ABSTRACT

LIN, JIANG. Topics in Application of Nonparametric Smoothing Splines. (Under the direction of Drs. Daowen Zhang and Marie Davidian)

There are two topics in this dissertation. The first topic is "Smoothing Parameter Selection in Nonparametric Generalized Linear Models via Sixth-order Laplace Approximation" and the second topic is "Smoothing Spline-based Score Tests for Proportional Hazards Models".

We present a new approach for the automatic selection of the smoothing parameter in nonparametric smoothing spline Generalized Linear Models (GLMs), using the Restricted Maximum Likelihood (REML) method and the sixth-order Laplace approximation of Raudenbush et al. (2000). The proposed approach is compared with Generalized Additive Mixed Model (GAMM, Lin and Zhang 1999) and Generalized Approximate Cross-Validation (GACV, Gu and Xiang 2001) through simulations and is shown to be effective.

We propose "score-type" tests for the proportional hazards assumption and for covariate effects in the Cox model, using the natural smoothing spline representation of the corresponding nonparametric functions of time or covariate. The tests are based on the penalized partial likelihood. By treating the inverse of the smoothing parameter as a variance component, we derive the score tests by testing an equivalent null hypothesis that the corresponding variance component is zero. The tests are shown to have size close to the nominal level and to provide good power against general alternatives in simulations. We apply the proposed tests to data from a cancer clinical trial.

Topics in Application of Nonparametric Smoothing Splines

by

Jiang Lin

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

STATISTICS

Raleigh

2004

Approved By:

Dr. Daowen Zhang Chair of Advisory Committee	Dr. Marie Davidian Co-Chair of Advisory Committee
Dr. John Monahan	Dr. Hao (Helen) Zhang

To my parents, my brother and my partner

Biography

Jiang Lin was born on January 1, 1975 in Wushan, Sichuan province, P.R.China. He completed his high school education in Wushan Middle School, Sichuan, P.R.China in 1993. Jiang entered the Department of Applied Mathematics at Tsinghua University, Beijing, P.R.China in 1993, where he obtained a B.S degree in Applied Mathematics in 1997. Since August 2000 Jiang has been studying in the graduate program of the Department of Statistics at North Carolina State University. He obtained a M.S degree in Statistics in May 2002 and continued to work on his Ph.D degree since. Upon graduation, Jiang will join GlaxoSmithKline at RTP, North Carolina, to start a position as a Senior Biostatistician.

Acknowledgements

First of all, I would like to thank my advisors Drs. Daowen Zhang and Marie Davidian for their guidance, encouragement and patience throughout the development of this dissertation. I would also like to thank the rest advisory committee memebers Dr. John Monahan and Dr. Hao (Helen) Zhang for their service and many valuable suggestions.

I would like to extend my thanks to the faculty in the Department of Statistics at NC State for their profound impact on my professional growth. I also thank my colleagues in the Statistics and Programming group at GlaxoSmithKline for their assistance and friendship.

Thanks to all my friends in the Sunday night movie group for making “life as a statistician” a more colorful one. Many thanks also go to David Dai and his parents for being the best hosts.

I thank my parents and my brother back at home in China for their constant support. Finally, I want to thank my partner, for always being with me.

Contents

List of Tables

List of Figures

Chapter 1

Motivation and Introduction to Smoothing Splines

In recent years, there has been much interest in the statistical literature in extending classical, parametric statistical models to nonparametric models. By "parametric", we mean that the model is characterized by a finite number of parameters; while in a "nonparametric" model, we do not have such restrictions; e.g., in a nonparametric regression model, we may only require that the regression function is smooth.

Nonparametric smoothing is one of the most popular techniques for such purposes. One of the main reasons for its popularity is: smoothing splines arise naturally as solutions to optimization problems in a roughness penalty approach, which provides a bridge towards classical, parametric statistical models. Also because of this reason, nonparametric smoothing technique can be applied to extending a wide variety of parametric models.

In this dissertation, we consider the application of smoothing splines to Generalized Linear Models and proportional hazards models, both of which have a "parametric regression model" flavor. We want to extend these models to nonparametric ones using smoothing splines.

The properties of smoothing splines have been studied extensively. We first give a brief introduction to smoothing splines as well as define some notation.

The univariate, natural polynomial spline, $s(x) = s_n^m(x)$, as defined in the Foreword of Wahba (1990), is a real-valued function on $[a, b]$ defined with the aid of n so-called knots $-\infty \leq a < x_1 < x_2 < \cdots < x_n < b \leq \infty$ with the following properties:

(i) $s \in \pi^{m-1},\ x \in [a, x_i],\ x \in [x_n, b],$

(ii) $s \in \pi^{2m-1},\ x \in [x_i, x_{i+1}],\ i = 1, 2, \cdots, n-1,$

(iii) $s \in C^{2m-2},\ x \in (-\infty, \infty),$

where π^k is the polynomials of degree k or less, and C^k is the class of functions with k continuous derivatives. The integer m (≥ 1) is called the order of the spline.

In words, $s(\cdot)$ is a piecewise polynomial in each interval $[a, x_i]$, $[x_i, x_{i+1}]$ $i = 1, 2, \cdots, n-1$, $[x_n, b]$ with the pieces joined at the knots so that $s(\cdot)$ has $2m-2$ continuous derivatives.

It takes m coefficients to define $s(\cdot)$ to the left of x_1, m coefficients to define $s(\cdot)$ to the right of x_n, and $(n-1) \times 2m$ coefficients to define $s(\cdot)$ in the $(n-1)$

interior intervals for a total of $2mn$ coefficients. The continuity conditions (iii) provide $(2m-1)n$ coefficients. It can then be shown that the values of $s(\cdot)$ at the n points $(x_1, \cdots, x_n)$ provide the remaining n coefficients to define $s(\cdot)$ uniquely.

Therefore, suppose $f(\cdot)$ satisfies conditions (i)-(iii), with an additional condition given by

$$\text{(iv) } f(x_i) = f_i,\ i = 1, 2, \cdots, n,$$

we can uniquely determine a nature polynomial spline satifying(i)-(iv).

Natural polynomial splines are closely related to the roughness penalty functionals, which are used to measure the roughness of a curve. One of the most popular such functionals is the quadratic penalty functional given by

$$\int_a^b \{f^{(m)}(x)\}^2 dx, \tag{1.1}$$

where $f^{(m)}(x)$ is the mth derivative of $f(x)$ with respect to x.

First consider the following interpolating problem: find $f(\cdot)$ satisfying condition (iv) that minimizes the penalty functional (1.1). Because $f(\cdot)$ is infinite-dimensional, we restrict our attention to finding such a function in a reproducing kernel Hilbert space (r.k.h.s) of smooth functions, given by

$$W_m = \left\{ f : f, f', \cdots, f^{(m-1)} \text{ are absolutely continuous, } \int_a^b \{f^m(x)\}^2 dx < \infty \right\}.$$

It can be shown that, subject to condition (iv), the minimizer of (1.1) in W_m is an mth order natural polynomial spline satisfying conditions (i)-(iv).

Statisticians are generally interested in smoothing rather than interpolating data. In this dissertation we are especially interested in the application of natural polynomial splines in nonparametric regression and related fields.

In the nonparametric regression context, we have a data model given by

$$(y, x, f),$$

where $y = (y_1, y_2, \cdots, y_n)^T$ is the vector of observed responses, $x = (x_1, x_2, \cdots, x_n)^T$ is the vector of covariates, f is a nonparametric function that relates known functions of y or some other model parameters to x; in addition, distributional assumptions, either parametric or semi-parametric, about the responses are often made to reflect the fact that we have noisy instead of exact data. We end up with an objective function $l(f; y, x)$ which we want to maximize. As a few examples, $l(f; y, x)$ is the log-likelihood function in a linear or generalized linear regression context; it is the negative residual sum of squares in a least squares problem; and it is the partial log-likelihood function in a proportional hazards model fitting problem, etc.

Without any constraint on f, we would always find the perfect fit to the data. A smoothing problem is hence to maximize the following penalized objective function

$$l_p(f; y, x) = l(f; y, x) - \frac{\lambda}{2} \int_a^b \{f^{(m)}(x)\}^2 dx, \tag{1.2}$$

subject to that $f \in W_m$, where $\lambda \geq 0$ is the smoothing parameter controlling the goodness-of-fit to the data, measured by $l(f; y, x)$, and the smoothness of the curve f, measured by $\int_a^b \{f^{(m)}(x)\}^2 dx$. If $\lambda = 0$, then we have an interpolating problem; if

$\lambda = \infty$, then f is forced to be an $(m-1)$th order polynomial.

It is easy to show that the maximizer of (1.2) is a natural polynomial spline of order m, with the knots at $x_1, x_2, \cdots, x_n$. In smoothing problems like this, this maximizer is also called a natural smoothing spline. To show this result, the argument, as detailed in Section 2.3.1 of Green and Silverman (1994), goes as follows: suppose $\widetilde{f}$ is a function in W_m that maximizes (1.2), and $\widetilde{f}(x_i) = \widetilde{f}_i$; then we can always find a unique natural smoothing spline, say, $g(\cdot)$, to interpolate the points $(x_i, \widetilde{f}_i)$ so that $g(x_i) = \widetilde{f}_i$. Therefore $l(g; y, x) = l(\widetilde{f}; y, x)$. But $\int_a^b \{g^{(m)}(x)\}^2 dx \leq \int_a^b \{\widetilde{f}^{(m)}(x)\}^2 dx$ by the property of natural smoothing splines, thus $l_p(g; y, x) \geq l_p(\widetilde{f}; y, x)$. But $\widetilde{f}$ is the maximizer of $l_p(f; y, x)$, so it must be that $\widetilde{f} = g$.

There are many essentially equivalent ways of specifying the solution, a natural smoothing spline, to the variational problem (1.2). We emphasize that a natural smoothing spline can be uniquely determined by its values evaluated at the knots. So what we want essentially is a representation of the vector of such values. One such representation is based on the r.k.h.s theory and is discussed in detail in Kimeldorf and Wahba (1971) and Section 1.3 of Wahba (1990). We discuss and use this representation in Chapter 3.

When $m = 2$, we have the so-called natural cubic splines. Natural cubic splines are probably the most considered splines in the statistical literature. For natural cubic splines, an alternative representation exits. Such a representation has proved convenient in the context of nonparametric Generalized Linear Models. A good reference for this representation is Chapter 2 of Green and Silverman (1994). We discuss and use this representation in Chapter 2.

One consequence of the solution to (1.2) being a natural smoothing spline is that the quadratic penalty functional can be represented as an equivalent quadratic form in some vector, say, a. From a Bayesian perspective, such a quadratic form can be treated as the kernel of a multivariate normal distribution. Consequently, a can be treated as random effects in a mixed model framework, and the inverse of the smoothing parameter can be treated and estimated as a variance component. We will take advantage of this result in both Chapter 2 and Chapter 3.

In this dissertation we consider two problems of interest to us, both related to smoothing splines in a roughness penalty approach. In the first problem we are concerned with the automatic selection of the smoothing parameter when using smoothing splines to fit a nonparametric Generalized Linear Model. This problem is discussed in Chapter 2. In the second problem our concern is testing the proportional hazards assumption and covariate effects in the Cox model, using smoothing splines to estimate the nonparametric functions of time or covariate. This problem is discussed in Chapter 3.

Chapter 2

Smoothing Parameter Selection in Nonparametric Generalized Linear Models via Sixth-order Laplace Approximation

2.1 Introduction

Suppose that we have data of the form

$$(y_i, x_i), \quad i = 1, 2, \cdots, n,$$

where y_i are independent scalar response variables each from a one-parameter exponential family depending on the covariates x_i (possibly vector-valued). The density of y_i has the form

$$p(y_i; \theta_i, \phi) = \exp\left[\{y_i\theta_i - b(\theta_i)\}/a_i(\phi) + c(y_i, \phi)\right],$$

where $a_i(> 0)$, b and c are known functions, with b a strictly convex function of θ_i on any bounded set. Often $a_i(\phi) = \phi m_i^{-1}$ with the prior weights, m_i, known; ϕ is a nuisance or scale parameter that is independent of θ_i; while θ_i are the natural parameters related to the covariates x_i.

By the exponential family theory, the mean (μ_i) and variance (V_i) of y_i are given by

$$\mathrm{E}(y_i) = b^{'}(\theta_i) \triangleq \mu_i,$$

$$\mathrm{var}(y_i) = a_i(\phi) b^{''}(\theta_i) \triangleq V_i.$$

Here $'$ and $''$ denote the first and second order differentiation, respectively.

Generalized Linear Models (GLMs; see, for example, McCullagh and Nelder 1989) provide a unified likelihood framework for modeling such data. In the usual parametric GLMs, the mean μ_i is related to the linear predictor $x_i^T\beta$ via the link function $g(\cdot)$ such that $g(\mu_i) = x_i^T\beta$ with the unknown parameter β to be estimated from the data. Here $g(\cdot)$ is assumed to be monotone and differentiable. If $g(\cdot)$ is chosen such that $g(\mu_i) = \theta_i$, then it is the so-called "canonical" link function. With such a parametric assumption, maximum likelihood theory may be used for estimating parameters and making inference.

However, a parametric model may not always be desirable in practice since the form of the dependence of the response variable on the covariates may not be well known in advance. Nonparametric GLMs hence have been proposed to allow for more flexible functional dependence. Instead of using a linear predictor, these models assume that

$$g(\mu_i) = f(x_i) \tag{2.1}$$

and try to estimate the unknown function $f(\cdot)$ nonparametrically.

To estimate $f(\cdot)$ which is infinite-dimensional, assumptions about the smoothness of $f(\cdot)$ are often made. One popular way is to assume that $f(\cdot)$ is an element of some reproducing kernel Hilbert space of smooth functions, and estimate $f(\cdot)$ by maximizing a penalized log-likelihood. Following O'Sullivan et al. (1986), the penalized log-likelihood function is given by

$$l_p\{f(\cdot); y\} = \sum_{i=1}^{n} l_i\{f(x_i); y_i\} - \frac{\lambda}{2} J(f), \tag{2.2}$$

where $l_i\{f(x_i); y_i\} = y_i f(x_i) - b\{f(x_i)\}$ is the log-likelihood of y_i under model (2.1), with $g(\cdot)$ being the canonical link and $a_i(\phi)$ absorbed into λ. $J(f)$ is a quadratic penalty functional that qualifies the smoothness of $f(\cdot)$, and λ is the smoothing parameter that controls the trade-off between the goodness of fit to the data and the smoothness of $f(\cdot)$.

Consider the case where x_i are one-dimensional. Without loss of generality, we assume that $0 < x_1 < x_2 < \cdots < x_n < 1$. $x_i, i = 1, \cdots, n$, are the so-called "knots". Let $f \in W_2^{(2)} = \{f : f, f'$ are absolutely continuous, $\int_0^1 \{f''(x)\}^2 dx < \infty\}$ and $J(f) = \int_0^1 \{f''(x)\}^2 dx$. Let $f = (f_1, f_2, \cdots, f_n)^T$ be the vector of values of $f(\cdot)$ evaluated at the distinct knots $x_1, x_2, \cdots, x_n$. One can show that the minimizer of the penalized log-likelihood (2.2) is a natural cubic smoothing spline which can be uniquely determined by its values evaluated at the distinct knots, and $J(f) = \int_0^1 \{f''(x)\}^2 dx = f^T K f$, where K is the corresponding non-negative definite smoothing matrix. $K = QR^{-1}Q^T$, and Q, R are $n \times (n-2)$ and $(n-2) \times (n-2)$ band matrices given in Section 2.1.2 of Green and Silverman (1994), as specified below.

