Econometric Theory, **24**, 2008, 749–794. Printed in the United States of America.
doi: 10.1017/S0266466608080316

SEMI-NONPARAMETRIC INTERVAL-CENSORED MIXED PROPORTIONAL HAZARD MODELS: IDENTIFICATION AND CONSISTENCY RESULTS

HERMAN J. BIERENS
Pennsylvania State University

In this paper I propose estimating distributions on the unit interval semi-nonparametrically using orthonormal Legendre polynomials. This approach will be applied to the interval-censored mixed proportional hazard (ICMPH) model, where the distribution of the unobserved heterogeneity is modeled semi-nonparametrically. Various conditions for the nonparametric identification of the ICMPH model are derived. I will prove general consistency results for *M*-estimators of (partly) non-euclidean parameters under weak and easy-to-verify conditions and specialize these results to sieve estimators. Special attention is paid to the case where the support of the covariates is finite.

1. INTRODUCTION

Given a continuous distribution function $G(x)$ with support $\Xi \subset \mathbb{R}$, any distribution function $F(x)$ with support contained in Ξ can be written as $F(x) = H(G(x))$, where $H(u) = F(G^{-1}(u))$ is a distribution function on $[0,1]$. Moreover, if F and G are absolutely continuous with densities f and g, respectively, then H is absolutely continuous with density $h(u)$, and $f(x) = h(G(x))g(x)$. Therefore, $f(x)$ can be estimated semi-nonparametrically by estimating $h(u)$ semi-nonparametrically. The role of G is twofold. First, G determines the support of f. Second, G acts as an initial guess of the unknown distribution function F. Obviously, in the latter case the initial guess is right if H is the cumulative distribution function (c.d.f.) of the uniform distribution on $[0,1]$.

Any density h on the unit interval can be written as $h(u) = \varphi(u)^2$, where φ is a Borel measurable real function on $[0,1]$. It will be shown that square-integrable Borel measurable functions on the unit interval have an infinite series expan-

Previous versions of this paper have been presented at the Universidade Federal do Ceara, Brazil, Texas Econometrics Camp 2005, the 2005 World Congress of the Econometric Society, Ohio State University, Cemapre, Lisbon, and Brown University. The helpful comments of Aleksandr Vashchilko, Yuichi Kitamura, and two referees are gratefully acknowledged. Address correspondence to Herman J. Bierens, Department of Economics, Pennsylvania State University, 608 Kern Graduate Building, University Park, PA 16802, USA; e-mail: hbierens@psu.edu.

© 2008 Cambridge University Press 0266-4666/08 $15.00

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

sion in terms of orthonormal Legendre polynomials. Therefore, the density $h(u)$ and its c.d.f. $H(u)$ can be modeled semi-nonparametrically in a similar way as proposed by Gallant and Nychka (1987), except that instead of Hermite polynomials the Legendre polynomials are used.

This approach will be applied to the mixed proportional hazard (MPH) model, which was proposed by Lancaster (1979), where the duration involved is only observed in the form of intervals. The interval-censored mixed proportional hazard (ICMPH) model involved is motivated by the data used in Bierens and Carvalho (2007b), where the durations involved, job search and recidivism, are only observed in the form of intervals.[1]

Bierens and Carvalho (2007a) have applied the results in this paper to model semi-nonparametrically the common unobserved heterogeneity in a bivariate MPH model of misdemeanor and felony recidivism, where the two durations involved are right-censored only.

The survival function of the MPH model takes the form $H(S(t))$, where $S(t)$ is the survival function of the proportional hazard model without unobserved heterogeneity and H is an absolutely continuous distribution function on the unit interval corresponding to the unobserved heterogeneity distribution. The density h of this distribution function H will be modeled semi-nonparametrically. I will set forth conditions under which the parameters of the systematic and baseline hazards are nonparametrically identified and the sieve maximum likelihood (ML) estimators involved are consistent.

The plan of the paper is as follows. In Section 2 the Legendre polynomials will be introduced and their use motivated. In Section 3 I will show how density and distribution functions on the unit interval can be represented by linear combinations of Legendre polynomials. In Section 4 I will discuss the ICMPH model, and in Section 5 I will derive conditions for nonparametric identification under interval-censoring. Because of the latter, the identification conditions and their derivations in Elbers and Ridder (1982) and Heckman and Singer (1984) are not directly applicable and have to be rederived for the interval-censored case. Heckman and Singer (1984) derive identification conditions by verifying the more general conditions in Kiefer and Wolfowitz (1956). As shown by Meyer (1995), the results in the latter paper can also be used to prove identification and consistency in the case of interval-censoring, provided that the systematic hazard has support $(0,\infty)$, which is the case if at least one covariate has as support the whole real line and has a nonzero coefficient. However, I will in this case derive the identification and consistency conditions directly without using the results of Kiefer and Wolfowitz (1956), for two related reasons. First, as is apparent from Meyer (1995), it is very complicated to link the conditions in Kiefer and Wolfowitz (1956) to the ICMPH model. Second, it is actually much easier and more transparent to derive these conditions directly.

In Section 6 I will sketch the requirements for consistency of the semi-nonparametric (SNP) ML sieve estimators. One of the requirements is that the space of density functions h involved is compact. Therefore, in Section 7 I will

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

show how to construct compact metric spaces of densities and distribution functions on the unit interval. In Section 8 I will prove general consistency results for *M*-estimators of (partly) non-euclidean parameters under weak and easy-to-verify conditions and specialize these results to ICMPH models.

One of the key conditions for nonparametric identification[2] in the interval-censored case is that at least one covariate has the whole real line $\mathbb{R}$ as support and has a nonzero coefficient, so that the systematic hazard has support $(0,\infty)$. However, in practice this condition is often not satisfied. Therefore, in Section 9 I will discuss the case where the covariates have finite support. In Section 10 the main contributions of this paper are summarized, and avenues for future research are indicated. Finally, proofs are only presented in the main text if they are essential for the flow of the argument. All other proofs are given in the Appendix.

2. ORTHONORMAL POLYNOMIALS

2.1. Hermite Polynomials

Gallant and Nychka (1987) consider SNP estimation of the Heckman (1979) sample selection model, where the bivariate error distribution of the latent variable equations is modeled semi-nonparametrically using a Hermite expansion of the error density. In the case of a density $f(x)$ on $\mathbb{R}$ this Hermite expansion takes the form

$$f(x) = \phi(x)\left(\sum_{k=0}^{\infty} \gamma_k \mu_k(x)\right)^2, \tag{1}$$

with $\sum_{k=0}^{\infty} \gamma_k^2 = 1$, where $\phi(x)$ is the standard normal density and the $\mu_k(x)$'s are Hermite polynomials satisfying $\int_{-\infty}^{\infty} \mu_k(x)\mu_m(x)\phi(x)\,dx = I(k = m)$, where $I(\cdot)$ is the indicator function. These polynomials can easily be generated via the recursive relation $\sqrt{n}\,\mu_n(x) = x\cdot\mu_{n-1}(x) - \sqrt{n-1}\,\mu_{n-2}(x)$, starting from $\mu_0(x) = 1$, $\mu_1(x) = x$. See, e.g., Hamming (1973, p. 457). The densities (1) can be approximated arbitrarily closely by SNP densities of the type

$$f_n(x) = \phi(x)\left(\sum_{k=0}^{n} \gamma_{k,n} \mu_k(x)\right)^2, \tag{2}$$

where $\sum_{k=0}^{n} \gamma_{k,n}^2 = 1$.

Because $f_0(x) = \phi(x)$, the Hermite expansion is particularly suitable for generalizations of the normal density. For example, a natural generalization of the probit model is $P[Y = 1|X] = F_n(\beta'X)$, where

$$F_n(x) = \int_{-\infty}^{x} f_n(z)\,dz = \sum_{k=0}^{n}\sum_{m=0}^{n} \gamma_{k,n}\gamma_{m,n} \int_{-\infty}^{x} \phi(z)\mu_k(z)\mu_m(z)\,dz, \tag{3}$$

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

which yields the standard probit model as a special case, corresponding to the null hypothesis $\gamma_{k,n} = 0$ for $k = 1, \ldots, n$. If the actual distribution function $F(x)$ in the general binary response model $P[Y = 1|X] = F(\beta'X)$ does not deviate too much from the probit function, a low value of n may suffice to give a good approximation. The same applies to generalizations of the ordered probit model based on Hermite polynomials.

Of course, the SNP density $f_n(x)$ can be transformed to a density $h_n(u)$ on the unit interval such that $h_0(u) = 1$, namely,

$$h_n(u) = f_n(F_0^{-1}(u))/\phi(F_0^{-1}(u)),$$

with corresponding distribution function

$$H_n(u) = F_n(F_0^{-1}(u)), \tag{4}$$

where f_n is the SNP density (2), F_n is the corresponding c.d.f. (3), and $F_0^{-1}(u)$ is the inverse of the probit function $F_0(x) = \int_{-\infty}^{x} \phi(z)\, dz$. However, the probit function does not have a closed form, and neither does its inverse.

One of the advantages of the Hermite polynomials is that the integrals in (3) can easily be computed recursively. However, SNP distribution functions on the unit interval based on ordinary polynomials are even easier to compute, as

$$H_n(u) = \frac{\eta' \Pi_{n+1}(u) \eta}{\eta' \Pi_{n+1}(1) \eta}, \qquad \eta \in \mathbb{R}^{n+1}, \tag{5}$$

where $\Pi_{n+1}(u)$ is an $(n+1) \times (n+1)$ matrix with typical element $\int_0^1 u^{i+j}\, du$:

$$\Pi_{n+1}(u) = \left(\frac{u^{i+j+1}}{i+j+1} \, ; i,j = 0,1,\ldots,n \right). \tag{6}$$

This SNP distribution function is a reparametrization of an equivalent one based on orthonormal Legendre polynomials, similar to (3). The Legendre polynomials play mainly a theoretical role, namely, to show that for an arbitrary absolutely continuous distribution function $H(u)$ on $[0,1]$ there exists a sequence of SNP distribution functions of the type (5) such that $\lim_{n\to\infty} \sup_{0\le u\le 1} |H_n(u) - H(u)| = 0$ and to derive a compact metric space of such distribution functions $H(u)$, endowed with the "sup" metric.

2.2. Legendre Polynomials

A convenient way to construct orthonormal polynomials on $[0,1]$ is to base them on Legendre polynomials $P_n(z)$ on $[-1,1]$. For $n \ge 2$ these polynomials can be constructed recursively by

$$P_n(z) = \frac{(2n-1)zP_{n-1}(z) - (n-1)P_{n-2}(z)}{n}, \tag{7}$$

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

starting from $P_0(z) = 1$, $P_1(z) = z$. They are orthogonal but not orthonormal:

$$\int_{-1}^{1} P_m(z)P_n(z)\,dz = \begin{cases} 0 & \text{if } n \neq m, \\ 2/(2n+1) & \text{if } n = m. \end{cases} \tag{8}$$

See, e.g., Hamming (1973, p. 455).

Now define for $u \in [0,1]$, $\rho_n(u) = \sqrt{2n+1}P_n(2u-1)$. Then it follows from (8) that the polynomials $\rho_n(u)$ are orthonormal:

$$\int_0^1 \rho_k(u)\rho_m(u)\,du = \begin{cases} 0 & \text{if } k \neq m, \\ 1 & \text{if } k = m \end{cases} \tag{9}$$

and from (7) that for $n \geq 2$ they can be computed recursively by

$$\rho_n(u) = \frac{\sqrt{4n^2-1}}{n}(2u-1)\rho_{n-1}(u) - \frac{(n-1)\sqrt{2n+1}}{n\sqrt{2n-3}}\rho_{n-2}(u), \tag{10}$$

starting from

$$\rho_0(u) = 1, \qquad \rho_1(u) = \sqrt{3}(2u-1). \tag{11}$$

3. DENSITY AND DISTRIBUTION FUNCTIONS ON THE UNIT INTERVAL

3.1. Polynomial Representation

Every density function $h(u)$ on $[0,1]$ can be written as $h(u) = f(u)^2$, where $\int_0^1 f(u)^2\,du = 1$. In Theorem 1, which follows, I will focus on the characterization of square-integrable functions on $[0,1]$ in terms of the Legendre polynomials $\rho_k(u)$ and then specialize the result involved to densities on $[0,1]$.

THEOREM 1. *Let $f(u)$ be a Borel measurable function on $[0,1]$ such that $\int_0^1 f(u)^2\,du < \infty$[3] and let $\gamma_k = \int_0^1 \rho_k(u)f(u)\,du$. Then $\sum_{k=0}^{\infty}\gamma_k^2 < \infty$, and the set $\{u \in [0,1]: f(u) \neq \sum_{k=0}^{\infty}\gamma_k\rho_k(u)\}$ has Lebesgue measure zero. In other words, the Legendre polynomials $\rho_k(u)$ on $[0,1]$ form a complete orthonormal basis for the Hilbert space $L^2_{\mathcal{B}}(0,1)$ of Borel measurable square-integrable real functions on $[0,1]$.*

Recall[4] that, more generally, the real Hilbert space $L^2(0,1)$ is the space of square-integrable Lebesgue measurable real functions on $[0,1]$; i.e., $f \in L^2(0,1)$ implies $\int_0^1 f(u)^2\,du < \infty$, endowed with the inner product $\langle f, g\rangle = \int_0^1 f(u)g(u)\,du$ and associated metric

$$\|f-g\|_2 = \sqrt{\int_0^1 (f(u)-g(u))^2\,du} \tag{12}$$

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

and norm $\|f\|_2$. Moreover, recall that Borel measurable functions are Lebesgue measurable because Borel sets are Lebesgue measurable sets.[5] Therefore, the subspace $L^2_{\mathcal{B}}(0,1)$ of Borel measurable real functions in $L^2(0,1)$ is a Hilbert space itself.

Proof of Theorem 1. The proof of Theorem 1 is based on the following straightforward corollary of Theorem 2 in Bierens (1982).

LEMMA 1. *Let $f_1(u)$ and $f_2(u)$ be Borel measurable real functions on $[0,1]$ such that*

$$\int_0^1 |f_1(u)|\,du < \infty, \qquad \int_0^1 |f_2(u)|\,du < \infty. \tag{13}$$

Then the set $\{u \in [0,1]: f_1(u) \neq f_2(u)\}$ has Lebesgue measure zero if and only if for all nonnegative integers k,

$$\int_0^1 u^k f_1(u)\,du = \int_0^1 u^k f_2(u)\,du. \tag{14}$$

Each u^k can be written as a linear combination of $\rho_0(u), \rho_1(u), \ldots, \rho_k(u)$ with Fourier coefficients $\int_0^1 u^k \rho_m(u)\,du$, $m = 0,1,\ldots,k$; hence condition (14) is equivalent to $\int_0^1 \rho_k(u) f_1(u)\,du = \int_0^1 \rho_k(u) f_2(u)\,du$ for $k = 0,1,2,\ldots$.

Now in Lemma 1 let $f_1(u) = f(u)$ and $f_2(u) = \sum_{k=0}^{\infty} \gamma_k \rho_k(u)$, where $\gamma_k = \int_0^1 \rho_k(u) f(u)\,du$. Then

$$\sum_{k=0}^{\infty} \gamma_k^2 < \infty \tag{15}$$

because $\int_0^1 (f(u) - \sum_{k=0}^{n} \gamma_k \rho_k(u))^2\,du$ is minimal for $\gamma_k = \int_0^1 \rho_k(u) f(u)\,du$, so that for all natural numbers n, $\sum_{k=0}^{n} \gamma_k^2 \leq \int_0^1 f(u)^2\,du < \infty$. The existence of $f_2(u)$ follows from the fact that, because of (15), $f_{2,n}(u) = \sum_{k=0}^{n} \gamma_k \rho_k(u)$ is a Cauchy sequence in the Hilbert space $L^2_{\mathcal{B}}(0,1)$ and therefore has a limit $f_2 \in L^2_{\mathcal{B}}(0,1)$: $\lim_{n\to\infty} \int_0^1 (f_{2,n}(u) - f_2(u))^2\,du = 0$. Hence, we can write $f_2(u) = \sum_{k=0}^{n} \gamma_k \rho_k(u) + r_n(u)$, where $\lim_{n\to\infty} \int_0^1 r_n(u)^2\,du = 0$. Next, choose a subsequence n_m such that $\sum_{m=1}^{\infty} \int_0^1 r_{n_m}(u)^2\,du < \infty$. Then it follows from the Borel–Cantelli lemma[6] and Chebyshev's inequality that $f_2(u) = \lim_{m\to\infty} \sum_{k=0}^{n_m} \gamma_k \rho_k(u)$ almost everywhere (a.e.) on $(0,1)$.

