Semi-Nonparametric Modeling of Densities on the Unit Interval, with Application to Interval-Censored Mixed Proportional Hazard Models: Identification and Consistency Results

Herman J. Bierens*
Pennsylvania State University

February 6, 2006

Abstract

In this paper I propose to estimate densities with possibly restricted support semi-nonparametrically (SNP) using SNP densities on the unit interval based on 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.
JEL Codes: C14, C21, C24, C25, C41

*Correspondence address: Department of Economics, Penn State University, 608 Kern Graduate Building, University Park, PA 16802. E-mail: hbierens@psu.edu. 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, and Ohio State University. The helpful comments of Aleksandr Vashchilko, Yuichi Kitamura and two referees are gratefully acknowledged.

Key words: Semi-nonparametric, Legendre polynomials, Duration models, Interval-censoring, Identification, Consistency, Discrete covariates.

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 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 expansion in terms of orthonormal Legendre polynomials. Therefore, the density $h(u)$ 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), but where now 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 (2006), where the durations involved, job search and recidivism, are only observed in the form of intervals.[1]

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

[1]However, since only a few discrete covariates affected the two durations, the presence of unobserved heterogeneity could not be detected.

of the systematic and baseline hazards are nonparametrically identified, and the sieve maximum likelihood estimators involved are consistent.

The plan of the paper is the following. In Section 2 the Legendre polynomials are be introduced and their use motivated, and 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 interval-censored MPH model and in Section 5 I will derive conditions for nonparametric identification under interval-censoring. Due to 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 re-derived 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 non-zero 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 SNP maximum likelihood sieve estimators. One of the requirements is that the space of density functions h involved is compact. Therefore, in Section 7 I will show how to construct a compact metric space of densities 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 that the covariates have finite support. In Section 10 the main contribution of this paper are summarized, and avenues

[2] In the sense that the parameters as well as the unobserved heterogeneity distribution are identified.

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 Heckman's (1979) sample selection model, where the bivariate error distribution of the latent variable equations is modeled semi-nonparametrically using an 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 $I(.)$ is the indicator function. These polynomials can easily be generated via the recursive relation $\sqrt{n}\mu_n(x) = x.\mu_{n-1}(x) - \sqrt{n-1}\mu_{n-2}(x)$, starting from $\mu_0(x) = 1$, $\mu_1(x) = x$. See for example Hamming (1973, p. 457). The densities (1) can be approximated arbitrarily close 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

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

which yields the standard Probit model as a special case, corresponding to the null hypothesis $\gamma_{k,n} = 0$ for $k = 1, ..., n$. 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.

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

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

with corresponding distribution function $H_n(u)=F_n\left(F_0^{-1}(u)\right)$, 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. Therefore, it is more convenient to define SNP densities on the unit interval directly on the basis of orthonormal polynomials on the unit interval than indirectly using Hermite polynomials.

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 $n\geq 2$ these polynomials can be constructed recursively by

$$P_n(z)=\frac{(2n-1)z.P_{n-1}(z)-(n-1)P_{n-2}(z)}{n} \tag{4}$$

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

$$\int_{-1}^{1}P_m(z)P_n(z)dz=\begin{cases}0 & if \quad n\neq m,\\ 2/(2n+1) & if \quad n=m.\end{cases} \tag{5}$$

See for example 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 (5) that the polynomials $\rho_n(u)$ are orthonormal:

$$\int_{0}^{1}\rho_k(u)\rho_m(u)du=\begin{cases}0 & if \quad k\neq m,\\ 1 & if \quad k=m,\end{cases} \tag{6}$$

and from (4) that for $n\geq 2$ they can be computed recursively by

$$\rho_n(u)=\frac{\sqrt{4n^2-1}}{n}(2u-1)\rho_{n-1}(u)-\frac{(n-1)\sqrt{2n+1}}{n\sqrt{2n-3}}\rho_{n-2}(u), \tag{7}$$

starting from

$$\rho_0(u)=1,\ \rho_1(u)=\sqrt{3}\,(2u-1)\,. \tag{8}$$

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 below 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 $\{u \in [0,1]\colon\ f(u) \neq \sum_{k=0}^{\infty} \gamma_k \rho_k(u)\}$ has Lebesgue measure zero. In other words, the Legendre polynomials $\rho_k(u)$ on $[0,1]$ form an complete orthonormal basis for the Hilbert space $L^2_{\mathcal{B}}(0,1)$ of Borel measurable 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

$$\|f - g\|_2 = \sqrt{\int_0^1 (f(u) - g(u))^2 \, du} \tag{9}$$

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.

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*

$$\int_0^1 |f_1(u)| \, du < \infty, \int_0^1 |f_2(u)| \, du < \infty. \tag{10}$$

[3] Note that this integral is the Lebegue integral.

[4] See for example Young (1988, pp. 24-25).

[5] See for example Royden (1968, pp. 59 and 66)

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

$$\int_0^1 u^k f_1(u)du = \int_0^1 u^k f_2(u)du. \tag{11}$$

Each u^k can be written as a linear combination of $\rho_0(u), \rho_1(u), ..., \rho_k(u)$ with Fourier coefficients $\int_0^1 u^k \rho_m(u)du$, $m = 0, 1, .., k$, hence condition (11) is equivalent to $\int_0^1 \rho_k(u) f_1(u)du = \int_0^1 \rho_k(u) f_2(u)du$ for $k = 0, 1, 2, ...$

Now let in Lemma 1, $f_1(u) = f(u)$ 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

$$\sum_{k=0}^{\infty} \gamma_k^2 < \infty \tag{12}$$

because $\int_0^1 \left(f(u) - \sum_{k=0}^n \gamma_k \rho_k(u)\right)^2 du$ 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, due to (12), $f_{2,n}(u) = \sum_{k=0}^n \gamma_k \rho_k(u)$ is a Cauchy sequence in the Hilbert space $L^2_{\mathcal{B}}(0,1)$ and therefore has a limit $f_2 \in L^2_{\mathcal{B}}(0,1)$: $\lim_{n\to\infty} \int_0^1 \left(f_{2,n}(u) - f_2(u)\right)^2 du = 0$. Hence, we can write $f_2(u) = \sum_{k=0}^n \gamma_k \rho_k(u) + r_n(u)$, were $\lim_{n\to\infty} \int_0^1 r_n(u)^2 du = 0$. 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 Chebishev's inequality that $f_2(u) = \lim_{m\to\infty} \sum_{k=0}^{n_m} \gamma_k \rho_k(u)$ 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.

Every density function $h(u)$ on $[0,1]$ is Borel measurable because, with H the corresponding distribution function, $h(u) = \lim_{k\to\infty} k\left(H(u + k^{-1}) - H(u)\right)$, 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 $\int_0^1 f(u)^2 du = 1$.

[6] See for example, Chung (1974, Section 4.2).

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 $\widetilde{B} = [0,1]\backslash B$,

$$\phi_B(u) = I(u \in B) - I(u \in \widetilde{B}). \tag{13}$$

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 (13). Consequently, any density $h(u)$ on $[0,1]$ can be represented by

$$h(u) = \left(\sum_{k=0}^{\infty} \gamma_k \rho_k(u)\right)^2, \text{ with } \gamma_k = \int_0^1 \rho_k(u)\phi_B(u)\sqrt{h(u)}du. \tag{14}$$

We can always choose B is such that

$$\gamma_0 = \int_0^1 \phi_B(u)\sqrt{h(u)}du = \int_B \sqrt{h(u)}du - \int_{\widetilde{B}} \sqrt{h(u)}du > 0. \tag{15}$$

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:

$$\begin{aligned} \gamma_k &= \frac{\delta_k}{\sqrt{1 + \sum_{k=1}^{\infty} \delta_k^2}}, \ k = 1, 2, 3, ..., \\ \gamma_0 &= \frac{1}{\sqrt{1 + \sum_{k=1}^{\infty} \delta_k^2}}, \end{aligned} \tag{16}$$

where $\sum_{k=1}^{\infty} \delta_k^2 < \infty$. However, because there are uncountable many Borel subsets B of $[0,1]$ for which (15) holds, there are also uncountable many of such reparametrizations. Thus,

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

$$h(u) = \frac{\left(1 + \sum_{k=1}^{\infty} \delta_k \rho_k(u)\right)^2}{1 + \sum_{k=1}^{\infty} \delta_k^2} \ a.e. \ on \ [0,1]. \tag{17}$$

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

$$h_n(u) = h_n(u|\delta) = \frac{\left(1 + \sum_{k=1}^n \delta_k \rho_k(u)\right)^2}{1 + \sum_{k=1}^n \delta_k^2}, \ \delta = (\delta_1, ..., \delta_n)'. \tag{18}$$

It is straightforward to verify that

Theorem 3. *For each density $h(u)$ on $[0,1]$ there exists a sequence of densities $h_n(u)$ of the type* (18) *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 $\lim_{n\to\infty} \sup_{0\leq u\leq 1} |H(u) - H_n(u)| = 0$.*

Following Gallant and Nychka (1987), the density functions of the type (18) 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 (17) have uncountable many equivalent series representations. This is no longer the case for SNP densities:

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 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 (18), can be written as

$$\begin{aligned} H_n(u|\delta) &= \frac{\int_0^u \left(1 + \sum_{m=1}^n \delta_m \rho_m(v)\right)^2 dv}{1 + \sum_{m=1}^n \delta_m^2} = \frac{(1, \delta')\, A_{n+1}(u) \binom{1}{\delta}}{1 + \delta'\delta}, \\ u &\in [0,1], \ \delta = (\delta_1, ..., \delta_n)', \end{aligned} \tag{19}$$

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

$$A_{n+1}(u) = \left(\int_0^u \rho_i(v)\rho_j(v)dv \ ; \ i,j = 0,1,....,n \right). \tag{20}$$

Let $\rho_m(u) = \sum_{k=0}^m \ell_{m,k} u^k$. Then it follows from (7) and (8) that

$$\ell_{0,0} = 1, \ \ell_{1,0} = -\sqrt{3}, \ \ell_{1,1} = 2\sqrt{3}, \tag{21}$$

and for $m \geq 2$,

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

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

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

starting from (21). 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 m-th row $(\ell_{m,0}, ..., \ell_{m,n})$.