Let $h_i = x_{i+1} - x_i$ for $i = 1, \cdots, n-1$. Then Q is the $n \times (n-2)$ matrix with entries q_{ij}, for $i = 1, \cdots, n$ and $j = 2, \cdots, n-1$, given by

$$q_{j-1,j} = h_{j-1}^{-1},\ q_{jj} = -h_{j-1}^{-1} - h_j^{-1}, \text{and } q_{j+1,j} = h_j^{-1}$$

for $j = 2, \cdots, n-1$, and $q_{ij} = 0$ for $|i-j| \geq 2$. Note that the columns of Q are numbered in a non-standard way, starting at $j = 2$, so that the top left element of Q is q_{12}.

The symmetric matrix R is $(n-2) \times (n-2)$ with elements r_{ij}, for i and j running from 2 to $(n-1)$, given by

$$r_{ii} = \frac{1}{3}(h_{i-1} + h_i) \text{ for } i = 2, \cdots, n-1,$$
$$r_{i,i+1} = r_{i+1,i} = \frac{1}{6} h_i \text{ for } i = 2, \cdots, n-2,$$

and $r_{ij} = 0$ for $|i-j| \geq 2$.

The matrix R is strictly diagonal dominant, in the sense that $|r_{ii}| > \sum_{j \neq i} |r_{ij}|$ for each i. R can be shown to be strictly positive definite by using standard arguments in numerical linear algebra. It is hence justified to take the inverse of R in the definition of the matrix K.

Thus (2.2) can be written as

$$l_p\{f(\cdot); y\} = \sum_{i=1}^{n}\{y_i f_i - b(f_i)\} - \frac{\lambda}{2} f^T K f. \tag{2.3}$$

If λ is given, maximizing (2.3) to get the Maximum Penalized Likelihood Estimator (MPLE) of f is a trivial optimization problem. General methods such as Newton-Raphson or Fisher-Scoring can be used in this context. However, λ is usually unknown in advance and often an automatic procedure for determining an appropriate amount of smoothing from the data is needed.

By far in the literature there are two main strategies for the automatic selection of λ in the context of nonparametric smoothing spline regression.

The first strategy has evolved around the theme of Cross-Validation (CV), Generalized Cross-Validation (GCV) and their variations. In such approaches, a score that qualifies the distance between the true function f and the estimated function f_λ given λ is defined. Because the true curve is unknown, a certain type of cross-validated approximation is always involved by leaving out one subject's data at a time. And the λ that minimizes this score is chosen to be the optimal smoothing parameter.

The second strategy is based on the Bayesian formulation of smoothing spline estimators. Under such a formulation estimation of f and the smoothing parameter can be unified under a likelihood framework via an equivalent mixed model representation of the smoothing spline estimators. Under such a representation, the inverse of the smoothing parameter can be treated as an extra variance component in the mixed model. Marginal likelihood method has been proposed to be used for estimating the smoothing parameter, which is equivalent to the Restricted Maximum Likelihood (REML) approach. However, under such a mixed model representation, intractable numerical integration is often involved. Second-order Laplace approximation has been used for approximate inference and proves to be effective in simulation studies (Zhang et al. 1998; Lin and Zhang 1999). However, there are considerable interests for developing more accurate approximations. We propose a new approach in this chapter based on the second strategy but using higher-order Laplace approximation to fulfill such a requirement. We are also interested in comparing the effectiveness of the two strategies for Non-Gaussian data, which has not yet been addressed in the literature.

The rest of this chapter is organized as follows. In Section 2.2 we review GCV, GACV and their variations. GAMM is reviewed in Section 2.3. We derive the proposed new method for the automatic selection of the smoothing parameter using REML via high-order Laplace approximation in Section 2.4. Simulation results of comparing the three approaches above are reported in Section 2.5. And finally, discussion and further work are presented in Section 2.6.

2.2 Generalized Approximate Cross-Validation

The use of CV and GCV to choose the smoothing parameter for Gaussian nonparametric smoothing spline regression was proposed by Wahba and Wold (1975) and Craven and Wahba (1979), respectively. They argued that GCV is often preferred over CV. In the GLM context, O'Sullivan et al. (1986) were the first to adapt GCV to non-Gaussian data. Gu and Xiang (2001) gave a detailed review of recent developments in this setting. They found that there are two basic approaches, which they termed as the direct approach and the indirect approach. The former evaluates a GCV-type score at the convergence of the iteration of finding the estimate of f, while the later as the iteration proceeds.

Some important GCV-type scores include the indirect GCV score of Gu (1990), the direct score of Cox and Chang (1990) and the indirect score of Gu (1992) which is similar to the Unbiased Risk (UBR) estimate in Craven and Wahba (1979) for Gaussian data.

Xiang and Wahba (1996) developed the direct Generalized Approximate Cross-Validation (GACV) score through a series of first-order Taylor expansions. They showed via simulations that GACV is the most effective one among all the direct and indirect scores.

Gu and Xiang (2001) also provided an alternative derivation of the GACV score of Xiang and Wahba (1996), resulting in a score which is essentially equivalent to GACV but can be computationally more convenient. In view of GACV having the best performance among GCV-type approaches, we choose it as a representative to be compared with the two REML-type approaches introduced in the next two sections.

In our setting of Section 2.1, the GACV score can be written as

$$GACV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[-y_i f_\lambda(x_i) + b\{f_\lambda(x_i)\}] + \frac{\mathrm{tr}(V)}{n}\frac{\sum_{i=1}^{n} y_i\{y_i - f_\lambda(x_i)\}}{n - \mathrm{tr}(VW)}, \qquad (2.4)$$

where f_λ is the smoothing spline estimator of f given λ, and V, W are to be given in Section 2.4.2.

We calculate f_λ on a grid of λ values and evaluate the GACV score for each λ. The λ that minimizes the GACV is chosen as the smoothing parameter estimate.

2.3 Generalized Additive Mixed Models

Lin and Zhang (1999) proposed Generalized Additive Mixed Models (GAMMs) for over-dispersed and correlated data. The nonparametric GLM in this chapter is a special case of their models. They explored the Generalized Linear Mixed Model (GLMM) representation of the smoothing spline estimators and estimated the smoothing parameter using REML by treating $\tau = 1/\lambda$ as an extra variance component. Specifically, starting with the penalized log-likelihood (2.3), they reparameterized f in terms of β (2×1) and a $((n-2) \times 1)$ via a one-to-one transformation as

$$f = X\beta + Ba, \qquad (2.5)$$

where $B = L(L^TL)^{-1}$, L is an $n \times (n-2)$ full rank matrix satisfying $K = LL^T$, and X is $n \times 2$ such that $L^TX = 0$.

It is easy to show that $f^T K f = a^T a$. Thus (2.3) can be written as

$$l_p\{f(\cdot); y\} = \sum_{i=1}^{n} l_i(f_i; y_i) - \frac{\lambda}{2} a^T a = \sum_{i=1}^{n} l_i(f_i; y_i) - \frac{1}{2\tau} a^T a, \tag{2.6}$$

where $l_i(f_i; y_i) = y_i f_i - b(f_i), i = 1, 2, \cdots, n$. Equation (2.6) suggests that we can estimate (f, λ) by estimating (β, a, λ) in the following GLMM

$$g(\mu) = X\beta + Ba, \tag{2.7}$$

where $\mu = (\mu_1, \mu_2, \cdots, \mu_n)^T$, the random effects $a \sim N(0, \tau I)$ and τ can be treated as a variance component.

Following Breslow and Clayton (1993), Lin and Zhang (1999) estimated (β, a) by maximizing the Double Penalized Quasi-Likelihood (DPQL), and estimated τ by REML. Note that in our case, the quasi-likelihood is the same as the likelihood, thus the REML of τ is given by

$$\exp\{l_M(\tau; y)\} = \tau^{-\frac{n-2}{2}} \int \exp\{\sum_{i=1}^{n} l_i(\beta, a; y_i) - \frac{1}{2\tau} a^T a\} d\beta da. \tag{2.8}$$

A similar approach was used by Zhang et al. (1998) for Gaussian correlated data. f is estimated by the BLUP estimator, i.e., $\hat{f} = X\hat{\beta} + B\hat{a}$, where $\hat{\beta}$ and $\hat{a}$ are the BLUP estimators of model (2.7). The motivation behind this REML approach will be explained in detail in the next section.

Note that in using DPQL, a second-order Laplace approximation is implicitly adopted to tackle the often intractable numerical integration in equation (2.8). How good this approximation is compared to GCV-type methods and the higher-order

Laplace approximation proposed in the next section will be addressed through simulations in Section 2.5.

2.4 REML via Sixth-order Laplace Approximation

2.4.1 Motivation Behind the REML Approach

For independent Gaussian data, we have the classical nonparametric regression model

$$y_i = f(x_i) + \epsilon_i, \;\; i = 1, 2, \cdots, n. \tag{2.9}$$

Here the ϵ_i are independent and follow $N(0, \sigma^2)$, where $N(\cdot)$ stands for normal distribution hereafter, and $f(\cdot)$ is to be estimated nonparametrically by a smoothing spline in the same spirit as stated in Section 2.1. Under this model, Wahba (1985) and Kohn et al. (1991) proposed estimating the smoothing parameter using Generalized Maximum Likelihood (GML) by assuming $f(x)$ has a partially improper integrated Wiener prior

$$f(x) = \delta_0 + \delta_1 x + \lambda^{-1/2} \int_0^x W(s) ds, \tag{2.10}$$

where δ_0 and δ_1 have improper uniform distributions on $(-\infty, \infty)$ and $W(s)$ is the standard Wiener process. Note that the prior specification in (2.10) is equivalent to assuming f takes the form in (2.5) with $a \sim N(0, \tau I)$ and a flat prior for β, and $B = \Sigma^{1/2}$, where Σ is the covariance matrix of the integrated Wiener process evaluated at X. The smoothing parameter estimate was found by maximizing a marginal likelihood given model (2.9) and the prior (2.10). Speed (1991) and Thompson (1985) pointed out that the GML estimator of τ is identical to the Restricted Maximum Likelihood (REML) estimator of τ under the linear mixed model

$$y = X\beta + Ba + \epsilon,$$

where $a \sim N(0, \tau I)$, $\epsilon \sim N(0, \sigma^2 I)$, and (X, B) can take the form either in this section or in Section 2.3.

Motivated by these results, Zhang et al. (1998) and Lin and Zhang (1999) extended the REML approach for estimating the smoothing parameter to correlated Gaussian and non-Gaussian data, respectively. Following Harville (1977), the REML-type estimator of the smoothing parameter can also be formulated from a Bayesian perspective. Specifically, in the case of nonparametric GLM, the REML of τ is the marginal likelihood obtained by assuming a flat prior for β and a Gaussian prior $N(0, \tau I)$ for a in $l(\beta, a; y) = \sum_{i=1}^{n} l_i(\beta, a; y_i)$ and then integrating out β and a completely from the joint likelihood, resulting in a marginal likelihood of τ as in expression (2.8).

The new approach proposed in this chapter is based on the finite-dimensional Bayesian formulation of smoothing splines given in Section 3.8.4 of Green and Silverman (1994); see also Section 4 of Green (1987). Specifically, f is assumed to have a partially improper Gaussian prior whose log density has kernel $-\lambda f^T K f/2$. That is, the prior is a multivariate normal distribution for f with mean 0 and inverse variance matrix λK. Here impropriety of the prior is equivalent to rank deficiency in the matrix K. One can easily see from equation (2.3) that the natural cubic smoothing spline estimator of f in model (2.1) is the posterior mode of the integrated log-likelihood function $l(f; y) = \sum_{i=1}^{n} l_i(f_i; y_i)$. Hence we consider estimating τ by maximizing the following marginal likelihood

$$\exp\{l_M(\tau; y)\} = \tau^{-\frac{n-2}{2}} \int \exp\{\sum_{i=1}^{n} l_i(f_i; y_i) - \frac{1}{2\tau} f^T K f\} df, \tag{2.11}$$

where $l_i(f_i; y_i)$ is defined in Section 2.3, and K is given in Section 2.1.

Recall from expression (2.5) that there is a one-to-one relationship between f and (β, a). Thus this marginal likelihood, ignoring multiplicative constants, is identical to the REML (2.8). Therefore the suggested approach is also a REML-type method.

2.4.2 Derivation of the New Approach

Consider the marginal likelihood of τ, ignoring multiplicative constants, given by equation (2.11). Let $h(f) = \sum_{i=1}^{n} l_i(f_i; y_i) - (1/2\tau) f^T K f$, then the integration we are interested in is of the form $\int \exp\{h(f)\} df$. Since this integration is often intractable except when y_i follow a normal distribution, we consider approximating it using the sixth-order Laplace approximation approach of Raudenbush et al. (2000) as follows.

For given τ, denote by $\tilde{f} = \tilde{f}(\tau)$ the mode of $h(f)$. Using a Taylor series expansion about $\tilde{f}$, we get

$$\begin{aligned} h(f) &= h(\tilde{f}) + h^{(1)}(\tilde{f})(f - \tilde{f}) + \frac{1}{2}(f - \tilde{f})^T h^{(2)}(\tilde{f})(f - \tilde{f}) \\ &\quad + \sum_{k=3}^{\infty} \frac{1}{k!} \{ \overset{k-1}{\otimes} (f - \tilde{f})^T \} h^{(k)}(\tilde{f})(f - \tilde{f}), \end{aligned} \tag{2.12}$$

where $h^{(k)}(\tilde{f}) = (\partial vec\{h^{(k-1)}(f)\}/\partial f^T)|_{f=\tilde{f}}$ is the kth derivative of $h(f)$ with respect to f, evaluated at $\tilde{f}$; and $\overset{k}{\otimes} u = u \otimes u \otimes \cdots \otimes u$, there being k u's in the Kronecker product.

In Appendix 2.7.1, we show that the derivatives of $h(f)$ with respect to f are given by

$$\begin{aligned} h^{(1)}(f) &= (y-\mu)^T - \frac{1}{\tau} f^T K, \\ h^{(2)}(f) &= -W - \frac{1}{\tau} K, \text{ and} \\ h^{(k)}(f) &= -\sum_{i=1}^{n} (\overset{k-1}{\otimes} z_i) m_i^{(k)} z_i^T, \text{ for } k \geq 3. \end{aligned}$$

Here $y = (y_1, y_2, \cdots, y_n)^T$, $\mu = (\mu_1, \mu_2, \cdots, \mu_n)^T$, $W = diag\{w_1, w_2, \cdots, w_n\}$ with $w_i = \partial \mu_i / \partial f_i$, z_i is a $n \times 1$ column vector with all elements equal to 0 except that the ith is 1, and $m_i^{(k)} = \partial^{(k-1)} \mu_i / \partial f_i^{(k-1)}, k = 1, 2, \cdots$.

Let $y^* = \tilde{W}^{-1}(y - \tilde{\mu}) + \tilde{f}$ be the working vector, where we use a tilde to denote that the matrix, vector or scalar is evaluated at $\tilde{f}$ hereafter. Since $\tilde{f}$ is the mode of $h(f)$, setting $h^{(1)}(\tilde{f}) = (y^* - \tilde{f})^T \tilde{W} - (1/\tau)\tilde{f}^T K = 0$, one obtains that

$$\tilde{f} = (\tilde{W} + \frac{1}{\tau} K)^{-1} \tilde{W} y^*. \tag{2.13}$$

Note that this gives the Fisher-Scoring iterating formula for obtaining the maximum penalized likelihood estimator of f when τ is known; see, p.100, Green and Silverman (1994).