Moreover, it follows from Liapounov's inequality and the orthonormality of the $\rho_k(u)$'s that $\int_0^1 |f_2(u)|\,du \leq \sqrt{\int_0^1 f_2(u)^2\,du} = \sqrt{\sum_{k=0}^{\infty} \gamma_k^2} < \infty$. Similarly, it follows from the condition $\int_0^1 f(u)^2\,du < \infty$ in Theorem 1 that $\int_0^1 |f_1(u)|\,du < \infty$. Therefore, all the conditions of Lemma 1 are satisfied, which proves Theorem 1. ■

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

Every density function $h(u)$ on $[0,1]$ is Borel measurable because with H the corresponding distribution function $h(u) = \lim_{k\to\infty} k(H(u + k^{-1}) - H(u))$, which is a pointwise limit of a sequence of continuous (hence Borel measurable) functions and therefore Borel measurable itself. Consequently, every density function h on $[0,1]$ can be written as $h(u) = f(u)^2$, where $f(u)$ is a Borel measurable real function on $[0,1]$ satisfying $\int_0^1 f(u)^2\,du = 1$.

Of course, this representation is not unique, as we may replace $f(u)$ by $f(u)\phi_B(u)$, where for arbitrary Borel subsets B of $[0,1]$ with complement $\tilde{B} = [0,1]\backslash B$,

$$\phi_B(u) = I(u \in B) - I(u \in \tilde{B}). \tag{16}$$

This is a simple function; hence $f(u)\phi_B(u)$ is Borel measurable. Therefore, any Borel measurable function f on $[0,1]$ for which $h(u) = f(u)^2$ is a density on $[0,1]$ can be written as $f(u) = \phi_B(u)\sqrt{h(u)}$, with $\phi_B(u)$ a simple function of the type (16). Consequently, any density $h(u)$ on $[0,1]$ can be represented by

$$h(u) = \left(\sum_{k=0}^{\infty} \gamma_k \rho_k(u)\right)^2, \quad \text{with } \gamma_k = \int_0^1 \rho_k(u)\phi_B(u)\sqrt{h(u)}\,du. \tag{17}$$

We can always choose B such that

$$\gamma_0 = \int_0^1 \phi_B(u)\sqrt{h(u)}\,du = \int_B \sqrt{h(u)}\,du - \int_{\tilde{B}} \sqrt{h(u)}\,du > 0. \tag{18}$$

This is useful, because it allows us to get rid of the restriction $\sum_{k=0}^{\infty} \gamma_k^2 = 1$ by reparametrizing the γ_k's as

$$\gamma_k = \frac{\delta_k}{\sqrt{1 + \sum_{k=1}^{\infty} \delta_k^2}}, \qquad k = 1,2,3,\ldots, \tag{19}$$

$$\gamma_0 = \frac{1}{\sqrt{1 + \sum_{k=1}^{\infty} \delta_k^2}},$$

where $\sum_{k=1}^{\infty} \delta_k^2 < \infty$. However, because there are uncountably many Borel subsets B of $[0,1]$ for which (18) holds, there are also uncountably many of such reparametrizations. Thus, we have the following result.

THEOREM 2. *For every density function $h(u)$ on $[0,1]$ there exist uncountably many infinite sequences $\{\delta_k\}_1^{\infty}$ satisfying $\sum_{k=1}^{\infty} \delta_k^2 < \infty$ such that*

$$h(u) = \frac{\left(1 + \sum_{k=1}^{\infty} \delta_k \rho_k(u)\right)^2}{1 + \sum_{k=1}^{\infty} \delta_k^2} \quad \textit{a.e. on } [0,1]. \tag{20}$$

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

3.2. SNP Density Functions on the Unit Interval

For a density $h(u)$ with one of the associated sequences $\{\delta_k\}_1^\infty$, let

$$h_n(u) = h_n(u|\delta) = \frac{\left(1 + \sum_{k=1}^{n} \delta_k \rho_k(u)\right)^2}{1 + \sum_{k=1}^{n} \delta_k^2}, \qquad \delta = (\delta_1, \ldots, \delta_n)'. \tag{21}$$

It is straightforward to verify the following result.

THEOREM 3. *For each density $h(u)$ on $[0,1]$ there exists a sequence of densities $h_n(u)$ of the type (21) such that $\lim_{n\to\infty} \int_0^1 |h(u) - h_n(u)|\,du = 0$. Consequently, for every absolutely continuous distribution function $H(u)$ on $[0,1]$ there exists a sequence of absolutely continuous distribution functions $H_n(u) = \int_0^u h_n(v)\,dv$ such that $\lim_{n\to\infty} \sup_{0\le u\le 1} |H_n(u) - H(u)| = 0$.*

Following Gallant and Nychka (1987), the density functions of the type (21) with a finite n will be called SNP density functions, and the corresponding distribution functions $H_n(u) = \int_0^u h_n(v)\,dv$ will be called SNP distribution functions.

As we have seen in Theorem 2, the densities (20) have uncountably many equivalent series representations. This is no longer the case for SNP densities: see Theorem 4, which follows.

THEOREM 4. *The parametrization of the SNP densities is unique in the sense that if for a pair $\delta_1, \delta_2 \in \mathbb{R}^n$, $h_n(u|\delta_1) = h_n(u|\delta_2)$ a.e. on a subset of $[0,1]$ with positive Lebesgue measure, then $\delta_1 = \delta_2$.*

This result follows easily from the fact that the number of roots of a polynomial of order n cannot exceed n; hence if two polynomials on $[0,1]$ are equal a.e. on a subset with positive Lebesgue measure, then they are equal a.e. on $[0,1]$.

3.3. Computation of SNP Distribution Functions on the Unit Interval

The distribution function $H_n(u|\delta) = \int_0^u h_n(v|\delta)\,dv$, with h_n given by (21), can be written as

$$H_n(u|\delta) = \frac{\int_0^u \left(1 + \sum_{m=1}^{n} \delta_m \rho_m(v)\right)^2 dv}{1 + \sum_{m=1}^{n} \delta_m^2} = \frac{(1, \delta')A_{n+1}(u)\begin{pmatrix}1\\ \delta\end{pmatrix}}{1 + \delta'\delta}, \tag{22}$$

$$u \in [0,1], \qquad \delta = (\delta_1, \ldots, \delta_n)',$$

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

where $A_{n+1}(u)$ is the $(n+1)\times(n+1)$ matrix

$$A_{n+1}(u)=\left(\int_0^u \rho_i(v)\rho_j(v)\,dv;\ i,j=0,1,\ldots,n\right). \tag{23}$$

Let $\rho_m(u)=\sum_{k=0}^m \ell_{m,k}u^k$. Then it follows from (10) and (11) that

$$\ell_{0,0}=1,\qquad \ell_{1,0}=-\sqrt{3},\qquad \ell_{1,1}=2\sqrt{3} \tag{24}$$

and for $m\geq 2$,

$$\sum_{k=0}^m \ell_{m,k}u^k=\frac{2\sqrt{4m^2-1}}{m}\sum_{k=1}^m \ell_{m-1,k-1}u^k-\frac{\sqrt{4m^2-1}}{m}\sum_{k=0}^{m-1}\ell_{m-1,k}u^k$$
$$-\frac{(m-1)\sqrt{2m+1}}{m\sqrt{2m-3}}\sum_{k=0}^{m-2}\ell_{m-2,k}u^k.$$

Hence, letting $\ell_{m,k}=0$ for $k>m$ and $k<0$, the coefficients $\ell_{m,k}$ can be computed recursively by

$$\ell_{m,k}=\frac{\sqrt{4m^2-1}}{m}(2\ell_{m-1,k-1}-\ell_{m-1,k})-\frac{(m-1)\sqrt{2m+1}}{m\sqrt{2m-3}}\ell_{m-2,k},$$

starting from (24). For $0\leq m\leq n$, $0\leq k\leq n$ the coefficients $\ell_{m,k}$ can be arranged as the elements of a lower triangular $(n+1)\times(n+1)$ matrix L_{n+1}, with mth row $(\ell_{m,0},\ldots,\ell_{m,n})$.

Next, observe that

$$\int_0^u \rho_k(v)\rho_m(v)\,dv=(\ell_{k,0},\ldots,\ell_{k,n})\Pi_{n+1}(u)\begin{pmatrix}\ell_{m,0}\\ \vdots\\ \ell_{m,n}\end{pmatrix},$$

where $\Pi_{n+1}(u)$ is the $(n+1)\times(n+1)$ matrix (6). Therefore,

$$H_n(u|\delta)=\frac{(1,\delta')L_{n+1}\Pi_{n+1}(u)L'_{n+1}\begin{pmatrix}1\\ \delta\end{pmatrix}}{1+\delta'\delta}, \tag{25}$$
$$u\in[0,1],\qquad \delta=(\delta_1,\ldots,\delta_n)'.$$

In practice, the lower triangular matrix L_{n+1} can only be computed with sufficient accuracy[7] up to about $n=15$.

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

However, rather than computing L_{n+1}, let

$$\eta = (\eta_0, \eta_1, \ldots, \eta_n)' = L'_{n+1}\begin{pmatrix}1\\ \delta\end{pmatrix}, \tag{26}$$

so that the reparametrized SNP distribution function (25) takes the form (5).

The equality (26) imposes a linear restriction on η, so that we cannot choose η freely. To determine this restriction, link δ to η via the equalities

$$\int_0^1 \left(1 + \sum_{m=1}^{n} \delta_m \rho_m(u)\right)\rho_k(u)\,du = \int_0^1 \left(\eta_0 + \sum_{m=1}^{n} \eta_m u^m\right)\rho_k(u)\,du,$$

$$k = 0, 1, \ldots, n,$$

which yields

$$\eta_0 = 1 - \sum_{m=1}^{n} \frac{\eta_m}{1+m}, \tag{27}$$

$$\delta_k = \sum_{m=0}^{n-k} \eta_{k+m}\,\mu_k(m), \qquad k = 1, \ldots, n, \tag{28}$$

where

$$\mu_n(m) = \int_0^1 u^{n+m}\rho_n(u)\,du. \tag{29}$$

Because of (27), (5) can now be written as

$$H_n(u|\eta) = \frac{(1 - \pi'_n\eta, \eta')\Pi_{n+1}(u)\begin{pmatrix}1-\pi'_n\eta\\ \eta\end{pmatrix}}{(1 - \pi'_n\eta, \eta')\Pi_{n+1}(1)\begin{pmatrix}1-\pi'_n\eta\\ \eta\end{pmatrix}}, \qquad \eta = (\eta_1, \ldots, \eta_n)', \tag{30}$$

where $\pi_n = (\frac{1}{2}, \frac{1}{3}, \ldots, 1/(n+1))'$.

4. THE INTERVAL-CENSORED MIXED PROPORTIONAL HAZARD MODEL

4.1. The Mixed Proportional Hazard Model

Let T be a duration and let X be a vector of covariates. As is well known,[8] the conditional hazard function is defined as $\lambda(t|X) = f(t|X)/(1 - F(t|X))$, where

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

$F(t|X) = P[T \leq t|X]$, $f(t|X)$ is the corresponding conditional density function, and $\int_0^\infty \lambda(\tau|X)\,d\tau = \infty$. Then the conditional survival function is

$$S(t|X) = 1 - F(t|X) = \exp\left(-\int_0^t \lambda(\tau|X)\,d\tau\right).$$

The MPH model assumes that the conditional survival function takes the form

$$S(t|X,\alpha,\beta) = S(t|X) = \mathrm{E}\left[\exp\left(-\exp(\beta'X + U)\int_0^t \lambda(\tau|\alpha)\,d\tau\right)\middle| X\right], \quad \textbf{(31)}$$

where U represents unobserved heterogeneity, which is independent of X, $\lambda(t|\alpha)$ is the baseline hazard function depending on a parameter (vector) α, and $\exp(\beta'X)$ is the systematic hazard function. See Lancaster (1979). Denoting the distribution function of $V = \exp(U)$ by $G(v)$ and the integrated baseline hazard by $\Lambda(t|\alpha) = \int_0^t \lambda(\tau|\alpha)\,d\tau$, we have

$$\begin{aligned} S(t|X,\alpha,\beta,H) &= \int_0^\infty \exp(-v.\exp(\beta'X)\Lambda(t|\alpha))\,dG(v) \\ &= \int_0^\infty (\exp(-\exp(\beta'X)\Lambda(t|\alpha)))^v\,dG(v) \\ &= H(\exp(-\exp(\beta'X)\Lambda(t|\alpha))), \end{aligned} \quad \textbf{(32)}$$

where

$$H(u) = \int_0^\infty u^v\,dG(v), \qquad u \in [0,1], \quad \textbf{(33)}$$

is a distribution function on $[0,1]$.

If the unobserved heterogeneity variable V satisfies $\mathrm{E}[V] < \infty$ then for $u \in (0,1]$,

$$\int_0^\infty v u^{v-1}\,dG(v) \leq u^{-1}\int_0^\infty v\,dG(v) < \infty, \quad \textbf{(34)}$$

so that by the mean value and dominated convergence theorems, $H(u)$ is differentiable on $(0,1)$ with density function

$$h(u) = \int_0^\infty v u^{v-1}\,dG(v). \quad \textbf{(35)}$$

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

Moreover, (34) implies that $h(u)$ is finite and continuous[9] on $(0,1]$. Furthermore, note that absence of unobserved heterogeneity, i.e., $P[V = 1] = 1$, is equivalent to the case $h(u) \equiv 1$.

Let the true conditional survival function be

$$S(t|X,\alpha_0,\beta_0,H_0) = \int_0^\infty \exp(-v\cdot \exp(\beta_0' X)\Lambda(t|\alpha_0))\, dG_0(v)$$
$$= H_0(\exp(-\exp(\beta_0' X)\Lambda(t|\alpha_0))), \tag{36}$$

where

$$H_0(u) = \int_0^u h_0(v)\, dv = \int_0^\infty u^v\, dG_0(v). \tag{37}$$

In the expressions (32) and (36), H and H_0 should be interpreted as unknown parameters contained in a parameter space $\mathcal{H}(0,1)$, say, of absolutely continuous distribution functions on $[0,1]$.

Elbers and Ridder (1982) have shown that if X does not contain a constant, $\Lambda(t|\alpha) = \Lambda(t|\alpha_0)$ for all $t > 0$ implies $\alpha = \alpha_0$, and $\int_0^\infty v dG_0(v) = \int_0^\infty v dG(v) = 1$ (which by (35) is equivalent to confining the parameter space $\mathcal{H}(0,1)$ to a space of distribution functions H on $[0,1]$ with density h satisfying $h(1) = 1$), then the MPH model is nonparametrically identified, in the sense that $S(T|X,\alpha,\beta,H) = S(T|X,\alpha_0,\beta_0,H_0)$ almost surely (a.s.) implies $\alpha = \alpha_0$, $\beta = \beta_0$, and $G = G_0$, hence $H = H_0$. Heckman and Singer (1984) provide an alternative identification proof based on the results of Kiefer and Wolfowitz (1956) and propose parametrizing G_0 as a discrete distribution: $G_0(v) = \sum_{i=1}^q I(v \le \theta_i)p_i$, with $I(\cdot)$ the indicator function, where $\theta_i > 0$, $p_i > 0$, and $\sum_{i=1}^q p_i = 1$. Thus, they implicitly specify $h_0(u) = \sum_{i=1}^q \theta_i u^{\theta_i - 1} p_i$.

Magder and Zeger (1996) propose modeling mixing distributions in models with unobserved heterogeneity as mixtures of normal distributions. They also use the results of Kiefer and Wolfowitz (1956). Although these authors do not explicitly consider MPH models, their approach is applicable to MPH models also.

The nonparametric identification of the MPH model hinges on the assumption that T is observed directly if T is not right-censored. In this paper I will consider the case that T is only observed in the form of intervals, so that the identification results in Elbers and Ridder (1982) and Heckman and Singer (1984) are not directly applicable.

Recall that the interval-censored case has been considered before by Meyer (1995), who derived identification and consistency conditions based on the results of Kiefer and Wolfowitz (1956). However, in this paper I will derive these conditions directly without using the Kiefer and Wolfowitz (1956) results, because that is much easier and more transparent, as comparison of the results that follow with Meyer (1995) will reveal.

