Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/uasa20

Inference for a Class of Transformed Hazards Models

Donglin Zeng[a], Guosheng Yin[a] & Joseph G Ibrahim[a]

[a] Donglin Zeng is Assistant Professor and Joseph G. Ibrahim is Professor , Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599. Guosheng Yin is Assistant Professor, Department of Biostatistics and Applied Mathematics, M. D. Anderson Cancer Center, The University of Texas, Houston, TX 77030 . The authors thank the two referees for their critical and helpful comments.
Published online: 01 Jan 2012.

To cite this article: Donglin Zeng, Guosheng Yin & Joseph G Ibrahim (2005) Inference for a Class of Transformed Hazards Models, Journal of the American Statistical Association, 100:471, 1000-1008, DOI: 10.1198/016214504000001637

To link to this article: http://dx.doi.org/10.1198/016214504000001637

Inference for a Class of Transformed Hazards Models

Donglin ZENG, Guosheng YIN, and Joseph G. IBRAHIM

A new class of transformed hazard rate models is considered that contains both the multiplicative hazards model and the additive hazards model as special cases. The sieve maximum likelihood estimators are derived for the model parameters, and the estimators for the regression coefficients are shown to be consistent and asymptotically normal with variance achieving the semiparametric efficiency bound. Simulation studies are conducted to examine the small-sample properties of the proposed estimates, and a real dataset is used to illustrate our approach.

KEY WORDS: Box–Cox transformation; Sieve estimation; Transformed hazard rate; Wavelet approximation.

1. INTRODUCTION

In survival analysis, the Cox multiplicative hazards model (Cox 1972) has been used extensively. In this model, the hazard rate function of the survival time given an external (possibly time-dependent) covariate vector $\mathbf{Z}(t)$ is assumed to be

$$\lambda(t|\mathbf{Z}(t)) = \lambda(t)\exp\{\boldsymbol{\beta}^T\mathbf{Z}(t)\},$$

where $\lambda(t)$ is an unknown and unspecified baseline hazard function and $\boldsymbol{\beta}$ is the regression coefficient for $\mathbf{Z}(t)$. An efficient estimate for $\boldsymbol{\beta}$ can be obtained by maximizing a partial likelihood function (Cox 1975; Andersen and Gill 1982). Because the proportionality in the multiplicative hazards model does not hold in many applications, one alternative form of modeling the hazard rate function is to assume that the hazard risks are additive across covariates, that is,

$$\lambda(t|\mathbf{Z}(t)) = \mu(t) + \boldsymbol{\beta}^T\mathbf{Z}(t),$$

where $\mu(t)$ is an unknown baseline hazard function. The additive hazards model has been studied by Lin and Ying (1994). Furthermore, to accommodate both the multiplicative and additive hazards structures, Lin and Ying (1995) proposed a multiplicative-additive hazards model where the hazard function takes the form

$$\lambda\big(t|\mathbf{Z}_1(t), \mathbf{Z}_2(t)\big) = \lambda(t)\exp\{\boldsymbol{\beta}_1^T\mathbf{Z}_1(t)\} + \boldsymbol{\beta}_2^T\mathbf{Z}_2(t),$$

where $\mathbf{Z}_1(t)$ and $\mathbf{Z}_2(t)$ are different covariates of $\mathbf{Z}(t)$. But all of these hazard-based regression models are restrictive in practice, because they may not be flexible enough to entertain situations where hazard risks are neither multiplicative nor additive among groups. Therefore, it is desirable to obtain a class of hazard-based models that allows a wide range of hazard structures while at the same time retaining the simple structures of the multiplicative and additive hazards models.

In this article we propose a unified family of hazard-based regression models. We propose a class of transformed hazards models by imposing both an additive structure and a known transformation $G(\cdot)$ on the hazard function. In this class, the hazard function for the survival times given covariate $\mathbf{Z}(t)$ takes the form

$$G\{\lambda(t|\mathbf{Z}(t))\} = \mu(t) + \boldsymbol{\beta}^T\mathbf{Z}(t), \tag{1}$$

where $\boldsymbol{\beta}$ is the unknown regression coefficient vector, $\mu(t)$ is an unknown baseline hazard function, and $G(\cdot)$ is a known and increasing transformation function. Essentially, model (1) can be considered a partial linear regression model for the transformed hazard function. One example of the transformation $G(\cdot)$ is the Box–Cox transformation (Box and Cox 1964), in which $G(x)$ is given by

$$G(x) = (x^s - 1)/s \tag{2}$$

for $s > 0$ and we define $G(x) = \log(x)$ if $s = 0$. Within the Box–Cox transformation family, when $s = 1$ in (2), (1) is the additive hazards model, and if $s = 0$, then (1) becomes the multiplicative hazards model. Thus the transformed model in (1) with $G(\cdot)$ given by (2) can be considered a smoothed class of hazards models linking the additive and multiplicative hazards models, which are the extremes of this class if s is restricted to the range of $[0, 1]$. Because our proposed class (1) allows a much broader class of hazard patterns than are allowed in the proportional hazards and additive hazards models, it provides us with more flexible models for analyzing survival data.

Our goal in this article is to provide a unified framework for deriving an efficient estimate for $\boldsymbol{\beta}$ in model (1) for any given transformation G, where G^{-1} is continuously three times differentiable. In particular, we use the sieve maximum likelihood estimation approach to construct an estimate of $\boldsymbol{\beta}$. We then examine the asymptotic properties of the resulting estimator.

The rest of this article is organized as follows. In Section 2 we present a general framework of sieve maximum likelihood estimation. In Section 3 we derive the asymptotic properties of the estimator, including consistency and asymptotic normality. In Section 4 we report on simulation studies that we conducted to examine the numerical properties of the proposed method in small samples. In Section 5 we analyze a lung cancer dataset using the proposed class of models and estimation procedure. We present a brief discussion in Section 6, and provide proofs of all theorems in the Appendix.

2. INFERENCE PROCEDURE

Suppose that we observe survival data with n iid observations in a study with termination time τ. We denote the observation for subject i by $(Y_i = T_i \wedge C_i, \Delta_i = I(T_i \le C_i), \{\mathbf{Z}_i(t) : t \in [0, \tau]\})$, where T_i is the failure time of subject i, C_i is the censoring time, $\{\mathbf{Z}_i(t) : t \in [0, \tau]\}$ denotes the external covariate process, "$\wedge$" denotes the minimum of two values, and $I(\cdot)$ is the indicator function.

Donglin Zeng is Assistant Professor (E-mail: *dzeng@bios.unc.edu*) and Joseph G. Ibrahim is Professor (E-mail: *ibrahim@bios.unc.edu*), Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599. Guosheng Yin is Assistant Professor, Department of Biostatistics and Applied Mathematics, M. D. Anderson Cancer Center, The University of Texas, Houston, TX 77030 (E-mail: *gyin@odin.mdacc.tmc.edu*). The authors thank the two referees for their critical and helpful comments.

Journal of the American Statistical Association
September 2005, Vol. 100, No. 471, Theory and Methods
DOI 10.1198/016214504000001637

We assume that C_i is independent of T_i conditional on the covariates. Under the assumption that the transformation $G(\cdot)$ in the model (1) is strictly increasing and differentiable, the observed likelihood function of the parameters $(\boldsymbol{\beta}, \mu)$ can be written as

$$L_n(\boldsymbol{\beta}, \mu) = \prod_{i=1}^{n} \{H(\mu(Y_i) + \boldsymbol{\beta}^T \mathbf{Z}(Y_i))\}^{\Delta_i} \times \exp\left\{ - \int_0^{Y_i} H(\mu(t) + \boldsymbol{\beta}^T \mathbf{Z}(t))\, dt \right\}, \quad (3)$$

where $H(\cdot)$ is the inverse function of $G(\cdot)$.

To obtain estimates for $\boldsymbol{\beta}$ and $\mu(t)$, we wish to maximize $L_n(\boldsymbol{\beta}, \mu)$ in (3). But such a maximum does not exist, because one can always find some function $\mu(t)$ such that $L_n(\boldsymbol{\beta}, \mu) = \infty$. Therefore, we must restrict $\mu(t)$ to some smaller functional space to ensure that the maximum of $L_n(\boldsymbol{\beta}, \mu)$ exists. One important method of doing this is sieve maximum likelihood estimation, which has been used in many semiparametric estimation problems (Shen and Wong 1994; Shen 1997, 1998). In the sieve estimation method, the infinite-dimensional functional parameter $\mu(t)$ is restricted to a functional space with finite dimension, which is called the sieve space for $\mu(t)$. Moreover, the size of this sieve space increases with increasing sample size n, and as $n \to \infty$, the sieve space approximates the whole space for $\mu(t)$. However, for fixed sample size n, the choice of the sieve space for $\mu(t)$ cannot be arbitrary; the space should be chosen large enough so that the bias of the sieve estimate for $\mu(t)$ does not dominate. On the other hand, the space cannot be chosen too large so that the variation in estimating $\mu(t)$ dominates the variation in estimating $\boldsymbol{\beta}$, which is the main parameter of interest. Once a sieve space is chosen, maximizing the likelihood function can be carried out on this space, which contains only a finite number of parameters.