With $h^{(1)}(\tilde{f}) = 0$, the second term on the right hand side of (2.12) disappears. Define $V = \{-h^{(2)}(f)\}^{-1} = (W + \frac{1}{\tau}K)^{-1}$, $\tilde{V} = \{-h^{(2)}(\tilde{f})\}^{-1} = (\tilde{W} + \frac{1}{\tau}K)^{-1}$, and let

$R = \sum_{k=3}^{\infty} T_k$, where $T_k = \frac{1}{k!}\{\overset{k-1}{\otimes}(f - \tilde{f})^T\}h^{(k)}(\tilde{f})(f - \tilde{f}), k = 3, 4, \cdots$. Substituting the expression (2.12) into the marginal likelihood (2.11), we get

$$\exp\{l_M(\tau; y)\} = \tau^{-\frac{n-2}{2}} \exp\{h(\tilde{f})\} \int \exp\{-\frac{1}{2}(f - \tilde{f})^T \tilde{V}^{-1}(f - \tilde{f})\} \cdot \exp(R)\ df$$

Realizing that $\exp\{-\frac{1}{2}(f - \tilde{f})^T \tilde{V}^{-1}(f - \tilde{f})\}$ is the kernel of a multivariate normal distribution with covariance matrix $\tilde{V}$, we can further write the marginal likelihood as, ignoring multiplicative constants,

$$\exp\{l_M(\tau; y)\} = \tau^{-\frac{n-2}{2}} \exp\{h(\tilde{f})\}|\tilde{V}|^{1/2}\mathrm{E}\{\exp(R)\}, \qquad (2.14)$$

where the expectation is taken with respect to the multivariate normal distribution stated above.

Since $\exp(R) = 1 + R + (1/2)R^2 + \cdots$, we have $\mathrm{E}\{\exp(R)\} = 1 + \mathrm{E}(\sum_{k=3}^{\infty} T_k) + (1/2)\mathrm{E}(\sum_{k=3}^{\infty} T_k)^2 + \cdots$. Following Raudenbush et al. (2000), we first realize that $\mathrm{E}(T_k) = 0$ for odd k, $k > 2$ and $\mathrm{E}(T_k T_l) = 0$ for odd $(k+l)$, k and l both > 2. Therefore the full expansion of the terms in $\mathrm{E}\{\exp(R)\}$ involves T_4, T_6, $T_3^2/2$, T_8, $T_3T_5/2$, $T_4^2/2$, $\cdots$. As suggested by Raudenbush et al. (2000), we use the approximation with $\mathrm{E}\{\exp(R)\} \approx 1 + \mathrm{E}(T_4) + \mathrm{E}(T_6) + (1/2)\mathrm{E}(T_3^2)$, which proves to be highly accurate in their simulation studies.

Consequently, the marginal likelihood (2.14) can be approximated as

$$\exp\{l_M(\tau; y)\} \approx \tau^{-\frac{n-2}{2}} \exp\{h(\tilde{f})\}|\tilde{V}|^{1/2}\{1 + \mathrm{E}(T_4) + \mathrm{E}(T_6) + (1/2)\mathrm{E}(T_3^2)\}.$$

Define $\tilde{B}_i = z_i^T \tilde{V} z_i = \tilde{V}_{ii}$, where $\tilde{V}_{ii}$ is the ith diagonal element of $\tilde{V}$, and $\tilde{c} = \sum_{i=1}^{n} \tilde{m}_i^{(3)} z_i \tilde{B}_i = \sum_{i=1}^{n} \tilde{m}_i^{(3)} z_i \tilde{V}_{ii}$. Using Theorem 2 in Raudenbush et al. (2000), one can show that the approximate log marginal likelihood is

$$\begin{aligned} l_M(\tau; y) &= -\frac{n-2}{2}\log\tau - \frac{1}{2}\log|\tilde{W} + \frac{1}{\tau}K| \\ &\quad + \sum_{i=1}^{n}\{y_i\tilde{f}_i - b(\tilde{f}_i)\} - (1/2\tau)\tilde{f}^TK\tilde{f} + \log(\tilde{A}), \end{aligned} \tag{2.15}$$

where $\tilde{A} = 1 - \frac{1}{8}\sum_{i=1}^{n}\tilde{m}_i^{(4)}\tilde{V}_{ii}^2 - \frac{1}{48}\sum_{i=1}^{n}\tilde{m}_i^{(6)}\tilde{V}_{ii}^3 + \frac{15}{72}\tilde{c}^T\tilde{V}\tilde{c}$.

Following Raudenbush et al. (2000), we use an algorithm similar to the approximate Fisher-Scoring of Green (1984) to maximize the approximate log marginal likelihood (2.15), which requires only the first derivative of (2.15) with respect to τ. In getting the score of τ from (2.15), we shall take into consideration that $h(\tilde{f})$ is evaluated at $\tilde{f} = \tilde{f}(\tau) = (\tilde{W} + \frac{1}{\tau}K)^{-1}\tilde{W}y^* = \tilde{V}\tilde{W}y^*$. After solving this interdependence with implicit differentiation of $\tilde{f}$ with respect to τ, we have

$$\begin{aligned} \frac{\partial\tilde{f}}{\partial\tau} &= \left(\frac{\partial\tilde{f}_1}{\partial\tau}, \frac{\partial\tilde{f}_2}{\partial\tau}, \cdots, \frac{\partial\tilde{f}_n}{\partial\tau}\right)^T \\ &= -(\tilde{W} + \frac{1}{\tau}K)^{-1}\left\{\frac{\partial(\tilde{W} + \frac{1}{\tau}K)}{\partial\tau}\right\}(\tilde{W} + \frac{1}{\tau}K)^{-1}\tilde{W}y^* \\ &= \frac{1}{\tau^2}(\tilde{W} + \frac{1}{\tau}K)^{-1}K(\tilde{W} + \frac{1}{\tau}K)^{-1}\tilde{W}y^* \\ &= \frac{1}{\tau^2}\tilde{V}K\tilde{V}\tilde{W}y^*, \end{aligned}$$

where we have used the following result

$$\frac{\partial A^{-1}(\theta)}{\partial\theta_k} = -A^{-1}(\theta)\left(\frac{\partial A(\theta)}{\partial\theta_k}\right)A^{-1}(\theta).$$

After tedious matrix algebra (see Appendix 2.7.2), we obtain the score of τ given by

$$S_\tau = -\frac{n-2}{2}\frac{1}{\tau} + \frac{1}{2\tau^2}\text{trace}(\tilde{V}K) - \frac{1}{2}\sum_{i=1}^{n}\tilde{m}_i^{(3)}(\partial\tilde{f}_i/\partial\tau)\tilde{V}_{ii} + \frac{1}{2\tau^2}\tilde{f}^TK\tilde{f}$$

$$+\frac{1}{\tilde{A}}\left[-\frac{1}{8}\sum_{i=1}^{n}\{\tilde{m}_i^{(5)}\tilde{V}_{ii}^2(\partial\tilde{f}_i/\partial\tau)+2\tilde{m}_i^{(4)}\tilde{V}_{ii}\tilde{a}_{ii}\}\right.$$
$$-\frac{1}{48}\sum_{i=1}^{n}\{\tilde{m}_i^{(7)}\tilde{V}_{ii}^3(\partial\tilde{f}_i/\partial\tau)+3\tilde{m}_i^{(6)}\tilde{V}_{ii}^2\tilde{a}_{ii}\}$$
$$\left.+\frac{15}{72}\{\tilde{c}^T\tilde{a}\tilde{c}+2\tilde{c}^T\tilde{V}(\partial\tilde{c}/\partial\tau)\}\right], \tag{2.16}$$

where $\tilde{a}=\partial\tilde{V}/\partial\tau=-\tilde{V}\left[diag\{\tilde{m}_1^{(3)}\partial\tilde{f}_1/\partial\tau,\tilde{m}_2^{(3)}\partial\tilde{f}_2/\partial\tau,\cdots,\tilde{m}_n^{(3)}\partial\tilde{f}_n/\partial\tau\}-\frac{1}{\tau^2}K\right]\tilde{V}$, and $\tilde{a}_{ii}$ is the ith diagonal element of $\tilde{a}$. $\partial\tilde{c}/\partial\tau=\sum_{i=1}^{n}(\tilde{m}_i^{(4)}\tilde{V}_{ii}\partial\tilde{f}_i/\partial\tau+\tilde{m}_i^{(3)}\tilde{a}_{ii})z_i$.

Different from the situation in Raudenbush et al. (2000), since we do not have repeated measurements here, the approximate Fisher-Scoring algorithm is not completely applicable. Using a superscript (k) to denote estimates from the kth iteration, our algorithm is as follows. At the kth iteration , $k\geq 2$, we first estimate the Hessian using the scores as follows

$$H^{(k)}=\frac{s_\tau^{(k-1)}-s_\tau^{(k-2)}}{\tau^{(k-1)}-\tau^{(k-2)}},$$

while in the first iteration, we simply take $H^{(1)}=1$. We then update τ by

$$\tau^{(k)}=\tau^{(k-1)}-\{H^{(k)}\}^{-1}s_\tau^{(k-1)}. \tag{2.17}$$

We alternate between the iterations of f using (2.13) and those of τ via (2.17) until both of their values stabilize. In updating τ, we use a step-halving approach to make sure that its values are always positive.

Since the sixth-order Laplace approximation is a key component of this new approach, we call it LAP6 as an abbreviation hereafter.

2.5 A Simulation Study

As we stated in Section 2.1, there are two main strategies for the automatic selection of the smoothing parameter. For Gaussian data under model (2.9), Kohn et al. (1991) showed through an extensive simulation study that the marginal likelihood or the REML-type estimate of the smoothing parameter has similar and often better performance compared to the GCV-type estimate in estimating the nonparametric function. No work has been done in the literature to compare the two strategies for non-Gaussian data. We hence carried out a Monte Carlo simulation study to compare the performance of the three approaches introduced in the previous three sections. We considered the setting of cubic smoothing spline logistic regression. Binary data $B\{1, p(x)\}$ and binomial data $B\{8, p(x)\}$, where $B(\cdot)$ stands for the binomial distribution, were generated according to the logistic model

$$\log \frac{p(x)}{1-p(x)} = f(x),$$

where the true curve $f(x)$ is one of the following

$$\begin{aligned}
(a) \quad f_1(x) &= \frac{1}{3}\{2F_{8,8}(x) + F_{5,5}(x)\} - 1, \\
(b) \quad f_2(x) &= \frac{1}{10}\{6F_{30,17}(x) + 4F_{3,11}(x)\} - 1, \\
(c) \quad f_3(x) &= 3\{10^5 x^{11}(1-x)^6 + 10^3 x^3 (1-x)^{10}\} - 2, \\
(d) \quad f_4(x) &= 2\sin(10x), \\
(e) \quad f_5(x) &= (-1.6x + 0.9)\mathrm{I}_{[x \le 0.5]} + (1.6x - 0.7)\mathrm{I}_{[x > 0.5]}, \\
(f) \quad f_6(x) &= 2\sin(2\pi x), \qquad (2.18)
\end{aligned}$$

where $F_{p,q}(x) = \Gamma(p+q)x^{p-1}(1-x)^{q-1}/\{\Gamma(p)\Gamma(q)\}$ and $\Gamma(\cdot)$ is the gamma function. The first two curves were used by Lin and Zhang (1999) in their simulation study, and we used the same knots as theirs, namely $x_i = (i-1)/100, i = 1, \cdots, 100$. The rest curves were chosen from Gu and Xiang (2001) with the same knots $x_i = (i-.5)/100, i = 1, \cdots, 100$. Plots of the curves are given in Figure 2.1. A total of 500 data sets were generated for each curve.

We chose the Kullback-Leibler loss function to measure the performance of $f_\lambda(x)$ as an estimate of $f(x)$. The same loss was used by Xiang and Wahba (1996) and Gu and Xiang (2001) and is given by

$$L(f, f_\lambda) = \frac{1}{n}\sum_{i=1}^{n}\{f(x_i) - f_\lambda(x_i)\}\{\mu(x_i) - \mu_\lambda(x_i)\}.$$

SAS macro was developed for our new approach and is available upon request. The LAP6 estimates were found by using this macro. Note that for binomial distribution with denominator n_i, some important quantities needed for evaluating the score of τ in LAP6 are given by

$$\begin{aligned}
&\mu_i = n_i \exp(f_i)/\{1+\exp(f_i)\},\ w_i = m_i^{(2)} = \mu_i(1-\mu_i/n_i),\\
&m_i^{(3)} = w_i(1-2\mu_i/n_i),\ m_i^{(4)} = w_i(1-6w_i/n_i),\\
&m_i^{(5)} = m_i^{(3)}(1-12w_i/n_i),\ m_i^{(6)} = m_i^{(4)}(1-12w_i/n_i) - 12{m_i^{(3)}}^2/n_i,\\
&m_i^{(7)} = m_i^{(5)}(1-12w_i/n_i) - 36m_i^{(3)}m_i^{(4)}/n_i,\ i = 1, 2, \cdots, n.
\end{aligned}$$

For a derivation of these equations, see Appendix 2.7.3.

The GAMM estimates of the smoothing parameters were found by using the

GAMM SAS macro developed by Lin and Zhang (1999). For the GACV approach, since the GACV score (2.4) can not be maximized directly, we first found the estimate f_λ for f on a grid of λ values. Following Xiang and Wahba (1996), we took $\log_{10}\lambda$ to be equally spaced on the interval $[-6, 0]$ with step 0.08. We then calculated the GACV score for each f_λ and the λ giving the smallest GACV score was identified as the GACV estimate. Since the true f is known in simulation, we also evaluated $L(f, f_\lambda)$ for each f_λ on the same grid of λ and identified the one that gave the smallest loss as the optimal estimate of the smoothing parameter. For the estimate of λ obtained from each approach, we also recorded the corresponding $L(f, f_\lambda)$. To measure the effectiveness of the three approaches, we calculated the ratio of the minimal loss to the loss achieved by the respective approaches. Note that this ratio is always less than one and the closer the ratio to one, the better the corresponding approach.

Figure 2.2 gives the boxplots of 500 loss ratios of GACV, GAMM, LAP6 for estimating the six curves in (2.18) where data were generated from binomial distribution with denominator 8. From these plots we see that GAMM and LAP6 generally outperform GACV for the binomial case in that they often have higher medians and shorter tails.

Figure 2.3 gives the same boxplots of loss ratios as in Figure 2.2 except that the data were generated from binary distribution. Since binary data are sparse, the normal theory based Laplace approximations may not perform as well. And we see in several cases GACV performs better than GAMM and LAP6.

There generally was not much difference between GAMM and LAP6, suggesting that the second-order Laplace approximation will suffice in most situations. The advantage of higher-order Laplace approximation over the second-order GAMM is more prominent for sparse data, as can be seen from Figure 2.3.

We note that unlike in Gaussian case, where the marginal likelihood of $\tau = 1/\lambda$ is exact, approximations are unavoidable in Non-Gaussian case to obtain the marginal likelihood. The performance of the REML-type approaches will in a large part depend on how good the approximations are. One nice feature of the proposed approach is that the Taylor expansion is allowed to go as far as we wish, although the computation involved will increase accordingly. Nonetheless, the sixth-order expansion used in our simulation generally performs very well, especially when the data are not too sparse.

