

Semi-Nonparametric Modeling of Densities on the Unit Interval, with Applications to Censored Mixed Proportional Hazard Models and Ordered Probability Models*

Herman J. Bierens[†]

January 30, 2005

Abstract

In this paper I propose to estimate densities with possibly restricted support semi-nonparametrically via semi-nonparametric (SNP) estimation of densities on the unit interval. The latter will be done similarly to the approach of Gallant and Nychka (1987), but instead of using Hermite polynomials I propose to use orthonormal Legendre polynomials on the unit interval. This approach will be applied to the mixed proportional hazard (MPH) model, where the duration involved is only observed within brackets, and the distribution of the unobserved heterogeneity is modeled semi-nonparametrically. Under mild conditions the MPH model is nonparametrically identified. It appears that the identification conditions involved also apply to generalized ordered probability models. I will set forth conditions such that for both types of models the SNP maximum likelihood estimators are consistent.

*This is a preliminary and incomplete paper. Please do not quote.

[†]I am grateful to Aleksandr Vashchilko for pointing out an error in a previous version of this paper.

1 Introduction

In this paper it will be shown that any density h on the unit interval can be written as $h(u) = f(u)^2$, where f is Borel measurable and has an infinite series expansion in terms of orthonormal Legendre polynomials. This approach can be used to approximate more general densities as well: 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 and its density g is to fix the support of f .

This approach will be applied to the mixed proportional hazard (MPH) model, where the duration is heavily censored in that it is only observed within brackets. 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. The density h of this distribution function H will be modeled semi-nonparametrically. I will set forth conditions under which the parameters of the systematic and baseline hazards are nonparametrically identified, and the maximum likelihood estimators involved are consistent.

In Section 2 the Legendre polynomials will be introduced, 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 MPH model and its nonparametric identification under bracketing. It appears that the identification conditions involved also apply to generalized ordered probability models. See Section 5. In Section 6 I will sketch the requirements for consistency of the SNP maximum likelihood 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. In Section 8 I will prove general consistency results for M estimators of (partly) non-Euclidean parameters, and specialize these results to MPH models and ordered probability models. Finally, in Section 9 I will briefly discuss what can be done if the covariates have finite support.

2 Orthonormal polynomials

2.1 Hermite polynomials

Gallant and Nychka (1987) consider SNP estimation of Heckman's 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 $f(x) = \phi(x) (\sum_{k=0}^{\infty} \gamma_k \mathcal{H}_k(x))^2$, with $\sum_{k=0}^{\infty} \gamma_k^2 = 1$, where $\phi(x)$ is the standard normal density and the $\mathcal{H}_k(x)$'s are orthonormal Hermite polynomials¹. These densities can be approximated arbitrarily close by semi-nonparametric (SNP) densities of the type $f_n(x) = \phi(x) (\sum_{k=0}^n \gamma_{k,n} \mathcal{H}_k(x))^2$, where $\sum_{k=0}^n \gamma_{k,n}^2 = 1$. In principle we can transform the densities $f(x)$ and $f_n(x)$ involved to densities on the unit interval. For example, let for $u \in [0, 1]$, $h(u) = (u(1-u))^{-1} f(\ln(u/(1-u)))$ and $h_n(u) = (u(1-u))^{-1} f_n(\ln(u/(1-u)))$. However, the problem is that in this case the uniform density $h(u) = 1$ can only be represented as a limit of $h_n(u)$. The uniform case is of interest because if $G(x)$ is an initial guess of the unknown distribution $F(x)$ then the guess is right if $h(u) = 1$. In particular, in the MPH case the uniform distribution corresponds to absence of unobserved heterogeneity. Therefore, I will use a different approach based on orthonormal Legendre polynomials on the unit interval.

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

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

starting from

$$P_0(z) = 1, \quad P_1(z) = z.$$

¹Hermitian polynomials $P_n(x)$, $x \in \mathbb{R}$, can be generated recursively by $P_{n+1}(x) = 2xP_n(x) - 2nP_{n-2}(x)$ for $n \geq 2$, starting from $P_0(x) = 1$, $P_1(x) = 2x$, and satisfy $\int_{-\infty}^{\infty} \exp(-x^2) P_n(x)P_m(x)dx = I(n=m) 2^n n! \sqrt{\pi}$, where $I(\cdot)$ is the indicator function. See, e.g., Hamming (1973). Denoting $\mathcal{H}_n(x) = P_n(x/\sqrt{2})/\sqrt{2^n n!}$, it follows that $\int_{-\infty}^{\infty} \phi(x)\mathcal{H}_n(x)\mathcal{H}_m(x)dx = I(n=m)$.

They are orthogonal, but not orthonormal:

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

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

Now define for $u \in [0, 1]$,

$$\rho_n(u) = \sqrt{2n+1}P_n(2u-1). \quad (3)$$

Then it follows from (2) that the polynomials $\rho_n(u)$ are orthonormal:

$$\int_0^1 \rho_k(u)\rho_m(u)du = \begin{cases} 0 & \text{if } k \neq m, \\ 1 & \text{if } k = m, \end{cases} \quad (4)$$

and from (1) that for $n \geq 2$ they can be computed recursively by

$$\rho_n(u) = \frac{\sqrt{2n-1}\sqrt{2n+1}}{n}(2u-1)\rho_{n-1}(u) - \frac{(n-1)\sqrt{2n+1}}{n\sqrt{2n-3}}\rho_{n-2}(u), \quad (5)$$

starting from

$$\rho_0(u) = 1, \quad \rho_1(u) = \sqrt{3}(2u-1). \quad (6)$$

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 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 $\int_0^1 f(u)^2 du < \infty$,² and let $\gamma_k = \int_0^1 \rho_k(u)f(u)du$. Then $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$, and the*

²Note that this integral is the Lebegue integral.

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

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

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

$$\int_0^1 |f_1(u)| du < \infty, \int_0^1 |f_2(u)| du < \infty. \quad (7)$$

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

$$\int_0^1 u^k f_1(u) du = \int_0^1 u^k f_2(u) du. \quad (8)$$

Now let in Lemma 1, $f_1(u) = f(u)$ and $f_2(u) = \sum_{k=0}^{\infty} \gamma_k \rho_k(u)$, where $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$. Then 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 $\int_0^1 |f_1(u)| du < \infty$.

Each u^k can be written as a linear combination of $\rho_0(u), \rho_1(u), \dots, \rho_k(u)$ with Fourier coefficients $\int_0^1 u^k \rho_m(u) du$, $m = 0, 1, \dots, k$, hence condition (8) in Lemma 1 is equivalent to

$$\int_0^1 \rho_k(u) f_1(u) du = \int_0^1 \rho_k(u) f_2(u) du$$

for $k = 0, 1, 2, \dots$. The latter integral is equal to γ_k . Therefore, choose $\gamma_k = \int_0^1 \rho_k(u) f(u) du$. The condition $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$ follows from the fact that

$\int_0^1 (f(u) - \sum_{k=0}^n \gamma_k \rho_k(u))^2 du$ is minimal for $\gamma_k = \int_0^1 \rho_k(u) f(u) du$, so that for all natural numbers n , $\sum_{k=0}^n \gamma_k^2 \leq \int_0^1 f(u)^2 du < \infty$. Therefore, the conditions of Lemma 1 are satisfied.

Recall³ 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

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

and norm $\|f\|_2$. Moreover, recall that Borel measurable functions are Lebesgue measurable because Borel sets are Lebesgue measurable sets.⁴ Therefore, the subspace $L_B^2(0, 1)$ of Borel measurable real functions in $L^2(0, 1)$ is a Hilbert space itself.

Every density function $h(u)$ on $[0, 1]$ is Borel measurable because, with H the corresponding distribution function,

$$h(u) = \lim_{k \rightarrow \infty} k (H(u + k^{-1}) - H(u)),$$

which is a pointwise limit of a sequence of continuous (hence Borel measurable) functions and therefore Borel measurable itself. Consequently, every density function h on $[0, 1]$ can be written as $h(u) = f(u)^2$, where $f(u)$ a Borel measurable real function on $[0, 1]$ satisfying $\int_0^1 f(u)^2 du = 1$.

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

$$\phi_B(u) = I(u \in B) - I(u \in \tilde{B}), \quad (10)$$

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

$$f(u) = \phi_B(u) \sqrt{h(u)},$$

³See for example Young (1988, pp. 24-25).

