

# Polynomial Spline Estimation and Inference of Proportional Hazards Regression Models with Flexible Relative Risk Form

**Jianhua Z. Huang\***

Department of Statistics, Texas A & M University, College Station, TX 77843-3143,  
U.S.A.

and

**Linxu Liu**

Department of Biostatistics, Mailman School of Public Health, Columbia University,  
New York, NY, 10032, U.S.A.

**SUMMARY.** The Cox proportional hazards model usually assumes an exponential form for the dependence of the hazard function on covariate variables. However, in practice this assumption may be violated and other relative risk forms may be more appropriate. In this paper, we consider the proportional hazards model with an unknown relative risk form. Issues in model interpretation are addressed. We propose a method to estimate the relative risk form and the regression parameters simultaneously by first approximating the logarithm of the relative risk form by a spline and then employing the maximum partial likelihood estimation. An iterative alternating optimization procedure is developed for efficient implementation. Statistical inference of the regression coefficients and of the relative risk form based on parametric asymptotic theory is discussed. The proposed methods are illustrated using simulation and

---

\* *Corresponding author's email:* jianhua@stat.tamu.edu

an application to the Veteran’s Administration lung cancer data.

KEY WORDS: Nonparametric regression; Partial likelihood; Proportional hazards model; Single index model; Spline.

## 1. Introduction

In the widely used Cox proportional hazards model (Cox, 1972) for analysis of failure and survival data, it is assumed that the hazard function conditional on the covariate vector  $x$  has the following form

$$\lambda(t|x) = \lambda_0(t) \exp\{\beta_0^T x\} \tag{1}$$

where  $\beta_0 \in R^p$  is a vector of regression coefficients and  $\lambda_0(t)$  is an arbitrary and unspecified baseline hazard function. An important property for this model is that the baseline hazard doesn’t need to be estimated before the regression coefficients are estimated. Partial likelihood method (Peto, 1972; Cox, 1975; Kalbfleisch and Prentice, 2002) can be used to estimate the regression coefficients  $\beta_0$ . After we get an estimate of  $\beta_0$ , the baseline hazard or equivalently, the baseline cumulative hazard function  $\Lambda_0(t) = \int_0^t \lambda_0(s) ds$  can be estimated using Breslow’s method (Breslow, 1972).

An important assumption of the Cox model is that the covariate variables have a linear effect on the log-hazard function. However, this assumption may be violated and misleading conclusion may result. As a remedy, nonparametric function estimation has been proposed to estimate the conditional hazard function, which in turn is specified as

$$\lambda(t|x) = \lambda_0(t) \exp\{\eta(x)\},$$

where  $\eta(x)$  is assumed to be a smooth function of  $x$ . See, for example, Tibshirani and Hastie (1987), Gentleman and Crowley (1991), O’Sullivan (1993), Gu (1996), and Fan, Gijbels and King (1997). However, unstructured nonparametric function estimation is subject to the “curse of dimensionality” and thus is not practically useful when

$x$  has many components. The problem is that accurate point estimation is difficult with the sample sizes usually available in practice because of the sparsity of the data in even moderately large dimensions. To overcome the curse of dimensionality, structured nonparametric models are usually considered. Sleeper and Harrington (1990) used additive models to model the nonlinear covariate effects in the Cox model. They modeled the log-hazard as an additive function of each covariate and then approximated each of the additive components as a polynomial spline. Gray (1992) applied penalized splines to additive models and time-varying coefficient models of the log-hazard function. Using functional ANOVA decompositions, Huang et al. (2000) studied a general class of structured models for proportional hazards regression that includes additive models as a special case; polynomial splines are used as building blocks for fitting the functional ANOVA models. Kooperberg, Stone and Truong (1995) proposed a flexible method for hazard regression using splines that does not require the proportional hazards assumption.

In this paper, we consider an alternative way to model the possible nonlinearity of the covariate effects to the log-hazard in the proportional hazards model (Wang, 2001; Wang, Wang, and Wang, 2001). Specifically, by extending (1), the conditional hazard function is modeled as

$$\lambda(t|x) = \lambda_0(t) \exp\{\psi(\beta_0^T x)\} \quad (2)$$

where  $\psi(\cdot)$ , referred to as the link function, is an unknown smooth function. Since  $\psi(\cdot)$  is not specified, the relative risk function has a flexible form. This model specification is closely related to the single index models (Ichimura, 1993) and generalized linear models with unknown links (Weisberg and Welsh, 1994; Chiou and Müller, 1998, 1999). It can model a possible departure from the standard Cox model that can not be captured by the additive, time-varying coefficient, or more general functional ANOVA models. It is possible to test the validity of the Cox model within this

modeling framework by testing whether the function  $\psi(\cdot)$  is linear or not. Fitting model (2) to the VA lung cancer data (Kalbfleisch and Prentice, 2002) has revealed a nonlinear link function and yielded results that are qualitatively different than those from previous analysis by the standard Cox model. Some covariates previously thought to be nonprognostic become statistically significant in explaining survival when a flexible nonlinear relative risk form is used. See Figure 1 for the fitted link function and Section 5 for details of analysis.

[Figure 1 about here.]

Estimation of an unknown  $\psi(\cdot)$  using a local polynomial smoother has been studied by Wang (2001) and Wang, Wang and Wang (2001); see also Wang (2004) for a description of the estimation procedure and statements of related asymptotic results. In this paper we try to contribute in two aspects: firstly we develop a practically simple yet flexible solution to the problem based on parametric spline approximations of the link function, secondly we provide detailed discussion on interpretation of the model. Such discussion is practically important since, in the presence of unknown link function, the regression coefficients do not have the usual interpretation as in the standard Cox model.

Our proposed estimation method for (2) is to first approximate the unknown function  $\psi(\cdot)$  by a polynomial spline and then employ the maximum partial likelihood estimation. Splines are well-known for their ability to provide good approximations to smooth functions (de Boor 1978, Schumaker 1981) and their application in nonparametric smoothing is broad (Stone et al., 1997). Application of splines in the current context is in particular convenient. After a spline basis is chosen, the spline approximated unknown function is totally characterized by the coefficients (referred to as spline coefficients) in the basis expansion. The spline coefficients and the regression coefficients are then simultaneously estimated by maximizing the partial likelihood.

The asymptotic variance-covariance matrix of these coefficients is derived following the theory of partial likelihood. This matrix is used to develop estimates of standard errors of the regression coefficients and of the unknown function. As we will see in Section 2, when the regression parameters are fixed, the partial likelihood is concave in the spline coefficients and its maximum is unique and can be found by a modified Newton-Raphson method. Taking this into account, we develop an iterative algorithm that alternates between optimization over the spline coefficients and the regression coefficients. The simplicity of our estimation and inference procedure relies on the nice parametric spline basis representation of the unknown link function.

The rest of the paper is organized as follows. Section 2 gives details of the model specification, proposed methods of estimation and inference, and numerical implementation. Issues for interpretation of the fitted model is discussed in Section 3. We present some simulation results in Section 4 and apply our method to the VA lung cancer data in Section 5. Discussion of some issues in asymptotic theory is given in Section 6.

## 2. The method

Let  $X$  be a vector of predictor variables that does not have 1 as one of its components and  $T$  be the survival time. We consider the proportional hazards regression model with unknown relative risk form as specified in (2).

### 2.1 Identifiability constraints

It is easy to see that any constant in  $\psi(\cdot)$  can be absorbed in  $\lambda_0(\cdot)$ , hence  $\psi(\cdot)$  itself is not identifiable and we impose  $\psi(0) = 0$  for identifiability. Similarly, we can only estimate the direction of  $\beta_0$  and the scale of it is not identifiable since any scale constant can be absorbed in  $\psi(\cdot)$ , so we standardize  $\beta_0$  by requiring  $\|\beta_0\| = 1$  to identify it, where  $\|a\| = (a^T a)^{1/2}$  denotes the Euclidean norm. Moreover, the

sign of the regression coefficients is not identified, since if we define  $\psi^*(s) = \psi(-s)$  and  $\beta^* = -\beta$  then we have  $\psi^*(x^T\beta^*) = \psi(x^T\beta)$ . Thus we require the first non-zero component of  $\beta_0$  to be positive for identifiability purpose. Note that the interpretation of the model parameters for the flexible model is not as straightforward as for the standard Cox model. See Section 3 for detailed discussion.