2.6 Discussion and Further Work

In this chapter we have developed a new approach for the automatic selection of the smoothing parameter in nonparametric smoothing spline GLMs, based on the REML approach and the sixth-order Laplace approximation. Through simulations, the proposed method is shown to be effective compared to GACV and GAMM. Furthermore, the approach can be extended to the degree of accuracy required by using even higher-order expansions in the Taylor series.

Compared to GACV, one advantage of REML-type approaches is that they do not need a subjective choice of the grid of values for λ. An iteration process exists for obtaining the maximum marginal likelihood estimate of the smoothing parameter. The computing time of both GAMM and LAP6 is comparable to that of GACV using a moderately dense grid, like the one used in our simulation study.

Another advantage of REML-type approaches is that they can be naturally incorporated in modeling correlated data, for example, Lin and Zhang (1999) use marginal quasi-likelihood method to choose the smoothing parameter in GAMM, and Zhang et al. (1998) use REML to estimate the smoothing parameter in modeling Gaussian longitudinal data. The GCV score has not yet been well defined for correlated data to date. More work needs to be done on developing and comparing smoothing parameter selection methods for correlated Gaussian and Non-Gaussian data.

For sparse data, REML-type approaches do not perform as well as GACV in some situations, suggesting that caution must be taken regarding using GAMM and LAP6 in this scenario.

In this chapter we only consider the case where the covariates are one-dimensional, more research is needed on developing REML-type approaches of choosing the smoothing parameter for data with multi-dimensional covariates.

2.7 Appendices

2.7.1 Derivation of the Partial Derivatives of $h(f)$ with respect to f

We have

$$h(f) = l(f;y) - \frac{1}{2\tau} f^T K f,$$

where

$$l(f;y) = \sum_{i=1}^{n} l_i(f_i;y_i) = \sum_{i=1}^{n} \{y_i f_i - b(f_i)\}.$$

Note

$$\frac{\partial l_i(f_i;y_i)}{\partial f_i} = y_i - b^{'}(f_i) = y_i - \mu_i,$$

thus

$$\frac{\partial l(f;y)}{\partial f^T} = (y-\mu)^T.$$

It is easy to see that

$$\frac{\partial(-\frac{1}{2\tau} f^T K f)}{\partial f^T} = -\frac{1}{\tau} f^T K,$$

therefore

$$h^{(1)}(f) = \frac{\partial h(f)}{\partial f^T} = (y-\mu)^T - \frac{1}{\tau} f^T K. \tag{2.19}$$

Note that by using the working vector $y^* = \tilde{W}^{-1}(y - \tilde{\mu}) + \tilde{f}$ defined in Section 2.4.2, we can write

$$h^{(1)}(\tilde{f}) = (y^* - \tilde{f})^T \tilde{W} - \frac{1}{\tau} \tilde{f}^T K.$$

Now since

$$\frac{\partial vec\{(y-\mu)^T\}}{\partial f^T} = -diag\left\{\frac{\partial \mu_1}{\partial f_1}, \cdots, \frac{\partial \mu_n}{\partial f_n}\right\} = -W,$$

and

$$\frac{\partial vec(-\frac{1}{\tau} f^T K)}{\partial f^T} = -\frac{1}{\tau} K,$$

we have

$$h^{(2)}(f) = \frac{\partial vec\{h^{(1)}(f)\}}{\partial f^T} = -W - \frac{1}{\tau} K. \qquad (2.20)$$

For $k \geq 3$, because $h(f) = l(f; y) - (1/2\tau) f^T K f$, we have $h^{(k)}(f) = l^{(k)}(f; y)$. First consider the situation where $k = 3$. Let z_i be a $n \times 1$ vector with all elements equal to 0 except that the ith is 1, $i = 1, 2 \cdots, n$. Then it is easy to see that

$$W = diag\{w_1, w_2, \cdots, w_n\} = \sum_{i=1}^{n} z_i w_i z_i^T.$$

Therefore

$$\begin{aligned} h^{(3)}(f) &= \frac{\partial vec\{h^{(2)}(f)\}}{\partial f^T} \\ &= \frac{\partial vec\{-\sum_{i=1}^{n} z_i w_i z_i^T - \frac{1}{\tau} K\}}{\partial f^T} \\ &= -\sum_{i=1}^{n} \frac{\partial vec\{z_i w_i z_i^T\}}{\partial f^T} \\ &= -\sum_{i=1}^{n} vec(z_i z_i^T) \frac{\partial w_i}{\partial f^T}, \end{aligned}$$

where note that the last equation is because w_i is a scalar.

Using the result $vec(ab^T) = b \otimes a$ for any two compatible column vectors a and b, we get

$$vec(z_i z_i^T) = z_i \otimes z_i = \overset{2}{\otimes} z_i.$$

Furthermore, because $w_i = \partial \mu_i / \partial f_i$ depends only on f_i, we have

$$\frac{\partial w_i}{\partial f^T} = \frac{\partial w_i}{\partial f_i} z_i^T = \frac{\partial^2 \mu_i}{\partial f_i^2} z_i^T = m_i^{(3)} z_i^T,$$

where $m_i^{(k)} = \partial^{(k-1)} \mu_i / \partial f_i^{(k-1)}, k \geq 3$. Therefore

$$h^{(3)}(f) = -\sum_{i=1}^{n} (\overset{2}{\otimes} z_i) m_i^{(3)} z_i^T.$$

Similarly,

$$\begin{aligned} h^{(4)}(f) &= \frac{\partial vec\{h^{(3)}(f)\}}{\partial f^T} \\ &= -\sum_{i=1}^{n} vec(\overset{2}{\otimes} z_i \cdot z_i^T) \frac{\partial m_i^{(3)}}{\partial f^T} \\ &= -\sum_{i=1}^{n} (\overset{3}{\otimes} z_i) m_i^{(4)} z_i^T. \end{aligned}$$

In general, for $k \geq 3$, we have

$$h^{(k)}(f) = -\sum_{i=1}^{n} (\overset{k-1}{\otimes} z_i) m_i^{(k)} z_i^T. \tag{2.21}$$

Equations (2.19), (2.20) and (2.21) give all partial derivatives of $h(f)$ with respect to f.

2.7.2 Derivation of the Score of τ, S_τ

In this appendix, we derive the score of τ from the approximate log marginal likelihood (2.15). First we have

$$\frac{\partial}{\partial \tau} \left(-\frac{n-2}{2} \log \tau \right) = -\frac{n-2}{2} \frac{1}{\tau}. \tag{2.22}$$

Now

$$\frac{\partial}{\partial\tau}\left(-\frac{1}{2}\log|\tilde{W}+\frac{1}{\tau}K|\right)=-\frac{1}{2}\text{trace}\left[(\tilde{W}+\frac{1}{\tau}K)^{-1}\frac{\partial(\tilde{W}+\frac{1}{\tau}K)}{\partial\tau}\right].$$

Note that $\tilde{V}=(\tilde{W}+\frac{1}{\tau}K)^{-1}$ as defined in Section 2.4.2. Because $\tilde{w}_i=\partial\tilde{\mu}_i/\partial\tilde{f}_i$, we have

$$\frac{\partial\tilde{w}_i}{\partial\tau}=\frac{\partial\tilde{w}_i}{\partial\tilde{f}_i}\cdot\frac{\partial\tilde{f}_i}{\partial\tau}=\frac{\partial^2\tilde{\mu}_i}{\partial\tilde{f}_i^2}\cdot\frac{\partial\tilde{f}_i}{\partial\tau}=\tilde{m}_i^{(3)}\frac{\partial\tilde{f}_i}{\partial\tau}.$$

Note $\tilde{W}=diag\{\tilde{w}_1,\cdots,\tilde{w}_n\}$, thus

$$\frac{\partial\tilde{W}}{\partial\tau}=diag\{\tilde{m}_1^{(3)}\frac{\partial\tilde{f}_1}{\partial\tau},\cdots,\tilde{m}_n^{(3)}\frac{\partial\tilde{f}_n}{\partial\tau}\}.$$

We also have

$$\frac{\partial(\frac{1}{\tau}K)}{\partial\tau}=-\frac{1}{\tau^2}K.$$

Therefore

$$\begin{aligned}&\frac{\partial}{\partial\tau}\left(-\frac{1}{2}\log|\tilde{W}+\frac{1}{\tau}K|\right)\\ =\;&-\frac{1}{2}\text{trace}\left[\tilde{V}\left\{diag\{\tilde{m}_1^{(3)}\frac{\partial\tilde{f}_1}{\partial\tau},\cdots,\tilde{m}_n^{(3)}\frac{\partial\tilde{f}_n}{\partial\tau}\}-\frac{1}{\tau^2}K\right\}\right]\\ =\;&\frac{1}{2\tau^2}\text{trace}(\tilde{V}K)-\frac{1}{2}\sum_{i=1}^n\tilde{m}_i^{(3)}(\partial\tilde{f}_i/\partial\tau)\tilde{V}_{ii},\end{aligned}\tag{2.23}$$

where $\tilde{V}_{ii}$ is the ith diagonal element of $\tilde{V}$.

Note that $h(\tilde{f})=\sum_{i=1}^n\{y_i\tilde{f}_i-b(\tilde{f}_i)\}-(1/2\tau)\tilde{f}^TK\tilde{f}$, therefore

$$\begin{aligned}&\frac{\partial\left[\sum_{i=1}^n\{y_i\tilde{f}_i-b(\tilde{f}_i)\}-(1/2\tau)\tilde{f}^TK\tilde{f}\right]}{\partial\tau}\\ =\;&\frac{\partial h(\tilde{f})}{\partial\tilde{f}^T}\frac{\partial\tilde{f}}{\partial\tau}+\frac{\partial\{-(1/2\tau)\tilde{f}^TK\tilde{f}\}}{\partial\tau}\\ =\;&0\cdot\frac{\partial\tilde{f}}{\partial\tau}+\frac{1}{2\tau^2}\tilde{f}^TK\tilde{f}\\ =\;&\frac{1}{2\tau^2}\tilde{f}^TK\tilde{f}.\end{aligned}\tag{2.24}$$

Because $\tilde{A}$ is a scalar, it is easy to see that

$$\frac{\partial \log(\tilde{A})}{\partial \tau} = \frac{1}{\tilde{A}}\frac{\partial \tilde{A}}{\partial \tau}.$$

We also have

$$\frac{\partial \tilde{m}_i^{(4)}}{\partial \tau} = \frac{\partial \tilde{m}_i^{(4)}}{\partial \tilde{f}_i}\frac{\partial \tilde{f}_i}{\partial \tau} = \tilde{m}_i^{(5)}\frac{\partial \tilde{f}_i}{\partial \tau},$$

and

$$\frac{\partial \tilde{m}_i^{(6)}}{\partial \tau} = \frac{\partial \tilde{m}_i^{(6)}}{\partial \tilde{f}_i}\frac{\partial \tilde{f}_i}{\partial \tau} = \tilde{m}_i^{(7)}\frac{\partial \tilde{f}_i}{\partial \tau}.$$

Recall that $z_i^T \tilde{V} z_i = \tilde{V}_{ii}$ is the ith diagonal element of $\tilde{V}$, therefore $\partial \tilde{V}_{ii}/\partial \tau$ is the ith diagonal element of $\partial \tilde{V}/\partial \tau$. And

$$\begin{aligned}\frac{\partial \tilde{V}}{\partial \tau} &= \frac{\partial (\tilde{W} + \frac{1}{\tau}K)^{-1}}{\partial \tau} \\ &= -\tilde{V}\left[diag\{\tilde{m}_1^{(3)}\partial \tilde{f}_1/\partial \tau, \cdots, \tilde{m}_n^{(3)}\partial \tilde{f}_n/\partial \tau\} - \frac{1}{\tau^2}K\right]\tilde{V}.\end{aligned}$$

Define $\tilde{a} = \partial \tilde{V}/\partial \tau$, then $\partial \tilde{V}_{ii}/\partial \tau = \tilde{a}_{ii}$.

For $\tilde{c}^T \tilde{V} \tilde{c}$, we have

$$\frac{\partial (\tilde{c}^T \tilde{V} \tilde{c})}{\partial \tau} = \tilde{c}^T \frac{\partial \tilde{V}}{\partial \tau}\tilde{c} + 2\tilde{c}^T \tilde{V}\frac{\partial \tilde{c}}{\partial \tau},$$

where $\tilde{c} = \sum_{i=1}^n \tilde{m}_i^{(3)} z_i \tilde{V}_{ii}$. Therefore,

$$\frac{\partial \tilde{c}}{\partial \tau} = \sum_{i=1}^n \left\{\tilde{m}_i^{(4)}\frac{\partial \tilde{f}_i}{\partial \tau} z_i \tilde{V}_{ii} + \tilde{m}_i^{(3)} z_i \frac{\partial \tilde{V}_{ii}}{\partial \tau}\right\}.$$

We have showed earlier that $\partial \tilde{V}_{ii}/\partial \tau = \tilde{a}_{ii}$, therefore

$$\frac{\partial \tilde{c}}{\partial \tau} = \sum_{i=1}^n \left\{\tilde{m}_i^{(4)}\tilde{V}_{ii}\frac{\partial \tilde{f}_i}{\partial \tau} + \tilde{m}_i^{(3)}\tilde{a}_{ii}\right\} z_i.$$

Note $\partial\tilde{V}/\partial\tau = \tilde{a}$, thus

$$\frac{\partial(\tilde{c}^T\tilde{V}\tilde{c})}{\partial\tau} = \tilde{c}^T\tilde{a}\tilde{c} + 2\tilde{c}^T\tilde{V}\frac{\partial\tilde{c}}{\partial\tau}.$$

We thus have

$$\begin{aligned}\frac{\partial\log(\tilde{A})}{\partial\tau} &= \frac{1}{\tilde{A}}\left[-\frac{1}{8}\sum_{i=1}^{n}\{\tilde{m}_i^{(5)}\tilde{V}_{ii}^2(\partial\tilde{f}_i/\partial\tau) + 2\tilde{m}_i^{(4)}\tilde{V}_{ii}\tilde{a}_{ii}\}\right. \\ &\quad -\frac{1}{48}\sum_{i=1}^{n}\{\tilde{m}_i^{(7)}\tilde{V}_{ii}^3(\partial\tilde{f}_i/\partial\tau) + 3\tilde{m}_i^{(6)}\tilde{V}_{ii}^2\tilde{a}_{ii}\} \\ &\quad \left.+\frac{15}{72}\{\tilde{c}^T\tilde{a}\tilde{c} + 2\tilde{c}^T\tilde{V}(\partial\tilde{c}/\partial\tau)\}\right]. \end{aligned} \tag{2.25}$$

Combining equations (2.22), (2.23), (2.24) and (2.25), we have the formula for the score of τ, S_τ, given in equation (2.16).