⁴See for example Royden (1968, pp. 59 and 66)

with $\phi_B(u)$ a simple function of the type (10). Hence, any density $h(u)$ on $[0, 1]$ can be represented by

$$h(u) = \left(\sum_{k=0}^{\infty} \gamma_k \rho_k(u) \right)^2, \text{ with } \sum_{k=0}^{\infty} \gamma_k^2 = 1. \quad (11)$$

Since by Theorem 1 $f(u)$ can be written as $f(u) = \sum_{k=0}^{\infty} \gamma_k \rho_k(u)$, where $\gamma_k = \int_0^1 \rho_k(u) f(u) du$, we can always choose B is such that

$$\begin{aligned} \gamma_0 &= \int_0^1 f(u) du = \int_0^1 \phi_B(u) \sqrt{h(u)} du \\ &= \int_B \sqrt{h(u)} du - \int_{\bar{B}} \sqrt{h(u)} du > 0. \end{aligned} \quad (12)$$

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:

$$\begin{aligned} \gamma_k &= \frac{\delta_k}{\sqrt{1 + \sum_{k=1}^{\infty} \delta_k^2}}, k = 1, 2, 3, \dots, \\ \gamma_0 &= \frac{1}{\sqrt{1 + \sum_{k=1}^{\infty} \delta_k^2}}, \end{aligned} \quad (13)$$

where $\sum_{k=1}^{\infty} \delta_k^2 < \infty$. However, because there are uncountable many Borel subsets B of $[0, 1]$ for which (12) 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*

$$h(u) = \frac{(1 + \sum_{k=1}^{\infty} \delta_k \rho_k(u))^2}{1 + \sum_{k=1}^{\infty} \delta_k^2} \text{ a.e. on } [0, 1]. \quad (14)$$

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

$$h_n(u) = h_n(u|\delta) = \frac{(1 + \sum_{k=1}^n \delta_k \rho_k(u))^2}{1 + \sum_{k=1}^n \delta_k^2}, \delta = (\delta_1, \dots, \delta_n)'. \quad (15)$$

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 (15) such that $\lim_{n \rightarrow \infty} \int_0^1 |h(u) - h_n(u)| du = 0$. 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 $\lim_{n \rightarrow \infty} \sup_{0 \leq u \leq 1} |H(u) - H_n(u)| = 0$.*

The density functions of the type (15) 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 (14) 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 $[0, 1]$, 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 .

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 (15) can be written as

$$H_n(u|\delta) = \frac{\int_0^u (1 + \sum_{m=1}^n \delta_m \rho_m(v))^2 dv}{1 + \sum_{m=1}^n \delta_m^2} = \frac{(1, \delta') A_{n+1}(u) \binom{1}{\delta}}{1 + \delta' \delta}, \quad (16)$$

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

where

$$A_{n+1}(u) = \begin{pmatrix} \int_0^u \rho_0(v)\rho_0(v)dv & \int_0^u \rho_0(v)\rho_1(v)dv & \cdots & \int_0^u \rho_0(v)\rho_n(v)dv \\ \int_0^u \rho_1(v)\rho_0(v)dv & \int_0^u \rho_1(v)\rho_1(v)dv & \cdots & \int_0^u \rho_1(v)\rho_n(v)dv \\ \vdots & \vdots & \ddots & \vdots \\ \int_0^u \rho_n(v)\rho_0(v)dv & \int_0^u \rho_n(v)\rho_1(v)dv & \cdots & \int_0^u \rho_n(v)\rho_n(v)dv \end{pmatrix}.$$

Because $\rho_m(u)$ is a polynomial of order m in u , the integral $\int_0^u \rho_k(v)\rho_m(v)dv$ is a polynomial of order $k + m + 1$. Therefore, we can write

$$\int_0^u \rho_k(v)\rho_m(v)dv = \sum_{j=0}^{k+m+1} a_{k,m,j}\rho_j(u),$$

where

$$a_{k,m,j} = \int_0^1 \rho_j(u) \left(\int_0^u \rho_k(v)\rho_m(v)dv \right) du$$

with

$$\sum_{j=0}^{k+m+1} a_{k,m,j}\rho_j(0) = 0, \quad \sum_{j=0}^{k+m+1} a_{k,m,j}\rho_j(1) = I(k = m).$$

It follows from (6) that the latter two conditions imply

$$\begin{aligned} a_{k,m,0} - \sqrt{3}a_{k,m,1} + \sum_{j=2}^{k+m+1} a_{k,m,j}\rho_j(0) &= 0, \\ a_{k,m,0} + \sqrt{3}a_{k,m,1} + \sum_{j=2}^{k+m+1} a_{k,m,j}\rho_j(1) &= I(k = m), \end{aligned}$$

hence

$$a_{k,m,0} = \frac{1}{2}I(k = m) - \frac{1}{2} \sum_{j=2}^{k+m+1} a_{k,m,j} (\rho_j(1) + \rho_j(0)), \quad (17)$$

$$a_{k,m,1} = \frac{1}{2\sqrt{3}}I(k = m) - \frac{1}{2\sqrt{3}} \sum_{j=2}^{k+m+1} a_{k,m,j} (\rho_j(1) - \rho_j(0)). \quad (18)$$

We can now write

$$A_{n+1}(u) = \sum_{j=0}^{2n+1} A_{j,n+1}\rho_j(u),$$

where $A_{j,n+1}$ is the $(n+1) \times (n+1)$ matrix with elements $a_{k,m,j}$. Hence, $H_n(u|\delta)$ is a polynomial of order $2n+1$.

Next, partition the matrices $A_{j,n+1}$ as

$$A_{j,n+1} = \begin{pmatrix} a_{0,0,j} & d'_{j,n} \\ d_{j,n} & C_{j,n} \end{pmatrix},$$

and observe from (6) and the orthonormality of the $\rho_j(u)$'s that

$$a_{0,0,j} = \int_0^1 \rho_j(u) u du = \begin{cases} 1/2 & \text{if } j = 0, \\ \frac{1}{2\sqrt{3}} & \text{if } j = 1, \\ 0 & \text{if } j \geq 2. \end{cases}$$

Then $H_n(u|\delta)$ can be written as

$$H_n(u|\delta) = \frac{u + \sum_{j=0}^{2n+1} (2\delta' d_{j,n} + \delta' C_{j,n} \delta) \rho_j(u)}{1 + \delta' \delta}.$$

The coefficients $a_{k,m,j}$ for $j = 2, \dots, k+m+1$ can easily be computed in advance by Monte Carlo integration, as follows. For $i = 1, \dots, M$, draw U_i and V_i independently from the uniform $[0, 1]$ distribution. Then

$$a_{k,m,j} = E [\rho_j(U_i) I(V_i < U_i) \rho_k(V_i) \rho_m(V_i)].$$

hence, for $j = 2, \dots, k+m+1$,

$$\tilde{a}_{k,m,j} = \frac{1}{M} \sum_{i=1}^M \rho_j(U_i) I(V_i < U_i) \rho_k(V_i) \rho_m(V_i) \rightarrow a_{k,m,j} \text{ a.s.}$$

as $M \rightarrow \infty$. The estimates of $a_{k,m,0}$ and $a_{k,m,1}$ then follow from (17) and (18).

4 The mixed proportional hazard model with unobserved heterogeneity

4.1 The MPH model

Let T be a duration, and let X be a vector of covariates. As is well-known, the conditional hazard function is defined as

$$\frac{f(t|X)}{1 - F(t|X)} = \lambda(t, X),$$

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

$$\begin{aligned} S(t|X, \alpha, \beta) &= S(t|X) \\ &= E\left[\exp\left(-\exp(\beta'X + U)\int_0^t \lambda(\tau|\alpha)d\tau\right)\middle|X\right], \end{aligned} \quad (19)$$

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. Denoting the distribution function of $V = \exp(U)$ by $G(v)$, and the integrated baseline hazard by

$$\Lambda(t|\alpha) = \int_0^t \lambda(\tau|\alpha)d\tau,$$

we have

$$\begin{aligned} S(t|X, \alpha, \beta, h) &= \int_0^\infty \exp(-v \cdot \exp(\beta'X)\Lambda(t|\alpha)) dG(v) \\ &= \int_0^\infty (\exp(-\exp(\beta'X)\Lambda(t|\alpha)))^v dG(v) \\ &= H(\exp(-\exp(\beta'X)\Lambda(t|\alpha))), \end{aligned} \quad (20)$$

where

$$H(u) = \int_0^\infty u^v dG(v), \quad u \in [0, 1], \quad (21)$$

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

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

$$\int_0^\infty v u^{v-1} dG(v) \leq u^{-1} \int_0^\infty v dG(v) < \infty, \quad (22)$$

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

$$h(u) = \int_0^\infty v u^{v-1} dG(v). \quad (23)$$

This is the reason for the argument h in the left-hand side of (20). Moreover, (22) implies that $h(u)$ is finite and continuous⁵ 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

$$\begin{aligned} S(t|X, \alpha_0, \beta_0, h_0) &= \int_0^\infty \exp(-v \cdot \exp(\beta_0' X) \Lambda(t|\alpha_0)) dG_0(v) \quad (24) \\ &= H_0(\exp(-\exp(\beta_0' X) \Lambda(t|\alpha_0))) \end{aligned}$$

where $H_0(u) = \int_0^u h_0(v) dv = \int_0^\infty u^v dG_0(v)$. In the expressions (20) and (24), h and h_0 should be interpreted as unknown parameters contained in a parameter space $\mathcal{D}(0, 1)$, say, of density functions on $(0, 1]$.

