NONPARAMETRIC ESTIMATION IN THE COX MODEL

by

Finbarr O'Sullivan

TECHNICAL REPORT No. 169
August 1989

Department of Statistics, GN-22
University of Washington
Seattle, Washington 98195 USA
Nonparametric Estimation in the Cox Model

Finbarr O'Sullivan
Department of Biostatistics and Statistics
University of Washington
Seattle, WA 98195.

ABSTRACT

Nonparametric estimation of the relative risk in a generalized Cox model with multivariate time dependent covariates is considered. Estimation is based on a penalized partial likelihood. Asymptotic properties are analyzed by applying the technique of Cox and O’Sullivan[4]. Martingale representations of the kind employed by Andersen and Gill[2] play an important role. Upper bounds on rates of convergence in a variety of norms are obtained. These upper bounds match the optimal rates available for linear non-parametric regression and density estimation. The results are uniform in the smoothing parameter, this is an important step for the analysis of data dependent rules for the selection of the smoothing parameter.


Key words and phrases. Cox Model, Counting Process Formulation, Martingales, Penalized Partial Likelihood, Rates of Convergence, Relative Risk, Uniformity.

Running Head: Nonparametric Estimation in the Cox Model

February 27, 1989

1 Research supported by the National Science Foundation under Grant No. MCS-840-3239, a Regents Fellowship from the University of California at Berkeley and the Department of Energy under Grant No. DE-FG06-85ER25006.
1.1. Counting Process Formulation of the Cox Model

For each \( n \), \( N^{(n)} = (N_1^{(n)}, \ldots, N_k^{(n)}) \) is a multivariate counting process with a random intensity process \( \lambda^{(n)} = (\lambda_1^{(n)}, \ldots, \lambda_d^{(n)}) \) for which

\[
\lambda_i^{(n)}(t) = y_i^{(n)}(t) \cdot \exp(\theta_0(x_i^{(n)}(t))) \cdot \lambda_0(t) .
\]  

(1.2)

The underlying baseline hazard \( \lambda_0 \) and the relative risk \( \theta_0 : \mathbb{R}^d \to \mathbb{R} \) are fixed unknown quantities. Throughout this the time index \( t \) takes values on a bounded interval which without loss of generality is taken to be \([0,1]\). A family of right continuous non-decreasing sub \( \sigma \)-algebras \( \{ F_t^{(n)} : t \in [0,1] \} \) are defined on the \( n \)th sample space, with \( F_t^{(n)} \) representing the history of the \( n \)th process up to time \( t \). All processes are adapted to this family of \( \sigma \)-algebras. \( y_i^{(n)}(\cdot) \) is a predictable process taking values in \( \{0, 1\} \); \( y_i^{(n)}(t) = 1 \) if the \( i \)'th component of the process is under observation just before time \( t \), \( y_i^{(n)}(t) = 0 \) otherwise, see [8] for further discussion. The \( d \)-dimensional covariate process \( x_i^{(n)}(\cdot) \) is predictable and locally bounded, in fact we will assume that the collection of covariate processes take values in a fixed bounded subset of \( \mathbb{R}^d \).

Specification of \( \lambda^{(n)} \) as an intensity process means that the process

\[
M_i^{(n)}(t) = N_i^{(n)}(t) - \int_0^t \lambda_i^{(n)}(\tau) d\tau , \quad i=1,2,\ldots,n \quad \text{and} \quad t \in [0,1] ,
\]  

(1.3)

is a local martingale with mean zero, \( EM_i^{(n)}(t) = 0 \). The predictable covariation of \( M^{(n)} \) is given by

\[
<M_i^{(n)}, M_j^{(n)}>(t) = \int_0^t \lambda_i^{(n)}(\tau) d\tau \quad ; \quad <M_i^{(n)}, M_j^{(n)}> = 0 \quad i \neq j .
\]  

(1.4)

For ease of notation the superscript \( (n) \) will be dropped from \( N, \lambda, M, y, \) and \( x \).