Next, observe that

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

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

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

Therefore,

$$\begin{aligned} H_n(u|\delta) &= \frac{(1,\delta') \, L_{n+1} \Pi_{n+1}(u) L'_{n+1} \binom{1}{\delta}}{1+\delta'\delta}, \\ u &\in [0,1], \ \delta = (\delta_1, ..., \delta_n)'. \end{aligned} \tag{22}$$

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

4 The interval-censored mixed proportional hazard model

4.1 The MPH 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 $F(t|X) = P[T \leq t|X]$, $f(t|X)$ 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 mixed proportional hazard model assumes that the conditional survival function takes the form

$$\begin{aligned} & S(t|X,\alpha,\beta) = S(t|X) \\ = \; & E\left[\exp\left(-\exp\left(\beta' X + U\right)\int_0^t \lambda(\tau|\alpha)d\tau\right)\middle| X\right], \end{aligned} \tag{23}$$

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

$$\begin{aligned} S(t|X,\alpha,\beta,h) &= \int_0^\infty \exp\left(-v.\exp(\beta' X)\Lambda(t|\alpha)\right)dG(v) \\ &= \int_0^\infty \left(\exp\left(-\exp(\beta' X)\Lambda(t|\alpha)\right)\right)^v dG(v) \\ &= H\left(\exp\left(-\exp(\beta' X)\Lambda(t|\alpha)\right)\right), \end{aligned} \tag{24}$$

[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 $n > 20$ 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 for example Van den Berg (2000) and the references therein.

where

$$H(u)=\int_0^\infty u^v dG(v),\ u\in[0,1], \tag{25}$$

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

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

$$\int_0^\infty vu^{v-1}dG(v)\leq u^{-1}\int_0^\infty vdG(v)<\infty, \tag{26}$$

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

$$h(u)=\int_0^\infty vu^{v-1}dG(v). \tag{27}$$

This is the reason for the argument h in the left-hand side of (24). Moreover, (26) 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

$$\begin{aligned} S(t|X,\alpha_0,\beta_0,h_0) &= \int_0^\infty \exp\left(-v.\exp(\beta_0'X)\Lambda(t|\alpha_0)\right)dG_0(v) \\ &= H_0\left(\exp\left(-\exp(\beta_0'X)\Lambda(t|\alpha_0)\right)\right) \end{aligned} \tag{28}$$

where $H_0(u)=\int_0^u h_0(v)dv=\int_0^\infty u^v dG_0(v)$. In the expressions (24) and (28), h and h_0 should be interpreted as unknown parameters contained in a parameter space $\mathcal{D}(0,1)$, say, of density functions on $(0,1]$.

For the ease of reference I will call this model the Interval-Censored Mixed Proportional Hazard (ICMPH) model.

Elbers and Ridder (1982) have shown that if X does not contain a constant,

$$\Lambda(t|\alpha)=\Lambda(t|\alpha_0) \text{ for all } t>0 \text{ implies } \alpha=\alpha_0, \tag{29}$$

and

$$\int_0^\infty vdG_0(v)=\int_0^\infty vdG(v)=1 \tag{30}$$

[9] The continuity also follows from the dominated convergence theorem.

(which by (27) is equivalent to confining the parameter space $\mathcal{D}(0,1)$ to a space of densities h on $(0,1]$ satisfying $h(1)=1$), then the MPH model is nonparametrically identified, in the sense that

$$S(T|X,\alpha,\beta_0,h) = S(T|X,\alpha_0,\beta_0,h_0) \text{ a.s.}$$

implies $\alpha = \alpha_0$ and $G = G_0$, hence $h(u) = h_0(u)$ a.e. on $[0,1]$. Heckman and Singer (1984) provide an alternative identification proof based on the results of Kiefer and Wolfowitz (1956), and propose to parametrize G_0 as a discrete distribution: $G_0(v) = \sum_{i=1}^{q} I(v \leq \theta_i)\, p_i$, with $I(.)$ the indicator function, where $\theta_i > 0$, $p_i > 0$, and $\sum_{i=1}^{q} p_i = 1$. Thus, they implicitly specify $h_0(u) = \sum_{i=1}^{q} \theta_i u^{\theta_i - 1} p_i$.

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 below with Meyer (1995) will reveal.

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

$$P[T_j^* > t|X_j] = S(t|X_j,\alpha_0,\beta_0,h_0), \tag{31}$$

where $S(t|X,\alpha,\beta,h)$ is defined by (28), α_0 and β_0 are the true parameter vectors and h_0 is the true density (27). 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 $b_1 < b_2 < ... < b_M \leq \overline{T}$, and create the dummy variables

$$\begin{aligned} D_{1,j} &= I\left(T_j \leq b_1\right) \\ D_{2,j} &= I\left(b_1 < T_j \leq b_2\right) \\ &\vdots \\ D_{M,j} &= I\left(b_{M-1} < T_j \leq b_M\right) \end{aligned} \tag{32}$$

where $I(.)$ is the indicator function. See Meyer (1995). Also, in some cases the durations T_j are only observed in the form of intervals. See for example Bierens and Carvalho (2006).

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

$$\mu_i\left(\alpha, \beta' X_j\right) = \exp\left(-\exp(\beta' X_j)\Lambda\left(b_i|\alpha\right)\right). \tag{33}$$

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

$$\begin{aligned} P[D_{i,j} = 1|X_j] &= S(b_{i-1}|X_j, \alpha_0, \beta_0, h_0) - S(b_i|X_j, \alpha_0, \beta_0, h_0) \quad (34) \\ &= H_0\left(\mu_{i-1}\left(\alpha_0, \beta_0' X_j\right)\right) - H_0\left(\mu_i\left(\alpha_0, \beta_0' X_j\right)\right) \\ &\quad i = 1, 2, ..., M, \\ P\left[\sum_{i=1}^{M} D_{i,j} = 0 \middle| X_j\right] &= S(b_M|X_j, \alpha_0, \beta_0, h_0) = H_0\left(\mu_M\left(\alpha_0, \beta_0' X_j\right)\right), \end{aligned}$$

where $H_0(u) = \int_0^u h_0(v)dv$. The density h_0 will be treated as a parameter.

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

$$\begin{aligned} &\ln\left(L_N(\alpha, \beta, h)\right) \\ &= \sum_{j=1}^{N}\sum_{i=1}^{M} D_{i,j} \ln\left(H\left(\mu_{i-1}\left(\alpha, \beta' X_j\right)\right) - H\left(\mu_i\left(\alpha, \beta' X_j\right)\right)\right) \\ &+ \sum_{j=1}^{N}\left(1 - \sum_{i=1}^{M} D_{i,j}\right) \ln\left(H\left(\mu_M\left(\alpha, \beta' X_j\right)\right)\right) \end{aligned} \tag{35}$$

with $H(u) = \int_0^u h(v)dv$.

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 $t = b_1, ..., b_M$. Therefore we may without loss of generality parametrize $\Lambda(t|\alpha)$ as a piecewise linear function:

$$\begin{aligned} \Lambda(t|\alpha) &= \Lambda(b_{i-1}|\alpha) + \alpha_i (t - b_{i-1}) \\ &= \sum_{k=1}^{i-1} \alpha_k (b_k - b_{k-1}) + \alpha_i (t - b_{i-1}) \text{ for } t \in (b_{i-1}, b_i], \\ \alpha_m &> 0 \text{ for } m = 1, ..., M, \ \alpha = (\alpha_1, ..., \alpha_M)' \in \mathbb{R}^M. \end{aligned} \tag{36}$$

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

$$\Lambda(b_i|\alpha) = \sum_{m=1}^{i} \alpha_m, \ \alpha_m > 0 \text{ for } m = 1, ..., M, \tag{37}$$

or

$$\begin{aligned} \Lambda(b_i|\alpha) &= \exp(\alpha_i), \ \alpha_1 < \alpha_2 < ... < \alpha_M, \\ \Lambda(b_0|\alpha) &= \Lambda(0|\alpha) = 0. \end{aligned} \tag{38}$$

The advantage of the specification (36) is that the null hypothesis $\alpha_1 = ... = \alpha_M$ 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 (37) is useful for deriving identification conditions, as will be shown in the next subsection. The same applies to (38).

Note that under specification (38) the probability model (34) takes the form of a generalized ordered probability model, similarly to an ordered probit or logit model:

$$P[\Sigma_{i=1}^m D_{i,j} = 1 \mid X_j] = F_0(\beta_0' X_j + \alpha_{0,m}), \ m = 1, 2, ..., M, \tag{39}$$

where

$$F_0(x) = 1 - H_0(\exp(-\exp(x))), \tag{40}$$

with density

$$f_0(x) = h_0(\exp(-\exp(x))) \exp(-\exp(x)) \exp(x). \tag{41}$$

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 (39). Moreover, for the identification of α_0 and β_0 in (39) 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

5.1 Introduction

Let us assume that $\Lambda(b_i|\alpha)$ is specified as (37). It follows easily from the inequality $\ln(x) < x - 1$ if $x > 0$ and $x \neq 1$, and the equality

$$E\left[L_N(\alpha, \beta, h)/L_N(\alpha_0, \beta_0, h_0)|X_1, ..., X_N\right] = 1 \text{ a.s.}$$

that

$$E\left[N^{-1}\ln\left(L_N(\alpha, \beta, h)/L_N(\alpha_0, \beta_0, h_0)\right)\middle| X_1, ..., X_N\right] < 0$$

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

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

This implies that

$$E\left[N^{-1}\ln\left(L_N(\alpha, \beta, h)/L_N(\alpha_0, \beta_0, h_0)\right)\right] = 0 \tag{42}$$

if and only if

$$H\left(\exp\left(-\exp\left(\beta' X\right)\left(\sum_{m=1}^{i}\alpha_m\right)\right)\right) \tag{43}$$
$$= H_0\left(\exp\left(-\exp\left(\beta_0' X\right)\left(\sum_{m=1}^{i}\alpha_{0,m}\right)\right)\right)$$
$$\textit{a.s. for } i = 1, ..., M,$$

where $X = X_j$.

Now the question arises: Under what conditions does (43) 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:

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

For $i = 1$, (43) reads

$$H\left(\exp\left(-\exp\left(\beta' X\right)\alpha_1\right)\right) = H_0\left(\exp\left(-\exp\left(\beta_0' X\right)\alpha_{0,1}\right)\right)\ a.s. \tag{44}$$

To derive further condition such that (46) 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 $\beta = c.\beta_0$, so that (44) now reads

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

or equivalently,

$$H\left(\exp\left(-\alpha_1\alpha_{0,1}^{-c}\left(\ln\left(1/u\right)\right)^c\right)\right) = H_0\left(u\right) \tag{45}$$

for all u in the support S_1 of $U = \exp\left(-\exp\left(\beta_0 X\right)\alpha_{0,1}\right)$. Note that, due to the monotonicity of H and H_0, (45) implies that $c > 0$.