## 2.2 Estimation

Suppose we have an independent right censoring scheme in which i.i.d. censoring times  $C_1, \dots, C_n$  are independent of the survival times  $T_1, \dots, T_n$  given the covariate variables  $X_1, \dots, X_n$ . Suppose we observe for the  $i$ th ( $i = 1, 2, \dots, n$ ) subject an event time  $Z_i = \min(T_i, C_i)$ , a censoring indicator  $\delta_i = I\{T_i \leq C_i\}$ , as well as the  $p$ -variable covariate vector  $X_i$ . Then we can denote the observed data as  $(X_i, Z_i, \delta_i)$ ,  $i = 1, \dots, n$ , which is an i.i.d. sample of  $(X, Z = \min(T, C), \delta = I\{T \leq C\})$ . We also assume that the covariate  $X$  does not change with time.

We propose to approximate the derivative of the unknown function  $\psi(\cdot)$  as a spline function. Such an approximation can be represented by a basis expansion

$$\psi'(\beta^T x) = \sum_{j=1}^k \gamma_j B_j(\beta^T x) \quad (3)$$

where  $B_j, j = 1, \dots, k$ , are the B-spline basis functions (de Boor, 1978). Other choices of basis such as truncated power basis can in principle be used here. But B-splines are preferable for numerical stability. Using the identifiability constraint  $\psi(0) = 0$ ,

$$\psi(\beta^T x) = \int_0^{\beta^T x} \sum_{j=1}^k \gamma_j B_j(t) dt = \sum_{j=1}^k \gamma_j \tilde{B}_j(\beta^T x) \quad (4)$$

where  $\tilde{B}_j(u) = \int_0^u B_j(s) ds$ ,  $j = 1, \dots, k$ , are the integrals of the B-spline basis functions. Let  $B(u) = (B_1(u), \dots, B_k(u))^T$ ,  $\tilde{B}(u) = (\tilde{B}_1(u), \dots, \tilde{B}_k(u))^T$ , and  $\gamma = (\gamma_1, \dots, \gamma_k)^T$ . We can write (3) and (4) in matrix form as  $\psi'(\beta^T x) = \gamma^T B(\beta^T x)$  and  $\psi(\beta^T x) = \gamma^T \tilde{B}(\beta^T x)$ . In our implementation of the method, we used quadratic B-splines in the basis expansion of  $\psi'(\cdot)$  and as a result  $\psi(\cdot)$  is a cubic spline.

Let  $t_1 < t_2 < \dots < t_m$  be the  $m$  distinctive ordered event times. By using the above approximation of the unknown function, the partial likelihood (as a function of  $\gamma$  and  $\beta$ ) is

$$PL = \prod_{i=1}^m \frac{\exp\{\psi(x_i^T \beta)\}}{\sum_{l \in R_i} \exp\{\psi(x_l^T \beta)\}} = \prod_{i=1}^m \frac{\exp\{\gamma^T \tilde{B}(x_i^T \beta)\}}{\sum_{l \in R_i} \exp\{\gamma^T \tilde{B}(x_l^T \beta)\}},$$

where  $R_i = \{l : Z_l \geq t_i\}$  is the risk set at event time  $t_i, i = 1, \dots, m$ . Therefore, the log-partial likelihood is

$$l(\beta, \gamma) = \log(PL) = \sum_{i=1}^m \left( \gamma^T \tilde{B}(x_i^T \beta) - \log \left[ \sum_{l \in R_i} \exp\{\gamma^T \tilde{B}(x_l^T \beta)\} \right] \right) \quad (5)$$

This partial likelihood is maximized over possible values of  $\gamma$  and  $\beta$ . Denote the maximizer as  $\hat{\gamma}$  and  $\hat{\beta}$ . Then the spline estimate of the unknown function is  $\hat{\psi}(u) = \hat{\gamma}^T \tilde{B}(u)$  and the regression parameter estimate is  $\hat{\beta}$ . When there exist ties among event times, standard procedures (Kalbfleisch and Prentice, 2002) in handling ties for the Cox model can be used. We used Breslow's approximation in our implementation.

It is straightforward to see that the score function  $S_\beta$  and Hessian matrix  $H_{\beta, \beta}$  of  $l$  with respect to  $\beta$  are

$$S_\beta = \frac{\partial l}{\partial \beta} = \sum_{i=1}^m \left\{ \gamma^T B(x_i^T \beta) x_i - \sum_{l \in R_i} w_{li} \gamma^T B(x_l^T \beta) x_l \right\}$$

and  $H_{\beta, \beta} = \partial^2 l / (\partial \beta \partial \beta^T) = H_1 - H_2$ , where  $H_1 = \sum_{i=1}^m \gamma^T B'(x_i^T \beta) x_i x_i^T$  and

$$H_2 = \sum_{i=1}^m \left( \sum_{l \in R_i} w_{li} [\gamma^T B'(x_i^T \beta) + \{\gamma^T B(x_l^T \beta)\}^2] x_l x_l^T - \sum_{l \in R_i} w_{li} \gamma^T B(x_l^T \beta) x_l \sum_{l \in R_i} w_{li} \gamma^T B(x_l^T \beta) x_l^T \right)$$

with  $w_{li} = \exp\{\gamma^T \tilde{B}(x_l^T \beta)\} / \sum_{j \in R_i} \exp\{\gamma^T \tilde{B}(x_j^T \beta)\}$  and  $B'(u) = (B'_1(u), \dots, B'_k(u))^T$ .

Similarly, we obtain that the score function  $S_\gamma = \partial l / \partial \gamma$  and Hessian matrix  $H_{\gamma, \gamma} = \partial^2 l / \partial \gamma \partial \gamma^T$  of  $l$  with respect to  $\gamma$  are

$$S_\gamma = \sum_{i=1}^m \left\{ \tilde{B}(x_i^T \beta) - \sum_{l \in R_i} w_{li} \tilde{B}(x_l^T \beta) \right\}$$

and

$$H_{\gamma,\gamma} = - \sum_{i=1}^m \left\{ \sum_{l \in R_i} w_{li} \tilde{B}(x_l^T \beta) \tilde{B}^T(x_l^T \beta) - \sum_{l \in R_i} w_{li} \tilde{B}(x_l^T \beta) \sum_{l \in R_i} w_{li} \tilde{B}^T(x_l^T \beta) \right\}. \quad (6)$$

It is easily seen that  $H_{\gamma,\gamma}$  is negative semi-definite, which implies that the log-partial likelihood  $l$  is a concave function of  $\gamma$  for fixed  $\beta$ .

The Hessian  $H_{\beta,\gamma} = \partial^2 l / \partial \beta \partial \gamma^T$  of  $l$  at  $\beta$  and  $\gamma$  is

$$\begin{aligned} H_{\beta,\gamma} = & \sum_{i=1}^m \left[ x_i B^T(x_i^T \beta) - \sum_{l \in R_i} w_{li} \{ x_l B^T(x_l^T \beta) + \gamma^T B(x_l^T \beta) x_l \tilde{B}^T(x_l^T \beta) \} \right. \\ & \left. + \sum_{l \in R_i} w_{li} \gamma^T B(x_l^T \beta) x_l \sum_{l \in R_i} w_{li} \tilde{B}^T(x_l^T \beta) \right] \end{aligned}$$

with  $w_{li}$  defined as above. The full Hessian matrix is then given by

$$H = \begin{pmatrix} H_{\beta,\beta} & H_{\beta,\gamma} \\ H_{\beta,\gamma}^T & H_{\gamma,\gamma} \end{pmatrix}.$$

The joint score vector is  $S = (S_\beta^T, S_\gamma^T)^T$ .

