RANK ESTIMATORS FOR A TRANSFORMATION MODEL

Elena Asparouhova, 626-395-4093

Tihomir Asparouhov, 310-391-9971

Robert Golanski, 626-395-4093

Krzysztof Kasprzyk, 626-395-4093

Robert P. Sherman, 626-395-4337

All the authors, except Tihomir Asparouhov, have the same affiliation:

California Institute of Technology

Humanities and Social Sciences

Mail Stop 228-77

Pasadena, CA, 91125

Tihomir Asparouhov has the following affiliation:

Mplus

11965 Venice Blvd., Suite 407

Los Angeles, CA 90066

RANK ESTIMATORS FOR A TRANSFORMATION MODEL

Robert P. Sherman

California Institute of Technology

Humanities and Social Sciences

Mail Stop 228-77

Pasadena, CA, 91125

Abstract

We establish $\sqrt{n}$-consistency and asymptotically normality of Han's (1987a) estimator of the parameters characterizing the transformation function in a semiparametric transformation model. We verify a Vapnik-Cervonenkis (VC) condition for the parameterizations of Box and Cox (1964) and Bickel and Doksum (1981). The verification establishes the VC property for a class of sets where nonlinear functions of the transformation parameters are positive. We also introduce a new class of rank estimators for these parameters. These estimators require only $O(n^2 \log n)$ computations to evaluate the criterion function, compared to $O(n^4)$ computations for Han's estimator. We prove that these estimators are also $\sqrt{n}$-consistent and asymptotically normal. A simulation study compares two of the new estimators to Han's estimator, as well as to the fully parametric estimator of Bickel and Doksum (1981) and the nonlinear two-stage least squares estimator of Amemiya and Powell (1981).

1. Introduction

Han (1987a) analyzes a transformation model of the form

$$Z(Y, \lambda_0) = \alpha_0 + X'\beta_0 + U \tag{1}$$

where Y is a scalar response variable, X is a $k \times 1$ vector of nonconstant explanatory variables, and U is an unobserved error term. In this model, (Y, X) is observed, $Z(Y, \lambda_0)$ is an increasing function of Y known up to the $m \times 1$ vector parameter λ_0, α_0 is an intercept, and β_0 is a $k \times 1$ vector of slope parameters. The object of estimation is $\theta_0 = (\lambda_0, \alpha_0, \beta_0)$.

A leading special case of the transformation function $Z(Y, \lambda_0)$ in (1) is the power transformation studied by Box and Cox (1964). In this model, $m = 1$ and $Y > 0$. The transformation has the form

$$Z(Y, \lambda_0) = (Y^{\lambda_0} - 1)/\lambda_0 \qquad \lambda_0 > 0\,. \tag{2}$$

This transformation is generalized by Bickel and Doksum (1981). For $Y \in \mathbb{R}$, their transformation has the form

$$Z(Y, \lambda_0) = (|Y|^{\lambda_0}\text{sign}(Y) - 1)/\lambda_0 \qquad \lambda_0 > 0\,. \tag{3}$$

Model (1) is either parametric or semiparametric depending on whether or not the functional form of the distribution of U is assumed to be known. For example, in (3), one could assume that U is normally distributed with mean zero and finite variance, and estimate θ_0 by maximum likelihood. However, in a parametric approach such as this, if the distributional assumption is incorrect, and especially if some data is contaminated, then standard estimation procedures are susceptible to large biases. This motivates a semiparametric approach that is not sensitive to outliers in the data.

Han (1987a) estimates θ_0 semiparametrically in three stages. In the first stage, he estimates $\delta_0 = \beta_0/|\beta_0|$ with the maximum rank correlation (MRC) estimator developed in Han (1985). Han (1985, 1987b) proves consistency and Sherman (1993) proves $\sqrt{n}$-consistency and asymptotic normality of this estimator. In the second stage, using the estimator of δ_0, Han (1987a) develops a consistent rank-based optimization estimator of λ_0. In the third stage, using the estimators of δ_0 and λ_0, Han (1987a) constructs consistent estimators of α_0 and $|\beta_0|$ using either least squares or least absolute deviations regression.

This paper focuses on estimation of λ_0, the parameter vector that characterizes the transformation function in (1). A number of semiparametric estimators of λ_0 have been developed. These include the nonlinear two-stage least squares (NL2SLS) estimator of Amemiya and Powell (1981) and a rescaled version of the NL2SLS estimator developed by Powell (1996). These estimators are special cases of generalized method of moment (GMM) estimators of λ_0 (see Mittlehammer, Judge, and Miller, 2000). Other semiparametric estimators of λ_0 include the quantile regression estimator of Powell (1991) and the nonlinear weighted least squares estimator of Foster, Tian, and Wei (2000). A recent simulation study by Savin and Wurtz (2001) compares the three nonlinear least squares estimators just mentioned.

In related work, Horowitz (1996), Ye and Duan (1997), Chen (2001), and Klein and Sherman (2001) have developed $\sqrt{n}$-consistent and asymptotically normal semiparametric estimators of the transformation function in (1) without making parametric assumptions about its functional form. (See also Gorgens and Horowitz, 1999.) However, all these estimators require independence of errors and regressors. In addition, with the exception of Chen's estimator, they involve the use of kernel regression estimators requiring subjective bandwidth choices. The estimators of Han (1987a) (see Remark 1 in Section 4), Powell (1991), and certain GMM estimators (see Mittlehammer et al. 2000) can be consistent under general forms of heteroscedasticity. Also, no bandwidth choices are needed to compute either Han's (1987a) estimator, or the estimators of Amemiya and Powell (1981), Powell (1991), Foster et al. (2000), or the other GMM estimators.

There are two main drawbacks to Han's estimator of λ_0. First, the limiting distribution of his estimator has not been established. Han (1987a) conjectured that his estimator was $\sqrt{n}$-consistent with a nonnormal limiting distribution, but provided no proofs. Part of the difficulty in establishing the limiting distribution of this estimator lies in the fact that it maximizes a discontinuous sample objective function. This renders standard Taylor series arguments inapplicable and suggests an approach based on recent developments in empirical process theory. However, the sample objective function is a generalized average of indicator functions of sets where nonlinear functions of λ_0 are positive. The nonlinearity of the transformation functions precludes the use of "off the shelf" empirical process results. Secondly, $O(n^4)$ computations are required to evaluate the criterion function, where n is the sample size. This makes the estimator prohibitively expensive to compute for moderate to large sample sizes.

The goal of this paper is to repair both of the previously mentioned deficiencies. We establish a result that lets us deduce the key Vapnik-Cervonenkis (VC) property for the classes of sets associated with the leading special cases in (2) and (3). This property is sufficient to establish $\sqrt{n}$-consistency and asymptotic normality of Han's estimator, thus proving one part and disproving the other part of Han's conjecture. We will then introduce a new class of rank-based optimization estimators that are also $\sqrt{n}$-consistent and asymptotically normal, but only require $O(n^2 \log n)$ computations to evaluate the objective function, making the estimators practicable for larger sample sizes. These estimators are adaptations of the rank estimators of δ_0 developed by Cavanagh and Sherman (1998), and exploit Spearman's (1904) measure of rank correlation rather than Kendall's (1938) measure, on which Han's estimator is based. The former is computationally more efficient.

In the next section, we present Han's estimator of λ_0 and state conditions under which the estimator is $\sqrt{n}$-consistent and asymptotically normal. We also establish the VC property for the class of sets associated with the transformation functions in (2) and (3). In Section 3, we establish the limiting distribution of the estimator and show how to estimate its asymptotic variance-covariance matrix. In Section 4, we introduce the new class of rank estimators of λ_0 and prove consistency. We also prove $\sqrt{n}$-consistency and asymptotic normality and indicate how to estimate asymptotic variance-covariance matrices. Section 5 presents simulation results comparing two of the new estimators to those of Han's estimator, the Bickel and Doksum (1981) parametric estimator, and the semiparametric estimator of Amemiya and Powell (1981). Section 6 summarizes.

2. Han's Estimator of λ_0 and Conditions for Asymptotic Normality

Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ denote a sample of iid observations from model (1). Let $\Lambda \subseteq \mathbb{R}^m$ denote the parameter space for λ_0, and $\Delta \subseteq \mathbb{R}^k$ the parameter space for δ_0. Let $\hat{\delta}$ denote a consistent estimator of δ_0. Han (1987a) proposes estimating λ_0 with $\hat{\lambda}(\hat{\delta}) = \underset{\lambda \in \Lambda}{\operatorname{argmax}}\, Q_n(\lambda, \hat{\delta})$ where

$$Q_n(\lambda, \delta) = \sum_{\boldsymbol{i_4}} \{Z(Y_i, \lambda) - Z(Y_j, \lambda) > Z(Y_k, \lambda) - Z(Y_l, \lambda)\}\{X_i'\delta - X_j'\delta > X_k'\delta - X_l'\delta\}$$

with $\boldsymbol{i_4} = (i, j, k, l)$ ranging over the $(n)_4 = n(n-1)(n-2)(n-3)$ ordered 4-tuples of different integers from the set $\{1, \ldots, n\}$. Note that for fixed λ and δ, $Q_n(\lambda, \delta)$ is a fourth-order U-statistic.