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

$$\Lambda(t|\alpha) = \Lambda(t|\alpha_0) \text{ for all } t > 0 \text{ implies } \alpha = \alpha_0, \quad (25)$$

and

$$\int_0^\infty v dG_0(v) = \int_0^\infty v dG(v) = 1 \quad (26)$$

(which by (23) 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 $[0, 1]$. See also Heckman and Singer (1984) for an alternative identification proof.

For the ease of reference I will call this model the Semi-Nonparametric Mixed Proportional Hazard (SNP-MPH) model.

4.2 The SNP-MPH likelihood function under bracketing

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.

⁵The continuity also follows from the dominated convergence theorem.

The actual duration is a latent variable $T_j^* > 0$ with conditional survival function

$$P[T_j^* > t|X_j] = S(t|X_j, \alpha_0, \beta_0, h_0), \quad (27)$$

where $S(t|X, \alpha, \beta, h)$ is defined by (24), α_0 and β_0 are the true parameter vectors and h_0 is the true density (23). If $C_j = 0$ then T_j^* is observed: $T_j = T_j^*$, and if $C_j = 1$ then T_j^* is censored: $T_j = \bar{T}_j < T_j^*$, where $[1, \bar{T}_j]$ is the time interval over which individual j has been, or would have been, monitored. It will be assumed that \bar{T}_j is entirely determined by the setup of the survey, and may therefore be considered exogenous, and that

$$\bar{T} = \inf_{j \geq 1} \bar{T}_j > 0.$$

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 $b_1 < b_2 < \dots < b_M \leq \bar{T}$, and create the dummy variables

$$\begin{aligned} D_{1,j} &= I(T_j \leq b_1) \\ D_{2,j} &= I(b_1 < T_j \leq b_2) \\ &\vdots \\ D_{M,j} &= I(b_{M-1} < T_j \leq b_M) \end{aligned} \quad (28)$$

where $I(\cdot)$ is the indicator function. This procedure is known as bracketing.

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

$$\mu_i(\alpha, \beta' X_j) = \exp(-\exp(\beta' X_j) \Lambda(b_i | \alpha)). \quad (29)$$

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

$$\begin{aligned} P[D_{i,j} = 1|X_j] &= S(b_{i-1}|X_j, \alpha_0, \beta_0, h_0) - S(b_i|X_j, \alpha_0, \beta_0, h_0) \\ &= H_0(\mu_{i-1}(\alpha_0, \beta'_0 X_j)) - H_0(\mu_i(\alpha_0, \beta'_0 X_j)) \\ &\quad i = 1, 2, \dots, M, \end{aligned} \quad (30)$$

$$P \left[\sum_{i=1}^M D_{i,j} = 0 \middle| X_j \right] = S(b_M|X_j, \alpha_0, \beta_0, h_0) = H_0(\mu_M(\alpha_0, \beta'_0 X_j)),$$

where $H_0(u) = \int_0^u h_0(v) dv$. Thus, the conditional log-likelihood function takes the form

$$\ln(L_N(\alpha, \beta, h)) \quad (31)$$

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

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

4.3 Baseline hazard specification

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

$$\begin{aligned}
\Lambda(t|\alpha) &= \Lambda(b_{i-1}|\alpha) + \alpha_i (t - b_{i-1}) & (32) \\
&= \sum_{k=1}^{i-1} \alpha_k (b_k - b_{k-1}) + \alpha_i (t - b_{i-1}) \text{ for } t \in (b_{i-1}, b_i], \\
\alpha_m &> 0 \text{ for } m = 1, \dots, M, \alpha = (\alpha_1, \dots, \alpha_M)' \in \mathbb{R}^M.
\end{aligned}$$

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

$$\Lambda(b_i|\alpha) = \sum_{m=1}^i \alpha_m, \alpha_m > 0 \text{ for } m = 1, \dots, M, \quad (33)$$

or

$$\begin{aligned}
\Lambda(b_i|\alpha) &= \exp(\alpha_i), \alpha_1 < \alpha_2 < \dots < \alpha_M, & (34) \\
\Lambda(b_0|\alpha) &= \Lambda(0|\alpha) = 0.
\end{aligned}$$

The advantage of the specification (32) is that the null hypothesis $\alpha_1 = \dots = \alpha_M$ corresponds to the integrated Weibull hazard $\Lambda(t|\alpha) = \alpha_1 t$. 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 (33) is useful for deriving identification conditions, as will be shown in the next subsection. The same applies to (34).

Note that under specification (34) the probability model (30) takes the form of a generalized ordered probability model, similarly to an ordered probit or logit model:

$$P[D_{1,j} = 1|X_j] = F_0(\beta'_0 X_j + \alpha_{0,1})$$

$$\begin{aligned}
P[D_{i,j} = 1|X_j] &= F_0(\beta'_0 X_j + \alpha_{0,i}) - F_0(\beta'_0 X_j + \alpha_{0,i-1}), \quad i = 2, \dots, M, \\
P\left[\sum_{i=1}^M D_{i,j} = 0 \middle| X_j\right] &= 1 - F_0(\beta'_0 X_j + \alpha_{0,M}), \quad (35)
\end{aligned}$$

where

$$F_0(x) = 1 - H_0(\exp(-\exp(x))), \quad (36)$$

with density

$$f_0(x) = h_0(\exp(-\exp(x))) \exp(-\exp(x)) \exp(x). \quad (37)$$

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 (35). Moreover, for the identification of α_0 and β_0 in (35) it is necessary to normalize the location and scale of the distribution $F_0(x)$.

These normalization conditions will be considered in the next subsections. In addition, we also need to require that $\beta'_0 X_j = \beta' X_j$ a.s. implies $\beta = \beta_0$. A sufficient condition for the latter is that the matrix $E[X_j X'_j]$ is well-defined (which is the case if $E[X'_j X_j] < \infty$) and is non-singular:

Assumption 1. *The vector X_j of covariates does not contain a constant, and $E[X'_j X_j] < \infty$. The matrix $E[X_j X'_j]$ is nonsingular.*

4.4 Nonparametric identification via extreme values

First, let us assume that $\Lambda(b_i|\alpha)$ is specified as (33). It follows easily from the inequality $\ln(x) < x - 1$ if $x > 0$ and $x \neq 1$, and the equality

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

that

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

if and only if

$$\begin{aligned}
& P \left[H \left(\exp \left(- \exp (\beta' X_j) \left(\sum_{m=1}^i \alpha_m \right) \right) \right) \right. \\
& \left. = H_0 \left(\exp \left(- \exp (\beta'_0 X_j) \left(\sum_{m=1}^i \alpha_{0,m} \right) \right) \right) \middle| X_j \right] < 1.
\end{aligned}$$

This implies that

$$E \left[N^{-1} \ln (L_N(\alpha, \beta, h) / L_N(\alpha_0, \beta_0, h_0)) \right] = 0 \quad (38)$$

if and only if

$$\begin{aligned}
& H \left(\exp \left(- \exp (\beta' X) \left(\sum_{m=1}^i \alpha_m \right) \right) \right) \quad (39) \\
& = H_0 \left(\exp \left(- \exp (\beta'_0 X) \left(\sum_{m=1}^i \alpha_{0,m} \right) \right) \right) \\
& \text{a.s. for } i = 1, \dots, M,
\end{aligned}$$

where $X = X_j$.

Now the question arises: Does (39) imply that $\alpha = \alpha_0$, $\beta = \beta_0$ and $H = H_0$, and if not, which additional conditions are needed to achieve identification?