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

4.2. Interval Censoring

Let $\{T_j, C_j, X_j\}_{j=1}^N$ be a random sample of possibly censored durations T_j, with corresponding censoring dummy variable C_j and vector X_j of covariates. The actual duration is a latent variable $T_j^* > 0$ with conditional survival function

$$P[T_j^* > t|X_j] = S(t|X_j, \alpha_0, \beta_0, H_0), \tag{38}$$

where $S(t|X, \alpha, \beta, H)$ is defined by (36), α_0 and β_0 are the true parameter vectors, and H_0 is the true c.d.f. (37). If $C_j = 0$ then T_j^* is observed: $T_j = T_j^*$, and if $C_j = 1$ then T_j^* is censored: $T_j = \overline{T}_j < T_j^*$, where $[1, \overline{T}_j]$ is the time interval over which individual j has been, or would have been, monitored. It will be assumed that $\overline{T}_j$ is entirely determined by the setup of the survey, and may therefore be considered exogenous, and that $\overline{T} = \inf_{j\geq 1} \overline{T}_j > 0$.

In practice the observed durations T_j are always measured in discrete units (days, weeks, months, etc.), so that we should not treat them as continuous random variables. Therefore, pick M positive numbers $b_1 < b_2 < \cdots < b_M \leq \overline{T}$ and create the dummy variables

$$\begin{aligned} D_{1,j} &= I(T_j \leq b_1), \\ D_{2,j} &= I(b_1 < T_j \leq b_2), \\ &\vdots \\ D_{M,j} &= I(b_{M-1} < T_j \leq b_M), \end{aligned} \tag{39}$$

where $I(\cdot)$ is the indicator function. See Meyer (1995). Also, in some cases the durations T_j are only observed in the form of intervals. See, e.g., Bierens and Carvalho (2007b). Given that the unobserved durations T_j are distributed according to a MPH model, the resulting model for (39) will be called an interval-censored MPH (ICMPH) model.

For notational convenience, let $b_0 = 0$ and denote for $i = 0, 1, \ldots, M$

$$\mu_i(\alpha, \beta' X_j) = \exp(-\exp(\beta' X_j)\Lambda(b_i|\alpha)). \tag{40}$$

Note that $\mu_0(\alpha, \beta' X_j) = 1$. Then

$$\begin{aligned} P[D_{i,j} = 1|X_j] &= S(b_{i-1}|X_j, \alpha_0, \beta_0, H_0) - S(b_i|X_j, \alpha_0, \beta_0, H_0) \\ &= H_0(\mu_{i-1}(\alpha_0, \beta_0' X_j)) - H_0(\mu_i(\alpha_0, \beta_0' X_j)), \\ & \qquad\qquad i = 1, 2, \ldots, M, \end{aligned}$$

$$P\left[\sum_{i=1}^M D_{i,j} = 0|X_j\right] = S(b_M|X_j, \alpha_0, \beta_0, H_0) = H_0(\mu_M(\alpha_0, \beta_0' X_j)). \tag{41}$$

The c.d.f. H_0 will be treated as a parameter.

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

The conditional log-likelihood function of the ICMPH model takes the form

$$\ln(L_N(\alpha,\beta,H)) = \sum_{j=1}^{N}\sum_{i=1}^{M} D_{i,j}\ln(H(\mu_{i-1}(\alpha,\beta'X_j)) - H(\mu_i(\alpha,\beta'X_j))) + \sum_{j=1}^{N}\left(1 - \sum_{i=1}^{M} D_{i,j}\right)\ln(H(\mu_M(\alpha,\beta'X_j))). \quad (42)$$

4.3. Baseline Hazard Specification

Note that we do not need to specify the integrated baseline hazard $\Lambda(t|\alpha)$ completely for all $t > 0$. It suffices to specify $\Lambda(t|\alpha)$ only for $t = b_1,\ldots,b_M$. Therefore we may without loss of generality parametrize $\Lambda(t|\alpha)$ as a piecewise linear function:

$$\Lambda(t|\alpha) = \Lambda(b_{i-1}|\alpha) + \alpha_i(t - b_{i-1}) = \sum_{k=1}^{i-1}\alpha_k(b_k - b_{k-1}) + \alpha_i(t - b_{i-1}) \quad \text{for } t \in (b_{i-1}, b_i],$$

$$\alpha_m > 0 \quad \text{for } m = 1,\ldots,M, \qquad \alpha = (\alpha_1,\ldots,\alpha_M)' \in \mathbb{R}^M. \quad (43)$$

There are of course other equivalent ways to specify $\Lambda(b_i|\alpha)$. For example, let

$$\Lambda(b_i|\alpha) = \sum_{m=1}^{i}\alpha_m, \qquad \alpha_m > 0 \quad \text{for } m = 1,\ldots,M, \quad (44)$$

or

$$\Lambda(b_i|\alpha) = \exp(\alpha_i), \qquad \alpha_1 < \alpha_2 < \cdots < \alpha_M, \quad (45)$$

$$\Lambda(b_0|\alpha) = \Lambda(0|\alpha) = 0.$$

The advantage of the specification (43) is that the null hypothesis $\alpha_1 = \cdots = \alpha_M$ corresponds to the constant hazard $\lambda(t|\alpha) = \alpha_1$. In that case $\exp(\beta'X)$ $\Lambda(t|\alpha) = \exp(\ln(\alpha_1) + \beta'X)t$, so that $\ln(\alpha_1)$ acts as a constant term in the systematic hazard.

However, the specification (44) is useful for deriving identification conditions, as will be shown in the next section. The same applies to (45).

Note that under specification (45) the probability model (41) takes the form of a generalized ordered probability model, similarly to an ordered probit or logit model:

$$P[\Sigma_{i=1}^{m} D_{i,j} = 1|X_j] = F_0(\beta_0'X_j + \alpha_{0,m}), \qquad m = 1,2,\ldots,M, \quad (46)$$

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

where

$$F_0(x) = 1 - H_0(\exp(-\exp(x))), \tag{47}$$

with density

$$f_0(x) = h_0(\exp(-\exp(x)))\exp(-\exp(x))\exp(x). \tag{48}$$

This case makes clear that we cannot allow a constant in the vector X_j, because the constant can be absorbed by the $\alpha_{0,i}$'s in (46). Moreover, for the identification of α_0 and β_0 in (46) it is necessary to normalize the location and scale of the distribution $F_0(x)$.

5. NONPARAMETRIC IDENTIFICATION OF THE ICMPH MODEL WITH CONTINUOUSLY DISTRIBUTED COVARIATES

Let us assume that $\Lambda(b_i|\alpha)$ is specified as (44). It follows easily from the inequality $\ln(x) < x - 1$ if $x > 0$ and $x \neq 1$ and the equality

$$\mathrm{E}[L_N(\alpha,\beta,H)/L_N(\alpha_0,\beta_0,H_0)|X_1,\ldots,X_N] = 1 \quad \text{a.s.}$$

that

$$\mathrm{E}[N^{-1}\ln(L_N(\alpha,\beta,H)/L_N(\alpha_0,\beta_0,H_0))|X_1,\ldots,X_N] < 0$$

if and only if for some $i \in \{1,\ldots,M\}$

$$P\left[H\left(\exp\left(-\exp(\beta' X_j)\left(\sum_{m=1}^{i}\alpha_m\right)\right)\right)\right.$$
$$\left. = H_0\left(\exp\left(-\exp(\beta_0' X_j)\left(\sum_{m=1}^{i}\alpha_{0,m}\right)\right)\right)\middle| X_j\right] < 1.$$

This implies that

$$\mathrm{E}[N^{-1}\ln(L_N(\alpha,\beta,H)/L_N(\alpha_0,\beta_0,H_0))] = 0 \tag{49}$$

if and only if

$$H\left(\exp\left(-\exp(\beta' X)\left(\sum_{m=1}^{i}\alpha_m\right)\right)\right)$$
$$= H_0\left(\exp\left(-\exp(\beta_0' X)\left(\sum_{m=1}^{i}\alpha_{0,m}\right)\right)\right) \quad \text{a.s.} \quad \text{for } i = 1,\ldots,M, \tag{50}$$

where $X = X_j$.

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

Now the following question arises. Under what conditions does (50) imply that $\alpha = \alpha_0$, $\beta = \beta_0$, and $H = H_0$? Obviously, the ICMPH model involved is not identified if $\beta_0 = 0$ or if one of the components of X is a constant. Thus we make the following assumption.

Assumption 1(a). None of the covariates is constant, and at least one covariate has a nonzero coefficient.

For $i = 1$, (50) reads

$$H(\exp(-\exp(\beta' X)\alpha_1)) = H_0(\exp(-\exp(\beta_0' X)\alpha_{0,1})) \quad \text{a.s.} \tag{51}$$

To derive further conditions such that (51) implies $\beta = \beta_0$ and $\alpha_1 = \alpha_{0,1}$ we need to distinguish two cases.

The first case is that there is only one covariate: $X \in \mathbb{R}$. Without loss of generality we may assume that $\beta = c \cdot \beta_0$, so that (51) now reads

$$H(\exp(-\exp(c \cdot \beta_0 X)\alpha_1)) = H_0(\exp(-\exp(\beta_0 X)\alpha_{0,1})) \quad \text{a.s.}$$

or equivalently,

$$H(\exp(-\alpha_1 \alpha_{0,1}^{-c}(\ln(1/u))^c)) = H_0(u) \tag{52}$$

for all u in the support S_1 of $U = \exp(-\exp(\beta_0 X)\alpha_{0,1})$. Note that, because of the monotonicity of H and H_0, (52) implies that $c > 0$.

Next, consider the case that $X \in \mathbb{R}^k$ with $k \geq 2$ and make the following assumption.

Assumption 1(b). The $k \times k$ matrix $\Sigma_0 = \mathrm{E}[(X - \mathrm{E}[X|\beta_0' X])(X - \mathrm{E}[X|\beta_0' X])']$ has rank $k - 1$.

Note that $\beta_0' X - \mathrm{E}[\beta_0' X|\beta_0' X] = 0$, so that $\Sigma_0 \beta_0 = 0$. Thus β_0 is the eigenvector corresponding to the zero eigenvalue of Σ_0.

Referring to Assumptions 1(a) and (b) together as Assumption 1, the following result holds.

LEMMA 2. *Suppose that*

$$\begin{aligned} P[T > b_1|X] &= H_0(\exp(-\exp(\beta_0' X)\alpha_{0,1})) \\ &= H(\exp(-\exp(\beta' X)\alpha_1)) \quad \textit{a.s.} \end{aligned} \tag{53}$$

Under Assumption 1, (53) implies that there exists a constant $c > 0$ such that $\beta = c \cdot \beta_0$, so that for all u in the support S_1 of $U = \exp(-\exp(\beta_0' X)\alpha_{0,1})$,

$$H(\exp(-\alpha_1 \alpha_{0,1}^{-c}(\ln(1/u))^c)) = H_0(u). \tag{54}$$

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

Proof. See the Appendix.

Given β_0, for every $\alpha_{0,1} > 0$ there exists a distribution function H_0 such that the first equality in (53) is true. Moreover, given β_0, $\alpha_{0,1}$, and H_0, for every $c > 0$ there exist an $\alpha_1 > 0$ and a distribution function H such that (54) holds. Therefore, to establish first that $c = 1$ (so that $\beta = \beta_0$), and then that $H(u^{\alpha_1/\alpha_{0,1}}) = H_0(u)$ on S_1 implies $\alpha_1 = \alpha_{0,1}$, we need to normalize H_0 and H in some way. How to do that depends on the support of $\beta_0' X$. If the following condition holds, then there are various options for normalizing H_0 and H, as follows.

Assumption 2. The support of $\beta_0' X$ is the whole real line $\mathbb{R}$.[10]

5.1. Nonparametric Identification via Extreme Values

Taking the derivative of (54) to u yields

$$h_0(u) = h(\exp(-\alpha_1 \alpha_{0,1}^{-c}(\ln(1/u))^c))\exp(-\alpha_1 \alpha_{0,1}^{-c}(\ln(1/u))^c)$$

$$\times\, \alpha_1 \alpha_{0,1}^{-c} c(\ln(1/u))^{c-1} \frac{1}{u}, \quad \forall u \in S_1. \tag{55}$$

Now suppose that

$$\forall x \in \mathbb{R}, \qquad P[\beta_0' X \le x] > 0, \tag{56}$$

which is a weaker condition than Assumption 2, and

$$h_0(1) = h(1) = 1. \tag{57}$$

Recall that (57) corresponds to the condition that $\mathrm{E}[V] = 1$. Moreover, using the easy equality $\rho_k(1) = \sqrt{2k+1}$, it follows straightforwardly from (20) that condition (57) can be implemented by restricting h_0 and h to density functions of the type (20), with

$$\delta_1 = \frac{1}{2}\sqrt{2\left(1 + \sum_{k=2}^{\infty} \delta_k^2\right) + \left(1 + \sum_{k=2}^{\infty} \delta_k \sqrt{2k+1}\right)^2}$$

$$- \frac{\sqrt{3}}{2}\left(1 + \sum_{k=2}^{\infty} \delta_k \sqrt{2k+1}\right).$$

Condition (56) implies that S_1 contains a sequence u_n that converges to 1. Then

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

$$1 = \lim_{n\to\infty} h_0(u_n) = \lim_{n\to\infty} h(\exp(-\alpha_1\alpha_{0,1}^{-c}(\ln(1/u_n))^c))$$

$$\times \lim_{n\to\infty} \exp(-\alpha_1\alpha_{0,1}^{-c}(\ln(1/u_n))^c)$$

$$\times \alpha_1\alpha_{0,1}^{-c}c \lim_{n\to\infty} (\ln(1/u_n))^{c-1}$$

$$= \alpha_1\alpha_{0,1}^{-c}c \lim_{n\to\infty} (\ln(1/u_n))^{c-1} = \begin{cases} 0 & \text{if } c > 1, \\ \alpha_1/\alpha_{0,1} & \text{if } c = 1, \\ \infty & \text{if } c < 1. \end{cases} \tag{58}$$

Thus $c = 1$; hence $\beta = \beta_0$ and $\alpha_1 = \alpha_{0,1}$.

Because now β_0 and $\alpha_{0,1}$ are identified, the next question is as follows. Does

$$H(\exp(-\exp(\beta_0'X)(\alpha_{0,1} + \alpha_2)))$$

$$= H_0(\exp(-\exp(\beta_0'X)(\alpha_{0,1} + \alpha_{0,2}))) \quad \text{a.s.} \tag{59}$$

imply $\alpha_2 = \alpha_{0,2}$? Let S_2 be the support of $U = \exp(-\exp(\beta_0'X)(\alpha_{0,1} + \alpha_{0,2}))$ and let $\eta = (\alpha_2 - \alpha_{0,2})/(\alpha_{0,1} + \alpha_{0,2})$. Then (59) implies that $H(u^{1+\eta}) = H_0(u)$ for all $u \in S_2$. Under condition (56) there exists a sequence u_n in S_2 that converges to 1. Therefore, similarly to (58) we have

$$1 = h_0(1) = \lim_{n\to\infty} \frac{H_0(1) - H_0(u_n)}{1 - u_n}$$

$$= \lim_{n\to\infty} \frac{H(1) - H(u_n^{1+\eta})}{1 - u_n} = h(1) \lim_{n\to\infty} \frac{1 - u_n^{1+\eta}}{1 - u_n} = 1 + \eta.$$

Hence, under the conditions (56) and (57), $\eta = 0$, and thus $\alpha_2 = \alpha_{0,2}$. Repeating this argument for $i = 3, \ldots, M$, it follows that $\alpha = \alpha_0$ and

$$H_0(\exp(-\exp(\beta_0'X)\alpha_{0,i})) = H(\exp(-\exp(\beta_0'X)\alpha_{0,i})) \quad \text{a.s.}$$

for $i = 1, \ldots, M$.

Because (43) and (44) are equivalent for $t = b_i$, the following result holds.

THEOREM 5. *Let the integrated baseline hazard be specified by (43). Under Assumption 1 and the conditions (56) and (57) the equality*

$$\mathrm{E}[\ln(L_N(\alpha, \beta, H)/L_N(\alpha_0, \beta_0, H_0))] = 0 \tag{60}$$

implies that $\alpha = \alpha_0$, $\beta = \beta_0$, *and*

$$H(\exp(-\exp(\beta_0'X_j)\Lambda(b_i|\alpha_0))) = H_0(\exp(-\exp(\beta_0'X_j)\Lambda(b_i|\alpha_0))) \tag{61}$$

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

a.s. for $i = 1, \ldots, M$. If in addition Assumption 2 holds, with $X = X_j$, then (61) implies that $H = H_0$.