A simple principle motivates the estimator. By (1), $Z(Y_1, \lambda_0) - Z(Y_2, \lambda_0) > Z(Y_3 \lambda_0) - Z(Y_4, \lambda_0)$ if and only if $(U_3 - U_4) - (U_1 - U_2) < (X_1 - X_2)'\delta_0 - (X_3 - X_4)'\delta_0$. Suppose in (1) that U is independent of X. Then the random variables $(U_3 - U_4) - (U_1 - U_2)$ and $(X_1 - X_2)'\delta_0 - (X_3 - X_4)'\delta_0$

are independent and symmetric about zero. It follows that given (Y_1, Y_2, Y_3, Y_4), if $Z(Y_1, \lambda_0) - Z(Y_2, \lambda_0) > Z(Y_3 \lambda_0) - Z(Y_4, \lambda_0)$, then it is more likely than not that $X_1'\delta_0 - X_2'\delta_0 > X_3'\delta_0 - X_4'\delta_0$. This identifies λ_0 as the maximizer of the expected value of $Q_n(\lambda, \delta_0)$. We also see that $\hat{\lambda}(\hat{\delta})$ maximizes Kendall's (1938) measure of rank correlation between the $Z(Y_i, \lambda) - Z(Y_j, \lambda)$'s and the $X_i'\hat{\delta} - X_j'\hat{\delta}$'s.

Write W for (Y, X) and let W have distribution P on the set $\mathcal{W} \subseteq \mathbb{R} \otimes \mathbb{R}^k$. For each $w \in \mathcal{W}$, $\lambda \in \Lambda$, and $\delta \in \Delta$, define

$$
\begin{aligned}
f(w, \lambda, \delta) &= g(w, P, P, P, \lambda, \delta) + g(P, w, P, P, \lambda, \delta) \\
&+ g(P, P, w, P, \lambda, \delta) + g(P, P, P, w, \lambda, \delta)
\end{aligned}
$$

where, for $w_i = (y_i, x_i) \in \mathcal{W}$, $i = 1, 2, 3, 4$,

$$
\begin{aligned}
g(w_1, w_2, w_3, w_4, \lambda, \delta) &= \{Z(y_1, \lambda) - Z(y_2, \lambda) > Z(y_3, \lambda) - Z(y_4, \lambda)\}\{x_1'\delta - x_2'\delta > x_3'\delta - x_4'\delta\} \\
&- \{Z(y_1, \lambda_0) - Z(y_2, \lambda_0) > Z(y_3, \lambda_0) - Z(y_4, \lambda_0)\}\{x_1'\delta - x_2'\delta > x_3'\delta - x_4'\delta\}
\end{aligned}
$$

and $g(w, P, P, P, \lambda, \delta)$, for example, is shorthand for $P \otimes P \otimes P g(w, \cdot, \cdot, \cdot, \lambda, \delta)$.

Write P_n for the empirical measure that puts mass $\frac{1}{n}$ on each $W_i = (Y_i, X_i)$. The term $P_n[f(\cdot, \lambda, \delta) - f(P, \lambda, \delta)]$ is the projection term in the Hoeffding decomposition (see, for example, Serfling, 1980) of $Q_n(\lambda, \delta) - Q_n(\lambda_0, \delta)$. This term, evaluated at $\delta = \hat{\delta}$, drives the asymptotic behavior of $\hat{\lambda}(\hat{\delta})$.

Let h denote an arbitrary function of λ and δ. Write h_λ for $\nabla_\lambda h$, $h_{\lambda\lambda}$ for $\nabla_\lambda [\nabla_\lambda h]$, $h_{\lambda\delta}$ for $\nabla_\delta [\nabla_\lambda h]$, and so on. Let $\|\cdot\|$ denote the matrix norm $\|(a_{ij})\| = (\sum_{i,j} a_{ij}^2)^{1/2}$. For $w \in \mathcal{W}$, let $\gamma(w, \delta_0)$ denote a function from $\mathcal{W}$ to $\mathbb{R}^m$, and let the symbol $\Longrightarrow$ denote convergence in distribution.

We now state conditions implying $\sqrt{n}$-consistency and asymptotic normality of $\hat{\lambda}(\hat{\delta})$.

A1. The U_i's, $i = 1, \ldots, n$, are iid.

A2. The X_i's are iid and independent of the U_i's.

A3. $X'\delta_0$ is continuously distributed and U has a nondegenerate distribution.

A4. (λ_0, δ_0) is an interior point of $\Lambda \times \Delta$, a compact subset of $\mathbb{R}^m \times \mathbb{R}^k$.

A5. For each $\lambda \in \Lambda$, $Z(\cdot, \lambda)$ is continuous and strictly increasing.

A6. With positive probability, $Z(\cdot, \lambda)$ is differentiable and nonlinear in $Z(\cdot, \lambda_0)$ for $\lambda \neq \lambda_0$.

A7. $\{\{(y_1, y_2, y_3, y_4) : Z(y_1, \lambda) - Z(y_2, \lambda) > Z(y_3, \lambda) - Z(y_4, \lambda)\} : \lambda \in \Lambda\}$ is a VC class of sets.

A8. $\sqrt{n}(\hat{\delta} - \delta_0) = \sqrt{n} P_n \gamma(\cdot, \delta_0) + o_p(1)$ as $n \to \infty$, where $\sqrt{n} P_n \gamma(\cdot, \delta_0) \Longrightarrow N(0, \mathbb{E}\gamma(\cdot, \delta_0)\gamma(\cdot, \delta_0)')$.

A9. Let $\mathcal{N} \subseteq \Lambda \times \Delta$ denote a nondegenerate convex neighborhood of (λ_0, δ_0).

(i) For each $w \in \mathcal{W}$, $f(w, \cdot, \cdot)$ has continuous mixed third partial derivatives on $\mathcal{N}$.

(ii) There is an integrable function $M(w)$ such that for each $w \in \mathcal{W}$ and $(\lambda, \delta) \in \mathcal{N}$

$$\|f_{\lambda\lambda}(w, \lambda, \delta) - f_{\lambda\lambda}(w, \lambda_0, \delta)\| \leq M(w)|\lambda - \lambda_0|\,.$$

(iii) $\mathbb{E}|f_\lambda(\cdot, \lambda_0, \delta_0)|^2 < \infty$.

(iv) $\mathbb{E}\|f_{\lambda\lambda}(\cdot, \lambda_0, \delta_0)\| < \infty$.

(v) The matrix $\mathbb{E} f_{\lambda\lambda}(\cdot, \lambda_0, \delta_0)$ is negative definite.

Assumptions A1 through A6 are sufficient to prove strong consistency of $\hat{\lambda}(\hat{\delta})$. These assumptions are slightly weaker than those used by Han (1987a) to prove strong consistency of $\hat{\lambda}(\hat{\delta})$. In fact, the assumption of independence between the X_i's and U_i's in A2 is much stronger than necessary. (See Remark 1 after the consistency proof in Section 4.) A7 is a key regularity condition used in the normality proof in Section 3. Below, we verify this property for the leading special cases given in (2) and (3). A8 requires that there be a $\sqrt{n}$-consistent and asymptotically normal first-stage estimator of δ_0. As mentioned previously, Han's (1985,1987b) maximum rank correlation estimator satisfies this condition. One could also use one of the computationally more efficient rank estimators of Cavanagh and Sherman (1998). These estimators require no subjective bandwidth choices. The conditions of A9 are standard and are used to support arguments based on Taylor expansions of $f(w, \cdot, \cdot)$ and its derivatives about λ_0 and δ_0. Note that if Z and the density of $\mathcal{W}$ are sufficiently smooth, then A9(i) through A9(iv) will hold.

Recall that a VC class of subsets of a set $\mathcal{S}$ can pick out at most a polynomial number of subsets from the 2^n possible subsets of an arbitrary set of n points in $\mathcal{S}$ (e.g., Pakes and Pollard, 1989). We will use the next result to verify A7 for the models described in (2) and (3). In the statement of this result, a real-valued function $f(\lambda)$, $\lambda \in \mathbb{R}$, is said to change sign at a point λ_0 if either (i) $f(\lambda) \leq 0$ for $\lambda < \lambda_0$ and $f(\lambda) > 0$ for $\lambda > \lambda_0$ or (ii) $f(\lambda) > 0$ for $\lambda < \lambda_0$ and $f(\lambda) \leq 0$ for $\lambda > \lambda_0$.