Next, consider the case that $X \in \mathbb{R}^k$ with $k \geq 2$. I will set forth a further condition under which $\beta' X = c.\beta_0' X$ for some constant $c > 0$, as follows. Denote $\Upsilon(u) = \ln\left(-\ln\left(H^{-1}\left(H_0\left(u\right)\right)\right)\right)$, and observe that $\Upsilon(u)$ is monotonic decreasing on $(0,1)$. Then it follows from (44) that

$$\begin{aligned} \beta' X &= \Upsilon\left(\exp\left(-\exp\left(\beta_0' X\right)\alpha_{0,1}\right)\right) - \ln\alpha_1 \\ &= \varphi\left(\beta_0' X\right), \end{aligned} \tag{46}$$

for example, where φ is a monotonic increasing function on $\mathbb{R}$. Now augment $||\beta_0||^{-1}\beta_0$ with a $k \times (k-1)$ matrix Q_1 to form an orthogonal matrix Q:

$$Q = \left(\frac{1}{||\beta_0||}\beta_0, Q_1\right) \tag{47}$$

and define

$$\begin{aligned} Z &= Q'X = \begin{pmatrix} ||\beta_0||^{-1}\beta_0'X \\ Q_1'X \end{pmatrix} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \\ \gamma &= Q'\beta = \begin{pmatrix} ||\beta_0||^{-1}\beta_0'\beta \\ Q_1'\beta \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} \end{aligned} \tag{48}$$

so that by (46),

$$\gamma_1 Z_1 + Z_2'\gamma_2 = \gamma' Z = \beta' X = \varphi\left(||\beta_0||.Z_1\right). \tag{49}$$

If $\gamma_2 = 0$ then $\beta'X = \gamma_1 Z_1 = \gamma_1 ||\beta_0||^{-1}\beta_0'X = c.\beta_0'X$, say. To establish that $\gamma_2 = 0$, observe from (49) that

$$Z_2'\gamma_2 = E[Z_2'\gamma_2|Z_1] = \varphi\left(||\beta_0||.Z_1\right) - \gamma_1 Z_1,$$

hence, denoting

$$W = Z_2 - E[Z_2|Z_1] \tag{50}$$

we have $W'\gamma_2 = 0$ and thus $E\left[WW'\right]\gamma_2 = 0$. Therefore, $\gamma_2 = 0$ if $E\left[WW'\right]$ is nonsingular, which is the case if and only if

Assumption 1(b): *The matrix* $\Sigma_0 = E\left[\left(X - E[X|\beta_0'X]\right)\left(X - E[X|\beta_0'X]\right)'\right]$ *has only one zero eigenvalue (corresponding to the eigenvector* β_0*).*

Note that $\beta_0'X - E[\beta_0'X|\beta_0'X] = 0$, so that $\Sigma_0\beta_0 = 0$.

To show that Assumption 1(b) is a necessary and sufficient condition for $\det\left(E\left[WW'\right]\right) > 0$, observe from (48) and (50) that $E\left[WW'\right] = Q_1'\Sigma_0 Q_1$. Without loss of generality we may assume that Q_1 in (47) is the matrix of orthonormal eigenvectors corresponding to the $k-1$ positive eigenvalues of Σ_0, so that

$$\Sigma_0 = \left(||\beta_0||^{-1}\beta_0, Q_1\right)\begin{pmatrix} 0 & 0' \\ 0 & \Lambda \end{pmatrix}\begin{pmatrix} ||\beta_0||^{-1}\beta_0' \\ Q_1' \end{pmatrix} = Q_1\Lambda Q_1',$$

where Λ is the diagonal matrix with the positive eigenvalues of Σ_0 on the diagonal. Thus $E\left[WW'\right] = Q_1'Q_1\Lambda Q_1'Q_1 = \Lambda$.

If Σ_0 has more than one zero eigenvalue then there exists a nonzero vector γ_2 such that $\gamma_2'Z_2 = \gamma_2'Q_1'X$ is a function of $\beta_0'X$. Since $\beta_1 = Q_1\gamma_2$ is orthogonal to β_0, Assumption 1(b) is therefore equivalent to the statement that

$$\beta_1 \neq 0,\ \beta_1'\beta_0 = 0 \Rightarrow P\left(\beta_1'X = E\left[\beta_1'X|\beta_0'X\right]\right) < 1.$$

Assumptions 1(a)-(b) together will now be referred to as Assumption 1. Thus, the following result has been shown.

Lemma 2. *Suppose that*

$$\begin{aligned} P\left[T > b_1|X\right] &= H_0\left(\exp\left(-\exp\left(\beta_0'X\right)\alpha_{0,1}\right)\right) \qquad (51) \\ &= H\left(\exp\left(-\exp\left(\beta'X\right)\alpha_1\right)\right)\ \ a.s. \end{aligned}$$

Under Assumption 1, (51) *implies that there exists a constant* $c > 0$ *such that* $\beta = c.\beta_0$*, so that for all* u *in the support* S_1 *of* $U = \exp\left(-\exp\left(\beta_0'X\right)\alpha_{0,1}\right)$,

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

Given β_0, for every $\alpha_{0,1} > 0$ there exists a distribution function H_0 such that the first equality in (51) is true. Moreover, given β_0, $\alpha_{0,1}$ and H_0, for every $c > 0$ there exists an $\alpha_1 > 0$ and a distribution function H such that (52) holds. Therefore, to establish first that $c = 1$ (so that $\beta = \beta_0$), and then that $H\left(u^{\alpha_1/\alpha_{0,1}}\right) = H_0\left(u\right)$ 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

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

then there are various options for normalizing H_0 and H, as follows.

5.2 Nonparametric identification via extreme values

Taking the derivative of (48) to u yields

$$\begin{aligned} h_0\left(u\right) &= h\left(\exp\left(-\alpha_1\alpha_{0,1}^{-c}\left(\ln\left(1/u\right)\right)^c\right)\right)\exp\left(-\alpha_1\alpha_{0,1}^{-c}\left(\ln\left(1/u\right)\right)^c\right) \qquad (53) \\ &\quad \times\alpha_1\alpha_{0,1}^{-c}c\left(\ln\left(1/u\right)\right)^{c-1}\frac{1}{u},\ \ \forall u \in S_1\ . \end{aligned}$$

[10]The results in Meyer (1995) are based on this assumption.

Now suppose that

$$\forall x \in \mathbb{R},\ P\left[\beta_0' X \leq x\right] > 0, \tag{54}$$

which is a weaker condition than Assumption 2, and

$$h_0(1) = h(1) = 1. \tag{55}$$

Recall that (55) corresponds to the condition that $E[V] = 1$. Moreover, using the easy equality $\rho_k(1) = \sqrt{2k+1}$, it follows straightforwardly from (17) that condition (55) can be implemented by restricting h_0 and h to density functions of the type (17), with

$$\begin{aligned} \delta_1 &= \frac{1}{2}\sqrt{2\left(1+\sum_{k=2}^{\infty}\delta_k^2\right)+\left(1+\sum_{k=2}^{\infty}\delta_k\sqrt{2k+1}\right)^2} \\ &\quad -\frac{\sqrt{3}}{2}\left(1+\sum_{k=2}^{\infty}\delta_k\sqrt{2k+1}\right). \end{aligned}$$

Condition (54) implies that S_1 contains a sequence u_n which converges to 1. Then

$$\begin{aligned} 1 &= \lim_{n\to\infty} h_0\left(u_n\right) = \lim_{n\to\infty} h\left(\exp\left(-\alpha_1\alpha_{0,1}^{-c}\left(\ln\left(1/u_n\right)\right)^c\right)\right) \\ &\quad \times \lim_{n\to\infty}\exp\left(-\alpha_1\alpha_{0,1}^{-c}\left(\ln\left(1/u_n\right)\right)^c\right) \\ &\quad \times\alpha_1\alpha_{0,1}^{-c}c\lim_{n\to\infty}\left(\ln\left(1/u_n\right)\right)^{c-1} \\ &= \alpha_1\alpha_{0,1}^{-c}c\lim_{n\to\infty}\left(\ln\left(1/u_n\right)\right)^{c-1} = \begin{cases} 0 & if \quad c>1, \\ \alpha_1/\alpha_{0,1} & if \quad c=1, \\ \infty & if \quad c<1. \end{cases} \end{aligned} \tag{56}$$

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

Since now β_0 and $\alpha_{0,1}$ are identified, the next question is: Does

$$\begin{aligned} &H\left(\exp\left(-\exp\left(\beta_0'X\right)\left(\alpha_{0,1}+\alpha_2\right)\right)\right) \\ &\quad = H_0\left(\exp\left(-\exp\left(\beta_0'X\right)\left(\alpha_{0,1}+\alpha_{0,2}\right)\right)\right)\ \ a.s. \end{aligned} \tag{57}$$

imply $\alpha_2 = \alpha_{0,2}$? Let S_2 be the support of $U = \exp\left(-\exp\left(\beta_0'X\right)\left(\alpha_{0,1}+\alpha_{0,2}\right)\right)$ and let

$$\eta = \frac{\alpha_2-\alpha_{0,2}}{\alpha_{0,1}+\alpha_{0,2}} \tag{58}$$

Then (57) implies that

$$H\left(u^{1+\eta}\right) = H_0\left(u\right) \text{ for all } u \in S_2. \tag{59}$$

Under condition (54) there exists a sequence u_n in S_2 which converges to 1. Therefore, similarly to (56) we have

$$\begin{aligned} 1 &= h_0(1) = \lim_{n\to\infty} \frac{H_0\left(1\right) - H_0\left(u_n\right)}{1 - u_n} \\ &= \lim_{n\to\infty} \frac{H\left(1\right) - H\left(u_n^{1+\eta}\right)}{1 - u_n} = h(1) \lim_{n\to\infty} \frac{1 - u_n^{1+\eta}}{1 - u_n} = 1 + \eta. \end{aligned}$$

Hence, under the conditions (54) and (55) , $\eta = 0$ and thus $\alpha_2 = \alpha_{0,2}$. Repeating this argument for $i = 3, .., M$, it follows that $\alpha = \alpha_0$ and

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

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

Since (36) and (37) are equivalent for $t = b_i$, the following result holds:

Theorem 5. *Let the integrated baseline hazard be specified by* (36). *Under Assumption 1 and the conditions* (54) *and* (55) *the equality*

$$E\left[\ln\left(L_N(\alpha,\beta,h)/L_N(\alpha_0,\beta_0,h_0)\right)\right] = 0 \tag{60}$$

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

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

a.s. for $i = 1, ..., M$. *If in addition Assumption 2 holds, with* $X = X_j$, *then* (61) *implies that* $h(u) = h_0(u)$ *a.e. on* $(0, 1]$.