For $i = 1$, (39) reads

$$H(\exp(-\exp(\beta' X) \alpha_1)) = H_0(\exp(-\exp(\beta'_0 X) \alpha_{0,1})) \text{ a.s.} \quad (40)$$

Let

$$\Upsilon(u) = \ln(H^{-1}(H_0(u))). \quad (41)$$

Then it follows from (40) that

$$\alpha_1 = -\exp(-\beta' X) \Upsilon(\exp(-\exp(\beta'_0 X) \alpha_{0,1})). \quad (42)$$

Taking the derivative to X' it follows that

$$\begin{aligned}
0 & = \beta \exp(-\beta' X) \Upsilon(\exp(-\exp(\beta'_0 X) \alpha_{0,1})) \\
& \quad + \alpha_{0,1} \exp(-\beta' X) \Upsilon'(\exp(-\exp(\beta'_0 X) \alpha_{0,1})) \\
& \quad \times \exp(-\exp(\beta'_0 X) \alpha_{0,1}) \exp(\beta'_0 X) \beta \\
& = -\beta + \alpha_{0,1} \exp((\beta_0 - \beta)' X) \Upsilon'(\exp(-\exp(\beta'_0 X) \alpha_{0,1})) \\
& \quad \times \exp(-\exp(\beta'_0 X)) \beta_0
\end{aligned}$$

hence

$$\begin{aligned} \beta &= \alpha_{0,1} \exp((\beta_0 - \beta)' X) \Upsilon'(\exp(-\exp(\beta'_0 X) \alpha_{0,1})) \\ &\quad \times \exp(-\exp(\beta'_0 X) \alpha_{0,1}) \beta_0. \end{aligned} \quad (43)$$

If $\beta_0 \neq 0$ then (43) implies that

$$\begin{aligned} c &= \alpha_{0,1} \exp((\beta_0 - \beta)' X) \Upsilon'(\exp(-\exp(\beta'_0 X) \alpha_{0,1})) \\ &\quad \times \exp(-\exp(\beta'_0 X) \alpha_{0,1}) \end{aligned} \quad (44)$$

is constant and thus

$$\beta = c \beta_0. \quad (45)$$

Note that $c > 0$ because Υ is monotonic increasing. Substituting (45) in (42) yields

$$\alpha_1 \exp(c \beta'_0 X) = -\Upsilon(\exp(-\exp(\beta'_0 X) \alpha_{0,1})), \quad (46)$$

which by (41) implies that

$$H_0(\exp(-\exp(\beta'_0 X) \alpha_{0,1})) = H(\exp(-\exp(c \beta'_0 X) \alpha_1)) \text{ a.s.}$$

or equivalently

$$H_0(u) = H(\exp(-\alpha_1 \alpha_{0,1}^{-c} (\ln(1/u))^c)) \quad (47)$$

for all u in the support S_1 of $U = \exp(-\exp(\beta'_0 X) \alpha_{0,1})$.

Now suppose that

$$\forall x \in \mathbb{R}, P[\beta'_0 X_j \leq x] > 0 \quad (48)$$

and

$$h_0(1) = h(1) = 1. \quad (49)$$

Recall that (49) corresponds to the condition that $E[V] = 1$. The condition (48) implies that S_1 contains a sequence u_n which converges to 1. Hence, it follows from (48) and (47) that

$$\begin{aligned} 1 &= h_0(1) = \lim_{n \rightarrow \infty} \frac{H_0(1) - H_0(u_n)}{1 - u_n} \\ &= \lim_{n \rightarrow \infty} \frac{H(1) - H(\exp(-\alpha_1 \alpha_{0,1}^{-c} (\ln(1/u_n))^c))}{1 - u_n} \end{aligned} \quad (50)$$

$$\begin{aligned}
&= h(1) \lim_{n \rightarrow \infty} \frac{1 - \exp(-\alpha_1 \alpha_{0,1}^{-c} (\ln(1/u_n))^c)}{1 - u_n} \\
&= \frac{c\alpha_1}{\alpha_{0,1}^c} \lim_{u \uparrow 1} \exp(-\alpha_1 \alpha_{0,1}^{-c} (\ln(1/u))^c) (\ln(1/u))^{c-1} \frac{1}{u} \\
&= \frac{c\alpha_1}{\alpha_{0,1}^c} \cdot \lim_{z \downarrow 0} z^{c-1} = \begin{cases} 0 & \text{if } c > 1, \\ \infty & \text{if } c < 1, \\ \alpha_1/\alpha_{0,1} & \text{if } c = 1. \end{cases}
\end{aligned}$$

Therefore, (48) and (49) imply that $c = 1$, $\alpha_1 = \alpha_{0,1}$, $\beta = \beta_0$ and

$$H(\exp(-\exp(\beta'_0 X) \alpha_{0,1})) = H_0(\exp(-\exp(\beta'_0 X) \alpha_{0,1})) \text{ a.s.}$$

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

$$\begin{aligned}
&H(\exp(-\exp(\beta'_0 X) (\alpha_{0,1} + \alpha_2))) & (51) \\
&= H_0(\exp(-\exp(\beta'_0 X) (\alpha_{0,1} + \alpha_{0,2}))) \text{ a.s.}
\end{aligned}$$

imply $\alpha_2 = \alpha_{0,2}$? Let S_2 be the support of $U = \exp(-\exp(\beta'_0 X) (\alpha_{0,1} + \alpha_{0,2}))$ and let

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

Then (51) implies that $H(u^{1+\eta}) = H_0(u)$ for all $u \in S_2$. Under condition (48) there exists a sequence u_n in S_2 which converges to 1. Therefore, similarly to (50) we have

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

Hence, under the conditions (48) and (49), $\eta = 0$ and thus $\alpha_2 = \alpha_{0,2}$. Repeating this argument for $i = 3, \dots, M$, it follows that $\alpha = \alpha_0$ and

$$H_0(\exp(-\exp(\beta'_0 X) \alpha_{0,i})) = H(\exp(-\exp(\beta'_0 X) \alpha_{0,i})) \text{ a.s.}$$

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

Since (32) and (33) are equivalent for $t = b_i$, it follows now that:

Theorem 5. *Let the integrated baseline hazard be specified by (32). Then under Assumption 1 and the conditions (48) and (49),*

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

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

$$H(\exp(-\exp(\beta'_0 X_j) \Lambda(b_i|\alpha_0))) = H_0(\exp(-\exp(\beta'_0 X_j) \Lambda(b_i|\alpha_0))) \quad (52)$$

a.s. for $i = 1, \dots, M$. If in addition the support of $\beta'_0 X_j$ is the whole real line \mathbb{R} then (52) implies that $h(u) = h_0(u)$ a.e. on $(0, 1]$.

Admittedly, condition (48) is 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 MPH model may not be identified. The case where the distribution of the covariates is discrete will be dealt with later, in Section 9

Since the result that $h(u) = h_0(u)$ a.e. on $(0, 1]$ will be needed to prove consistency of the SNP maximum likelihood estimators of α_0 , β_0 and h_0 , I will assume that

Assumption 2. *The support of $\beta'_0 X_j$ is the whole real line \mathbb{R} .*

The condition (49) is only effective in pinning down α_0 and β_0 if $U_{i,j} = \exp(-\exp(\beta'_0 X_j) \Lambda(b_i|\alpha_0))$ can get close enough to 1. Thus, the identification hinges on the **extreme values** of $\beta'_0 X_j$. Under Assumption 2 it follows that $p \lim_{N \rightarrow \infty} \min_{j=1, \dots, N} \beta'_0 X_j = -\infty$, hence $p \lim_{N \rightarrow \infty} \max_{j=1, \dots, N} U_{i,j} = 1$, but in finite samples $\max_{j=1, \dots, N} U_{i,j}$ may not get close enough to 1 for condition (49) to be effective. Therefore, I will now derive alternative identification conditions.