LEMMA 1: *Let $h(s, \lambda)$ be a real-valued function of $s \in \mathcal{S} \subseteq \mathbb{R}^q$ and $\lambda \in \Lambda \subseteq \mathbb{R}$. Let $N(s)$ denote the number of points at which $h(s, \cdot)$ changes sign. If $\sup_{s \in \mathcal{S}} N(s) \leq B < \infty$, then $\{\{s : h(s, \lambda) > 0\} : \lambda \in \Lambda\}$ is a VC class of sets.*

PROOF. Let $s_1, \ldots, s_n$ denote n arbitrary points in $\mathcal{S}$. Write $\mathcal{H}$ for $\{\{s : h(s, \lambda) > 0\} : \lambda \in \Lambda\}$. We say, for example, that $\mathcal{H}$ picks out the subset $\{s_2, s_3, s_7\}$ from $s_1, \ldots, s_n$ if there exists a $\lambda \in \Lambda$ such that $h(s_i, \lambda) > 0$ for $i \in \{2, 3, 7\}$ and $h(s_i, \lambda) \leq 0$ for $i \notin \{2, 3, 7\}$. We will show that $\mathcal{H}$ can pick out no more than $Bn + 1$ of the 2^n possible subsets of $s_1, \ldots, s_n$. This will prove the result.

By assumption, for each i, there are at most B points at which $h(s_i, \cdot)$ changes sign. The union of these points contains at most Bn points and partitions Λ into at most $Bn + 1$ intervals. Each interval corresponds to an n-tuple of $+$'s and $-$'s, where the ith component is a $+$ if $h(s_i, \lambda) > 0$ for each λ in the interval, and a $-$ if $h(s_i, \lambda) \leq 0$ for each λ in the interval. Thus, the number of such n-tuples bounds the number of subsets of $s_1, \ldots, s_n$ that $\mathcal{H}$ can pick out. □

COROLLARY: *A7 holds for the models described in (2) and (3).*

PROOF. Consider model (2). Take $s = (y_1, y_2, y_3, y_4)$, $\mathcal{S} = \mathbb{R}^4_+$, $\Lambda = \mathbb{R}_+$, and $h(s, \lambda) = y_1^\lambda - y_2^\lambda - y_3^\lambda + y_4^\lambda$. Simple calculus shows that for each $s \in \mathcal{S}$, $h(s, \cdot)$ changes sign at most $B = 2$ times. The proof for (3) is similar. □

3. The Limiting Distribution of Han's Estimator

In this section, we prove that $\hat{\lambda}(\hat{\delta})$ is $\sqrt{n}$-consistent and asymptotically normal, and show how to estimate its asymptotic variance-covariance matrix.

Write $\Gamma_n(\lambda, \delta)$ for $(n)_4^{-1} [Q_n(\lambda, \delta) - Q_n(\lambda_0, \delta)]$ and $\Gamma(\lambda, \delta)$ for $\mathbb{E}\Gamma_n(\lambda, \delta)$. Note that $\hat{\lambda}(\hat{\delta}) = \underset{\lambda \in \Lambda}{\operatorname{argmax}}\, \Gamma_n(\lambda, \hat{\delta})$ and $\lambda_0 = \underset{\lambda \in \Lambda}{\operatorname{argmax}}\, \Gamma(\lambda, \delta_0)$. Finally, write $\tau(\lambda, \delta)$ for $\frac{1}{4}\mathbb{E} f_{\lambda\delta}(\cdot, \lambda, \delta)$.

THEOREM 2: *If A1 through A9 hold, then*

$$\sqrt{n}(\hat{\lambda}(\hat{\delta}) - \lambda_0) \Longrightarrow N(0, V^{-1}\Sigma V^{-1})$$

where

$$\begin{aligned} V &= \frac{1}{4}\mathbb{E} f_{\lambda\lambda}(\cdot, \lambda_0, \delta_0) \\ \Sigma &= \mathbb{E}\left[f_\lambda(\cdot, \lambda_0, \delta_0) + \tau(\lambda_0, \delta_0)\gamma(\cdot, \delta_0)\right]\left[f_\lambda(\cdot, \lambda_0, \delta_0) + \tau(\lambda_0, \delta_0)\gamma(\cdot, \delta_0)\right]' . \end{aligned}$$

PROOF. Under A1 through A6, Han (1987a) proved that $|\hat{\lambda}(\hat{\delta}) - \lambda_0| = o_p(1)$ as $n \to \infty$. We will show that uniformly in $o_p(1)$ neighborhoods of (λ_0, δ_0),

$$\Gamma_n(\lambda, \hat{\delta}) = \frac{1}{2}(\lambda - \lambda_0)'V(\lambda - \lambda_0) + \frac{1}{\sqrt{n}}(\lambda - \lambda_0)'W_n + o_p(|(\lambda - \lambda_0)|^2) + o_p(1/n) \tag{4}$$

where W_n converges in distribution to a $N(\mathbf{0}, \Sigma)$ random vector. Then A9(v), (4), and Theorem 1 in Sherman (1993) will imply that

$$|(\hat{\lambda}(\hat{\delta}) - \lambda_0)| = O_p(1/\sqrt{n}). \tag{5}$$

The result will then follow from (4), (5), and Theorem 2 in Sherman (1993).

Write $U_{k,n}$ for the probability measure that puts mass $1/(n)_k$ on each k-tuple $(W_{i_1}, \ldots, W_{i_k})$, $k = 2, 3, 4$. Apply a version of the Hoeffding decomposition (see Sherman, 1994, p.449) to write

$$\begin{aligned}\Gamma_n(\lambda, \delta) &= \Gamma(\lambda, \delta) + P_n\left[f(\cdot, \lambda, \delta) - f(P, \lambda, \delta)\right] \\ &+ U_{2,n} f_2(\cdot, \cdot, \lambda, \delta) + U_{3,n} f_3(\cdot, \cdot, \cdot, \lambda, \delta) + U_{4,n} f_4(\cdot, \cdot, \cdot, \cdot, \lambda, \delta)\end{aligned}$$

where f_k is a degenerate U-statistic of order k, $k = 2, 3, 4$.

Apply A3, A7, Lemma 2.12 in Pakes and Pollard (1989), Lemma 6 and Corollary 8 in Sherman (1994), and argue as in Theorem 4 of Sherman (1993) to see that the degenerate U-processes of order two, three, and four can be neglected. That is, uniformly over $o_p(1)$ neighborhoods of (λ_0, δ_0),

$$\Gamma_n(\lambda, \delta) = \Gamma(\lambda, \delta) + P_n\left[f(\cdot, \lambda, \delta) - f(P, \lambda, \delta)\right] + o_p(1/n).$$

Deduce that uniformly over $o_p(1)$ neighborhoods of (λ_0, δ_0),

$$\Gamma_n(\lambda, \hat{\delta}) = \Gamma(\lambda, \hat{\delta}) + P_n\left[f(\cdot, \lambda, \hat{\delta}) - f(P, \lambda, \hat{\delta})\right] + o_p(1/n)\,. \tag{6}$$

Next, we show that uniformly over $o_p(1)$ neighborhoods of (λ_0, δ_0),

$$\Gamma(\lambda, \hat{\delta}) = \frac{1}{\sqrt{n}}(\lambda - \lambda_0)' \left[\sqrt{n} P_n \tau(\lambda_0, \delta_0)\gamma(\cdot, \delta_0) + o_p(1)\right] + \frac{1}{2}(\lambda - \lambda_0)' V(\lambda - \lambda_0) + o_p(|(\lambda - \lambda_0)|^2)\,. \tag{7}$$

The term $\sqrt{n} P_n \tau(\lambda_0, \delta_0)\gamma(\cdot, \delta_0)$ quantifies the penalty paid for having to estimate δ_0.