Usually, the sieve space for $\mu(t)$ is constructed from a linear space with a finite number of basis functions. Many basis functions can be used for this purpose. The most commonly used basis functions include B-splines and wavelet basis functions. In this article, we use wavelet basis functions to construct a sieve space for $\mu(t)$ for both mathematical and computational convenience, as is demonstrated in the subsequent arguments. A sequence of wavelet basis functions can be obtained from a single function $\phi(t)$, which is called the "father" wavelet and satisfies the following conditions:

(a) $\{\phi(t-k) : k \in \mathcal{Z}\}$ is an orthonormal system in $L_2(R)$, where $\mathcal{Z}$ consists of all the integers.

(b) Denote $V_j = \{\sum_k c_k \phi(2^j t - k) : \sum_k |c_k|^2 < \infty\}$ for any $j \geq 0$; then $V_0 \subset V_1 \subset \cdots$, and $\bigcup_{j \geq 0} V_j$ is dense in $L_2(R)$.

The sequence $\{V_j : j = 0, 1, \ldots\}$ is called a multiresolution approximation in the wavelet analysis (Mallat 1998, sec. 7.1). From (b), the basis functions $\{\phi(2^j t - k)\}$ from V_j for some suitable j can be candidates for constructing a sieve space. Furthermore, the orthogonality given in (a) concludes that the L_2 distance between any two functions in the sieve space can be expressed as the summed square difference of the coefficients of the basis functions, which does not hold for B-spline sieves. We note that V_j is still of infinite dimension. However, because our function $\mu(t)$ is of interest only for $t \in [0, \tau]$, the basis functions in V_j whose supports do not overlap with $[0, \tau]$ can be discarded. Thus the number of those remaining basis functions is finite, particularly if we choose $\phi(t)$ to have a compact support. Furthermore, $\phi(t)$ needs to be smooth to ensure the approximation of the sieve space to the whole space for $\mu(t)$. In summary, we assume that the father wavelet $\phi(t)$ satisfies the following:

(c) $\phi(t)$ has a finite support $[0, \tau]$ and $\phi \in W^{3,2}[0, \tau]$, where $W^{3,2}[0, \tau]$ is a Sobolev space containing all of the functions whose third derivatives are L_2-integrable in $[0, \tau]$ (cf. Adams 1975, chap. 1).

Typical choices of $\phi(t)$ satisfying (c) are the Daubechies wavelets (Daubechies 1992), after suitable shifting and scaling. In the commercial package MATLAB, the Wavelet toolbox provides a number of these choices.

After $\phi(t)$ is given, we can approximate the function $\mu(t), t \in [0, \tau]$, using the functions in the K_n-level multiresolution V_{K_n}. We choose the basis functions from $\{\phi(2^{K_n} t - k + 1) : 1 - \tau \leq k \leq 2^{K_n}\tau + 1\}$ whose supports overlap with $[0, \tau]$. Let $B_1(t), \ldots, B_{m_n}(t)$ denote these basis functions, where m_n is the number of integers between $1 - \tau$ and $2^{K_n}\tau + 1$. In addition, we impose an upper bound M_n for the summation of absolute values of all of the wavelet coefficients, to prevent the divergence of these coefficients in the maximization. As a result, a sieve space for the parameters $(\boldsymbol{\beta}, \mu)$ is proposed as

$$\mathcal{S}_n = \left\{ (\boldsymbol{\beta}, \mu(t)) : \mu(t) = \sum_{k=1}^{m_n} \alpha_k B_k(t), \; B_k(t) = \phi(2^{K_n} t - k + 1), \; \sum_{k=1}^{m_n} |\alpha_k| \leq M_n, \; \boldsymbol{\beta} \in \mathcal{B}_0, \mathcal{B}_0 \text{ is a known bounded open set containing the true value of } \boldsymbol{\beta} \right\},$$

where M_n is a constant depending on n. The choice of M_n is discussed in Section 3.

We thus maximize the likelihood function $L_n(\boldsymbol{\beta}, \mu)$ over $\mathcal{S}_n$. The maximization is carried out by an optimum search over the space

$$\left\{ (\boldsymbol{\beta}, \alpha_1, \ldots, \alpha_{m_n}) : \boldsymbol{\beta} \in \mathcal{B}_0, \sum_{k=1}^{m_n} |\alpha_k| \leq M_n \right\}.$$

Many optimization algorithms for estimating the parameters can be implemented. In particular, in the numerical computations of Section 4, we use the algorithm for searching the optimum in MATLAB. Details of the computational procedure are discussed in Section 4.

We denote the sieve maximum likelihood estimate for $(\boldsymbol{\beta}, \mu)$ by $(\hat{\boldsymbol{\beta}}, \hat{\mu})$. Our subsequent results show that $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ has an asymptotically normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$, which is equal to the semiparametric efficiency bound for $\boldsymbol{\beta}$. Unfortunately, $\boldsymbol{\Sigma}$ does not have an explicit expression. Thus, to estimate the asymptotic covariance of $\hat{\boldsymbol{\beta}}$, we propose the following sieve profile likelihood function. We define

$$pl_n(\boldsymbol{\beta}) = \max_{\mu \in \mathcal{S}_n} \log L_n(\boldsymbol{\beta}, \mu).$$

Then for any constant vector $\mathbf{e}$, we can approximate $\mathbf{e}^T\boldsymbol{\Sigma}^{-1}\mathbf{e}$ by

$$-\frac{1}{nh_n^2}\{pl_n(\hat{\boldsymbol{\beta}}+h_n\mathbf{e})-2pl_n(\hat{\boldsymbol{\beta}})+pl_n(\hat{\boldsymbol{\beta}}-h_n\mathbf{e})\},$$

where h_n is a constant of order $1/\sqrt{n}$. The sieve profile likelihood function imitates the profile likelihood function investigated by Murphy and van der Vaart (2000), and has been discussed by Fan and Wong (2000). Additionally, likelihood ratio inference based on the sieve likelihood function has been recently studied by Shen and Shi (2004) and Fan and Zhang (2004). Our simulation study in Section 4 shows that for moderate sample sizes, the profile sieve likelihood approach gives valid estimates of the variance.

3. ASYMPTOTIC PROPERTIES

We obtain the asymptotic properties for $\hat{\boldsymbol{\beta}}$ in this section. In particular, we show that the sieve maximum likelihood estimate $(\hat{\boldsymbol{\beta}}, \hat{\mu})$ is consistent under some suitable metric. Next we show that $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$ converges in distribution to a normal distribution and the asymptotic variance attains the semiparametric efficiency bound (cf. Bickle, Klaassen, Ritov, and Wellner 1993, chap. 3). All proofs are given in the Appendix.

To establish these results, we assume that the following conditions hold:

(C.1) With probability 1, $\{\mathbf{Z}(t): t\in[0,\tau]\}$ is a bounded process. Moreover, if there exists some vector $\tilde{\boldsymbol{\beta}}$ such that $\tilde{\boldsymbol{\beta}}^T\mathbf{Z}(t)=c(t)$ for some deterministic function $c(t)$, then $\tilde{\boldsymbol{\beta}}=\mathbf{0}$ and $c(t)=0$.