Note that this result and the corresponding results in Theorems 6 and 7, which follows, are conditional on a *common* vector X_j of covariates, because different sets of covariates may yield observationally equivalent ICMPH models. For example, let $X_j^* \neq X_j$. Then there may exist a pair α, β and a c.d.f. H such that $H(\exp(-\exp(\beta' X_j^*)\Lambda(b_i|\alpha))) = H_0(\exp(-\exp(\beta_0' X_j)\Lambda(b_i|\alpha_0)))$ a.s.[11]

Admittedly, Assumption 2 and even the weaker condition (56) are often not satisfied in practice. In most applications the covariates are bounded and discrete, and often quite a few of them are dummy variables. In unemployment duration studies one of the key covariates is the age of the respondent, which is expected to have a negative coefficient. But even age is a bounded variable and is usually measured in discrete units (e.g., years). In that case the ICMPH model may not be identified. Nevertheless, I will maintain Assumption 2 for the time being. In Section 9 I will consider the more realistic case that the covariates have finite support.

The condition (57) is only effective in pinning down α_0 and β_0 if $U_{i,j} = \exp(-\exp(\beta_0' X_j)\Lambda(b_i|\alpha_0))$ can get close enough to one. Thus, the identification hinges on the *extreme negative values* of $\beta_0' X_j$. Under Assumption 2 it follows that $p\lim_{N\to\infty} \min_{j=1,\ldots,N} \beta_0' X_j = -\infty$, hence $p\lim_{N\to\infty} \max_{j=1,\ldots,N} U_{i,j} = 1$, but in finite samples $\max_{j=1,\ldots,N} U_{i,j}$ may not get close enough to one for condition (57) to be effective. Therefore, I will now derive alternative identification conditions.

5.2. Nonparametric Identification via Quantile Restrictions

Suppose that the distribution functions H and H_0 are confined to distributions with two common quantiles. For a given pair u_1, u_2 of distinct points in (0,1), let

$$H(u_i) = H_0(u_i) = u_i, \qquad i = 1, 2. \tag{62}$$

Again, the latter equality facilitates the benchmark case of absence of unobserved heterogeneity: $V = 1$ a.s., which is equivalent to $H(u) = H_0(u) = u$ a.e. on $[0, 1]$.

Under Assumption 2, $u_1, u_2 \in S_1 = (0, 1)$. Therefore, it follows straightforwardly from the quantile restrictions (62) and Lemma 2 that under Assumptions 1 and 2, $(\ln(1/u_1))^{c-1} = (\ln(1/u_2))^{c-1} = \alpha_{0,1}^c/\alpha_1$. Because $u_1 \neq u_2$, the first equality is only possible if $c = 1$, hence $\beta = \beta_0$, and the second equality then implies that $\alpha_1 = \alpha_{0,1}$. Similar to Theorem 5 the next result now follows.

THEOREM 6. *Under Assumptions 1 and 2 and the quantile conditions (62) the ICMPH model is nonparametrically identified.*

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

In other words, under Assumptions 1 and 2 the equality (60) implies that $\alpha = \alpha_0$, $\beta = \beta_0$, and $H = H_0$.

5.3. Nonparametric Identification via Moment Conditions

Consider the ordered probability model form (46) of the ICMPH model. Let $F(x)$ be a distribution function of the type (47),

$$F(x) = 1 - H(\exp(-\exp(x))) \tag{63}$$

with density

$$f(x) = h(\exp(-\exp(x)))\exp(-\exp(x))\exp(x), \tag{64}$$

where $H(u)$ is a distribution function on $[0,1]$ with density $h(u)$, and suppose that for some constants $\sigma > 0$ and $\mu \in \mathbb{R}$,

$$F(\sigma x + \mu) \equiv F_0(x). \tag{65}$$

I will set forth moment conditions such that (65) implies $\mu = 0$ and $\sigma = 1$, so that then model (46) is nonparametrically identified.

Taking the derivative of (65) it follows that (65) implies

$$\begin{aligned} f_0(x) &= h_0(\exp(-\exp(x)))\exp(-\exp(x))\exp(x) \\ &= \sigma h(\exp(-\exp(\sigma x + \mu)))\exp(-\exp(\sigma x + \mu))\exp(\sigma x + \mu) \\ &= \sigma f(\sigma x + \mu) \quad \text{a.e.;} \end{aligned}$$

hence it follows from (48) and (64) that for any function φ on $\mathbb{R}$ for which $\int_{-\infty}^{\infty} \varphi(x) f_0(x)\, dx$ is well defined,

$$\begin{aligned} &\int_0^1 \varphi(\ln(\ln(1/u)))h_0(u)\, du \\ &\quad = \int_{-\infty}^{\infty} \varphi(x) f_0(x)\, dx \\ &\quad = \sigma \int_{-\infty}^{\infty} \varphi(x) f(\sigma x + \mu)\, dx = \int_{-\infty}^{\infty} \varphi\left(\frac{x-\mu}{\sigma}\right) f(x)\, dx \\ &\quad = \int_0^1 \varphi\left(\frac{\ln(\ln(1/u)) - \mu}{\sigma}\right) h(u)\, du. \end{aligned} \tag{66}$$

If we choose $\varphi(x) = x$ then (66) implies

$$\int_0^1 \ln(\ln(1/u))h(u)\, du = \sigma \int_0^1 \ln(\ln(1/u))h_0(u)\, du + \mu, \tag{67}$$

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

and if we choose $\varphi(x) = x^2$ then (66) implies

$$\sigma^2 \int_0^1 (\ln(\ln(1/u)))^2 h_0(u)\,du = \int_0^1 (\ln(\ln(1/u)))^2 h(u)\,du - 2\mu \int_0^1 \ln(\ln(1/u))h(u)\,du + \mu^2. \tag{68}$$

Now assume that

$$\int_0^1 \ln(\ln(1/u))h(u)\,du = \int_0^1 \ln(\ln(1/u))h_0(u)\,du, \tag{69}$$

$$\int_0^1 (\ln(\ln(1/u)))^2 h(u)\,du = \int_0^1 (\ln(\ln(1/u)))^2 h_0(u)\,du. \tag{70}$$

Then it follows from (67) and (69) that

$$\mu = (1-\sigma)\int_0^1 \ln(\ln(1/u))h_0(u)\,du \tag{71}$$

and from (68) through (71) that

$$(\sigma^2 - 1)\int_0^1 (\ln(\ln(1/u)))^2 h_0(u)\,du = (\sigma^2 - 1)\left(\int_0^1 \ln(\ln(1/u))h_0(u)\,du\right)^2. \tag{72}$$

The equality (72) implies $\sigma = 1$ because

$$\int_0^1 (\ln(\ln(1/u)))^2 h_0(u)\,du > \left(\int_0^1 \ln(\ln(1/u))h_0(u)\,du\right)^2,$$

so that by (71), $\mu = 0$.

Note that the values of the integrals in (69) and (70) do not matter for this result, provided that the integrals involved are finite. However, to accommodate the benchmark case $h_0(u) = h(u) \equiv 1$, which corresponds to absence of unobserved heterogeneity, I will assume that the c.d.f. H in the log-likelihood function (42) is confined to a space of distribution functions satisfying the moment conditions

$$\int_0^1 \ln(\ln(1/u))\,dH(u) = \int_0^1 \ln(\ln(1/u))\,du, \tag{73}$$

$$\int_0^1 (\ln(\ln(1/u)))^2\,dH(u) = \int_0^1 (\ln(\ln(1/u)))^2\,du. \tag{74}$$

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

It is obvious from the easy equalities

$$\int_0^1 (\ln(\ln(1/u)))^p\, du = \int_{-\infty}^{\infty} x^p \cdot \exp(x)\exp(-\exp(x))\, dx$$

for $p = 1, 2$ that the right-hand-side integrals in (73) and (74) are finite. Their values are

$$\int_0^1 \ln(\ln(1/u))\, du = -0.577189511,$$

$$\int_0^1 (\ln(\ln(1/u)))^2\, du = 1.981063818,$$

which have been computed by Monte Carlo integration.[12]

THEOREM 7. *Let H and H_0 be confined to distribution functions satisfying the moment conditions (73) and (74). Then under Assumptions 1 and 2 the ICMPH model is nonparametrically identified.*

5.4. Implementation of Moment and Quantile Conditions

The moment conditions (73) and (74) can be implemented by penalizing the log-likelihood function (42) for deviations from the moment conditions involved by augmenting the log-likelihood function $\ln(L_N(\alpha, \beta, H))$ with two penalty terms:

$$\begin{aligned}\ln(L_N^*(\alpha,\beta,H)) = \ln(L_N(\alpha,\beta,H)) &- N\left(\int_0^1 \ln(\ln(1/u))\, dH(u) - \int_0^1 \ln(\ln(1/u))\, du\right)^{2\ell} \\ &- N\left(\int_0^1 (\ln(\ln(1/u)))^2\, dH(u) - \int_0^1 (\ln(\ln(1/u)))^2\, du\right)^{2\ell}\end{aligned} \tag{75}$$

for some integer $\ell \geq 1$.

Similarly, also the quantile restrictions (62) can be implemented by penalizing the log-likelihood function:

$$\begin{aligned}\ln(L_N^*(\alpha,\beta,H)) = \ln(L_N(\alpha,\beta,H)) &- N(H(u_1) - u_1)^{2\ell} - N(H(u_2) - u_2)^{2\ell}\end{aligned} \tag{76}$$

for some integer $\ell \geq 1$.

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

For SNP density functions (21) the moment conditions (73) and (74) read

$$\left(1+\sum_{k=1}^{n}\delta_k^2\right)\int_0^1(\ln(\ln(1/u)))^p h_n(u)\,du$$
$$=\int_0^1(\ln(\ln(1/u)))^p\,du+2\sum_{k=1}^{n}\delta_k\int_0^1(\ln(\ln(1/u)))^p\rho_k(u)\,du$$
$$+\sum_{k=1}^{n}\sum_{m=1}^{n}\delta_k\left(\int_0^1(\ln(\ln(1/u)))^p\rho_k(u)\rho_m(u)\,du\right)\delta_m$$
$$=\left(1+\sum_{k=1}^{n}\delta_k^2\right)\int_0^1(\ln(\ln(1/u)))^p\,du$$

for $p=1$ and $p=2$, respectively. Hence, denoting

$$a'_{n,p}=\left(\int_0^1(\ln(\ln(1/u)))^p\rho_1(u)\,du,\ldots,\int_0^1(\ln(\ln(1/u)))^p\rho_n(u)\,du\right)$$

and

$$B_{n,p}=\left(\int_0^1(\ln(\ln(1/u)))^p\rho_{i_1}(u)\rho_{i_2}(u)\,du;\,i_1,i_2=1,2,\ldots,n\right)$$
$$-\int_0^1(\ln(\ln(1/u)))^p\,du\cdot I_n,$$

the conditions (73) and (74) with h replaced by (21) are equivalent to $2\delta' a_{n,1}+\delta' B_{n,1}\delta=0$, $2\delta' a_{n,2}+\delta' B_{n,2}\delta=0$, respectively, where $\delta=(\delta_1,\ldots,\delta_n)'$. Therefore, if we replace H in (75) by $H_n(\cdot|\delta)$, the penalized log-likelihood can be written as

$$\ln(L_N^*(\alpha,\beta,H_n(\cdot|\delta)))$$
$$=\ln(L_N(\alpha,\beta,H_n(\cdot|\delta)))$$
$$-N\left(\frac{2\delta' a_{n,1}+\delta' B_{n,1}\delta}{1+\delta'\delta}\right)^{2\ell}-N\left(\frac{2\delta' a_{n,2}+\delta' B_{n,2}\delta}{1+\delta'\delta}\right)^{2\ell}\qquad(77)$$

for some integer $\ell\geq 1$.

Note that for $p=1,2$ the vectors $a_{n,p}$ and matrices $B_{n,p}$ can easily be computed in advance by Monte Carlo integration.

If we would assume that for some fixed n, $H_0(u)=H_n(u|\delta_0)$, so that $H_n(u|\delta_0)$ is treated as a parametric specification of the c.d.f. $H_0(u)$, and if we choose $\ell\geq 2$, then

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

$$\lim_{N\to\infty} \mathrm{Var}\left(\frac{1}{\sqrt{N}}\,\frac{\partial \ln(L_N^*(\alpha_0,\beta_0,H_n(\cdot|\delta_0)))}{\partial(\alpha_0',\beta_0',\delta_0')}\right)$$

$$= \lim_{N\to\infty} \mathrm{Var}\left(\frac{1}{\sqrt{N}}\,\frac{\partial \ln(L_N(\alpha_0,\beta_0,H_n(\cdot|\delta_0)))}{\partial(\alpha_0',\beta_0',\delta_0')}\right)$$

and

$$\lim_{N\to\infty} \mathrm{E}\left(\frac{-1}{N}\,\frac{\partial^2 \ln(L_N^*(\alpha_0,\beta_0,H_n(\cdot|\delta_0)))}{\partial(\alpha_0',\beta_0',\delta_0')'\partial(\alpha_0',\beta_0',\delta_0')}\right)$$

$$= \lim_{N\to\infty} \mathrm{E}\left(\frac{-1}{N}\,\frac{\partial^2 \ln(L_N(\alpha_0,\beta_0,H_n(\cdot|\delta_0)))}{\partial(\alpha_0',\beta_0',\delta_0')'\partial(\alpha_0',\beta_0',\delta_0')}\right).$$

Consequently, the penalized ML estimators of α_0, β_0, and δ_0 are then asymptotically efficient. Therefore, I advocate choosing $\ell = 2$.

Similarly, if we replace H in (76) by $H_n(\cdot|\delta)$, the penalized log-likelihood becomes

$$\ln(L_N^*(\alpha,\beta,H_n(\cdot|\delta))) = \ln(L_N(\alpha,\beta,H_n(\cdot|\delta))) - N(H_n(u_1|\delta)-u_1)^{2\ell} - N(H_n(u_2|\delta)-u_2)^{2\ell},$$

with $H_n(u|\delta)$ defined by (25). For the same reason as before I recommend choosing $\ell = 2$.

6. REQUIREMENTS FOR CONSISTENCY OF SNP MAXIMUM LIKELIHOOD ESTIMATORS

We can write the (penalized) log-likelihood as

$$\ln(L_N^*(\alpha,\beta,H)) = \sum_{j=1}^{N} \Psi(Y_j,\alpha,\beta,H),$$

where $Y_j = (D_{1,j},\ldots,D_{M,j},X_j')'$. In the cases (75) and (76),

$$\Psi(Y_j,\alpha,\beta,H) = \sum_{i=1}^{M} D_{i,j}\ln(H(\mu_{i-1}(\alpha,\beta'X_j)) - H(\mu_i(\alpha,\beta'X_j))) + \left(1-\sum_{i=1}^{M} D_{i,j}\right)\ln(H(\mu_M(\alpha,\beta'X_j))) - \Pi(H), \quad \textbf{(78)}$$

where $\Pi(H)$ represents the two penalty terms.

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

The ML estimators of α_0, β_0, and H_0 are

$$(\hat{\alpha}, \hat{\beta}, \hat{H}) = \arg \max_{\alpha \in A, \beta \in B, H \in \mathcal{H}(0,1)} N^{-1} \ln(L_N^*(\alpha, \beta, H)), \tag{79}$$

where we make the following assumption.

Assumption 3. A and B are given compact parameter spaces for α and β, respectively, containing the true parameters $\alpha_0 \in A$, $\beta_0 \in B$, and the space $\mathcal{H}(0,1)$ is a compact metric space of distribution functions on $[0,1]$ endowed with the sup metric

$$\|H_1 - H_2\| = \sup_{0 \le u \le 1} |H_1(u) - H_2(u)|. \tag{80}$$

The space $\mathcal{H}(0,1)$ corresponds to a metric space $\mathcal{D}(0,1)$ of densities on $(0,1)$ endowed with the L_1 metric

$$\|h_1 - h_2\|_1 = \int_0^1 |h_1(u) - h_2(u)| \, du. \tag{81}$$

Let

$$\bar{\Psi}(\alpha, \beta, H) = \mathrm{E}[\Psi(Y_j, \alpha, \beta, H)]. \tag{82}$$

To prove the consistency of the ML estimators, we need to show first that

$$p \lim_{N \to \infty} \bar{\Psi}(\hat{\alpha}, \hat{\beta}, \hat{H}) = \bar{\Psi}(\alpha_0, \beta_0, H_0). \tag{83}$$

Similar to the standard consistency proof for M-estimators it can be shown that if $\bar{\Psi}$ is continuous and (α_0, β_0, H_0) is unique then (83) implies that $p \lim_{N \to \infty} \hat{\alpha} = \alpha_0$, $p \lim_{N \to \infty} \hat{\beta} = \beta_0$, and $p \lim_{N \to \infty} \|\hat{H} - H_0\| = 0$.