4.5 Nonparametric identification via moment conditions

Consider the generalized probability model form (35) of the SNP-MPH model under review. Let $F(x)$ be a distribution function of the type (36),

$$F(x) = 1 - H(\exp(-\exp(x))) \quad (53)$$

with density

$$f(x) = h(\exp(-\exp(x))) \exp(-\exp(x)) \exp(x), \quad (54)$$

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

$$F(\sigma x + \mu) \equiv F_0(x). \quad (55)$$

Clearly, under Assumptions 1-2 the SNP-MPH model in generalized probability model form (35) is nonparametrically identified if (55) implies $\mu = 0$, $\sigma = 1$. Taking derivatives of (55) it follows that (55) implies

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

hence it follows from (37) and (54) that for any function φ on \mathbb{R} for which $\int_{-\infty}^{\infty} \varphi(x) f_0(x) dx$ is well-defined,

$$\begin{aligned} &\int_0^1 \varphi(\ln(\ln(1/u))) h_0(u) du \\ &= \int_{-\infty}^{\infty} \varphi(x) f_0(x) dx = \sigma \int_{-\infty}^{\infty} \varphi(x) f(\sigma x + \mu) dx \\ &= \int_{-\infty}^{\infty} \varphi\left(\frac{x - \mu}{\sigma}\right) f(x) dx \\ &= \int_0^1 \varphi\left(\frac{\ln(\ln(1/u)) - \mu}{\sigma}\right) h(u) du. \end{aligned} \quad (56)$$

If we choose $\varphi(x) = x$ then (56) implies

$$\int_0^1 \ln(\ln(1/u)) h(u) du = \sigma \int_0^1 \ln(\ln(1/u)) h_0(u) du + \mu, \quad (57)$$

and if we choose $\varphi(x) = x^2$ then (56) implies

$$\begin{aligned} \sigma^2 \int_0^1 (\ln(\ln(1/u)))^2 h_0(u) du &= \int_0^1 (\ln(\ln(1/u)))^2 h(u) du \\ &\quad - 2\mu \int_0^1 \ln(\ln(1/u)) h(u) du + \mu^2. \end{aligned} \quad (58)$$

Now assume that

$$\int_0^1 \ln(\ln(1/u)) h(u) du = \int_0^1 \ln(\ln(1/u)) h_0(u) du, \quad (59)$$

$$\int_0^1 (\ln(\ln(1/u)))^2 h(u) du = \int_0^1 (\ln(\ln(1/u)))^2 h_0(u) du. \quad (60)$$

Then it follows from (57) and (59) that

$$\mu = (1 - \sigma) \int_0^1 \ln(\ln(1/u)) h_0(u) du \quad (61)$$

and from (58) through (61) that

$$\begin{aligned} (\sigma^2 - 1) \int_0^1 (\ln(\ln(1/u)))^2 h_0(u) du \\ = (\sigma^2 - 1) \left(\int_0^1 \ln(\ln(1/u)) h_0(u) du \right)^2. \end{aligned} \quad (62)$$

The latter equality implies $\sigma = 1$ because

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

and it follows now from (61) that $\mu = 0$.

Note that the values of the integrals in (59) and (60) do not matter for this result, provided that the integrals involved are finite. In order to accommodate the benchmark case $h_0(u) = h(u) \equiv 1$, which corresponds to absence of unobserved heterogeneity, let us now assume that the density h in the log-likelihood function (31) is confined to a space of density functions h on $(0, 1]$ satisfying the moment conditions

$$\int_0^1 \ln(\ln(1/u)) h(u) du = \int_0^1 \ln(\ln(1/u)) du, \quad (63)$$

$$\int_0^1 (\ln(\ln(1/u)))^2 h(u) du = \int_0^1 (\ln(\ln(1/u)))^2 du. \quad (64)$$

It is obvious from the easy equalities

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

$$p = 1, 2,$$

that the right-hand side integrals in (63) and (64) are finite. Their values are

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

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

which have been computed by Monte Carlo integration.⁶

The moment conditions (63) and (64) can, at least in theory, be implemented by penalizing the log-likelihood function (31) for deviations from the moment conditions involved by augmenting the log-likelihood function $\ln(L_N(\alpha, \beta, h))$ with two penalty terms:

$$\begin{aligned} \ln(L_N^*(\alpha, \beta, h)) &= \ln(L_N(\alpha, \beta, h)) & (65) \\ &- N \left(\int_0^1 \ln(\ln(1/u)) h(u) du - \int_0^1 \ln(\ln(1/u)) du \right)^{2\ell} \\ &- N \left(\int_0^1 (\ln(\ln(1/u)))^2 h(u) du - \int_0^1 (\ln(\ln(1/u)))^2 du \right)^{2\ell} \end{aligned}$$

for some integer $\ell \geq 1$. Then, similar to Theorem 5, we have:

Theorem 6. *Let the integrated baseline hazard be specified by (32) and the penalized log-likelihood function $L_N^*(\alpha, \beta, h)$ by (65). Under Assumptions 1-2, $E[N^{-1} \ln(L_N^*(\alpha, \beta, h)/L_N^*(\alpha_0, \beta_0, h_0))] = 0$ implies that $\alpha = \alpha_0$, $\beta = \beta_0$, and $h(u) = h_0(u)$ a.e. on $(0, 1]$.*

⁶Using one million random drawings from the uniform $[0, 1]$ distribution.

4.6 Moment conditions for SNP density functions

For SNP density functions (15) the moment conditions (63) and (64) read

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

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

$$\begin{aligned}
a'_{n,p} &= \left(\int_0^1 (\ln(\ln(1/u)))^p \rho_1(u) du, \dots, \int_0^1 (\ln(\ln(1/u)))^p \rho_n(u) du \right) \\
B_{n,p} &= \begin{pmatrix} \int_0^1 (\ln(\ln(1/u)))^p \rho_1(u) \rho_1(u) du & \cdots & \int_0^1 (\ln(\ln(1/u)))^p \rho_1(u) \rho_n(u) du \\ \vdots & \ddots & \vdots \\ \int_0^1 (\ln(\ln(1/u)))^p \rho_n(u) \rho_1(u) du & \cdots & \int_0^1 (\ln(\ln(1/u)))^p \rho_n(u) \rho_n(u) du \end{pmatrix} \\
&\quad - \int_0^1 (\ln(\ln(1/u)))^p du \cdot I_n,
\end{aligned}$$

the conditions (63) and (64) with h replaced by (15) are equivalent to

$$\begin{aligned}
2\delta' a_{n,1} + \delta' B_{n,1} \delta &= 0, \\
2\delta' a_{n,2} + \delta' B_{n,2} \delta &= 0,
\end{aligned}$$

respectively, where $\delta = (\delta_1, \dots, \delta_n)'$. Therefore, if we replace h in (65) by $h_n(\cdot|\delta)$, the penalized log-likelihood can be written as

$$\begin{aligned}
\ln(L_N^*(\alpha, \beta, h_n(\cdot|\delta))) &= \ln(L_N(\alpha, \beta, h_n(\cdot|\delta))) \\
&\quad - N \left(\frac{2\delta' a_{n,1} + \delta' B_{n,1} \delta}{1 + \delta' \delta} \right)^{2\ell} - N \left(\frac{2\delta' a_{n,2} + \delta' B_{n,2} \delta}{1 + \delta' \delta} \right)^{2\ell}
\end{aligned} \tag{66}$$

for some integer $\ell \geq 1$.

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

$$\begin{aligned} & \lim_{N \rightarrow \infty} \text{Var} \left(\frac{1}{\sqrt{N}} \frac{\partial \ln(L_N^{**}(\alpha_0, \beta_0, h_n(\cdot|\delta_0)))}{\partial (\alpha'_0, \beta'_0, \delta'_0)} \right) \\ &= \lim_{N \rightarrow \infty} \text{Var} \left(\frac{1}{\sqrt{N}} \frac{\partial \ln(L_N(\alpha_0, \beta_0, h_n(\cdot|\delta_0)))}{\partial (\alpha'_0, \beta'_0, \delta'_0)} \right) \end{aligned}$$

and

$$\begin{aligned} & \lim_{N \rightarrow \infty} E \left(\frac{-1}{N} \frac{\partial^2 \ln(L_N^{**}(\alpha_0, \beta_0, h_n(\cdot|\delta_0)))}{\partial (\alpha'_0, \beta'_0, \delta'_0)' \partial (\alpha'_0, \beta'_0, \delta'_0)} \right) \\ &= \lim_{N \rightarrow \infty} E \left(\frac{-1}{N} \frac{\partial^2 \ln(L_N(\alpha_0, \beta_0, h_n(\cdot|\delta_0)))}{\partial (\alpha'_0, \beta'_0, \delta'_0)' \partial (\alpha'_0, \beta'_0, \delta'_0)} \right). \end{aligned}$$

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

Finally, note that the vectors $a_{n,p}$ and matrices $B_{n,p}$ can easily be computed in advance by Monte Carlo integration.