(C.2) C is independent of T given $\{\mathbf{Z}(t): t\in[0,\tau]\}$. Moreover, with probability 1,

$$\begin{aligned}&\inf_{\mathbf{z}(t),t\in[0,\tau]} P\big(C\ge\tau\,|\,\mathbf{Z}(t)=\mathbf{z}(t),t\in[0,\tau]\big)\\ &\quad=\inf_{\mathbf{z}(t),t\in[0,\tau]} P\big(C=\tau\,|\,\mathbf{Z}(t)=\mathbf{z}(t),t\in[0,\tau]\big)\\ &\quad>0.\end{aligned}$$

(C.3) Denote the true values of $(\boldsymbol{\beta},\mu)$ by $(\boldsymbol{\beta}_0,\mu_0)$. Assume that $\boldsymbol{\beta}_0\in\mathcal{B}_0$ and that $\mu_0(t)$ is a continuously three times differentiable function in $[0,\tau]$. Moreover, assume that with probability 1,

$$\inf_{t\in[0,\tau]} H\big(\mu_0(t)+\boldsymbol{\beta}_0^T\mathbf{Z}(t)\big)>0,\qquad \sup_{t\in[0,\tau]}|\mu_0'''(t)|<\infty.$$

Condition (C.1) ensures the identifiability of $\boldsymbol{\beta}$ in model (1). Condition (C.2) implies that the distribution for the censoring times is not informative, and thus $L_n(\boldsymbol{\beta},\mu)$ is the only part of the full likelihood function that we need to maximize. The second part of (C.2) is equivalent to saying that any subjects surviving to at least τ are considered right-censored at τ. Both (C.1) and (C.2) are standard assumptions in the Cox proportional hazards model. Condition (C.3) implies that the true conditional hazard rate for T given the covariates is bounded away from 0.

We also need assumptions for the choices of m_n (or K_n) and M_n. Specifically, we assume that the number of basis functions in the sieve space increases with sample size n, but at a low rate. Moreover, we assume that the upper bound M_n in the sieve space should tend toward infinity at an appropriate rate depending on the transformation function H. The details are given in the following theorem.

Theorem 1. In addition to conditions (C.1)–(C.3), for each $M_n>0$, define

$$\begin{aligned}\gamma_1(M_n)&=2H(M_n+B),\\ \gamma_2(M_n)&=\sup_{x\in[-M_n-B,M_n+B]} H'(x),\\ \gamma_3(M_n)&=\Big\{\inf_{x\in[-M_n-B,M_n+B]} H'(x)\Big\}^{-1},\end{aligned}$$

where B is the upper bound of $|\boldsymbol{\beta}^T\mathbf{Z}(t)|$. Assume that the following condition holds:

(C.4) m_n satisfies that $m_n\to\infty$ and $m_n^7/n\to0$. Moreover, M_n satisfies that

$$M_n^{1/3}\xi(M_n)^{2/3}\left(\sqrt{\frac{m_n^{7/6}}{n^{1/6}}}+\frac{1}{m_n}\right)\to0,$$

where $\xi(M_n)=M_n^2\gamma_1(M_n)^2\gamma_2(M_n)^4\gamma_3(M_n)^2$.

Then $\hat{\boldsymbol{\beta}}$ and $\hat{\mu}(\cdot)$ are consistent in the sense that $|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0|+\|\hat{\mu}-\mu_0\|_{L_2[0,\tau]}\to0$ in probability.

The first part of (C.4) stipulates that the number of basis functions in the sieve space, m_n, increases at a lower rate than $n^{1/7}$. We also remark that M_n satisfying (C.4) always exists for a given m_n and n. We specify some particular choices of m_n and M_n for the class of Box–Cox transformations at the end of this section. The convergence rates of $(\hat{\boldsymbol{\beta}},\hat{\mu})$ are obtained explicitly in the following theorem.

Theorem 2. Under conditions (C.1)–(C.4),

$$|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0|^2+\|\hat{\mu}-\mu_0\|^2_{L_2[0,\tau]}\le o_p\left(\frac{1}{\sqrt{n}}\right)+O_p\left(\frac{1}{m_n^6}\right).$$

Finally, the asymptotic distribution for $\hat{\boldsymbol{\beta}}$ can be summarized in the following theorem.

Theorem 3. In addition to conditions (C.1)–(C.4), suppose that with probability 1, $\mathbf{Z}(t)$ is continuously three times differentiable in $[0,\tau]$, and with respect to some dominating measure, the conditional density of C given $\{\mathbf{Z}(t): t\in[0,\tau]\}$ is three times continuously differentiable. Moreover, $H=G^{-1}$ is continuously three times differentiable, and m_n satisfies the following condition

(C.5) $\sqrt{n}/m_n^6\to0$.

Then $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$ converges weakly to a normal distribution with mean 0, and its asymptotic variance attains the semiparametric efficiency bound.

The regularity condition for $\mathbf{Z}(t)$ in Theorem 3 holds when $\mathbf{Z}(t)$ is time-independent. Because, from Theorem 2, the bias of the sieve estimate $\hat{\mu}(t)$ is of order m_n^{-12}, condition (C.5) implies that the square of this bias does not dominate the variation of $\hat{\boldsymbol{\beta}}$, which is of order $n^{-1/2}$.

The choices of K_n and M_n satisfying (C.4) and (C.5) exist. For large n, we can choose $K_n=\theta\log n/\log 2$ (thus $m_n=n^\theta$), where θ is a constant in the interval $(1/12,1/7)$. If G is the Box–Cox transformation, then M_n can be chosen to be particularly of order $\log n$.

4. SIMULATION STUDY

We conducted simulation studies to examine the small-sample performance of our proposed estimators. In the simulation we generated two independent covariates, Z_1 and Z_2, from Uniform(0, 1) and Bernoulli(.5). We generated the failure time from the model

$$\frac{\{\lambda(t|Z_1, Z_2)\}^s - 1}{s} = \frac{t}{2} + \beta_1 Z_1 + \beta_2 Z_2,$$

where $\beta_1 = .7$ and $\beta_2 = .2$. We varied the choices of s using the values of 0, .25, .5, .75, and 1. Thus when $s = 0$, the failure time was generated from a proportional hazards regression model with baseline hazard $\exp(t/2)$, and when $s = 1$, the failure time was generated from an additive hazards model with baseline hazard $t/2 + 1$. The censoring time was taken as the minimum of 1 and C^*, where $C^* \sim$ Uniform(.5, 1.5), and the censoring rates varied from 20% to 25% for $s = 0$ to $s = 1$.

For each s, we simulated 500 datasets, and for each data realization, we used the proposed sieve maximum likelihood estimation approach to estimate the regression coefficients. In sieve estimation, we chose the Db4-father wavelet (Daubechies 1992) for $\phi(t)$ and used resolution level $K_n = 3$ to estimate the nuisance parameter $\mu(t)$. We obtained the sieve maximum likelihood estimates by the algorithm for searching the optimum in the Optimization toolbox in MATLAB. This algorithm is a subspace trust region method and is based on the interior-reflective Newton method (Coleman and Li 1994, 1996), after both gradients and Hessian derivatives of the objective function are provided. Because the objective function may not be concave in the parameters, choosing initial values can be very important. In our experience, when the initial values were chosen not too far away from the true values, the estimates at convergence were very similar. In the simulation study, the optimum search usually converged within a few iterations when either the step size of the search or the gradient of the function was very small. We used the sieve profile likelihood function to estimate the asymptotic variance of $\hat{\boldsymbol{\beta}}$, where we chose $h_n = n^{-1/2}$. In the simulation study, we also used $K_n = 4, 5$ and $h_n = .1n^{-1/2}, 5n^{-1/2}$, and found the results to be fairly robust with respect to these choices.

Table 1 summarizes the simulation results for different choices of s values for $n = 200$ and $n = 400$. The columns after the true value correspond to the average values of the estimates, the standard errors of the estimates, the average estimates of the asymptotic standard errors, and the coverage proportions of the 95% confidence intervals, based on the normal distribution. The results in Table 1 indicate that the sieve maximum likelihood estimates for the regression coefficients have a small bias, the estimated standard errors based on the sieve profile likelihood function are close to the empirical standard errors, and the coverage proportions of 95% confidence intervals are accurate. Increasing the sample size from 200 to 400 decreases both the bias and the standard errors of the estimates.

