Nonparametric Estimation
in the
Cox Proportional Hazards Model

by

Finbarr O'Sullivan[1]

Department of Statistics
University of California
Berkeley, CA 94720

Technical Report No. 64
May 1986

*Research Supported by National Science Foundation
Grant No. MCS-8403239.

Department of Statistics
University of California
Berkeley, California

Nonparametric Estimation in the Cox Proportional Hazards Model.

Finbarr O'Sullivan[1]

Department of Statistics
University of California
Berkeley. CA 94720.

ABSTRACT

The modern counting process formulation of the Cox model with multivariate time dependent covariates is considered. Nonparametric estimation of the proportional hazards regression function is based on a penalized partial likelihood. The paper addresses the asymptotic behavior of this estimator. Rates of convergence in a variety of norms are obtained. The analysis makes essential use of martingale representations. Spectral norms associated with the limiting penalized partial likelihood are used extensively. The results obtained match those already known for familiar nonparametric regression estimators.

AMS 1980 subject classifications. Primary, 62-G05, Secondary, 62-J05, 41-A35, 41-A25, 47-A53, 45-L10, 45-M05.

Key words and phrases. Counting Process, Penalized Partial Likelihood, Penalty Information Scale, Linearization, Rates of Convergence.

Running Head: Nonparametric Estimation in the Cox Model

May 31, 1986

[1] Research supported by the National Science Foundation under Grant No. MCS-840-3239.

Nonparametric Estimation

in the

Cox Proportional Hazards Model.

Finbarr O'Sullivan[1]

Department of Statistics

University of California ,

Berkeley. CA 94720.

1. Introduction

One of the most widely used techniques in survival analysis is the Cox proportional hazards model [7]. In this model the hazard rate or intensity of failure for the survival time for an individual with covariate vector X which may depend on time t is expressed as

$$\lambda(t\,;X(t)) = \lambda_0(t)\exp\left\{\theta_0(X(t))\right\} \quad , \quad t \geq 0 \,. \tag{1.1}$$

θ_0 is the regression function and λ_0 is the underlying baseline hazard. Both θ_0 and λ_0 are unknown. Usually θ_0 is specified as a parametric linear function, i.e. $\theta_0(X) = \beta' X$. In practical data analysis (exploratory or confirmatory) the nonparametric estimation of θ_0 could be of interest. A method of estimation based on a penalized partial likelihood is proposed here.

If $t_1, t_2, \cdots t_n$ are a set of (possibly right censored) survival times on n individu-

[1] Research supported by the National Science Foundation under Grant No. MCS-840-3239.

als with corresponding covariate vectors $X_1, X_2, \cdots, X_n$, where X_i is observed on $[0, t_i]$, then the estimate of θ_0 is defined as the minimizer of the functional

$$l_{n\mu}(\theta) = -\log L_n(\theta) + \mu J(\theta) \quad , \quad \mu > 0 \, , \tag{1.2}$$

where

$$L_n(\theta) = \prod_{i=1}^{n} \left\{ \frac{e^{\theta(X_i(t_i))}}{\sum\limits_{j \in R_i} e^{\theta(X_j(t_i))}} \right\}^{\delta_i} \tag{1.3}$$

with $R_i = \left\{ j : t_j \geq t_i \right\}$ and $1-\delta_i$ is a censoring indicator. The penalty functional $J(\theta)$ is designed to incorporate prior notions about the smoothness of θ into the estimation. The term ***Penalized Likelihood*** is due to Good and Gaskins[9]. The technique, which is closely related to Tikhonov's ***Method of Regularization*** can be used to generate a broad range of nonparametric estimators. For example, in the usual nonparametric regression context the method can be used to define the smoothing spline estimators of Wahba[21]. For further examples, see Anderson and Senthilselvan[4], Cox and O'Sullivan[6], Leonard[13], and Silverman[17]. The use of penalized partial likelihood for nonparametric estimation in the Cox model seems quite natural. Alternative approaches based on additive approximations to the regression function have been proposed by Hastie and Tibshirani[10] and also Stone[18].

The numerical computation of the penalized partial likelihood estimate in (1.2) along with an efficient strategy for selecting the smoothing parameter (μ) is an important practical issue. This will be discussed in a future paper. By and large the methods described in O'Sullivan[14] can be applied. The purpose of the present paper is to study the asymptotic behavior of the penalized likelihood estimator. The modern treatment of the large sample properties of the regression parameter in the Cox model uses a counting process formulation of the model, combined with certain martingale convergence theorems, see Andersen and Gill[3]. The analysis in this paper also uses martingale representations, however, only rudimentary properties of martingales are needed in the analysis. The main theoretical result shows that the penalized likelihood estimate converges in probability, for a variety of norms, at a rate characteristic of other non-parametric regression estimates (see, for example, Cox and O'Sullivan[6]).

Some basic definitions, assumptions and a statement of the main result are given in §2. Convergence is studied in a variety of norms that are related to the structure of the problem. These norms which are equivalent to certain Sobolev norms are described in §3. The behavior of the penalized likelihood estimator is studied via certain asymptotic linearizations; these are described and justified in §4 and §5. The results depend on showing that the third derivative of the penalized partial likelihood is bounded and that the sample Hessian (the second derivative of the penalized partial likelihood) converges at an appropriate rate to a limiting form.

Acknowledgement

I am indebted to Professors John Crowley and Grace Wahba for stimulating my interest in this topic.

2. Basic Definitions and the Main Asymptotic Result

2.1. Counting Process Formulation of the Cox Model

The multivariate counting process formulation of the Cox model as described by Andersen and Gill[3] and Gill[8], will be used. A brief overview of this formulation is given next. For a detailed account, including proper definitions, see Andersen and Gill[3] and the references cited therein. Throughout the paper the time index t is assumed to take values on an interval [0,1].

For each n, $N^{(n)} = (N_1^{(n)}, N_2^{(n)}, \cdots, N_n^{(n)})$, is a multivariate counting process with a random intensity process $\lambda^{(n)} = (\lambda_1^{(n)}, \lambda_2^{(n)}, \cdots, \lambda_n^{(n)})$ for which

$$\lambda_i^{(n)}(t) = Y_i^{(n)}(t) \cdot \exp\{\theta_0(X_i^{(n)}(t))\} \cdot \lambda_0(t) \quad . \tag{2.1}$$

The underlying baseline hazard λ_0 and the regression function $\theta_0 : \mathbb{R}^d \rightarrow \mathbb{R}$ are fixed. A family of right continuous non-decreasing sub σ-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 defined relative to this sequence of σ-algebras. $Y_i^{(n)}(\cdot)$ is a predictable process taking values in $\{0,1\}$, with $Y_i^{(n)} = 1$ whenever the i'th component of the process is under study. The d-dimensional covariate process $X_i^{(n)}(\cdot)$ is predictable and locally bounded. Specification of $\lambda^{(n)}$ as an intensity process means that the process

$$M_i^{(n)}(t) = N_i^{(n)}(t) - \int_0^t \lambda_i^{(n)}(\tau) d\tau, \quad i=1,2, \cdots n \quad \text{and} \quad t \in [0,1] \, , \tag{2.2}$$

is a local martingale. The predictable covariation of $M^{(n)}$ is described as

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

For ease of notation the superscript (n) will be dropped from N, λ, M, Y, and X.

2.2. Definition of the Penalized Partial Likelihood

Inferences for the regression function θ_0 are to be based on a penalized partial likelihood functional. Assuming the process is stopped at time t, this functional is defined by:

$$\int_0^t \log\left[\frac{1}{n}\sum_{i=1}^n Y_i(\tau)e^{\theta(X_i(\tau))}\right] d\overline{N}(\tau) - \frac{1}{n}\sum_{i=1}^n \int_0^t \theta(X_i(\tau))dN_i(\tau) + \mu J(\theta) \quad \mu > 0\ , \qquad (2.3)$$

where $\overline{N}(t) = \frac{1}{n}\sum_{i=1}^n N_i(t)$. In the standard Cox model framework, the first two terms represent the negative logarithm of the partial likelihood given in (1.3). J is a penalty functional. Before going further, assumptions on the measurement model and the nature of the parameter space will be made explicit. The assumptions in this paper combine the standard kinds of conditions set out in the analysis of the Cox model and in the analysis of nonparametric regression estimators, see §4 of Andersen and Gill[3] and assumptions A-C of Cox and O'Sullivan[6].

Assumption A. (Measurement Model)

(i) $(N_i^{(n)}, Y_i^{(n)}, X_i^{(n)})$ $i = 1,2, \cdots ,n$ *are iid replicates of a fixed triple of random processes* (N, Y, X) *defined on* $[0,1]$.

(ii) $\{X(t)\ ,\ t \in [0,1]\} \subseteq \underline{X} \subseteq R^d$ *where* $\underline{X}$ *is a bounded open simply connected set with* C^∞ *- boundary (See Definition 3.2.1.2. of Triebel[20]).*

(iii) For each $t \in [0,1]$, *the random variable* $X(t)$ *has density* $h(\cdot \mid t)$ *and for all* $z \in \underline{X}$

$$0 < k_1 \leq h(z \mid t) \leq k_2 < \infty \ , \qquad (2.4)$$

where the constants k_1 *and* k_2 *are independent of* t *and* z. *The marginal distribution of the process* $X(\cdot)$ *is denoted* $h(\cdot)$. *From (2.4),* h *is bounded away from zero and infinity.*

(iv) For all $x \in \underline{X}$,

$$p_Y(t \mid x) = P[\ Y(t) = 1 \mid X(t) = x\ ]\ . \tag{2.5}$$

$p_Y(t \mid x)$ *is continuous in* t *and* x, *and is bounded away from zero uniformly in* t *and* x.

(v) $\Lambda_0(1) = \int_0^1 \lambda_0(t)dt$, *and* $\Lambda_0(1) < \infty$.

For the assumptions on the parameter space, let $W_2^p(\underline{X},\mathbb{R})$ denote the Sobolev space of real valued L_2 functions defined on $\underline{X}$ whose p'th derivative is square integrable, see Adams[1] (fractional derivatives are allowed). $\underline{W}_2^p(\underline{X},\mathbb{R})$ is the subspace of $W_2^p(\underline{X},\mathbb{R})$ which consists of functions with mean zero. Thus $\int_{\underline{X}} \theta(x)dx = 0$, for all $\theta \in \underline{W}_2^p(\underline{X},\mathbb{R})$.

Assumption B. (Parameter Space)

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

(ii) For some $m > 3d/2$, $\Theta = \underline{W}_2^m(\underline{X}\ ;\ \mathbb{R})$, *as sets and they have equivalent norms.*

(iii) The penalty functional $J(\theta) = <\theta, W\theta>$ *where* W *is a bounded linear operator on* Θ, *which is self-adjoint and nonnegative definite.*

(iv) For some positive constants k_1, k_2,

$$k_1||\theta||^2 \leq <\theta, W\theta> + ||\theta \mid L_2(\underline{X}\ ;\ \mathbb{R})||^2 \leq k_2||\theta||^2\ , \tag{2.6}$$

for all $\theta \in \Theta$.

(v) The true function parameter θ_0 *is in* W_2^s *for some* $s > 3d/2$.

Where necessary, norms etc.. are indexed to indicate exactly how the norm is defined. Thus for $\theta : \underline{X} \to \mathbb{R}$,

$$||\theta||^2_{L_2} = \int_x [\theta(x)]^2 dx \qquad , \qquad ||\theta||^2_{sup} = \sup_x \; |\,\theta(x)\,|^2 \text{ etc.}$$

Whenever subscripting gets too cumbersome, a notation like $||\theta \mid L_2(\underline{X};R)||$ (as in (2.6) above) will be used. If A is a real Banach space then $S(R,A)$ denotes the centered ball of radius R in A. $S_x(R,A)$ is the ball of radius R about x. The centered sup-norm ball is denoted $S(R,C)$:

$$S(R,C) = \left\{ \theta : \mathbb{R}^d \to \mathbb{R} \text{ such that } ||\theta||_{sup} \leq R \right\}.$$

$S_{\theta_0}(R,C)$ is similarly defined. (Here C stands for the Banach space $C(\underline{X};\mathbb{R})$ of real valued continuous functions on $\underline{X}$).

Asymptotic Notation

The following asymptotic notation will be used. If f and g are real valued functions on a metric space U and $u_0 \in U$, then

$$f(u) \lesssim g(u) \qquad \text{as} \qquad u \to u_0$$

means for some K and some neighborhood N of u_0,

$$\left| \frac{f(u)}{g(u)} \right| \leq K \quad , \quad \text{for all } u \in N,$$

where the numerator is required to be zero whenever the denominator is zero. If there is an additional variable v and $V(u)$ is a set of values of v for each u, then

$$f(u,v) \lesssim g(u,v) \qquad \text{as} \qquad u \to u_0$$

uniformly in $v \in V(u)$ means

$$\sup_{v \in V(u)} \left\{ \frac{f(u,v)}{g(u,v)} \right\} \le K \qquad \text{as} \quad u \to u_0 \ .$$

The notation

$$f(u) \approx g(u) \qquad \text{as} \quad u \to u_0$$

means $f(u) \lesssim g(u)$ and $g(u) \lesssim f(u)$.

