_n' \Sigma_n x$ is positive and bounded away from zero for all unit vectors $x \in I\!R^{k-1}$. Thus, $wp \to 1$ as $n \to \infty$, the maximum eigenvalue of A_n is strictly less than unity. This proves the result. □

PROOF OF LEMMA 6. Note that $|\hat{M}_n(\phi) - M_n(\phi)|$ is bounded by

$$\left| n^{-1} \mathbf{X}_n' \left[\hat{\mathbf{u}}_n(\phi) - \tilde{\mathbf{u}}_n(\phi)\right] \right| + \left| n^{-1} \mathbf{X}_n' \left[\tilde{\mathbf{u}}_n(\phi) - \bar{\mathbf{u}}_n(\phi)\right] \right| + \left| n^{-1} \mathbf{X}_n' \left[\bar{\mathbf{u}}_n(\phi) - \mathbf{u}_n(\phi)\right] \right| \tag{11}$$

where $\hat{\mathbf{u}}_n(\phi) = (\hat{u}_1(\phi), \ldots, \hat{u}_n(\phi))'$ with $\hat{u}_j(\phi) = \left[Y_j \hat{\nu}(X_j'\phi, \phi) + (1 - Y_j)\hat{\pi}(X_j'\phi, \phi)\right] \tau_n(X_j'\phi)$, $\tilde{\mathbf{u}}_n(\phi) = (\tilde{u}_1(\phi), \ldots, \tilde{u}_n(\phi))'$ with $\tilde{u}_j(\phi) = \left[Y_j \nu(X_j'\phi, \phi) + (1 - Y_j)\pi(X_j'\phi, \phi)\right] \tau_n(X_j'\phi)$, $\bar{\mathbf{u}}_n(\phi) = (\bar{u}_1(\phi), \ldots, \bar{u}_n(\phi))'$ with $\bar{u}_j(\phi) = \left[F(X_j'\beta_0)\nu(X_j'\phi, \phi) + (1 - F(X_j'\beta_0))\pi(X_j'\phi, \phi)\right] \tau_n(X_j'\phi)$, and $\mathbf{u}_n(\phi) = (u_1(\phi), \ldots, u_n(\phi))'$ with $u_j(\phi) = \left[F(X_j'\beta_0)\nu(X_j'\phi, \phi) + (1 - F(X_j'\beta_0))\pi(X_j'\phi, \phi)\right]$.

Each component of the second term in (11) is an average of zero mean iid random variables. Deduce from this, A11, $I\!E|X|^2 < \infty$, and Lemma 2.13 in Pakes and Pollard (1989) that this term has order $O_p(n^{-1/2})$ uniformly over $\mathcal{N}$.

Each component of the third term in (11) can be decomposed into its mean plus a term that is an average of zero mean iid random variables. This latter term has order $O_p(n^{-1/2})$ uniformly over $\mathcal{N}$. This follows from the same argument used to handle the second term in (11). The

mean of the ith component is $\mathbb{E}W_i\left[F(X'\beta_0)\nu(X'\phi,\phi)+(1-F(X'\beta_0))\pi(X'\phi,\phi)\right]\left[\tau_n(X'\phi)-1\right]$ which converges to zero as $n\to\infty$ by A2, A11, and dominated convergence, since $\tau_n(t)\to 1$ as $n\to\infty$ for all $t\in\mathbb{R}$.

Consider the first term in (11). The result will follow if $wp\to 1$ as $n\to\infty$, uniformly over j for which $\tau_n(X_j'\phi)=1$ and $\phi\in\mathcal{N}$, $\hat{\nu}(X_j'\phi,\phi)$ converges to $\nu(X_j'\phi,\phi)$ and $\hat{\pi}(X_j'\phi,\phi)$ converges to $\pi(X_j'\phi,\phi)$. We will prove the former. Proof of the latter is similar. Note that $\tau_n(X_j'\phi)\left|\hat{\nu}(X_j'\phi,\phi)-\nu(X_j'\phi,\phi)\right|$ equals

$$\tau_n(X_j'\phi)\left|\sum_{i=1}^{j}\frac{\sigma_n(X_i'\phi)\hat{F}(X_i'\phi,\phi)\,\Delta(X_i'\phi)}{\hat{F}(X_j'\phi,\phi)}-\frac{\int_{-\infty}^{X_j'\phi}F(u,\phi)\,du}{F(X_j'\phi,\phi)}\right| \tag{12}$$