5. APPLICATION

We applied our proposed approach to a lung cancer dataset from a recent phase III clinical trial (Socinski et al. 2002) of nonsmall-cell lung cancer (NSCLC), the leading cause of cancer-related mortality. In the year 2001, among approximately 170,000 patients newly diagnosed, more than 90% died from NSCLC, and approximately 35% of all new cases were disease stage IIIB/IV (malignant pleural effusion). A randomized, two-armed, multicenter trial was initiated in 1998 with the aim of determining the optimal duration of chemotherapy by comparing four cycles of therapy versus continuous therapy in advanced NSCLC. Patients were randomized to two treatment arms: four cycles of carboplatin at an area under the curve of 6 and paclitaxel 200 mg/m^2 every 21 days (arm A), or continuous treatment with carboplatin/paclitaxel until progression (arm B). At progression, all patients on both arms received second-line weekly paclitaxel at 80 mg/m^2. One of the primary endpoints was survival, which could be right-censored due to loss to follow-up. The original dataset comprised 230 NSCLC patients; 4 cases were missing follow-up times, and hence our analysis is based on $n = 226$ cases, of which 113 were in

Table 1. Simulation Results From 500 Repetitions

s	n	Coefficient	True value	Estimate	SE	Estimated SE	95% coverage proportion
0	200	β_1	.7	.703	.236	.234	.948
		β_2	.2	.207	.156	.151	.942
	400	β_1	.7	.701	.160	.165	.966
		β_2	.2	.192	.108	.107	.936
.25	200	β_1	.7	.691	.260	.278	.968
		β_2	.2	.203	.182	.179	.960
	400	β_1	.7	.708	.190	.194	.956
		β_2	.2	.195	.126	.125	.956
.5	200	β_1	.7	.708	.327	.317	.930
		β_2	.2	.210	.207	.203	.948
	400	β_1	.7	.691	.224	.222	.936
		β_2	.2	.194	.136	.142	.956
.75	200	β_1	.7	.678	.356	.349	.936
		β_2	.2	.191	.212	.222	.964
	400	β_1	.7	.693	.251	.249	.950
		β_2	.2	.208	.166	.158	.950
1	200	β_1	.7	.735	.384	.379	.944
		β_2	.2	.172	.254	.241	.928
	400	β_1	.7	.695	.286	.273	.934
		β_2	.2	.203	.170	.173	.960

NOTE: Estimated SE is the average of the profile likelihood estimated standard errors.

arm A and 113 were in arm B. The censoring rate was approximately 32%.

We illustrate the proposed additive transformation hazards models with these NSCLC data and demonstrate the flexibility and generality of this class of models. The covariates included in the model were treatment (0, arm A; 1, arm B), sex (0, female; 1, male), and age at entry. In this population, 63% of the patients were male, and the age at entry ranged from 32 to 82 years (mean, 62 years). In the analysis, we rescaled the time axis to the interval [0, 1].

We fit a class of Box–Cox transformed hazard models to the NSCLC data. The parameter s in the transformation was chosen as 0, .25, .5, .75, or 1; the multiresolution level K_n was chosen from 2, 3, 4, and 5. The Akaike information criteria (AIC), defined as twice the negative log-likelihood function plus twice the number of the parameters, was used as a criterion to select the best-fitting model. Using the AIC by varying s and K_n ensured the best model choice in terms of both model structure and parsimony, although it is difficult to determine whether the best fit is due to the transformation or to the choice of basis functions. We also penalized those choices of s and K_n for which the estimated parameters induced negative predicted values for the hazard function. If the estimated hazard rate was negative, then we set the objective function that needed to be maximized to be a very small negative number. Thus the best model using AIC always ensures that the predicted hazard function is positive. From the analysis, we found that increasing the number of basis functions dramatically increased the value of AIC, and the model with $s = .5$ and $K_n = 2$ yielded the minimal AIC value. The estimates and standard errors for the coefficients of the three covariates were $\hat{\beta}_{\text{treat}} = -.1176$ (.2841), $\hat{\beta}_{\text{sex}} = .7086$ (.2966), and $\hat{\beta}_{\text{age}} = .6568$ (.5332). Thus only the covariate sex was significantly predictive of hazard risk. The male patients had a higher risk than the females. Neither treatment nor age was significant. We also plotted the predicted survival curves versus the Kaplan–Meier survival curves in Figure 1. Each plot in Figure 1 represents the predicted survival curves and the Kaplan–Meier curves stratified by treatment and sex, where the age value is substituted with its median value 63. The plots indicate that the best model ($s = .5,\ K = 2$) indeed provides a good fit to the data.

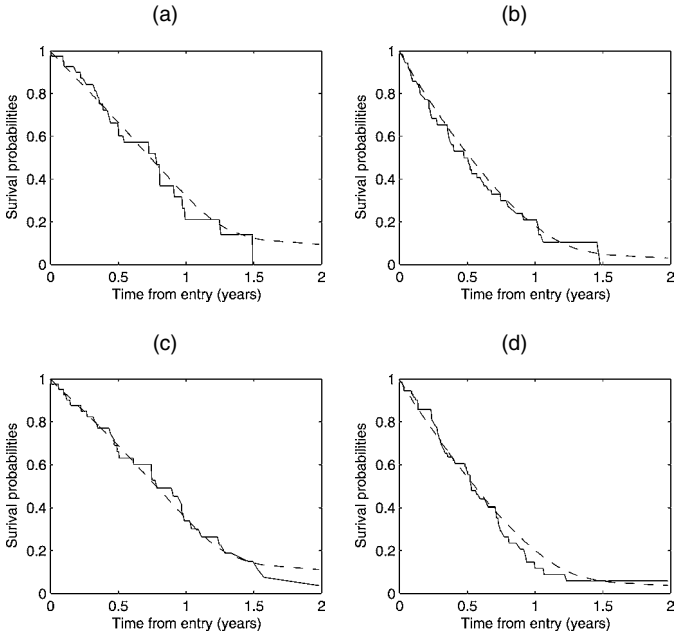

Figure 1. Predicated Survival Curves (- - - -) Based on the Best Model versus the Kaplan–Meier Curves (——). (a) Arm A, female; (b) arm A, male; (c) arm B, female; (d) arm B, male.

6. DISCUSSION

We have proposed a class of transformation models for modeling the hazard function. This class of models contains both multiplicative and additive hazards models as special cases. We have propose a unified estimation procedure in which the sieve maximum likelihood estimates are obtained by maximizing the observed likelihood function over a sieve space of wavelets. The resulting estimators for the regression coefficients have been shown to be asymptotically normal. Simulation studies indicated that the proposed estimates performed well for sample sizes of 200 and 400. Applying the Box–Cox transformed hazards model to the lung cancer data demonstrated that the best model might not be either the multiplicative or the additive hazards model.

In the optimization for computing the maximum likelihood estimates, choosing the initial values is an important issue. Although our numerical studies indicate that convergence is often satisfactory if initial values are not far from true values, one must guess an initial value in practice. One possible way to do this is to use the estimates from the proportional hazards model, which corresponds to transformation $H(x) = \exp\{x\}$ and has the concave log-likelihood function, as the initial values. Another, more general solution is to choose a few widespread points in the parameter domain as initial values, and from among all of the estimates starting from these initial values, consider the one with the maximal likelihood function to be the maximum likelihood estimate.

Although in our theoretical derivations a high-order smooth father wavelet is needed to ensure that the asymptotic results hold for the regression parameters, our simulation study and data application showed that using a low-order smooth father wavelet (e.g., the Db4 wavelet) works quite well. In practice, if one is interested only in inference on the regression coefficients, then a low-order smooth wavelet basis such as the Db4 wavelet may be used, whereas a high-order smooth wavelet should be used to obtain a smooth predicted function of the hazard rate.

In many other nonparametric estimation contexts, it is important to choose a suitable smoothing parameter. In the sieve maximum likelihood estimation that we have proposed, such a parameter is the multiresolution level K_n (thus m_n). In data applications, we used the AIC criterion to choose K_n, but other criteria can be used to choose K_n; one possibility is to replace the negative log-likelihood function in the AIC criterion by a distance measure, which is defined as the L_2 distance between the predicted survival function based on the model and the Kaplan–Meier survival function. The AIC criterion or the just-proposed criterion can also be used to choose the model that best fits the data from a class of transformed hazards models, as we did in the data application. In all of these model selection procedures, the variation in choosing the best model is not accounted for in our inference for the regression parameters. One possibility for

accounting for such variation is to treat the transformation G, indexed by the parameter s, as another model parameter; then we maximize the observed likelihood function over all model parameters, including the transformation G. However, the asymptotic properties of the estimators for the regression coefficients are not yet available.

APPENDIX: PROOFS

A.1 Proof of Theorem 1

The consistency proof contains the following steps, where $r = 3$.