Admittedly, Assumption 2 and even the weaker condition (54) 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 below I will consider the more realistic case that the covariates have finite support.

The condition (55) is only effective in pinning down α_0 and β_0 if $U_{i,j} = \exp\left(-\exp\left(\beta_0' X_j\right)\Lambda(b_i|\alpha_0)\right)$ can get close enough to 1. Thus, the identification hinges on the **extreme negative values** of $\beta_0' X_j$. Under Assumption 2 it follows that $p\lim_{N\to\infty}\min_{j=1,...,N}\beta_0' X_j = -\infty$, hence $p\lim_{N\to\infty}\max_{j=1,...,N} U_{i,j} = 1$, but in finite samples $\max_{j=1,...,N} U_{i,j}$ may not get close enough to 1 for condition (55) to be effective. Therefore, I will now derive alternative identification conditions.

5.3 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

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

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 1-2, $\left(\ln\left(1/u_1\right)\right)^{c-1} = \left(\ln\left(1/u_2\right)\right)^{c-1} = \alpha_{0,1}^c/\alpha_1$. Since $u_1 \neq u_2$, the first equality is only possible if $c = 1$, hence $\beta = \beta_0$, and the second equality then implies that $\alpha_1 = \alpha_{0,1}$.

Similarly, it follows from (58), (59) and (62) that $\alpha_2 = \alpha_{0,2}$. Therefore, similar to Theorem 5 we have that

Theorem 6. *Under Assumption 1-2 and the quantile conditions* (62) *the ICMPH model is nonparametrically identified.*

In other words, under Assumptions 1-2 the equality (60) implies that $\alpha = \alpha_0$, $\beta = \beta_0$, and $h(u) = h_0(u)$ a.e. on $(0,1]$.

5.4 Nonparametric identification via moment conditions

Consider the ordered probability model form (39) of the ICMPH model. Let $F(x)$ be a distribution function of the type (40),

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

with density

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

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

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

Clearly, under Assumptions 1-2 model (39) is nonparametrically identified if (65) implies that $\mu = 0$ and $\sigma = 1$. Taking derivatives of (65) it follows that (65) implies

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

hence it follows from (41) and (64) that for any function φ on $\mathbb{R}$ for which $\int_{-\infty}^{\infty}\varphi(x)f_0(x)dx$ is well-defined,

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

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

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

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

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

The latter equality 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, in order to accommodate the benchmark case $h_0(u) = h(u) \equiv 1$, which corresponds to absence of unobserved heterogeneity, I will assume that the density h in the log-likelihood function (35) is confined to a space of density functions h on $(0, 1]$ satisfying the moment conditions

$$\int_0^1 \ln(\ln(1/u))\,h(u)\,du = \int_0^1 \ln(\ln(1/u))\,du, \tag{73}$$
$$\int_0^1 (\ln(\ln(1/u)))^2 h(u)\,du = \int_0^1 (\ln(\ln(1/u)))^2\,du. \tag{74}$$

It is obvious from the easy equalities

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

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

$$\begin{aligned} \int_0^1 \ln(\ln(1/u)) \, du &= -0.577189511, \\ \int_0^1 (\ln(\ln(1/u)))^2 \, du &= 1.981063818, \end{aligned}$$

which have been computed by Monte Carlo integration.[11]

Theorem 7. *Let h and h_0 be confined to density functions satisfying the moment conditions* (73) *and* (74)*. Then under Assumptions 1-2 the ICMPH model is nonparametrically identified.*

5.5 Implementation of moment and quantile conditions

The moment conditions (73) and (74) can be implemented by penalizing the log-likelihood function (35) for deviations from the moment conditions involved by augmenting the log-likelihood function $\ln(L_N(\alpha, \beta, h))$ with two penalty terms:

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

for some integer $\ell \geq 1$.

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

$$\begin{aligned} \ln(L_N^*(\alpha, \beta, h)) &= \ln(L_N(\alpha, \beta, h)) \\ &- N\left(\int_0^{u_1} h(v) \, dv - u_1\right)^{2\ell} - N\left(\int_0^{u_2} h(v) \, dv - u_2\right)^{2\ell} \end{aligned} \tag{76}$$

[11] Using one million random drawings from the uniform $[0, 1]$ distribution.

for some integer $\ell \geq 1$.

For SNP density functions (18) 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

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

and

$$\begin{aligned} B_{n,p} &= \left(\int_0^1 (\ln(\ln(1/u)))^p\, \rho_{i_1}(u)\rho_{i_2}(u)du\ ;\ i_1, i_2 = 1, 2, ..., n\right) \\ &\quad - \int_0^1 (\ln(\ln(1/u)))^p\, du.I_n, \end{aligned}$$

the conditions (73) and (74) with h replaced by (18) 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 $\delta = (\delta_1, ..., \delta_n)'$. Therefore, if we replace h in (75) by $h_n(.|\delta)$, the penalized log-likelihood can be written as

$$\begin{aligned} &\ln(L_N^*(\alpha, \beta, h_n(.|\delta))) = \ln(L_N(\alpha, \beta, h_n(.|\delta))) \\ &- N\left(\frac{2\delta' a_{n,1} + \delta' B_{n,1}\delta}{1+\delta'\delta}\right)^{2\ell} - N\left(\frac{2\delta' a_{n,2} + \delta' B_{n,2}\delta}{1+\delta'\delta}\right)^{2\ell} \end{aligned} \tag{77}$$

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 density $h_0(u)$, and if we choose $\ell \geq 2$, then

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

and

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

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

Similarly, if we replace h in (76) by $h_n(.|\delta)$, the penalized log-likelihood becomes

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

with $H_n(u|\delta)$ defined by (22). For the same reason as before I recommend to choose $\ell = 2$.

6 Requirements for consistency of SNP maximum likelihood estimators

We can write the (penalized) log-likelihood as

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

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

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

$$+\left(1-\sum_{i=1}^{M}D_{i,j}\right)\ln\left(H\left(\mu_M\left(\alpha,\beta'X_j\right)\right)\right)-\Pi(h),$$

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

The maximum likelihood estimators of α_0, β_0 and h_0 are

$$\left(\widehat{\alpha},\widehat{\beta},\widehat{h}\right)=\arg\max_{\alpha\in A,\beta\in B,h\in\mathcal{D}(0,1)}N^{-1}\ln\left(L_N^*(\alpha,\beta,h)\right), \tag{79}$$

where

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 $\mathcal{D}(0,1)$ is a compact metric space of density functions on $[0,1]$, containing the true density h_0. The space $\mathcal{D}(0,1)$ will be endowed with the metric

$$\|h_1-h_2\|_1=\int_0^1|h_1(u)-h_2(u)|\,du. \tag{80}$$

Let

$$\overline{\Psi}(\alpha,\beta,h)=E\left[\Psi(Y_j,\alpha,\beta,h)\right]. \tag{81}$$

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

$$p\lim_{N\to\infty}\overline{\Psi}(\widehat{\alpha},\widehat{\beta},\widehat{h})=\overline{\Psi}(\alpha_0,\beta_0,h_0). \tag{82}$$

Similar to the standard consistency proof for M estimators it can be shown that if $\overline{\Psi}$ is continuous and (α_0,β_0,h_0) is unique then (82) implies that $p\lim_{N\to\infty}\widehat{\alpha}=\alpha_0$, $p\lim_{N\to\infty}\widehat{\beta}=\beta_0$ and $p\lim_{N\to\infty}\left\|\widehat{h}-h_0\right\|_1=0$.

In general it will be impossible to compute (79) because it requires to maximize the log-likelihood function over a space of density functions. However, there exists an increasing sequence $\mathcal{D}_{n_N}(0,1)$ of compact subspaces of $\mathcal{D}(0,1)$ such that the densities in $\mathcal{D}_{n_N}(0,1)$ can be parametrized by a finite (but increasing) number of parameters, namely a space of densities of the type (18), where $n=n_N$ is a subsequence of N, and the δ's are confined to a compact subset of $\mathbb{R}^{n_N}$. Then

$$\left(\widetilde{\alpha},\widetilde{\beta},\widetilde{h}\right)=\arg\max_{\alpha\in A,\beta\in B,h\in\mathcal{D}_{n_N}(0,1)}N^{-1}\ln\left(L_N^*(\alpha,\beta,h)\right) \tag{83}$$

is feasible. This is known as sieve estimation. See Chen (2005) for a review of sieve estimation. Moreover, it follows from Theorem 3 that we can choose a sequence of densities $h_N \in \mathcal{D}_{n_N}(0,1)$ such that $\lim_{N\to\infty} ||h_N - h_0||_1 = 0$. This result can be used to prove that $p\lim_{N\to\infty} \widetilde{\alpha} = \alpha_0$ and $p\lim_{N\to\infty} \widetilde{\beta} = \beta_0$, and $p\lim_{N\to\infty} ||\widetilde{h} - h_0||_1 = 0$.

The crux of the consistency problem is twofold, namely: (1) how to make the metric space $\mathcal{D}(0,1)$ compact; and (2) how to prove (82). These problems will be addresses in the next sections.

7 Compactness of the density space

Consider the space of density functions of the type (17), subject to the condition $\sum_{k=1}^{\infty} \delta_k^2 < \infty$. This condition can easily be imposed, for example by restricting the δ_k's such that for some constant $c > 0$,

$$|\delta_k| \leq \frac{c}{1 + \sqrt{k}\ln(k)}, \tag{84}$$

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 (84) also play a key-role in proving compactness:

Theorem 8. *Let $\mathcal{D}(0,1)$ be the space of densities of the type* (17) *subject to the restrictions* (84) *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.*

Proof: Appendix.

Of course, the result of Theorem 8 is only useful for our purpose if the constant c in (84) is chosen so large that

Assumption 4. *The true density h_0 is contained in $\mathcal{D}(0,1)$.*

Finally, it follows now straightforwardly from Theorem 3 that the following result holds.