$$\leq\ \tau_n(X_j'\phi)\left|\sum_{i=1}^{j}\frac{\sigma_n(X_i'\phi)\hat{F}(X_i'\phi,\phi)\,\Delta(X_i'\phi)}{\hat{F}(X_j'\phi,\phi)}-\sum_{i=1}^{j}\frac{\sigma_n(X_i'\phi)F(X_i'\phi,\phi)\,\Delta(X_i'\phi)}{F(X_j'\phi,\phi)}\right| \tag{13}$$

$$+\ \tau_n(X_j'\phi)\left|\sum_{i=1}^{j}\frac{\sigma_n(X_i'\phi)F(X_i'\phi,\phi)\,\Delta(X_i'\phi)}{F(X_j'\phi,\phi)}-\frac{\int_{-\infty}^{X_j'\phi}F(u,\phi)\,du}{F(X_j'\phi,\phi)}\right|. \tag{14}$$

Consider (13) in the last display. A4 and the fact that $\hat{F}(X_i'\phi,\phi)$ is bounded between zero and unity imply that for all $\delta>0$,

$$\sum_{i=1}^{j}\sigma_n(X_i'\phi)\hat{F}(X_i'\phi,\phi)\Delta(X_i'\phi)\ll n^{\delta}\,. \tag{15}$$

Deduce from (15), Lemma 3A(iii), and A6, that for all $\delta>0$, $wp\to 1$ as $n\to\infty$, uniformly over j and ϕ,

$$\tau_n(X_j'\phi)\left|\sum_{i=1}^{j}\sigma_n(X_i'\phi)\hat{F}(X_i'\phi,\phi)\,\Delta(X_i'\phi)\left[\frac{1}{\hat{F}(X_j'\phi,\phi)}-\frac{1}{F(X_j'\phi,\phi)}\right]\right|=o_p(n^{-1/4+\delta})\,. \tag{16}$$

It then follows from (15), (16), Lemma 3A(iii), and A6, that for all $\delta>0$, $wp\to 1$ as $n\to\infty$, uniformly over j and ϕ, the term in (13) has order $o_p(n^{-1/4+\delta})$.

Next, consider (14). Write $X_\sigma'\phi$ for the smallest index value for which $\sigma_n(X_i'\phi)=1$. If $\tau_n(X_j'\phi)=1$, then $X_j'\phi\geq -c_n$. By A7, $F(t,\phi)\searrow 0$ as $t\to-\infty$ for each $\phi\in\mathcal{N}$. This and the

triangle inequality imply that $wp \to 1$ as $n \to \infty$ the term in (14) is bounded by

$$\frac{1}{F(X'_j\phi,\phi)} \sum_{i=\sigma}^{j} \left| F(X'_i\phi,\phi)\Delta(X'_i\phi) - \int_{X'_{i-1}\phi}^{X'_i\phi} F(u,\phi)\,du \right| + \frac{\int_{-\infty}^{X'_{\sigma-1}\phi} F(u,\phi)\,du}{F(-c_n,\phi)}\,. \tag{17}$$

By A5 and Lemma 4A, $wp \to 1$ as $n \to \infty$, $X'_{\sigma-1}\phi$ can be made arbitrarily close to $-c_n^\alpha$. Deduce that the second term in (17) is bounded by $\int_{-\infty}^{-c_n^\alpha} F(u,\phi)\,du/F(-c_n,\phi)$ which converges to zero by A7. Consider the first term in (17). Define $H(\epsilon) = \int_{t-\epsilon}^{t} F(u,\phi)\,du$. By A9 and a Taylor expansion of $H(\epsilon)$ about $\epsilon = 0$ followed by a Taylor expansion of $F(t-\epsilon,\phi)$ about $\epsilon = 0$, we see that $H(\epsilon) = \epsilon F(t,\phi) + O(\epsilon^2)$ uniformly over $t \in I\!R$ and $\phi \in \mathcal{N}$. Deduce from this and A5 that the first term in (17) has order $o(n^{1+\delta}\Delta_n^2)$ for all $\delta > 0$ where $\Delta_n = \sup_\phi \Delta_n(\phi)$ with $\Delta_n(\phi) = \max_i\{\Delta(X'_i\phi)\sigma_n(X'_i\phi)\}$, the norm of the partition of the interval $[-c_n^\alpha, c_n^\alpha]$. By A5 and Lemma 4A, $wp \to 1$ as $n \to \infty$, $\Delta_n \ll n^\delta \log n/n$ for all $\delta > 0$. Deduce that as $n \to \infty$, the first term in (17) has order $o_p(n^{\delta-1})$ for all $\delta > 0$. This proves the result. □