Step 1. We first choose $\tilde{\mu}(t)$ as the approximate function in the K_nth multiresolution to $\mu_0(t)$ such that $(\boldsymbol{\beta}_0, \tilde{\mu}) \in \mathcal{S}_n$. According to the results of the wavelet analysis (Härdle, Kerkyacharian, Picard, and Tsybakov 2000, sec. 9.4),

$$\|\tilde{\mu} - \mu_0\|_{W^{1,\infty}} \le O(1)\frac{\|\mu_0\|_{W^{r,\infty}}}{m_n^{r-1}},$$

$$\|\tilde{\mu} - \mu_0\|_{L^\infty} \le O(1)\frac{\|\mu_0\|_{W^{r,\infty}}}{m_n^{r}},$$

where $\|\mu\|_{W^{r,\infty}} = \sup_{l\le r}\sup_{t\in[0,\tau]}|\mu^{(j)}(t)|$ for $k = 0, \ldots, r$. Moreover, the wavelet coefficients for $\tilde{\mu}(t) = \sum_{j=1}^{m_n}\tilde{\alpha}_j B_j(t)$ satisfy that $\sum_{j=1}^{m_n}|\tilde{\alpha}_j| < \infty$. Thus $(\boldsymbol{\beta}_0, \tilde{\mu}) \in \mathcal{S}_n$.

Step 2. We obtain a bound for the distance

$$d\big((\hat{\boldsymbol{\beta}}, \hat{\mu}), (\boldsymbol{\beta}_0, \mu_0)\big) \equiv |\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0| + \|\hat{\mu} - \mu_0\|_{L_2},$$

where $\|\mu\|_{L_2} = \{\int_0^\tau |\mu(t)|^2\,dt\}^{1/2}$. From the construction of $\tilde{\mu}$, we immediately obtain that

$$L_n(\hat{\boldsymbol{\beta}}, \hat{\mu}) \ge L_n(\boldsymbol{\beta}_0, \tilde{\mu}). \tag{A.1}$$

If we let $\mathbf{P}_n$ denote the empirical measure based on the n iid observations and $\mathbf{P}$ denote the corresponding expectation, then, after taking the log on both sides of (A.1) and dividing by n, we have that

$$\mathbf{P}_n\left\{\Delta\log H\big(\hat{\mu}(Y) + \hat{\boldsymbol{\beta}}^T\mathbf{Z}(Y)\big) - \int_0^Y H\big(\hat{\mu}(t) + \hat{\boldsymbol{\beta}}^T\mathbf{Z}(t)\big)\,dt\right\}$$
$$\ge \mathbf{P}_n\left\{\Delta\log H\big(\tilde{\mu}(Y) + \boldsymbol{\beta}_0^T\mathbf{Z}(Y)\big) - \int_0^Y H\big(\tilde{\mu}(t) + \boldsymbol{\beta}_0^T\mathbf{Z}(t)\big)\,dt\right\}.$$

Note that the function $\Delta\log H(\cdot) - \int_0^Y H(\cdot)\,dt$ is concave in $H(\cdot)$. Thus, for any $\delta_n > 0$, if we define

$$H_{\delta_n}(t) = \delta_n H\big(\hat{\mu}(t) + \hat{\boldsymbol{\beta}}^T\mathbf{Z}(t)\big) + (1-\delta_n)H\big(\tilde{\mu}(t) + \boldsymbol{\beta}_0^T\mathbf{Z}(t)\big),$$

then we have

$$\mathbf{P}_n\left\{\Delta\log H_{\delta_n}(Y) - \int_0^Y H_{\delta_n}(t)\,dt\right\}$$
$$\ge \mathbf{P}_n\left\{\Delta\log H\big(\tilde{\mu}(Y) + \boldsymbol{\beta}_0^T\mathbf{Z}(Y)\big) - \int_0^Y H\big(\tilde{\mu}(t) + \boldsymbol{\beta}_0^T\mathbf{Z}(t)\big)\,dt\right\}.$$

Thus

$$n^{-1/2}\mathbf{G}_n\left\{\Delta\log H_{\delta_n}(Y) - \int_0^Y H_{\delta_n}(t)\,dt\right.$$
$$\left. - \Delta\log H\big(\tilde{\mu}(Y) + \boldsymbol{\beta}_0^T\mathbf{Z}(Y)\big) + \int_0^Y H\big(\tilde{\mu}(t) + \boldsymbol{\beta}_0^T\mathbf{Z}(t)\big)\,dt\right\}$$
$$\ge -\mathbf{P}\left\{\Delta\log H_{\delta_n}(Y) - \int_0^Y H_{\delta_n}(t)\,dt - \Delta\log H\big(\tilde{\mu}(Y) + \boldsymbol{\beta}_0^T\mathbf{Z}(Y)\big)\right.$$
$$\left. + \int_0^Y H\big(\tilde{\mu}(t) + \boldsymbol{\beta}_0^T\mathbf{Z}(t)\big)\,dt\right\}, \tag{A.2}$$

where $\mathbf{G}_n$ denotes the empirical process $\sqrt{n}(\mathbf{P}_n - \mathbf{P})$.

We now want to bound the left side of (A.2) using the results of the empirical process theory. Toward this goal, we choose δ_n such that for some small constant δ_0, $\delta_n = \delta_0/\{M_n\gamma_1(M_n)\gamma_2(M_n)\}$, where $\gamma_1(M_n) = 2H(M_n + B)$ and $\gamma_2(M_n) = \sup_{x\in[-M_n-B,M_n+B]} H'(x)$. Hence,

$$\big\|H_{\delta_n}(t) - H\big(\tilde{\mu}(t) + \boldsymbol{\beta}_0^T\mathbf{Z}(t)\big)\big\|_{L_\infty}$$
$$\le \delta_n\big\|H\big(\hat{\mu}(t) + \hat{\boldsymbol{\beta}}^T\mathbf{Z}(t)\big) - H\big(\tilde{\mu}(t) + \boldsymbol{\beta}_0^T\mathbf{Z}(t)\big)\big\|_{L_\infty}$$
$$\le \delta_0.$$

Moreover, we define a class of functions

$$\mathcal{H}_n = \big\{\delta_n H\big(\mu(t) + \boldsymbol{\beta}^T\mathbf{Z}(t)\big)$$
$$+ (1-\delta_n)H\big(\tilde{\mu}(t) + \boldsymbol{\beta}_0^T\mathbf{Z}(t)\big) : (\boldsymbol{\beta}, \mu) \in \mathcal{S}_n\big\}.$$

By the property of the father wavelet, for any $(\boldsymbol{\beta}, \mu) \in \mathcal{S}_n$,

$$|\mu'(t)| \le \sum_{j=1}^{m_n}|\alpha_j||B_j'(t)| \le c_0 m_n M_n$$

for some constant c_0, so the ϵ-bracket covering number for the class of such μ with respect to $L_2(P)$-norm is of the order $\exp\{O(M_n m_n/\epsilon)\}$ (van der Vaart and Wellner 1996, corollary 2.7.2). By the monotonicity of $H(\cdot)$, we thus can construct the $\exp\{O(M_n m_n/\epsilon)\}$ brackets to cover $\mathcal{H}_n$ such that within each bracket, any two functions, indexed by $(\boldsymbol{\beta}_1, \mu_1)$ and $(\boldsymbol{\beta}_2, \mu_2)$, satisfy $|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2| + \|\mu_1 - \mu_2\|_{L_2(P)} \le \epsilon$. But because

$$\delta_n H'(x)|_{x=\mu(t)+\boldsymbol{\beta}^T\mathbf{Z}(t)} \le \delta_n O(\gamma_2(M_n)) \le \frac{O(1)}{M_n},$$

for these two functions,

$$\big\|\delta_n H\big(\mu_1(t) + \boldsymbol{\beta}_1^T\mathbf{Z}(t)\big) - \delta_n H\big(\mu_2(t) + \boldsymbol{\beta}_2^T\mathbf{Z}(t)\big)\big\|_{L_2(P)} \le O(\epsilon/M_n).$$

We thus conclude that

$$N_{[]}(\epsilon, \mathcal{H}_n, L_2(P)) \le O\big(\exp\{O(m_n/\epsilon)\}\big).$$

Consequently, another class of functions, defined as

$$\mathcal{F}_n = \left\{\Delta\log H_{\delta_n}(Y) - \int_0^Y H_{\delta_n}(t)\,dt : H_{\delta_n} \in \mathcal{H}_n\right\},$$

has a bracket covering number of order

$$N_{[\cdot]}(\epsilon, \mathcal{F}_n, L_2(P)) \le O\big(\exp\{O(m_n/\epsilon)\}\big).$$

Note that $\mathcal{F}_n$ has a bounded covering function. According to lemma 19.38 of van der Vaart (1998), we obtain that

$$E_p^*\|\mathbf{G}_n\|_{\mathcal{F}_n} \le \int_0^{O(1)}\sqrt{\log N_{[]}(\epsilon, \mathcal{F}_n, L_2(P))}\,d\epsilon$$
$$\le O(\sqrt{m_n}).$$

This implies that the left side of (A.2) is bounded by $O_p(\sqrt{m_n}/\sqrt{n})$.