Theorem 9. *For a subsequence $n = n_N$ of N, let $\mathcal{D}_n(0,1)$ be the space of densities of the type* (18) *subject to the restrictions* (84), *with c the same as for $\mathcal{D}(0,1)$. For each N, $\mathcal{D}_{n_N}(0,1)$ is a compact subset of $\mathcal{D}(0,1)$, and* *for each* $h \in \mathcal{D}(0,1)$ *there exists a sequence* $h_N \in \mathcal{D}_{n_N}(0,1)$ *such that* $\lim_{N\to\infty}\int_0^1 |h(u) - h_N(u)|\,du = 0.$

8 Consistency of M-estimators in the presence of non-Euclidean parameters

I will now address the problem how to prove (82). To relate (82) to Theorem 10 below, denote $\Theta = \{(\alpha,\beta,h) : \alpha \in A,\ \beta \in B,\ h \in \mathcal{D}(0,1)\}$, and define a metric $d(.,.)$ on Θ by combining the metrics on A, B and $\mathcal{D}(0,1)$. For example, for $\theta_1 = (\alpha_1,\beta_1,h_1) \in \Theta$, $\theta_2 = (\alpha_2,\beta_2,h_2) \in \Theta$, let

$$d(\theta_1,\theta_2) = \max\left[\sqrt{(\alpha_1-\alpha_2)'(\alpha_1-\alpha_2)},\right. \\ \left.\sqrt{(\beta_1-\beta_2)'(\beta_1-\beta_2)},\ \int_0^1 |h_1(u)-h_2(u)|\,du\right]. \tag{85}$$

Theorem 10. *Let* Y_j, $j = 1,...,N$, *be a sequence of i.i.d. random vectors in a Euclidean space, defined on a 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* $\theta \in \Theta$,

$$E\left[|g(Y_1,\theta)|\right] < \infty, \tag{86}$$

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

$$E\left[\max\left(\sup_{\theta\in\Theta} g(Y_1,\theta), -K_0\right)\right] < \infty. \tag{87}$$

Denote $\widehat{\theta} = \arg\max_{\theta\in\Theta} N^{-1}\sum_{j=1}^N g(Y_j,\theta)$ *and* $\theta_0 = \arg\max_{\theta\in\Theta}\overline{g}(\theta)$. *Then* $p\lim_{N\to\infty}\left(\frac{1}{N}\sum_{j=1}^N g(Y_j,\widehat{\theta}) - \overline{g}(\widehat{\theta})\right) = 0$ *and consequently,*

$$p\lim_{N\to\infty}\overline{g}(\widehat{\theta}) = \overline{g}(\theta_0). \tag{88}$$

(a) *If* θ_0 *is unique then* (88) *implies* $p\lim_{N\to\infty} d(\widehat{\theta},\theta_0) = 0$.

(*b*) *Suppose that* θ_0 *is partially unique, in the following sense. Let* $\theta_0 = \left(\theta_{0,1}', \theta_{0,2}'\right)' \in \Theta_1 \times \Theta_2 = \Theta$, *where* Θ_1 *and* Θ_2 *are compact metric spaces with metrics* d_1 *and* d_2 *respectively. There exists a subset* $\Theta_2^* \subset \Theta_2$, *possibly containing more than one point, such that for all* $\theta_2 \in \Theta_2^*$, $\overline{g}\left(\left(\theta_{0,1}', \theta_{0,2}\right)'\right) = \overline{g}\left(\left(\theta_{0,1}', \theta_2'\right)'\right)$. *On the other hand, let for all* $\theta_1 \in \Theta_1 \backslash \{\theta_{0,1}\}$,

$$\sup_{\theta_2 \in \Theta_2} \overline{g}\left((\theta_1', \theta_2')'\right) < \overline{g}\left(\left(\theta_{0,1}', \theta_{0,2}\right)'\right). \tag{89}$$

Partition $\widehat{\theta}$ *accordingly as* $\widehat{\theta} = \left(\widehat{\theta}_1', \widehat{\theta}_2'\right)' \in \Theta_1 \times \Theta_2$. *Then* $p\lim_{N\to\infty} d_1\left(\widehat{\theta}_1, \theta_{0,1}\right) = 0$ *and* $p\lim_{N\to\infty} \overline{g}\left(\left(\theta_{0,1}', \widehat{\theta}_2'\right)'\right) = \overline{g}\left(\left(\theta_{0,1}', \theta_{0,2}'\right)'\right)$.

Proof: 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, $H\left(\mu_i(\alpha, \beta' x)\right)$ is continuous in $\alpha \in A$, $\beta \in B$ and all x, and H itself is uniformly continuous with respect to the metric (80), $\sup_{0\leq u\leq 1} |H_1(u) - H_2(u)| \leq \|h_1 - h_2\|_1$. Moreover, the penalty term $-\Pi(h)$ in (78) is continuous in h. Therefore, $\Psi(y, \alpha, \beta, h)$ is continuous on $\mathcal{Y} \times A \times B \times \mathcal{D}(0,1)$, where $\mathcal{Y}$ is the Euclidean space with dimension the dimension of $Y_j = \left(D_{1,j}, ..., D_{M,j}, X_j'\right)$. Then $\overline{\Psi}(\alpha, \beta, h)$ is also continuous on $A \times B \times \mathcal{D}(0,1)$.

It is easy to verify from (78) that $\Psi(Y_j, \alpha, \beta, h) \leq 0$, hence condition (87) holds, and condition (86) holds if

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

Thus, under Assumptions 1-5, the conditions of part (a) of Theorem 10 hold, hence (82) 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* $\widetilde{\theta} = \arg\max_{\theta \in \Theta_N} N^{-1} \sum_{j=1}^N g(Y_j, \theta)$ *is feasible.*

If each Θ_N *contains an element* θ_N *such that* $\lim_{N\to\infty} d(\theta_N, \theta_0) = 0$*, then* $p\lim_{N\to\infty} \overline{g}(\widetilde{\theta}) = \overline{g}(\theta_0)$*. Consequently, the results* (a) *and* (b) *in Theorem 10 carry over to* $\widetilde{\theta}$*.*

Proof: Appendix.

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

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*

$$\left(\widetilde{\alpha}, \widetilde{\beta}, \widetilde{h}\right) = \arg \max_{\alpha\in A, \beta\in B, h\in\mathcal{D}_{n_N}(0,1)} \ln\left(L_N^*(\alpha, \beta, h)\right)$$

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

$$p\lim_{N\to\infty} \int_0^1 \left|\widetilde{h}(u) - h_0(u)\right| du = 0.$$

Note that the speed of convergence n_N of $h_N \in \mathcal{D}_{n_N}(0,1)$ to $h_0 \in \mathcal{D}(0,1)$ [see Theorem 9] does not matter for this result. Therefore, as far as consistency is concerned the space $\mathcal{D}_{n_N}(0,1)$ may be selected adaptively, by using for example the well-known Hannan-Quinn (1979) or Schwarz (1978) information criteria.

9 The ICMPH model with finite-valued covariates

As said before, Assumption 2 if 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:* $S = \{x_1, x_2, ..., x_K\}$, $P[X_j \in S] = 1$.

9.1 Lack of identification

In this case the ICMPH model is no longer identified. Instead of a single true parameter vector $\theta_0 = \left(\alpha_0', \beta_0'\right)'$ there now exists a set Θ_0 of "true" parameters (henceforth called *admissible* rather than true), in the sense that for each $\left(\alpha_0', \beta_0'\right)' \in \Theta_0$ there exist uncountable many distribution function H_0 on $[0,1]$ for which the model is correct. I will show this for the ICMPH model (39) in the form

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

where ω_m is column m of the $M \times M$ upper-triangular matrix $\Omega = (\omega_{i_1,i_2})$ with typical element $\omega_{i_1,i_2} = I(i_1 \leq i_2)$, and $\alpha_0 = (\alpha_{0,1}, \alpha_{0,2}, ..., \alpha_{0,M})'$ satisfying $\alpha_{0,m} > 0$ for $m = 2, 3, ..., M$. As part of the model specification I will choose closed hypercubes for the parameter spaces of α and β:

$$A = \times_{i=1}^M \left[\underline{\alpha}_i, \overline{\alpha}_i\right],\ \underline{\alpha}_i > 0 \text{ for } i = 2, ..., M,\ B = \times_{i=1}^k \left[\underline{\beta}_i, \overline{\beta}_i\right], \tag{91}$$

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 (35), except that now $\mu_i(\alpha, \beta' X_j) = \exp(-\exp(\beta' X_j + \alpha' \omega_i))$ for $i = 1, ..., M$.

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

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

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

$$P_0 = \left(p_{i,\ell}^0\,;\ i = 1, ..., M,\ \ell = 1, ..., K\right) \tag{93}$$

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

$$\widehat{p}_{m,\ell} = \frac{\sum_{j=1}^{N} \left(1 - \Sigma_{i=1}^{m} D_{i,j}\right) I\left(X_j = x_\ell\right)}{\sum_{j=1}^{N} I\left(X_j = x_\ell\right)}. \tag{94}$$

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

Assumption 2*. *The probabilities* $p^0_{m,\ell} = P\left[\Sigma_{i=1}^{m} D_{i,j} = 0 \mid X_j = x_\ell\right]$, $m = 1, ..., M$, $\ell = 1, ..., K$, *are all different,*[12]

because then this set is the simplex:

$$\begin{aligned} \Theta_0 &= \cap_{p^0_{m_2,\ell_2} > p^0_{m_1,\ell_1}} \Big\{\theta = (\alpha', \beta')' \in A \times B : \\ &\qquad \beta'\left(x_{\ell_1} - x_{\ell_2}\right) + \alpha'\left(\omega_{m_1} - \omega_{m_2}\right) > 0\Big\}, \end{aligned} \tag{95}$$

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

$$\begin{aligned} p^0_{m,\ell} &= P\left[\Sigma_{i=1}^{m} D_{i,j} = 0 \mid X_j = x_\ell\right] \\ &= H\left(\exp\left(-\exp\left(\beta' x_\ell + \alpha' \omega_m\right)\right)\right). \end{aligned} \tag{96}$$

as well.