We apply an iterative alternating optimization procedure to calculate the maximum partial likelihood estimate that employs the specific structure of the problem and is numerically stable. Note that for fixed  $\beta$  the partial likelihood  $l(\beta, \gamma)$  is concave as a function of  $\gamma$  whose maximum is uniquely defined if it exists. We solve the maximization problem by iteratively maximizing  $l(\beta, \gamma)$  over  $\beta$  and  $\gamma$ . More specifically, for the fixed current value  $\hat{\beta}_c$  of  $\beta$ , we update the estimate of  $\gamma$  by maximizing  $l(\hat{\beta}_c, \gamma)$ , and for the fixed current value  $\hat{\gamma}_c$  of  $\gamma$ , we update the estimate of  $\beta$  by maximizing  $l(\beta, \hat{\gamma}_c)$ , the process is iterated until some convergence criterion is met. We find that the proposed procedure is easy to implement and the algorithm usually converges fast in our simulation study.

While it is possible to maximize the log-partial likelihood in (5) simultaneously w.r.t  $\beta$  and  $\gamma$ , we find the iterative alternating optimization more appealing. The iterative procedure is numerically more stable and computationally simpler, as there is

no need to calculate a bigger Hessian matrix as would do if simultaneous optimization were implemented with the Newton-Raphson algorithm. The idea of iterative optimization is not new, which has been well-studied in the numerical analysis literature (Ruhe and Wedin, 1980). Wang et al. (2001) has developed an iterative optimization procedure for model (2) when the local smoothing method is used to estimate the link function; see also Weisberg and Welsh (1994) and Chiou and Müller (1998, 1999) for application of similar ideas. Since the local smoother does not have a parsimonious representation, an iterative algorithm is necessary for implementing their method. This is different from our approach, where our estimator is defined as the maximizer of a global criterion and the iterative algorithm is only a computational device for efficient numerical optimization.

### 2.3 Inference

As we discussed in section 2, we need a constraint  $\|\beta\|^2 = 1$  for identifiability. To get the variance-covariance matrix of  $(\hat{\beta}, \hat{\gamma})$ , we reparametrize by writing  $\beta = ((1 - \|\alpha\|^2)^{1/2}, \alpha_1, \dots, \alpha_{p-1})^T$  with  $\alpha = (\alpha_1, \dots, \alpha_{p-1})^T$ . The observed Fisher information  $I(\alpha, \gamma)$  of  $(\alpha, \gamma)$  equals  $-H(\alpha, \gamma)$ , the negative of the Hessian of  $(\alpha, \gamma)$ . By a standard application of the martingale theory of partial likelihood (e.g., an extension of the arguments by Prentice and Self, 1983), we can show that  $(\hat{\alpha}, \hat{\gamma})$  is asymptotically normal with an estimated asymptotic variance-covariance matrix  $\{I(\hat{\alpha}, \hat{\gamma})\}^{-1}$ . By the delta method,  $(\hat{\beta}, \hat{\gamma})$  is also asymptotically normal with an estimated asymptotic variance-covariance matrix

$$\begin{aligned} \text{var}(\hat{\beta}, \hat{\gamma}) &= \begin{pmatrix} \frac{\hat{\beta}_2}{\hat{\beta}_1} & \dots & \frac{\hat{\beta}_p}{\hat{\beta}_1} & 0 & \dots & 0 \\ & & I_{p+k-1} & & & \end{pmatrix} \\ &\quad \times \{-H(\hat{\alpha}, \hat{\gamma})\}^{-1} \begin{pmatrix} \frac{\hat{\beta}_2}{\hat{\beta}_1} & \dots & \frac{\hat{\beta}_p}{\hat{\beta}_1} & 0 & \dots & 0 \\ & & I_{p+k-1} & & & \end{pmatrix}^T, \end{aligned}$$

where  $I_{p+k-1}$  is the  $(p+k-1) \times (p+k-1)$  identity matrix, and the formula of  $H(\alpha, \gamma)$  is given in the Appendix A. The appropriate diagonal elements of the above matrix will give  $\text{var}(\hat{\beta}_i)$ ,  $i = 1, \dots, p$ , and thus an approximate 95% confidence interval for  $\beta_i$  is  $\hat{\beta}_i \pm 1.96 \{\text{var}(\hat{\beta}_i)\}^{1/2}$ . For a fixed  $u$ , the variance of estimated function evaluated at  $u$  can be estimated as  $\text{var}(\hat{\psi}(u)) = \tilde{B}(u)^T \text{var}(\hat{\gamma}) \tilde{B}(u)$ , where  $\text{var}(\hat{\gamma})$  is given by the appropriate submatrix of  $\text{var}(\hat{\beta}, \hat{\gamma})$ . An approximate 95% confidence interval for  $\psi(u)$  is  $\hat{\psi}(u) \pm 1.96 \{\text{var}(\hat{\psi}(u))\}^{1/2}$ .

Since the standard Cox model is nested in our more general model (2), we can perform the likelihood ratio test to test whether the standard Cox model is appropriate. Specifically, the test statistic is defined as

$$LR = -2(\log\{\text{likelihood}_{\text{Cox}}\} - \log\{\text{likelihood}_{\text{Spline}}\})$$

where  $\text{likelihood}_{\text{Cox}}$  and  $\text{likelihood}_{\text{Spline}}$  denote respectively the values of the partial likelihood for the fitted standard Cox model and the Cox model with spline estimated link. Under the null hypothesis that the standard Cox model holds, the  $LR$  statistic approximately has a  $\chi^2$  distribution with  $r$  degrees of freedom, where  $r = nknot + d - 2$ , with  $nknot$  being the number of knots and  $d$  being the degree of the spline. We reject the standard Cox model at significance level  $\alpha$  when the observed  $LR$  statistic is larger than  $\chi_{r,\alpha}^2$ , where  $\chi_{r,\alpha}^2$  is the upper  $\alpha$ -th percentile of the  $\chi^2$  distribution with  $r$  degrees of freedom. It is also straightforward to develop a Wald-type test, details of which are omitted.

Rigorously speaking, the asymptotic inference developed above is only valid conditional on the knots of spline. In practice, the number of knots may be selected using the data. It is advisable to check the sensitivity of the result with respect to the number of knots.

**Remark.** In this paper we only consider splines with equally spaced knots (see Section 2.5). The method works well in our simulated and real examples. For methods

involving adaptive knot selection (Stone, et al., 1997), straightforward application of above asymptotic inference would certainly underestimate the variability of parameter estimators, while bootstrapping has been used as an alternative.

## 2.4 Implementation

An algorithm for implementing the proposed iterative alternating optimization procedure is described as below.

Step 1. Start with an initial guess  $\hat{\beta}^{(0)}$  and  $\hat{\gamma}^{(0)}$ .

Step 2. Given the current values  $\hat{\beta}^{(s)}$  and  $\hat{\gamma}^{(s)}$  of  $\beta$  and  $\gamma$ , we update the estimate of  $\beta$  using one step of the Newton-Raphson method, in which  $\hat{\beta}^{(s+1)}$  is obtained according to

$$\hat{\beta}^{(s+1)} = \hat{\beta}^{(s)} - \{H_{\hat{\beta}^{(s)}, \hat{\beta}^{(s)}}(\hat{\beta}^{(s)}, \hat{\gamma}^{(s)})\}^{-1} S_{\hat{\beta}^{(s)}}(\hat{\beta}^{(s)}, \hat{\gamma}^{(s)}).$$

Standardize  $\hat{\beta}^{(s+1)}$  such that  $\|\hat{\beta}^{(s+1)}\| = 1$  and its first component is positive.