In contrast, the right side of (A.2) can be written as

$$-\mathbf{P}\left\{\Delta\log H_{\delta_n}(Y) - \Delta\log H\big(\mu_0(Y) + \boldsymbol{\beta}_0^T\mathbf{Z}(Y)\big)\right.$$
$$\left. - \int_0^Y H_{\delta_n}(t)\,dt + \int_0^Y H\big(\mu_0(t) + \boldsymbol{\beta}_0^T\mathbf{Z}(t)\big)\,dt\right\}$$
$$-\mathbf{P}\left\{\Delta\log H\big(\mu_0(Y) + \boldsymbol{\beta}_0^T\mathbf{Z}(Y)\big) - \Delta\log H\big(\mu_0(Y) + \boldsymbol{\beta}_0^T\mathbf{Z}(Y)\big)\right.$$
$$\left. - \int_0^Y H\big(\tilde{\mu}(t) + \boldsymbol{\beta}_0^T\mathbf{Z}(t)\big)\,dt + \int_0^Y H\big(\tilde{\mu}(t) + \boldsymbol{\beta}_0^T\mathbf{Z}(t)\big)\,dt\right\}. \tag{A.3}$$

We denote the two terms in (A.3) by (I) and (II) and denote $H_0(Y)$ by $H(\mu_0(Y)+\boldsymbol{\beta}_0^T\mathbf{Z}(t))$. Then applying the mean value theorem to the term (I) yields

$$\text{(I)} = -\mathbf{P}\left\{\frac{\Delta}{H_0(Y)}\big(H_{\delta_n}(Y)-H_0(Y)\big)-\int_0^Y\big(H_{\delta_n}(t)-H_0(t)\big)\,dt\right\} + \mathbf{P}\left\{\frac{\Delta}{\tilde{H}(Y)^2}\big(H_{\delta_n}(Y)-H_0(Y)\big)^2\right\},$$

where $\tilde{H}$ is a function between H_0 and H_{δ_n}. Because $(\boldsymbol{\beta}_0,\mu_0)$ maximizes $\mathbf{P}\{\Delta\log H(\mu(Y)+\boldsymbol{\beta}^T\mathbf{Z}(Y))-\int_0^Y H(\mu(t)+\boldsymbol{\beta}^T\mathbf{Z}(t))\,dt\}$, the derivative of the previous function along the submodel $\boldsymbol{\beta}=\boldsymbol{\beta}_0$, $\mu(t)=\mu_0(t)+\epsilon q(t)$, $\epsilon\in(0,\epsilon_0)$, where ϵ_0 is a small positive constant and $q(t)$ is any measurable function in $L_2(P)$, should be 0. This gives that

$$\mathbf{P}\left\{\frac{\Delta}{H_0(Y)}H'\big(\mu_0(Y)+\boldsymbol{\beta}_0^T\mathbf{Z}(Y)\big)q(Y) - \int_0^Y H'\big(\mu_0(t)+\boldsymbol{\beta}_0^T\mathbf{Z}(t)\big)q(t)\,dt\right\}=0.$$

Thus the first part of the right side in (I) is 0. Because $\tilde{H}(Y)$ is smaller than some constant and $H'(x)\ge 1/\gamma_3(M_n)$ for $x\in[-M_n-B,M_n+B]$, we have that

$$\begin{aligned}\text{(I)} &\ge O(1)\mathbf{P}\left\{\frac{\Delta}{\tilde{H}(Y)^2}\big(H_{\delta_n}(Y)-H_0(Y)\big)^2\right\}\\ &\ge O(1)\frac{\delta_n^2}{\gamma_2(M_n)^2}E\big[\{(\hat{\mu}(Y)-\mu_0(Y))+(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\mathbf{Z}(Y)\}^2\big]\\ &\quad - O\big(\gamma_3(M_n)^2\|\tilde{\mu}(t)-\mu_0\|_{L_2}^2\big).\end{aligned}$$

Similarly, we apply the expansion to the second term (II) of (A.3) around the true parameter $(\boldsymbol{\beta}_0,\mu_0)$. The first order in the expansion vanishes, and the second order is bounded by $O(1)\int_0^\tau(\tilde{\mu}(t)-\mu_0(t))^2\,dt\le O(1/m_n^{2r})$ from the construction of $\tilde{\mu}(t)$. Thus the term (II) is at least $-c_0/m_n^{2r}$ for some positive constant c_0.

Hence we obtain that

$$\begin{aligned}&E\big[\{(\hat{\mu}(Y)-\mu_0(Y))+(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\mathbf{Z}(Y)\}^2\big]\\ &\le O(1)\left\{\frac{M_n^2\gamma_1(M_n)^2\gamma_2(M_n)^4\sqrt{m_n}}{\sqrt{n}}+\frac{\gamma_1(M_n)^2\gamma_2(M_n)^4\gamma_3(M_n)^2}{m_n^{2r}}\right\}\\ &\le O(1)\xi(M_n)^2\left\{\frac{\sqrt{m_n}}{\sqrt{n}}+\frac{1}{m_n^{2r}}\right\}.\end{aligned}$$

Because $\{\mathbf{Z}(t):t\in[0,\tau]\}$ is external and linearly independent with the constant, we obtain that

$$E\big\{(\hat{\mu}(Y)-\mu_0(Y))^2\big\}+(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^TE\{\mathbf{Z}(Y)\mathbf{Z}(Y)^T\}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) \le O(1)\xi(M_n)^2\left(\frac{\sqrt{m_n}}{\sqrt{n}}+\frac{1}{m_n^{2r}}\right).$$

Furthermore, by assumption (C.4), $E\{\mathbf{Z}(Y)^T\mathbf{Z}(Y)\}>0$. It then follows that

$$\int_0^\tau(\hat{\mu}(t)-\mu_0(t))^2\,dt+|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0|^2\le O(1)\xi(M_n)^2\left(\frac{\sqrt{m_n}}{\sqrt{n}}+\frac{1}{m_n^{2r}}\right).$$

Thus, by the choices of M_n and K_n in (C.4), Theorem 1 holds.

A.2 Proof of Theorem 2

To prove Theorem 2, we need a consistency result of $\hat{\mu}$ under a stronger norm than the L_2 norm. First, from the construction of $\mathcal{S}_n$, we have that

$$\|\hat{\mu}-\mu_0\|_{W^{r,2}}\le O\left(\sum_{j=1}^{m_n}|\hat{\alpha}_j|\|B_j(t)\|_{W^{r,2}}\right)\le O(M_n)m_n^r.$$

Then, according to the Sobolev interpolation theorem (Adams 1975), it holds that

$$\|\hat{\mu}'(t)-\mu_0'(t)\|_{L_2}\le c_1\|\hat{\mu}-\mu_0\|_{W^{r,2}}^{1/r}\|\hat{\mu}-\mu_0\|_{L_2}^{1-1/r}$$

for some constant c_1. Then

$$\begin{aligned}\|\hat{\mu}'(t)-\mu_0'(t)\|_{L_2}&\le O(1)m_nM_n^{1/r}\left[\xi(M_n)\left\{\frac{m_n^{1/4}}{n^{1/4}}+\frac{1}{m_n^r}\right\}\right]^{1-1/r}\\ &\le O(1)M_n^{1/r}\xi(M_n)^{1-1/r}\left[\frac{m_n^{5/4-1/4r}}{n^{1/4-1/4r}}+\frac{1}{m_n^{r-2}}\right].\end{aligned}$$

Based on the choice of M_n and K_n in (C.4), this term converges to 0. We thus conclude that, in probability,

$$|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0|\to 0,\qquad \|\hat{\mu}-\mu_0\|_{W^{1,2}}\to 0.$$