2.3. Derivatives of $l_{n\mu}$ and a Representation for the Estimate

The penalized partial likelihood estimator, $\hat{\theta}_{n\mu}$, is defined to be a minimizer over Θ of

$$l_{n\mu}(\theta) \equiv l_{n\mu}(\theta;1) = \int_0^1 \log\left[\frac{1}{n}\sum_{i=1}^{n} Y_i(t)e^{\theta(X_i(t))}\right] d\bar{N}(t)$$
$$- \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \theta(X_i(t))dN_i(t) + \mu <\theta, W\theta> \ , \qquad (2.7)$$

where $\mu > 0$. $l_{n\mu}$ is a mapping from Θ into $\mathbb{R}$. In order to describe the asymptotic behavior of $\hat{\theta}_{n\mu}$, various derivatives of $l_{n\mu}$ must be defined. By assumption B(ii), evaluation at any point $x \in X$ is a continuous linear functional in Θ. Hence for $x \in \underline{X}$, the Riesz representer of evaluation at x is well defined, this is denoted $\xi(x)$. Thus for all θ in Θ

$$\theta(x) = <\theta, \xi(x)> \ .$$

As in Andersen and Gill[3], the quantities:

$$S^{(0)}(\theta,t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)e^{\theta(X_i(t))}$$
$$S^{(1)}(\theta,t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\xi(X_i(t))e^{\theta(X_i(t))}$$

$$S^{(2)}(\theta,t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\xi(X_i(t))\times\xi(X_i(t))e^{\theta(X_i(t))}$$

$$S^{(3)}(\theta,t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\xi(X_i(t))\times\xi(X_i(t))\times\xi(X_i(t))e^{\theta(X_i(t))}, \qquad (2.8)$$

come up frequently. $S^{(1)} \in \Theta$, $S^{(2)} \in \Theta\times\Theta$, and $S^{(3)} \in \Theta\times\Theta\times\Theta$. For any $\phi \in \Theta$,

$$<\xi\times\xi,\phi> \equiv \xi<\xi,\phi>$$

$$<\xi\times\xi\times\xi,\phi> \equiv \xi\times\xi<\xi,\phi> \ .$$

The limiting versions of $S^{(0)}$, $S^{(1)}$ and so on, are also of interest. These are defined by:

$$s^{(0)}(\theta,t) = E_z[S^{(0)}(\theta,t)] = \int_z p_Y(t \mid z)e^{\theta(z)}h(z \mid t)dz$$

$$s^{(1)}(\theta,t) = E_z[S^{(1)}(\theta,t)] = \int_z p_Y(t \mid z)\xi(z)e^{\theta(z)}h(z \mid t)dz$$

$$s^{(2)}(\theta,t) = E_z[S^{(2)}(\theta,t)] = \int_z p_Y(t \mid z)\xi(z)\times\xi(z)e^{\theta(z)}h(z \mid t)dz \qquad (2.9)$$

$$s^{(3)}(\theta,t) = E_z[S^{(3)}(\theta,t)] = \int_z p_Y(t \mid z)\xi(z)\times\xi(z)\times\xi(z)e^{\theta(z)}h(z \mid t)dz \ .$$

For ϕ , $\psi \in \Theta$, the notation

$$S^{(2)}(\theta,t)\phi\psi = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\phi(X_i(t))\ \psi(X_i(t))e^{\theta(X_i(t))},$$

$$S^{(3)}(\theta,t)\phi\psi = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\phi(X_i(t))\ \psi(X_i(t))\xi(X_i(t))e^{\theta(X_i(t))},$$

will be used. The definitions of $s^{(2)}(\theta,t)\phi\psi$ and $s^{(3)}(\theta,t)\phi\psi$ are similar. An elementary but extremely useful lemma is the following:

Lemma 2.1. *Let $\theta,\theta_* \in S(R,C)$*

(a) $\dfrac{S^{(0)}(\theta,t)}{S^{(0)}(\theta_*,t)} \approx 1$ *and* $\dfrac{s^{(0)}(\theta,t)}{s^{(0)}(\theta_*,t)} \approx 1$, *uniformly in t.*

(b) $s^{(0)}(\theta,\cdot)$ *is bounded away from zero and infinity.*

(c) *if $\phi,\psi \in \Theta$ then $\{<s^{(1)}(\theta,t),\phi>\}^2 \lesssim ||\phi||_{L_2}^2$ and $\{s^{(2)}(\theta,t)\phi\psi\}^2 \lesssim ||\phi||_{L_2}^2\,||\psi||_{L_2}^2$.*

The results are uniform in θ and θ_*.

Proof: (a) follows from the definition of $S^{(0)}$ and $s^{(0)}$ and the uniform boundedness of θ and θ_*. (b) follows from Assumptions A(iii) and A(iv) and the boundedness of θ. The Cauchy Schwartz inequality, A(iii) and A(iv), and the uniform boundedness of θ implies (c). *Q.E.D.*

With these definitions the first and second derivatives of the penalized partial likelihood are:

$$\begin{aligned} Z_{n\mu}(\theta) &\equiv Dl_{n\mu}(\theta) \\ &= \int_0^1 \frac{S^{(1)}(\theta,t)}{S^{(0)}(\theta,t)} d\overline{N}(t) - \frac{1}{n}\sum_{i=1}^{n}\int_0^1 \xi(X_i(t))dN_i(t) + 2\mu W\theta \quad , \end{aligned} \tag{2.10}$$

$$\begin{aligned} I_{n\mu}(\theta) &\equiv D^2 l_{n\mu}(\theta) \\ &= \int_0^1 \left\{ \frac{S^{(2)}(\theta,t)}{S^{(0)}(\theta,t)} - \frac{S^{(1)}(\theta,t)}{S^{(0)}(\theta,t)} \times \frac{S^{(1)}(\theta,t)}{S^{(0)}(\theta,t)} \right\} d\overline{N}(t) + 2\mu W \quad . \end{aligned} \tag{2.11}$$

But with $p_i(t) = \dfrac{\frac{1}{n} Y_i(t) e^{\theta(X_i(t))}}{\frac{1}{n}\sum_{j=1}^{n} Y_j(t) e^{\theta(X_j(t))}}$ for $i=1,2,\cdots$, and $\overline{\phi}_t = \sum_{j=1}^{n} \phi(X_j(t))p_j(t)$,

$$\frac{S^{(2)}(\theta,t)\phi\phi}{S^{(0)}(\theta,t)} - \frac{S^{(1)}(\theta,t)\phi}{S^{(0)}(\theta,t)}\frac{S^{(1)}(\theta,t)\phi}{S^{(0)}(\theta,t)} = \sum_{i=1}^{n} p_i(t)[\phi(X_i(t)) - \overline{\phi}_t]^2 \ .$$

It follows from this and the fact that W is nonnegative definite, that for any $\phi \in \Theta$, $<I_{n\mu}(\theta)\phi,\phi> \geq 0$. Thus the penalized partial likelihood is convex. A straightforward argument, along the lines given in the appendix of O'Sullivan, Yandell and Raynor[15], shows that the penalized partial likelihood estimator must lie in the subspace $\Theta_n = N(W) \oplus Sp\{\xi(X_i(t_j)), i,j=1,2,\cdots,n\}$, where $N(W)$ is the null space of the linear operator W, Sp stands for the span for the given set, and t_j, for $j=1,2,\cdots,n$, are the survival times of the n individuals under study. If W corresponds to the usual Laplacian penalty functional used to generate thin plate smoothing splines (see O'Sullivan, Yandell and Raynor[15] and Wahba[21]) then the penalized partial likelihood estimator can be represented as a generalized Laplacian smoothing spline. From Theorem 3.2 in Cox and O'Sullivan[6] 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. *The minimizer of the penalized partial likelihood in (2.7) lies in the subspace* $\Theta_n = N(W) \oplus Sp\{\xi(X_i(t_j))\ ,\ i,j=1,2,\cdots,n\ \}$. *Moreover a sufficient condition for the existence of a unique minimizer of the penalized partial likelihood is that there exist a unique minimizer of the negative logarithm of the partial likelihood over* $N(W)$.

Letting $n \rightarrow \infty$, the limiting version of the penalized partial likelihood becomes:

$$l_\mu(\theta) = \int_0^1 \log[s^{(0)}(\theta,t)]s^{(0)}(\theta_0,t)\lambda_0(t)dt - \int_0^1\int_X \theta(x)p_Y(t \mid x)h(x \mid t)e^{\theta_0(x)}dx\,\lambda_0(t)dt + \mu< W\theta,\theta>$$

and the first and second derivatives are:

$$Z_\mu(\theta) = \int_0^1 \frac{s^{(1)}(\theta,t)}{s^{(0)}(\theta,t)}s^{(0)}(\theta_0,t)\lambda_0(t)dt - \int_0^1\int_x \xi(x)p_Y(t \mid x)h(x \mid t)e^{\theta_0(x)}dx\,\lambda_0(t)dt + 2\mu W\theta$$

$$= \int_0^1 \frac{s^{(1)}(\theta,t)}{s^{(0)}(\theta,t)}s^{(0)}(\theta_0,t)\lambda_0(t)dt - \int_0^1 s^{(1)}(\theta_0,t)\lambda_0(t)dt + 2\mu W\theta \tag{2.12}$$

$$G_\mu(\theta) = \int_0^1 \left\{ \frac{s^{(2)}(\theta,t)}{s^{(0)}(\theta,t)} - \frac{s^{(1)}(\theta,t)}{s^{(0)}(\theta,t)} \times \frac{s^{(1)}(\theta,t)}{s^{(0)}(\theta,t)} \right\} s^{(0)}(\theta_0,t)\lambda_0(t)dt + 2\mu W$$

$$= U(\theta) + 2\mu W\ . \tag{2.13}$$

It is easily verified that θ_0 satisfies $Z_0(\theta_0) \equiv 0$. From Lemma 3.1 in the next section, if $\phi \in \Theta$ and $<G_0(\theta)\phi,\phi> = 0$ then $\phi = 0$. Therefore l_0 is strictly convex so θ_0 is the uniquely defined as the solution to $Z_0 \equiv 0$.

2.4. Main Asymptotic Result

The large sample behavior of the penalized partial likelihood estimator $\hat{\theta}_{n\mu}$ is of interest. Since $l_{n\mu}$ is convex, $\hat{\theta}_{n\mu}$ can be thought of as a solution to the variational equation, $Z_{n\mu} \equiv 0$. (The existence of a unique $\hat{\theta}_{n\mu}$, for n sufficiently large and μ sufficiently small, is shown in §5). The asymptotic behavior of $\theta_{n\mu}$ is studied via two linearizations. Heuristically, as $n \rightarrow \infty$, $Z_{n\mu} \rightarrow Z_\mu$ so $\hat{\theta}_{n\mu}$ should be close to θ_μ where $Z_\mu(\theta_\mu) = 0$. The bias in $\hat{\theta}_{n\mu}$ is approximated by $[\theta_\mu - \theta_0]$, the random error is studied by considering $[\hat{\theta}_{n\mu} - \theta_\mu]$. A one-step linearization of Z_μ about θ_μ gives

$$Z_\mu(\theta_0) = - G_\mu(\theta_0)[\bar{\theta}_\mu - \theta_0] \, .$$

Letting $\bar{\theta}_\mu = \theta_0 - G_\mu^{-1}(\theta_0)Z_\mu(\theta_0)$, it is shown that

$$\mathit{dist}(\bar{\theta}_\mu,\theta_0) = \mathit{dist}(\theta_\mu,\theta_0)[1 + o(1)] \qquad \text{as } \mu \rightarrow 0. \tag{2.14}$$

(Here "*dist*" stands for any appropriate norm, see §3 for examples). A one-step linearization of $Z_{n\mu}$ about $\hat{\theta}_{n\mu}$ gives

$$\bar{\theta}_{n\mu} = \theta_\mu - G_\mu^{-1}(\theta_\mu)Z_{n\mu}(\theta_\mu)$$

(Note that $G_\mu(\theta_\mu)$ rather than $I_{n\mu}(\theta_\mu)$ appears in this formula). Again it can be shown that

$$\mathit{dist}(\bar{\theta}_{n\mu},\theta_\mu) = \mathit{dist}(\hat{\theta}_{n\mu},\theta_\mu)[1 + o_p(1)] \qquad \text{as } n \rightarrow \infty. \tag{2.15}$$

The above results say that in order to analyze the the penalized partial likelihood estimator it is enough to understand the behavior of the linearized ***estimators*** $\bar{\theta}_\mu$ and $\bar{\theta}_{n\mu}$. However the asymptotic behavior of these linearized estimators is relatively easy to determine. Rates of convergence in probability are given in a variety of norms. For $0 \leq b \leq 1$, let $||\cdot||_b$ be the Sobolev W_2^{bm} - norm. The following asymptotic convergence result is obtained.

Theorem 2.3. (Asymptotic Convergence Result)

Suppose Assumptions A and B hold. Let $p = s/m$, where m is as in B (ii) and s is in B (v). Let b satisfy

$$0 \leq b < \min(2-d/2m, (p-d/2m)/2) \ .$$

If $\mu = \mu_n$ is a deterministic sequence such that for some arbitrarily small $\epsilon > 0$

$$\mu_n \to 0 \qquad \text{and} \qquad n^{-1}\mu_n^{-(3d/m+\epsilon)} \to 0 \ ,$$

then

$$(i) \qquad ||\theta_\mu - \theta_0||_b^2 \, , \lesssim \mu^{-[p-b]}$$

and

$$(ii) \qquad ||\hat{\theta}_{n\mu} - \theta_\mu||_b^2 \approx O_p(n^{-1}\mu^{-[b+d/2m]}) \ .$$

Moreover, the optimal rate of convergence for $||\hat{\theta}_{n\mu} - \theta_0||_b^2$ is $O_p(n^{-2m(p-b)/(2mp+d)})$ which is obtained by letting $\mu_n \approx n^{-2m/(2mp+d)}$.