Step 3. Given the current values  $\hat{\beta}^{(s+1)}$  and  $\hat{\gamma}^{(s)}$  of  $\beta$  and  $\gamma$ , we update the estimate of  $\gamma$  using one step of the Newton-Raphson method with step-halving, in which  $\hat{\gamma}^{(s+1)}$  is determined according to

$$\hat{\gamma}^{(s+1)} = \hat{\gamma}^{(s)} - 2^{-k} \{H_{\hat{\gamma}^{(s)}, \hat{\gamma}^{(s)}}(\hat{\beta}^{(s+1)}, \hat{\gamma}^{(s)})\}^{-1} S_{\hat{\gamma}^{(s)}}(\hat{\beta}^{(s+1)}, \hat{\gamma}^{(s)})$$

where  $k$  is the smallest nonnegative integer such that

$$l(\hat{\beta}^{(s+1)}, \hat{\gamma}^{(s)} - 2^{-k} \{H_{\hat{\gamma}^{(s)}, \hat{\gamma}^{(s)}}(\hat{\beta}^{(s+1)}, \hat{\gamma}^{(s)})\}^{-1} S_{\hat{\gamma}^{(s)}}(\hat{\beta}^{(s+1)}, \hat{\gamma}^{(s)})) \geq l(\hat{\beta}^{(s+1)}, \hat{\gamma}^{(s)}).$$

Repeat steps 2 and 3 until some convergence criterion is met.

The log-partial likelihood (5) is concave in  $\gamma$  but not necessarily in  $\beta$ . Thus this algorithm is not guaranteed to converge to the maximum partial likelihood estimate. We recommend repeating the algorithm using a variety of (usually randomly chosen) initial values and choose the final estimate by maximizing the log-partial likelihood. Our experiments show that our implementation of the algorithm performs very well.

The algorithm may not converge if the initial values are far from the truth, but it usually converges within a few steps for reasonable initial values. In our simulation study, we run the program with 10 different random starts and the program always converges for more than half of the 10 choices of initial values. (In our implementation we terminate the program if it fails to converge in 30 iterations.)

### 2.5 *Choosing the number and position of knots*

In our implementation, for a given number of knots, we put the knots equally-spaced between the smallest and the largest values of  $\beta^T X$ . For a smooth function, 3 to 10 knots are usually adequate and the results are quite stable when we vary the number of knots. For those multimodal functions which have many local maxima or minima, more knots may be needed. In our implementation, we vary the number of knots in a relatively large range and choose the one which minimizes the AIC or BIC. Here AIC and BIC are defined as

$$\text{AIC} = -2\log(\text{likelihood}) + 2(nknot + d + p - 2)$$

$$\text{BIC} = -2\log(\text{likelihood}) + \log(n)(nknot + d + p - 2)$$

where  $nknot$  and  $d$  denote respectively the number of knots and degree of the spline,  $p$  is the number of covariates and  $n$  is the number of subjects. We also check the sensitivity of the results to different number of knots. The knot positions change in each iteration of the program since  $\beta$  is updated, but should stabilize when the program converges. For a large number of knots, if there is not enough data on the boundary, we may need to put knots unequally spaced to avoid singularity problems. For example, we can put more knots within the first and last deciles of observed values of  $\beta^T X$  and less knots outside of them.

## 3. Interpretation of the fitted model

Since the fitted link function may be nonlinear and even nonmonotone, the interpretation of covariate effects using the fitted model is not as straightforward as the standard Cox model. In particular, the effect of a particular covariate may be dependent on the values of other covariates. However, using average derivatives (Chaudhuri, Doksum and Samarov, 1997), the regression coefficients in the fitted model can still be given an interpretation similar to the usual intuitive idea of what regression coefficients are.

In the standard Cox model the partial derivatives  $\partial \log \lambda(t|x)/\partial x_i$ ,  $x = (x_1, \dots, x_p)$ , are assumed to be constant (i.e., independent of  $t$  and  $x$ ) and are called regression coefficients. They measure how much the log conditional hazard function is changed as the  $i$ th covariate is perturbed while other covariates are held fixed. In model (2), the partial derivatives need not be a constant. To give a concise summary of covariate effect, we consider the average gradient vector

$$\nu = (\nu_1, \dots, \nu_p) = E[\nabla \log \lambda(t|X)]$$

which gives the average change in the log conditional hazard function as the  $i$ th covariate is perturbed while other covariates are held fixed. Note that model (2) implies that  $\nu = c\beta_0$ , where the constant  $c = E[\psi'(x^T\beta_0)]$ . It follows that  $\beta_0$  can be interpreted as  $\nu$ . Note further that only the direction of  $\beta_0$  is identifiable and  $\beta_{0i}/\beta_{0j} = \nu_i/\nu_j$  so that  $\beta_{0i}$  and  $\beta_{0j}$  give the relative importance of the covariates  $X_i$  and  $X_j$ .

The regression coefficient in the fitted model can also be interpreted using average derivatives of a quantile regression. To this end, note that model (2) can be written in the form of a transformation model

$$h(T) = \psi(\beta_0^T X) + \epsilon,$$

where  $h(t) = \log\{-\log[1 - F_0(t)]\}$ , and the distribution  $F_0$  of  $\epsilon$  is equal to the extreme

value distribution  $1 - \exp[-\exp(t)]$ . For a given  $0 < \alpha < 1$ , the  $\alpha$ -th conditional quantile of  $T$  given  $X = x$  is

$$\theta_\alpha(x) = h^{-1}\{\psi(\beta_0^T x + e_\alpha)\}$$

where  $e_\alpha$  is the  $\alpha$ -th quantile of  $\epsilon$ . The average derivatives of the quantile function  $\theta_\alpha(x)$  are given by

$$\nu_\alpha = E[\nabla\theta_\alpha(X)] = c_\alpha\beta_0, \text{ where } c_\alpha = E\{[(h^{-1})'(\psi(\beta_0^T X + e_\alpha))]\psi'(\beta_0^T X + e_\alpha)\}.$$

Thus, the components of  $\beta_0$  give the relative importance of the covariates for interpreting the average change in a quantile (e.g., median) survival time as the  $i$ th covariate is perturbed while the others are held fixed.

The usual interpretation of  $\beta$  in the standard Cox model as log hazard ratios does not involve the covariate distribution. The interpretation of  $\nu$  and  $\nu_\alpha$  as average derivatives depends heavily on the distribution of covariates, which may vary across populations. This dependence limits the generalizability of the results on  $\nu$  and  $\nu_\alpha$  across studies/populations, where covariate distributions may differ. In practice, however, one can compute by definition the average derivative  $\nu$  or  $\nu_\alpha$  for any given population, where the expectation in the definition can be approximated by the sample average when a random sample of covariates is available. On the other hand, the interpretation of  $\beta_0$  does not involve the covariate distribution, since the dependence on covariate distribution in  $\nu$  or  $\nu_\alpha$  is totally captured by  $c = E[\psi'(x^T\beta_0)]$  or  $c_\alpha = E\{[(h^{-1})'(\psi(\beta_0^T X + e_\alpha))]\psi'(\beta_0^T X + e_\alpha)\}$ .

Sometimes covariate effects can be better understood using graphical tools. Suppose  $x = (x_1, x_2)$ , where  $x_1 = 0$  or  $1$  is a treatment indicator and  $x_2$  represents other covariates collectively. Model (2) implies that the treatment effect in terms of log hazard at  $x_2$  is

$$\log \lambda(t|1, x_2) - \log \lambda(t|0, x_2) = \psi(\beta_{01} + \beta_{02}^T x_2) - \psi(\beta_{02}^T x_2), \quad (7)$$

which does not depend on  $t$ , but in general depends on the value of  $x_2$ . To see the dependence of treatment effect on the covariate values, one can plot the above log hazard ratio as a function of  $\beta_{02}^T x_2$ , with  $\psi(\cdot)$  and  $\beta$  replaced by fitted values, and  $x_2$  random draws from a given distribution that reflects the target population. A loess smooth with a large span can aid in understanding the treatment effect. The overall treatment effect can be measured using the log hazard ratio averaged over  $x_2$ ; a positive or negative average log hazard ratio will indicate respectively a beneficial or harmful overall effect of treatment. The idea is illustrated in the data analysis in Section 5.

When the estimated relative risk form is monotone, covariate effects can be interpreted qualitatively using the sign of components of  $\beta$ . If  $\psi$  is monotone increasing (decreasing), then a positive sign for the coefficient of a particular covariate suggests increased (decreased) risk at larger values of the covariate, and vice versa for a negative sign. Our methodology can be extended to incorporate monotone constraint on the relative risk form. For example, one could model  $\log \psi'(\beta^T x)$  instead of  $\psi(\beta^T x)$  using a spline in (3). Such an extension would be useful if one had prior knowledge that the relative risk form is monotone. However, such knowledge is usually not available and it is the advantage of the proposed method to let the data to pick the right form. If the actual relative risk form is monotone, the estimated form should be so or closely so.