In addition, from the Sobolev embedding theorem (Adams 1975), we have that, in probability,

$$\|\hat{\mu}-\mu_0\|_{L_\infty}\to 0.$$

We further improve the convergence rate of $\hat{\boldsymbol{\beta}}$ and $\hat{\mu}$. We simply repeat Step 2 in proving Theorem 1 and obtain a similar inequality as (A.2), but set δ_n to 1. Then the left side of (A.2) belongs to the process $n^{-1/2}\mathbf{G}_n(\mathcal{F}_n^*)$, where

$$\begin{aligned}\mathcal{F}_n^*=\Big\{&\Delta\log H\big(\mu(Y)+\boldsymbol{\beta}^T\mathbf{Z}(Y)\big)-\Delta\log H\big(\tilde{\mu}(Y)+\boldsymbol{\beta}_0^T\mathbf{Z}(Y)\big)\\ &-\int_0^Y H\big(\mu(t)+\boldsymbol{\beta}^T\mathbf{Z}(t)\big)\,dt+\int_0^Y H\big(\tilde{\mu}(t)+\boldsymbol{\beta}_0^T\mathbf{Z}(t)\big)\,dt:\\ &|\boldsymbol{\beta}-\boldsymbol{\beta}_0|<\epsilon,\|\mu-\mu_0\|_{W^{1,2}}<\epsilon\Big\}\end{aligned}$$

for any small number ϵ. Hence $\mathcal{F}_n$ is P-Donsker, and thus the left side is bounded by $o_p(n^{-1/2})$. We again apply Taylor's series expansion to the right side of (A.2), but in this case the bounds $\gamma_1(M_n)$, $\gamma_2(M_n)$, and $\gamma_3(M_n)$ can all be replaced by constants independent of n, due to the fact that $\|\hat{\mu}-\mu_0\|_{L_\infty}\to 0$. Thus we conclude that

$$|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0|^2+\|\hat{\mu}-\mu_0\|_{L_2[0,\tau]}^2\le o_p\left(\frac{1}{\sqrt{n}}\right)+O_p\left(\frac{1}{m_n^{2r}}\right).$$

A.3 Proof of Theorem 3

The proof of asymptotic normality is outlined as follows. We first obtain the least-favorable direction for $\boldsymbol{\beta}_0$, then expand the score equation for $\hat{\boldsymbol{\beta}}$ and $\hat{\mu}$ along an approximate least-favorable model. Here the least-favorable direction for $\boldsymbol{\beta}_0$ is defined as a tangent function at μ_0, denoted by $q(t)$, such that $l_\mu^* l_\mu[q(t)]=l_\mu^* l_{\boldsymbol{\beta}}$, where $l_{\boldsymbol{\beta}}$ is the score function for $\boldsymbol{\beta}_0$, $l_\mu[q(t)]$ is the score function for μ_0 along the submodel $\mu_0(t)+\epsilon q(t)$, and l_μ^* is the dual operator of l_μ. Thus we in turn prove the following steps:

Step 1. We first show that the least-favorable direction $q(t)$ exists. Recall that $H_0(Y)=H(\mu_0(Y)+\boldsymbol{\beta}_0^T\mathbf{Z}(Y))$ and $\Psi(Y)=H'(x)|_{x=\mu_0(Y)+\boldsymbol{\beta}_0^T\mathbf{Z}(Y)}$. By simple algebraic manipulations, we obtain that

$$l_{\boldsymbol{\beta}}=\frac{\Delta\Psi(Y)}{H_0(Y)}\mathbf{Z}(Y)-\int_0^Y\Psi(t)\mathbf{Z}(t)\,dt$$

and

$$l_\mu[q(t)]=\frac{\Delta\Psi(Y)}{H_0(Y)}q(Y)-\int_0^Y\Psi(t)q(t)\,dt.$$

Moreover, the closed linear space spanned by the score functions for μ in $L_2(\nu)$, where ν is the dominating measure, is given by

$$\left\{\frac{\Delta\Psi(Y)}{H_0(Y)}q(Y)-\int_0^Y\Psi(t)q(t)\,dt:q(t)\in L_2[0,\tau]\right\}.$$

Thus l_μ is a linear operator from $L_2[0,\tau]$ to $L_2(\nu)$. Its dual operator l_μ^* satisfies that for any $q \in L_2[0,\tau]$ and a measurable function $g(\Delta, Y, \mathbf{Z})$ (where $\mathbf{Z}$ denotes the covariate process $\{\mathbf{Z}(t): t \in [0,\tau]\}$),

$$E\big[l_\mu[q]g(\Delta,Y,\mathbf{Z})\big] = \int_0^\tau l_\mu^*[g(\Delta,Y,\mathbf{Z})]q(t)\,dt.$$

We expand both sides and, after comparison, obtain that

$$\begin{aligned} l_\mu^*[g(\Delta,Y,\mathbf{Z})] &= E_{\mathbf{Z}}\left[\frac{\Psi(t)}{H_0(t)}g(1,t,\mathbf{Z})S_C(t|\mathbf{Z})f_T(t|\mathbf{Z})\right] \\ &\quad - \big\{E_{T,\mathbf{Z}}\big[I(T\ge t)\Psi(t)g(1,T,\mathbf{Z})S_C(T|\mathbf{Z})\big] \\ &\quad + E_{C,\mathbf{Z}}\big[I(C\ge t)\Psi(t)g(0,C,\mathbf{Z})S_T(C|\mathbf{Z})\big]\big\}, \end{aligned}$$

where $S_T(\cdot|\mathbf{Z})$ and $S_C(\cdot|\mathbf{Z})$ are the conditional survival functions for T and C given $\mathbf{Z}$. Therefore,

$$l_\mu^* l_\mu[q] = q(t)E\left[\frac{\Psi(t)^2}{H_0(t)^2}S_C(t|\mathbf{Z})f_T(t|\mathbf{Z})\right] + \int q(s)k(s,t)\,ds,$$

where

$$\begin{aligned} k(s,t) &= -E_{\mathbf{Z}}\left[f_T(t|\mathbf{Z})S_C(t|\mathbf{Z})I(t\ge s)\Psi(s)\frac{\Psi(t)}{H_0(t)}\right] \\ &\quad - E_{\mathbf{Z}}\left[f_T(s|\mathbf{Z})S_C(s|\mathbf{Z})I(s\ge t)\Psi(t)\frac{\Psi(s)}{H_0(s)}\right] \\ &\quad + E_{Y,\mathbf{Z}}[\Psi(t)\Psi(s)I(Y\ge t)I(Y\ge s)]. \end{aligned}$$

Note that $k(s,t)$ is a continuous function of (t,s) based on (C.5). Therefore, $l_\mu^* l_\mu[q] = l_\mu^* l_{\boldsymbol{\beta}}$ is a Fredholm-type equation, and the existence of the solution is equivalent to showing that $l_\mu^* l_\mu[\tilde{q}] = 0$ has a trivial solution. The latter is clear from the following argument: If $l_\mu^* l_\mu[\tilde{q}] = 0$, then $E\{l_\mu[\tilde{q}]l_\mu[\tilde{q}]\} = 0$. Thus $l_\mu[\tilde{q}] = 0$, so it is clear that $\tilde{q}(t) \equiv 0$. We conclude that there exists a solution $q(t)$ such that $l_\mu^* l_\mu[q(t)] = l_\mu^* l_{\boldsymbol{\beta}}$. Clearly, from the equation for $q(t)$ and condition (C.5), as well as the smoothness condition in Theorem 2, $q(t)$ is continuously three times differentiable in $[0,\tau]$.

Step 2. We choose an approximate submodel $(\hat{\boldsymbol{\beta}} + \epsilon\mathbf{b}, \hat{\mu} + \epsilon\hat{q})$, where $\hat{q}$ is the approximate wavelet function for q in the sieve space $\mathcal{S}_n$, and thus $\hat{q} \in W^{r,2}$ and $\|\hat{q} - q\|_{L_2} \le O(1/m_n^r)$. Because $(\hat{\boldsymbol{\beta}}, \hat{\mu})$ maximizes the observed likelihood function along this submodel, we immediately obtain that

$$\mathbf{P}_n\big\{l_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}},\hat{\mu}) + l_\mu(\hat{\boldsymbol{\beta}},\hat{\mu})[\hat{q}]\big\} = 0,$$

where $l_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}},\hat{\mu})$ is the score function for $\boldsymbol{\beta}$ evaluated at $(\hat{\boldsymbol{\beta}},\hat{\mu})$ and $l_\mu(\hat{\boldsymbol{\beta}},\hat{\mu})[\hat{q}]$ is the score function for μ evaluated at $(\hat{\boldsymbol{\beta}},\hat{\mu})$. Thus

$$\mathbf{G}_n\big\{l_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}},\hat{\mu}) + l_\mu(\hat{\boldsymbol{\beta}},\hat{\mu})[\hat{q}]\big\} = -\sqrt{n}\mathbf{P}\big\{l_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}},\hat{\mu}) + l_\mu(\hat{\boldsymbol{\beta}},\hat{\mu})[\hat{q}]\big\}.$$