5 Ordered probability models

Ordered models apply to discrete dependent variables Y representing an ordering of items. For example, let Y be the outcome of a taste test, coded like $Y = 0 \Leftrightarrow \text{disgusting}$, $Y = 1 \Leftrightarrow \text{bad}$, $Y = 2 \Leftrightarrow \text{OK}$, $Y = 3 \Leftrightarrow \text{good}$, $Y = 4 \Leftrightarrow \text{delicious}$. In this case Y is not a quantity, but nevertheless a larger value of Y means more, or better. This type of data is usually modeled conditional on covariates via a latent variable model:

$$Y^* = -\beta'_0 X - \gamma_0 + \varepsilon,$$

where X is a vector of covariates, exclusive a constant, and ε is an independent error term with distribution function $F_0(\cdot)$. The conditional probabilities $P[Y \leq m|X]$ for $m = 0, 1, \dots, M$, are assumed to be related to the conditional distribution of the unobserved latent variable Y^* by

$$\begin{aligned} P[Y = 0|X] &= P[Y^* \leq 0|X] = F_0(\gamma_0 + \beta'_0 X), \\ P[Y = 1|X] &= P[0 < Y^* \leq \delta_1|X] \end{aligned} \tag{67}$$

$$\begin{aligned}
&= F_0(\delta_1 + \gamma_0 + \beta'_0 X) - F_0(\gamma_0 + \beta'_0 X), \\
P[Y = m|X] &= P[\delta_{m-1} < Y^* \leq \delta_m|X] \\
&= F_0(\delta_m + \gamma_0 + \beta'_0 X) - F_0(\delta_{m-1} + \gamma_0 + \beta'_0 X) \\
&\quad \text{for } m = 2, \dots, M-1, \\
P[Y = M|X] &= P[Y^* > \delta_{M-1}|X] = 1 - F_0(\delta_{M-1} + \gamma_0 + \beta'_0 X),
\end{aligned}$$

where $0 < \delta_1 < \dots < \delta_{M-1}$. If $F_0(\cdot)$ is specified as the c.d.f. of the standard normal distribution or the logistic distribution we get the well-known ordered probit and logit models, respectively.

Now let

$$F_0(x) = 1 - H_0(1 - G(x))$$

where $G(x)$ is a given distribution function on \mathbb{R} , for example, let $G(x)$ be the Logit or Probit function, and the distribution function $H_0(u)$ on $[0, 1]$ is left unspecified, except for normalization conditions. This model will be referred to as the Semi-Nonparametric Ordered Probability (SNPOP) model.

Denoting $D_{m,j} = I(Y_j = m - 1)$ for $m = 1, \dots, M$, where $I(\cdot)$ is the indicator function, and

$$\begin{aligned}
\alpha_{0,1} &= \delta_0, \quad \alpha_{0,m} = \delta_{m-1} + \gamma_0 \quad \text{for } m = 2, \dots, M, \\
\alpha_{0,1} &< \alpha_{0,2} < \dots < \alpha_{0,M},
\end{aligned} \tag{68}$$

the SNPOP model involved takes the form (35), hence the log-likelihood function involved takes the form

$$\begin{aligned}
&\ln(L_N(\alpha, \beta, h)) \\
&= \sum_{j=1}^N \sum_{i=1}^M I(Y_j = i - 1) \ln(H(\mu_{i-1}(\alpha, \beta' X_j)) - H(\mu_i(\alpha, \beta' X_j))) \\
&\quad + \sum_{j=1}^N I(Y_j = M) \ln(H(\mu_M(\alpha, \beta' X_j)))
\end{aligned} \tag{69}$$

where again $\mu_0(\alpha, \beta' X_j) = 1$ but now

$$\mu_i(\alpha, \beta' X_j) = 1 - G(\alpha_i + \beta' X_j).$$

The main difference with the SNP-MPH model is that moment conditions (59) and (60) need to be adjusted, because these moment conditions corresponded to the special case $G(x) = 1 - \exp(-\exp(x))$. These conditions now

become

$$\begin{aligned}\int_0^1 G^{-1}(1-u) h_0(u) du &= \int_0^1 G^{-1}(1-u) h(u) du, \\ \int_0^1 (G^{-1}(1-u))^2 h_0(u) du &= \int_0^1 (G^{-1}(1-u))^2 h(u) du.\end{aligned}$$

Again, in order to accommodate the uniform case $h_0(u) = h(u) = 1$, we should confine the densities h_0 and h to the space of densities satisfying

$$\int_0^1 G^{-1}(1-u) h(u) du = \int_0^1 G^{-1}(1-u) du = \int_{-\infty}^{\infty} x dG(x), \quad (70)$$

$$\begin{aligned}\int_0^1 (G^{-1}(1-u))^2 h(u) du &= \int_0^1 (G^{-1}(1-u))^2 du \\ &= \int_{-\infty}^{\infty} x^2 dG(x),\end{aligned} \quad (71)$$

and similar to the SNP-MPH model case we can impose these conditions by penalizing the log-likelihood:

$$\begin{aligned}\ln(L_N^*(\alpha, \beta, h)) &= \ln(L_N(\alpha, \beta, h)) \\ &- N \left(\int_0^1 G^{-1}(1-u) h(u) du - \int_{-\infty}^{\infty} x dG(x) \right)^{2\ell} \\ &- N \left(\int_0^1 (G^{-1}(1-u))^2 h(u) du - \int_{-\infty}^{\infty} x^2 dG(x) \right)^{2\ell}.\end{aligned} \quad (72)$$

In particular, if G is the logistic distribution function,

$$G(x) = 1 / (1 + \exp(-x)), \quad (73)$$

then $G^{-1}(1-u) = \ln((1-u)/u)$, so that in that case,

$$\begin{aligned}\int_0^1 \ln\left(\frac{1-u}{u}\right) du &= \int_{-\infty}^{\infty} x dG(x) = 0 \\ \int_0^1 \left(\ln\left(\frac{1-u}{u}\right)\right)^2 du &= \int_{-\infty}^{\infty} x^2 dG(x) = 3.293257190.\end{aligned}$$

The latter value has been computed by Monte Carlo integration.

Thus, similar to Theorem 6 we have:

Theorem 7. *Let $L_N^*(\alpha, \beta, h)$ be the penalized log-likelihood (72) of the SNPOP model (67), with α_0 defined by (68). Under Assumptions 1-2,*

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

implies that $\alpha = \alpha_0$, $\beta = \beta_0$, and $h(u) = h_0(u)$ a.e. on $(0, 1]$.

6 Requirements for consistency of SNP maximum likelihood estimators

In both the SNP-MPH and SNPOP cases we can write the (penalized) log-likelihood as

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

where $Y_j = (D_{1,j}, \dots, D_{M,j}, X_j')'$. In the case (31),

$$\begin{aligned} & \Psi(Y_j, \alpha, \beta, h) \\ &= \sum_{i=1}^M D_{i,j} \ln (H(\mu_{i-1}(\alpha, \beta' X_j)) - H(\mu_i(\alpha, \beta' X_j))) \\ &+ \left(1 - \sum_{i=1}^M D_{i,j} \right) \ln (H(\mu_M(\alpha, \beta' X_j)) - \Pi(h)), \end{aligned} \tag{74}$$

where $\Pi(h)$ represents the two penalty terms. The function $\Psi(Y_j, \alpha, \beta, h)$ in the SNPOP case takes a similar form.

The maximum likelihood estimators of α_0 , β_0 and h_0 are

$$\left(\hat{\alpha}, \hat{\beta}, \hat{h} \right) = \arg \max_{\alpha \in A, \beta \in B, h \in \mathcal{D}(0,1)} N^{-1} \ln (L_N^*(\alpha, \beta, h)), \tag{75}$$

where A and B are 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

$$\|h_1 - h_2\|_1 = \int_0^1 |h_1(u) - h_2(u)| du. \quad (76)$$

Let

$$\bar{\Psi}(\alpha, \beta, h) = E[\Psi(Y_j, \alpha, \beta, h)]. \quad (77)$$

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

$$p \lim_{N \rightarrow \infty} \bar{\Psi}(\hat{\alpha}, \hat{\beta}, \hat{h}) = \bar{\Psi}(\alpha_0, \beta_0, h_0). \quad (78)$$

Similar to the standard consistency proof for M estimators it can be shown that if $\bar{\Psi}$ is continuous and (α_0, β_0, h_0) is unique then (78) implies that $p \lim_{N \rightarrow \infty} \hat{\alpha} = \alpha_0$, $p \lim_{N \rightarrow \infty} \hat{\beta} = \beta_0$ and $p \lim_{N \rightarrow \infty} \left\| \hat{h} - h_0 \right\|_1 = 0$.

In general it will be impossible to compute (75) because it requires to maximize the log-likelihood function over a space of density functions. However, there exists an increasing sequence $\mathcal{D}_N(0, 1)$ of compact subspaces of $\mathcal{D}(0, 1)$ such that the densities in $\mathcal{D}_N(0, 1)$ can be parametrized by a finite (but increasing) number of parameters, namely a space of densities of the type (15), 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

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

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

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 (78). These problems will be addressed in the next sections.