1.2. Definition of the Penalized Partial Likelihood and Some Assumptions

Inferences for the relative risk \( \theta_0 \) will be based on the penalized partial likelihood functional

\[
l_{p\mu}(\theta) = l_n(\theta) + \mu J(\theta) , \quad \mu > 0 .
\]  

(1.5)
For each \( t \in [0,1] \), the random variable \( x(t) \) has density \( h(\cdot | t) \). Let

\[
p(t,x) = P \left[ y(t) = 1 \mid x(t) = x \right],
\]

and let \( q(x,t) = p(x,t)h(x | t) \). We suppose that there are strictly positive constants, \( k_1 \) and \( k_2 \) (independent of \( x \) and \( t \)), such that

\[
k_1 < q(x,t) < k_2 \quad \text{and} \quad \frac{\partial}{\partial t} q(x,t) < k_2.
\]

(iv) \( \{x(t), t \in [0,1]\} \subseteq X \subseteq \mathbb{R}^d \) where \( X \) is a bounded open simply connected set with \( C^\infty \) boundary (see Definition 3.2.1.2. of Triebel[14]).

Assumptions (i) and (ii) should be familiar, (iii) is used used in proving results concerning derivatives of a continuous version of the partial likelihood and (iv) is used is to obtain growth behavior on eigenvalues which in turn determine the ultimate rates of convergence. For the assumptions on the parameter space, let \( W^k_2(X,\mathbb{R}) \) denote the Sobolev space of real valued \( L_2 \) functions defined on \( X \) whose \( k \)'th derivative is square integrable, see Adams[1]. Sobolev spaces may also be defined for non-integer \( k \) (again see Adams[1]) and throughout this paper the order \( k \) can assume any positive real value. \( W^k_2(X,\mathbb{R}) \) is the subspace of \( W^k_2(X,\mathbb{R}) \) which consists of functions with mean zero. Thus \( \int_X \theta(x) dx = 0 \), for all \( \theta \in W^k_2(X,\mathbb{R}) \).

Assumption B. (Parameter Space)

(i) \( \Theta \) is a Hilbert space of functions \( \theta : X \rightarrow \mathbb{R} \) with inner product \( <\cdot,\cdot> \) and norm \( ||\cdot|| \). The elements of \( \Theta \) are constrained to have mean zero.

(ii) For some \( m > 3d/2 \), \( \Theta = W^m_2(X,\mathbb{R}) \) (meaning the spaces are equal as sets and they have equivalent norms). The true function parameter \( \theta_0 \) is in \( W^m_p \) for some \( p > 3d/2m \).

(iii) The penalty functional \( J(\theta) = \frac{1}{2} <\theta, W \theta> \) where \( W \) is a bounded linear operator on \( \Theta \), which is self-adjoint and nonnegative definite. and there are positive constants \( k_1, k_2 \) such that
\[ \| \theta_{n \mu_n} - \theta_{\mu_n} \|_2^2 \leq M \, n^{-1 \mu_n^{(b+d)/2m}} \quad \text{and} \quad \sup_{\mu \in [\mu_n, \mu_0]} \left\{ \frac{\| \theta_{n \mu} - \theta_{\mu} \|_2^2}{n^{-1 \mu^{-1}(b+d)/2m}} \right\} \leq M \log(\mu_n^{-1}) \]

occurs with probability approaching unity as \( n \to \infty \).

Proof: The results follow from Theorem 3.1 and 3.3 of §3. \( \square \)

1.4. Discussion and Outline of the Paper

The first part of the Theorem gives the order of the systematic error and the second part gives the order of the stochastic error. From the theorem it follows that if \( \mu_n^* \) is \( O(n^{-2m/(2mp+d)}) \), then \( \| \theta_{n \mu_n^*} - \theta_{\mu_n^*} \|_2^2 \) is bounded by \( O_p(n^{-2m(p-b)/(2mp+d)}) \). In particular, if \( \theta_0 \in W^2 \) (\( i.e., \ p = 1 \)), the integrated square error \( \| \theta_{n \mu_n^*} - \theta_{\mu_n^*} \|_2^2 \) is bounded by \( O_p(n^{-2m/(2mp+d)}) \). Thus in the analogue of the standard 1-dimensional cubic smoothing spline setup (\( m = 2, d = 1 \) and \( p = 1 \)) the integrated squared error converges we get the familiar rate of \( O_p(n^{-4/5}) \). Uniformity in the smoothing parameter is noteworthy. We expect that this will be most useful in the analysis of data dependent rules for choosing the smoothing parameter. Theorem 1 yields upper bounds on the rates of convergence. These upper bounds match the optimal rates of convergence obtained by Stone[12] for non-parametric regression and density estimation in Holder spaces. We conjecture that the rates obtained are optimal.