Fix $w \in \mathcal{W}$ and $(\lambda, \delta) \in \mathcal{N}$. Invoke A9(i) and expand $f(w, \lambda, \delta)$ about $\lambda = \lambda_0$ to get

$$f(w, \lambda, \delta) = (\lambda - \lambda_0)' f_\lambda(w, \lambda_0, \delta) + \frac{1}{2}(\lambda - \lambda_0)' f_{\lambda\lambda}(w, \lambda^*, \delta)(\lambda - \lambda_0) \tag{8}$$

for λ^* between λ and λ_0. By A9(ii), for each $w \in \mathcal{W}$ and each $(\lambda, \delta) \in \mathcal{N}$

$$\|(\lambda - \lambda_0)'[f_{\lambda\lambda}(w, \lambda, \delta) - f_{\lambda\lambda}(w, \lambda_0, \delta)](\lambda - \lambda_0)\| \le M(z)|(\lambda - \lambda_0)|^3\,. \tag{9}$$

Invoke (9) and the integrability of M, then take expectations in (8) and evaluate at $\delta = \hat{\delta}$ to get that uniformly over $o_p(1)$ neighborhoods of (λ_0, δ_0),

$$4\Gamma(\lambda, \hat{\delta}) = (\lambda - \lambda_0)' \mathbb{E} f_\lambda(\cdot, \lambda_0, \hat{\delta}) + \frac{1}{2}(\lambda - \lambda_0)' 4V(\lambda - \lambda_0) + o(|(\lambda - \lambda_0)|^2)\,.$$

Since $\Gamma(\lambda, \delta_0)$ is maximized at $\lambda = \lambda_0$, $\mathbb{E} f_\lambda(\cdot, \lambda_0, \delta_0) = 0$. By a Taylor expansion about $\delta = \delta_0$, $\mathbb{E} f_\lambda(\cdot, \lambda_0, \hat{\delta}) = \mathbb{E} f_{\lambda\delta}(\cdot, \lambda_0, \delta^*)(\hat{\delta} - \delta_0)$ where δ^* is between $\hat{\delta}$ and δ_0. Divide through by 4 and apply A8 and A9(i) to establish (7).

Next, we show that uniformly over $o_p(1)$ neighborhoods of (λ_0, δ_0),

$$P_n\left[f(\cdot,\lambda,\hat{\delta}) - f(P,\lambda,\hat{\delta})\right] = \frac{1}{\sqrt{n}}(\lambda-\lambda_0)'\left[\sqrt{n}P_n f_\lambda(\cdot,\lambda_0,\delta_0) + o_p(1)\right] + o(|\lambda-\lambda_0|^2)\,. \tag{10}$$

Note that $f(\cdot,\lambda_0,\delta) = 0$ for all δ. Also, since $f(P,\lambda,\delta) = 4\Gamma(\lambda,\delta)$, $f_\lambda(P,\lambda_0,\delta_0) = 0$. Condition (10) then follows from A9(i) and a Taylor expansion about $\lambda = \lambda_0$ followed by a Taylor expansion about $\delta = \delta_0$.

Conditions (6), (7), and (10) imply condition (4). This proves the result. □

We now develop consistent estimators of V and Σ in Theorem 2. We use numerical derivatives, as in Pakes and Pollard (1989). (For ease of notation, we present one-sided difference quotient estimators. Since the criterion function is a step function, centered difference quotient estimators perform better in practice, especially for small to moderate sample sizes.) Alternatively, one could develop expressions for V and Σ in terms of model primitives and estimate components nonparametrically, as is done in Cavanagh and Sherman (1998). However, we do not do so here.

Recall the definition of $f(w,\lambda,\delta)$ given in Section 2. Also, recall that $(n)_3 = n(n-1)(n-2)$ and $U_{3,n}$ denotes the probability measure that puts mass $1/(n)_3$ on each 3-tuple (W_i, W_j, W_k), where (i,j,k) ranges over the $(n)_3$ ordered 3-tuples of different integers from the set $\{1,\ldots,n\}$. For each $w \in \mathcal{W}$, $\lambda \in \Lambda$, and $\delta \in \Delta$, define

$$f_n(w,\lambda,\delta) = U_{3,n}\left[g(w,\cdot,\cdot,\cdot,\lambda,\delta) + g(\cdot,w,\cdot,\cdot,\lambda,\delta) + g(\cdot,\cdot,w,\cdot,\lambda,\delta) + g(\cdot,\cdot,\cdot,w,\lambda,\delta)\right]\,.$$

Note that $\mathbb{E} f_n(w,\lambda,\delta) = f(w,\lambda,\delta)$. Standard U-process arguments, such as those given in the proof of Theorem 2, imply that as $n \to \infty$,

$$\sup_{\mathcal{W}\times\Lambda\times\Delta} |f_n(w,\lambda,\delta) - f(w,\lambda,\delta)| = O_p(1/\sqrt{n})\,. \tag{11}$$

Write $f^i_\lambda(w,\lambda,\delta)$ for the ith component of $f_\lambda(w,\lambda,\delta)$, $f^{ij}_{\lambda\delta}(w,\lambda,\delta)$ for the ijth component of $f_{\lambda\delta}(w,\lambda,\delta)$, and $f^{ij}_{\lambda\lambda}(w,\lambda,\delta)$ for the ijth component of $f_{\lambda\lambda}(w,\lambda,\delta)$. Assume, for simplicity, that all these partial derivatives are bounded in a neighborhood of (λ_0,δ_0). Let $\{q_n\}$, $\{r_n\}$, $\{s_n\}$, and $\{t_n\}$ denote sequences of real numbers converging to zero as $n \to \infty$. Let $\{u_1,\ldots,u_m\}$ and $\{v_1,\ldots,v_k\}$ denote standard bases for $\mathbb{R}^m$ and $\mathbb{R}^k$, respectively. Define

$$\begin{aligned}
\hat{f}^i_\lambda(w,\lambda,\delta) &= q_n^{-1}\left[f_n(w,\lambda+q_n u_i,\delta) - f_n(w,\lambda,\delta)\right] \\
\hat{f}^{ij}_{\lambda\delta}(w,\lambda,\delta) &= r_n^{-1}s_n^{-1}\left[f_n(w,\lambda+r_n u_i,\delta+s_n v_j) - f_n(w,\lambda,\delta+s_n v_j) - f_n(w,\lambda+r_n u_i,\delta) + f_n(w,\lambda,\delta)\right] \\
\hat{f}^{ij}_{\lambda\lambda}(w,\lambda,\delta) &= t_n^{-2}\left[f_n(w,\lambda+t_n(u_i+u_j),\delta) - f_n(w,\lambda+t_n u_j,\delta) - f_n(w,\lambda+t_n u_i,\delta) + f_n(w,\lambda,\delta)\right]\,.
\end{aligned}$$

Deduce from this, (11), and a one-term Taylor expansion of the population difference quotients about (λ_0,δ_0) using the boundedness and continuity (A9(i)) of the partial derivatives near (λ_0,δ_0) and the $\sqrt{n}$-consistency of $(\hat\lambda(\hat\delta),\hat\delta)$, that as $n\to\infty$,

$$\begin{aligned}
\hat{f}^i_\lambda(w,\hat\lambda(\hat\delta),\hat\delta) &= f^i_\lambda(w,\lambda_0,\delta_0) + o(1) + q_n^{-1}O_p(1/\sqrt{n}) \\
\hat{f}^{ij}_{\lambda\delta}(w,\hat\lambda(\hat\delta),\hat\delta) &= f^{ij}_{\lambda\delta}(w,\lambda_0,\delta_0) + o(1) + r_n^{-1}s_n^{-1}O_p(1/\sqrt{n}) \\
\hat{f}^{ij}_{\lambda\lambda}(w,\hat\lambda(\hat\delta),\hat\delta) &= f^{ij}_{\lambda\lambda}(w,\lambda_0,\delta_0) + o(1) + t_n^{-2}O_p(1/\sqrt{n})
\end{aligned}$$

where the $o(1)$ terms and the $O_p(1/\sqrt{n})$ terms are uniformly bounded. Choose the sequences $\{q_n\}$,

$\{r_n\}$, $\{s_n\}$, and $\{t_n\}$ so that as $n \to \infty$, $q_n\sqrt{n} \to \infty$, $r_n s_n \sqrt{n} \to \infty$, and $t_n^2\sqrt{n} \to \infty$. Define

$$
\begin{aligned}
\hat{f}_\lambda(w, \lambda, \delta) &= (\hat{f}_\lambda^1(w, \lambda, \delta), \ldots, \hat{f}_\lambda^m(w, \lambda, \delta))' \\
\hat{f}_{\lambda\delta}(w, \lambda, \delta) &= (\hat{f}_{\lambda\delta}^{ij}(w, \lambda, \delta))_{m \times k} \\
\hat{f}_{\lambda\lambda}(w, \lambda, \delta) &= (\hat{f}_{\lambda\lambda}^{ij}(w, \lambda, \delta))_{m \times m} \\
\hat{\tau}(\lambda, \delta) &= \frac{1}{4} P_n \hat{f}_{\lambda\delta}(\cdot, \lambda, \delta)\,.
\end{aligned}
$$

Finally, define

$$
\begin{aligned}
\hat{V} &= \frac{1}{4} P_n \hat{f}_{\lambda\lambda}(\cdot, \hat{\lambda}(\hat{\delta}), \hat{\delta}) \\
\hat{\Sigma} &= P_n \left[\hat{f}_\lambda(\cdot, \hat{\lambda}(\hat{\delta}), \hat{\delta}) + \hat{\tau}(\hat{\lambda}(\hat{\delta}), \hat{\delta})\gamma(\cdot, \hat{\delta})\right] \left[\hat{f}_\lambda(\cdot, \hat{\lambda}(\hat{\delta}), \hat{\delta}) + \hat{\tau}(\hat{\lambda}(\hat{\delta}), \hat{\delta})\gamma(\cdot, \hat{\delta})\right]'\,.
\end{aligned}
$$

Deduce from the previous arguments and the consistency of $(\hat{\lambda}(\hat{\delta}), \hat{\delta})$ for (λ_0, δ_0) that $\hat{V}$ and $\hat{\Sigma}$ consistently estimate their population counterparts.