Proof: Part (i) follows from Theorem 5.4 and Theorem 4.1. Part (ii) follows from Theorem 5.5 and Theorem 4.2. The ***optimal*** rate is obtained by equating the bounds on the order of magnitude of terms (i) and (ii). *Q.E.D.*

Remarks.

(i) The condition that $n^{-1}\mu_n^{-(3d/m+\epsilon)} \rightarrow 0$ for some $\epsilon > 0$ means that in order for the optimal rate to apply $mp > 5d/2$. This is a rather stringent lower bound. One would imagine that the result still holds even if this lower bound is replaced by $3d/2$, see §5 for more discussion.

(ii) An interesting generalization of the above result would be to include an asymptotic distribution for the estimator. For this one would need a more sophisticated version of the central limit theorems for local martingales, see Robelledo[16]. Such a result would also open up the possibility of studying the weak convergence of a Breslow type estimator for the baseline cumulative hazard,

$$\Lambda_\mu(t) = \int_0^t \frac{d\overline{N}(\tau)}{\frac{1}{n}\sum_{i=1}^{n} Y_i(\tau)e^{\dot{\theta}_{n,\mu}(X_i(\tau))}} \quad . \tag{2.16}$$

3. The Penalty Information Scale

It is convenient to study rates of convergence in norms derived from the limiting penalized partial likelihood:

$$l_\mu(\theta) = \int_0^1 \log\left[s^{(0)}(\theta,t)\right] s^{(0)}(\theta_0,t)\lambda_0(t)dt \qquad (3.1)$$
$$- \int_0^1\int_x \theta(x)e^{\theta_0(x)} p_Y(t \mid x)h(x \mid t)dx\,\lambda_0(t)dt \; + \mu<\theta, W\theta> \quad .$$

In this section a family of such norms is defined and some of their basic properties are established. A brief overview is given first. Whereas the information matrix plays a role in the asymptotic analysis of finite dimensional parameters, the spectral properties of the information operator G_μ, turn out to be important here. From §2.2, G_μ is represented as:

$$G_\mu(\theta) = U(\theta) + 2\mu W \quad , \qquad (3.2)$$

where $U(\theta)$ is the Hessian of the limiting partial log-likelihood (i.e. $U(\theta)$ is the second derivative of $l_\mu(\theta)$ - $\mu<\theta,W\theta>$). Convergence properties are studied in norms related to the spectral decomposition of W relative to $U(\theta)$. These norms and associated Hilbert spaces are obtained as follows: For each $\theta_\bullet$ sufficiently close to θ_0 (in sup-norm) there exist sequences of eigenvalues $\{\gamma_{\bullet\nu\,;\,\nu=1,2,\cdots}\}$ and eigenfunctions $\{\phi_{\bullet\nu}\,;\,\nu = 1,2,\cdots\}$ satisfying:

$$<\phi_{\bullet\nu}, U(\theta_\bullet)\phi_{\bullet\mu}> = \delta_{\nu\mu}$$
$$<\phi_{\bullet\nu}, W\phi_{\bullet\mu}> = \gamma_{\bullet\nu}\,\delta_{\nu\mu} \,, \qquad (3.3)$$

where $\delta_{\nu\mu}$ is Kronecker's delta. For $b \geq 0$, let

$$||\theta||_{\bullet b} = \{\sum_{\nu=1}^{\infty}[1 + \gamma_{\bullet\nu}^{b}]<\theta, U(\theta_\bullet)\phi_{\bullet\nu}>^2\}^{1/2} \quad ,$$

and let $\Theta_{\bullet b}$ be the Hilbert space obtained by completing $\{\theta \in \Theta : ||\theta||_{\bullet b} < \infty\}$ in the

$||\cdot||_{*b}$-norm, with inner product

$$<\theta,\varsigma>_{*b} = \sum_{\nu=1}^{\infty}[1+\gamma_{*\nu}^{b}]<\theta,U(\theta_*)\phi_{*\nu}><\varsigma,U(\theta_*)\phi_{*\nu}> .$$

Appealing to the K-method of interpolation (see Triebel[20]), the $||\cdot||_{*b}$ norms can be shown to be equivalent to $\underline{W}_2^{bm}$ - norm for $0 \leq b \leq 1$, uniformly in θ_* in sufficiently small sup-norm balls about θ_0. In Cox and O'Sullivan[6] the collection of Hilbert spaces $\{\Theta_{*b}, 0 \leq b \leq 1\}$ is referred to as the ***Penalty Information*** (P.I.) scale of Hilbert spaces. If $\theta_* = \theta_\mu$ then the μ will be used in place of "$*$" in the definition of the P.I. scale associated with θ_μ. For notational convenience the "$*$" or μ index will be dropped altogether at times; thus $||\cdot||_b$ is used in place of $||\cdot||_{*b}$.

The existence of the eigensystem in (3.3) has to be established. The main technical result needed for this is a continuity theorem of Kato[12] on the perturbation of the spectrum of a self-adjoint operator. A separation theorem in Weinberger[23] is used to obtain estimates on the asymptotic behavior of the eigenvalues. Some additional properties of the eigensystem, which are used repeatedly in later sections, are recorded in §3.2.

3.1. Spectral Decomposition

The next lemma gives conditions under which the quadratic form obtained from $U(\theta)$ is equivalent to the L_2 norm on Θ.

Lemma 3.1. *If R is sufficiently small then there exist finite positive constants k_1 and k_2 such that for all θ_* in $S_{\theta_0}(R,C)$*

$$k_1||\theta||_{L_2}^2 \leq <\theta,U(\theta_*)\theta> \leq k_2||\theta||_{L_2}^2 \quad \textit{for all } \theta \in \Theta \quad . \tag{3.4}$$

Proof: Let

$$k(x \mid t,\theta_*) = \frac{p_Y(t \mid x)h(x)}{s^{(0)}(t,\theta_*)} \quad .$$

$$<\theta,U(\theta_*)\vartheta> = \int_0^1\int_x \left[\theta(x) - K_*\theta[t]\right]^2 k(x \mid t,\theta_*)dx \;\; s^{(0)}(t,\theta_0)\lambda_0(t)dt$$

where $K_*\theta[t] = \int_x k(x \mid t,\theta_*)\theta(x)dx$. By Lemma 2.1 (a), $\dfrac{s^{(0)}(t \mid \theta_0)}{s^{(0)}(t \mid \theta_*)} \approx 1$ uniformly for $\theta_* \in S_{\theta_0}(R,C)$ and $t \in [0,1]$, so

$$<\theta,U(\theta_*)\theta> \approx \int_0^1\int_x \left[\theta(x) - K_*\theta[t]\right]^2 p_Y(t \mid x)e^{\theta_*(x)}h(x)dx \;\; \lambda_0(t)dt \quad .$$

By assumptions A(iii), A(iv), A(v), and B(v), it follows that h, p_Y and $e^{\theta_*(x)}$ are uniformly bounded away from zero and infinity, so

$$\approx \int_0^1\int_x \left[\theta(x) - K_*\theta[t]\right]^2 dx\frac{\lambda_0(t)}{\Lambda_0(1)}dt \quad . \tag{3.5}$$

Letting m_x is the Lebesgue measure of $\underline{X}$ (since θ has mean zero), this becomes

$$= \int_x \theta^2(x)dx \; - m_x \int_0^1 \left[K_*\theta[t]\right]^2 \frac{\lambda_0(t)}{\Lambda_0(1)}dt \quad ,$$

$$= ||\theta||_{L_2}^2 - m_x(K_*\theta, K_*\theta) \quad ,$$

where $(\phi,\phi) = \int_0^1 [\phi(t)]^2\frac{\lambda_0(t)}{\Lambda_0(1)}dt$. $k(x \mid t,\theta_*)$ is uniformly continuous in x and t on $\overline{\underline{X}} \times [0,1]$ ($\overline{\underline{X}}$ is the closure of $\underline{X}$). Thus, from example 4.1 on page 159 of Kato[12], the integral operator K_* is a compact mapping from $\underline{L}_1(\overline{\underline{X}};\mathbf{R})$ to $C([0,1],\mathbf{R})$, and indeed K_* is a compact operator from $L_2(\overline{\underline{X}};\mathbf{R})$ into the space $\{\phi:[0,1] \to R$: such that $\int_0^1 [\phi(t)]^2\frac{\lambda_0(t)}{\Lambda_0(1)}dt < \infty$ $\}$. Consequently, if K_*^T is the adjoint of K_*, then $K_*^TK_*$ is a compact self-adjoint operator on $\underline{L}_2(\overline{\underline{X}};\mathbf{R})$.

$$(K_\bullet\theta, K_\bullet\theta) = \int_x \theta(x) K_\bullet^T K_\bullet \theta(x) dx \; = \; <\theta, K_\bullet^T K_\bullet \theta>_{L_2}$$

(The constant $m_x^{1/2}$ is adsorbed into $K_\bullet$ from here on). The largest eigenvalue of $K_\bullet^T K_\bullet$ can be characterized as the maximum of the Raleigh quotient:

$$\lambda_1 = \max_{\theta \in \underline{L}_2(\underline{X};\mathbf{R})} \frac{<\theta, K_\bullet^T K_\bullet \theta>_{L_2}}{<\theta,\theta>_{L_2}} = \frac{<\theta_1, K_\bullet^T K_\bullet \theta_1>_{L_2}}{<\theta_1,\theta_1>_{L_2}} \quad ,$$

where $\theta_1 \in \underline{L}_2(\underline{X};\mathbf{R})$ is the non-zero eigenfunction corresponding to λ_1. From (3.5), $\lambda_1 \leq 1$. If $\lambda_1 = 1$ then (3.4) would imply that $\theta_1 = 0$ almost surely, hence $\lambda_1 < 1$. λ_1 depends on $\theta_\bullet$, so we write $\lambda_1 = \lambda_1(\theta_\bullet)$. Let K_0 be the value of the operator $K_\bullet$ corresponding to $\theta_\bullet = \theta_0$. For any $u, v \in \underline{L}_2(\underline{X}, \mathbf{R})$, it is easy to show that

$$<u, (K_\bullet^T K_\bullet - K_0^T K_0) v>_{L_2}^2 \; \leq \; k_1 \|u\|_{L_2}^2 \|v\|_{L_2}^2 M(R)^2 \quad ,$$

where $M(R) = \sup_{\theta_\bullet \in S_{\theta_0}(R)} \{ \max_{x,t} | k(x \mid t, \theta_\bullet) - k(x \mid t, \theta_0) | \}$, and k_1 is a positive constant independent of u and v. $M(R)$ tends to zero as R tends to zero, so $K_\bullet^T K_\bullet$ tends to $K_0^T K_0$ in operator norm as R tends to zero. It follows from Theorem 4.10 on page 291 of Kato[12] that

$$\lambda_1(\theta_\bullet) \rightarrow \lambda_1(\theta_0) \quad \text{as} \quad R \rightarrow 0 \quad \text{uniformly in } \theta_\bullet \; .$$

Thus there exists R such that for all $\theta_\bullet \in S_{\theta_0}(R, C)$, $\lambda_1(\theta_\bullet)$ satisfies

$$\lambda_1(\theta_\bullet) < [1 + \lambda_1(\theta_0)]/2 < 1 \; .$$

From this for all $\theta \in \underline{L}_2(\underline{X};\mathbf{R})$ and $\theta_\bullet \in S_{\theta_0}(R, C)$

$$\|\theta\|_{L_2}^2 \geq \|\theta\|_{L_2}^2 - <\theta, K_\bullet^T K_\bullet \theta>_{L_2}^2 = \|\theta\|_{L_2}^2 \{1 - \frac{<\theta, K_\bullet^T K_\bullet \theta>_{L_2}^2}{\|\theta\|_{L_2}^2} \}$$

$$\geq \|\theta\|_{L_2}^2 [1 - \lambda_1(\theta_0)]/2 \quad .$$

Since, $||\theta||^2_{L_2} - <\theta,K_*^T K_*\theta>^2_{L_2} \approx <\theta,U(\theta_*)\theta>$ for $\theta \in \underline{L}_2(\overline{X};\mathbb{R})$, uniformly in $\theta_* \in S_{\theta_0}(R,C)$, this implies that there are constants k_1 and k_2 such that

$$k_1||\theta||^2_{L_2} \leq <\theta,U(\theta_*)\theta> \leq k_2||\theta||^2_{L_2}$$

for all $\theta \in \Theta$ and $\theta_* \in S_{\theta_0}(R,C)$. *Q.E.D.*

With this lemma the main theorem of this section can be proved.

Theorem 3.2. (Spectral Decomposition)

(i) *For some* $R > 0$, *for all* $\theta_* \in S_{\theta_0}(R,C)$ *there exist sequences of eigenfunctions* $\{\phi_{*\nu}\,;\,\nu = 1,2,\cdots\} \subseteq \Theta$ *and corresponding eigenvalues* $\{\gamma_{*\nu}\,;\,\nu = 1,2,\cdots\} \subseteq [0,\infty)$ *satisfying:*

$$<\phi_{*\nu},U(\theta_*)\phi_{*\mu}> = \delta_{\nu\mu}$$
$$<\phi_{*\nu},W\phi_{*\mu}> = \gamma_{*\nu}\delta_{\nu\mu}$$

(ii) *The asymptotic behavior of the eigenvalues is given by:*

$$\gamma_{*\nu} \approx \nu^{2m/d}$$

uniformly for $\theta_* \in S_{\theta_0}(R,C)$.