On the other hand, the distribution functions H and H_0 can be confined to SNP distribution functions:

Lemma 3. *Given K points $(u_i, v_i) \in (0,1) \times (0,1)$ satisfying $u_1 < u_2 < ... < u_K$, $v_1 < v_2 < ... < v_K$, there exists an SNP distribution function*

[12] Note that this assumption is stronger a condition than Assumption 1(a), as it implies that the elements of S can be ordered such that $\beta_0' x_1 < \beta_0' x_2 < .... < \beta_0' x_K$, 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 $E[X_j | \beta_0' X_j] = E[X_j | X_j] = X_j$. Therefore, Σ_0 in Assumption 1(b) is now a zero matrix!

$H_n(u|\delta) = \int_0^u h_n(v|\delta)dv$, with h_n defined by (18), *such that $H_n(u_i|\delta) = v_i$ for $i = 1, ..., K$. However, δ may not be unique, even for the minimal value of n.*

Proof: Appendix.

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

Neither the quantile conditions (62) nor the moment condition (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 $p_{m,\ell}^0$ using for h a low-order SNP density $h_q(u|\delta)$. Suppose that

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

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

is unique.

If in addition,

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

then it follows similar to Theorem 11 that the QML estimators

$$\left(\widetilde{\alpha}_q, \widetilde{\beta}_q, \widetilde{h}_q\right) = \arg \max_{(\alpha,\beta,h)\in A\times B\times\mathcal{D}_q(0,1)} \ln L_N(\alpha,\beta,h) \tag{98}$$

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

Note that the QML estimation problem (98) is fully parametric. Denoting

$$\Delta_q = \left\{\delta = (\delta_1, ..., \delta_q)' \text{ s.t. } (84)\right\},$$

(98) is equivalent to

$$\left(\widetilde{\alpha}_q, \widetilde{\beta}_q, \widetilde{\delta}_q\right) = \arg \max_{(\alpha,\beta,\delta)\in A\times B\times\Delta_q} \ln\left(L_N\left(\alpha,\beta,h_q(.|\delta)\right)\right)$$

and (97) is equivalent to

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

Therefore, the standard QML asymptotics applies [c.f. White (1982)]. In particular, if

Assumption 5*. $\left(\alpha_q^{'}, \beta_q^{'}, \delta_q^{\prime}\right)^{\prime}$ *is an interior point of* $\Theta_0 \times \Delta_q$,

then

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

in distribution, where

$$\begin{aligned} \Omega_1 &= \mathrm{Var}\left(\frac{\partial \ln L_N\left(\alpha_q,\beta_q,h_q(.|\delta_q)\right)/\sqrt{N}}{\partial(\alpha_q^{\prime},\beta_q^{\prime},\delta_q^{\prime})}\right), \\ \Omega_2 &= -E\left[\frac{\partial^2 \ln L_N\left(\alpha_q,\beta_q,h_q(.|\delta_q)\right)/N}{\partial(\alpha_q^{\prime},\beta_q^{\prime},\delta_q^{\prime})^{\prime}\partial(\alpha_q^{\prime},\beta_q^{\prime},\delta_q^{\prime})}\right]. \end{aligned}$$

Next, suppose that:

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

where H_q is the c.d.f. of h_q. If so, then the ranking of the QML estimates

$$\widetilde{p}_{m,\ell}(q) = \widetilde{H}_q\left(\exp\left(-\exp\left(\widetilde{\beta}_q^{\prime}x_\ell + \widetilde{\alpha}_q^{\prime}\omega_m\right)\right)\right) \tag{99}$$

of the $p_{m,\ell}(q)$'s will be equal to the ranking of the $p_{m,\ell}^0$'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

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

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

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

with corresponding QML estimators $\widehat{\alpha}=\widetilde{\alpha}_q$, $\widehat{\beta}=\widetilde{\beta}_q$, satisfying

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

(due to Assumption 5*), and

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

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 follows now from Lemma 3 that there exists a smallest polynomial order n_0 and a $h_{n_0}\in\mathcal{D}_{n_0}(0,1)$ with c.d.f. H_{n_0} such that

$$\begin{aligned}P\left[\Sigma_{i=1}^m D_{i,j}=0\mid X_j=x_\ell\right]&=H_{n_0}\left(\exp\left(-\exp\left(\beta_0'x_\ell+\alpha_0'\omega_m\right)\right)\right),\\ m&=1,...,M,\ \ \ell=1,...,K.\end{aligned} \tag{103}$$

I will show that n_0 can be estimated consistently via a new information criterion:

$$C_N(n)=\frac{-1}{N}\sup_{h\in\mathcal{D}_n(0,1)}\ln\left(L_N\left(\widehat{\alpha},\widehat{\beta},h\right)\right)+(M+k+n)\,.\frac{\varphi(N)}{N}, \tag{104}$$

where $\lim_{N\to\infty}\varphi(N)/N=0$ and $\lim_{N\to\infty}\varphi(N)/\sqrt{N}=\infty$.

Recall that the Hannan-Quinn (1979) criterion corresponds to the case $\varphi(N) = \ln(\ln(N))$ and the Schwarz (1978) criterion to the case $\varphi(N) = \ln(N)/2$, so that these criteria are not suitable for our purpose.

Theorem 13. *Given the 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*

$$p\lim_{N\to\infty} C_N(n+1) < p\lim_{N\to\infty} C_N(n) \quad if \quad n < n_0, \tag{105}$$

$$p\lim_{N\to\infty} \frac{N}{\varphi(N)} \left(C_N(n) - C_N(n-1)\right) = 1 \quad if \quad n > n_0. \tag{106}$$

Proof: Appendix.

Consequently, under the conditions of Theorem 13,

$$\widehat{n} = \max_{s.t.\ C_N(n)<C_N(n-1),\ n\geq 2} n,$$

is a consistent estimator of n_0:

$$\lim_{N\to\infty} P\left[\widehat{n} = n_0\right] = 1. \tag{107}$$

9.4 Consistency of the SNP probability estimators

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

$$\widehat{p}^*_{m,\ell} = \widehat{H}\left(\exp\left(-\exp\left(\widehat{\beta}^{'} x_\ell + \widehat{\alpha}'\omega_m\right)\right)\right) \tag{108}$$

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

Theorem 14. *Under the conditions of Theorem 13 the SNP probability estimators* (108) *are consistent:* $p\lim_{N\to\infty} \widehat{p}^*_{m,\ell} = P\left[\Sigma_{i=1}^m D_{i,j} = 0 \mid X_j = x_\ell\right].$

Proof: 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 $\widehat{p}^*_{m,\ell}$'s matches the ranking of the estimates $\widetilde{p}_{m,\ell}(q)$ 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 (39), the results in this paper are straightforwardly applicable to more general ordered probability models as well, simply by replacing (40) 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 (39) becomes a semi-nonparametric 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 $G(x) = 1 - \exp(-\exp(x))$.

The idea to model densities semi-nonparametrically via semi-nonparametric densities on the unit interval is the most straightforward part of this paper. The main contributions in this paper are (1) the application of this idea to interval-censored mixed proportional hazard models; (2) the derivation of the conditions under which these models are nonparametrically identified; (3) the construction of the compact metric space of densities on the unit interval; (4) the weak consistency results for sieve M-estimators under weak and verifiable conditions, even under partial identification, and (5) 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 semi-nonparametric parameter estimators [see for example Chen (2005) and the references therein], this literature is not directly applicable to the cases considered in this paper. The reason is that the semi-nonparametric argument in the log-likelihood of the models considered in this paper involve distribution functions $H(u)$ on the unit interval only, and that the L^1 metric $\|h_1 - h_2\|_1 = \int_0^1 |h_1(u) - h_2(u)|\, du$ on the space $\mathcal{D}(0,1)$ of density functions $h(u)$ on $[0,1]$ implies the sup metric on the corresponding space of distribution function $H(u)$ on $[0,1]$. The latter metric plays a crucial role in proving consistency in the case of continuously distributed covariates. For deriving asymptotic normality, however, we need to work with the scores of the log-likelihood, which involved ratios of densities and distribution functions on the unit interval and their derivatives. In view of the asymptotic normality conditions in Chen (2005, Section 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 mixed proportional hazard 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. Also this problem will be left for future research.

11 References

Bierens, H. J. (1982), "Consistent Model Specification Tests", *Journal of Econometrics,* 20, 105-134.

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

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

Bierens, H. J. and J. R. Carvalho (2006), "Job Search, Conditional Treatment and Recidivism: The Employment Services for Ex-Offenders Program", working paper (http://econ.la.psu.edu/~hbierens/PAPERS.HTM).

Chen, X. (2005), "Large Sample Sieve Estimation of Semi-nonparametric Models", *Handbook of Econometrics, Vol. 6* (forthcoming).

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

Elbers, C., and 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., and D. W. Nychka (1987), "Semi-Nonparametric Maximum Likelihood Estimation", *Econometrica,* 55, 363-390.

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

Hannan, E. J., and B. G. Quinn (1979), "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society, B*, 41, 190-195.

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

Heckman, J. J., and 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., and J. Wolfowitz (1956), "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters", *The Annals of Mathematical Statistic*, 27, 887-906.
Lancaster, T. (1979), "Econometric Methods for the Duration of Unemployment", *Econometrica*, 47, 939-956.
Meyer, B. D. (1995), "Semiparametric Estimation of Hazard Models", unpublished paper (http://www.faculty.econ.northwestern.edu/faculty/meyer/SEMIR2.pdf).
Royden, H. L. (1968), *Real Analysis*, London: Macmillan.
Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461-464.
Van den Berg, G., J. (2000), "Duration Models: Specification, Identification, and Multiple Durations", *Handbook of Econometrics, Vol. 5.*
White, H. (1982), "Maximum Likelihood Estimation of MisspecifiedModels", *Econometrica*, 50, 1-25.
Young, N. (1988), *An Introduction to Hilbert Space*, Cambridge, UK: Cambridge University Press.

12 Appendix

12.1 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*

$$\xi_0 > 1, \ \sum_{k=1}^{\infty} \xi_k^2 < \infty, \tag{109}$$

and let $\mathcal{F}_\xi$ be the set of functions $f(u) = \sum_{k=0}^{\infty} \gamma_k \rho_k(u)$ in $L^2_{\mathcal{B}}(0,1)$ for which $\gamma_k \in [-\xi_k, \xi_k]$, $k = 0, 1, 2, ...$, endowed with the metric (9). Then $\mathcal{F}_\xi$ is compact.