The paper represents an application of a general approach based on Taylor series expansions to the asymptotic analysis of penalized likelihood-type estimators developed in Cox and O'Sullivan[4] (hereafter abbreviated CO[4]). The theoretical framework is elaborated in §2, and some useful bounds on derivatives of the partial likelihood are obtained. The main results are proved in §3. The uniformity result requires a strengthening of a Theorem in CO[4], this result is proved in the appendix along with a more technical lemma used in obtaining bounds on derivatives.
third order Frechet derivatives, of \( s_n(\theta, t) \) are given by

\[
D_s n(\theta, t) \phi = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i(t)) y_i(t) e^{\theta(x_i(t))}
\]

\[
D^2 s_n(\theta, t) \phi \psi = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i(t)) \psi(x_i(t)) y_i(t) e^{\theta(x_i(t))}
\]

\[
D^3 s_n(\theta, t) \phi \psi \xi = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i(t)) \psi(x_i(t)) \xi(x_i(t)) y_i(t) e^{\theta(x_i(t))}
\]

where \( \phi, \psi, \xi \) are functions. Since evaluations is a continuous linear functional in \( W_2^{m, n}(X; R) \) the above quantities are the Frechet derivatives of \( s_n(\theta, t) \) in \( W_2^{m, n}(X; R) \). We note that the Frechet derivative is the function space generalization of the total derivative used in standard multivariate calculus, see Rall[11] for example. The continuity of the derivatives follows since \( e^{\theta(x)} \) is continuous in \( \theta \) for \( \theta \in S(R, \alpha) \) because \( \alpha > d/2m \).

Fixing \( \theta \in W_2^{m, n} \) by the strong law of large numbers \( s_n(\theta, t) \to s(\theta, t) \) almost surely where

\[
s(\theta, t) = E\left[ y(t) e^{\theta(t)} \right] = \int e^{\theta(x)} q(x, t) dx.
\]

The Frechet derivatives of \( s(\theta, t) \) are given by

\[
D s(\theta, t) \phi = \int \phi(x) e^{\theta(x)} q(x, t) dx
\]

\[
D^2 s(\theta, t) \phi \psi = \int \phi(x) \psi(x) e^{\theta(x)} q(x, t) dx
\]

\[
D^3 s(\theta, t) \phi \psi \xi = \int \phi(x) \psi(x) \xi(x) e^{\theta(x)} q(x, t) dx
\]

The derivatives are clearly continuous. The following Lemma will be useful.

**Lemma 2.1.** For any \( R > 0 \) and \( \alpha > d/2m \) there are constants \( 0 < m_R \leq M_R < \infty \) such that for all \( \theta, \theta* \in S(R, \alpha) \) and \( t \in [0, 1] \).

(a) \( m_R < s_n(\theta, t)/s_n(\theta*, t) < M_R \)

(b) \( m_R < s(\theta, t) < M_R \) and \( s_n(\theta, t) < M_R \)

(c) if \( \phi, \psi \in \Theta \) then \( \{D_s(\theta, t) \phi\}^2 \leq M_R \|\phi\|\|^2, \) and \( \{D^2 s(\theta, t) \phi \psi\}^2 \leq M_R \|\phi\|\|\psi\|\|^2, \)

**Proof:** (a) follows from the definition of \( s_n \), A(iv) and the uniform boundedness of \( \theta \) and \( \theta* \) (they are both elements of \( S(R, \alpha) \) and \( \alpha > d/2m \)). (b) follows from Assumptions A(ii), A(iii) and
\(\theta_0\) is the global minimizer of \(l(\theta)\) over \(W_{\mathbb{R}^d}(X; R)\). This result along with existence of continuous derivatives up to order three guarantees that Assumption A.3 of CO[4] holds.

The first and second order Frechet derivatives of \(l_n\) are given by

\[
Dl_n(\theta)\phi = \frac{1}{n} \int_0 t \frac{D_s \phi}{s_n(\theta,t)} d\bar{N}(t) - \frac{1}{n} \sum_{i=1}^n \phi(x_i(t)) dN_i(t)
\]

\[
D^2 l_n(\theta)\psi = \frac{1}{n} \int_0 \left( \frac{D^2 s \phi}{s_n(\theta,t)} - \frac{D s \phi}{s_n(\theta,t)} \frac{D s \psi}{s_n(\theta,t)} \right) d\bar{N}(t).
\]  \quad (2.11)