2.7.3 Derivation of $m_i^{(k)}$ for the Binomial Distribution

For the binomial distribution under the nonparametric GLM (2.1) with the canonical link, we have

$$\mu_i = n_i\frac{\exp(f_i)}{1+\exp(f_i)}.$$

$$\begin{aligned} w_i &= \frac{\partial\mu_i}{\partial f_i} = n_i\frac{\exp(f_i)}{[1+\exp(f_i)]^2} \\ &= \frac{\mu_i(n_i-\mu_i)}{n_i}. \end{aligned}$$

$$\begin{aligned} m_i^{(3)} &= \frac{\partial w_i}{\partial f_i} = \frac{w_i(n_i-\mu_i)+\mu_i(-w_i)}{n_i} \\ &= \frac{w_i(n_i-2\mu_i)}{n_i}. \end{aligned}$$

$$\begin{aligned}
m_i^{(4)} &= \frac{\partial m_i^{(3)}}{\partial f_i} = \frac{m_i^{(3)}(n_i - 2\mu_i) + w_i(-2w_i)}{n_i} \\
&= \frac{1}{n_i} w_i \left[\frac{(n_i - 2\mu_i)^2}{n_i} - 2w_i \right] \\
&= \frac{1}{n_i} w_i \left[n_i - \frac{4\mu_i(n_i - \mu_i)}{n_i} - 2w_i \right] \\
&= \frac{1}{n_i} w_i (n_i - 4w_i - 2w_i) \\
&= \frac{w_i(n_i - 6w_i)}{n_i}.
\end{aligned}$$

$$\begin{aligned}
m_i^{(5)} &= \frac{\partial m_i^{(4)}}{\partial f_i} = \frac{m_i^{(3)}(n_i - 6w_i) + w_i(-6m_i^{(3)})}{n_i} \\
&= \frac{m_i^{(3)}(n_i - 12w_i)}{n_i}.
\end{aligned}$$

$$\begin{aligned}
m_i^{(6)} &= \frac{\partial m_i^{(5)}}{\partial f_i} = \frac{m_i^{(4)}(n_i - 12w_i) + m_i^{(3)}(-12m_i^{(3)})}{n_i} \\
&= \frac{m_i^{(4)}(n_i - 12w_i) - 12{m_i^{(3)}}^2}{n_i}.
\end{aligned}$$

$$\begin{aligned}
m_i^{(7)} &= \frac{\partial m_i^{(6)}}{\partial f_i} = \frac{m_i^{(5)}(n_i - 12w_i) + m_i^{(4)}(-12m_i^{(3)}) - 24m_i^{(3)}m_i^{(4)}}{n_i} \\
&= \frac{m_i^{(5)}(n_i - 12w_i) - 36m_i^{(3)}m_i^{(4)}}{n_i}.
\end{aligned}$$

For the binary distribution, it is just a special case of the binomial distribution with $n_i = 1$.

2.8 Figures

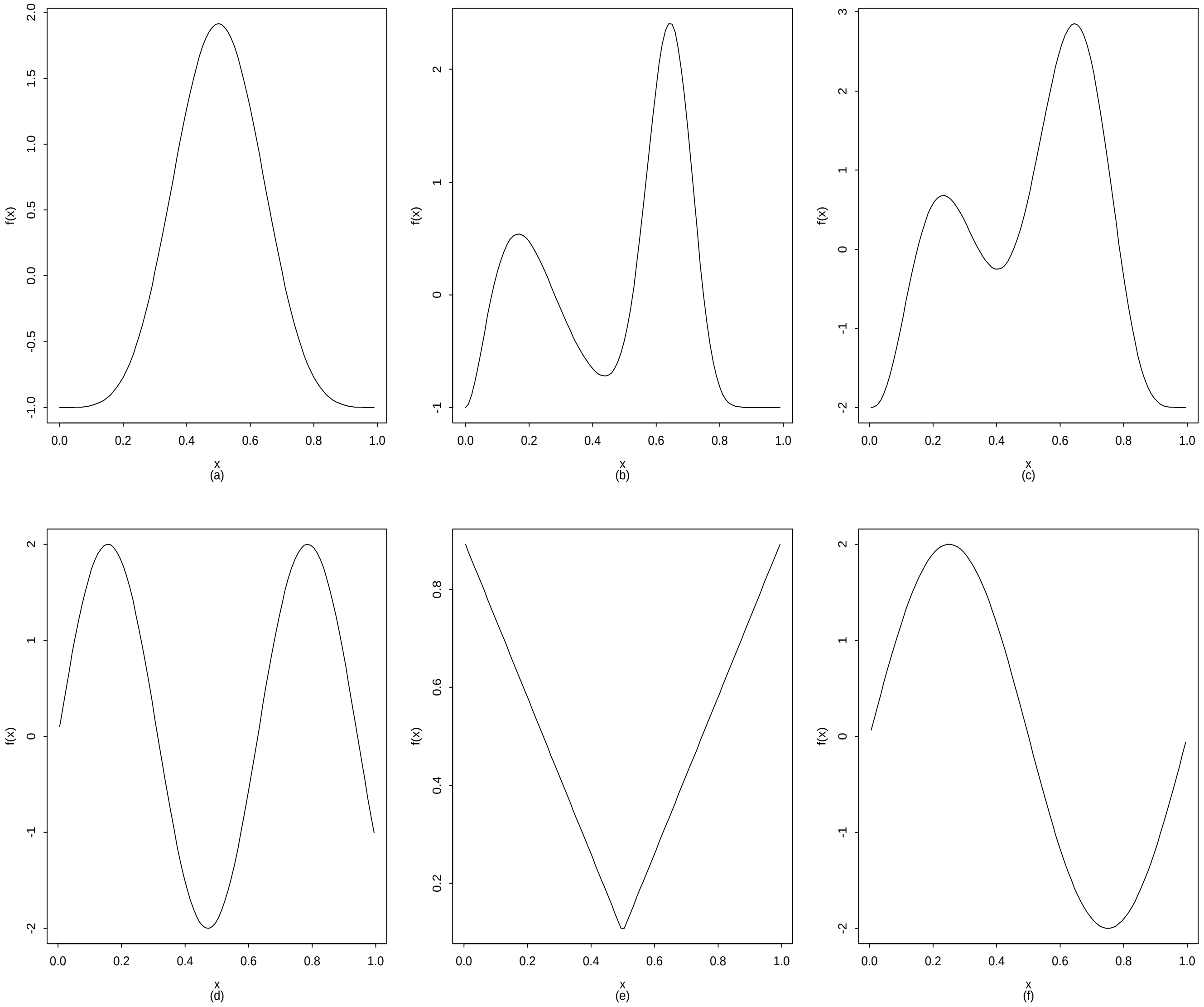

Figure 2.1: Plots of curves used in the simulation study: $(a) f_1(x) = \frac{1}{3}\{2F_{8,8}(x) + F_{5,5}(x)\} - 1$; $(b) f_2(x) = \frac{1}{10}\{6F_{30,17}(x) + 4F_{3,11}(x)\} - 1$; $(c) f_3(x) = 3\{10^5 x^{11}(1-x)^6 + 10^3 x^3 (1-x)^{10}\} - 2$; $(d) f_4(x) = 2\sin(10x)$; $(e) f_5(x) = (-1.6x + 0.9)\mathrm{I}_{[x \leq 0.5]} + (1.6x - 0.7)\mathrm{I}_{[x > 0.5]}$; $(f) f_6(x) = 2\sin(2\pi x)$.

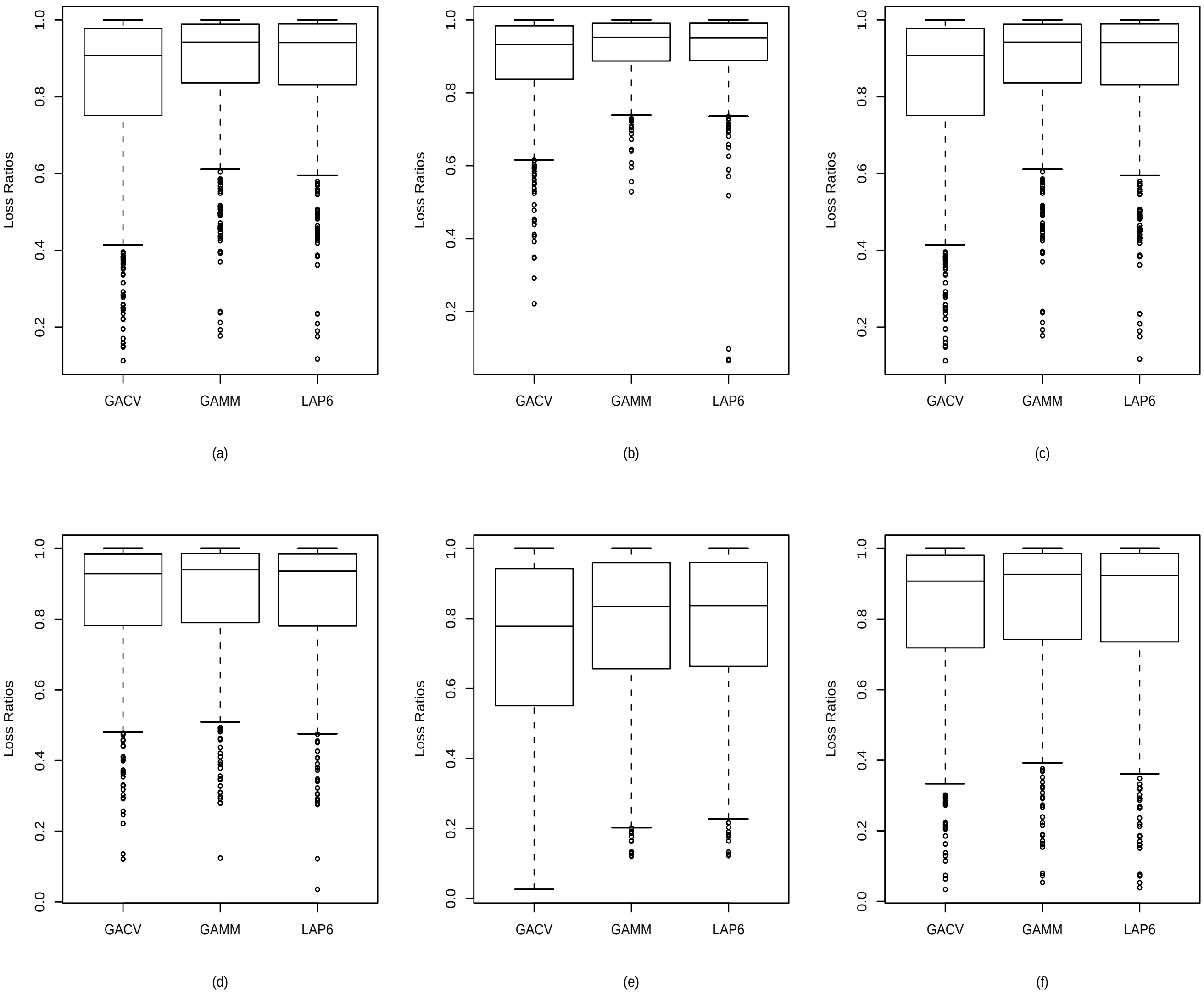

Figure 2.2: Loss ratios of GACV, GAMM and LAP6 for estimating curves in (2.18) with data generated from binomial 8 distribution.

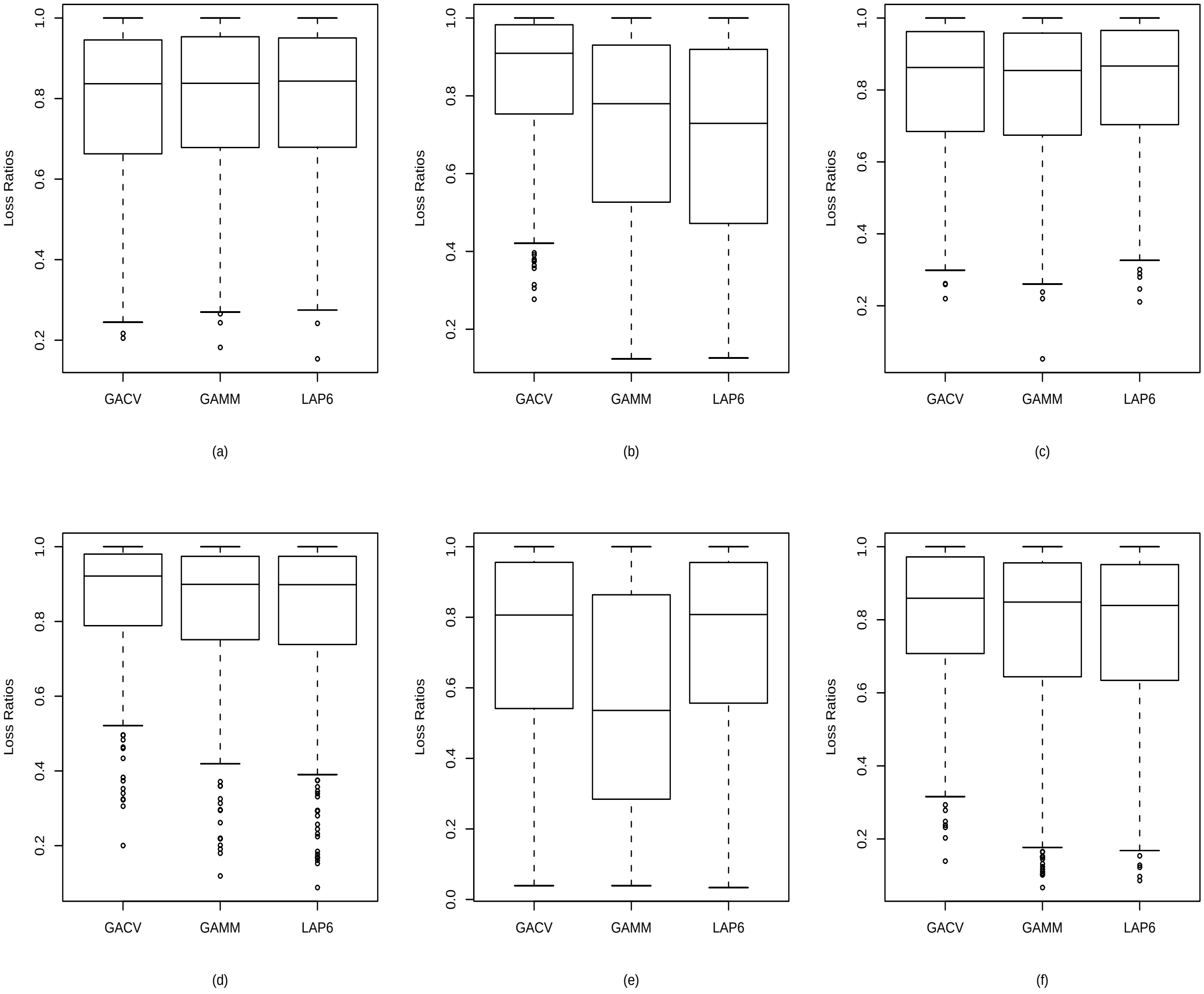

Figure 2.3: Loss ratios of GACV, GAMM and LAP6 for estimating curves in (2.18) with data generated from binary distribution.

Chapter 3

Smoothing Spline-based Score Tests for Proportional Hazards Models

3.1 Introduction

Censored survival data arise routinely in biomedical applications. For the regression analysis of such data, Cox's proportional hazards model (Cox, 1972) is unquestionably the most popular platform. The assumption of proportional hazards may not always be realistic, however; e.g., Gray (2000) notes that effects of prognostic factors in cancer often do not exhibit proportional hazards, and we have found the assumption questionable in a number of cancer and cardiovascular disease data analyses.

Accordingly, good data-analytic practice dictates that the assumption be critically evaluated and alternative models considered if necessary.