7 Compactness

Consider the space of density functions of the type (14), 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 $c > 0$,

$$|\delta_k| \leq \frac{c}{1 + \sqrt{k} \ln(k)}, \quad (80)$$

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

Theorem 8. *Let $\mathcal{D}(0, 1)$ be the space of densities of the type (14) subject to the restrictions (80) 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.*

Theorem 8 follows from the following lemmas:

Lemma 2. *Let $\xi = \{\xi_k\}_{k=0}^{\infty}$ be a given sequence of positive numbers satisfying*

$$\xi_0 > 1, \quad \sum_{k=1}^{\infty} \xi_k^2 < \infty, \quad (81)$$

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

Proof: It suffices to prove that \mathcal{F}_ξ 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}_ξ . Since $f_n(u)$ is a Cauchy sequence in the Hilbert space $L_{\mathcal{B}}^2(0, 1)$ it converges to a function $f(u) = \sum_{k=0}^{\infty} \gamma_k \rho_k(u)$ in $L_{\mathcal{B}}^2(0, 1)$. Now \mathcal{F}_ξ 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 $\|f_n - f\|_2 = \sqrt{\sum_{k=0}^{\infty} (\gamma_{n,k} - \gamma_k)^2} \rightarrow 0$ implies that for each k , $\gamma_{n,k} \rightarrow \gamma_k$. 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 $k = 0, 1, \dots, m$, $\gamma_{n,k} \rightarrow \gamma_k$, it follows that $0 \leq 1 - \sum_{k=0}^m \gamma_k^2 < \varepsilon$. Thus $\lim_{m \rightarrow \infty} \sum_{k=0}^m \gamma_k^2 = 1$. Hence \mathcal{F}_ξ 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 $\gamma_k \in [-\xi_k, \xi_k]$, $k = 0, 1, 2, \dots, n$. 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 $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_n)'$ satisfying $\gamma \in \times_{k=0}^n [-\xi_k, \xi_k]$, $\gamma' \gamma \leq 1$ is a closed and bounded subset of \mathbb{R}^{n+1} and is therefore compact, and consequently, $\mathcal{F}_{\xi,n}$ is compact. Therefore, there exists a finite number of functions $f_1, \dots, f_M \in \mathcal{F}_{\xi,n}$ such that

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

This implies that

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

hence \mathcal{F}_ξ is totally bounded. Q.E.D.

Lemma 3. *Under condition (81) 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 (13) and (81) that

$$\begin{aligned} \gamma_k^2 &= \frac{\delta_k^2}{1 + \sum_{k=1}^{\infty} \delta_k^2} \leq \xi_k^2, \quad k \geq 1, \\ \gamma_0^2 &= \frac{1}{1 + \sum_{k=1}^{\infty} \delta_k^2} \leq 1 < \xi_0^2 \\ \gamma_0^2 &\geq \frac{1}{1 + \sum_{k=1}^{\infty} \xi_k^2} > 0, \end{aligned} \tag{82}$$

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

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

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

Lemma 4. *The space $\mathcal{D}_\xi(0, 1) = \{h : h = f^2, f \in \mathcal{F}_\xi^*\}$ of density functions on $[0, 1]$ endowed with the metric (76) is compact.*

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

$$\int_0^1 |f(u)^2 - g(u)^2| du \tag{83}$$

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

Let $h_n = f_n^2$ be an infinite sequence in $\mathcal{D}_\xi(0, 1)$. Because \mathcal{F}_ξ^* is compact, there exists a subsequence f_{m_n} which converges to a limit f in \mathcal{F}_ξ^* , hence it follows from (83) 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 $\mathcal{D}(0, 1) \subset \mathcal{D}_{\xi_k}(0, 1)$. It is now easy to verify that $\mathcal{D}(0, 1)$ is sequentially compact and therefore compact. Q.E.D.

Of course, the result of Theorem 8 is only useful for our purpose if the constant c in (80) is chosen so large that

Assumption 3. *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 (15) subject to the restrictions (80), with c the same as for $\mathcal{D}(0, 1)$. For each N , $\mathcal{D}_N(0, 1)$ is a compact subset of $\mathcal{D}(0, 1)$, and for each $h \in \mathcal{D}(0, 1)$ there exists a sequence $h_N \in \mathcal{D}_N(0, 1)$ such that $\lim_{N \rightarrow \infty} \int_0^1 |h(u) - h_N(u)| du = 0$.*

8 Consistency of M-estimators of non-Euclidean parameters

I will now address the problem how to prove (78). To relate (78) to Theorem 10 below, denote

$$\Theta = \{(\alpha, \beta, h) : \alpha \in A, \beta \in B, h \in \mathcal{D}(0, 1)\},$$

where

Assumption 4. *A and B are compact sets containing the true parameter vectors α_0 and β_0 , respectively,*

and define a metric $d(\cdot, \cdot)$ 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

$$d(\theta_1, \theta_2) = \max \left[\sqrt{(\alpha_1 - \alpha_2)'(\alpha_1 - \alpha_2)}, \right. \quad (84) \\ \left. \sqrt{(\beta_1 - \beta_2)'(\beta_1 - \beta_2)}, \int_0^1 |h_1(u) - h_2(u)| du \right].$$

Theorem 10. *Let Y_j , $j = 1, \dots, N$, 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 $\theta \in \Theta$,*

$$E [|g(Y_1, \theta)|] < \infty, \quad (85)$$

so that

$$\bar{g}(\theta) = E [g(Y_1, \theta)]$$

is defined and finite, and let for some constant $K_0 > 0$,

$$E \left[\max \left(\sup_{\theta \in \Theta} g(Y_1, \theta), -K_0 \right) \right] < \infty. \quad (86)$$

Denote

$$\begin{aligned}\widehat{\theta} &= \arg \max_{\theta \in \Theta} N^{-1} \sum_{j=1}^N g(Y_j, \theta), \\ \theta_0 &= \arg \max_{\theta \in \Theta} \bar{g}(\theta).\end{aligned}$$

Then

$$p \lim_{N \rightarrow \infty} \left(\frac{1}{N} \sum_{j=1}^N g(Y_j, \widehat{\theta}) - \bar{g}(\widehat{\theta}) \right) = 0 \quad (87)$$

and consequently,

$$p \lim_{N \rightarrow \infty} \bar{g}(\widehat{\theta}) = \bar{g}(\theta_0). \quad (88)$$

Proof: 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 (85) and (86) replaced by

$$E \left[\sup_{\theta \in \Theta} |g(Y_1, \theta)| \right] < \infty \quad (89)$$

we have

$$\lim_{N \rightarrow \infty} \sup_{\theta \in \Theta} \left| \frac{1}{N} \sum_{j=1}^N g(Y_j, \theta) - \bar{g}(\theta) \right| = 0 \text{ a.s.} \quad (90)$$

However, the condition (89) is too difficult to verify in the log-likelihood case. Therefore I will use the weaker conditions (85) and (86).

Originally the uniform strong law (90) was derived for the case that Θ is a compact subset of a Euclidean space, but it is easy to verify from the proof 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

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

Then it follows from (90) with $g(Y_j, \theta)$ replaced by $\max(g(Y_j, \theta), -K)$ that

$$\lim_{N \rightarrow \infty} \sup_{\theta \in \Theta} \left| \frac{1}{N} \sum_{j=1}^N \max(g(Y_j, \theta), -K) - \bar{g}_K(\theta) \right| = 0 \text{ a.s.}, \quad (91)$$

where

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

As is well-known, (91) is equivalent to the statement that for all $\varepsilon > 0$,

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

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

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

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

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

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

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

hence, for arbitrary $\varepsilon > 0$,

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

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

$$\sup_{\theta \in \Theta} \left| \frac{1}{n_N} \sum_{j=1}^{n_N} \max(g(Y_j, \theta), -K_N) - \bar{g}_{K_N}(\theta) \right| \rightarrow 0 \text{ a.s.}, \quad (92)$$

and the same applies if we would have replaced N first by an arbitrary subsequence. Thus, every subsequence of N contains a further subsequence n_N such that (92) 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, (92) implies that there exists a sequence K_N converging to infinity with N such that

$$p \lim_{N \rightarrow \infty} \sup_{\theta \in \Theta} \left| \frac{1}{N} \sum_{j=1}^N \max(g(Y_j, \theta), -K_N) - \bar{g}_{K_N}(\theta) \right| = 0. \quad (93)$$