Again, with \(p_i(t) = \left[ \frac{1}{n} y_i(t) e^{\phi(x_i(t))} \right] s_n(\theta,t) \) for \(i=1,2, \ldots, n\), and \(\bar{\phi}_i = \sum_{i=1}^n \phi(x_i(t)) p_i(t)\),

\[
\frac{D^2 s \phi}{s_n(\theta,t)} - \frac{D s \phi}{s_n(\theta,t)} \frac{D s \psi}{s_n(\theta,t)} = \sum_{i=1}^n p_i(t) \left[ \phi(x_i(t)) - \bar{\phi}_i \right]^2,
\]  \quad (2.12)

and it follows that \(l_n\) and \(l_{n\mu}\) are convex. A straightforward argument, along the lines given in the appendix of [9], shows that the penalized partial likelihood estimator must lie in the subspace \(\Theta_n = N(W) + \text{Span} \{ \xi(x_i(t_i)) \}\), where \(\xi(x)\) is the Riesz representer of evaluation at \(x\) and \(N(W)\) is the null space of the linear operator \(W\). \(\text{Span}\) stands for the span for the given set where \(t_{ij}\) ranges over the event times of the counting process \(N_i(t)\) for \(t \in [0,1]\) and \(i = 1,2, \ldots, n\). If \(N(W)\) is finite then \(\Theta_n\) is a finite dimensional space, although its dimension will in general be larger than \(n\). When \(W\) corresponds to the usual Laplacian penalty functional used to generate thin plate smoothing splines[9, 15], the penalized partial likelihood estimator can be represented as a generalized Laplacian smoothing spline. Following the argument in[9] a sufficient condition for the existence of a unique minimizer of the penalized partial likelihood in (2.7) is that there exist a unique minimizer of the negative logarithm of the partial likelihood over \(N(W)\). These results are summarized in the following theorem.

**Theorem 2.2.** Under Assumptions A and B, \(\theta_0\) is the unique root of \(Dl(\theta)\) in \(W_{\mathbb{R}^d}\) for \(\alpha > d/2m\). If the null space of \(W\) is finite, the minimizer of the \(l_{n\mu}\) must lie in the finite dimensional subspace, \(\Theta_n\). A sufficient condition for the existence of a unique minimizer of \(l_{n\mu}\) is that
It is easily shown using standard interpolation theory, that $\Theta_b$ is equivalent to $W^{b}_{b}(X;R)$ for $b \in [0,1]$. We denote $W^{b}_{b}(X;R)$ by $\Theta_b$ from here on. The operators $U$ and $W$ extend to linear operators on $\Theta_b$ for $b \in [0,1]$ (see equation (2.6) of CO[4]).

Let $U(\theta)$ be defined by
\[
\langle \psi, U(\theta) \phi \rangle = D^{2l}(\theta) \phi \psi \quad (2.17)
\]

By definition of the derivative $U(\theta)$ is a bounded linear operator on $\Theta_\alpha$ since $\alpha > d/2m$. From (2.10) we have the next Lemma which implies assumption A.4 of CO[4].

**Lemma 2.3.** There are constants $0 < c_1 \leq c_2 < \infty$, such that for all $\theta \in S(R,\alpha)$
\[
c_1 \langle \theta, U \theta \rangle \leq \langle \theta, U(\theta^*) \theta \rangle \leq c_2 \langle \theta, U \theta \rangle
\]
for all $\theta \in \Theta$.

Replacing $U$ by $U(\theta^*)$ in equation (2.11) we obtain for each $\theta^* \in N_{b}$ sequences of eigenvalues $\{\gamma_v ; v = 1,2, \cdots \}$ and corresponding eigenfunctions $\{\phi_v ; v = 1,2, \cdots \}$. This leads to a norm
\[
\|\theta\|_{b^*} = (\sum_{v=1}^{\infty} [1 + \gamma_v^2] \langle \theta, U(\theta^*) \phi_v \rangle^2)^{1/2}, \quad (2.18)
\]
and corresponding Hilbert space $\Theta_{b^*}$. These spaces are uniformly equivalent for $\theta^* \in N_{b^*}$.

$U(\theta^*)$ extends to a bounded linear operator on $\Theta_{b^*}$ for $b \in [0,1]$ and the linear operator
\[
G_{\mu}(\theta^*) = U(\theta^*) + \mu W
\]
is bounded and invertible on $\Theta_{b^*}$. From Lemma 2.1 of CO[4] we have

**Lemma 2.4.** For $R > 0$ and $\theta^* \in S(R,\alpha)$, $b > 0$ and $v = 1,2, \cdots$

(i) \[
\|\phi_v\|_{b^*}^2 = 1 + \gamma_v^2 \quad \text{and} \quad \|\phi_v\|_{b^*}^2 = 1 + \gamma_v^2
\]
Lemma 2.5.