#### 4. Simulation study

To evaluate how our method works, we have conducted simulation studies and found our method in general performs well. Here we report the results of one of the simulation studies. We use the angle between the true direction and the estimated direction to measure the performance of the estimate. Specifically, the angle between  $\hat{\beta}$  and  $\beta_0$

is defined as

$$\angle(\hat{\beta}, \beta_0) = \arccos\left(\frac{\langle \hat{\beta}, \beta_0 \rangle}{\|\hat{\beta}\| \|\beta_0\|}\right).$$

Let the baseline hazard  $\lambda_0$  be a constant,  $\lambda_0 = 1$ . We use exponential distribution to generate the independent censoring time and, partly by trial and error, select the exponential rate parameter to achieve appropriate censoring levels. The iterative algorithm of our spline method starts from an initial value of  $\beta$  randomly drawn from  $N(0, I_p)$  (standardized to have norm 1) and  $\gamma$  randomly drawn from  $N(0, I_k)$ , where  $I_p$  and  $I_k$  are the  $p \times p$  and  $k \times k$  identity matrices. We do ten random starts and then decide on the final estimate based on maximizing the partial log-likelihood.

In this example, the true log relative risk function is  $\psi(\beta_0^T X) = 5 \sin\{(1/2)\beta_0^T X\}$ , where  $X_i \sim N(1, 4), i = 1, \dots, 5$ , and  $\beta_0 = (1, -1, 1, -1, 1)^T / \sqrt{5}$ . We report here mainly results for sample size  $n = 300$  and censoring rate 30%. The number of simulations is 500. For this example, we use splines with five equally spaced knots to approximate the  $\psi(\cdot)$  in our estimation procedure. Results not reported here also show that the estimate is not very sensitive to the number of knots. We compare the performance of the proposed estimate with estimated unknown relative risk form to the maximum partial likelihood estimates with the relative risk form being as in the standard Cox model and with the true known form. Note that the true known relative risk form is not available in real applications, but the performance of the estimate using the true form information serves as a benchmark to gauge the performance of our proposed estimate.

Summary statistics for the angles between the estimated and the true directions are reported in Table 1 (results for 20% censoring rate,  $n = 300$ , and for 40% censoring rate,  $n = 150$  are also shown in this table). We observe that, when the relative risk form is misspecified, the estimate based on the wrong form can give seriously biased estimate of the true direction of  $\beta$ . On the other hand, the proposed method is robust

to the violation. In fact, the proposed estimate of  $\beta$  when the relative risk form is unknown is close to the estimate based on the known true function form. Results for assessing the accuracy of the standard error formula are given in Table 2. We find the proposed standard error estimate works well: the sample means of estimated standard errors of  $\beta$ s are reasonably close to the Monte Carlo standard deviations of the estimates; the Monte Carlo coverage probabilities of the 95% confidence intervals are reasonably close to the nominal level. Figure 2 shows the average of fitted link functions and 95% pointwise Monte Carlo intervals based on 500 simulation runs. (The 2.5% and 97.5% quantiles of the fitted values from the 500 simulation runs are the two end points of a 95% Monte Carlo interval.) The fitted function follows the true function closely. Experimentation with other smooth functions, censoring rates, and sample sizes yields the same conclusions.

[Table 1 about here.]

[Table 2 about here.]

[Figure 2 about here.]

## 5. Data analysis

The Veteran's Administration lung cancer data was used by Kalbfleisch and Prentice (2002) to illustrate the Cox model. In this clinical trial, males with advanced inoperable lung cancer were randomized to either a standard or test chemotherapy. The primary end point for therapy comparison was time to death. Only 9 of the 137 survival times were censored. As is common in such studies, there was much heterogeneity between patients in disease extent and pathology, previous treatment of the disease, demographic background, and initial health status. The data set includes 6 covariate variables measuring this heterogeneity: Treatment, Age, Karnofsky score,

Diagnosis time, Cell type, and Prior therapy. See Kalbfleisch and Prentice (2002) for an explanation of these variables.

We have applied the proposed method to fit a Cox model with flexible relative risk form and compared the result with the standard Cox model. The logarithm of the unknown risk form is fitted as a spline function with 4 knots which is chosen by both the AIC and BIC criteria. The Cox model with flexible relative risk form yields a larger log-likelihood, which is  $-464.16$  compared with  $-475.18$  for the standard Cox model. The fitted link function and corresponding 95% pointwise confidence interval are given in Figure 1. The identity function lies outside the 95% confidence interval for most of the range of the data, indicating that the proportional hazards model is not appropriate for this data set. The likelihood ratio test also rejects the standard Cox model ( $p$  value = 0.0002). Our conclusion is not sensitive to the number of knots. As an illustration, the fitted link function and corresponding pointwise confidence intervals for 8 knots are also plotted in Figure 1; results for other numbers (i.e., 5-7) of knots are similar.

[Table 3 about here.]

We present in Table 3 the estimated parameters and corresponding standard errors. To compare the results of the standard Cox model and the model with flexible relative risk form, we rescaled the estimates of the model with unknown link such that the coefficient vector has the same norm as that for the Cox model with identity link. The last column of Table 3 shows the rescaled estimates and corresponding standard errors of the parameters; the calculation of the standard errors is given in Appendix B. The results for the standard Cox model, which agree with those reported in Kalbfleisch and Prentice (2002, page 120), are quite different from those for the model with flexible relative risk form. The angle between the vectors of regression

coefficients from the two models is 79.67°. It is interesting that the estimated covariate effects from the model with flexible relative risk form are rather different from the standard Cox model. Treatment, age, diagnostic time, squamous vs large, and small versus large are statistically significant with unknown link, but not with identity link. Adeno versus large is significant with identity link, but not in the model with unknown link. As noted in the Section 3, the fitted regression coefficients give the relative importance of the covariates in explaining survival and can not be interpreted in the same way as in the standard Cox model, particularly since the estimated link is bimodal. Because  $\hat{\psi}$  is not monotone increasing (decreasing), a positive sign for the coefficient of a particular covariate need not suggest increased (decreased) risk at larger values of the covariate, and vice versa for a negative sign.

To compare the test and standard treatments using the model with flexible relative risk form, we compute the log hazard ratio as defined in (7) with  $\psi(\cdot)$  and  $\beta$  replaced by fitted values, using  $x_1 = 1$  for the test treatment and  $x_1 = 0$  for the standard treatment. We plot in Figure 3 the log hazard ratio as a function of  $\beta_{02}^T x_2$  where  $x_2$  takes all values in the sample. The average log hazard ratio is 0.432, indicating that overall the test treatment is worse than the standard treatment for the population where the sample is from. Note the average log hazard ratio should be re-calculated for a different population, which is likely to remain positive as suggested by the loess smooth in Figure 3.

[Figure 3 about here.]

## 6. Discussion

In this paper we employ a working spline model on the unknown link function in the log conditional hazard function. The working model involves only finite number of parameters and associated estimation and inference procedures work well even when

the link function is not a spline function. Rigorously speaking the partial likelihood theory used for asymptotic inference requires that the number and locations of knots are fixed and the corresponding spline model for  $\psi(\cdot)$  is correctly specified. We have ignored the bias caused by spline approximation in developing the asymptotic theory. In many practical situations, however, such a bias is of a relatively small magnitude compared with variance and its effects on the inference for the regression coefficients  $\hat{\beta}$  could be negligible. Moreover, the bias caused by spline approximation is often of much smaller magnitude compared with that by using the misspecified Cox model (with identity link). These points have been demonstrated in our simulation studies and illustrated in the reported simulated example.