Proof: Consider the Raleigh quotient $\dfrac{B_*(\theta,\theta)}{A_*(\theta,\theta)}$,

where

$$B_*(\theta,\theta) = <\theta,U(\theta_*)\theta>\ ,$$

and

$$A_*(\theta,\theta) = <\theta,U(\theta_*)\theta> + <\theta,W\theta> \quad . \tag{3.6}$$

If B_* is ***completely continuous*** (see Weinberger[23] for the definition) with respect to A_*, then the existence of the eigensystem follows as in the construction in Proposition 2.2 of Cox[5]. From Lemma 3.1, for some $R > 0$, $A_*(\theta,\theta)$ is equivalent to $\|\theta\|^2_{L_2}$, uniformly for $\theta_* \in S_{\theta_0}(R,C)$. By Sobolev's Imbedding Theorem and assumptions B(ii) and B(iv), $\|\theta\|^2_{L_2}$ is completely continuous with respect to $\|\theta\|^2_{L_2} + <\theta, W\theta>$. Therefore Lemma 2 on page 61 of Weinberger[22] implies that B_* is completely continuous with respect to A_*, and the existence of the eigensystem is established.

For part (ii); since A_* is uniformly equivalent to L_2-norm on Θ and, by B(ii) and B(iv), B_* is uniformly equivalent to W_2^m-norm on Θ, the Mapping Principle in Weinberger[23] implies that

$$\gamma_{*\nu} \approx \bar{\gamma}_\nu \ ,$$

where $\bar{\gamma}_\nu$ are the eigenvalues of a $2m$'th order elliptic differential operator obtained from equation (A1.3) in Cox and O'Sullivan[6] - see the discussion following equation (3.1) of the same paper. The elements of Θ are constrained to have mean zero but, by Corollary 1 of Theorem 9.1 on page 63 of Weinberger[23], upper and lower bounds on $\bar{\gamma}_\nu$ can be obtained. If $\{\overset{-}{\gamma}_\nu\}$ are the eigenvalues in equation (3.1) of Cox and O'Sullivan[6], then the result in Weinberger[23] implies

$$\overset{-}{\gamma}_\nu \leq \bar{\gamma}_\nu \leq \overset{-}{\gamma}_{\nu+1} \cdot$$

The results of Agmon[2] give that $\overset{-}{\gamma}_\nu \approx \nu^{2m/d}$, (see Cox and O'Sullivan[6] for a discussion), so the claimed estimate on the asymptotic behavior of $\bar{\gamma}_\nu$ (and consequently $\gamma_{*\nu}$) holds. *Q.E.D*

3.2. Useful Properties of the Penalty Information Scale

The following lemmas give some important properties of the penalty information scale which are used repeatedly later on. A similar set of properties are discussed in Appendix 1 of Cox and O'Sullivan[6]. The proofs of the lemmas are omitted since they amount to trivial restatements of the corresponding results in [6].

Lemma 3.3.

If $b \geq 0$ and $c \geq 0$ are such that $b + c < 2 - d/2m$ then

$$\sum_{\nu=1}^{\infty} (1+\gamma_{*\nu}^{b})(1+\gamma_{*\nu}^{c})(1+2\mu\gamma_{*\nu})^{-2} \approx \mu^{-(b+c+d/2m)} \tag{3.7}$$

as $\mu \to 0$ uniformly in $\theta_* \in S_{\theta_0}$(R,C).

Lemma 3.4.

(a) Let $b \in [0,1]$. For for all $R > 0$, $\exists K_1(R), K_2(R)$ such that for all $\theta_* \in S_{\theta_0}(1,C)$, $\Theta_{*b} = \underline{W}_2^{bm}(\underline{X};\boldsymbol{R})$ as sets and for all $\theta \in \Theta_{*b}$,

$$K_1(R)\|\theta \mid W_2^{bm}\| \leq \|\theta\|_{*b} \leq K_2(R)\|\theta \mid W_2^{bm}\| . \tag{3.8}$$

(b) Let $b \in [1,2]$. Assume $J(\theta) = \sum_{i=1}^{q} \int_x \sum_{|\alpha|=m} [D_x^{\alpha}\theta_i(x)]^2 dx$. Then there exist differential operators $B_0, B_1, \cdots, B_{m-1}$ defined on the boundary $\partial\underline{X}$ such that: (i) B_i is of order $m+i$; (ii) $\{B_i : 0 \leq i \leq m\}$ is a normal system (Definition 4.3.3.1 of Triebel[20]); (iii) if $d=1$ then $B_i = (d/dx)^{m+i}$; (iv) Assuming $bm-1/2$ is not an integer, Θ_{*b} is the closed subspace of W_2^{bm} given by $\{\theta \in W_2^{bm} : B_i\theta \equiv 0$ on $\partial\underline{X}$ for all *$i < (b-1)m-1/2\}$, and (3.8) holds.*

Lemma 3.5 *Let $\theta_* \in S_{\theta_0}(R,C)$ and $b > 0$.*

(i) $||\phi_{\bullet\nu}||_b^2 = 1 + \gamma_{\bullet\nu}^b$ *for* $\nu=1,2,\cdots$

(ii) $G_\mu^{-1}(\theta_\bullet)U(\theta_\bullet)\phi_{\bullet\nu} = (1 + 2\mu\gamma_{\bullet\nu})^{-1}\phi_{\bullet\nu}$ *for* $\nu=1,2,\cdots$ *and* $\mu > 0$.

(iii) *If* $\mu_0 > 0$ *is sufficiently small and* $b \le 1$*, then for all* $x \in X$, $\epsilon > 0$, $1 \le j \le q$, *and* $\mu \in (0,\mu_0]$,

$$||G_\mu^{-1}(\theta_\bullet)\xi_j(x)||_b^2 \le K(R,\mu_0)\mu^{-(b+(1+\epsilon)d/m)} .$$

4. Rates of Convergence for the Linearized Estimators

Convergence properties are studied by means of two linearizations. These are given by:

(i) Continuous Linearization:

$$\overline{\theta}_{\mu} = \theta_0 - G_{\mu}^{-1}(\theta_0)\, Z_{\mu}(\theta_0)$$

(ii) Discrete Linearization:

$$\overline{\theta}_{n\,\mu} = \theta_{\mu} - G_{\mu}^{-1}(\theta_{\mu})\, [Z_{n\,\mu}(\theta_{\mu}) - Z_{\mu}(\theta_{\mu})] \ . \tag{4.1}$$

For norms $||\cdot||_b$, the asymptotic bias of the penalized partial likelihood estimator is on the order of $||\overline{\theta}_{\mu} - \theta_0||_b$, while the asymptotic variability is on the order of $\mathrm{E}||\overline{\theta}_{n\,\mu} - \theta_{\mu}||_b^2$. The justification of this is rather technical; the argument amounts to showing that $I_{n\,\mu}(\theta_{\mu}) - G_{\mu}(\theta_{\mu}) \rightarrow 0$ as $n \rightarrow \infty$ and that the third order derivatives of the penalized partial likelihood can be bounded in an appropriate manner, see Theorems 5.4 and 5.5 in §5. For the moment consider the asymptotic behavior of the linearized *estimators* in (4.1). A result on the behavior of the bias follows immediately from Theorem 2.3(c) of Cox[5].

Theorem 4.1. (Bias)

If $0 \leq b < (p-d/2m)/2$ *where* $p=s/m$ *and* s *is given in B(v), then as* $\mu \rightarrow 0$

$$||\overline{\theta}_{\mu} - \theta_0||_b \lesssim \mu^{(p-b)/2}$$

where $p = s/m$ *.*

Proof: From §3 the operators $U(\theta_0)$ and W generate norms and associated *Penalty Information* scale with the same structure as those discussed in Cox and O'Sullivan[6].

Therefore, as in Theorem 5.1 (i) of Cox and O'Sullivan[6] the result of Cox[5] applies. *Q.E.D.*

More computation is required to analyze the discrete linearization.

Theorem 4.2. (Variability)

For $0 \leq b < 2\text{-}d/2m$

$$\mathrm{E}\, \|\bar{\theta}_{n\mu} - \theta_\mu\|_b^2 \lesssim n^{-1}\mu^{-(b+d/2m)} \quad \text{as} \quad n \rightarrow \infty ,$$

uniformly for $\mu \in (0,\mu_0]$ ***and*** μ_0 ***sufficiently small.***

Proof: The continuous linearization result to be proved in the next section (Theorem 5.4: let $b = d/2m$ and use Sobolev's Imbedding Theorem), shows that θ_μ converges in sup-norm to θ_0 as $\mu \rightarrow 0$. From the results of the last section, let μ_0 be chosen so that for all $\mu < \mu_0$ the norms and ***Penalty Information*** scale associated with $U(\theta_\mu)$ are uniformly equivalent to those associated with $U(\theta_0)$.

$$\bar{\theta}_{n\mu} - \theta_\mu = G_\mu^{-1}(\theta_\mu)[Z_{n\mu}(\theta_\mu) - Z_\mu(\theta_\mu)] \quad .$$

Using the series expansion for the $\|\cdot\|_{\mu b}$-norm,

$$\|\bar{\theta}_{n\mu} - \theta_\mu\|_b^2 = \sum_{nu=1}^{\infty} [1 + \gamma_{\mu\nu}^b][1 + 2\mu\gamma_{\mu\nu}]^{-2} \cdot \{<Z_{n\mu}(\theta_\mu) - Z_\mu(\theta_\mu) \, , \phi_{\mu\nu}>\}^2 \, .$$

This expansion uses the fact, proved in Lemma 3.5(ii), that $G_\mu^{-1}(\theta_\mu)U(\theta_\mu)\phi_{\mu\nu} = [1+2\mu\gamma_{\mu\nu}]^{-1}\phi_{\mu\nu}$. A similar expansion appears many times in Cox and O'Sullivan[6]. It will be shown that for any $\phi \in \Theta$

$$\mathrm{E}\{<Z_{n\mu}(\theta_\mu) - Z_\mu(\theta_\mu),\phi>\}^2 \lesssim \frac{1}{n}\, \|\phi\|_{L_2}^2 , \tag{4.2}$$

uniformly in $\mu < \mu_0$. With this, Lemma 3.5(i) implies $\|\phi_{\mu\nu}\|_{L_2}^2 = 2$, so

$$E\|\theta_{\mu n}-\theta_\mu\|^2_{\mu b} \lesssim n^{-1}\sum_\nu[1+\gamma^b_{\mu\nu}][1+2\mu\gamma_{\mu\nu}]^{-2}\|\phi_{\mu\nu}\|^2_{L_2}$$

$$\approx n^{-1}\sum_\nu[1+\gamma^b_{\mu\nu}][1+2\mu\gamma_{\mu\nu}]^{-2}$$

$$\approx n^{-1}\mu^{-(b+d/2m)} \qquad \text{by Lemma 3.3 .}$$

Turning to the verification of equation (4.2),

$$<Z_{n\mu}(\theta_\mu)-Z_\mu(\theta_\mu),\phi> = \left\{\int_0^1\frac{<S^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)}\,d\bar{N}(t)-\frac{1}{n}\sum_{i=1}^n\int_0^1\phi(X_i(t))dN_i(t)\right\}$$
$$-\left\{\int_0^1\frac{<s^{(1)}(\theta_\mu,t),\phi>}{s^{(0)}(\theta_\mu,t)}\,s^{(0)}(\theta_0,t)\lambda_0(t)dt-\int_0^1<s^{(1)}(\theta_0,t),\phi>\lambda_0(t)dt\right\} .$$

Rewriting in terms of the local martingale $\bar{M}(t)=\frac{1}{n}\sum_{i=1}^n M_i(t)$,

$$<Z_{n\mu}(\theta_\mu)-Z_\mu(\theta_\mu),\phi> = \int_0^1\left[\frac{<S^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)}S^{(0)}(\theta_0,t)-\frac{<s^{(1)}(\theta_\mu,t),\phi>}{s^{(0)}(\theta_\mu,t)}s^{(0)}(\theta_0,t)\right]\lambda_0(t)dt$$
$$+\int_0^1\frac{<S^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)}\,d\bar{M}(t)-\frac{1}{n}\sum_{i=1}^n\phi(X_i(t))dM_i(t)$$
$$-\int_0^1<S^{(1)}(\theta_0,t)-s^{(1)}(\theta_0,t),\phi>\lambda_0(t)dt .$$

Therefore

$$\{<Z_{n\mu}(\theta_\mu)-Z_\mu(\theta_\mu),\phi>\}^2$$
$$\leq\left\{\int_0^1\left[\frac{<S^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)}S^{(0)}(\theta_0,t)-\frac{<s^{(1)}(\theta_\mu,t),\phi>}{s^{(0)}(\theta_\mu,t)}s^{(0)}(\theta_0,t)\right]\lambda_0(t)dt\right\}^2$$
$$+\left\{\int_0^1\frac{<S^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)}\,d\bar{M}(t)\right\}^2+\left\{\frac{1}{n}\sum_{i=1}^n\phi(X_i(t))dM_i(t)\right\}^2 \qquad (4.3)$$

$$+\left\{\int_0^1 <S^{(1)}(\theta_0,t) - s^{(1)}(\theta_0,t),\phi>\lambda_0(t)\,dt\right\}^2 .$$

It will be shown that the order of magnitude of the expected value of each term on the right hand side is $\frac{1}{n}\ ||\phi||_{L_2}^2$ - this will prove (4.2). The last two terms in equation (4.3) are easy to handle. Since $<S^{(1)}(\theta_\mu,t),\phi> = \frac{1}{n}\sum_{i=1}^{n}\phi(X_i(t))Y_i(t)e^{\theta(X_i(t))}$ and, by A(i), $X_i(\cdot)$, $Y_i(\cdot)$ $i=1,2,\cdots,n$ are iid replicates of fixed processes $X(\cdot)$ and $Y(\cdot)$,

$$\mathrm{E}\left\{\int_0^1 <S^{(1)}(\theta_\mu,t) - s^{(1)}(\theta_\mu,t),\phi>\lambda_0(t)\,dt\right\}^2 \leq \int_0^1 \mathrm{E}\left\{<S^{(1)}(\theta_\mu,t) - s^{(1)}(\theta_\mu,t),\phi>\right\}^2 dt \cdot \int_0^1 \lambda_0(t)\,dt$$

$$\lesssim \frac{1}{n}\int_0^1 \mathrm{Var}\{\phi(X(t))Y(t)e^{\theta_\mu(X(t))}\}\lambda_0(t)\,dt \quad .$$

But for all $\mu < \mu_0$, θ_μ is uniformly bounded in sup-norm (Theorem 5.4) $h(\cdot\,|\,t)$ is uniformly bounded by A(iii), and $Y(\cdot)$ is bounded, so

$$\mathrm{Var}\{\phi(X(t))Y(t)e^{\theta_\mu(X(t))}\} \lesssim ||\phi||_{L_2}^2 \qquad (4.4)$$

uniformly in t and $\mu \leq \mu_0$. This gives the order $n^{-1}\ ||\phi||_{L_2}^2$ bound on the last term in (4.3).