There is a constant $0 < M < \infty$ and a random variable $A_n = O_p(1)$ such that for

$b \leq 2 - \alpha - d/2m$

(i) $K_3(\mu, b) \leq M \mu^{-(b+d/2m)}$

(ii) $K_{2n}(\mu, b) \leq A_n M n^{-1} \mu^{-(b+d/2m)} \mu^{-\alpha}$

(iii) $K_{3n}(\mu, b) \leq A_n M \mu^{-(b+d/2m)} (1 + n^{-1} \mu^{-\alpha})$

Both $M$ and $A_n$ are independent of $\mu$ and $b$.

Proof: The results follow from Lemma A.1 in the appendix because $N_{\theta_0} \subset S(R, \alpha)$ for some $R$.

We consider part (iii) which illustrates the technicalities. Let $\theta_* = \theta_1$ and $\theta = \theta_2$

$$
\|G_{\mu}(\theta_*)^{-1}D^3 l_\mu(\theta)uv\|_b \leq M \|G_{\mu}(\theta_*)^{-1}D^3 l_\mu(\theta)uv\|_{\mu^2} \tag{2.21}
= M \sum_{\gamma=1}^\infty [1 + \gamma \gamma] \|D^3 l_\mu(\theta)uv\|_{\mu^2}^2
$$

But from Lemma A.1(iii)

$$
\leq A_n M \|u\|_{\hat{a}_{\gamma}} \|v\|_{\hat{a}_{\gamma}} \sum_{\gamma=1}^\infty [1 + \gamma \gamma] \|D^3 l_\mu(\theta)uv\|_{\mu^2}^2
$$

and the from Lemma 2.4 (i) and (iii) (the condition that $b < 2 - d/2m - \alpha$ is used here)

$$
\leq A_n M \|u\|_{\hat{a}_{\gamma}} \|v\|_{\hat{a}_{\gamma}} \mu^{-(b+d/2m)} (1 + n^{-1} \mu^{-\alpha}) \tag{2.22}
$$

This proves part (iii). The arguments for parts (i) and (ii) are very similar, just replace the application of Lemma A.1 (iii) by Lemma A.1 (i) and (iii) respectively. □
\[ 2\alpha < p - d/2m, \ r(\mu, \alpha) \to 0 \text{ as } \mu \to 0. \] Thus Assumption A.5 of CO[4] holds and form Theorem 3.1 of CO[4]. The conclusion of the theorem follows. \[ \square \]

Before considering the stochastic component of the error we analyze \( \| \vec{\theta}_{n\mu} - \theta_{\mu} \|_b \) for \( b \leq \alpha \).

Lemma 3.2.

Let \( \mu_n \) is any sequence tending to zero such that for some \( \alpha \) satisfying \( d/2m < \alpha < (p - d/2m)/2 \), \n^{-1}\mu^{-2(\alpha + d/2m)} \to 0. \) Let \( 0 \leq b \leq \min(2 - \alpha - d/2m, \alpha) \) and let \( \mu_{*n} \) be a deterministic sequence in \([\mu_n, \mu_0]\). We have

\[ \| \theta_{n\mu_n} - \theta_{\mu_n} \|_b \leq O_p (n^{-1}\mu_n^{-2(b+d/2m)}) \quad \text{and} \quad \sup_{\mu \in [\mu_n, \mu_0]} \left\{ \frac{\| \theta_{n\mu} - \theta_{\mu} \|_b}{n^{-1}\mu_n^{-2(b+d/2m)}} \right\} \leq O_p (\log(\mu_n^{-1})) \]

Proof: By first order Taylor series expansion about \( \theta_0 \) (see Rall[11])

\[ \bar{\theta}_{n\mu} - \theta_{\mu} = G_{\mu}(\theta_0)^{-1}[Dl_n(\theta_0) - Dl(\theta_0)] + \frac{1}{\theta}{G_{\mu}(\theta_0)^{-1}[D^2l_n(\theta_0) - D^2l(\theta_0)]}(\theta_{\mu} - \theta_0)sds \] \tag{3.5}

where \( \theta_s = \theta_0 + s(\theta_{\mu} - \theta_0) \). Thus

\[ \| \bar{\theta}_{n\mu} - \theta_{\mu} \|_b \leq \| G_{\mu}(\theta_0)^{-1}[Dl_n(\theta_0) - Dl(\theta_0)] \|_b + \frac{1}{2}K_{2n}(\mu, b)^2/\| \theta_{\mu} - \theta_0 \|_b \] \tag{3.6}

From Theorem 3.1 and Lemma 2.5 (iii) the second term is bounded above by \( A_n \ n^{-1}\mu^{-2(b+d/2m)} \mu^{-\alpha} \mu^{-\alpha} \) where \( A_n = O_p(1) \). Since \( \alpha < (p - d/2m)/2 \) and \( b \leq \alpha \), this is less than \( A_n \ n^{-1}\mu^{-2(b+d/2m)} \). To prove the theorem we need to develop bounds for the first term.