4. A New Class of Rank Estimators of λ_0

In this section, we introduce a new class of rank estimators of λ_0 that require only $O(n^2 \log n)$ computations to evaluate the objective function. These estimators are adaptations of the rank estimators of δ_0 developed by Cavanagh and Sherman (1998). We prove consistency, and then establish $\sqrt{n}$-consistency and asymptotic normality and indicate how to estimate the variance-covariance matrices.

Let M denote an increasing function on $I\!R$. For real numbers a_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, n$, $i \neq j$, let $R_n(a_{ij}) = \sum_{s \neq t}\{a_{st} < a_{ij}\}$, the rank of a_{ij}. Let $\hat{\delta}$ denote a consistent estimator of δ_0. We propose estimating λ_0 with $\hat{\lambda}(\hat{\delta}) = \underset{\lambda \in \Lambda}{\operatorname{argmax}}\, Q_n(\lambda, \hat{\delta})$ where

$$Q_n(\lambda, \delta) = \sum_{i \neq j} R_n(Z(Y_i, \lambda) - Z(Y_j, \lambda)) M(X_i'\delta - X_j'\delta)\,. \tag{12}$$

For ease of notation, we suppress the dependence of $\hat{\lambda}(\hat{\delta})$ on M. Note that when $M(a_{ij}) = R_n(a_{ij})$, then $Q_n(\lambda, \delta)$ is a linear function of Spearman's (1904) measure of rank correlation between the $Z(Y_i, \lambda) - Z(Y_j, \lambda)$'s and the $X_i'\delta - X_j'\delta$'s. (See also Lehmann 1975, Chapter 7.) Since sorting m numbers requires only $O(m \log m)$ computations (e.g., Aho et al., 1976, Section 3.4), we see that only $O(n^2 \log n)$ computations are needed to evaluate $Q_n(\lambda, \delta)$.

Maximum robustness is achieved by choosing $M(a_{ij}) = R_n(a_{ij})$. However, more efficiency may be obtained by choosing a deterministic specification like $M(z) = z$. An intermediate choice is the winsorized function $M(z) = a\{z < a\} + z\{a \leq z \leq b\} + b\{z > b\}$ for $a < b$. See Cavanagh and Sherman (1998) for more discussion on specifying M.

When M is deterministic, we make the following assumptions:

A10. M is a continuous, nonconstant, increasing function on the support of $X_1'\delta - X_2'\delta$.

A11. $\sup_{\delta \in \Delta} \mathbb{E}\,[M(X_1'\delta - X_2'\delta)]^2 < \infty$.

Continuity in A10 can be relaxed to almost-sure continuity. Also, note that A11 is trivially satisfied when M is bounded, as when $M(z)$ is the winsorized function defined above.

We now prove consistency of $\hat{\lambda}(\hat{\delta})$ when M is deterministic. In a remark after the proof of Theorem 3, we show how the proof easily extends to cover the case $M(a_{ij}) = R_n(a_{ij})$. The idea behind our identification proof is similar to Han's (1987a). However, we give our own proof since we could not follow the details of Han's proof.

THEOREM 3: *Suppose M is deterministic and $|\hat{\delta} - \delta_0| = o_p(1)$ as $n \to \infty$. If A1 through A7,* *A10, and A11 hold, then* $|\hat{\lambda}(\hat{\delta}) - \lambda_0| = o_p(1)$ *as* $n \to \infty$.

PROOF. Expand the rank function in (12) into a sum and throw away terms with equal indices. These terms are negligible, asymptotically. Divide by $(n)_4$ and, abusing notation slightly, let $Q_n(\lambda, \delta)$ denote this new objective function. Define $Q(\lambda, \delta) = \mathbb{E} Q_n(\lambda, \delta) = \mathbb{E}\{Z(Y_1, \lambda) - Z(Y_2, \lambda) > Z(Y_3, \lambda) - Z(Y_4, \lambda)\} M(X_1'\delta - X_2'\delta)$.

We will show

(i) $Q(\lambda, \delta_0)$ is uniquely maximized at λ_0.

(ii) $\sup_{\lambda \in \Lambda} |Q_n(\lambda, \hat{\delta}) - Q(\lambda, \delta_0)| = o_p(1)$ as $n \to \infty$.

(iii) $Q(\lambda, \delta_0)$ is continuous on Λ.

Consistency then follows from standard arguments (e.g., Amemiya 1985, pp.106-107).

Write $\boldsymbol{Y}$ for (Y_1, Y_2, Y_3, Y_4). Define

$$\begin{aligned} \Delta X_{ij}(\delta) &= X_i'\delta - X_j'\delta \\ H_{12}(\boldsymbol{Y}, \delta_0) &= \mathbb{E}\left[M(\Delta X_{12}(\delta_0)) \mid \boldsymbol{Y}\right] \\ H_{34}(\boldsymbol{Y}, \delta_0) &= \mathbb{E}\left[M(\Delta X_{34}(\delta_0)) \mid \boldsymbol{Y}\right] \\ \Delta Z_{ij}(\lambda) &= Z(Y_i, \lambda) - Z(Y_j, \lambda)\,. \end{aligned}$$

By symmetry,

$$Q(\lambda, \delta_0) = \frac{1}{2}\mathbb{E}\left[\{\Delta Z_{12}(\lambda) > \Delta Z_{34}(\lambda)\} H_{12}(\boldsymbol{Y}, \delta_0) + \{\Delta Z_{12}(\lambda) < \Delta Z_{34}(\lambda)\} H_{34}(\boldsymbol{Y}, \delta_0)\right]. \qquad (13)$$

Write ΔU for $(U_3 - U_4) - (U_1 - U_2)$ and ΔX for $\Delta X_{12}(\delta_0) - \Delta X_{34}(\delta_0)$. Recall from (1) that

$$\Delta Z_{12}(\lambda_0) > \Delta Z_{34}(\lambda_0) \Longleftrightarrow \Delta U < \Delta X\,.$$

By A1 and A2, ΔU and ΔX are independent and symmetric about zero. By Bayes' Theorem,

$$\begin{aligned}\mathbb{P}\{\Delta X > 0 \mid \Delta U < \Delta X\} &= \mathbb{P}\{\Delta U < \Delta X \mid \Delta X > 0\}\mathbb{P}\{\Delta X > 0\}/\mathbb{P}\{\Delta U < \Delta X\}\\ \mathbb{P}\{\Delta X < 0 \mid \Delta U < \Delta X\} &= \mathbb{P}\{\Delta U < \Delta X \mid \Delta X < 0\}\mathbb{P}\{\Delta X < 0\}/\mathbb{P}\{\Delta U < \Delta X\}\,.\end{aligned}$$

By symmetry, $\mathbb{P}\{\Delta X > 0\} = \mathbb{P}\{\Delta X < 0\}$. By independence and symmetry,

$$\mathbb{P}\{\Delta U < \Delta X \mid \Delta X > 0\} > \mathbb{P}\{\Delta U < \Delta X \mid \Delta X < 0\}\,. \tag{14}$$

Deduce that

$$\mathbb{P}\{\Delta X > 0 \mid \Delta U < \Delta X\} > \mathbb{P}\{\Delta X < 0 \mid \Delta U < \Delta X\}\,. \tag{15}$$

It follows from condition (15) and A3 that on a set of probability one,

$$\Delta Z_{12}(\lambda_0) > \Delta Z_{34}(\lambda_0) \Longleftrightarrow \mathbb{P}\left[\Delta X_{12}(\delta_0) > \Delta X_{34}(\delta_0) \mid \boldsymbol{Y}\right] > \mathbb{P}\left[\Delta X_{12}(\delta_0) < \Delta X_{34}(\delta_0) \mid \boldsymbol{Y}\right]\,.$$

In other words, on a set of probability one, $\Delta Z_{12}(\lambda_0) > \Delta Z_{34}(\lambda_0)$ if and only if the distribution of ΔX given $\boldsymbol{Y}$ puts more probability mass on points for which $\Delta X_{12}(\delta_0) > \Delta X_{34}(\delta_0)$ than on points for which $\Delta X_{12}(\delta_0) < \Delta X_{34}(\delta_0)$. Deduce from this, A3, and the fact that M is nonconstant and increasing, that

$$\Delta Z_{12}(\lambda_0) > \Delta Z_{34}(\lambda_0) \Longleftrightarrow H_{12}(\boldsymbol{Y}, \delta_0) > H_{34}(\boldsymbol{Y}, \delta_0)\,. \tag{16}$$

When $\lambda = \lambda_0$, the indicator functions in (13) pick out the larger of $H_{12}(\boldsymbol{Y}, \delta_0)$ and $H_{34}(\boldsymbol{Y}, \delta_0)$. This happens with probability one by A3 and A5. Thus, $Q(\lambda, \delta_0)$ is maximized at λ_0.