Because the function $l_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}},\hat{\mu}) + l_\mu(\hat{\boldsymbol{\beta}},\hat{\mu})[\hat{q}]$ belongs to a P-Donsker class, the foregoing equation becomes

$$\begin{aligned} &\mathbf{G}_n\big\{l_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0,\mu_0) + l_\mu(\boldsymbol{\beta}_0,\mu_0)[q]\big\} + o_p(1) \\ &\qquad = -\sqrt{n}\mathbf{P}\big\{l_{\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}},\hat{\mu}) + l_\mu(\hat{\boldsymbol{\beta}},\hat{\mu})[\hat{q}]\big\}. \end{aligned}$$

We perform Taylor's series expansion of the right side at $(\boldsymbol{\beta}_0,\mu_0)$, and obtain

$$\begin{aligned} &\mathbf{G}_n\big\{l_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0,\mu_0) + l_\mu(\boldsymbol{\beta}_0,\mu_0)[q]\big\} + o_p(1) \\ &\quad = -\sqrt{n}\mathbf{P}\big\{l_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0,\mu_0) + l_{\boldsymbol{\beta}\mu}(\boldsymbol{\beta}_0,\mu_0)[q]\big\}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ &\qquad - \sqrt{n}\mathbf{P}\big\{l_{\boldsymbol{\beta}\mu}(\boldsymbol{\beta}_0,\mu_0)[\hat{\mu}-\mu_0] + l_{\mu\mu}(\boldsymbol{\beta}_0,\mu_0)[q,\hat{\mu}-\mu_0]\big\} \\ &\qquad + \sqrt{n}O\big(|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|^2 + \|\hat{\mu}-\mu_0\|_{L_2}^2 + \|\hat{q}-q\|_{L_2}^2\big). \end{aligned} \tag{A.4}$$

Here $l_{\boldsymbol{\beta}\mu}(\boldsymbol{\beta}_0,\mu_0)[\hat{\mu}-\mu_0]$ is the derivative of $l_{\boldsymbol{\beta}}$ along the path $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, $\mu = \mu_0 + \epsilon(\hat{\mu}-\mu_0)$, and $l_{\mu\mu}(\boldsymbol{\beta}_0,\mu_0)[q,\hat{\mu}-\mu_0]$ is the derivative of $l_\mu[q]$ along the path $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, $\mu = \mu_0 + \epsilon(\hat{\mu}-\mu_0)$. The second term on the right side of (A.4) is 0, because $q(t)$ satisfies $l_\mu^* l_\mu[q(t)] = l_\mu^* l_{\boldsymbol{\beta}}$; the third term on the right side of (A.4) is $o_p(1)$ based on the results of the convergence rate for $(\hat{\boldsymbol{\beta}},\hat{\mu})$ and the condition that $\sqrt{n}/m_n^{2r} \to 0$. Hence,

$$\begin{aligned} &-\sqrt{n}\mathbf{P}\big\{l_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0,\mu_0) + l_{\boldsymbol{\beta}\mu}(\boldsymbol{\beta}_0,\mu_0)[q]\big\}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ &\qquad = \mathbf{G}_n\big\{l_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0,\mu_0) + l_\mu(\boldsymbol{\beta}_0,\mu_0)[q]\big\} + o_p(1). \end{aligned} \tag{A.5}$$

Step 3. We show that the matrix $\mathbf{P}\{l_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0,\mu_0) + l_{\boldsymbol{\beta}\mu}(\boldsymbol{\beta}_0,\mu_0)[q]\}$ is nonsingular. If it is not, then there exists a non-0 vector $\mathbf{b}$ such that

$$\mathbf{b}^T\mathbf{P}\big\{l_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0,\mu_0) + l_{\boldsymbol{\beta}\mu}(\boldsymbol{\beta}_0,\mu_0)[q]\big\}\mathbf{b} = 0;$$

that is, $\mathbf{P}[\{\mathbf{b}^T l_{\boldsymbol{\beta}} + \mathbf{b}^T l_\mu[q]\}^2] = \mathbf{0}$. Then $\mathbf{b}^T l_{\boldsymbol{\beta}} + \mathbf{b}^T l_\mu[q] = \mathbf{0}$. It is easy to see that $\mathbf{b}^T\mathbf{Z}(t) + q(t) = 0$. We thus obtain a contradiction.

Step 4. Finally, from (A.5), we obtain that

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &= -\big[\mathbf{P}\big\{l_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0,\mu_0) + l_{\boldsymbol{\beta}\mu}(\boldsymbol{\beta}_0,\mu_0)[q]\big\}\big]^{-1} \\ &\quad \times \mathbf{G}_n\big\{l_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0,\mu_0) + l_\mu(\boldsymbol{\beta}_0,\mu_0)[q]\big\} \\ &\quad + o_p(1). \end{aligned}$$

Therefore, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ converges to a normal distribution and has influence function given by

$$\begin{aligned} &\big[\mathbf{P}\big\{l_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0,\mu_0) + l_{\boldsymbol{\beta}\mu}(\boldsymbol{\beta}_0,\mu_0)[q]\big\}\big]^{-1} \\ &\qquad \times \big\{l_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0,\mu_0) + l_\mu(\boldsymbol{\beta}_0,\mu_0)[q]\big\}. \end{aligned}$$

Because this influence function is on the linear space spanned by the score functions $l_{\boldsymbol{\beta}}$ and $l_\mu[q]$, the influence function is the same as the efficient influence function for $\boldsymbol{\beta}_0$. Hence the asymptotic variance of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ attains the semiparametric efficiency bound.

[Received April 2004. Revised September 2004.]

REFERENCES

Adams, R. A. (1975), *Sobolev Spaces*, New York: Academic Press.

Andersen, P. K., and Gill, R. D. (1982), "Cox's Regression Model for Counting Processes: A Large-Sample Study," *The Annals of Statistics*, 10, 1100–1120.

Bickle, P. J., Klaassen, C. A. I., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore, MD: Johns Hopkins University Press.

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 26, 211–252.

Coleman, T. F., and Li, Y. (1994), "On the Convergence-Reflective Newton Methods for Large-Scale Nonlinear Minimization Subject to Bounds," *Mathematical Programming*, 67, 189–224.

——— (1996), "An Interior, Trust Region Approach for Nonlinear Minimization Subject to Bounds," *SIAM Journal on Optimization*, 6, 418–445.

Cox, D. R. (1972), "Regression Models and Life-Tables" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 34, 187–220.

——— (1975), "Partial Likelihood," *Biometrika*, 62, 269–276.

Daubechies, I. (1992), *Ten Lectures on Wavelets*, Notes from the 1990 CBMS–NSF Conference on Wavelets and Applications at Lowell, Philadelphia, MA: SIAM.

Fan, J., and Wong, W. H. (2000), Comment on "On Profile Likelihood," by S. A. Murphy and A. W. van der Vaart, *Journal of the American Statistical Association*, 92, 968–976.

Fan, J., and Zhang, J. (2004), "Sieve Empirical Likelihood Ratio Tests for Nonparametric Functions," *The Annals of Statistics*, 32, 1858–1907.

Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (2000), *Wavelets, Approximation, and Statistical Applications*, New York: Springer-Verlag.

Lin, D. Y., and Ying, Z. (1994), "Semiparametric Analysis of the Additive Risk Model," *Biometrika*, 81, 61–71.

——— (1995), "Semiparametric Analysis of General Additive-Multiplicative Hazard Models for Counting Processes," *The Annals of Statistics*, 23, 1712–1734.

Mallat, S. (1998), *A Wavelet Tour of Signal Processing*, New York: Academic Press.

Murphy, S. A., and van der Vaart, A. W. (2000), "On Profile Likelihood," *Journal of the American Statistical Association*, 95, 449–465.

Shen, X. (1997), "On Methods of Sieves and Penalization," *The Annals of Statistics*, 25, 2555–2591.

——— (1998), "Proportional Odds Regression and Sieve Maximum Likelihood Estimation," *Biometrika*, 85, 165–177.

Shen, X., and Shi, J. (2004), "Sieve Likelihood Ratio Inference on General Parameter Space," *Science China*, to appear.

Shen, X., and Wong, W. H. (1994), "Convergence Rate of Sieve Estimates," *The Annals of Statistics*, 22, 580–615.

Socinski, M. A., Schell, M. J., Peterman, A., Bakri, K., Yates, S., Gitten, R., Unger, P., Lee, J., Lee, J. H., Tynan, M., Moore, M., and Kies, M. S. (2002), "Phase III Trial Comparing a Defined Duration of Therapy versus Continuous Therapy Followed by Second-Line Therapy in Advanced-Stage IIIB/IV Non-Small-Cell Lung Cancer," *Journal of Clinical Oncology*, 20, 1335–1343.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, New York: Cambridge University Press.

van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer-Verlag.