Let \( R_{\mu} = G_{\mu}(\theta_0)^{-1}[D^2l(\theta_{\mu}) - D^2l(\theta_0)] \) and \( x_n(\mu) = G_{\mu}(\theta_0)^{-1}[Dl_n(\theta_0) - Dl(\theta_0)] \) so

\[ G_{\mu}(\theta_0)^{-1}[Dl_n(\theta_0) - Dl(\theta_0)] = [I + R_{\mu}]^{-1}x_n(\mu) \] \tag{3.7}

\( R_{\mu} \) is clearly a bounded linear operator on \( \Theta_{\alpha} \). Let

\[ 1R_{\mu}1_{\alpha, b} = \sup_{\| \theta_{\mu} \|_b = 1} \| R_{\mu} \theta \|_b \] \tag{3.8}

By Taylor series expansion, \( R_{\mu} = \frac{1}{\theta}{G_{\mu}(\theta_0)^{-1}[D^3l(\theta_s)(\theta_{\mu} - \theta_0)sds} \) where
The expected values of the squares of the last three terms are easily computed and shown to be bounded by $M^{-1}/2\theta_0$. For the first term

$$E\left[\frac{D_s(\theta_0, t)\phi_{ov}}{s(\theta_0, t)} - \frac{D_s(\theta_0, t)\phi_{ov}}{s(\theta_0, t)}d\bar{N}(t)\right]^2$$

Applying arguments used in Lemma A.1(i) in the appendix gives bounds of $O(n^{-1}/2\theta_0)$ for the expectations of both terms on the right hand side of the latter expression. Thus

$$EB^{(n)} \leq M^{-1}/2\theta_0$$

where $M$ does not depend on $n$ or $v$. The first part of (3.12) now follows directly by Markov's inequality because

$$E|x_n(\mu)| \leq \sum_{v=1}^{\infty} [1 + \bar{\gamma}_v]^b [1 + \bar{\mu}\bar{\gamma}_v]^{-2}EB^{(n)}$$

$$\leq M^{-1} \sum_{v=1}^{\infty} [1 + \bar{\gamma}_v]^b [1 + \bar{\mu}\bar{\gamma}_v]^{-2} \leq M^{-1} \mu^{-\left(b+2d/2m\right)}$$

as $\mu \to 0$. The last inequality comes from Lemma 2.4(iii). For the second part of (3.12) consider

$$h_n(\mu) = \frac{\sum_{v=1}^{\infty} [1 + \bar{\gamma}_v]^b [1 + \bar{\mu}\bar{\gamma}_v]^{-2}B^{(n)}}{n^{-1}\mu^{-\left(b+2d/2m\right)}}$$

Clearly

$$\sup_{\mu \in [\mu_a, \mu_2]} |h_n(\mu)| \leq |h_n(\mu_a)| + \int |h_n'(\mu)| d\mu$$

The condition on $\alpha$ and the previous argument implies $|h_n(\mu_n)| \leq M O_p(1)$. For the second
This tends to zero in probability. Using Lemma 2.5

$$\sup_{\mu \in [\mu_n, \mu_0]} r_n(\mu, \alpha)^2 \leq O_P(n^{-1} \mu_n^{-2(2\alpha+d/2m)} + \log(\mu_n^{-1} \mu_n^{-2(\alpha+d/2m)})) \xrightarrow{P} 0 \quad (3.22)$$