For the second to last term, computing the predictable variation of the martingale gives:

$$\mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n}\int_0^1 \phi(X_i(t))\,dM_i(t)\right\}^2 = \mathrm{E}\left\{\frac{1}{n}\int_0^1 \frac{1}{n}\sum_{i=1}^{n}\phi(X_i(t))^2 Y_i(t)e^{\theta_0(X_i(t))}\lambda_0(t)\,dt\right\}$$

$$= \frac{1}{n}\int_0^1 \mathrm{E}\left\{\phi(X(t))^2 Y(t)e^{\theta_0(X(t))}\right\}\lambda_0(t)\,dt$$

$$\lesssim \frac{1}{n} \|\phi\|_{L_2}^2 \ .$$

The last line follows from the boundedness of $Y(\cdot)$, $h(\cdot \mid \cdot)$, and θ_0.

This leaves the first two terms of equation (4.3) to analyze. Again computing the predictable variation of the martingale term, gives:

$$E\left\{\int_0^1 \frac{<S^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)}\, d\overline{M}(t)\right\}^2 = \frac{1}{n}E\int_0^1 \left[\frac{<S^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)}\right]^2 S^{(0)}(\theta_0,t)\lambda_0(t)dt \ .$$

By the Cauchy-Schwartz inequality

$$\left\{\frac{<S^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)}\right\}^2 \leq \frac{\left\{\frac{1}{n}\sum_{i=1}^n Y_i(t)e^{\theta_\mu(X_i(t))}\right\}\left\{\frac{1}{n}\sum_{i=1}^n Y_i(t)\phi(X_i(t))^2 e^{\theta_\mu(X_i(t))}\right\}}{S^{(0)}(\theta_\mu,t)^2} \ ,$$

$$\lesssim \frac{\left\{\frac{1}{n}\sum_{i=1}^n Y_i(t)\phi(X_i(t))^2 e^{\theta_\mu(X_i(t))}\right\}}{S^{(0)}(\theta_\mu,t)} \ , \qquad \text{by Lemma 2.1(a)} \ .$$

Applying Lemma 2.1(a) again,

$$\left\{\frac{<S^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)}\right\}^2 S^{(0)}(\theta_0,t) \lesssim \frac{1}{n}\sum_{i=1}^n Y_i(t)\phi(X_i(t))^2 e^{\theta_\mu(X_i(t))} \ .$$

Taking expectations and using the same bounds as in equation (4.4):

$$E\int_0^1 \left\{\frac{<S^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)}\right\}^2 S^{(0)}(\theta_0,t)\lambda_0(t)dt \lesssim \|\phi\|_{L_2}^2 \ ,$$

uniformly in $\mu < \mu_0$. For the remaining term in equation (4.3) it is necessary to study the convergence of

$$\left[\frac{<S^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)}S^{(0)}(\theta_0,t) - \frac{<s^{(1)}(\theta_\mu,t),\phi>}{s^{(0)}(\theta_\mu,t)}s^{(0)}(\theta_0,t)\right] \ . \tag{4.5}$$

The arguments in Andersen and Gill[3] could be adapted to show that this term converges to zero, uniformly in t and μ, for fixed ϕ in $C(\underline{X};\mathbf{R})$. A result of a different character is needed here; basically, the result must yield an order $n^{-1/2}$ rate of convergence uniform in ϕ and μ (but not necessarily uniform in t). To prove this, note

$$\frac{<S^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)} = \left\{\frac{<S^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)} - \frac{<s^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)}\right\}$$
$$+ <s^{(1)}(\theta_\mu,t),\phi>\left\{\frac{1}{S^{(0)}(\theta_\mu,t)} - \frac{1}{s^{(0)}(\theta_\mu,t)}\right\}$$
$$+ \frac{<s^{(1)}(\theta_\mu,t),\phi>}{s^{(0)}(\theta_\mu,t)}$$

$$= \left\{\frac{<S^{(1)}(\theta_\mu,t) - s^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)}\right\}$$
$$+ \left\{\frac{<s^{(1)}(\theta_\mu,t),\phi>}{s^{(0)}(\theta_\mu,t)}\right\}\left\{\frac{s^{(0)}(\theta_\mu,t) - S^{(0)}(\theta_\mu,t)}{S^{(0)}(\theta_\mu,t)}\right\}$$
$$+ \frac{<s^{(1)}(\theta_\mu,t),\phi>}{s^{(0)}(\theta_\mu,t)} .$$

So

$$\left\{\frac{<S^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)} S^{(0)}(\theta_0,t) - \frac{<s^{(1)}(\theta_\mu,t),\phi>}{s^{(0)}(\theta_\mu,t)} s^{(0)}(\theta_0,t)\right\}$$
$$= \left[<S^{(1)}(\theta_\mu,t) - s^{(1)}(\theta_\mu,t),\phi>\right]\left\{\frac{S^{(0)}(\theta_0,t)}{S^{(0)}(\theta_\mu,t)}\right\}$$
$$- \left[S^{(0)}(\theta_\mu,t) - s^{(0)}(\theta_\mu,t)\right]\left\{\frac{<s^{(1)}(\theta_\mu,t),\phi>}{s^{(0)}(\theta_\mu,t)}\right\}\left\{\frac{S^{(0)}(\theta_0,t)}{S^{(0)}(\theta_\mu,t)}\right\}$$
$$+ \left[S^{(0)}(\theta_0,t) - s^{(0)}(\theta_0,t)\right]\left\{\frac{<s^{(1)}(\theta_\mu,t),\phi>}{s^{(0)}(\theta_\mu,t)}\right\} .$$

Using the fact that $\dfrac{S^{(0)}(\theta_0,t)}{S^{(0)}(\theta_\mu,t)} \approx 1$ uniformly in t and $\mu < \mu_0$, the Cauchy-Schwartz inequality, and the finiteness of $\int_0^1 \lambda_0(t)\,dt$,

$$\left\{ \int_0^1 \left[\frac{<S^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)} S^{(0)}(\theta_0,t) - \frac{<s^{(1)}(\theta_\mu,t),\phi>}{s^{(0)}(\theta_\mu,t)} s^{(0)}(\theta_0,t) \right] \lambda_0(t)dt \right\}^2$$

$$\lesssim \int_0^1 [<S^{(1)}(\theta_\mu,t) - s^{(1)}(\theta_\mu,t),\phi>]^2 \lambda_0(t)dt \tag{4.6}$$

$$+ \left\{ \int_0^1 \{[S^{(0)}(\theta_0,t) - s^{(0)}(\theta_0,t)]^2 + [S^{(0)}(\theta_\mu,t) - s^{(0)}(\theta_\mu,t)]^2\} \lambda_0(t)dt \right.$$

$$\left. \cdot \int_0^1 \left[\frac{<s^{(1)}(\theta_\mu,t),\phi>}{s^{(0)}(\theta_\mu,t)} \right]^2 \lambda_0(t)dt \right\} .$$

Computing expectations an order n^{-1} $||\phi||^2_{L_2}$ bound (uniform in μ) is obtained for the first term. By Lemma 2.1 (b) and (c),

$$\left\{ \frac{<s^{(1)}(\theta_\mu,t),\phi>}{s^{(0)}(\theta_\mu,t)} \right\}^2 \lesssim ||\phi||^2_{L_2} .$$

Finally, $\int_0^1 \mathrm{E}\left\{ S^{(0)}(\theta_\mu,t) - s^{(0)}(\theta_\mu,t) \right\}^2 \lambda_0(t)dt$ is order n^{-1} uniformly for $\mu < \mu_0$, so the correct bound is also obtained for the second term in equation (4.6). This completes the proof. *Q.E.D.*

5. Validity of the Linearizations

The final step is to verify that the linearized estimators, analyzed in §4, mimic the first order asymptotic bias and variability characteristics of the penalized partial likelihood estimator. For this it suffices to show that the hypotheses of the continuous and discrete linearization theorems in Cox and O'Sullivan[6] are satisfied. These linearization theorems are based on general fixed point arguments. They also yield existence results. The ideas were motivated by the work of Huber[11] on the asymptotics of M-estimators, however, whereas Huber's analysis takes place in a finite dimensional setting, the analysis here takes place in infinite dimensional parameter space. Because of this the analysis is more complicated.

In the linearization theorems a pair of norms, $\|\cdot\|_b$ and $\|\cdot\|_{b^*}$, are employed. The fixed point property is established in the $\|\cdot\|_{b^*}$-norm and from this information on the efficacy of the linearizations is deduced for the other norm. The argument (see Lemma 5.2), makes critical use of a lower bound of d/m on b^* and this lower bound ultimately effects the final convergence result stated in Theorem 2.3. If the restriction that $mp > 5d/2$, which is needed to include the optimal rate of convergence, is to be relaxed, then the analysis in Lemma 5.2 of this section would have to be refined. The proofs of the theorems rely on the third order derivatives of the penalized partial likelihood functional being suitably bounded, and that the operator $I_{n\mu}(\theta_\mu)$ converges to $G_\mu(\theta_\mu)$ at an appropriate rate.

5.1. Additional Notation and the Main Theorems

Some additional notation must be defined. The Banach space obtained by completing Θ under the norm $\|\cdot\|_{\mu b}$ is denoted $\Theta_{\mu b}$. As usual the subscript μ will typically be dropped. Let $d^*(\mu) \approx \|\bar{\theta}_\mu - \theta_0\|_{b^*}$ and let $d^*(n,\mu)$ be such that

$O_p(d^*(n,\mu)) \approx ||\overline{\theta}_{n\mu} - \theta_\mu||_{b^*}$. The asymptotic behavior of $d^*(\mu)$ and $d^*(n,\mu)$ is described in Theorems 4.1 and 4.2. Theorem 4.1 gives that $d^*(\mu) \approx \mu^{(p-b^*)/2}$, provided $0 \leq b^* < (p - d/2m)/2$; Theorem 4.2 combined with Markov's inequality gives the order $n^{-1/2}\mu^{-(b^*+d/2m)/2}$ for $d^*(n,\mu)$, provided $0 \leq b < 2 - d/2m$.

The magnitude of the third order derivative of the penalized partial likelihood is measured by:

$$K_3(\mu,R) = \sup_{\substack{\phi_3 \in S(1,\Theta_{b^*}) \\ \phi_1,\phi_2 \in S(R,\Theta_{b^*})}} ||G_\mu^{-1}(\theta_0)D^3 l_\mu(\theta_0+\phi_1)\phi_2\phi_3||_b$$

in the continuous case, and by

$$K_3(n,\mu,R) = \sup_{\substack{\phi_3 \in S(1,\Theta_{b^*}) \\ \phi_1,\phi_2 \in S(R,\Theta_{b^*})}} ||G_\mu^{-1}(\theta_\mu)D^3 l_{n\mu}(\theta_\mu+\phi_1)\phi_2\phi_3||_b$$

in the discrete case (D is the differentiation operator). The difference between $I_{n\mu}$ and $G_\mu(\theta_\mu)$ is measured by

$$K_2(n,\mu) = \sup_{\phi \in S(1,\Theta_{b^*})} ||G_\mu^{-1}(\theta_\mu)(I_{n\mu} - G_\mu(\theta_\mu))\phi||_b \quad .$$

The quantities $K_3^*(\mu,R)$, $K_3^*(n,\mu,R)$ and $K_2^*(n,\mu)$ are defined in a similar manner - the $||\cdot||_b$-norm is replaced by the $||\cdot||_{b^*}$-norm. The next two lemmas are crucial.

Lemma 5.1. (Bounds on the 3'rd order derivatives)

Let $0 \leq b < 2-d/2m$ *and suppose* $d/2m < b^* < (p-d/2m)/2$ *where* $p = s/m$ *and* s *is given in B(v), then*

(a) $K_3(\mu,R) \lesssim R \cdot \mu^{-(b+d/2m)/2}$

(b) $K_3(n,\mu,R) \lesssim R \cdot B_1(n,\mu)$, where $\mathrm{E}B_1^2(n,\mu) \lesssim \mu^{-(b+d/2m)}$, uniformly in

$\mu \in (0,\mu_0]$.

Lemma 5.2. (Convergence of $I_{n\mu}(\theta_\mu)$)

Let $0 \leq b < 2-d/2m$ *and suppose* $d/m < b^* < 2-3d/2m$ *then*

$$K_2(n,\mu) \lesssim B_2(n,\mu) \ ,$$

where $\mathrm{E}B_2^2(n,\mu) \lesssim n^{-1}\mu^{-(b+d/2m)}$.