In general it will be impossible to compute (79) because it requires maximizing the log-likelihood function over a space of distribution functions. However, it follows from Theorem 3 that for each subsequence $n = n_N$ there exists a sequence of compact subspaces $\mathcal{H}_{n_N}(0,1) \subset \mathcal{H}(0,1)$ of distribution functions of the type (5) or (25), containing a c.d.f. H_{n_N} such that $\lim_{N \to \infty} \|H_{n_N} - H_0\| = 0$, so that

$$(\tilde{\alpha}, \tilde{\beta}, \tilde{H}) = \arg \max_{\alpha \in A, \beta \in B, H \in \mathcal{H}_{n_N}(0,1)} N^{-1} \ln(L_N^*(\alpha, \beta, H)) \tag{84}$$

is feasible. This is known as sieve estimation. See Chen (2007) for a review of sieve estimation. It will be shown that under suitable conditions, $p \lim_{N \to \infty} \tilde{\alpha} = \alpha_0$, $p \lim_{N \to \infty} \tilde{\beta} = \beta_0$, and $p \lim_{N \to \infty} \|\tilde{H} - H_0\| = 0$.

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

The crux of the consistency problem is twofold, namely, (a) how to make the metric space $\mathcal{H}(0,1)$ compact and (b) how to prove (83). These problems will be addressed in the next sections.

7. COMPACTNESS OF THE DENSITY AND DISTRIBUTION FUNCTION SPACES

Consider the space of density functions of the type (20), subject to the condition $\sum_{k=1}^{\infty} \delta_k^2 < \infty$. This condition can easily be imposed, e.g., by restricting the δ_k's such that for some constant $c > 0$,

$$|\delta_k| \leq \frac{c}{1 + \sqrt{k}\ln(k)}, \tag{85}$$

because then $\sum_{k=1}^{\infty} \delta_k^2 < c^2 + c^2 \sum_{k=2}^{\infty} k^{-1}(\ln(k))^{-2} < c^2 + c^2/\ln(2) < \infty$.

The conditions (85) also play a key role in proving compactness, as follows.

THEOREM 8. *Let $\mathcal{D}(0,1)$ be the space of densities of the type (20) subject to the restrictions (85) for some constant $c > 0$, endowed with the metric $\|h_1 - h_2\|_1 = \int_0^1 |h_1(u) - h_2(u)|\, du$. Then $\mathcal{D}(0,1)$ is compact. Consequently, the space*

$$\mathcal{H}(0,1) = \left\{ H(u) = \int_0^u h(v)\, dv,\ h \in \mathcal{D}(0,1) \right\},$$

endowed with the "sup" metric (80), is compact also.

Proof. See the Appendix.

Of course, the result of Theorem 8 is only useful for our purpose if the constant c in (85) is chosen so large that we can make the following assumption.

Assumption 4. The true c.d.f. H_0 is contained in $\mathcal{H}(0,1)$.

Moreover, it follows now straightforwardly from Theorem 3 that the following results hold.

THEOREM 9. *Let $\mathcal{D}_n(0,1)$ be the space of densities of the type (21) subject to the restrictions (85), with c the same as for $\mathcal{D}(0,1)$, with corresponding space of distribution functions*

$$\mathcal{H}_n(0,1) = \left\{ H(u) = \int_0^u h(v)\, dv,\ h \in \mathcal{D}_n(0,1) \right\}.$$

For each n, $\mathcal{H}_n(0,1)$ is a compact subset of $\mathcal{H}(0,1)$, and for each $H \in \mathcal{H}(0,1)$ there exists a c.d.f. $H_n \in \mathcal{H}_n(0,1)$ such that $\lim_{n\to\infty} \sup_{0\leq u\leq 1} |H(u) - H_n(u)| = 0$.

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

Recall that the SNP distribution functions $H_n(u)$ take two equivalent forms, namely, the form $H_n(u|\delta)$ defined by (25) and the reparametrized form $H_n(u|\eta)$ defined by (30), where the parameters η in the latter are linked to δ via (28). Hence, the restrictions (85) imply the following restriction on the parameter vector η:

$$\left| \sum_{m=0}^{n-k} \eta_{k+m}\, \mu_k(m) \right| \leq \frac{c}{1+\sqrt{k}\ln(k)}, \qquad k = 1, \ldots, n, \tag{86}$$

where $\mu_k(m)$ is defined by (29). Thus, $\mathcal{H}_n(0,1)$ is also the space of distribution functions of the type (30) subject to the restrictions (86).

It is not too hard to verify from (10) and (11) that the integrals $\mu_n(m) = \int_0^1 u^{n+m}\rho_n(u)\,du$ for $m = 0,1,2,\ldots$ can be computed recursively, namely,

$$\mu_0(m) = 1/(m+1),$$

$$\mu_n(m) = \kappa_n \cdot \mu_{n-1}(m) + 0.5I(m \geq 1)\cdot \mu_n(m-1) - \omega_n I(m \geq 2)\cdot \mu_{n+1}(m-2) \quad \text{for } n \geq 1,$$

where

$$\kappa_n = \frac{n\sqrt{2n+3}}{2\sqrt{4(n+1)^2-1}\sqrt{2n-1}}, \qquad \omega_n = \frac{n+1}{2\sqrt{4(n+1)^2-1}}.$$

The restrictions (85) or (86) can be implemented in practice by penalizing the log-likelihood, in particular if the simplex iteration of Nelder and Mead (1965) is used for maximizing the log-likelihood. If the bounds (86) or (85) are exceeded, set the log-likelihood to a large negative value, e.g., -10^{30}, so that the next simplex iteration will bounce away from the bounds.

8. CONSISTENCY OF *M*-ESTIMATORS IN THE PRESENCE OF NON-EUCLIDEAN PARAMETERS

I will now address the problem of how to prove (83). To relate (83) to Theorem 10, which follows, denote $\Theta = \{(\alpha,\beta,h) : \alpha \in A, \beta \in B, H \in \mathcal{H}(0,1)\}$ and define a metric $d(.,.)$ on Θ by combining the metrics on A, B, and $\mathcal{H}(0,1)$. For example, for $\theta_1 = (\alpha_1,\beta_1,H_1) \in \Theta$, $\theta_2 = (\alpha_2,\beta_2,H_2) \in \Theta$, let

$$d(\theta_1,\theta_2) = \max\Big[\sqrt{(\alpha_1-\alpha_2)'(\alpha_1-\alpha_2)}, \sqrt{(\beta_1-\beta_2)'(\beta_1-\beta_2)}, \times \sup_{0\leq u\leq 1} |H_1(u) - H_2(u)|\Big]. \tag{87}$$

THEOREM 10. *Let Y_j, $j = 1,\ldots,N$, be a sequence of independent and identically distributed (i.i.d.) random vectors in a euclidean space, defined on a*

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

common probability space $\{\Omega, \mathcal{F}, P\}$*, with support contained in an open set* $\mathcal{Y}$*. Let* Θ *be a compact metric space with metric* $d(\theta_1, \theta_2)$*. Let* $g(y, \theta)$ *be a continuous real function on* $\mathcal{Y} \times \Theta$ *such that for each* $\theta \in \Theta$*,*

$$\mathrm{E}[|g(Y_1, \theta)|] < \infty, \tag{88}$$

so that $\bar{g}(\theta) = \mathrm{E}[g(Y_1, \theta)]$ *is defined and finite, and let for some constant* $K_0 > 0$

$$\mathrm{E}\left[\max\left(\sup_{\theta \in \Theta} g(Y_1, \theta), -K_0\right)\right] < \infty. \tag{89}$$

Denote $\hat{\theta} = \arg\max_{\theta \in \Theta} N^{-1} \sum_{j=1}^{N} g(Y_j, \theta)$ *and* $\theta_0 = \arg\max_{\theta \in \Theta} \bar{g}(\theta)$*. Then* $p\lim_{N \to \infty} \bar{g}(\hat{\theta}) = \bar{g}(\theta_0)$*. If* θ_0 *is unique then the latter implies* $p\lim_{N \to \infty} d(\hat{\theta}, \theta_0) = 0$*.*

Proof. See the Appendix.

In the penalized log-likelihood case, let $g(Y_j, \theta) = \Psi(Y_j, \alpha, \beta, H)$, where the latter is defined by (78). Clearly, $H(\mu_i(\alpha, \beta' x))$ is continuous in $\alpha \in A$, $\beta \in B$, and all x, and H itself is uniformly continuous with respect to the "sup" metric. Moreover, the penalty term $-\Pi(H)$ in (78) is continuous in H. Therefore, $\Psi(y, \alpha, \beta, H)$ is continuous on $\mathcal{Y} \times A \times B \times \mathcal{H}(0,1)$, where $\mathcal{Y}$ is the euclidean space with dimension the dimension of $Y_j = (D_{1,j}, \ldots, D_{M,j}, X_j')$. Then $\bar{\Psi}(\alpha, \beta, H)$ is also continuous on $A \times B \times \mathcal{H}(0,1)$.

It is easy to verify from (78) that $\Psi(Y_j, \alpha, \beta, H) \leq 0$; hence condition (89) holds, and condition (88) holds if we make the following assumption.

Assumption 5. For all $(\alpha, \beta, H) \in A \times B \times \mathcal{H}(0,1)$, $\mathrm{E}[\ln(L_N^*(\alpha, \beta, H))] > -\infty$.

Thus, under Assumptions 1–5, the conditions of Theorem 10 hold, and hence (83) is true.

As said before, maximizing a function over a non-euclidean metric space Θ is usually not feasible, but it may be feasible to maximize such a function over a subset $\Theta_N \subset \Theta$ such that under some further conditions the resulting feasible M-estimator is consistent.

THEOREM 11. *Let the conditions of Theorem 10 hold and let* $\Theta_N \subset \Theta$ *be such that the computation of* $\tilde{\theta} = \arg\max_{\theta \in \Theta_N} N^{-1} \sum_{j=1}^{N} g(Y_j, \theta)$ *is feasible. If each* Θ_N *contains an element* θ_N *such that* $\lim_{N \to \infty} d(\theta_N, \theta_0) = 0$*, then* $p\lim_{N \to \infty} \bar{g}(\tilde{\theta}) = \bar{g}(\theta_0)$*. Consequently, Theorem 10 carries over to* $\tilde{\theta}$*.*

Proof. See the Appendix.

We can now formulate the consistency results for the sieve ML estimators of the parameters of the SNP ICMPH model.

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

THEOREM 12. *Let α_0, β_0, H_0 be the true parameters of the ICMPH model. Let $L_N^*(\alpha,\beta,H)$ be the penalized likelihood function and let*

$$(\tilde{\alpha},\tilde{\beta},\tilde{H}) = \arg \max_{\alpha\in A,\,\beta\in B,\,H\in\mathcal{H}_{n_N}(0,1)} \ln(L_N^*(\alpha,\beta,H)),$$

where $\mathcal{H}_{n_N}(0,1)$ is the space of distribution functions defined in Theorem 9. Then under Assumptions 1–5, $p\lim_{N\to\infty}\tilde{\alpha}=\alpha_0$, $p\lim_{N\to\infty}\tilde{\beta}=\beta_0$, and

$$p\lim_{N\to\infty}\sup_{0\le u\le 1}|\tilde{H}(u)-H_0(u)|=0.$$

Note that the speed of convergence n_N of $H_{N_N}\in\mathcal{H}_{n_N}(0,1)$ to $H_0\in\mathcal{H}(0,1)$ does not matter for this result. Therefore, as far as consistency is concerned the space $\mathcal{H}_{n_N}(0,1)$ may be selected adaptively by using, e.g., the well-known Hannan and Quinn (1979) or Schwarz (1978) information criteria.

9. THE ICMPH MODEL WITH FINITE-VALUED COVARIATES

As said before, Assumption 2 is often not satisfied in practice. In this section I will therefore consider the more realistic case that the covariates are finite-valued.

Assumption 1*. None of the components of the vector $X_j\in\mathbb{R}^k$ of covariates is a constant. The support S of X_j is finite: $S=\{x_1,x_2,\ldots,x_K\}$, $P[X_j\in S]=1$.

9.1. Lack of Identification

In this case the ICMPH model is no longer identified. Instead of a single true parameter vector $\theta_0=(\alpha_0',\beta_0')'$ there now exists a set Θ_0 of "true" parameters (henceforth called *admissible* rather than true), in the sense that for each $(\alpha_0',\beta_0')'\in\Theta_0$ there exist uncountably many distribution functions H_0 on $[0,1]$ for which the model is correct. I will show this for the ICMPH model (46) in the form

$$\begin{aligned} p_{m,\ell}^0 &= P[\Sigma_{i=1}^m D_{i,j}=0\,|\,X_j=x_\ell] \\ &= H_0(\exp(-\exp(\beta_0' x_\ell+\alpha_0'\omega_m))), \\ m &= 1,\ldots,M, \qquad \ell=1,\ldots,K, \end{aligned} \tag{90}$$

where ω_m is column m of the $M\times M$ upper triangular matrix $\Omega=(\omega_{i_1,i_2})$ with typical element $\omega_{i_1,i_2}=I(i_1\le i_2)$ and $\alpha_0=(\alpha_{0,1},\alpha_{0,2},\ldots,\alpha_{0,M})'$ satisfying

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

$\alpha_{0,m} > 0$ for $m = 2,3,\ldots,M$. As part of the model specification I will choose closed hypercubes for the parameter spaces of α and β:

$$A = \times_{i=1}^{M}[\underline{\alpha}_i, \bar{\alpha}_i], \qquad \underline{\alpha}_i > 0 \quad \text{for } i = 2,\ldots,M, \qquad B = \times_{i=1}^{k}[\underline{\beta}_i, \bar{\beta}_i], \quad \textbf{(91)}$$

where the intervals involved are wide enough such that $A \times B$ contains at least one admissible parameter vector $(\alpha_0', \beta_0')'$. Moreover, note that the log-likelihood $\ln(L_N(\alpha,\beta,H))$ involved is the same as in (42), except that now $\mu_i(\alpha,\beta' X_j) = \exp(-\exp(\beta' X_j + \alpha'\omega_i))$ for $i = 1,\ldots,M$.

If K is small relative to the sample size N, then we can treat the probabilities $p_{m,\ell}^0$ as parameters, with corresponding log-likelihood

$$\begin{aligned}\ln(L_N(P)) &= \sum_{j=1}^{N} D_{1,j} \sum_{\ell=1}^{K} I(X_j = x_\ell)\ln(1 - p_{1,\ell}) \\ &= \sum_{j=1}^{N}\sum_{i=2}^{M} D_{i,j} \sum_{\ell=1}^{K} I(X_j = x_\ell)\ln(p_{i-1,\ell} - p_{i,\ell}) \\ &\quad + \sum_{j=1}^{N}\left(1 - \sum_{i=1}^{M} D_{i,j}\right)\sum_{\ell=1}^{K} I(X_j = x_\ell)\ln(p_{M,\ell}), \end{aligned} \quad \textbf{(92)}$$

where $P = (p_{i,\ell};\ i = 1,\ldots,M,\ \ell = 1,\ldots,K)$ is an $M \times K$ parameter matrix. It is easy to verify that the ML estimator of

$$P_0 = (p_{i,\ell}^0;\ i = 1,\ldots,M,\ \ell = 1,\ldots,K) \quad \textbf{(93)}$$

is an $M \times K$ matrix $\hat{P}$ with elements

$$\hat{p}_{m,\ell} = \frac{\sum_{j=1}^{N}(1 - \Sigma_{i=1}^{m} D_{i,j})I(X_j = x_\ell)}{\sum_{j=1}^{N} I(X_j = x_\ell)}. \quad \textbf{(94)}$$

Thus, in this case there is no need for a model. However, I will assume that K is too large for this approach.

The size of the set Θ_0 of admissible parameter vectors in $A \times B$ is maximal (in term of Lebesgue measure) if we make the following assumption.

Assumption 2*. The probabilities $p_{m,\ell}^0 = P[\Sigma_{i=1}^{m} D_{i,j} = 0 | X_j = x_\ell]$, $m = 1,\ldots,M$, $\ell = 1,\ldots,K$, are all different.[13]

Then the set Θ_0 is the simplex

$$\Theta_0 = \cap_{p_{m_2,\ell_2}^0 > p_{m_1,\ell_1}^0}\{\theta = (\alpha',\beta')' \in A \times B : \beta'(x_{\ell_1} - x_{\ell_2}) + \alpha'(\omega_{m_1} - \omega_{m_2}) > 0\}, \quad \textbf{(95)}$$

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

where $\ell_1, \ell_2 = 1, \ldots, K$ and $m_1, m_2 = 1, \ldots, M$. Any point $\theta = (\alpha', \beta')' \in \Theta_0$ is an admissible parameter vector because there exist uncountably many continuous distribution functions H on $[0,1]$ that fit through the $M \times K$ points $(\exp(-\exp(\beta' x_\ell + \alpha' \omega_m)), p_{m,\ell})$, so that then

$$p^0_{m,\ell} = P[\Sigma_{i=1}^m D_{i,j} = 0 | X_j = x_\ell]$$

$$= H(\exp(-\exp(\beta' x_\ell + \alpha' \omega_m))) \tag{96}$$

also.