Thus Assumption A.6' in the appendix is satisfied and so by Theorem A.2 the event that $\theta_{n\mu}$ uniquely exists for all $\mu \in [\mu_n, \mu_0]$ with

$$\|\theta_{n\mu} - \overline{\theta}_{n\mu}\|/b \leq r_n(\mu, b)d_n(\mu, \alpha)$$

occurs with probability approaching unity as $n \to \infty$. On this event

$$\|\theta_{n\mu} - \theta_{\mu}\|/b \leq \|\theta_{n\mu} - \overline{\theta}_{n\mu}\|/b + \|\theta_{n\mu} - \overline{\theta}_{n\mu}\|/b$$

$$\leq d_n(\mu, b) + r_n(\mu, b)d_n(\mu, \alpha) \quad (3.23)$$

Using Lemma 3.2

$$\sup_{\mu \in [\mu_n, \mu_0]} \left\{ \frac{r_n(\mu, b)^2 d_n(\mu, \alpha)^2}{n^{-1} \mu_n^{-1} \mu_n^{-2(\alpha+d/2m)}} \right\} \leq O_P(n^{-1} \log(\mu_n^{-1} \mu_n^{-2(\alpha+d/2m)})) \xrightarrow{P} 0 \quad (3.24)$$

Thus the convergence rate of $\|\theta_{n\mu} - \theta_{\mu}\|/b$ is determined by the convergence of $d_n(\mu, b)$. From here the results follow by Lemma 3.2. $\square$. 

Each term is this expression is analyzed separately, fortunately the analysis is very similar for the different terms. For the first term, direct application of the Cauchy-Schwartz inequality and Lemma 2.2 gives

\[
\left( \int \frac{D^3 s(\theta, t) u \nu w}{s(\theta, t)} \right)^2 s(\theta_0, t) \lambda_0(t) dt \leq M_R \left( \int [\mu(x) \nu(x) e^{\alpha \cdot x} dx] \right)^2 \lambda_0(t) dt \\
\leq M_R \sup_{x \in X} |u(x)|^2 \sup_{x \in X} |\nu(x)|^2 \|w\| \|L_2\| \\
\leq M_R \|u\|_2^2 \|\nu\|_2^2 \|w\|_2^2 
\]  
(A.3)

The last inequality follows from Sobolev’s Imbedding Theorem because \( \alpha > d/2m \). Similar analysis of the other terms leads to the bound in (i).

For part (ii),

\[
D^2 l_n(\theta) u \nu - D^2 l(\theta) u \nu 
\]
\[
\frac{D^2 s_n(\theta, t) u \nu}{s_n(\theta, t)} - \frac{D s_n(\theta, t) u}{s(\theta, t)} \frac{D s_n(\theta, t) \nu}{s(\theta, t)} dN(t) 
\]
\[
\frac{D^2 s_n(\theta, t) u \nu}{s(\theta, t)} \frac{D s_n(\theta, t) \nu}{s(\theta, t)} s(\theta_0, t) \lambda_0(t) dt
\]

We analyze a representative term

\[
\int D^2 s_n(\theta, t) u \nu \frac{D s_n(\theta, t) \nu}{s(\theta, t)} dN(t) - \int D^2 s_n(\theta, t) u \nu s(\theta_0, t) \lambda_0(t) dt 
\]
\[
= \int D^2 s_n(\theta, t) u \nu \frac{D s_n(\theta, t) \nu}{s(\theta, t)} dN(t) + \int D^2 s_n(\theta, t) u \nu \frac{[D s_n(\theta, t) - s(\theta_0, t) \lambda_0(t)] dN(t)}{s(\theta_0, t) \lambda_0(t)} dt
\]

Here write

\[
\int D^2 s_n(\theta, t) u \nu - D^2 s_n(\theta, t) u \nu \frac{D s_n(\theta, t) \nu}{s(\theta, t)} dN(t) 
\]
\[
= \int D^2 s_n(\theta, t) u \nu - D^2 s_n(\theta, t) u \nu \frac{D s_n(\theta, t) \nu}{s(\theta, t)} dN(t) + \int \frac{D^2 s_n(\theta, t) u \nu [s_n(\theta, t) - s(\theta_0, t)]}{s_n(\theta, t) \lambda_0(t)} dN(t)
\]

Let \( \psi(x) = e^{\alpha \cdot x} u(x) v(x) \). \( \psi \in W_2^{m}; \|\psi\|_2 = \|e^{\alpha \cdot x} u\|_2 \leq \|e^{\alpha \cdot x} u\|_2 \|u\|_2 \|v\|_2 \). For \( \alpha > d/2m \) (see appendix of CO(4)) and by series representation of the exponential \( \|e^{\alpha \cdot x}\|_2 \leq M_R e^{\|\alpha\|_2} \). Let

\[
\psi(x) = \sum_{i} \psi_i \phi_i(x) \quad \text{where} \quad \psi_i = \int \psi(x) \phi_i(x) dx. \quad \text{Also} \quad r_n(t) \phi_i = \frac{1}{n} \sum_{i=1}^{n} \psi_i(t) \phi_i(x_i(t)) \quad \text{and}
\]
\( \sup D^2 s(\theta *, t) uv \leq M_R \| u \|_2 \| v \|_2 n^{-1} \), so by repeating the previous argument leading to (A.9) we obtain an upper bound of \( \| u \|_2 \| v \|_2 n^{-1} M_R A_n \) for (A.10).