The remaining results follow from these lemmas. For $x(\mu) \approx d^*(\mu)$ and $x(n,\mu) \approx d^*(n,\mu)$, let $r^*(\mu) = K_3^*(\mu,x(\mu))$ and $r(\mu) = K_3(\mu,x(\mu))$; furthermore, let $r^*(n,\mu)$ and $r(n,\mu)$ be such that $O_p(r^*(n,\mu)) = K_2^*(n,\mu) + K_3^*(\mu,x(n,\mu))$ and $O_p(r(n,\mu)) = K_2(n,\mu) + K_3(\mu,x(n,\mu))$. The next theorem, describes the asymptotic behavior of these constants.

Theorem 5.3. *If* $0 \leq b \leq 2-d/2m$ *and* $d/m < b^* < \min(2-3d/2m,(p-d/2m)/2)$ *then*

(i) $r^*(\mu) \to 0$ *as* $\mu \to 0$,

$r(\mu) << \mu^{(b^*-b)/2}$ *as* $\mu \to 0$,

If $\{\mu_n \ ; n \geq 1\}$ *is a sequence such that for some* $d/m < b^* < 2-3d/2m$, $n^{-1}\mu_n^{-2(b^*+d/2m)} \to 0$. *then for any sequence of* μ*'s in* $[\mu_n,\mu_0]$.

(ii) $r^*(n,\mu) \to 0$, *as* $n \to \infty$.

$r(n,\mu) << \mu^{(b^*-b)/2}$, *as* $n \to \infty$.

Proof: Apply Lemma 5.1 and 5.2. together with the rate estimates on $d^*(\mu)$ and $d^*(n,\mu)$, discussed in the first paragraph of this subsection.

For part (i), from Lemma 5.1

$$\begin{aligned} r(\mu) &\lesssim d^*(\mu)\,\mu^{-(b^* + d/2m)/2} \\ &\lesssim \mu^{(p - b^*)/2}\,\mu^{-(b^* + d/2m)/2} \\ &= \mu^{(p-d/2m)/2 - b^*}\,\mu^{(b^* - b)/2} \\ &<< \mu^{(b^* - b)/2} \quad \text{,as } \mu \to 0 \text{ , since } \quad b^* < (p-d/2m)/2 \quad . \end{aligned}$$

For part (ii), from Lemma 5.1 and Lemma 5.2

$$\begin{aligned} r^2(n,\mu) &\lesssim n^{-1}\mu^{-(b+d/2m)} + d^*(n,\mu)\,\mu^{-(b+d/2m)} \\ &\lesssim n^{-1}\mu^{-(b+d/2m)} + n^{-1}\mu^{-(b^*+d/2m)}\,\mu^{-(b+d/2m)} \\ &\approx n^{-1}\mu^{-2(b^*+d/2m)}\,\mu^{(b^* - b)} \\ &<< \mu^{(b^* - b)} \quad \text{since} \quad n^{-1}\mu^{-2(b^*+d/2m)} \to 0 \quad \text{as} \quad n \to \infty \,. \end{aligned}$$

Q.E.D.

The main results can now be proved.

Theorem 5.4. (Continuous Linearization)

Let $d/2m < b^* < (p-d/2m)/2$ where $p = s/m$ and s is given in $B(v)$. There are constants k_0, k_1 and μ_0 such that, for all $\mu \in [0,\mu_0]$, $\exists\,!\ \phi_\mu \in S(\frac{1}{2}K_0 d^*(\mu),\Theta_{\mu b^*})$ such that if $\theta_\mu = \phi_\mu + \theta_0$ then $Z_\mu(\theta_\mu) = 0$, and

$$||\overline{\theta}_\mu - \theta_\mu||_{b^*} \le k_1\, r^*(\mu)\, d^*(\mu)$$

$$||\overline{\theta}_\mu - \theta_\mu||_b \le k_1\, r(\mu)\, d^*(\mu)$$

where $\overline{\theta}_\mu = \theta_0 - G_\mu^{-1}(\theta_\mu) Z_\mu(\theta_0)$.

In particular for $0 \le b < 2-d/2m$ then $||\theta_\mu - \theta_0||_b \approx ||\overline{\theta}_\mu - \theta_0||_b$ as $\mu \to 0$.

Proof: From Theorem 5.3(i) $r^*(\mu) \to 0$ as $\mu \to 0$ so the hypotheses oi Theorem 4.1 in Cox and O'Sullivan[6] hold. *Q.E.D.*

Theorem 5.5. (Discrete Linearization)

Let $\{\mu_n \, ; n \geq 1\}$ *be a sequence such that for some* $d/m < b^* < 2\text{–}3d/2m$, $n^{-1}\mu_n^{-2(b^*+d/2m)} \to 0$. *For any sequence of* μ*'s in* $[\mu_n, \mu_0]$ *and* $0 \leq b < 2\text{–}d/2m$. *Consider the event* $\mathrm{E}(n,\mu)$ *given by*

$$\mathrm{E}(n,\mu) : \exists! \ \textit{solution to } Z_{n\mu} \equiv 0 \ , \hat{\theta}_{n\mu} = \theta_\mu + \hat{\phi}_{n\mu} \, , \, \hat{\phi}_{n\mu} \in S(\frac{1}{2}K_0 d^*(n,\mu), \Theta_{\mu b^*}),$$

$$\textit{satisfying } \|\hat{\theta}_{n\mu} - \overline{\theta}_{n\mu}\|_{b^*} \leq k_1 \, r^*(n,\mu) \, d^*(n,\mu) \, ,$$

$$\textit{and } \ \|\hat{\theta}_{n\mu} - \overline{\theta}_{n\mu}\|_b \leq k_1 \, r(n,\mu) \, d^*(n,\mu) \, ,$$

then for all $\delta > 0$ *there is some* n_0 *and constants* k_0 *and* k_1 *such that this event occurs with probability* $> 1\text{–}\delta$, *for all* $n \geq n_0$ *and* $\mu \in [\mu_n, \mu_0]$.

In particular on $E(n,\mu)$, $\hat{\theta}_{n\mu}$ *exists and for* $0 \leq b < 2\text{–}d/2m$

$$\|\hat{\theta}_{n\mu} - \theta_\mu\|_b \approx \|\overline{\theta}_{n\mu} - \theta_\mu\|_b \quad ,$$

so $O_p(n^{-1}\mu^{-(b+d/2m)}) = \|\hat{\theta}_{n\mu} - \theta_\mu\|_b^2$.

Proof. The condition on b^* implies that $r^*(n,\mu) \to 0$, Theorem 5.3(ii), so Theorem 4.2 of Cox and O'Sullivan[6] gives the result. The last statement follows by Markov's inequality and the rate obtained in Theorem 4.2 of §4. *Q.E.D.*

5.2. Proofs of Lemmas 5.1 and 5.2

Proof of Lemma 5.1. (Bounds on the 3'rd order derivatives)

For part (a)

$$D^3 l_\mu(\theta_*)vw = G_\mu^{-1}(\theta_0)\dot{U}(\theta_*)vw$$

where $\dot{U}(\theta_*)vw$ is given by:

$$\begin{aligned}\dot{U}(\theta_*)vw = \int_0^1 \Bigg[& \frac{s^{(3)}(\theta_*,t)vw}{s^{(0)}(\theta_*,t)} \\ & - \frac{s^{(2)}(\theta_*,t)vw}{s^{(0)}(\theta_*,t)} \frac{s^{(1)}(\theta_*,t)}{s^{(0)}(\theta_*,t)} \\ & - \frac{<s^{(2)}(\theta_*,t),v>}{s^{(0)}(\theta_*,t)} \frac{<s^{(1)}(\theta_*,t),w>}{s^{(0)}(\theta_*,t)} \\ & - \frac{<s^{(2)}(\theta_*,t),w>}{s^{(0)}(\theta_*,t)} \frac{<s^{(1)}(\theta_*,t),v>}{s^{(0)}(\theta_*,t)} \\ & + \frac{<s^{(1)}(\theta_*,t),v>}{s^{(0)}(\theta_*,t)} \frac{<s^{(1)}(\theta_*,t),w>}{s^{(0)}(\theta_*,t)} \frac{s^{(1)}(\theta_*,t)}{s^{(0)}(\theta_*,t)} \Bigg] s^{(0)}(\theta_0,t)\lambda_0(t)dt\end{aligned}$$

Expanding in terms of the eigensystem $\{\phi_{0\nu}, \gamma_{0\nu}; \nu = 1,2, \cdots\}$,

$$\begin{aligned}||G_\mu^{-1}(\theta_0)D^3 l_\mu(\theta_*)vw||_b^2 = \sum_{\nu=1}^{\infty} [1 + \gamma_{0\nu}^b]\,[1 + 2\mu\gamma_{0\nu}]^{-2} & \qquad (5.1) \\ \cdot \{<\dot{U}(\theta_*)vw, \phi_{0\nu}>\}^2 \,. &\end{aligned}$$

It will be shown that for $\phi \in \Theta$,

$$\{<\dot{U}(\theta_*)vw, \phi>\}^2 \lesssim R^2 \cdot ||\phi||_{L_2}^2 \qquad (5.2)$$

uniformly for $\theta_* \in S_{\theta_0}(R,C)$. The argument involves a straightforward application of sup-norm bounds on θ_*, v, and w, combined with a couple of applications of Holder's inequality. Since $b^* > d/2m$, sup-norms of v and w are finite. There are several terms to handle, but as an illustration consider the term

$$\int_0^1 \left[\frac{<s^{(3)}(\theta_*,t)vw, \phi>}{s^{(0)}(\theta_*,t)} \right]^2 s^{(0)}(\theta_0,t)\lambda_0(t)dt \qquad (5.3)$$

$$= \int_0^1 \left[\left\{ \int_X \phi(x) v(x) w(x) e^{\theta_{\bullet}(x)} p_Y(t \mid x) h(x \mid t) dx \right\} / s^{(0)}(\theta_{\bullet}, t) \right]^2 s^{(0)}(\theta_0, t) \lambda_0(t) dt \quad .$$

Using Holder's inequality twice, A(iii), A(iv), and sup-norm bounds on $\theta_{\bullet}$, v and w,

$$\{ \int_X \phi(x) v(x) w(x) e^{\theta_{\bullet}(x)} p_Y(t \mid x) h(x \mid t) dx \}^2$$

$$\lesssim \sup_x \mid v(x) \mid \cdot \sup_x \mid w(x) \mid \cdot \sup_x \mid e^{\theta_{\bullet}(x)} \mid$$

$$\cdot \int_X \phi(x)^2 dx \cdot \int_X e^{\theta_{\bullet}(x)} p_Y(t \mid x) h(x \mid t) dx$$

$$\lesssim R^2 \cdot \|\phi\|^2_{L_2} \cdot s^{(0)}(\theta_{\bullet}, t) \quad .$$

But by Lemma 2.1(a) $\dfrac{s^{(0)}(\theta_\mu, t)}{s^{(0)}(\theta_0, t)} \approx 1$, also $\Lambda_0(1) < \infty$, therefore (5.3) is bounded by a constant times $R^2 \cdot \|\phi\|^2_{L_2}$. A similar type analysis works for the other terms and the result in equation (5.2) follows.

To complete the proof, Lemma 3.6(i) implies $\|\phi_{0\mu}\|^2_{L_2} = 2$, so

$$\sum_{\nu=1}^{\infty} [1 + \gamma_{0\nu}^b] \, [1 + 2\mu\gamma_{0\nu}]^{-2} \, \{R^2 \|\phi_{0\nu}\|^2_{L_2}\} \lesssim R^2 \sum_{\nu=1}^{\infty} [1 + \gamma_{0\nu}^b] \, [1 + 2\mu\gamma_{0\nu}]^{-2}$$

$$\approx R^2 \mu^{-(b + d/2m)} \quad ,$$

since $0 \leq b < 2 - d/2m$. The result is uniform in $\theta_{\bullet}$ in sup-norm balls about θ_0. This proves part (a).