On the other hand, the distribution functions H and H_0 can be confined to SNP distribution functions as indicated in the following lemma.

LEMMA 3. *Given K points $(u_i, v_i) \in (0,1) \times (0,1)$ satisfying $u_1 < u_2 < \cdots < u_K$, $v_1 < v_2 < \cdots < v_K$, there exists an SNP distribution function $H_n(u|\delta)$ defined by (25)[14] such that $H_n(u_i|\delta) = v_i$ for $i = 1, \ldots, K$. However, δ may not be unique, even for the minimal value of n.*

Proof. See the Appendix.

Therefore, the space $\mathcal{H}(0,1)$ of distribution functions H can be restricted to $\underline{\mathcal{H}}(0,1) = \cup_n^\infty \mathcal{H}_n(0,1)$, where $\mathcal{H}_n(0,1)$ is defined in Theorem 9. Thus, for each $(\alpha', \beta')' \in \Theta_0$ there exist a minimal polynomial order $\underline{n}(\alpha, \beta)$ and one or more distribution functions $H_{\underline{n}(\alpha,\beta)} \in \mathcal{H}_{\underline{n}(\alpha,\beta)}(0,1)$ such that (96) holds for $H(u) = H_{\underline{n}(\alpha,\beta)}(u)$.

Neither the quantile conditions (62) nor the moment conditions (73) and (74) are sufficient to solve this identification problem. The only way we can solve this problem is to design a procedure for selecting a unique point in the simplex Θ_0. I will propose to do that via quasi maximum likelihood (QML).

9.2. Quasi Maximum Likelihood

The idea is to approximate $p^0_{m,\ell}$ using for h a low-order SNP density $h_q(u|\delta)$. Suppose that the following assumption holds.

Assumption 3*. The polynomial order q is chosen such that

$$(\alpha_q, \beta_q, H_q) = \arg \max_{(\alpha,\beta,H) \in A \times B \times \mathcal{H}_q(0,1)} \mathrm{E}[N^{-1} \ln L_N(\alpha, \beta, H)] \tag{97}$$

is unique.

We also make the following assumption.

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

Assumption 4*. $\mathrm{E}[N^{-1}\ln L_N(\alpha,\beta,H)] > -\infty$ for all $(\alpha,\beta,H) \in A \times B \times (\cup_n^\infty \mathcal{H}_n(0,1))$.

If, in addition, Assumption 4* holds, then it follows from Theorem 11 that

$$(\tilde{\alpha}_q,\tilde{\beta}_q,\tilde{H}_q) = \arg \max_{(\alpha,\beta,H)\in A\times B\times \mathcal{H}_q(0,1)} \ln L_N(\alpha,\beta,H) \tag{98}$$

converges in probability to (α_q,β_q,H_q).

Note that the QML estimation problem (98) is fully parametric. Denoting $\Delta_q = \{\delta = (\delta_1,\ldots,\delta_q)' \text{ s.t. } (85)\}$, (98) is equivalent to

$$(\tilde{\alpha}_q,\tilde{\beta}_q,\tilde{\delta}_q) = \arg \max_{(\alpha,\beta,\delta)\in A\times B\times \Delta_q} \ln(L_N(\alpha,\beta,H_q(\cdot|\delta))),$$

and (97) is equivalent to

$$(\alpha_q,\beta_q,\delta_q) = \arg \max_{(\alpha,\beta,\delta)\in A\times B\times \Delta_q} \mathrm{E}[N^{-1}\ln(L_N(\alpha,\beta,H_q(\cdot|\delta)))].$$

Therefore, the standard QML asymptotics applies (cf. White, 1982).

Assumption 5*. $(\alpha_q',\beta_q',\delta_q')'$ is an interior point of $\Theta_0 \times \Delta_q$.

In particular, if Assumption 5* holds, then

$$\sqrt{N}\begin{pmatrix} \tilde{\alpha}_q - \alpha_q \\ \tilde{\beta}_q - \beta_q \\ \tilde{\delta}_q - \delta_q \end{pmatrix} \to N_{M+k+q}[0, \Omega_2^{-1}\Omega_1\Omega_2^{-1}]$$

in distribution, where

$$\Omega_1 = \mathrm{Var}\left(\frac{\partial \ln L_N(\alpha_q,\beta_q,H_q(\cdot|\delta_q))/\sqrt{N}}{\partial(\alpha_q',\beta_q',\delta_q')}\right),$$

$$\Omega_2 = -\mathrm{E}\left[\frac{\partial^2 \ln L_N(\alpha_q,\beta_q,H_q(\cdot|\delta_q))/N}{\partial(\alpha_q',\beta_q',\delta_q')'\partial(\alpha_q',\beta_q',\delta_q')}\right].$$

Next, suppose that the following condition holds.

Assumption 6*. The ranking of $p_{m,\ell}(q) = H_q(\exp(-\exp(\beta_q'x_\ell + \alpha_q'\omega_m)))$ is the same as the ranking of $p_{m,\ell}^0 = P[\Sigma_{i=1}^m D_{i,j} = 0 | X_j = x_\ell]$.

Then the ranking of the QML estimates

$$\tilde{p}_{m,\ell}(q) = \tilde{H}_q(\exp(-\exp(\tilde{\beta}_q'x_\ell + \tilde{\alpha}_q'\omega_m))) \tag{99}$$

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

of the $p_{m,\ell}(q)$'s will be equal to the ranking of the $p^0_{m,\ell}$'s with probability converging to one.

Note that a sufficient condition for Assumption 6* to hold is that it is possible to choose q such that

$$\max_{m,\ell} |p^0_{m,\ell} - p_{m,\ell}(q)| < \frac{1}{2} \min_{(m_1,\ell_1)\neq(m_2,\ell_2)} |p^0_{m_1,\ell_1} - p^0_{m_2,\ell_2}|. \tag{100}$$

Then it follows from (95) that $(\alpha_q', \beta_q')' \in \Theta_0$; hence we may interpret α_q and β_q as the "true" parameters,

$$\alpha_0 = \alpha_q, \qquad \beta_0 = \beta_q, \qquad (\alpha_0', \beta_0')' \in \Theta_0, \tag{101}$$

with corresponding QML estimators $\hat{\alpha} = \tilde{\alpha}_q$, $\hat{\beta} = \tilde{\beta}_q$, satisfying

$$\lim_{N\to\infty} P[(\hat{\alpha}', \hat{\beta}')' \in \Theta_0] = 1$$

(because of Assumption 5*), and

$$\sqrt{N}\begin{pmatrix} \hat{\alpha} - \alpha_0 \\ \hat{\beta} - \beta_0 \end{pmatrix} \to N_{M+k}[0, \Sigma] \tag{102}$$

in distribution, where $\Sigma = (I_{M+k}, O)\Omega_2^{-1}\Omega_1\Omega_2^{-1}(I_{M+k}, O)'$.

9.3. Consistent Estimation of the Minimal Polynomial Order

Given the reparametrization (101) it now follows from Lemma 3 that there exist a smallest polynomial order n_0 and a c.d.f. $H_{n_0} \in \mathcal{H}_{n_0}(0,1)$ such that

$$P[\Sigma_{i=1}^m D_{i,j} = 0 | X_j = x_\ell] = H_{n_0}(\exp(-\exp(\beta_0' x_\ell + \alpha_0' \omega_m))),$$
$$m = 1, \ldots, M, \qquad \ell = 1, \ldots, K. \tag{103}$$

I will show that n_0 can be estimated consistently by using an information criterion similar to the Hannan and Quinn (1979) and Schwarz (1978) information criteria:

$$C_N(n) = \frac{-2}{N} \sup_{H\in\mathcal{H}_n(0,1)} \ln(L_N(\hat{\alpha}, \hat{\beta}, H)) + (M + k + n)\cdot\frac{\varphi(N)}{N}, \tag{104}$$

$$\text{where} \quad \lim_{N\to\infty} \varphi(N)/N = 0, \qquad \lim_{N\to\infty} \varphi(N) = \infty. \tag{105}$$

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

Recall that the Hannan and Quinn (1979) criterion corresponds to the case $\varphi(N) = 2\ln(\ln(N))$ and the Schwarz (1978) criterion to the case $\varphi(N) = \ln(N)$, but any $\varphi(N) > 0$ satisfying (105) will do.

THEOREM 13. *Given Assumptions 1*–6* and the reparametrization (101), let n_0 be the smallest polynomial order for which there exists an SNP distribution function H_{n_0} such that (103) is true. Let $C_N(n)$ be defined by (104). Then*

$$p\lim_{N\to\infty} C_N(n+1) < p\lim_{N\to\infty} C_N(n) \quad \text{if } n < n_0, \tag{106}$$

$$p\lim_{N\to\infty} \frac{N}{\varphi(N)}\,(C_N(n) - C_N(n-1)) = 1 \quad \text{if } n > n_0. \tag{107}$$

Consequently, $\hat{n} = \max_{s.t.\ C_N(n)<C_N(n-1),\, n\geq 2} n$ is a consistent estimator of n_0:

$$\lim_{N\to\infty} P[\hat{n} = n_0] = 1. \tag{108}$$

Proof. See the Appendix.

9.4. Consistency of the SNP Probability Estimators

The partial optimization problem $\sup_{h\in\mathcal{D}_{\tilde{n}}(0,1)} \ln(L_N(\hat{\alpha},\hat{\beta},h))$ yields SNP estimates

$$\hat{p}^*_{m,\ell} = \hat{H}(\exp(-\exp(\hat{\beta}' x_\ell + \hat{\alpha}'\omega_m))) \tag{109}$$

of the probabilities $p^0_{m,\ell}$ in (90). Although $\hat{H}(u)$ may not be unique on $[0,1]$, we nevertheless have the following result.

THEOREM 14. *Under the conditions of Theorem 13 the SNP probability estimators (109) are consistent: $p\lim_{N\to\infty} \hat{p}^*_{m,\ell} = P[\Sigma_{i=1}^m D_{i,j} = 0 | X_j = x_\ell]$.*

Proof. See the Appendix.

It is clear that Assumption 6* is the crux of the results in this section. Although designing a test for the validity of Assumption 6* is beyond the scope of this paper, in the absence of such a test one should at least check whether the ranking of the $\hat{p}^*_{m,\ell}$'s matches the ranking of the estimates $\tilde{p}_{m,\ell}(q)$ in (99) based on the QML results. If not, increase q and redo the estimation until the rankings match.

10. CONCLUDING REMARKS

Because the ICMPH model under review is equivalent to an ordered probability model of the form (46), the results in this paper are straightforwardly appli-

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

cable to more general ordered probability models also, simply by replacing (47) with $F_0(x) = 1 - H_0(1 - G(x))$, where $G(x)$ is a given distribution function. For example, let $G(x) = 1/(1 + \exp(-x))$. Then (46) becomes an SNP ordered logit model. The only difference with the ICMPH model is that moment conditions (73) and (74) need to be adjusted, because these moment conditions correspond to the special case $G(x) = 1 - \exp(-\exp(x))$.

Moreover, it seems that the results in this paper can easily be adapted to ordered probability models based on Hermite polynomials.

The idea to model SNP densities and distribution functions using Legendre polynomials on the unit interval is the most straightforward part of this paper. The main contributions in this paper are (a) the application of this idea to interval-censored mixed proportional hazard models; (b) the derivation of the conditions under which these models are nonparametrically identified; (c) the construction of the compact metric space of density and distribution functions on the unit interval; (d) the weak consistency results for sieve M-estimators under weak and verifiable conditions; and (e) the two-step approach to get around the lack of identification problem in the case where the support of the covariates is finite.

Although there is a fair amount of literature on asymptotic normality of SNP parameter estimators (see, e.g., Chen, 2007 and the references therein), this literature is not directly applicable to the cases considered in this paper. For deriving asymptotic normality, however, we need to work with the scores of the log-likelihood, which involves ratios of densities and distribution functions on the unit interval and their derivatives. In view of the asymptotic normality conditions in Chen (2007, Sect. 4.2.1) the L^1 metric on $\mathcal{D}(0,1)$ will not be sufficient to derive asymptotic normality results for the models under review. The asymptotic normality problem will be left for future research.

The current topology of the space $\mathcal{D}(0,1)$ is also not sufficient for consistent sieve estimation of MPH models with right censoring only, because then the log-likelihood function involves the log of a density $h(u)$ on $[0,1]$, for which the L^1 metric is too weak to prove consistency along the lines in this paper. This problem also will be left for future research.

NOTES

1. However, it appears that only a few discrete covariates affect the two durations. Therefore, the presence of unobserved heterogeneity could not be detected.

2. In the sense that the parameters and also the unobserved heterogeneity distribution are identified.

3. Note that this integral is the Lebesgue integral.

4. See, e.g., Young (1988, pp. 24–25).

5. See, e.g., Royden (1968, pp. 59, 66).

6. See, e.g., Chung (1974, Sect. 4.2).

7. Note that $L_{n+1}\Pi_{n+1}(1)L'_{n+1} = I_{n+1}$. With $n = 15$, the computed elements of the matrix $L_{n+1}\Pi_{n+1}(1)L'_{n+1} - I_{n+1}$ are smaller in absolute value than 0.000000001. Moreover, for $n > 20$

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

some of the elements $\ell_{n,k}$ become too big (more than 29 digits, including the decimal point) to be stored in the memory of a PC.

8. See, e.g., van den Berg (2001) and the references therein.
9. The continuity also follows from the dominated convergence theorem.
10. The results in Meyer (1995) are based on this assumption.
11. This was pointed out by a referee.
12. Using one million random drawings from the uniform $[0,1]$ distribution.
13. Note that this assumption is a stronger condition than Assumption 1(a), as it implies that the elements of S can be ordered such that $\beta_0'x_1 < \beta_0'x_2 < \cdots < \beta_0'x_K$, which may not be possible if some of the components of β_0 are zero. Moreover, Assumption 1(b) is no longer applicable, because $\beta_0'x$ is now a one-to-one mapping on S, which implies that $\mathrm{E}[X_j|\beta_0'X_j] = \mathrm{E}[X_j|X_j] = X_j$. Therefore, Σ_0 in Assumption 1(b) is now a zero matrix!
14. Or $H_n(u|\eta)$ defined by (30).

REFERENCES

Bierens, H.J. (1982) Consistent model specification tests. *Journal of Econometrics* 20, 105–134.

Bierens, H.J. (1994) *Topics in Advanced Econometrics*. Cambridge University Press.

Bierens, H.J. (2004) *Introduction to the Mathematical and Statistical Foundations of Econometrics*. Cambridge University Press.

Bierens, H.J. & J.R. Carvalho (2007a) Semi-nonparametric competing risks analysis of recidivism. *Journal of Applied Econometrics* 22, 971–993.

Bierens, H.J. & J.R. Carvalho (2007b) Job Search, Conditional Treatment and Recidivism: The Employment Services for Ex-Offenders Program Reconsidered. Manuscript, Department of Economics, Pennsylvania State University (http://econ.la.psu.edu/~hbierens/PAPERS.HTM).

Chen, X. (2007) Large sample sieve estimation of semi-nonparametric models. In J. Heckman & E. Leamer (eds.), *Handbook of Econometrics*, vol. 6B, ch. 76. Elsevier.

Chung, K.L. (1974) *A Course in Probability Theory*. Academic Press.

Elbers, C. & G. Ridder (1982) True and spurious duration dependence: The identifiability of the proportional hazard model. *Review of Economic Studies* 49, 403–409.

Gallant, A.R. & D.W. Nychka (1987) Semi-nonparametric maximum likelihood estimation. *Econometrica* 55, 363–390.

Hamming, R.W. (1973) *Numerical Methods for Scientists and Engineers*. Dover Publications.

Hannan, E.J. & B.G. Quinn (1979) The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B* 41, 190–195.

Heckman, J.J. (1979) Sample selection bias as a specification error. *Econometrica* 47, 153–161.

Heckman, J.J. & B. Singer (1984) A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* 52, 271–320.

Jennrich, R.I. (1969) Asymptotic properties of nonlinear least squares estimators. *Annals of Mathematical Statistics* 40, 633–643.

Kiefer, J. & J. Wolfowitz (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics* 27, 887–906.

Lancaster, T. (1979) Econometric methods for the duration of unemployment. *Econometrica* 47, 939–956.

Magder, L.S. & S.L. Zeger (1996) A smooth nonparametric estimate of a mixing distribution using a mixture of Gaussians. *Journal of the American Statistical Association* 91, 1141–1151.