To deal with the remaining term in (A.5) let \( g(t) = \frac{D^2 s(\theta *, t) uv}{s(\theta *, t)} \). From assumption A(iii),

\[
\| g \|_2^2 + \| \frac{dg}{dt} \|_2^2 \leq M_R \| u \|_2^2 \| v \|_2 n \quad (A.11)
\]

Therefore \( g \in W_2^1[0,1] \). Thus \( g \) has the representation \( g(t) = \sum_v g_v b_v(t) \) where \( b_v \) are \( L_2[0,1] \)-orthonormal functions and \( \| g \|_2^2 = \sum_v [1+v]^2 g_v^2 \). Substituting the series expansion for \( g \) and applying the Cauchy-Schwartz inequality to the sum over \( v \) (as in the argument before equation (A.7)) gives

\[
\int \frac{D^2 s(\theta *, t) uv}{s(\theta *, t)} [dN(t) - s(\theta_0, t) \lambda_0(t) dt]^2
\leq (\sum_v [1+v]^2 g_v^2) \cdot (\sum_v [1+v]^{-2} \int [b_v(t) [dN(t) - s(\theta_0, t) \lambda_0(t) dt]^2]
\]

But

\[
E[\int b_v(t) [dN(t) - s(\theta_0, t) \lambda_0(t) dt]^2]
\leq E[\int b_v(t) (dN(t))^2] + E[\int b_v(t) [s(\theta_0, t) - s(\theta_0, t) \lambda_0(t) dt]^2]
\leq Mn^{-1} \| b_v \|_{L_2}^2 = Mn^{-1}
\]

Combining results gives the required bound for (A.5). Similar arguments are applied for the other part of (A.4) and this gives the bound in part (ii). The proof of part (iii) uses techniques already encountered in the analysis of parts (i) and (ii). □

**Uniform Linearization Result**

A generalization of a linearization result in CO[4] for penalized likelihood estimators is now proved. Consider a penalized likelihood functional \( I_{n\mu} \) defined on a real Hilbert space \( \Theta \) by

\[
I_{n\mu}(\theta) = I_n(\theta) + \frac{\mu}{2} <\theta, W \theta >
\quad (A.14)
\]
\[ m < 1 \text{ so that } A.6' \text{ holds. Now we restrict to the event } E_?^n. \text{ Here } S_{a_\alpha}(t_{n\mu}, \alpha) \subset N_{a_\alpha} \text{ for all } n > n_0 \]

and \( \sup_{\mu \in [\mu_n, \mu_0]} r_\alpha(\mu, \alpha) < 1/2. \) Let \( F_{n\mu}(\phi) = \phi - [U(\theta_\mu) + \mu W]^{-1}Z_{n\mu}(\theta_\mu + \phi) \) for \( \phi \in \Theta_\alpha. \) Repeating

the computations in Theorem 3.2 of CO[4] gives

\[
\|F_{n\mu}(\phi)\|_{\alpha} \leq [r_\alpha(\mu, \alpha) + \frac{1}{2}] t_{n\mu} \leq [\frac{1}{2} + \frac{1}{2}] t_{n\mu} = t_{n\mu} \tag{A.18}
\]

and

\[
\|F_{n\mu}(\phi_1) - F_{n\mu}(\phi_2)\|_{b} \leq \frac{1}{2} \|\phi_1 - \phi_2\|_{b} \tag{A.19}
\]

which hold for all \( \mu \in [\mu_n, \mu_0]. \) Thus \( F_{n\mu} \) is a contraction on \( S_{a_\alpha}(t_{n\mu}, \alpha) \) for all \( \mu \in [\mu_n, \mu_0]. \) From here the argument of Theorem 3.2 in CO[4] is repeated to obtain parts (i) and (ii) of the theorem.

\( \Box \)

Acknowledgement

I am very grateful to Professors John Crowley and Grace Wahba for stimulating discussions related to this paper.