The proof of part (b) is very similar.

$$D^3 l_{n\mu}(\theta_{\bullet}) vw = G_\mu^{-1}(\theta_0) \dot{U}_n(\theta_{\bullet}) vw \quad ,$$

where $\dot{U}_n(\theta_{\bullet}) vw$ is given by:

$$\dot{U}_n(\theta_{\bullet}) vw = \int_0^1 \left[\frac{S^{(3)}(\theta_{\bullet}, t) vw}{S^{(0)}(\theta_{\bullet}, t)} \right.$$

$$- \frac{S^{(2)}(\theta_{\bullet}, t) vw}{S^{(0)}(\theta_{\bullet}, t)} \frac{S^{(1)}(\theta_{\bullet}, t)}{S^{(0)}(\theta_{\bullet}, t)}$$

$$- \frac{< S^{(2)}(\theta_*, t), v >}{S^{(0)}(\theta_*, t)} \frac{< S^{(1)}(\theta_*, t), w >}{S^{(0)}(\theta_*, t)}$$
$$- \frac{< S^{(2)}(\theta_*, t), w >}{S^{(0)}(\theta_*, t)} \frac{< S^{(1)}(\theta_*, t), v >}{S^{(0)}(\theta_*, t)}$$
$$+ \frac{< S^{(1)}(\theta_*, t), v >}{S^{(0)}(\theta_*, t)} \frac{< S^{(1)}(\theta_*, t), w >}{S^{(0)}(\theta_*, t)} \frac{S^{(1)}(\theta_*, t)}{S^{(0)}(\theta_*, t)} \Bigg] d\overline{N}(t)$$

Again expanding in terms of the eigensystem $\{\phi_{0\nu}, \gamma_{0\nu} ; \nu = 1,2, \cdots \}$,

$$||G_\mu^{-1}(\theta_0) D^3 l_{n\mu}(\theta_*) vw ||_b^2 = \sum_{\nu=1}^{\infty} [1 + \gamma_{0\nu}^b] \, [1 + 2\mu\gamma_{0\nu}]^{-2} \cdot \{ < \dot{U}_n(\theta_*) vw, \phi_{0\nu} > \}^2 \tag{5.4}$$

Using sup-norm bounds on $\theta_* \in S_{\theta_0}(R, C(\underline{X}, \boldsymbol{R}))$, v and w and Holder's inequality, it can be shown that for $\phi \in \Theta$,

$$\{ < \dot{U}_n(\theta_*) vw, \phi > \}^2 \lesssim R^2 \cdot B^2(n, \mu, \phi) \tag{5.5}$$

where

$$B(n, \mu, \phi) = \int_0^1 \left[\left\{ \frac{1}{n} \sum_{i=1}^{n} Y_i(t) \,|\, \phi(X_i(t) \,|\, e^{\theta_0(X_i(t))} \right\} / S^{(0)}(\theta_\mu, t) \right] d\overline{N}(t) .$$

A direct computation, using the martingale structure, the Cauchy-Schwartz inequality and Lemma 2.1(a), yields:

$$\mathrm{E}B^2(n, \mu, \phi) \lesssim \frac{1}{n} \mathrm{E} \int_0^1 \left[\frac{1}{n} \sum_{i=1}^{n} Y_i(t) \phi^2(X_i(t)) e^{\theta_0(X_i(t))} \right] \lambda_0(t) dt$$
$$+ \mathrm{E} \int_0^1 \left[\frac{1}{n} \sum_{i=1}^{n} Y_i(t) \phi^2(X_i(t)) e^{2\theta_0(X_i(t))} \right] \lambda_0(t) dt .$$
$$\approx \frac{1}{n} ||\phi||_{L_2}^2 + ||\phi||_{L_2}^2 \approx ||\phi||_{L_2}^2 . \tag{5.6}$$

So with $B^2(n, \mu) = \sum_{\nu=1}^{\infty} [1 + \gamma_{\mu\nu}^b][1 + 2\mu\gamma_{\mu\nu}]^{-2} B^2(n, \mu, \phi_{\nu\mu})$, and $0 \le b < 2 - d/2m$,

$K_3(n,\mu,R) \lesssim R^2\, B^2(n,\mu)$ where $\mathrm{E}B^2(n,\mu) \lesssim \mu^{-(b+d/2m)}$. *Q.E.D.*

Proof of Lemma 5.2. (Convergence of $I_{n\,\mu}(\theta_\mu)$.)

$$I_{n\,\mu}(\theta_\mu) - G_\mu(\theta_\mu) = \int_0^1 \left\{ \frac{S^{(2)}(\theta_\mu,t)}{S^{(0)}(\theta_\mu,t)} - \frac{S^{(1)}(\theta_\mu,t)}{S^{(0)}(\theta_\mu,t)} \times \frac{S^{(1)}(\theta_\mu,t)}{S^{(0)}(\theta_\mu,t)} \right\} d\overline{N}(t) \tag{5.7}$$

$$- \int_0^1 \left\{ \frac{s^{(2)}(\theta_\mu,t)}{s^{(0)}(\theta_\mu,t)} - \frac{s^{(1)}(\theta_\mu,t)}{s^{(0)}(\theta_\mu,t)} \times \frac{s^{(1)}(\theta_\mu,t)}{s^{(0)}(\theta_\mu,t)} \right\} s^{(0)}(\theta_0,t)\lambda_0(t)\,dt \;.$$

Writing this in terms of $\overline{M}(t)$,

$$I_{n\,\mu}(\theta_\mu) - G_\mu(\theta_\mu) = \int_0^1 \left\{ \frac{S^{(2)}(\theta_\mu,t)}{S^{(0)}(\theta_\mu,t)} - \frac{S^{(1)}(\theta_\mu,t)}{S^{(0)}(\theta_\mu,t)} \times \frac{S^{(1)}(\theta_\mu,t)}{S^{(0)}(\theta_\mu,t)} \right\} d\overline{M}(t)$$

$$+ \int_0^1 \left\{ \frac{S^{(2)}(\theta_\mu,t)}{S^{(0)}(\theta_\mu,t)} - \frac{S^{(1)}(\theta_\mu,t)}{S^{(0)}(\theta_\mu,t)} \times \frac{S^{(1)}(\theta_\mu,t)}{S^{(0)}(\theta_\mu,t)} \right\} S^{(0)}(\theta_0,t)\lambda_0(t)\,dt$$

$$- \int_0^1 \left\{ \frac{s^{(2)}(\theta_\mu,t)}{s^{(0)}(\theta_\mu,t)} - \frac{s^{(1)}(\theta_\mu,t)}{s^{(0)}(\theta_\mu,t)} \times \frac{s^{(1)}(\theta_\mu,t)}{s^{(0)}(\theta_\mu,t)} \right\} s^{(0)}(\theta_0,t)\lambda_0(t)\,dt \;.$$

To analyze $K_2(n,\mu)$, let $u = \sum_{\nu=1}^{\infty} u_\nu \phi_{\mu\nu}$ and $v = \sum_{\nu^*=1}^{\infty} v_{\nu^*} \phi_{\mu\nu^*}$ be unit elements in Θ_b and Θ_{b^*} respectively, i.e.

$$\sum_{\nu=1}^{\infty} u_\nu^2\, [1+\gamma_{\mu\nu}^{b}] = \sum_{\nu^*=1}^{\infty} v_{\nu^*}^2\, [1+\gamma_{\mu\nu^*}^{b^*}] = 1 \quad .$$

A direct computation using the Penalty Information scale gives

$$< u, G_\mu^{-1}(\theta_\mu)[I_{n\,\mu}(\theta_\mu) - G_\mu(\theta_\mu)] v >_b \; = \sum_{\nu=1}^{\infty} \sum_{\nu^*=1}^{\infty} u_\nu [1+\gamma_{\mu\nu}^{b}]^{1/2} \{ \Delta_{\nu\nu^*}^{(1)} + \Delta_{\nu\nu^*}^{(2)} \} v_{\nu^*} [1+\gamma_{\mu\nu^*}^{b^*}]^{1/2} \;.$$

where (to simplify notation, let $\phi = \phi_{\mu\nu}$ and $\psi = \phi_{\mu\nu^*}$)

$$\Delta_{\nu\nu^*}^{(1)} = [1+\gamma_{\mu\nu}^{b}]^{1/2}\, [1+2\mu\gamma_{\mu\nu}]^{-1}\, [1+\gamma_{\mu\nu^*}^{b^*}]^{-1/2} \tag{5.8}$$

$$\cdot \left\{ \int_0^1 \left[\frac{S^{(2)}(\theta_\mu,t)\phi\psi}{S^{(0)}(\theta_\mu,t)} - \frac{<S^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)} \frac{<S^{(1)}(\theta_\mu,t),\psi>}{S^{(0)}(\theta_\mu,t)} \right] d\overline{M}(t) \right\} .$$

and

$$\Delta_{\nu\nu^*}^{(2)} = [1+\gamma_{\mu\nu}^{b}]^{1/2}\, [1+2\mu\gamma_{\mu\nu}]^{-1}\, [1+\gamma_{\mu\nu^*}^{b^*}]^{-1/2} \tag{5.9}$$

$$\cdot \left\{ \int_0^1 \left[\frac{S^{(2)}(\theta_\mu,t)\phi\psi}{S^{(0)}(\theta_\mu,t)} - \frac{<S^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)} \frac{<S^{(1)}(\theta_\mu,t),\psi>}{S^{(0)}(\theta_\mu,t)} \right] S^{(0)}(\theta_0,t)\lambda_0(t)\,dt \right.$$

$$\left. - \int_0^1 \left[\frac{s^{(2)}(\theta_\mu,t)\phi\psi}{s^{(0)}(\theta_\mu,t)} - \frac{<s^{(1)}(\theta_\mu,t),\phi>}{s^{(0)}(\theta_\mu,t)} \frac{<s^{(1)}(\theta_\mu,t),\psi>}{s^{(0)}(\theta_\mu,t)} \right] s^{(0)}(\theta_0,t)\lambda_0(t)\,dt \right\} .$$

Applying the Cauchy-Schwartz inequality and the fact that u and v have unit norms

$$<u, G_\mu^{-1}(\theta_\mu)[I_{n\mu}(\theta_\mu) - G_\mu(\theta_\mu)]v>_b^2 \lesssim \sum_{\nu=1}^{\infty} u_\nu^2 [1+\gamma_{\mu\nu}^{b}] \sum_{\nu=1}^{\infty}\sum_{\nu^*=1}^{\infty} \{\Delta_{\nu\nu^*}^{(1)} + \Delta_{\nu\nu^*}^{(2)}\}^2 \sum_{\nu^*=1}^{\infty} v_{\nu^*}^2 [1+\gamma_{\mu}^{b^*}$$

$$\leq \sum_{\nu=1}^{\infty}\sum_{\nu^*=1}^{\infty} \{[\Delta_{\nu\nu^*}^{(1)}]^2 + [\Delta_{\nu\nu^*}^{(2)}]^2\} . \tag{5.10}$$

It will be shown that

(i) $\quad \mathrm{E}\sum_{\nu=1}^{\infty}\sum_{\nu^*=1}^{\infty} [\Delta_{\nu\nu^*}^{(1)}]^2 \lesssim n^{-1}\mu^{-(b+d/2m)}$;

(ii) $\quad \mathrm{E}\sum_{\nu=1}^{\infty}\sum_{\nu^*=1}^{\infty} [\Delta_{\nu\nu^*}^{(2)}]^2 \lesssim n^{-1}\mu^{-(b+d/2m)}$.

From this it will follow that, with $B_2(n,\mu) = \sum_{\nu=1}^{\infty}\sum_{\nu^*=1}^{\infty} [\Delta_{\nu\nu^*}^{(1)} + \Delta_{\nu\nu^*}^{(2)}]$, $K_2(n\,\mu) \lesssim B_2(n,\mu)$ and $\mathrm{E}B_2^2(n,\mu) \lesssim n^{-1}\mu^{-(b+d/2m)}$.

The proofs of (i) and (ii) are accomplished as follows:

$$\Delta_{\nu\nu^*}^{(1)} = [1+\gamma_{\mu\nu}^{b}]^{1/2}\, [1+2\mu\gamma_{\mu\nu}]^{-1}\, [1+\gamma_{\mu\nu^*}^{b^*}]^{-1/2}\, \{\delta_{\nu\nu^*}^{(1)}\}$$

$$\Delta_{\nu\nu^*}^{(2)} = [1+\gamma_{\mu\nu}^{b}]^{1/2}\, [1+2\mu\gamma_{\mu\nu}]^{-1}\, [1+\gamma_{\mu\nu^*}^{b^*}]^{-1/2}\, \{\delta_{\nu\nu^*}^{(2)}\}$$

where the $\delta_{\nu\nu^*}$'s are the terms in curly brackets in equations (5.8) and (5.9). Suppose

$$\mathrm{E}\{\delta_{\nu\nu^*}^{(1)}\}^2 \lesssim \frac{1}{n}\, ||\phi||_{L_2}^2\, ||\psi||_{sup}^2 \tag{5.11}$$

and

$$\mathrm{E}\{\delta_{\nu\nu^*}^{(2)}\}^2 \lesssim \frac{1}{n}\, ||\phi||_{L_2}^2\, ||\psi||_{sup}^2 \quad , \tag{5.12}$$

then since $||\phi_{\mu\nu}||_{L_2}^2 = 2$ and $||\phi_{\mu\nu}||_{sup}^2 = [1+\gamma_{\mu\nu^*}^{d/2m}]$, (the latter follows from Sobolev's Imbedding Theorem and Lemma 3.5(i)), $\mathrm{E}\sum_{\nu=1}^{\infty}\sum_{\nu^*=1}^{\infty}[\Delta_{\nu\nu^*}^{(1)}]^2$ and $\mathrm{E}\sum_{\nu=1}^{\infty}\sum_{\nu^*=1}^{\infty}[\Delta_{\nu\nu^*}^{(2)}]^2$ are both bounded above by a constant multiple of

$$n^{-1}\sum_{\nu=1}^{\infty}[1+\gamma_{\mu\nu}^{b}][1+2\mu\gamma_{\mu\nu}]^{-2} \cdot \sum_{\nu^*=1}^{\infty}[1+\gamma_{\mu\nu^*}^{b^*}]^{-1}[1+\gamma_{\mu\nu^*}^{d/2m}] \quad .$$

But $b^* > d/m$ and $\gamma_{\mu\nu} \approx \nu^{2m/d}$, so $\sum_{\nu^*=1}^{\infty}[1+\gamma_{\mu\nu^*}^{b^*}]^{-1}[1+\gamma_{\mu\nu^*}^{d/2m}] < \infty$. Hence, since $0 \leq b \leq 2\text{-}d/2m$, $\mathrm{E}B_2^2(n,\mu) \lesssim n^{-1}\mu^{-(b+d/2m)}$, the result is proved.