A situation in which the proportional hazards assumption may be suspect is in the analysis of covariate effects on survival in Cancer and Leukemia Group B (CALGB) Protocol 8541. CALGB 8541 was a randomized clinical trial comparing three doses (high, moderate, and low) of chemotherapy (cyclophosphamide, doxorubicin, also known as adriamycin, and 5 fluorouracil, abbreviated CAF) in women with early stage, node-positive breast cancer. The primary analysis found no difference in survival between high and moderate doses, both of which were superior to the low dose. Based on long-term follow-up, subsequent interest focused on whether certain patient characteristics are prognostic for survival. Figure 3.1 shows estimated survival curves and the log-negative-log of survival curves for the 1437 patients for whom Estrogen Receptor (ER) status was available; the plot shows these for the 520 ER-negative and 917 ER-positive women, respectively. If the proportional hazards assumption were valid, the two log-negative-log of survival curves should be parallel. This is obviously not the case; in fact, the two curves cross on the interval $(0, 1)$ year. Figure 3.2 shows a plot of the Schoenfeld (1982) residuals. If proportional hazards were adequate, then, on average, the residuals should be zero. The noticeable trend away from zero further calls into question the relevance of the proportional hazards assumption. Formal evidence in support of the visual impression in the figures would be valuable to the data analyst assessing whether the Cox model is an appropriate framework for inference.

Many approaches have been advocated for assessing the relevance of the assumptions; e.g., Fleming and Harrington (1991, sec. 4.5), Klein and Moeschberger (1997, secs. 9.2 and 11.4), and Therneau and Grambsch (2000, Chap. 6) discuss procedures such as including a function of time [e.g., $\log(t)$] as a time-dependent covariate in the linear predictor, plots of and smoothing of Schoenfeld (1982) residuals (e.g., based on assumed time-dependent coefficient models), partitioning the time axis into disjoint intervals in each of which the model is fitted and the results compared, and various other techniques. There is furthermore a large literature on formal approaches to testing (e.g., Pettitt and Bin Daud, 1990; Gray, 1994). O'Sullivan (1988), Hastie and Tibshirani (1990), Zucker and Karr (1990) and authors referenced therein discuss estimation in the proportional hazards model with nonparametric covariate or time-varying coefficient effects using smoothing splines in a penalized partial likelihood approach. Gray (1992, 1994) proposes spline-based tests for covariate and time effects using fixed knot splines. Numerical results suggest that the tests perform well in moderate samples. However, the testing procedure requires the smoothing parameter to be finely tuned according to the true alternative to achieve good power properties, which may not be realistic in practice.

Recently, Zhang and Lin (2003) proposed a penalized likelihood approach to deriving a score test for nonparametric covariate effects in generalized additive mixed effects models, based on regarding the inverse of the smoothing parameter as a variance component. This yields a test with low degrees of freedom that, moreover, does not require fitting of the model under the alternative, which can be computationally intensive. Zhang and Lin (2003) show that the test enjoys valid size and high power properties in practical settings. The success of this approach suggests that it may be fruitful in other problems. We hence propose adapting this strategy to testing departures from proportional hazards.

Another problem of interest is testing for covariate effects in Cox models; specifically, testing whether the appropriate functional form that represents the effect of a covariate on survival time is a fixed degree polynomial. We show that this can also be addressed by adapting the strategy in Zhang and Lin (2003).

In Section 3.2, we discuss the proposed score tests for proportional hazards in detail. The proposed score tests for covariate effects are given in Section 3.3. We report empirical results for the tests of proportional hazards and covariate effects in Section 3.4. The methods are applied to the data from CALGB 8541 in Section 3.5.

3.2 Score Test for Proportional Hazards

For the ith of n subjects, let T_i and C_i be survival and censoring times; X_i a p-dimensional vector of covariates; and S_i a scalar covariate of interest, where T_i and C_i are independent given $(X_i^T, S_i)^T$. The observed data are $V_i = \min(T_i, C_i)$, $\Delta_i = I(T_i \leq C_i)$. Cox's proportional hazards model (Cox, 1972) is given by

$$\lambda(t|X_i, S_i) = \lambda_0(t) \exp\{X_i^T \beta + S_i \theta\}, \tag{3.1}$$

where β $(p \times 1)$ and θ (scalar) are regression coefficients, $\lambda(t|X_i, S_i)$ is the hazard function given $(X_i^T, S_i)^T$, and $\lambda_0(t)$ is the unspecified baseline hazard. Note that model (3.1) implies for any X that $\lambda(t|X, S_k)/\lambda(t|X, S_l) = \exp\{(S_k - S_l)\theta\}$ independent of time t, the so-called "proportional hazards" assumption. As suggested by Cox (1972), evaluation of this assumption may be addressed by including in the model a time-dependent covariate that is the product of the covariate of interest and a function of time and then testing if the coefficient of this covariate is different from 0. Rather than considering a known such function, which limits the scope of possible departures from model (3.1), we consider the alternative

$$\lambda(t|X_i, S_i) = \lambda_0(t) \exp\{X_i^T \beta + S_i \gamma(t)\}, \tag{3.2}$$

where now $\gamma(\cdot)$ is an arbitrary smooth function of time. Because $\gamma(\cdot)$ is infinite-dimensional, we follow Gray (1994) and consider estimating it along with β by maximizing the penalized partial log-likelihood

$$l_p\{\beta, \gamma(\cdot), \eta\} = l_c\{\beta, \gamma(\cdot)\} - (\eta/2) \int \{\gamma^{(m)}(t)\}^2 dt, \tag{3.3}$$

where $l_c\{\beta, \gamma(\cdot)\}$ is the usual Cox partial log-likelihood, $m \geq 1$ is an integer, and $\eta > 0$ is a smoothing parameter controlling the roughness of $\gamma(t)$ and the goodness-of-fit of the model.

Following Zhang and Lin (2003), we consider the smoothing spline representation of $\gamma(t)$ of Kimeldorf and Wahba (1971). Denote by $t^0 = (t_1^0, \cdots, t_r^0)^T$ an $(r \times 1)$ vector of ordered, distinct V_i's with $\Delta_i = 1$ (i.e., all failure times) and by γ the corresponding vector of $\gamma(t)$ evaluated at each element of t^0. Without loss of generality, assume $0 < t_1^0 < \cdots < t_r^0 < 1$. As $l_c\{\beta, \gamma(\cdot)\}$ depends on $\gamma(\cdot)$ only through γ, it is well-known that maximizing $l_p\{\beta, \gamma(\cdot), \eta\}$ leads to a natural smoothing spline of order m for the estimator of $\gamma(t)$, expressed as

$$\gamma(t) = \sum_{k=1}^{m} \delta_k \phi_k(t) + \sum_{l=1}^{r} a_l R(t, t_l^0), \tag{3.4}$$

where $\{\delta_k\}$ and $\{a_l\}$ are constants; $\{\phi_k(t)\}_{k=1}^m$ is a basis for the space of $(m-1)$th order polynomials; and $R(t,s) = \int_0^1 (t-u)_+^{m-1}(s-u)_+^{m-1}/\{(m-1)!\}^2$, where $x_+ = x$ if $x > 0$ and 0 otherwise. Writing $\delta = (\delta_1, \cdots, \delta_m)^T$ and $a = (a_1, \cdots, a_r)^T$, we have $\int \{\gamma^{(m)}(t)\}^2 dt = a^T \Sigma a$ and $\gamma = H\delta + \Sigma a$, where H $(r \times m)$ has (k,l) element $\phi_l(t_k^0)$, and Σ is positive definite with (k,l) element $R(t_k^0, t_l^0)$. This quadratic representation of the penalty term in (3.3) suggests that a can be viewed as a random vector with $a \sim N(0, \tau\Sigma^{-1})$, where $\tau = 1/\eta$ is a variance component. Then (3.3) may be represented as $l_p(\beta, \delta, \tau, a) = l_c\{\beta, \gamma(\delta, a)\} - a^T \Sigma a/(2\tau)$, where the Cox partial log-likelihood is now given by

$$\begin{aligned} l_c\{\beta, \gamma(\delta, a)\} &= \sum_{i=1}^{n} \Delta_i \left[X_i^T \beta + S_i c_i^T (H\delta + \Sigma a) \right. \\ &\quad \left. - \log \left\{ \sum_{j \in \mathcal{R}(t_i^0)} \exp\{X_j^T \beta + S_j c_i^T (H\delta + \Sigma a)\} \right\} \right]. \end{aligned} \tag{3.5}$$

Here, $\mathcal{R}(t)$ is the risk set at time t; and c_i is an $(r \times 1)$ vector of all 0's except when $\Delta_i = 1$, when it has a 1 in the position corresponding to the failure time t_i^0 for subject i. Thus, viewing (3.5) as a "conditional (on a) log-likelihood", we may write a "marginal likelihood" for $(\beta^T, \delta^T, \tau)^T$ as

$$L(\beta, \delta, \tau) = \int \exp\left[l_c\{\beta, \gamma(\delta, a)\}\right] \varphi_r(a; 0, \tau\Sigma^{-1}) da, \tag{3.6}$$

where φ_r represents the density of an r-dimensional normal distribution.

The natural spline representation of $\gamma(t)$ in (3.4) implies that $\gamma(t)$ is an $(m-1)$th order polynomial if and only if $a = 0$, which in (3.6) is equivalent to $H_0 : \tau = 0$. Thus, testing whether $\gamma(t)$ is a constant as in (3.1) versus the broad alternative (3.2) may be addressed by setting $m = 1$ and testing H_0. Following Zhang and Lin (2003), we propose a "score-type" test for H_0 as follows. Writing $l(\beta, \delta, \tau) = \log\{L(\beta, \delta, \tau)\}$, making the transformation $u = \tau^{-1/2}\Sigma^{1/2}a$ in (3.6), and using L'Hôpital's rule, some algebra shows that the "score" of τ based on (3.6) takes the form

$$\begin{aligned} \left.\frac{\partial l(\beta, \delta, \tau)}{\partial \tau}\right|_{\widehat{\beta}, \widehat{\delta}, \tau=0} &= \frac{1}{2}\left\{ \frac{\partial l_c\{\beta, \gamma(\delta, 0)\}}{\partial \gamma^T} \Sigma \frac{\partial l_c\{\beta, \gamma(\delta, 0)\}}{\partial \gamma} \right. \\ &\quad \left.\left. + \mathrm{tr}\left(\frac{\partial^2 l_c\{\beta, \gamma(\delta, 0)\}}{\partial \gamma \partial \gamma^T} \Sigma \right) \right\}\right|_{\widehat{\beta}, \widehat{\delta}}, \end{aligned} \tag{3.7}$$

where $\widehat{\beta}, \widehat{\delta}$ are the usual maximum partial likelihood estimators for β, δ found by maximizing (3.5) under $H_0 : a = 0$. For a derivation of this equation, see Appendix 3.7.1.

The asymptotic distribution of the "score" of τ under the null hypothesis is of interest in deriving the score test statistic. It is shown heuristically in Appendix 3.7.2 that, under H_0, the second term on the right hand side of (3.7) converges in probability to a constant. Denote by $U_\tau\{\widehat{\beta}, \gamma(\widehat{\delta}, 0)\}$ the first term, and note that it is a quadratic form in $S_\gamma\{\widehat{\beta}, \gamma(\widehat{\delta}, 0)\} = \partial l_c\{\widehat{\beta}, \gamma(\widehat{\delta}, 0)\}/\partial\gamma$. In Appendix 3.7.2 we provide a heuristic argument that, under H_0, $S_\gamma\{\widehat{\beta}, \gamma(\widehat{\delta}, 0)\}$ is asymptotically normal with mean 0 and variance of the form of $nWVW^T$, which can be consistently estimated from the data. Because of the special structure of the matrix Σ, Zhang and Lin (2003) argue that the standardized version of U_τ may not have a normal distribution. Instead, asymptotically, U_τ follows a weighted chi-square distribution given by $\sum_{i=1}^r \psi_i \chi_{1i}^2$, where χ_{1i}^2 are independent random variables following a chi-square distribution with one degree-of-freedom, and ψ_i are all distinct eigenvalues of the matrix $nWVW^T\Sigma/2$. Because calculation of the ψ_i is often computationally intensive and that of the exact probability associated with a weighted chi-square distribution is difficult, following Zhang and Lin (2003), we approximate the distribution of U_τ by a scaled chi-square $k\chi_v^2$ using the Satterthwaite method. The mean and variance of U_τ are given by $e = \text{tr}(nWVW^T\Sigma)/2$ and $I_{\tau\tau} = \text{tr}\{(nWVW^T\Sigma)^2\}/2$, respectively. Matching these with the mean and variance of $k\chi_v^2$, we obtain $k = I_{\tau\tau}/2e$, $v = 2e^2/I_{\tau\tau}$. The test statistic is $S_\tau = U_\tau/k$, and we reject H_0 at nominal level α if $S_\tau > \chi_{v,1-\alpha}^2$, where $\chi_{v,1-\alpha}^2$ is the $100(1-\alpha)\%$ percentile of the χ_v^2 distribution.

3.3 Score Test for Covariate Effects

We use the same setup as in Section 3.2 but consider the different general alternative

$$\lambda(t|X_i, S_i) = \lambda_0(t)\exp\{X_i^T\beta + \gamma(S_i)\}.$$

Here the unknown function $\gamma(\cdot)$ represents the effect of covariate S_i on the outcome. We are interested in testing the functional form of $\gamma(\cdot)$; specifically, our null hypothesis is $H_0: \ \gamma(\cdot)$ is an $(m-1)$th order polynomial. Two cases of special interest are that of $m = 1$, corresponding to a test for no effect, and $m = 2$, the situation of a linear effect of S_i.

Using the same smoothing spline technique employed in Section 3.2, we estimate $\gamma(\cdot)$ along with β by maximizing the penalized partial log-likelihood

$$l_p\{\beta, \gamma(\cdot), \eta\} = l_c\{\beta, \gamma(\cdot)\} - (\eta/2) \int \{\gamma^{(m)}(s)\}^2 ds. \tag{3.8}$$

Denote by $s^0 = (s_1^0, \cdots, s_r^0)^T$ an $(r \times 1)$ vector of ordered, distinct S_i's and by γ the corresponding vector of $\gamma(s)$ evaluated at each element of s^0. Again assume $0 < s_1^0 < \cdots < s_r^0 < 1$, then maximizing $l_p\{\beta, \gamma(\cdot), \eta\}$ leads to a natural smoothing spline of order m for the estimator of $\gamma(s)$. We again have $\int \{\gamma^{(m)}(s)\}^2 ds = a^T \Sigma a$ and $\gamma = H\delta + \Sigma a$, where H $(r \times m)$ has (k, l) element $\phi_l(s_k^0)$, and Σ is positive definite with (k, l) element $R(s_k^0, s_l^0)$. Equation (3.8) can be represented as $l_p\{\beta, \delta, \tau, a\} = l_c\{\beta, \gamma(\delta, a)\} - a^T \Sigma a/(2\tau)$, where the Cox partial log-likelihood now has a different form given by

$$\begin{aligned} l_c\{\beta, \gamma(\delta, a)\} \quad = \quad & \sum_{i=1}^n \Delta_i \left[X_i^T \beta + c_i^T (H\delta + \Sigma a) \right. \\ & \left. - \log \left\{ \sum_{j \in \mathcal{R}(V_i)} \exp\{X_j^T \beta + c_j^T (H\delta + \Sigma a)\} \right\} \right]. \end{aligned}$$

Here c_i is an $(r \times 1)$ vector of all 0's with the exception of a 1 in the position corresponding to the covariate value s_i^0 for subject i.