Proof: It suffices to prove that $\mathcal{F}_\xi$ is complete and totally bounded. See Royden (1968, Proposition 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 $\mathcal{F}_\xi$. Since $f_n(u)$ is a Cauchy sequence in the Hilbert space $L^2_{\mathcal{B}}(0,1)$ it converges to a function $f(u) = \sum_{k=0}^{\infty} \gamma_k\rho_k(u)$ in $L^2_{\mathcal{B}}(0,1)$. 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, $\gamma_{n,k} \to \gamma_k$. Since $\gamma_{n,k} \in [-\xi_k, \xi_k]$ 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. Since $\sum_{k=0}^{m} \gamma_{n,k}^2 = 1 - \sum_{k=m+1}^{\infty} \gamma_{n,k}^2 \geq 1 - \sum_{k=m+1}^{\infty} \xi_k^2$ we can choose m so large that uniformly in n, $0 \leq 1 - \sum_{k=0}^{m} \gamma_{n,k}^2 < \varepsilon$. Since for $k = 0, 1, ..., m$, $\gamma_{n,k} \to \gamma_k$, 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 $\gamma_k \in [-\xi_k, \xi_k]$, $k = 0, 1, 2, ..., n$. 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 $\gamma = (\gamma_0, \gamma_1, ..., \gamma_n)'$ satisfying $\gamma \in \times_{k=0}^{n} [-\xi_k, \xi_k]$, $\gamma'\gamma \leq 1$ is a closed an 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 $f_1, ..., f_M \in \mathcal{F}_{\xi,n}$ such that

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

This implies that

$$\begin{aligned} \mathcal{F}_\xi &\subset \cup_{j=1}^{M} \left\{ f \in \mathcal{F}_\xi(0,1) : \|f - f_j\|_2 < 2\varepsilon \right\} \\ &\subset \cup_{j=1}^{M} \left\{ f \in L^2_{\mathcal{B}}(0,1) : \|f - f_j\|_2 < 2\varepsilon \right\}, \end{aligned}$$

hence $\mathcal{F}_\xi$ is totally bounded. Q.E.D.

Lemma A.2. *Under condition* (109) *the space*

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

endowed with the metric (9) *is compact.*

Proof: It follows from (16) and (109) that

$$\begin{aligned}
\gamma_k^2 &= \frac{\delta_k^2}{1+\sum_{k=1}^{\infty}\delta_k^2} \leq \xi_k^2,\ k\geq 1, \\
\gamma_0^2 &= \frac{1}{1+\sum_{k=1}^{\infty}\delta_k^2} \leq 1 < \xi_0^2 \\
\gamma_0^2 &\geq \frac{1}{1+\sum_{k=1}^{\infty}\xi_k^2} > 0,
\end{aligned} \tag{110}$$

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

For a metric space the notions of compactness and sequential compactness are equivalent. See Royden (1968, Corollary 14, p. 163). Sequential compactness means that any infinite sequence in the metric space has a convergent subsequence which 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

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

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

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

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. Q.E.D.

Lemma A.3. *The space* $\mathcal{D}_\xi(0,1) = \left\{h : h = f^2,\ f \in \mathcal{F}_\xi^*\right\}$ *of density functions on* $[0,1]$ *endowed with the metric* (80) *is compact.*

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

$$\begin{aligned} &\int_0^1 \left|f(u)^2 - g(u)^2\right| du \qquad (111) \\ &\leq \int_0^1 |f(u) - g(u)|\,|f(u)|\,du + \int_0^1 |f(u) - g(u)|\,|g(u)|\,du \\ &\leq \sqrt{\int_0^1 (f(u) - g(u))^2\,du}\left(\sqrt{\int_0^1 f(u)^2 du} + \sqrt{\int_0^1 g(u)^2 du}\right) \\ &= 2\sqrt{\int_0^1 (f(u) - g(u))^2\,du}. \end{aligned}$$

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} which converges to a limit f in $\mathcal{F}_\xi^*$, hence it follows from (111) 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. Q.E.D.

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.

12.2 Proof of Theorem 10

It follows now from Jennrich's (1969) uniform strong law of large numbers, in the version in Bierens (1994, Section 2.7) or Bierens (2004, Appendix to Chapter 6) that under the conditions of Theorem 10, with the conditions (86) and (87) replaced by

$$E\left[\sup_{\theta\in\Theta} |g(Y_1, \theta)|\right] < \infty \qquad (112)$$

we have

$$\lim_{N\to\infty} \sup_{\theta\in\Theta} \left|\frac{1}{N}\sum_{j=1}^N g(Y_j, \theta) - \overline{g}(\theta)\right| = 0 \ a.s. \qquad (113)$$

However, the condition (112) is too difficult to verify in the log-likelihood case. Therefore I will use the weaker conditions (86) and (87).

Originally the uniform strong law (113) 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, Section 2.7) or Bierens (2004, Appendix to Chapter 6) that this law carries over to random functions on compact metric spaces.

Let $K > K_0$ and note that

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

hence $E\left[\sup_{\theta\in\Theta}\left|\max\left(g(Y_1,\theta),-K\right)\right|\right] < \infty$. Then it follows from (113) with $g(Y_j,\theta)$ replaced by $\max\left(g(Y_j,\theta),-K\right)$ that

$$\lim_{N\to\infty}\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{j=1}^{N}\max\left(g(Y_j,\theta),-K\right) - \overline{g}_K(\theta)\right| = 0 \ a.s., \tag{114}$$

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

As is well-known, (114) is equivalent to the statement that for all $\varepsilon > 0$,

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

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 $N \geq N(K,k,m)$,

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

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

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

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

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

hence, for arbitrary $\varepsilon>0$,

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

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

$$\sup_{\theta\in\Theta}\left|\frac{1}{n_N}\sum_{j=1}^{n_N}\max\left(g(Y_j,\theta),-K_N\right)-\overline{g}_{K_N}(\theta)\right|\to 0 \text{ a.s.,} \tag{115}$$

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 (115) 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, (115) implies that there exists a sequence K_N converging to infinity with N such that

$$p\lim_{N\to\infty}\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{j=1}^{N}\max\left(g(Y_j,\theta),-K_N\right)-\overline{g}_{K_N}(\theta)\right|=0. \tag{116}$$

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

$$\begin{aligned}\overline{g}_K(\theta) &= E\left[\max\left(g(Y_j,\theta),-K\right)\right]\geq\max\left(E\left[g(Y_j,\theta)\right],-K\right) \\ &= \max\left(\overline{g}(\theta),-K\right)\geq\overline{g}(\theta)\end{aligned} \tag{117}$$

and similarly

$$\frac{1}{N}\sum_{j=1}^{N}\max\left(g(Y_j,\theta),-K\right)\geq\max\left(\frac{1}{N}\sum_{j=1}^{N}g(Y_j,\theta),-K\right)\geq\frac{1}{N}\sum_{j=1}^{N}g(Y_j,\theta). \tag{118}$$

It follows from (87), (117) and the dominated convergence theorem that $\lim_{K\to\infty}\sup_{\theta\in\Theta}|\overline{g}_K(\theta)-\overline{g}(\theta)|=0$, hence (116) now becomes

$$p\lim_{N\to\infty}\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{j=1}^{N}\max\left(g(Y_j,\theta),-K_N\right)-\overline{g}(\theta)\right|=0. \tag{119}$$

Finally, observe from (118) that

$$\frac{1}{N}\sum_{j=1}^{N}\max\left(g(Y_j,\widehat{\theta}),-K_N\right)-\overline{g}(\widehat{\theta})\geq\frac{1}{N}\sum_{j=1}^{N}g(Y_j,\widehat{\theta})-\overline{g}(\widehat{\theta}) \tag{120}$$
$$\geq\frac{1}{N}\sum_{j=1}^{N}g(Y_j,\theta_0)-\overline{g}(\theta_0)+\overline{g}(\theta_0)-\overline{g}(\widehat{\theta})\geq\frac{1}{N}\sum_{j=1}^{N}g(Y_j,\theta_0)-\overline{g}(\theta_0).$$

By Kolmogorov's strong law of large numbers, the lower bound in (120) converges a.s. to zero, and by (119) the upper bound in (120) converges in probability to zero, hence $p\lim_{N\to\infty}\left(\frac{1}{N}\sum_{j=1}^{N}g(Y_j,\widehat{\theta})-\overline{g}(\widehat{\theta})\right)=0$, which implies (88).

(a) If θ_0 is unique then by the continuity of $\overline{g}(\theta)$ and the compactness of Θ there exists a $\overline{\varepsilon}>0$ such that for all $\varepsilon\in(0,\overline{\varepsilon}]$, $\sup_{\theta\in\Theta,d(\theta,\theta_0)\geq\varepsilon}\overline{g}(\theta)<\overline{g}(\theta_0)$. See, for example, Bierens (2004, Appendix II, Theorem II.6). It follows therefore from (88) that

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

(b) It follows from (88) that $p\lim_{N\to\infty}\overline{g}\left(\left(\widehat{\theta}_1',\widehat{\theta}_2'\right)'\right)=\overline{g}\left(\left(\theta_{0,1}',\theta_{0,2}'\right)'\right)$. Recall that convergence in probability is equivalent to a.s. convergence along a further subsequence of an arbitrary subsequence. Thus, every subsequence of N contains a further subsequence N_m with corresponding estimators $\left(\widehat{\theta}_{1,N_m}',\widehat{\theta}_{2,N_m}'\right)'$ such that, for $m\to\infty$,

$$\overline{g}\left(\left(\widehat{\theta}_{1,N_m}',\widehat{\theta}_{2,N_m}'\right)'\right)\to\overline{g}\left(\left(\theta_{0,1}',\theta_{0,2}'\right)'\right)\text{ a.s.}$$

In other words, for all ω in a set with probability 1,

$$\lim_{m\to\infty}\overline{g}\left(\left(\widehat{\theta}_{1,N_m}(\omega)',\widehat{\theta}_{2,N_m}(\omega)'\right)'\right)=\overline{g}\left(\left(\theta_{0,1}',\theta_{0,2}'\right)'\right) \tag{121}$$