It has been shown in many contexts of function estimation that, by letting the number of knots increase with the sample size at an appropriate rate, spline estimate of an unknown function can achieve the optimal nonparametric rate of convergence (see for example, Stone 1994, Huang 1998, 2001, 2003, and Huang et al. 2000). Such asymptotic results take into account the bias caused by spline approximation. Developing similar results in the current context is important and practically relevant but beyond the scope of this paper.

#### ACKNOWLEDGEMENTS

Jianhua Z. Huang's work is partially supported by National Science Foundation grant DMS-0204556. We are grateful to an AE and a referee for thoughtful comments. A suggestion from the AE has led to the inclusion of Section 3. The authors thank Professors Tony Cai, Edward George, Daniel Heitjan and Paul Shaman for helpful discussions. We would also like to thank Dr. Wei Wang for making her unpublished paper available to us.

#### REFERENCES

Breslow, N.E. (1972). Contribution to the discussion of “Regression models and life-tables” by D. R. Cox. *Journal of the Royal Statistical Society, Series B* **34**, 216-217.

Chaudhuri, P., Doksum, K. and Samarov, A. (1997). On average derivative quantile regression. *The Annals of Statistics* **25**, 715–744.

Chiou, J.-M. and Müller, H.-G. (1998). Quasi-likelihood regression with unknown link and variance functions. *Journal of the American Statistical Association* **93**, 1376-1387.

Chiou, J.-M. and Müller, H.-G. (1999). Nonparametric quasi-likelihood. *Annals of Statistics* **27**, 36-64.

Cox, D. R. (1972). Regression models and life-table (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187-220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269-276. In a rigorous treatment of asymptotic theory that can , one should let

de Boor, C. (1978). *A Practical Guide to Splines*. New York: Springer.

Fan, J., Gijbels, I. and King, M. (1997). Local likelihood and local partial likelihood in hazard regression. *Annals of Statistics* **25**, 1661-1690.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.

Gentleman, R. and Crowley, J. (1991). Local full likelihood estimation for the proportional hazards model. *Biometrics* **47**, 1283-1296.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association* **87**, 942-951.

Gu, C. (1996). Penalized likelihood hazard estimation: a general procedure. *Statistica Sinica* **6**, 861-876.

- Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *The Annals of Statistics*, **26**, 242-272.
- Huang, J. Z. (2001). Concave extended linear modeling: A theoretical synthesis. *Statistica Sinica*, **11**, 173-197.
- Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics*, **31**, 1600-1635.
- Huang, J. Z., Kooperberg, C., Stone, C. J., Truong, Y. K. (2000). Functional ANOVA modeling for proportional hazards regression. *Annals of Statistics* **28**, 960-999.
- Liu, L. (2004). Semiparametric and nonparametric models for survival data. Ph. D. thesis, Department of Statistics, University of Pennsylvania.
- Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* **58**, 71-120.
- Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd Edition. New York: Wiley.
- Kooperberg, C., Stone, C. J. and Truong, Y. K. (1995). Hazard regression. *Journal of the American Statistical Association* **90**, 78-94.
- O'Sullivan, F. (1993). Nonparametric estimation in the Cox model. *Annals of Statistics* **21**, 124-145.
- Peto, R. (1972). Contribution to the discussion of "Regression models and life-tables" by D. R. Cox. *Journal of the Royal Statistical Society, Series B* **34**, 205-207.
- Prentice, R. L. and Self, S. (1983). Asymptotic distribution theory for Cox-type regression models with general relative risk form. *Annals of Statistics* **11**, 804-813.
- Ruhe, A. and Wedin, P. (1980). Algorithms for separable nonlinear least squares problems. *SIAM Review* **22**, 318-337.
- Schumaker, L. L. (1981). *Spline Functions: Basic Theory*. New York: Wiley.

Sleeper, L. A. and Harrington, D. P. (1990). Regression splines in the Cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association* **85**, 941-949.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Annals of Statistics* **22**, 118–171.

Stone, C. J., Hansen, M., Kooperberg, C. and Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *Annals of Statistics* **25**, 1371-1470.

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association* **82**, 559-567.

Wang, W. (2004). Proportional hazards regression models with unknown link function and time-dependent covariates. *Statistica Sinica* **14**, 885–905.

Wang, W. (2001). Proportional hazards model with unknown link function and applications to longitudinal time-to-event data. Ph.D. dissertation of the University of California at Davis.

Wang, W., Wang, J.-L. and Wang Q. (2001). Proportional hazards regression with unknown link function. Manuscript.

Weisberg, S. and Welsh, A.H. (1994). Adapting for the missing link. *Annals of Statistics* **22**, 1674-1700.

## APPENDIX A

### *The expression of $H(\alpha, \gamma)$*

Now we give the expression of  $H(\alpha, \gamma)$ . Write

$$H(\alpha, \gamma) = \begin{pmatrix} H_{\alpha,\alpha} & H_{\alpha,\gamma} \\ H_{\alpha,\gamma}^T & H_{\gamma,\gamma} \end{pmatrix},$$

where  $H_{\alpha,\alpha} = \partial^2 l / (\partial \alpha \partial^T \alpha)$  etc. Note that  $H_{\gamma,\gamma}$  is given in (6). Let  $\tilde{x}_i = (x_{i2}, \dots, x_{ip})^T$ ,  $\xi_i = -x_{i1}\alpha / (1 - \|\alpha\|^2)^{1/2} + \tilde{x}_i$ ,  $i = 1, \dots, n$ . Let  $A = (a_{ij})$  be a  $(p-1) \times (p-1)$  matrix with entries  $a_{ii} = 1 + \alpha_i^2 / (1 - \|\alpha\|^2)$  and  $a_{ij} = \alpha_i \alpha_j / (1 - \|\alpha\|^2)$ ,  $i \neq j$ ,  $i, j = 1, \dots, p-1$ .

Then it is easy to derive that

$$\begin{aligned} H_{\alpha,\alpha} &= \sum_{i=1}^m \gamma^T B(x_i^T \beta) \left( -\frac{x_{i1}}{\sqrt{1 - \|\alpha\|^2}} \right) A + \sum_{i=1}^m \gamma^T B'(x_i^T \beta) \xi_i \xi_i^T \\ &\quad - \sum_{i=1}^m \sum_{l \in R_i} w_{li} \left[ \gamma^T B(x_l^T \beta) \left( -\frac{x_{l1}}{\sqrt{1 - \|\alpha\|^2}} \right) A + \gamma^T B'(x_l^T \beta) \xi_l \xi_l^T \right] \\ &\quad + \{ \gamma^T B(x_l^T \beta) \}^2 \xi_l \xi_l^T + \sum_{i=1}^m \left\{ \sum_{l \in R_i} w_{li} \gamma^T B(x_l^T \beta) \xi_l \sum_{l \in R_i} w_{li} \gamma^T B(x_l^T \beta) \xi_l^T \right\} \end{aligned}$$

and

$$\begin{aligned} H_{\alpha,\gamma} &= \sum_{i=1}^m \xi_i B^T(x_i^T \beta) - \sum_{i=1}^m \sum_{l \in R_i} w_{li} \{ \xi_l B^T(x_l^T \beta) + \gamma^T B(x_l^T \beta) \xi_l \tilde{B}^T(x_l^T \beta) \} \\ &\quad + \sum_{i=1}^m \left\{ \sum_{l \in R_i} w_{li} \gamma^T B(x_l^T \beta) \xi_l \sum_{l \in R_i} w_{li} \tilde{B}^T(x_l^T \beta) \right\}, \end{aligned}$$

where  $\beta = ((1 - \|\alpha\|^2)^{1/2}, \alpha_1, \dots, \alpha_{p-1})^T$ .