To show that λ_0 is the unique maximizer of $Q(\lambda, \delta_0)$, we show that $Q(\lambda_0, \delta_0) > Q(\lambda, \delta_0)$ for $\lambda \neq \lambda_0$. From (13), we get that

$$
\begin{aligned}
Q(\lambda_0, \delta_0) - Q(\lambda, \delta_0) = \frac{1}{2}\mathbb{E}[(\{\Delta Z_{12}(\lambda_0) > \Delta Z_{34}(\lambda_0)\} \quad &- \quad \{\Delta Z_{12}(\lambda) > \Delta Z_{34}(\lambda)\})H_{12}(\boldsymbol{Y}, \delta_0) \\
+(\{\Delta Z_{12}(\lambda_0) < \Delta Z_{34}(\lambda_0)\} \quad &- \quad \{\Delta Z_{12}(\lambda) < \Delta Z_{34}(\lambda)\})H_{34}(\boldsymbol{Y}, \delta_0)]\,.
\end{aligned}
$$

Write (A, B, C, D) for the value of the vector of four indicator functions in the last expression. That is, $A = \{\Delta Z_{12}(\lambda_0) > \Delta Z_{34}(\lambda_0)\}$, $B = \{\Delta Z_{12}(\lambda) > \Delta Z_{34}(\lambda)\})$, $C = \{\Delta Z_{12}(\lambda_0) < \Delta Z_{34}(\lambda_0)\}$, and $D = \{\Delta Z_{12}(\lambda) < \Delta Z_{34}(\lambda)\})$. On a set of probability one, there are four possible values for (A, B, C, D): $(1, 0, 0, 1)$, $(1, 1, 0, 0)$, $(0, 0, 1, 1)$, and $(0, 1, 1, 0)$. The contribution to $Q(\lambda_0, \delta_0) - Q(\lambda, \delta_0)$ from the second and third 4-tuples is zero. The contribution from the first and fourth 4-tuples is nonnegative by (16). To show that $Q(\lambda_0, \delta_0) > Q(\lambda, \delta_0)$, it is enough to show that the contribution from the first 4-tuple is strictly positive. This will hold if there exists a set $V = V_1 \times V_2 \times V_3 \times V_4$ such that $\prod_{i=1}^{4} \mathbb{P}\{Y_i \in V_i\} > 0$ and $A = D = 1$ on V.

For concreteness, first consider model (2). The general proof is similar, as we will show. We must find a set with positive probability on which the following two conditions hold:

$$
\begin{aligned}
y_1^{\lambda_0} - y_2^{\lambda_0} \quad &> \quad y_3^{\lambda_0} - y_4^{\lambda_0} \\
y_1^{\lambda} - y_2^{\lambda} \quad &< \quad y_3^{\lambda} - y_4^{\lambda}\,.
\end{aligned}
$$

Equivalently, we seek a set with positive probability on which

$$v_1^\alpha - v_2^\alpha > v_3^\alpha - v_4^\alpha$$

$$v_1 - v_2 < v_3 - v_4$$

where $v_i = y_i^\lambda$ and $\alpha = \lambda_0/\lambda$. Note that $v_i > 0$, $i = 1, 2, 3, 4$ and $\alpha > 0$.

Choose $0 < r_2 < r_1 < r_4 < r_3$ such that $r_1 - r_2 = r_3 - r_4$. Fix $\epsilon > 0$ satisfying $\epsilon < r_4 - r_1$ and $\epsilon < (r_1 - r_2)/2$. Fix $\alpha \in (0, 1)$. (A similar argument will work for $\alpha > 1$, as we will show.) Since r^α is increasing and concave in r, for all $s \in (r_2, r_1)$ and $t \in (r_4 - \epsilon, r_3 + \epsilon)$,

$$\frac{d}{dr} r^\alpha \mid_{r=s} > \frac{d}{dr} r^\alpha \mid_{r=t} . \tag{17}$$

Take $V_1 = (r_1 - \epsilon, r_1)$, $V_2 = (r_2, r_2 + \epsilon)$, $V_3 = (r_3, r_3 + \epsilon)$, and $V_4 = (r_4 - \epsilon, r_4)$. By construction, for all $v_i \in V_i$, $i = 1, 2, 3, 4$,

$$v_1 - v_2 < v_3 - v_4 .$$

By the Mean Value Theorem and (17), for all $v_i \in V_i$, $i = 1, 2, 3, 4$,

$$\frac{v_1^\alpha - v_2^\alpha}{v_1 - v_2} > \frac{v_3^\alpha - v_4^\alpha}{v_3 - v_4} .$$

Now choose ϵ small enough so that

$$v_1^\alpha - v_2^\alpha > v_3^\alpha - v_4^\alpha .$$

If $\alpha > 1$, let r_1, r_2, V_1, and V_2 change respective roles with r_3, r_4, V_3, and V_4 in the previous argument. This proves condition (i) for model (2).

Consider the general case. Write $Z_\lambda(y)$ for $Z(y,\lambda)$. We must find a set with positive probability on which the following two conditions hold:

$$\begin{aligned} Z_{\lambda_0}(y_1) - Z_{\lambda_0}(y_2) &> Z_{\lambda_0}(y_3) - Z_{\lambda_0}(y_4) \\ Z_\lambda(y_1) - Z_\lambda(y_2) &< Z_\lambda(y_3) - Z_\lambda(y_4)\,. \end{aligned}$$

Equivalently, we must find a set with positive probability on which

$$\begin{aligned} Z_{\lambda_0}(Z_\lambda^{-1}(v_1)) - Z_{\lambda_0}(Z_\lambda^{-1}(v_2)) &> Z_{\lambda_0}(Z_\lambda^{-1}(v_3)) - Z_{\lambda_0}(Z_\lambda^{-1}(v_4)) \\ v_1 - v_2 &< v_3 - v_4 \end{aligned}$$

where $v_i = Z_\lambda(y_i)$, $i = 1,2,3,4$.

Note that

$$\frac{d}{dv} Z_{\lambda_0}(Z_\lambda^{-1}(v)) = Z'_{\lambda_0}(Z_\lambda^{-1}(v))/Z'_\lambda(Z_\lambda^{-1}(v))\,.$$

Thus, $\frac{d}{dv} Z_{\lambda_0}(Z_\lambda^{-1}(v))$ is constant on an interval I if and only if $Z'_{\lambda_0}(y) \propto Z'_\lambda(y)$ for all y for which $Z_\lambda(y) \in I$. By A3 and A6, there is an interval of positive probability on which this does not happen. Choose $r_2 < r_1 < r_4 < r_3$ from this interval and argue as before to get the general result. This proves condition (i) for the general case.

Turn to condition (ii). Since $\hat{\delta}$ consistently estimates δ_0, there exists a sequence $\{\epsilon_n\}$ of positive real numbers satisfying $\epsilon_n = o(1)$ as $n \to \infty$ for which $\mathbb{P}\{|\hat{\delta} - \delta_0| > \epsilon_n\} \to 0$. It follows that as $n \to \infty$,

$$\sup_{\lambda \in \Lambda} |Q_n(\lambda, \hat{\delta}) - Q(\lambda, \delta_0)| \le \sup_{\lambda \in \Lambda, |\delta - \delta_0| \le \epsilon_n} |Q_n(\lambda, \delta) - Q(\lambda, \delta_0)| + o_p(1)\,.$$

The first term on the right is bounded by

$$\sup_{\lambda\in\Lambda,|\delta-\delta_0|\le\epsilon_n} |Q_n(\lambda,\delta)-Q(\lambda,\delta)| + \sup_{\lambda\in\Lambda,|\delta-\delta_0|\le\epsilon_n} |Q(\lambda,\delta)-Q(\lambda,\delta_0)|\,.$$

The first term has order $O_p(1/\sqrt{n})$ as $n\to\infty$. This follows from the fact that $Q_n(\lambda,\delta)-Q(\lambda,\delta)$ is a zero-mean U-process of order 4, and from standard U-process and empirical process results: apply A7, the argument in Section 5 of Sherman (1993), and Lemma 2.12 and Lemma 2.14(ii) in Pakes and Pollard (1989) to see that the kernel of $Q_n(\lambda,\delta)-Q(\lambda,\delta)$ has the requisite Euclidean properties, then apply A11 and Corollary 7 in Sherman (1994). The second term in the last expression has order $o_p(1)$ as $n\to\infty$. This follows from the Cauchy-Schwarz inequality, A10, A11, and a dominated convergence argument. This proves condition (ii).