We proceed by treating a as a normal random vector, obtaining the "marginal likelihood," deriving the score of τ, and defining the test statistic in the same fashion as in Section 3.2. For reasons of identifiability, the first component of the vector δ must be absorbed into the baseline hazard so that only the rest of the components need be estimated under H_0.

For $m > 1$, all results in Section 3.2 apply here, with the only difference being that the form of l_c is different. In addition, a more special case is testing for no effect of S_i. The null model is $\lambda(t|X_i, S_i) = \lambda_0(t)\exp(X_i^T\beta)$. We take $m = 1$, and because δ has only one component it is absorbed into $\lambda_0(t)$, which is equivalent to $\delta = 0$. The null hypothesis is $H_0 : a = 0$ so that we only need to estimate β. The score of τ under H_0 takes the same form as in (3.7) except now the expression is evaluated at $(\widehat{\beta}, 0, 0)$. The second term can again be shown to converge in probability to a constant. The first term is given by $U_\tau\{\widehat{\beta}, \gamma(0,0)\} = S_\gamma^T\{\widehat{\beta}, \gamma(0,0)\}\Sigma S_\gamma\{\widehat{\beta}, \gamma(0,0)\}/2$, where $S_\gamma\{\widehat{\beta}, \gamma(0,0)\} = \partial l_c\{\widehat{\beta}, \gamma(0,0)\}/\partial\gamma$. Following the same procedure and using the notation in Appendix 3.7.2, we can show that $n^{-1/2}S_\gamma\{\widehat{\beta}, \gamma(0,0)\} \xrightarrow{d} N(0, WVW^T)$, where now $W = (-V_{\gamma\beta}V_{\beta\beta}^{-1}\ \mathcal{I}_r)$. Some algebra yields that $WVW^T = V_{\gamma\gamma} - V_{\gamma\beta}V_{\beta\beta}^{-1}V_{\beta\gamma}$ so $n\widehat{W}\widehat{V}\widehat{W}^T = \widehat{I}_{\gamma\gamma} - \widehat{I}_{\gamma\beta}\widehat{I}_{\beta\beta}^{-1}\widehat{I}_{\beta\gamma}$, which is exactly the efficient (observed) information matrix $\widehat{I}_{\gamma\gamma|\beta}$. If S_i is the only covariate in the Cox model, then $W = (0_{r\times p}\ \mathcal{I}_r)$, and $WVW^T = V_{\gamma\gamma}$ so $n\widehat{W}\widehat{V}\widehat{W}^T = \widehat{I}_{\gamma\gamma}$, and the test statistic is again defined in the same way.

3.4 Simulation Evidence

3.4.1 Test for Proportional Hazards

We carried out a simulation to evaluate the performance of the proposed test for the proportional hazards assumption. The cases we considered are similar to those in Gray (1994).

To evaluate size of the test, failure times were generated under the null model $\lambda(t|S_i) = \lambda_0(t)\exp\{S_i\delta_0\}$, $i = 1, 2, \cdots, n$, with $\lambda_0(t) = 1$ and $\delta_0 = 0$, 1 or 2. Values of S_i were equally spaced on the interval $(0, 1)$ with an equal number of subjects having each distinct S_i value; e.g., if "number of distinct covariate values" is 2, then half subjects had $S_i = 0$, while the other half had $S_i = 1$. We considered two different censoring distributions: a unit exponential distribution and a uniform distribution on $(0, 2)$; the former gave a minimum censoring probability of 0.119 and a maximum of 0.500, while the latter gave a minimum of 0.068 and a maximum of 0.432. Sample sizes were $n = 100$ and 200, and $N = 2000$ samples were generated for each scenario. Empirical size was estimated in each case as the proportion of 2000 samples rejected by the nominal 0.05-level score test.

Results are given in Table 3.1 and show that the empirical sizes of our test are very close to the nominal level for all scenarios we considered. In most cases the empirical sizes are within sampling error of the nominal level. Larger differences from the nominal level are mostly seen when the censoring distribution was unit exponential, as censoring probability in that case is higher.

To evaluate power, failure times were generated under the alternative $\lambda(t|S_i) = \lambda_0(t)\exp\{S_i\gamma(t)\}, i = 1, 2, \cdots, n$. Here, S_i was a single binary covariate defining two groups of equal size, and the true log hazard ratios for the two groups, $\gamma(t)$, were given by

$$
\begin{aligned}
&\text{Curve 1}: \quad \gamma(t) = \log\{.75t\} \\
&\text{Curve 2}: \quad \gamma(t) = \log\{2/(1+5t)\} \\
&\text{Curve 3}: \quad \gamma(t) = \log\{e^t\} = t \\
&\text{Curve 4}: \quad \gamma(t) = \log\{(t-.75)^2\} \\
&\text{Curve 5}: \quad \gamma(t) = \log\{e^{I(t\geq 1)}\} = I(t \geq 1),
\end{aligned}
$$

where $I(\cdot)$ is the indicator function. Curves 1, 2 and 4 were considered by Gray (1994) with the same setup of generating failure and censoring times. Plots of these curves are given in Figure 3.3. Again $\lambda_0(t) = 1$; thus, failure times in the baseline group ($S_i = 0$) were unit exponential. Failure times in the other group ($S_i = 1$) were generated by using the appropriate transformation to obtain the required hazard ratio. Censoring was uniform on (0, 2), which gave a censoring probability in the baseline group of 0.432. We took $n = 200$, and $N = 1000$ samples were generated for each scenario. Empirical power was estimated as the proportion of 1000 samples rejected by the nominal 0.05-level score test.

For comparison, we also computed powers for several 1-degree-of-freedom score tests as follows. Under the model $\lambda(t|S_i) = \lambda_0(t)\exp\{\beta_0 S_i + \beta_1 S_i g(t)\}$, the “linear”, “quadratic”, “log” and “optimal” tests are the score tests of $H_0 : \beta_1 = 0$ from this model with $g(t) = t$, t^2, $\log(t)$ and $\gamma(t)$, respectively. Because the “optimal” test is based on the true alternative $\gamma(\cdot)$, it provides an upper bound on the power of the other tests.

Results from the power simulation are given in Table 3.2. For smooth monotone alternatives (curves 1, 2 and 3), the power of our test is very close to that of the “optimal” test. These alternatives are either linear or close to linear, hence the “linear” test also provides good power for detecting them. For non-monotone (curve 4) or non-smooth (curve 5) alternatives, the power of the proposed test is not as good as that of the “optimal” test. However, for curve 4 our test out-performs all other tests, while for curve 5 has power close to those of the “linear” and the “quadratic” and much higher than that of the “log” test. The proposed test is based on the penalized partial likelihood, thus considers broader alternatives than any specific parametric tests. The penalty function penalizes non-smooth alternatives more than smooth ones, hence the power of the proposed test is focused toward smoother alternatives. We see the proposed test gives some power for non-monotone or non-smooth alternatives, while providing good power for very smooth alternatives. So in a “robust” sense the proposed test can provide good protection against a wide variety of alternatives.

Gray (1992, 1994) proposes spline-based tests in Cox models using fixed knots splines. We now compare our results to those in Section 4 of Gray (1994) under the same simulation setup. Both tests have empirical sizes close to the nominal level. However, in contrast to our test, Gray's test yields several empirical sizes that are far from nominal. For smooth monotone alternatives, the power of our test is comparable to that of Gray. For non-monotone or non-smooth alternatives, Gray's test can have better power, provided that an optimal degree-of-freedom (df) is used. However, this optimal df often needs to be tuned based on the unknown true alternative, which is unrealistic in practice, while our test requires no such tuning.

3.4.2 Test for Covariate Effects

A simulation was also carried out to evaluate performance of the proposed score test for covariate effects. We considered testing both for no covariate effect and for a linear effect.

For size, failure times were generated under the null model $\lambda(t|S_i) = \lambda_0(t)$ (no covariate effect) and $\lambda(t|S_i) = \lambda_0(t)\exp\{S_i\}$ (linear effect), $i = 1, 2, \cdots, n$. In both cases, the S_i values were the same as those used in the size simulation in Section 3.4.1, and $\lambda_0(t) = 1$. Censoring was unit exponential and uniform on $(0, 1.5)$; thus for the former the censoring probability was 0.500 for testing no effect and between 0.269 and 0.500 for testing the linear effect, while for the latter the censoring probability was 0.518 for testing no effect and between 0.241 and 0.518 for testing the linear effect. Sample sizes were $n = 100$ and 200, and $N = 2000$ samples were generated for each scenario.

Table 3.3 shows the size simulation results. The sizes of the proposed test are again very close to the nominal 0.05-level for testing both no and linear effect. In fact, with $n = 200$, all sizes are within the binomial standard error (0.49%) of the nominal level.

For the power simulation, we used the same setup as in the simulation study of Gray (1994). Failure times were generated under the alternative $\lambda(t|S_i) = \lambda_0(t)\exp\{\gamma(S_i)\}, i = 1, 2, \cdots, n$, where $n = 200$, and we were interested in testing $H_0 : \gamma(\cdot) = 0$ and $H_0 : \gamma(\cdot)$ is a linear function, respectively. The following six curves for $\gamma(\cdot)$ were used for both cases:

Curve 1 – exponential (E) :	$\gamma(s) = .25\exp\{.8s\}$
Curve 2 – logistic (L):	$\gamma(s) = .6\exp\{3.5s\}/(1 + \exp\{3.5s\})$
Curve 3 – step 1 (S1):	$\gamma(s) = .9I(s > 1.1)$
Curve 4 – quadratic (Q):	$\gamma(s) = .3s^2$
Curve 5 – cosine (C):	$\gamma(s) = .5\cos(3.5s)$
Curve 6 – step 2 (S2):	$\gamma(s) = .7I(\|s\| < .5)$

Plots of these curves are given in Figure 3.4. The covariate S_i values were equally spaced on the interval $[-1.719, 1.719]$ with step 0.0173 (hence standardized to have mean 0 and variance 1). Censoring times were generated from a uniform distribution on $(0, 1.5)$. $N = 1000$ simulation runs were performed for each scenario.

For testing no effect, we also calculated empirical powers of the usual 1-, 2-, 3-

degree-of-freedom score tests based on adding linear, quadratic, and cubic terms to the null model. For example, the cubic test is the score test of $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ in the model $\lambda(t|S_i) = \lambda_0(t) \exp\{\beta_1 S_i + \beta_2 S_i^2 + \beta_3 S_i^3\}$. Similarly, for testing a linear effect, empirical powers of the usual 1-, 2-degree-of-freedom score tests based on adding quadratic and cubic terms to the null model were provided. For example, the cubic test is the score test of $H_0 : \beta_2 = \beta_3 = 0$ in the model $\lambda(t|S_i) = \lambda_0(t) \exp\{\beta_1 S_i + \beta_2 S_i^2 + \beta_3 S_i^3\}$.

In both cases, the optimal test is the 1-degree-of-freedom score test for the true alternative, thus providing an upper bound on the power of the other tests. For testing no effect, this is the score test of $H_0 : \beta = 0$ in the model $\lambda(t|S_i) = \lambda_0(t) \exp\{\beta\gamma(S_i)\}$; For testing a linear effect, this is the score test of $H_0 : \beta_2 = 0$ in the model $\lambda(t|S_i) = \lambda_0(t) \exp\{\beta_1 S_i + \beta_2 \gamma(S_i)\}$, where $\gamma(\cdot)$ is the true curve used to generate the data.

Power simulation results are given in Table 3.4. For testing no effect, under smooth monotone alternatives (E, L) the proposed test provides good power that is close to that of the optimal test. Results are similar for the linear test because these alternatives are close to linear. For the 2-step alternative (S1), our test is better than the linear and is close to the quadratic and the cubic. For the other three alternatives, which are non-monotone (Q, C) and non-smooth (S2), our test provides some power and is better than the linear but not as good as the other tests. Note no test except the optimal has good power for alternative (C) because of the special shape of the curve. For testing linear effect, alternatives (E, L) are close to linear so none of the tests have good power for detecting them. Our test has better power than the quadratic and the cubic for the other four alternatives except for alternative (Q) for which the quadratic is the optimal test. Even in that case the proposed test has power very close to the that of the optimal. The spline test generally has better power for testing linear effect than for testing no effect, because higher order ($m = 2$) smoothing splines are used for testing linear effect, in contrast to that $m = 1$ for testing no effect. Therefore we have better approximation to the nonparametric function when testing linear effect, consequently increasing the power of the test. Again, because the proposed test is based on the penalized partial likelihood, power of the proposed test is focused toward smoother alternatives. Overall, for testing covariate effects, the proposed test provides good protection against very general alternatives.

The comparison of our results to those in Section 3 of Gray (1994) shows a similar pattern as discussed in the last paragraph of Section 3.4.1, so the comments there apply here as well.

3.5 Application to CALGB 8541

We apply the proposed score tests to the data from CALGB 8541. Data on 1479 eligible patients were available to us after long-term follow-up.

As discussed in Section 3.1, the proportional hazards assumption for evaluation of whether the binary covariate ER status is prognostic for survival time is suspect. Among the 1437 patients who had known ER status, 917 were censored, resulting in a censoring percentage of 63.8%. A proportional hazards fit of time-to-death on ER gives an estimated hazard ratio of 0.768 with a p-value of 0.003.

Results from application of the proposed formal testing procedures confirm the observations in Figures 3.1 and 3.2. For testing the proportional hazards assumption on ER, the score test yields a p-value < 0.001. The “linear”, “quadratic” and “log” test also give p-values significant at nominal level 0.05. Based on these results, modification of the model is thus required to achieve a valid analysis. As the hazard ratio appears to be fairly constant within the interval $[1, 8)$, we may fit a piecewise constant hazard ratio model with three pieces: $[0, 1)$, $[1, 8)$, and $[8, \infty)$. Such a fit gives a significant (level 0.05) p-value for non-proportional hazards on ER ($p = 0.003$). At nominal level 0.05, the effect of ER is significant on the interval $[0, 1)$ (hazard ratio $= 0.263$; $p = 0.004$) and $[1, 8)$ (hazard ratio $= 0.747$; $p = 0.003$) but not significant on the interval $[8, \infty)$ (hazard ratio $= 1.589$; $p = 0.137$), which is another indication that the hazards are not proportional.

Another covariate of interest is menopausal status (pre- or post-menopausal), abbreviated MENO. All 1479 patients had known MENO. Among them, 947 were censored, resulting in a censoring percentage of 64.0%. A proportional hazards fit of time-to-death on MENO gives an estimated hazard ratio of 0.921 with a p-value of 0.347, which is not significant at level 0.05. Figures 3.5 and 3.6 show the survival and log-negative-log of survival curves by MENO and the Schoenfeld residuals of MENO for 638 pre-menopausal and 841 post-menopausal patients, respectively. We see a similar pattern as that described in Section 3.1 on ER, hence the proportional hazards assumption on MENO is also suspect. For testing the proportional hazards assumption on MENO, the score test yields a p-value of 0.011; while the “linear”, “quadratic” and “log” test have a p-value of 0.032, 0.023, and 0.175, respectively. Had we used the “log” test, we would have not rejected the null hypothesis at level 0.05. To get a better understanding of the effect of MENO, we again consider a piecewise constant hazard ratio model. The hazard ratio shows a dramatic change on the interval $[2, 3.5)$ but otherwise appears to be fairly constant, hence we consider such a model with three pieces: $[0, 2)$, $[2, 3.5)$, and $[3.5, \infty)$. Such a fit gives a significant (level 0.05) p-value for non-proportional hazards on MENO ($p = 0.002$). At level 0.05, the effect of MENO is not significant on the interval $[0, 2)$ (hazard ratio = 0.975; $p = 0.905$) and $[3.5, \infty)$ (hazard ratio = 1.148; $p = 0.240$) but significant on the interval $[2, 3.5)$ (hazard ratio = 0.549; $p = 0.001$). This model gives more insight into how MENO influences the outcome than does an overall proportional hazards model.