## APPENDIX B

### *Standard errors of the rescaled coefficients used in Table 3*

Let  $\hat{\beta}_{\text{Cox}}$ ,  $\hat{\beta}_{\text{spline}}$  and  $\hat{\beta}_{\text{rescaled}}$  denote the estimated parameters of the standard Cox model, the model with flexible relative risk form and the rescaled estimates. Assume  $\hat{\beta}_{\text{rescaled}} = c \hat{\beta}_{\text{spline}}$  where  $c = \|\hat{\beta}_{\text{rescaled}}\| = \|\hat{\beta}_{\text{Cox}}\|$  and denote  $\hat{\beta}_{\text{spline}}^j$  and  $\hat{\beta}_{\text{rescaled}}^j$  as the  $j$ -th components of  $\hat{\beta}_{\text{spline}}$  and  $\hat{\beta}_{\text{rescaled}}$ . Then, by the delta-method,

$$\text{AVar}(\hat{\beta}_{\text{spline}}^j) = \text{AVar} \left\{ \frac{\hat{\beta}_{\text{rescaled}}^j}{\|\hat{\beta}_{\text{rescaled}}\|} \right\} = \left\{ \frac{1}{\|\hat{\beta}_{\text{rescaled}}\|} - \frac{(\hat{\beta}_{\text{rescaled}}^j)^2}{\|\hat{\beta}_{\text{rescaled}}\|^3} \right\}^2 \text{AVar}(\hat{\beta}_{\text{rescaled}}^j),$$

where  $\text{AVar}(\cdot)$  denotes the asymptotic variance. Plugging in  $\|\hat{\beta}_{\text{rescaled}}\| = \|\hat{\beta}_{\text{Cox}}\|$  and  $\hat{\beta}_{\text{rescaled}}^j = \|\hat{\beta}_{\text{Cox}}\| \hat{\beta}_{\text{spline}}^j$ , we obtain

$$\text{AVar}(\hat{\beta}_{\text{rescaled}}^j) = \frac{\|\hat{\beta}_{\text{Cox}}\|^2}{[1 - (\hat{\beta}_{\text{spline}}^j)^2]^2} \text{AVar}(\hat{\beta}_{\text{spline}}^j).$$

## APPENDIX C

### *Asymptotic result*

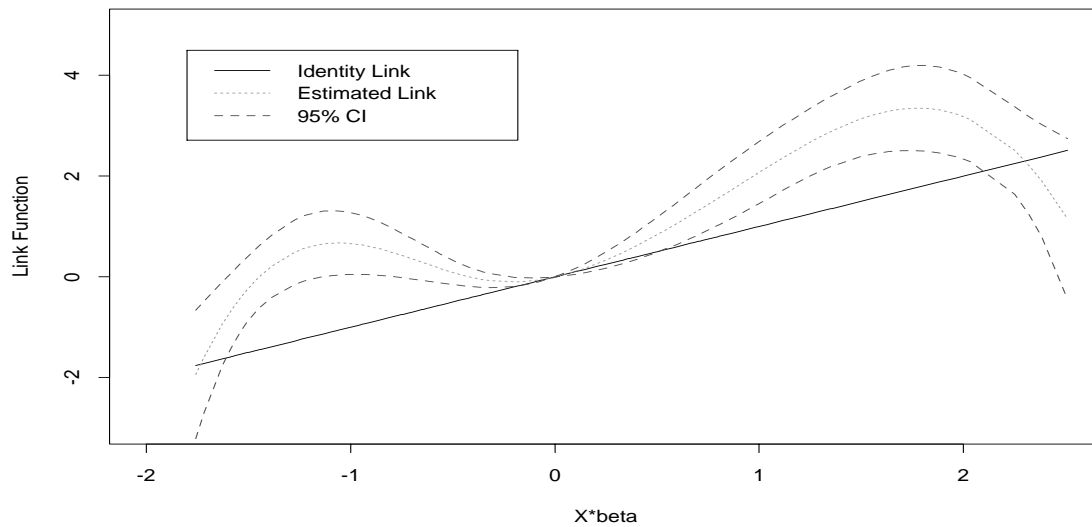
We show that the arguments by Prentice and Self (1983) can be used to derive the large sample property of the proposed estimates when the link function is a spline with pre-specified knots. Such a development ignores the approximation error of the spline approximation, but appears to provide a reasonable good approximation in our simulation study.

Prentice and Self (1983) specifies the relative risk form as  $r(X^T \beta)$  where  $r$  is a known function on  $R$ . Their arguments can be extended easily to the case where the relative risk form is  $g(X; \theta)$  with  $g$  defined on a multivariate domain. The difference is that we need do a Taylor expansion of  $g(x; \theta)$  around  $\theta_0$ , while they do a Taylor expansion of  $r(x^T \beta)$  around  $x^T \beta_0$ . In our context,

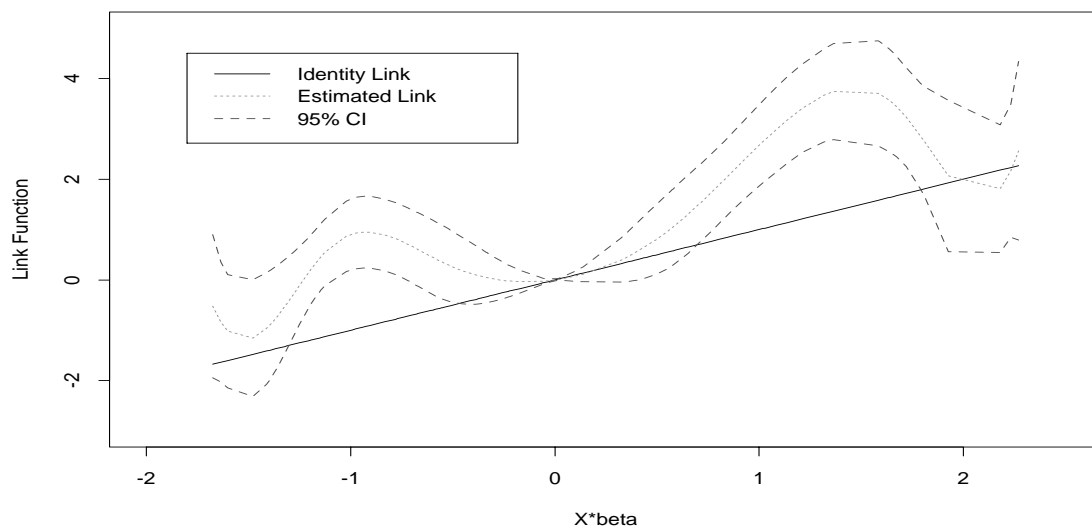
$$g(X; \theta) = \sum_{j=1}^k \gamma_j \tilde{B}_j(X^T \beta(\alpha)), \quad \theta = (\gamma, \alpha), \quad \gamma = (\gamma_1, \dots, \gamma_k),$$

where  $\beta(\alpha) = ((1 - \|\alpha\|^2)^{1/2}, \alpha_1, \dots, \alpha_{p-1})^T$  with  $\alpha = (\alpha_1, \dots, \alpha_{p-1})^T$  and  $\tilde{B}_j$ ,  $j = 1, \dots, k$ , are integrals of B-spline basis functions as defined following (4). Details can be found in Liu (2004). Note that the partial likelihood is no longer concave, so we can only claim that there is a consistent root of the likelihood equation that is asymptotically normal, just as we do for the standard parametric asymptotic for MLE when the likelihood is not concave.