Meyer, B.D. (1995) Semiparametric Estimation of Hazard Models. Manuscript, Department of Economics, Northwestern University (http://www.faculty.econ.northwestern.edu/faculty/meyer/SEMIR2.pdf).

Nelder, J.A. & R. Mead (1965) A simplex method for function minimization. *Computer Journal* 7, 308–311.

Royden, H.L. (1968) *Real Analysis*. Macmillan.

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.
van den Berg, G.J. (2001) Duration models: Specification, identification, and multiple durations. In J. Heckman & E. Leamer (eds.), *Handbook of Econometrics*, vol. 5, ch. 55, pp. 3381–3460. Elsevier.
White, H. (1982) Maximum likelihood estimation of misspecified models. *Econometrica* 50, 1–25.
Young, N. (1988) *An Introduction to Hilbert Space*. Cambridge University Press.

APPENDIX: Proofs

Proof of Lemma 2. Denote $\Upsilon(u) = \ln(-\ln(H^{-1}(H_0(u))))$ and observe that $\Upsilon(u)$ is monotonic decreasing on $(0,1)$. Then it follows from (51) that

$$\beta' X = \Upsilon(\exp(-\exp(\beta_0' X)\alpha_{0,1})) - \ln \alpha_1 = \varphi(\beta_0' X),$$

e.g., where φ is a monotonic increasing function on $\mathbb{R}$. Hence $\mathrm{E}[\beta' X | \beta_0' X] = \varphi(\beta_0' X) = \beta' X$, and thus $\beta' \Sigma_0 \beta = 0$. The latter implies that β is an eigenvector of Σ_0 corresponding to its (only) zero eigenvalue. But so is β_0; hence $\beta = c\beta_0$ for some constant $c \neq 0$. Because now $c\beta_0' X = \varphi(\beta_0' X)$ and φ is monotonic increasing, c cannot be negative. Finally, the result (54) follows straightforwardly from $\beta = c\beta_0$. ■

Proof of Theorem 8. The theorem involved follows from the following lemmas.

LEMMA A.1. *Let $\xi = \{\xi_k\}_{k=0}^{\infty}$ be a given sequence of positive numbers satisfying*

$$\xi_0 > 1, \qquad \sum_{k=1}^{\infty} \xi_k^2 < \infty \tag{A.1}$$

and let $\mathcal{F}_\xi$ be the set of functions $f(u) = \sum_{k=0}^{\infty} \gamma_k \rho_k(u)$ in $L_{\mathcal{B}}^2(0,1)$ for which $\gamma_k \in [-\xi_k, \xi_k]$, $k = 0,1,2,\ldots$, endowed with the metric (12). Then $\mathcal{F}_\xi$ is compact.

Proof. It suffices to prove that $\mathcal{F}_\xi$ is complete and totally bounded. See Royden (1968, Prop. 15, p. 164).

To prove completeness, let $f_n(u) = \sum_{k=0}^{\infty} \gamma_{n,k} \rho_k(u)$ be an arbitrary Cauchy sequence in $\mathcal{F}_\xi$. Because $f_n(u)$ is a Cauchy sequence in the Hilbert space $L_{\mathcal{B}}^2(0,1)$ it converges to a function $f(u) = \sum_{k=0}^{\infty} \gamma_k \rho_k(u)$ in $L_{\mathcal{B}}^2(0,1)$. Now $\mathcal{F}_\xi$ is complete if $f \in \mathcal{F}_\xi$. Thus, we need to show that $\gamma_k \in [-\xi_k, \xi_k]$ for all k and $\sum_{k=0}^{\infty} \gamma_k^2 = 1$.

To prove $\gamma_k \in [-\xi_k, \xi_k]$, note that $\|f_n - f\|_2 = \sqrt{\sum_{k=0}^{\infty} (\gamma_{n,k} - \gamma_k)^2} \to 0$ implies that for each k, $\gamma_{n,k} \to \gamma_k$. Because $\gamma_{n,k} \in [-\xi_k, \xi_k]$ it follows that $\gamma_k \in [-\xi_k, \xi_k]$.

To prove $\sum_{k=0}^{\infty} \gamma_k^2 = 1$, let $\varepsilon \in (0,1)$ be arbitrary. Because $\sum_{k=0}^{m} \gamma_{n,k}^2 = 1 - \sum_{k=m+1}^{\infty} \gamma_{n,k}^2 \geq 1 - \sum_{k=m+1}^{\infty} \xi_k^2$ we can choose m so large that uniformly in n, $0 \leq 1 - \sum_{k=0}^{m} \gamma_{n,k}^2 < \varepsilon$. Because for $k = 0,1,\ldots,m$, $\gamma_{n,k} \to \gamma_k$, it follows that $0 \leq 1 - \sum_{k=0}^{m} \gamma_k^2 < \varepsilon$. Thus $\lim_{m\to\infty} \sum_{k=0}^{m} \gamma_k^2 = 1$. Hence $\mathcal{F}_\xi$ is complete.

To prove total boundedness, let $\varepsilon > 0$ be arbitrary and let $\mathcal{F}_{\xi,n}$ be the space of functions $f_n(u) = \sum_{k=0}^{n} \gamma_k \rho_k(u)$ such that $\sum_{k=0}^{n} \gamma_k^2 \leq 1$ and $\gamma_k \in [-\xi_k, \xi_k]$, $k = 0,1,2,\ldots,n$. Choose n so large that $\sum_{k=n+1}^{\infty} \xi_k^2 < \varepsilon$. Then for each $f \in \mathcal{F}_\xi$ there exists an $f_n \in \mathcal{F}_{\xi,n}$ such that $\|f - f_n\|_2 < \varepsilon$. The set of vectors $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_n)'$ satisfying $\gamma \in \times_{k=0}^{n} [-\xi_k, \xi_k]$, $\gamma'\gamma \leq 1$ is a closed and bounded subset of $\mathbb{R}^{n+1}$ and is therefore compact, and consequently, $\mathcal{F}_{\xi,n}$ is compact. Therefore, there exists a finite number of functions $f_1, \ldots, f_M \in \mathcal{F}_{\xi,n}$ such that

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

$$\mathcal{F}_{\xi,n} \subset \cup_{j=1}^{M}\{f \in \mathcal{F}_{\xi}(0,1) : \|f - f_j\|_2 < \varepsilon\}.$$

This implies that

$$\mathcal{F}_{\xi} \subset \cup_{j=1}^{M}\{f \in \mathcal{F}_{\xi}(0,1) : \|f - f_j\|_2 < 2\varepsilon\}$$

$$\subset \cup_{j=1}^{M}\{f \in L_{\mathcal{B}}^2(0,1) : \|f - f_j\|_2 < 2\varepsilon\};$$

hence $\mathcal{F}_{\xi}$ is totally bounded. ■

LEMMA A.2. *Under condition (A.1) the space*

$$\mathcal{F}_{\xi}^{*} = \left\{ f : f(u) = \frac{1 + \sum_{k=1}^{\infty} \delta_k \rho_k(u)}{\sqrt{1 + \sum_{k=1}^{\infty} \delta_k^2}}, \delta_k^2 \le \xi_k^2 \right\}$$

endowed with the metric (12) is compact.

Proof. It follows from (19) and (A.1) that

$$\gamma_k^2 = \frac{\delta_k^2}{1 + \sum_{k=1}^{\infty} \delta_k^2} \le \xi_k^2, \qquad k \ge 1,$$

$$\gamma_0^2 = \frac{1}{1 + \sum_{k=1}^{\infty} \delta_k^2} \le 1 < \xi_0^2, \tag{A.2}$$

$$\gamma_0^2 \ge \frac{1}{1 + \sum_{k=1}^{\infty} \xi_k^2} > 0;$$

hence $\mathcal{F}_{\xi}^{*} \subset \mathcal{F}_{\xi}$.

For a metric space the notions of compactness and sequential compactness are equivalent. See Royden (1968, Cor. 14, p. 163). Sequential compactness means that any infinite sequence in the metric space has a convergent subsequence that converges to an element in this space. Therefore, any infinite sequence $f_n \in \mathcal{F}_{\xi}^{*} \subset \mathcal{F}_{\xi}$ has a convergent subsequence

$$f_{m_n}(u) = \frac{1 + \sum_{k=1}^{\infty} \delta_{k,m_n} \rho_k(u)}{\sqrt{1 + \sum_{k=1}^{\infty} \delta_{k,m_n}^2}}$$

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

with limit

$$f(u) = \sum_{k=0}^{\infty} \gamma_k \rho_k(u) \in \mathcal{F}_\xi(0,1).$$

It is easy to verify that

$$\gamma_0 = \lim_{n\to\infty} \frac{1}{\sqrt{1 + \sum_{k=1}^{\infty} \delta_{k,m_n}^2}} \geq \frac{1}{\sqrt{1 + \sum_{k=1}^{\infty} \xi_k^2}} > 0,$$

$$\gamma_k = \lim_{n\to\infty} \frac{\delta_{k,m_n}}{\sqrt{1 + \sum_{k=1}^{\infty} \delta_{k,m_n}^2}} = \gamma_0 \lim_{n\to\infty} \delta_{k,m_n}, \qquad k \geq 1.$$

Denoting $\delta_k = \gamma_k/\gamma_0$ we can write $f(u)$ as

$$f(u) = \frac{1 + \sum_{k=1}^{\infty} \delta_k \rho_k(u)}{\sqrt{1 + \sum_{k=1}^{\infty} \delta_k^2}},$$

where $\delta_k^2 = \lim_{n\to\infty} \delta_{k,m_n}^2 \leq \xi_k^2$, so that $f \in \mathcal{F}_\xi^*$. Thus $\mathcal{F}_\xi^*$ is sequentially compact and hence compact. ■

LEMMA A.3. *The space* $\mathcal{D}_\xi(0,1) = \{h : h = f^2, f \in \mathcal{F}_\xi^*\}$ *of density functions on* $[0,1]$ *endowed with the metric (81) is compact.*

Proof. It follows from Schwarz inequality that for each pair of functions $f, g \in \mathcal{F}_\xi^*$

$$\begin{aligned}
&\int_0^1 |f(u)^2 - g(u)^2|\, du \\
&\quad \leq \int_0^1 |f(u) - g(u)||f(u)|\, du + \int_0^1 |f(u) - g(u)||g(u)|\, du \\
&\quad \leq \sqrt{\int_0^1 (f(u) - g(u))^2\, du}\left(\sqrt{\int_0^1 f(u)^2\, du} + \sqrt{\int_0^1 g(u)^2\, du}\right) \\
&\quad = 2\sqrt{\int_0^1 (f(u) - g(u))^2\, du}. \qquad \textbf{(A.3)}
\end{aligned}$$

Let $h_n = f_n^2$ be an infinite sequence in $\mathcal{D}_\xi(0,1)$. Because $\mathcal{F}_\xi^*$ is compact, there exists a subsequence f_{m_n} that converges to a limit f in $\mathcal{F}_\xi^*$; hence it follows from (A.3) that $h_{m_n} = f_{m_n}^2$ converges to $h = f^2$. Thus $\mathcal{D}_\xi(0,1)$ is sequentially compact and therefore compact. ■

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

We can choose ξ_k such that $\mathcal{D}(0,1) \subset \mathcal{D}_{\xi}(0,1)$. It is now easy to verify that $\mathcal{D}(0,1)$ is sequentially compact and therefore compact and that this implies that $\mathcal{H}(0,1)$ is compact.

Proof of Theorem 10. It follows from the Jennrich (1969) uniform strong law of large numbers, in the version in Bierens (1994, Sect. 2.7) or Bierens (2004, App. to Ch. 6), that under the conditions of Theorem 10, with the conditions (88) and (89) replaced by

$$\mathrm{E}\left[\sup_{\theta\in\Theta}|g(Y_1,\theta)|\right] < \infty \tag{A.4}$$

we have

$$\lim_{N\to\infty}\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{j=1}^{N} g(Y_j,\theta) - \bar{g}(\theta)\right| = 0 \qquad \text{a.s.} \tag{A.5}$$

However, the condition (A.4) is too difficult to verify in the log-likelihood case. Therefore I will use the weaker conditions (88) and (89).

Originally the uniform strong law (A.5) was derived by Jennrich (1969) for the case that Θ is a compact subset of a euclidean space, but it is easy to verify from the more detailed proofs in Bierens (1994, Sect. 2.7) or Bierens (2004, App. to Ch. 6) that this law carries over to random functions on compact metric spaces.

Let $K > K_0$ and note that

$$\mathrm{E}\left[\max\left(\sup_{\theta\in\Theta} g(Y_1,\theta), -K\right)\right] \le \mathrm{E}\left[\max\left(\sup_{\theta\in\Theta} g(Y_1,\theta), -K_0\right)\right] < \infty;$$

hence $\mathrm{E}[\sup_{\theta\in\Theta}|\max(g(Y_1,\theta),-K)|] < \infty$. Then it follows from (A.5) with $g(Y_j,\theta)$ replaced by $\max(g(Y_j,\theta),-K)$ that

$$\lim_{N\to\infty}\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{j=1}^{N}\max(g(Y_j,\theta),-K) - \bar{g}_K(\theta)\right| = 0 \quad \text{a.s.}, \tag{A.6}$$

where $\bar{g}_K(\theta) = \mathrm{E}[\max(g(Y_j,\theta),-K)]$.

As is well known, (A.6) is equivalent to the statement that for all $\varepsilon > 0$,

$$\lim_{N\to\infty} P\left[\sup_{n\ge N}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n}\max(g(Y_j,\theta),-K) - \bar{g}_K(\theta)\right| < \varepsilon\right] = 1.$$

In its turn this is equivalent to the statement that for arbitrary natural numbers k and m there exists a natural number $N(K,k,m)$ such that for all $N \ge N(K,k,m)$,

$$P\left[\sup_{n\ge N}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n}\max(g(Y_j,\theta),-K) - \bar{g}_K(\theta)\right| < \frac{1}{k}\right] > 1 - \frac{1}{m}.$$

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

Let $k \le K \le m$. Then there exists a natural number $N(K)$ such that for all $N \ge N(K)$,

$$P\left[\sup_{n\ge N}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n}\max(g(Y_j,\theta),-K)-\bar{g}_K(\theta)\right|<\frac{1}{K}\right]>1-\frac{1}{K}.$$

For given N, let K_N be the maximum K for which $N \ge N(K)$. Then

$$P\left[\sup_{n\ge N(K_N)}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n}\max(g(Y_j,\theta),-K_N)-\bar{g}_{K_N}(\theta)\right|<\frac{1}{K_N}\right]>1-\frac{1}{K_N};$$

hence, for arbitrary $\varepsilon > 0$,

$$\lim_{N\to\infty}P\left[\sup_{n\ge N(K_N)}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{j=1}^{n}\max(g(Y_j,\theta),-K_N)-\bar{g}_{K_N}(\theta)\right|<\varepsilon\right]=1.$$

This result implies that along the subsequence $n_N = N(K_N)$,

$$\sup_{\theta\in\Theta}\left|\frac{1}{n_N}\sum_{j=1}^{n_N}\max(g(Y_j,\theta),-K_N)-\bar{g}_{K_N}(\theta)\right|\to 0 \quad \text{a.s.}, \tag{A.7}$$

and the same applies if we had replaced N first by an arbitrary subsequence. Thus, every subsequence of N contains a further subsequence n_N such that (A.7) holds. As is well known, a sequence of random variables converges in probability if and only if every subsequence contains a further subsequence along which the sequence involved converges a.s. Thus, (A.7) implies that there exists a sequence K_N converging to infinity with N such that

$$p\lim_{N\to\infty}\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{j=1}^{N}\max(g(Y_j,\theta),-K_N)-\bar{g}_{K_N}(\theta)\right|=0. \tag{A.8}$$

Because the function $\max(x,-K)$ is convex, it follows from Jensen's inequality that

$$\begin{aligned}\bar{g}_K(\theta) &= \mathrm{E}[\max(g(Y_j,\theta),-K)] \ge \max(\mathrm{E}[g(Y_j,\theta)],-K)\\ &= \max(\bar{g}(\theta),-K) \ge \bar{g}(\theta)\end{aligned} \tag{A.9}$$

and similarly

$$\frac{1}{N}\sum_{j=1}^{N}\max(g(Y_j,\theta),-K)\ge\max\left(\frac{1}{N}\sum_{j=1}^{N}g(Y_j,\theta),-K\right)\ge\frac{1}{N}\sum_{j=1}^{N}g(Y_j,\theta). \tag{A.10}$$

It follows from (89), (A.9), and the dominated convergence theorem that $\lim_{K\to\infty}\sup_{\theta\in\Theta}|\bar{g}_K(\theta)-\bar{g}(\theta)|=0$; hence (A.8) now becomes

$$p\lim_{N\to\infty}\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{j=1}^{N}\max(g(Y_j,\theta),-K_N)-\bar{g}(\theta)\right|=0. \tag{A.11}$$