Condition (iii) also follows from A10, A11, and dominated convergence. This proves the theorem. □

REMARK 1. Condition (14) is the key identifying condition in the consistency proof. Note that independence of the U_i's and X_i's is not needed for this condition to hold. It is sufficient that ΔU and ΔX be either uncorrelated or negatively correlated. For example, if $U_i=\epsilon_i f(X_i)$, where f is an arbitrary real-valued function on $\mathbb{R}^k$ and ϵ_i is uncorrelated with $f(X_i)$, then ΔU and ΔX are uncorrelated. Thus, $\hat\lambda(\hat\delta)$ can be consistent under general forms of heteroscedasticity. In this setting, Powell's (1991) quantile regression estimator, for example, could be used to provide a $\sqrt{n}$-consistent, asymptotically normal first-stage estimator of δ_0.

REMARK 2. Suppose $M(a_{ij})=R_n(a_{ij})$. Expand the rank functions into sums and ignore all terms with equal indices to get

$$Q_n(\lambda,\delta) = \sum_{\boldsymbol{i_6}} \{Z(Y_i,\lambda) - Z(Y_j,\lambda) > Z(Y_k,\lambda) - Z(Y_l,\lambda)\}\{X_i'\delta - X_j'\delta > X_s'\delta - X_t'\delta\}$$

with $\boldsymbol{i_6} = (i,j,k,l,s,t)$ ranging over the $(n)_6 = n(n-1)(n-2)(n-3)(n-4)(n-5)$ ordered 6-tuples of different integers from the set $\{1,\dots,n\}$. For fixed λ and δ, $Q_n(\lambda,\delta)$ is a sixth-order U-statistic. Divide through by $(n)_6$ and take expectations. Denote the result by

$$Q(\lambda,\delta) = \mathbb{E}\{Z(Y_1,\lambda) - Z(Y_2,\lambda) > Z(Y_3,\lambda) - Z(Y_4,\lambda)\}\{X_1'\delta - X_2'\delta > X_5'\delta - X_6'\delta\}\,.$$

Define

$$\begin{aligned} H_{12}(\boldsymbol{Y},\delta_0) &= \mathbb{E}\left[F(\Delta X_{12}(\delta_0)) \mid \boldsymbol{Y}\right] \\ H_{34}(\boldsymbol{Y},\delta_0) &= \mathbb{E}\left[F(\Delta X_{34}(\delta_0)) \mid \boldsymbol{Y}\right] \end{aligned}$$

where F is the cumulative distribution function of $\Delta X_{56}(\delta_0)$. The consistency proof goes through as before with F playing the role of M. Note, however, that A10 and A11 are automatically satisfied for $M = F$.

We now establish the limiting distribution of $\hat{\lambda}(\hat{\delta})$. We consider two cases: M deterministic and $M(a_{ij}) = R_n(a_{ij})$.

We begin with the case M deterministic. Recall that $W = (Y,X)$ and W has distribution P on the set $\mathcal{W} \subseteq \mathbb{R} \otimes \mathbb{R}^k$. For each $w \in \mathcal{W}$, $\lambda \in \Lambda$, and $\delta \in \Delta$, define

$$f(w,\lambda,\delta) \quad = \quad g(w,P,P,P,\lambda,\delta) + g(P,w,P,P,\lambda,\delta)$$

$$+ \quad g(P, P, w, P, \lambda, \delta) + g(P, P, P, w, \lambda, \delta)$$

where, for $w_i = (y_i, x_i) \in \mathcal{W}$, $i = 1, 2, 3, 4$,

$$\begin{aligned} g(w_1, w_2, w_3, w_4, \lambda, \delta) &= \{Z(y_1, \lambda) - Z(y_2, \lambda) > Z(y_3, \lambda) - Z(y_4, \lambda)\} M(x_1'\delta - x_2'\delta) \\ &- \{Z(y_1, \lambda_0) - Z(y_2, \lambda_0) > Z(y_3, \lambda_0) - Z(y_4, \lambda_0)\} M(x_1'\delta - x_2'\delta) \end{aligned}$$

and, as before, $g(w, P, P, P, \lambda, \delta)$, for example, is shorthand for $P \otimes P \otimes P g(w, \cdot, \cdot, \cdot, \lambda, \delta)$.

THEOREM 4: *Suppose M is deterministic. If f is defined as in the previous paragraph and A1 through A11 hold, then*

$$\sqrt{n}(\hat{\lambda}(\hat{\delta}) - \lambda_0) \Longrightarrow N(0, V^{-1} \Sigma V^{-1})$$

where

$$\begin{aligned} V &= \frac{1}{4} \mathbb{E} f_{\lambda\lambda}(\cdot, \lambda_0, \delta_0) \\ \Sigma &= \mathbb{E} \left[f_\lambda(\cdot, \lambda_0, \delta_0) + \tau(\lambda_0, \delta_0)\gamma(\cdot, \delta_0) \right] \left[f_\lambda(\cdot, \lambda_0, \delta_0) + \tau(\lambda_0, \delta_0)\gamma(\cdot, \delta_0) \right]' . \end{aligned}$$

The proof of Theorem 4 is identical to the proof of Theorem 2.

Next, we consider the case $M(a_{ij}) = R_n(a_{ij})$. For each $w \in \mathcal{W}$, $\lambda \in \Lambda$, and $\delta \in \Delta$, define

$$\begin{aligned} f(w, \lambda, \delta) &= g(w, P, P, P, P, P, \lambda, \delta) + g(P, w, P, P, P, P, \lambda, \delta) + g(P, P, w, P, P, P, \lambda, \delta) \\ &+ g(P, P, P, w, P, P, \lambda, \delta) + g(P, P, P, P, w, P, \lambda, \delta) + g(P, P, P, P, P, w, \lambda, \delta) \end{aligned}$$

where, for $w_i = (y_i, x_i) \in \mathcal{W}$, $i = 1, 2, 3, 4, 5, 6$,

$$\begin{aligned} g(w_1, \ldots, w_6, \lambda, \delta) &= \{Z(y_1, \lambda) - Z(y_2, \lambda) > Z(y_3, \lambda) - Z(y_4, \lambda)\}\{x_1'\delta - x_2'\delta > x_5'\delta - x_6'\delta)\} \\ &- \{Z(y_1, \lambda_0) - Z(y_2, \lambda_0) > Z(y_3, \lambda_0) - Z(y_4, \lambda_0)\}\{x_1'\delta - x_2'\delta > x_5'\delta - x_6'\delta)\}\,. \end{aligned}$$

THEOREM 5: *Suppose $M(a_{ij}) = R_n(a_{ij})$. If f is defined as in the previous paragraph and A1 through A9 hold, then*

$$\sqrt{n}(\hat{\lambda}(\hat{\delta}) - \lambda_0) \Longrightarrow N(0, V^{-1}\Sigma V^{-1})$$

where

$$\begin{aligned} V &= \frac{1}{6}\mathbb{E} f_{\lambda\lambda}(\cdot, \lambda_0, \delta_0) \\ \Sigma &= \mathbb{E}\left[f_\lambda(\cdot, \lambda_0, \delta_0) + \tau(\lambda_0, \delta_0)\gamma(\cdot, \delta_0)\right]\left[f_\lambda(\cdot, \lambda_0, \delta_0) + \tau(\lambda_0, \delta_0)\gamma(\cdot, \delta_0)\right]'\,. \end{aligned}$$

Apart from cosmetic differences, the proof of Theorem 5 is identical to that of Theorem 2.

We close this section by noting that the variance-covariance matrices in the last two theorems can be estimated using numerical derivatives as prescribed in Section 3 for Han's estimator.

5. SIMULATIONS

In this section, we present simulation results comparing two of the new rank estimators of λ_0 to Han's estimator, the Bickel and Doksum (1981) parametric estimator under normality, and the semiparametric NL2SLS estimator of Amemiya and Powell (1981).