The Estimated Link Function for the Cancer Data with 4 Knots

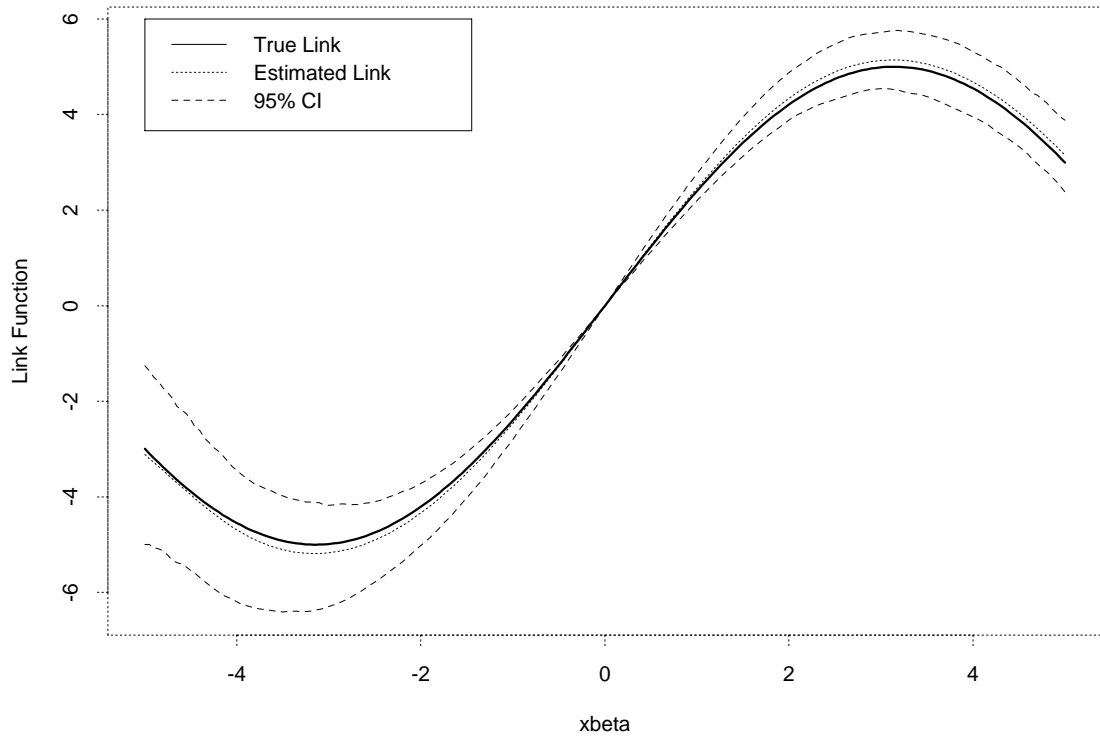


The Estimated Link Function for the Cancer Data with 8 Knots



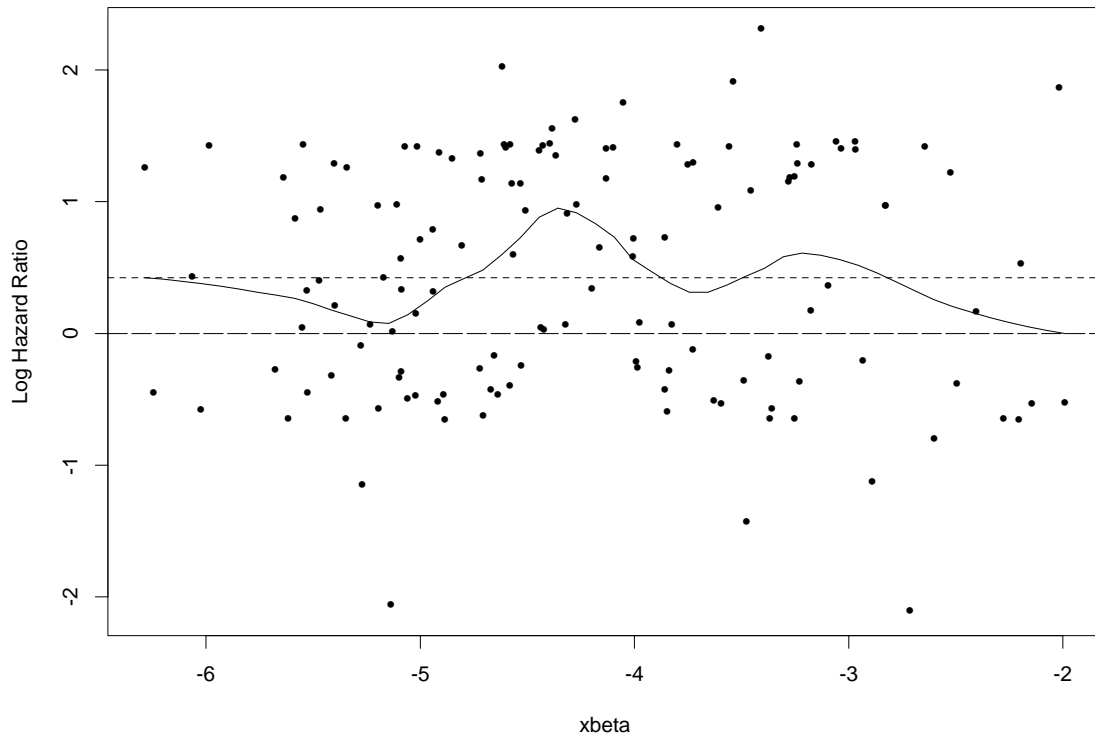
**Figure 1.** The estimated link function with 95% pointwise confidence interval for the VA Lung Cancer data for 4 knots and 8 knots. The solid line is the identity function. The dotted curve is the estimated link function and the dot dashed curves are the 95% pointwise confidence intervals.

### Estimate of the Link Function



**Figure 2.** The average of estimated link functions with 95% pointwise Monte Carlo intervals from 500 simulation runs. The 2.5% and 97.5% sample quantiles of the fitted values from the 500 simulation runs are the two end points of a 95% Monte Carlo interval. The solid curve is the true function. The dotted curve is the average of the estimated function. The dot dashed curves are 95% pointwise Monte Carlo intervals.

Estimated Treatment Effect Using Fitted Model for the Cancer Data



**Figure 3.** The log hazard ratio for treatment effect, defined as in (7), plotted as a function of  $\beta_{02}^T x_2$ . A loess smooth with span = 0.4 is superimposed (solid curve). The average log hazard ratio, which is 0.432, is plotted as a horizontal short-dashed line.

**Table 1**

*The summary statistics for the angles between the estimated and the true directions in the simulation example. The estimates considered are the maximum partial likelihood estimates based on the identity link, the true known link, and the proposed estimate with unknown link.*

---

---

Censoring rate	20% ( $n = 300$ )		30% ( $n = 300$ )		40% ( $n = 150$ )	
	mean	SD	mean	SD	mean	SD
Identity link	10.427°	4.463°	10.962°	4.444°	13.048°	5.952°
Known link	1.921°	0.681°	2.108°	0.754°	3.137°	1.239°
Unknown link	1.965°	0.724°	2.159°	0.789°	3.261°	1.305°

---

---

**Table 2**

*Results of the simulation study of the proposed method. ave.: sample average; SD: sample standard deviation; cov. prob.: empirical coverage probability of the 95% confidence interval. Based on 500 Monte Carlo simulations.  $n = 300$ . censoring rate = 30%.*

---

---

	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$	$\beta_5$
true $\beta$	0.447	-0.447	0.447	-0.447	0.447
ave. ( $\hat{\beta}$ )	0.447	-0.448	0.446	-0.445	0.448
SD ( $\hat{\beta}$ )	0.018	0.018	0.017	0.017	0.018
ave. {SE ( $\hat{\beta}$ )}	0.017	0.017	0.017	0.017	0.017
SD {SE ( $\hat{\beta}$ )}	0.002	0.002	0.002	0.002	0.002
cov. prob.	0.930	0.934	0.940	0.928	0.930

---

---

**Table 3**

*The VA lung cancer data. The estimates and corresponding standard errors (in the parentheses) for the standard Cox model (identity link), the model with flexible relative risk form (unknown link). Also shown are the rescaled fitted regression coefficients for the model with unknown link, where rescaling is such that the coefficient vector has the same Euclidean norm as that for the standard Cox model. Breslow's approximation is used in handling ties.*

Variable	Identity Link	Unknown Link	Unknown Link(Rescaled)
Treatment	0.290(0.207)	0.571(0.091)	0.592(0.140)
Age	-0.009(0.009)	-0.020(0.007)	-0.021(0.007)
Karnofsky score	-0.033(0.006)	-0.048(0.009)	-0.050(0.009)
Diagnosis time	-0.000(0.009)	0.013(0.003)	0.013(0.003)
Cell type			
Squamous vs. Large	-0.400(0.283)	-0.716(0.057)	-0.742(0.121)
Small vs. Large	0.457(0.266)	-0.373(0.103)	-0.387(0.124)
Adeno vs. Large	0.789(0.303)	-0.140(0.180)	-0.145(0.191)
Prior therapy	0.007(0.023)	-0.011(0.010)	-0.011(0.011)