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

$$\begin{aligned} \bar{g}_K(\theta) &= E[\max(g(Y_j, \theta), -K)] \geq \max(E[g(Y_j, \theta)], -K) \\ &= \max(\bar{g}(\theta), -K) \geq \bar{g}(\theta) \end{aligned} \quad (94)$$

and similarly

$$\frac{1}{N} \sum_{j=1}^N \max(g(Y_j, \theta), -K) \geq \max\left(\frac{1}{N} \sum_{j=1}^N g(Y_j, \theta), -K\right) \geq \frac{1}{N} \sum_{j=1}^N g(Y_j, \theta). \quad (95)$$

It follows from (86), (94) and the dominated convergence theorem that

$$\lim_{K \rightarrow \infty} \sup_{\theta \in \Theta} |\bar{g}_K(\theta) - \bar{g}(\theta)| = 0,$$

hence (93) now becomes

$$p \lim_{N \rightarrow \infty} \sup_{\theta \in \Theta} \left| \frac{1}{N} \sum_{j=1}^N \max(g(Y_j, \theta), -K_N) - \bar{g}(\theta) \right| = 0. \quad (96)$$

Finally, observe from (95) that

$$\begin{aligned} &\frac{1}{N} \sum_{j=1}^N \max(g(Y_j, \hat{\theta}), -K_N) - \bar{g}(\hat{\theta}) \geq \frac{1}{N} \sum_{j=1}^N g(Y_j, \hat{\theta}) - \bar{g}(\hat{\theta}) \\ &\geq \frac{1}{N} \sum_{j=1}^N g(Y_j, \theta_0) - \bar{g}(\theta_0) + \bar{g}(\theta_0) - \bar{g}(\hat{\theta}) \\ &\geq \frac{1}{N} \sum_{j=1}^N g(Y_j, \theta_0) - \bar{g}(\theta_0) \end{aligned} \quad (97)$$

By Kolmogorov's strong law of large numbers, the lower bound in (97) converges a.s. to zero, and by (96) the upper bound in (97) converges in probability to zero, hence (87) holds, and so does (88). Q.E.D.

In the penalized log-likelihood case, let $g(Y_j, \theta) = \Psi(Y_j, \alpha, \beta, h)$, where the latter is defined by (74). Clearly, $H(\mu_i(\alpha, \beta'x))$ is continuous in $\alpha \in A$, $\beta \in B$ and all x , and H itself is uniformly continuous with respect to the metric (76):

$$\sup_{0 \leq u \leq 1} |H_1(u) - H_2(u)| \leq \|h_1 - h_2\|_1.$$

Moreover, the penalty term $-\Pi(h)$ in (74) 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 $Y_j = (D_{1,j}, \dots, D_{M,j}, X'_j)$. Then $\bar{\Psi}(\alpha, \beta, h)$ is also continuous on $A \times B \times \mathcal{D}(0, 1)$.

It is easy to verify from (74) that $\Psi(Y_j, \alpha, \beta, h) \leq 0$, hence condition (86) holds, and condition (85) holds if

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

Thus, under Assumptions 1-4, (78) is true.

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

Theorem 11. Let the conditions of Theorem 10 hold, and let $\Theta_N \subset \Theta$ be such that the computation of $\tilde{\theta} = \arg \max_{\theta \in \Theta_N} N^{-1} \sum_{j=1}^N g(Y_j, \theta)$ is feasible. If each Θ_N contains an element θ_N such that $\lim_{N \rightarrow \infty} d(\theta_N, \theta_0) = 0$, then $p \lim_{N \rightarrow \infty} \bar{g}(\tilde{\theta}) = \bar{g}(\theta_0)$. If θ_0 is unique then the latter implies that $p \lim_{N \rightarrow \infty} d(\tilde{\theta}, \theta_0) = 0$.

Proof: Similar to (97) we have,

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

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

It follows from (96) that

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

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

$$\frac{1}{N} \sum_{j=1}^N g(Y_j, \theta_0) \rightarrow \bar{g}(\theta_0) \text{ a.s.} \quad (98)$$

Moreover,

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

because of the continuity of $E [|g(Y_1, \theta) - g(Y_1, \theta_0)|]$ in θ and $\lim_{N \rightarrow \infty} d(\theta_N, \theta_0) = 0$. Hence by Chebishev's inequality,

$$p \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{j=1}^N (g(Y_j, \theta_N) - g(Y_j, \theta_0)) = 0. \quad (99)$$

Thus,

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

which by (98) and (99) implies that $p \lim_{N \rightarrow \infty} \bar{g}(\tilde{\theta}) = \bar{g}(\theta_0)$.

If θ_0 is unique then by the continuity of $\bar{g}(\theta)$ there exists a $\bar{\delta} > 0$ such that for all $\delta \in (0, \bar{\delta}]$,

$$\sup_{\theta \in \Theta, d(\theta, \theta_0) \geq \delta} \bar{g}(\theta) < \bar{g}(\theta_0).$$

See, for example, Bierens (2004, Appendix II, Theorem II.6). Thus,

$$P \left[d(\tilde{\theta}, \theta_0) \geq \delta \right] \leq P \left[\bar{g}(\tilde{\theta}) \leq \sup_{\theta \in \Theta, d(\theta, \theta_0) \geq \delta} \bar{g}(\theta) \right] \rightarrow 0.$$

Q.E.D.

We can now formulate the consistency results for the sieve ML estimators of the parameters of the bracketed SNP-MPH model and the SNPOP model:

Theorem 12. *Let α_0, β_0, h_0 be the true parameters of the bracketed SNP-MPH model or the SNPOP model. Let $L_N^*(\alpha, \beta, h)$ be the penalized likelihood function, and let*

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

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

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

9 Discrete covariates

As admitted before, Assumption 2 is often violated in practice. To see what the problem is if the covariates are discrete, consider the SNP probability model (67) with $M = 1$:

$$\begin{aligned} P[Y = 0|X] &= F_0(\alpha_0 + \beta_0'X) = 1 - H_0(G(\alpha_0 + \beta_0'X)), \\ P[Y = 1|X] &= 1 - F_0(\alpha_0 + \beta_0'X) = H_0(G(\alpha_0 + \beta_0'X)). \end{aligned}$$

If G is the logistic distribution function (73) this model is a SNP-Logit model.

Now suppose that X is a single random variable with finite support, say $\{0, 1, \dots, K - 1\}$ and that $\beta_0 \neq 0$. Then $G(\alpha_0 + \beta_0 X)$ takes only K values

$$u_{0,i} = G(\alpha_0 + \beta_0(i - 1)), \quad i = 1, \dots, K,$$

with corresponding conditional probabilities $P[Y = 1|X = i - 1]$,

$$v_{0,i} = H_0(u_{0,i}) = P[Y = 1|X = i - 1], \quad i = 1, \dots, K.$$

There exist uncountable many distribution functions $H_0(u)$ on $[0, 1]$ that fit through the points $(u_{0,1}, v_{0,1}), \dots, (u_{0,K}, v_{0,K})$. Similarly, denoting, $u_i = G(\alpha + \beta(i - 1))$, $i = 1, \dots, K$, where α and β are arbitrary, except that $\beta \neq 0$, there exist uncountable many distribution functions H on $[0, 1]$ that fit through the points $(u_1, v_{0,1}), \dots, (u_K, v_{0,K})$, hence there are uncountable many pairs of distributions functions H_0 and H such that

$$H_0(G(\alpha_0 + \beta_0 X)) = H(G(\alpha + \beta X)) \quad \text{a.s.} \quad (100)$$

Clearly, the conditions (70) and (71) are not sufficient to solve this problem.

The only solution is to make parametric assumptions such that for distribution functions H_0 and H belonging to a particular parametric family \mathcal{H} , (100) implies $\alpha = \alpha_0$, $\beta = \beta_0$ and $H_0 = H$.

The distribution functions of the type (16) with not too large a fixed n may then still be useful as a specification of a (flexible) parametric family \mathcal{H} . Thus, in this case we need to **assume** that for given n , $H_0(u) = H_n(u|\delta_0)$, and that n is small enough to guarantee the identification of α_0 , β_0 and δ_0 . It is advisable though to implement the moment conditions (70) and (71) in the SNPOP case by penalizing the log-likelihood similar to (66), in order to strengthen the identification. The same applies to the SNP-MPH model.

10 An empirical application

To be done.

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.
- 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.

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.

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

Young, N. (1988), *An Introduction to Hilbert Space*, Cambridge, UK: Cambridge University Press.