We choose $M(a_{ij}) = R_n(a_{ij})$ for the first new rank estimator, and $M(t) = t$ for the second. We call these estimators $Rank_1$ and $Rank_2$, respectively. Since the focus of this paper is estimation of

λ_0, we take δ_0 as known and estimate λ_0 in model (3). Specifically, for $\lambda_0 > 0$, we take

$$(|Y|^{\lambda_0}\text{sign}(Y) - 1)/\lambda_0 = X + u\,.$$

We take $\lambda_0 = .5, 1, 2$, the sample size $n = 50, 100, 200$, and the distribution P of u to be normal (N), exponential (E), and gamma with parameters $(.5, 2)$ (G). Thus, G is distributed $\chi^2(1)$. All error distributions are independent of X, centered to zero mean, and rescaled to have standard error $\sqrt{.5}$. In each simulation, the distribution of X is normal with mean zero and standard error $\sqrt{5}$. Thus, the signal to noise ratio in each simulation is about 3 to 1. The number of replications in each simulation is 50. This parallels the simulations done by Han (1987a). In addition, we estimate a contamination model (C). In this model, $P = N$, and for each simulation, we choose one of the n Y values at random and replace it with $2\max_{1\leq i\leq n} Y_i$. For the NL2SLS estimator, we take the instrument matrix W to be the $n \times 2$ matrix consisting of a column of ones and a column of X values (see Amemiya and Powell, 1981, p.356). In each replication, we estimate λ_0 with a naive grid search: we lay down a grid of 500 points spaced .01 units apart on $[0, 5]$ and take $\hat{\lambda}$ to be the maximizer of the appropriate criterion function over these points[1]. When there is an interval of maximizers, we take $\hat{\lambda}$ to be the midpoint of the interval. For each simulation, we compute the mean of the estimators as well as the root mean squared error ($RMSE$) based on the 50 replications. We also compute computation times for each simulation[2]. The results for the estimators of Bickel and Doksum, Han, and the two new rank estimators are given in Table 1. The results for the NL2SLS estimator are given in Table 2.

In models N, E, and G, corresponding to the different error distributions, we see that for sample

[1]For all simulations in which $\lambda_0 = .5$, we lay down a grid of 400 points spaced .005 units apart on $[0, 2]$.

[2]Because of the computational burden involved in computing Han's estimator, we only simulated his estimator for $n = 200$ when $P = C$.

sizes of 100 and 200 and for all choices of λ_0, all the estimators do reasonably well in terms of bias and $RMSE$, with the parametric estimator and NL2SLS comparable and having a slight edge over Han's, Han's having a slight edge over $Rank_2$, and $Rank_2$ having a slight edge over $Rank_1$. Recall that the parametric estimator is calculated under the assumption of normal errors. We find its relatively good performance under all three error distributions surprising. Han (1987a) obtained similar results. Except for the NL2SLS estimator, there is a slight deterioration in performance in terms of bias for all the estimators as skewness in the error distribution increases. However, for the contamination model C, the performance of the parametric estimator degrades substantially for all choices of λ_0 and all sample sizes. The NL2SLS estimator performs better than the parametric estimator, but not as well overall as the rank-based estimators. The rank-based estimators perform well for all choices of λ_0 for sample sizes of 100 and 200. Overall, Han's estimator slightly outperforms $Rank_2$, and $Rank_2$ slightly outperforms $Rank_1$. Han's estimator does well even when $n = 50$.

In terms of computation time, the new rank estimators substantially dominate Han's estimator. For model C it took over 70 hours to perform one simulation for Han's estimator for a sample of size 200, whereas it took only about 12 minutes for either of the new rank estimators. The simulation programs were written in C and were run on a 450 Megahertz Toshiba/Satellite PC with 64 Megabytes of RAM.

We also generated histograms for four of the estimators when $n = 200$ and $P = N$. The results appear in Figure 1. It appears that the normal approximation may be reasonable for this model even for the semiparametric estimators at $n = 200$.

6. Summary

This paper establishes $\sqrt{n}$-consistency and asymptotically normality of Han's (1987a) estimator of the parameters characterizing the transformation function in a semiparametric transformation model. We verify a key Vapnik-Cervonenkis (VC) condition for the parameterizations of Box and Cox (1964) and Bickel and Doksum (1981). The verification establishes the VC property for a class of sets where nonlinear functions of the transformation parameters are positive.

We also introduce a new class of rank estimators for the transformation parameters and establish $\sqrt{n}$-consistency and asymptotic normality. These estimators require only $O(n^2 \log n)$ computations to evaluate the criterion function, compared to $O(n^4)$ computations for Han's estimator. This is achieved by exploiting Spearman's (1904) measure of rank correlation rather than Kendall's (1938) measure, on which Han's (1987a) estimator is based. The former is more computationally efficient.

A simulation study compares the new estimators to Han's estimator, as well as to the fully parametric estimator of Bickel and Doksum (1981) and the semiparametric NL2SLS estimator of Amemiya and Powell (1981). In these simulations, Han's estimator slightly outperforms the new rank estimators in terms of bias and root mean squared error, but takes over 400 times longer to compute for a sample of size 200.

REFERENCES

AHO, A. V., HOPCROFT, J. E., and J. D. ULLMAN (1976): *The Design and Analysis of Computer Algorithms.* Addison-Wesley, Reading, Mass.

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

AMEMIYA, T. (1985): *Advanced Econometrics.* Harvard Univ. Press, Cambridge, Mass.

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

Bickel, P. J. and K. A. Doksum (1981): "An Analysis of Transformations Revisited," *Journal of the American Statistical Association*, 76, 296–311.

Box, G. E. P. and D. R. Cox (1964): "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Series B, 34, 187–220.

Cavanagh, C. L. and R. P. Sherman (1998): "Rank Estimators for Monotonic Index Models," *Journal of Econometrics*, 84, 351–381.

Chen, S. (2001): "Rank Estimation of Transformation Models," *Econometrica*, to appear.

Gørgens, T. and J. Horowitz (1999): "Semiparametric Estimation of a Censored Regression Model with an Unknown Transformation of the Dependent Variable," *Journal of Econometrics*, 90, 155–191.

Han, A. K. (1985): "Nonparametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator," Discussion Paper 1173, Harvard Institute of Economic Research, Cambridge, MA.

Han, A. K. (1987a): "A Nonparametric Analysis of Transformations," *Journal of Econometrics*, 35, 191–209.

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

Horowitz, J. L. (1996): "Semiparametric Estimation of a Regression Model with an Unknown Transformation of the Dependent Variable," *Econometrica*, 64, 103–138.

Kendall, M. G. (1938): "A New Measure of Rank Correlation," *Biometrica*, 30, 81–93.

Klein, R. W. and R. P. Sherman (2001): "Shift Restrictions and Semiparametric Estimation in Ordered Response Models," *Econometrica*, to appear.

LEHMANN, E. L. (1975): *Nonparametrics: Statistical Methods Based on Ranks.* California: Holden Day.

MITTLEHAMMER, R. C., JUDGE, G. C., AND D. J. MILLER (2000): *Econometric Foundations*, Cambridge University Press, Cambridge, MA.

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

POWELL, J. L. (1991): "Estimation of Monotonic Regression Models under Quantile Restrictions," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, 357–384, Edited by Barnett, Powell, and Tauchen, Cambridge University Press, Cambridge, MA.

POWELL, J. L. (1996): "Rescaling Method-of-Moments Estimation for the Box-Cox Regression Model," *Economic Letters*, 51, 259–265.

SAVIN, N. E. AND A. H. WURTZ (2001): "Semiparametric Estimation of the Box-Cox Model," manuscript.

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

SHERMAN, R. P. (1993): "The Limiting Distribution of the Maximum Rank Correlation Estimator," *Econometrica*, 61, 123–137.

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

SPEARMAN, C. (1904): "The Proof and Measurement of Association between Two Things," *Am. J. Psychol.*, 15, 72–101.

YE, J. AND N. DUAN (1997): "Nonparametric $\sqrt{n}$-Consistent Estimation for the General Transformation Models," *Annals of Statistics*, 25, 2682–2717.

NL2SLS Simulation Results

λ_0	n	P	mean	RMSE
0.5	50	N	0.498	0.039
0.5	100	N	0.504	0.027
0.5	200	N	0.502	0.020
0.5	50	E	0.506	0.034
0.5	100	E	0.492	0.028
0.5	200	E	0.495	0.018
0.5	50	G	0.504	0.040
0.5	100	G	0.496	0.018
0.5	200	G	0.501	0.017
0.5	50	C	0.447	0.069
0.5	100	C	0.466	0.069
0.5	200	C	0.480	0.028
1	50	N	1.010	0.066
1	100	N	1.000	0.048
1	200	N	1.006	0.037
1	50	E	1.008	0.074
1	100	E	1.001	0.055
1	200	E	1.006	0.031
1	50	G	1.024	0.065
1	100	G	1.007	0.050
1	200	G	0.998	0.034
1	50	C	0.843	0.177
1	100	C	0.904	0.116
1	200	C	0.964	0.051
2	50	N	1.998	0.083
2	100	N	1.993	0.072
2	200	N	2.010	0.050
2	50	E	2.035	0.123
2	100	E	2.011	0.087
2	200	E	2.009	0.069
2	50	G	2.000	0.081
2	100	G	2.002	0.064
2	200	G	2.013	0.037
2	50	C	1.673	0.357
2	100	C	1.846	0.195
2	200	C	1.928	0.101

Table 2: In all simulations, the number of replications equals 50.