Since Θ_1 and Θ_2 are compact, the sequences $\widehat{\theta}_{1,N_m}(\omega)$ and $\widehat{\theta}_{2,N_m}(\omega)$ have limit points $\theta_1(\omega)$ and $\theta_2(\omega)$ in Θ_1 and Θ_2, respectively, hence

$$\overline{g}\left(\left(\theta_1(\omega)', \theta_2(\omega)'\right)'\right) = \overline{g}\left(\left(\theta'_{0,1}, \theta'_{0,2}\right)'\right)$$

and thus $\sup_{\theta_2 \in \Theta_2} \overline{g}\left(\left(\theta_1(\omega)', \theta'_2\right)'\right) \geq \overline{g}\left(\theta'_{0,1}, \theta'_{0,2}\right)$. It follows now from (89) that $\theta_1(\omega) \equiv \theta_{0,1}$, hence $d_1\left(\widehat{\theta}_{1,N_m}, \theta_{0,1}\right) \to \infty$ a.s., and consequently,

$$p \lim_{N\to\infty} d_1\left(\widetilde{\theta}_{1,N}, \theta_{0,1}\right).$$

Finally, recall that a continuous function on a compact set is uniformly continuous on that set. It follows now from (121) and the uniform continuity of $\overline{g}$ that $\lim_{m\to\infty} \overline{g}\left(\left(\theta'_{0,1}, \widehat{\theta}_{2,N_m}(\omega)'\right)'\right) = \overline{g}\left(\left(\theta'_{0,1}, \theta'_{0,2}\right)'\right)$, hence

$$p \lim_{N\to\infty} \overline{g}\left(\left(\theta'_{0,1}, \widehat{\theta}'_2\right)'\right) = \overline{g}\left(\left(\theta'_{0,1}, \theta'_{0,2}\right)'\right).$$

12.3 Proof of Theorem 11

Similar to (120) we have,

$$\begin{aligned}
&\frac{1}{N}\sum_{j=1}^{N} \max\left(g(Y_j, \widetilde{\theta}), -K_N\right) - \overline{g}(\widetilde{\theta}) \geq \frac{1}{N}\sum_{j=1}^{N} g(Y_j, \theta_N) - \overline{g}(\widetilde{\theta}) \\
&\geq \frac{1}{N}\sum_{j=1}^{N} g(Y_j, \theta_N) - \overline{g}(\theta_0) + \overline{g}(\theta_0) - \overline{g}(\widetilde{\theta}) \\
&\geq \frac{1}{N}\sum_{j=1}^{N} \left(g(Y_j, \theta_N) - g(Y_j, \theta_0)\right) + \frac{1}{N}\sum_{j=1}^{N} g(Y_j, \theta_0) - \overline{g}(\theta_0)
\end{aligned}$$

It follows from (119) that

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

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

$$\frac{1}{N}\sum_{j=1}^{N} g(Y_j,\theta_0) \to \overline{g}(\theta_0) \text{ a.s.} \tag{122}$$

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

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

Hence by Chebishev's inequality,

$$p\lim_{N\to\infty}\frac{1}{N}\sum_{j=1}^{N}\left(g(Y_j,\theta_N) - g(Y_j,\theta_0)\right) = 0. \tag{123}$$

Thus, $p\lim_{N\to\infty}\left(\frac{1}{N}\sum_{j=1}^{N} g(Y_j,\theta_N) - \overline{g}(\widetilde{\theta})\right) = 0$, which by (122) and (123) implies that $p\lim_{N\to\infty}\overline{g}(\widetilde{\theta}) = \overline{g}(\theta_0)$.

12.4 Proof of Lemma 3

12.4.1 The case $K = 1$

Let $\overline{H}(u)$ be a distribution function on $[0,1]$ such that $\overline{H}(u_1) > v_1$, 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 exists an n and a pair $\overline{\delta},\underline{\delta} \in \mathbb{R}^n$ such that $H_n(u_1|\overline{\delta}) > v_1$ and $H_n(u_1|\underline{\delta}) < v_1$. Since $\varphi_n(\lambda) = H_n(u_1|(1-\lambda)\underline{\delta} + \lambda\overline{\delta}) - v_1$ 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 $\delta = (1-\lambda)\underline{\delta} + \lambda\overline{\delta}$. Then $H_n(u_1|\delta) = v_1$.

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

12.4.2 The case $K = 2$

Again, it follows straightforwardly from Theorem 3 that there exists an n and a pair $\overline{\delta},\underline{\delta} \in \mathbb{R}^n$ such that

$$H_n(u_1|\underline{\delta}) < v_1, \; H_n(u_1|\overline{\delta}) > v_1,$$

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

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

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

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

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

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

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

for instance. If $\eta > v_1$ then

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

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

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

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

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

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

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

12.4.3 The case $K > 2$

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

$$\begin{aligned}
&\Phi_n^{(m)}(\lambda|M) = H_n(u_m|(1-\lambda)\underline{\delta} + \lambda\overline{\delta}) - v_m \\
&+ \sum_{i=1}^{m-1} \left(\left(\frac{H_n(u_i|(1-\lambda)\underline{\delta} + \lambda\overline{\delta})}{v_i} \right)^M + \left(\frac{v_i}{H_n(u_i|(1-\lambda)\underline{\delta} + \lambda\overline{\delta})} \right)^M \right) \\
&\times \left(H_n(u_i|(1-\lambda)\underline{\delta} + \lambda\overline{\delta}) - v_i \right)^2,
\end{aligned}$$

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

$$\begin{aligned}
\Phi_n^{(m)}(0|M) &= H_n(u_m|\underline{\delta}) - v_m < 0, \\
\Phi_n^{(m)}(1|M) &= H_n(u_m|\overline{\delta}) - v_m > 0,
\end{aligned}$$

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, $H_n(u_i|(1-\lambda_*)\underline{\delta} + \lambda_*\overline{\delta}) = v_i$, $i = 1, .., m$, hence $\cap_{i=1}^{m} \Delta_{i,n} \neq \emptyset$. The general result follows now by induction.

Finally, to show that δ may not be unique, consider the case $K = 1$, with $u_1 = 1/2$ and $v_1 < 1/2$. It is easy to verify that

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

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

12.5 Proof of Theorem 13

Part (105) of Theorem 13 follows from the inequality

$$p \lim_{N\to\infty} \frac{1}{N} \sup_{h \in \mathcal{D}_n(0,1)} \ln\left(L_N\left(\widehat{\alpha}, \widehat{\beta}, h\right)\right) < p \lim_{N\to\infty} \frac{1}{N} \sup_{h \in \mathcal{D}_{n+1}(0,1)} \ln\left(L_N\left(\widehat{\alpha}, \widehat{\beta}, h\right)\right)$$

for $n < n_0$, which in its turn follows trivially from Theorem 10. Part (106) is true if for $n > n_0$,

$$\sup_{h \in \mathcal{D}_n(0,1)} \ln\left(L_N\left(\widehat{\alpha}, \widehat{\beta}, h\right)\right) - \sup_{h \in \mathcal{D}_{n_0}(0,1)} \ln\left(L_N\left(\widehat{\alpha}, \widehat{\beta}, h\right)\right) = O_p(\sqrt{N}). \quad (124)$$

To show that (124) holds, observe first that for all n,

$$\sup_{h\in\mathcal{D}_n(0,1)} \ln\left(L_N\left(\widehat{\alpha},\widehat{\beta},h\right)\right) \leq \ln\left(L_N\left(\widehat{P}\right)\right),$$

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

$$2\left(\ln\left(L_N\left(\widehat{P}\right)\right) - \ln\left(L_N\left(P_0\right)\right)\right) \to \chi^2_{M.K}$$

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

$$\sup_{h\in\mathcal{D}_n(0,1)} \ln\left(L_N\left(\widehat{\alpha},\widehat{\beta},h\right)\right) \leq \ln\left(L_N\left(P_0\right)\right) + O_p(1). \tag{125}$$

Furthermore, it is trivial that.

$$\sup_{h\in\mathcal{D}_{n_0}(0,1)} \ln\left(L_N\left(\widehat{\alpha},\widehat{\beta},h\right)\right) \geq \ln\left(L_N\left(\widehat{\alpha},\widehat{\beta},h_{n_0}\right)\right). \tag{126}$$

Finally, it follows straightforwardly from (102), the mean value theorem, and the easy equality

$$\ln\left(L_N\left(\alpha_0,\beta_0,h_{n_0}\right)\right) = \ln\left(L_N\left(P_0\right)\right) \tag{127}$$

that

$$\ln\left(L_N\left(\widehat{\alpha},\widehat{\beta},h_{n_0}\right)\right) = \ln\left(L_N\left(P_0\right)\right) + O_p\left(\sqrt{N}\right). \tag{128}$$

Combining (125), (126) and (126) for $n > n_0$, (124) follows.

12.6 Proof of Theorem 14

Let $\overline{\Psi}(\alpha,\beta,h) = E\left[N^{-1}\ln\left(L_N\left(\alpha,\beta,h\right)\right)\right]$ and let $\widehat{h}$ be the density of $\widehat{H}$. It follows from Theorem 10 and (107) that

$$p\lim_{N\to\infty} \overline{\Psi}(\widehat{\alpha},\widehat{\beta},\widehat{h}) = \sup_{h\in\mathcal{D}_{n_0}(0,1)} \overline{\Psi}(\alpha_0,\beta_0,h) \tag{129}$$

Next, denote $\widehat{P}_* = \left(\widehat{p}^*_{m,\ell}\ ;\ m = 1, ..., M,\ \ \ell = 1, ..., K\right)$ and let

$$\begin{aligned} g(P) &= E\left[N^{-1}\ln\left(L_N\left(P\right)\right)\right] = \sum_{\ell=1}^{K}\left(1 - p^0_{1,\ell}\right)\ln\left(1 - p_{1,\ell}\right) \\ &= \sum_{i=2}^{M}\sum_{\ell=1}^{K}\left(p^0_{i-1,\ell} - p^0_{i,\ell}\right)\ln\left(p_{i-1,\ell} - p_{i,\ell}\right) + \sum_{\ell=1}^{K} p^0_{M,\ell}\ln\left(p_{M,\ell}\right). \end{aligned}$$

C.f. (92). Then $\overline{\Psi}(\widehat{\alpha}, \widehat{\beta}, \widehat{h}) = g(\widehat{P}_*)$ and $\sup_{h\in\mathcal{D}_{n_0}(0,1)}\overline{\Psi}(\alpha_0, \beta_0, h) = g\left(P_0\right).$ See (127) for the latter. Hence, it follows from (129) that $p\lim_{N\to\infty} g(\widehat{P}_*) = g(P_0)$, which by the continuity of $g(P)$ in the elements of P implies that $p\lim_{N\to\infty}\widehat{P}_* = P_0$.