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

Finally, observe from (A.10) that

$$\frac{1}{N}\sum_{j=1}^{N}\max(g(Y_j,\hat{\theta}),-K_N)-\bar{g}(\hat{\theta})$$

$$\geq \frac{1}{N}\sum_{j=1}^{N} g(Y_j,\hat{\theta})-\bar{g}(\hat{\theta})$$

$$\geq \frac{1}{N}\sum_{j=1}^{N} g(Y_j,\theta_0)-\bar{g}(\theta_0)+\bar{g}(\theta_0)-\bar{g}(\hat{\theta})$$

$$\geq \frac{1}{N}\sum_{j=1}^{N} g(Y_j,\theta_0)-\bar{g}(\theta_0). \tag{A.12}$$

By Kolmogorov's strong law of large numbers, the lower bound in (A.12) converges a.s. to zero, and by (A.11) the upper bound in (A.12) converges in probability to zero; hence $p\lim_{N\to\infty}((1/N)\sum_{j=1}^{N} g(Y_j,\hat{\theta})-\bar{g}(\hat{\theta}))=0$, which implies

$$p\lim_{N\to\infty}\bar{g}(\hat{\theta})=\bar{g}(\theta_0). \tag{A.13}$$

Because θ_0 is unique it follows from the continuity of $\bar{g}(\theta)$ and the compactness of Θ that there exists an $\bar{\varepsilon}>0$ such that for all $\varepsilon\in(0,\bar{\varepsilon}]$, $\sup_{\theta\in\Theta,\,d(\theta,\theta_0)\geq\varepsilon}\bar{g}(\theta)<\bar{g}(\theta_0)$. See, e.g., Bierens (2004, App. II, Thm. II.6). It follows therefore from (A.13) that

$$P[d(\hat{\theta},\theta_0)\geq\varepsilon]\leq P\left[\bar{g}(\hat{\theta})\leq\sup_{\theta\in\Theta,\,d(\theta,\theta_0)\geq\varepsilon}\bar{g}(\theta)\right]\to 0. \qquad \blacksquare$$

Proof of Theorem 11. Similar to (A.12) we have

$$\frac{1}{N}\sum_{j=1}^{N}\max(g(Y_j,\tilde{\theta}),-K_N)-\bar{g}(\tilde{\theta})$$

$$\geq \frac{1}{N}\sum_{j=1}^{N} g(Y_j,\theta_N)-\bar{g}(\tilde{\theta})$$

$$\geq \frac{1}{N}\sum_{j=1}^{N} g(Y_j,\theta_N)-\bar{g}(\theta_0)+\bar{g}(\theta_0)-\bar{g}(\tilde{\theta})$$

$$\geq \frac{1}{N}\sum_{j=1}^{N}(g(Y_j,\theta_N)-g(Y_j,\theta_0))+\frac{1}{N}\sum_{j=1}^{N} g(Y_j,\theta_0)-\bar{g}(\theta_0).$$

It follows from (A.11) that

$$p\lim_{N\to\infty}\left(\frac{1}{N}\sum_{j=1}^{N}\max(g(Y_j,\tilde{\theta}),-K_N)-\bar{g}(\tilde{\theta})\right)=0,$$

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

and it follows from Kolmogorov's strong law of large numbers that

$$\frac{1}{N}\sum_{j=1}^{N} g(Y_j,\theta_0) \to \bar{g}(\theta_0) \quad \text{a.s.} \tag{A.14}$$

Moreover, it follows from the continuity of $\mathrm{E}[|g(Y_1,\theta) - g(Y_1,\theta_0)|]$ in θ and $\lim_{N\to\infty} d(\theta_N,\theta_0) = 0$ that

$$\mathrm{E}\left|\frac{1}{N}\sum_{j=1}^{N}(g(Y_j,\theta_N) - g(Y_j,\theta_0))\right| \le \mathrm{E}|g(Y_1,\theta_N) - g(Y_1,\theta_0)| \to 0.$$

Hence by Chebyshev's inequality,

$$p\lim_{N\to\infty}\frac{1}{N}\sum_{j=1}^{N}(g(Y_j,\theta_N) - g(Y_j,\theta_0)) = 0. \tag{A.15}$$

Thus, $p\lim_{N\to\infty}((1/N)\sum_{j=1}^{N} g(Y_j,\theta_N) - \bar{g}(\tilde{\theta})) = 0$, which by (A.14) and (A.15) implies that $p\lim_{N\to\infty}\bar{g}(\tilde{\theta}) = \bar{g}(\theta_0)$. ■

Proof of Lemma 3. First, consider the case $K = 1$. Let $\bar{H}(u)$ be a distribution function on $[0,1]$ such that $\bar{H}(u_1) > v_1$ and let $\underline{H}(u)$ be a distribution function on $[0,1]$ such that $\underline{H}(u_1) < v_1$. Then it follows trivially from Theorem 3 that there exist an n and a pair $\bar{\delta},\underline{\delta} \in \mathbb{R}^n$ such that $H_n(u_1|\bar{\delta}) > v_1$ and $H_n(u_1|\underline{\delta}) < v_1$. Because $\varphi_n(\lambda) = H_n(u_1|(1-\lambda)\underline{\delta} + \lambda\bar{\delta}) - v_1$ is a continuous function of $\lambda \in [0,1]$, with $\varphi(0) < 0$, $\varphi(1) > 0$, there exists a $\lambda \in (0,1)$ such that $\varphi_n(\lambda) = 0$. For this λ, let $\delta = (1-\lambda)\underline{\delta} + \lambda\bar{\delta}$. Then $H_n(u_1|\delta) = v_1$.

This argument shows that n can be chosen so large that the sets $\Delta_{i,n} = \{\delta \in \mathbb{R}^n : H_n(u_i|\delta) = v_i\}$, $i = 1,\ldots,K$, are nonempty. It remains to show that there exist an n such that $\cap_{i=1}^{K}\Delta_{i,n} \neq \varnothing$.

Next, consider the case $K = 2$. Again, it follows straightforwardly from Theorem 3 that there exist an n and a pair $\bar{\delta},\underline{\delta} \in \mathbb{R}^n$ such that

$$H_n(u_1|\underline{\delta}) < v_1, \qquad H_n(u_1|\bar{\delta}) > v_1,$$

$$H_n(u_2|\underline{\delta}) < v_2, \qquad H_n(u_2|\bar{\delta}) > v_2.$$

For such a pair $\bar{\delta},\underline{\delta}$, consider the function

$$\psi_n(\lambda|M) = H_n(u_2|(1-\lambda)\underline{\delta} + \lambda\bar{\delta}) - v_2 + \left(\frac{H_n(u_1|(1-\lambda)\underline{\delta} + \lambda\bar{\delta})}{v_1}\right)^M - \left(\frac{v_1}{H_n(u_1|(1-\lambda)\underline{\delta} + \lambda\bar{\delta})}\right)^M,$$

where $M > 0$ is a natural number and $\lambda \in [0,1]$. For each $M > 0$, $\psi_n(\lambda|M)$ is continuous in $\lambda \in [0,1]$, and for sufficient large M, $\psi_n(0|M) < 0$, $\psi_n(1|M) > 0$. Consequently, for such an $M > 0$ there exists a $\lambda_M \in (0,1)$ for which $\psi_n(\lambda_M|M) = 0$, so that

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

$$H_n(u_2|(1-\lambda_M)\underline{\delta}+\lambda_M\bar{\delta})$$

$$= v_2 + \left(\frac{v_1}{H_n(u_1|(1-\lambda_M)\underline{\delta}+\lambda\bar{\delta}_M)}\right)^M - \left(\frac{H_n(u_1|(1-\lambda_M)\underline{\delta}+\lambda_M\bar{\delta})}{v_1}\right)^M.$$

Now λ_M is a sequence in the compact interval [0,1]; hence all the limit points of λ_M are contained in [0,1], and for each limit point λ_* there exists a subsequence M_j such that $\lim_{j\to\infty}\lambda_{M_j}=\lambda_*$. Thus, e.g.,

$$H_n(u_1|(1-\lambda_*)\underline{\delta}+\lambda_*\bar{\delta}) = \lim_{j\to\infty} H_n(u_1|(1-\lambda_{M_j})\underline{\delta}+\lambda_{M_j}\bar{\delta}) = \eta.$$

If $\eta > v_1$ then

$$\lim_{j\to\infty}\left(\frac{H_n(u_1|(1-\lambda_{M_j})\underline{\delta}+\lambda_{M_j}\bar{\delta})}{v_1}\right)^{M_j} = \infty,$$

$$\lim_{j\to\infty}\left(\frac{v_1}{H_n(u_1|(1-\lambda_{M_j})\underline{\delta}+\lambda_{M_j}\bar{\delta})}\right)^{M_j} = 0;$$

hence $H_n(u_2|(1-\lambda_*)\underline{\delta}+\lambda_*\bar{\delta})=\infty$, which is impossible. Similarly, if $\eta < v_1$ then $H_n(u_2|(1-\lambda_*)\underline{\delta}+\lambda_*\bar{\delta})=-\infty$, which is again impossible. Consequently, $\eta = v_1$, so that

$$H_n(u_1|(1-\lambda_*)\underline{\delta}+\lambda_*\bar{\delta}) = v_1, \qquad H_n(u_2|(1-\lambda_*)\underline{\delta}+\lambda_*\bar{\delta}) = v_2.$$

This argument shows that for large enough n, $\Delta_{1,n}\cap\Delta_{2,n}\neq\varnothing$. Also, this result implies that we can choose an n and a pair $\underline{\delta},\bar{\delta}\in\Delta_{1,n}$ such that

$$H_n(u_2|\underline{\delta}) < v_2, \qquad H_n(u_2|\bar{\delta}) > v_2.$$

The latter will be the basis for the proof of the general case.

Finally, consider the case $K>2$. Let $m\geq 2$ and suppose that for large enough n there exists a pair $\bar{\delta},\underline{\delta}\in\cap_{i=1}^{m-1}\Delta_{i,n}\neq\varnothing$ such that $H_n(u_m|\bar{\delta})>v_m$, $H_n(u_m|\underline{\delta})<v_m$. This assumption has been proved for $m=2$. Next, consider the function

$$\Phi_n^{(m)}(\lambda|M) = H_n(u_m|(1-\lambda)\underline{\delta}+\lambda\bar{\delta}) - v_m$$

$$+\sum_{i=1}^{m-1}\left(\left(\frac{H_n(u_i|(1-\lambda)\underline{\delta}+\lambda\bar{\delta})}{v_i}\right)^M + \left(\frac{v_i}{H_n(u_i|(1-\lambda)\underline{\delta}+\lambda\bar{\delta})}\right)^M\right)$$

$$\times (H_n(u_i|(1-\lambda)\underline{\delta}+\lambda\bar{\delta})-v_i)^2,$$

where again $M>0$ is a natural number and $\lambda\in[0,1]$. Note that

$$\Phi_n^{(m)}(0|M) = H_n(u_m|\underline{\delta}) - v_m < 0,$$

$$\Phi_n^{(m)}(1|M) = H_n(u_m|\bar{\delta}) - v_m > 0,$$

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

so that by the continuity of $\Phi_n^{(m)}(\lambda|M)$ in λ there exists a $\lambda_M \in (0,1)$ such that $\Phi_n^{(m)}(\lambda_M|M) = 0$. Similar to the case $K = 2$ it follows that for all limit points λ_* of λ_M, $H_n(u_i|(1-\lambda_*)\underline{\delta} + \lambda_*\bar{\delta}) = v_i$, $i = 1,\ldots,m$; hence $\cap_{i=1}^m \Delta_{i,n} \neq \varnothing$. The general result follows now by induction.

To show that δ may not be unique, consider the case $K = 1$, with $u_1 = \frac{1}{2}$ and $v_1 < \frac{1}{2}$. It is easy to verify that

$$v_1 = H_1(\tfrac{1}{2}|\delta) = \frac{\int_0^{1/2} (1+\delta\sqrt{3}(2u-1))^2\,du}{1+\delta^2} = \frac{1}{2}\left(\frac{1-2\delta^2}{1+\delta^2}\right);$$

hence $\delta = \pm\sqrt{1-2v_1}/\sqrt{2(1+v_1)}$. ■

Proof of Theorem 13. Part (106) of Theorem 13 follows from the inequality

$$p\lim_{N\to\infty}\frac{1}{N}\sup_{H\in\mathcal{H}_n(0,1)}\ln(L_N(\hat{\alpha},\hat{\beta},H)) < p\lim_{N\to\infty}\frac{1}{N}\sup_{H\in\mathcal{H}_{n+1}(0,1)}\ln(L_N(\hat{\alpha},\hat{\beta},H))$$

for $n < n_0$, which in its turn follows trivially from Theorem 10. Part (107) is true if for $n > n_0$,

$$\sup_{H\in\mathcal{H}_n(0,1)}\ln(L_N(\hat{\alpha},\hat{\beta},H)) - \sup_{H\in\mathcal{H}_{n_0}(0,1)}\ln(L_N(\hat{\alpha},\hat{\beta},H)) = O_p(1). \tag{A.16}$$

To show that (A.16) holds, observe first that for all n,

$$\sup_{H\in\mathcal{H}_n(0,1)}\ln(L_N(\hat{\alpha},\hat{\beta},H)) \le \ln(L_N(\hat{P})),$$

where $\hat{P}$ is the $M \times K$ matrix with elements $\hat{p}_{m,\ell}$ defined by (94) and $\ln(L_N(P))$ is defined by (92). Moreover, it follows from standard ML (ratio test) theory that

$$2(\ln(L_N(\hat{P})) - \ln(L_N(P_0))) \to \chi^2_{M\cdot K}$$

in distribution, where P_0 is matrix (93), so that for all n,

$$\sup_{H\in\mathcal{H}_n(0,1)}\ln(L_N(\hat{\alpha},\hat{\beta},H)) \le \ln(L_N(P_0)) + O_p(1). \tag{A.17}$$

Furthermore, it is trivial that

$$\sup_{H\in\mathcal{H}_{n_0}(0,1)}\ln(L_N(\hat{\alpha},\hat{\beta},H)) \ge \ln(L_N(\hat{\alpha},\hat{\beta},H_{n_0})). \tag{A.18}$$

Finally, it follows straightforwardly from (102) and the second-order Taylor expansion of $\ln(L_N(\hat{\alpha},\hat{\beta},H_{n_0}))$ around $\ln(L_N(\alpha_0,\beta_0,H_{n_0})) = \ln(L_N(P_0))$ that

$$\ln(L_N(\hat{\alpha},\hat{\beta},H_{n_0})) = \ln(L_N(P_0)) + O_p(1). \tag{A.19}$$

Combining (A.17)–(A.19) for $n > n_0$, (A.16) follows. ■

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press

Proof of Theorem 14. Let $\bar{\Psi}(\alpha,\beta,H) = \mathrm{E}[N^{-1}\ln(L_N(\alpha,\beta,H))]$. It follows from Theorem 10 and (108) that

$$p\lim_{N\to\infty}\bar{\Psi}(\hat{\alpha},\hat{\beta},\hat{H}) = \sup_{H\in\mathcal{H}_{n_0}(0,1)}\bar{\Psi}(\alpha_0,\beta_0,H). \tag{A.20}$$

Next, denote $\hat{P}_* = (\hat{p}^*_{m,\ell};\, m = 1,\ldots,M,\ \ell = 1,\ldots,K)$ and let

$$g(P) = \mathrm{E}[N^{-1}\ln(L_N(P))] = \sum_{\ell=1}^{K}(1-p^0_{1,\ell})\ln(1-p_{1,\ell})$$

$$= \sum_{i=2}^{M}\sum_{\ell=1}^{K}(p^0_{i-1,\ell}-p^0_{i,\ell})\ln(p_{i-1,\ell}-p_{i,\ell}) + \sum_{\ell=1}^{K}p^0_{M,\ell}\ln(p_{M,\ell}).$$

Cf. (92). Then $\bar{\Psi}(\hat{\alpha},\hat{\beta},\hat{H}) = g(\hat{P}_*)$ and $\sup_{H\in\mathcal{H}_{n_0}(0,1)}\bar{\Psi}(\alpha_0,\beta_0,H) = g(P_0)$. Hence, it follows from (A.20) that $p\lim_{N\to\infty} g(\hat{P}_*) = g(P_0)$, which by the continuity of $g(P)$ in the elements of P implies that $p\lim_{N\to\infty}\hat{P}_* = P_0$. ■

https://doi.org/10.1017/S0266466608080316 Published online by Cambridge University Press