It remains to establish (5.11) and (5.12). Computing the predictable variation in (5.11) gives:

$$\mathrm{E}\{\delta_{\nu\nu^*}^{(1)}\}^2 = \frac{1}{n}\,\mathrm{E}\int_0^1\left[\frac{S^{(2)}(\theta_\mu,t)\phi\psi}{S^{(0)}(\theta_\mu,t)} - \frac{<S^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)}\frac{<S^{(1)}(\theta_\mu,t),\psi>}{S^{(0)}(\theta_\mu,t)}\right]^2 S^{(0)}(\theta_0,t)\lambda_0(t)\,dt \quad .$$

With $p_i(t) = \dfrac{\frac{1}{n}Y_i(t)e^{\theta_\mu(X_i(t))}}{\frac{1}{n}\sum_{j=1}^{n}Y_j(t)e^{\theta_\mu(X_j(t))}}$ for $i=1,2,\cdots$, the term in square brackets can be written as a covariance i.e.

$$\left[\frac{S^{(2)}(\theta_\mu,t)\phi\psi}{S^{(0)}(\theta_\mu,t)} - \frac{<S^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)}\frac{<S^{(1)}(\theta_\mu,t),\psi>}{S^{(0)}(\theta_\mu,t)}\right]^2 = \{\sum_{i=1}^{n}p_i(t)[\phi(X_i(t))\text{-}\overline{\phi}_t]\psi(X_i(t))\}^2;$$

where $\overline{\phi}_t = \sum_{j=1}^{n}\phi(X_j(t))p_j(t)$. This in turn is bounded by

$$\sup_x \ | \psi(x) | \ \{ \sum_{i=1}^n p_i(t) \, | \phi(X_i(t)) | \}^2 \ .$$

However,

$$\{ \sum_{i=1}^n p_i(t) \, | \phi(X_i(t)) | \}^2 \cdot S^{(0)}(\theta_\mu, t) = \frac{\left[\frac{1}{n} \sum_{i=1}^n Y_i(t) \, | \phi(X_i(t)) | \, e^{\theta_\mu(X_i(t))} \right]^2}{S^{(0)}(\theta_\mu, t)^2} S^{(0)}(\theta_0, t)$$

$$\approx \left\{ \frac{1}{n} \sum_{i=1}^n Y_i(t) \, | \phi(X_i(t)) | \, e^{\theta_\mu(X_i(t))} \right\}^2 / S^{(0)}(\theta_\mu, t) \qquad \text{by Lemma 2.1(a).}$$

Using the Cauchy-Schwartz inequality and Lemma 2.1(a), this is bounded by

$$\frac{1}{n} \sum_{i=1}^n Y_i(t) \phi^2(X_i(t)) \quad .$$

Thus computing the expectation of this term and using the bounds on h, p_Y and λ_0, one has that

$$\mathrm{E}\{\delta_{\nu\nu^*}^{(1)}\}^2 \lesssim \frac{1}{n} \, ||\phi||_{L_2}^2 \, ||\psi||_{sup}^2 \ ,$$

which proves (5.11).

For (5.12)

$$\begin{aligned} \mathrm{E}\{\delta_{\nu\nu^*}^{(2)}\}^2 \le \mathrm{E} & \left\{ \int_0^1 \left[\frac{S^{(2)}(\theta_\mu, t) \phi \psi S^{(0)}(\theta_0, t)}{S^{(0)}(\theta_\mu, t)} - \frac{s^{(2)}(\theta_\mu, t) \phi \psi s^{(0)}(\theta_0, t)}{s^{(0)}(\theta_\mu, t)} \right] \lambda_0(t) dt \right\}^2 \\ & + \mathrm{E} \left\{ \int_0^1 \left[\frac{< S^{(1)}(\theta_\mu, t), \phi >}{S^{(0)}(\theta_\mu, t)} \cdot \frac{< S^{(1)}(\theta_\mu, t), \psi >}{S^{(0)}(\theta_\mu, t)} S^{(0)}(\theta_0, t) \right. \right. \qquad (5.13) \\ & \left. \left. - \ \frac{< s^{(1)}(\theta_\mu, t), \phi >}{s^{(0)}(\theta_\mu, t)} \cdot \frac{< s^{(1)}(\theta_\mu, t), \psi >}{s^{(0)}(\theta_\mu, t)} s^{(0)}(\theta_\mu, t) \right] \lambda_0(t) dt \right\}^2 \ . \end{aligned}$$

The first term in brackets is bounded by

$$\left\{\int_0^1 s^{(2)}(\theta_\mu,t)\phi\psi \left\{\frac{S^{(0)}(\theta_\mu,t)}{S^{(0)}(\theta_\mu,t)} - \frac{s^{(0)}(\theta_\mu,t)}{s^{(0)}(\theta_\mu,t)}\right\}\lambda_0(t)dt\right\}^2 + E\left\{\int_0^1 E^{(2)}(t,\mu)\phi\psi\left\{\frac{S^{(0)}(\theta_0,t)}{S^{(0)}(\theta_\mu,t)}\right\}\lambda_0(t)dt\right\}^2$$

where $E^{(2)}(t,\mu)\phi\psi = [S^{(2)}(\theta_\mu,t)\phi\psi - s^{(2)}(\theta_\mu,t)\phi\psi]$. Applying the Cauchy-Schwartz inequality and Lemma 2.1(a), this is bounded by

$$\int_0^1 [s^{(2)}(\theta_\mu,t)\phi\psi]^2\lambda_0(t)dt \cdot E\int_0^1\left\{\frac{S^{(0)}(\theta_\mu,t)}{S^{(0)}(\theta_\mu,t)} - \frac{s^{(0)}(\theta_\mu,t)}{s^{(0)}(\theta_\mu,t)}\right\}^2\lambda_0(t)dt + E\int_0^1 [E^{(2)}(t,\mu)\phi\psi]^2\lambda_0(t)dt \quad .$$

Further applications of the Cauchy Schwartz inequality, Lemma 2.1(c) and an analysis similar to that used for equation (4.5) in the previous section, results in the upper bound

$$n^{-1}\{\|\phi\|^2_{L_2}\|\psi\|^2_{L_2} + \|\phi\psi\|^2_{L_2}\} \quad .$$

But $\|\phi\psi\|^2_{L_2} \lesssim \|\phi\|^2_{L_2}\|\psi\|^2_{sup}$ so the first term in (5.13)

$$\lesssim n^{-1}\{\|\phi\|^2_{L_2}\|\psi\|^2_{sup}\} \quad .$$

The second term in (5.13) is

$$E\left\{\int_0^1\left[\frac{<S^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)}\cdot\frac{<S^{(1)}(\theta_\mu,t),\psi>}{S^{(0)}(\theta_\mu,t)}S^{(0)}(\theta_0,t)\right.\right.$$

$$\left.\left.-\frac{<s^{(1)}(\theta_\mu,t),\phi>}{s^{(0)}(\theta_\mu,t)}\cdot\frac{<s^{(1)}(\theta_\mu,t),\psi>}{s^{(0)}(\theta_\mu,t)}s^{(0)}(\theta_\mu,t)\right]\lambda_0(t)dt\right\}^2 \quad .$$

Adding and subtracting $\left\{\frac{<s^{(1)}(\theta_\mu,t),\phi>}{s^{(0)}(\theta_\mu,t)}\frac{<S^{(1)}(\theta_\mu,t),\psi>}{s^{(0)}(\theta_\mu,t)}\right\}S^{(0)}(\theta_0,t)$, this is bounded by

$$E\left\{\int_0^1\left[\frac{<s^{(1)}(\theta_\mu,t),\phi>}{s^{(0)}(\theta_\mu,t)}\right]\cdot\left[\frac{<S^{(1)}(\theta_\mu,t),\psi>}{S^{(0)}(\theta_\mu,t)}S^{(0)}(\theta_0,t) - \frac{<s^{(1)}(\theta_\mu,t),\psi>}{s^{(0)}(\theta_\mu,t)}s^{(0)}(\theta_0,t)\right]\lambda_0(t)dt\right\}^2$$

$$+ \mathrm{E}\left\{\int_0^1 \left[\frac{<S^{(1)}(\theta_\mu,t),\psi>}{S^{(0)}(\theta_\mu,t)}S^{(0)}(\theta_0,t)\right]\left[\frac{<S^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)} - \frac{<s^{(1)}(\theta_\mu,t),\phi>}{s^{(0)}(\theta_\mu,t)}\right]\lambda_0(t)dt\right\}^2 .$$

By a familiar argument,

$$\lesssim \|\phi\|_{L_2}^2 \mathrm{E}\int_0^1 \left[\frac{<S^{(1)}(\theta_\mu,t),\psi>}{S^{(0)}(\theta_\mu,t)}S^{(0)}(\theta_0,t) - \frac{<s^{(1)}(\theta_\mu,t),\psi>}{s^{(0)}(\theta_\mu,t)}s^{(0)}(\theta_0,t)\right]^2 \lambda_0(t)dt$$

$$+ \mathrm{E}\int_0^1 [<S^{(1)}(\theta_*,t),\psi>]^2\lambda_0(t)dt \int_0^1 \left[\frac{<S^{(1)}(\theta_\mu,t),\phi>}{S^{(0)}(\theta_\mu,t)} - \frac{<s^{(1)}(\theta_\mu,t),\phi>}{s^{(0)}(\theta_\mu,t)}\right]^2 \lambda_0(t)dt$$

But $<S^{(1)}(\theta_*,t),\psi>^2 \lesssim \|\psi\|_{sup}^2$ uniformly in $t \in [0,1]$ and $\theta_* \in S_{\theta_0}(R,C)$, so following an analysis similar to that used for (4.5) yields

$$\lesssim \|\phi\|_{L_2}^2 \{\frac{1}{n}\|\psi\|_{L_2}^2\} + \|\psi\|_{sup}^2\{\frac{1}{n}\|\phi\|_{L_2}^2\} \lesssim \frac{1}{n}\|\phi\|_{L_2}^2 \|\psi\|_{sup}^2 .$$

This completes the proof of the result. *Q.E.D.*

References

1. Adams, R., ***Sobolev Spaces,*** Academic Press, New York, 1975.

2. Agmon, S., ***Elliptic Boundary Value Problems,*** Van Nostrand, 1965.

3. Andersen, P. K. and Gill, R. D., "Cox's regression model for counting processes: a large sample study," ***Ann. Statist.***, vol. 10, no. 4, pp. 1100-1120, 1982.

4. Anderson, J. A. and Senthilselvan, A., "Smooth estimates for the hazard function," ***J. R. Statist. Soc. B.***, vol. 42, no. 3, pp. 322-327, 1980.

5. Cox, D. D., "Approximation of the method of regularization estimators," Tech. Rep. No. 723, Statistics Dept., University of Wisconsin-Madison, 1983.

6. Cox, D. D. and O'Sullivan, F., "Analysis of penalized likelihood type estimators wih application to generalized smoothing in Sobolev Spaces," Tech. Rep. No. 51 (Submitted to the Annals of Statistics), Statistics Dept., University of California-Berkeley, 1985.

7. Cox, D. R., "Regression models and life tables (with discussion).," ***J. Roy. Statist. Soc. B.***, vol. 34, pp. 187-220, 1972.

8. Gill, R. D., "Understanding Cox's regression model: A martingale approach.," ***J. Amer. Statist. Assoc.***, vol. 79, pp. 441-447, 1984.

9. Good, I. J. and Gaskins, R. A., "Non-parametric roughness penalties for probability densities," ***Biometrika,*** vol. 58, pp. 255-277, 1971.

10. Hastie, T. J. and Tibshirani, R. J., "Generalized additive models (with discussion)," ***J. Statist. Sci. (to appear),*** 1986.

11. Huber, P. J., ***Robust Statistics,*** John Wiley & Sons, New York, 1981.

12. Kato, T., ***Perturbation Theory for Linear Operators,*** Die Grundlehren der mathematischen Wissenschaften 132, Springer-Verlag, New York, 1966.

13. Leonard, T., "An empirical Bayesian approach to the smooth estimation of unknown functions," Tech. Rep. No. 2339, Math. Research Center, University of Wisconsin-Madison, 1982.

14. O'Sullivan, F., "Estimation of densities and hazards by the method of penalized maximum likelihood," Tech. Rep. No. 58, Statistics Dept., University of California-Berkeley, 1986.

15. O'Sullivan, F., Yandell, B., and Raynor, W. J., "Automatic smoothing of regression functions in generalized linear models," *J. Amer. Statist. Assoc.*, vol. 81, no. 393, pp. 96-104, 1986.

16. Robelledo, R., "Central limit theorems for local martingales," *Z. Wahrsch. verw. Gebiete*, vol. 51, pp. 269-286, 1980.

17. Silverman, B. W., "On the estimation of a probability density function by the maximum penalised likelihood method," *Ann. Statist.*, vol. 10, no. 3, pp. 795-810, 1982.

18. Stone, C. J., "Contribution to the discussion of the paper by T. Hastie and R. Tibshirani (to appear)," *J. Statist. Sci.*, 1986.

19. Tikhonov, A. and Arsenin, V., *Solutions of Ill-Posed Problems,* Wiley, New York, 1977.

20. Triebel, H., *Interpolation Theory, Function Spaces, Differential Operators,* North-Holland, New York, 1978.

21. Wahba, G., "Cross-validated spline methods for the estimation of multivariate functions from data on functionals," in *Statistics: An Appraisal*, ed. H. A. David and H. T. David, pp. 205-233, The Iowa State University Press, 1984.

22. Weinberger, H. F., "Variational Methods for Eigenvalue Problems," in *Lecture Notes by G. P. Schwartz*, Department of Mathematics, University of Minnesota,

Minneapolis, 1962.

23. Weinberger, H. F., "Variational Methods for Eigenvalue Approximation," in *CBMS Regional conference series in applied mathematics*, SIAM, Philadelphia, 1974.