Other covariates available to us include treatment, size of breast cancer tumor (cm), number of histologically positive lymph nodes found. As noted in Section 3.1, the difference in survival between the two groups treated with a moderate or high dose was not significant at level 0.05 using the log-rank test ($p = 0.814$). We hence grouped these two doses as one treatment, so along with the low dose, we have a binary treatment covariate. After controlling for other covariates, a smoothing spline-based score test of proportional hazards of ER gives a significant (level 0.05) p-value of

0.012. Again we can fit a piecewise constant proportional hazards model on ER and assuming proportional hazards on other covariates. The flexibility of our approach allows other tests to be performed. For example, a score test of the null hypothesis that the effect of "number of positive lymph nodes" is linear gives a p-value of 0.457, which is not significant at level 0.05, suggesting a linear fit is adequate.

3.6 Discussion

We have developed score tests for the proportional hazards assumption and for covariate effects in Cox models, based on the penalized partial likelihood and natural smoothing spline representation. The tests achieve size close to nominal and provide good power for very general alternatives. The tests perform especially well for smooth monotone alternatives.

One advantage of the proposed tests is their simplicity. The test statistic is easy to calculate as we only need to fit the null model, which may be accomplished by maximizing the usual partial likelihood under the null hypothesis. Existing software such as SAS PROC PHREG or S-PLUS/R function *coxph()* can be used directly for this purpose.

We used the Satterthwaite method to approximate the null sampling distribution of the score statistic. If better precision is desired, methods are available for calculating the quantiles from a weighted chi-square distribution; e.g., see, Davies (1980).

If the proportional hazards assumption is rejected, we can include in the proportional hazards predictor interactions between functions of time and covariates to get a more suitable model. The difficulty with this approach is to identify the form of the interaction. Plotting and smoothing the Schoenfeld residuals may provide some insight. An alternative strategy is to use a stratified proportional hazards model. An advantage of this approach is that we do not have to assume a particular form of the interaction. A disadvantage is the resulting inability to examine the effects of the stratifying covariates.

3.7 Appendices

3.7.1 Derivation of Equation (3.7)

From equation (3.6), We have

$$\begin{aligned} L(\beta,\delta,\tau) &= \int \exp\left[l_c\{\beta,\gamma(\delta,a)\}\right]\varphi_r(a;0,\tau\Sigma^{-1})da \\ &= \int \exp\left[l_c\{\beta,\gamma(\delta,a)\}\right](2\pi)^{-r/2}|\tau\Sigma^{-1}|^{-1/2}\exp\{-\frac{1}{2}a^T(\tau\Sigma^{-1})^{-1}a\}da \\ &\propto \tau^{-r/2}\int \exp\left[l_c\{\beta,\gamma(\delta,a)\}\right]\exp\{-\frac{1}{2\tau}a^T\Sigma a\}da. \qquad (3.9) \end{aligned}$$

Let $u = \tau^{-1/2}\Sigma^{1/2}a$, then $a = \tau^{1/2}\Sigma^{-1/2}u$, and

$$-\frac{1}{2\tau}a^T\Sigma a = -\frac{1}{2\tau}(\tau^{1/2}u^T\Sigma^{-1/2})\Sigma(\Sigma^{-1/2}u\tau^{1/2}) = -\frac{1}{2}u^Tu.$$

Now

$$\gamma(\delta,a) \;=\; H\delta+\Sigma a$$

$$\begin{aligned} &= H\delta + \Sigma(\tau^{1/2}\Sigma^{-1/2}u) \\ &= H\delta + \tau^{1/2}\Sigma^{1/2}u \\ &\triangleq \gamma(\delta, \tau, u), \end{aligned}$$

thus

$$\frac{\partial \gamma}{\partial \tau} = \frac{1}{2}\tau^{-1/2}\Sigma^{1/2}u.$$

Note that the Jacob is given by

$$\left|\frac{\partial a}{\partial u^T}\right| = \left|\tau^{1/2}\Sigma^{-1/2}\right| = \tau^{r/2}|\Sigma|^{-1/2},$$

therefore

$$\begin{aligned} (3.9) \quad &= \tau^{-r/2}\int \exp\left[l_c\{\beta, \gamma(\delta, \tau, u)\}\right]\exp\{-\frac{1}{2}u^T u\}\tau^{r/2}|\Sigma|^{-1/2}du \\ &\propto \int \exp\left[l_c\{\beta, \gamma(\delta, \tau, u)\}\right]\exp\{-\frac{1}{2}u^T u\}du. \end{aligned}$$

Let $l(\beta, \delta, \tau) = \log\{L(\beta, \delta, \tau)\}$, then

$$\frac{\partial l(\beta, \delta, \tau)}{\partial \tau} = \frac{\partial L(\beta, \delta, \tau)/\partial \tau}{L(\beta, \delta, \tau)}.$$

Now

$$\begin{aligned} \frac{\partial L(\beta, \delta, \tau)}{\partial \tau} &= \int \exp\left[l_c\{\beta, \gamma(\delta, \tau, u)\}\right]\exp\{-\frac{1}{2}u^T u\}\frac{\partial l_c\{\beta, \gamma(\delta, \tau, u)\}}{\partial \gamma^T}(\frac{1}{2}\tau^{-1/2}\Sigma^{1/2}u)du \\ &= \frac{1}{2\tau^{1/2}}\int \exp\left[l_c\{\beta, \gamma(\delta, \tau, u)\}\right]\exp\{-\frac{1}{2}u^T u\}\frac{\partial l_c\{\beta, \gamma(\delta, \tau, u)\}}{\partial \gamma^T}\Sigma^{1/2}udu. \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{\partial l(\beta, \delta, \tau)}{\partial \tau} &= \int \exp\left[l_c\{\beta, \gamma(\delta, \tau, u)\}\right]\exp\{-\frac{1}{2}u^T u\} \\ &\quad \times \frac{\partial l_c\{\beta, \gamma(\delta, \tau, u)\}}{\partial \gamma^T}\Sigma^{1/2}udu/\{2\tau^{1/2}L(\beta, \delta, \tau)\}. \end{aligned}$$

For simplicity, let us define

$$\begin{aligned} n(\beta,\delta,\tau) &= \int \exp\left[l_c\{\beta,\gamma(\delta,\tau,u)\}\right]\exp\{-\frac{1}{2}u^Tu\}\frac{\partial l_c\{\beta,\gamma(\delta,\tau,u)\}}{\partial\gamma^T}\Sigma^{1/2}udu, \\ d(\beta,\delta,\tau) &= 2\tau^{1/2}L(\beta,\delta,\tau), \end{aligned}$$

then

$$\begin{aligned} \frac{\partial l(\beta,\delta,\tau)}{\partial\tau} &= \frac{n(\beta,\delta,\tau)}{d(\beta,\delta,\tau)}, \\ \frac{\partial L(\beta,\delta,\tau)}{\partial\tau} &= \frac{1}{2\tau^{1/2}}\cdot n(\beta,\delta,\tau). \end{aligned}$$

We want to evaulate $\partial l(\beta,\delta,\tau)/\partial\tau$ at $(\widehat{\beta},\widehat{\delta},\tau=0)$, it is easy to see that $d(\beta,\delta,\tau)$ equals to 0 when evaluated at $(\widehat{\beta},\widehat{\delta},\tau=0)$. For $n(\beta,\delta,\tau)$, because $u=\tau^{-1/2}\Sigma^{1/2}a$, $a\sim N(0,\tau\Sigma^{-1})$, and

$$\tau^{-1/2}\Sigma^{1/2}(\tau\Sigma^{-1})\Sigma^{1/2}\tau^{-1/2}=\mathcal{I},$$

where $\mathcal{I}$ is the identity matrix, we have

$$u\sim N(0,\mathcal{I}).$$

Hence the expectation of u is 0, which implies that $n(\beta,\delta,\tau)$ also equals to 0 when evaluated at $(\widehat{\beta},\widehat{\delta},\tau=0)$. Therefore we shall apply the L'Hôpital's rule.

We now have

$$\begin{aligned} \frac{\partial d(\beta,\delta,\tau)}{\partial\tau} &= \tau^{-1/2}L(\beta,\delta,\tau)+2\tau^{1/2}\frac{\partial L(\beta,\delta,\tau)}{\partial\tau} \\ &= \tau^{-1/2}L(\beta,\delta,\tau)+n(\beta,\delta,\tau), \end{aligned}$$

and

$$\frac{\partial n(\beta,\delta,\tau)}{\partial\tau} = \int \exp\left[l_c\{\beta,\gamma(\delta,\tau,u)\}\right]\exp\{-\frac{1}{2}u^Tu\}\times\{A+B\}du,$$

where

$$\begin{aligned} A &= \frac{\partial\{\frac{\partial l_c\{\beta,\gamma(\delta,\tau,u)\}}{\partial\gamma^T}\Sigma^{1/2}u\}}{\partial\tau} \\ &= (\Sigma^{1/2}u)^T\frac{\partial^2 l_c\{\beta,\gamma(\delta,\tau,u)\}}{\partial\gamma\partial\gamma^T}\frac{\partial\gamma}{\partial\tau} \\ &= \frac{1}{2\tau^{1/2}}u^T\Sigma^{1/2}\frac{\partial^2 l_c\{\beta,\gamma(\delta,\tau,u)\}}{\partial\gamma\partial\gamma^T}\Sigma^{1/2}u, \end{aligned}$$

and

$$\begin{aligned} B &= \frac{\partial l_c\{\beta,\gamma(\delta,\tau,u)\}}{\partial\gamma^T}\Sigma^{1/2}u\frac{\partial l_c\{\beta,\gamma(\delta,\tau,u)\}}{\partial\tau} \\ &= \frac{1}{2\tau^{1/2}}\frac{\partial l_c\{\beta,\gamma(\delta,\tau,u)\}}{\partial\gamma^T}\Sigma^{1/2}uu^T\Sigma^{1/2}\frac{\partial l_c\{\beta,\gamma(\delta,\tau,u)\}}{\partial\gamma}, \end{aligned}$$

where we use the fact that for any two compatible vectors a and b, where a depends on a scalar τ, while b is independent of τ,

$$\frac{\partial(a^Tb)}{\partial\tau} = b^T\frac{\partial a}{\partial\tau}.$$

Therefore,

$$\begin{aligned} & \left.\frac{\partial l(\beta,\delta,\tau)}{\partial\tau}\right|_{\widehat{\beta},\widehat{\delta},\tau=0} \\ = & \left.\frac{n(\beta,\delta,\tau)}{d(\beta,\delta,\tau)}\right|_{\widehat{\beta},\widehat{\delta},\tau=0} \\ = & \left.\frac{\partial n(\beta,\delta,\tau)/\partial\tau}{\partial d(\beta,\delta,\tau)/\partial\tau}\right|_{\widehat{\beta},\widehat{\delta},\tau=0} \\ = & \left.\frac{\int\exp\left[l_c\{\beta,\gamma(\delta,\tau,u)\}\right]\exp\{-\frac{1}{2}u^Tu\}\times Cdu}{2\tau^{1/2}\{\tau^{-1/2}L(\beta,\delta,\tau)+n(\beta,\delta,\tau)\}}\right|_{\widehat{\beta},\widehat{\delta},\tau=0} \\ = & \left.\frac{\int\exp\left[l_c\{\beta,\gamma(\delta,\tau,u)\}\right]\exp\{-\frac{1}{2}u^Tu\}\times Cdu}{2L(\beta,\delta,\tau)+2\tau^{1/2}n(\beta,\delta,\tau)}\right|_{\widehat{\beta},\widehat{\delta},\tau=0}, \end{aligned}$$

where

$$C = u^T\Sigma^{1/2}\frac{\partial^2 l_c\{\beta,\gamma(\delta,\tau,u)\}}{\partial\gamma\partial\gamma^T}\Sigma^{1/2}u + \frac{\partial l_c\{\beta,\gamma(\delta,\tau,u)\}}{\partial\gamma^T}\Sigma^{1/2}uu^T\Sigma^{1/2}\frac{\partial l_c\{\beta,\gamma(\delta,\tau,u)\}}{\partial\gamma}.$$

Note that when $\tau = 0$,

$$\gamma(\delta, \tau, u) = H\delta = \gamma(\delta, 0)$$

does not depend on u. Therefore,

$$\begin{aligned} & \left.\frac{\partial l(\beta, \delta, \tau)}{\partial \tau}\right|_{\widehat{\beta}, \widehat{\delta}, \tau=0} \\ = & \left.\frac{\exp\left[l_c\{\beta, \gamma(\delta, 0)\}\right] \int \exp\{-\frac{1}{2}u^T u\} \times C|_{\tau=0} du}{2L(\beta, \delta, 0)}\right|_{\widehat{\beta}, \widehat{\delta}}. \end{aligned}$$

Note that

$$\begin{aligned} & L(\beta, \delta, 0) \\ = & \int \exp\left[l_c\{\beta, \gamma(\delta, 0)\}\right] \varphi_r(a; 0, 0) da \\ = & \exp\left[l_c\{\beta, \gamma(\delta, 0)\}\right], \end{aligned}$$

and recall that

$$u \sim N(0, \mathcal{I}).$$

Using the following result: If $\mathrm{E}(X) = \mu$, $\mathrm{Var}(X) = V$ and A is symmetric, then $\mathrm{E}(X^T A X) = \mu^T A \mu + \mathrm{tr}(AV)$, where E and Var represent expectation and variance, respectively, and noting that $\mathrm{E}(uu^T) = \mathcal{I}$, we get

$$\begin{aligned} & \int \exp\{-\frac{1}{2}u^T u\} \times C|_{\tau=0} du \\ = & \frac{\partial l_c\{\beta, \gamma(\delta, 0)\}}{\partial \gamma^T} \Sigma \frac{\partial l_c\{\beta, \gamma(\delta, 0)\}}{\partial \gamma} + \mathrm{tr}\left(\frac{\partial^2 l_c\{\beta, \gamma(\delta, 0)\}}{\partial \gamma \partial \gamma^T} \Sigma\right). \end{aligned}$$

Therefore,

$$\left.\frac{\partial l(\beta, \delta, \tau)}{\partial \tau}\right|_{\widehat{\beta}, \widehat{\delta}, \tau=0} = \frac{1}{2}\left.\left\{\frac{\partial l_c\{\beta, \gamma(\delta, 0)\}}{\partial \gamma^T} \Sigma \frac{\partial l_c\{\beta, \gamma(\delta, 0)\}}{\partial \gamma} + \mathrm{tr}\left(\frac{\partial^2 l_c\{\beta, \gamma(\delta, 0)\}}{\partial \gamma \partial \gamma^T} \Sigma\right)\right\}\right|_{\widehat{\beta}, \widehat{\delta}}.$$

We have proved the result.

3.7.2 Heuristic Derivation of the Distribution of the Score of τ under H_0

Throughout, we assume that the null hypothesis $\mathrm{H}_0 : \tau = 0$ is true. Denote by β_0 and δ_0 the true values of the parameters β and δ. For simplicity, define

$$
\begin{aligned}
S_\beta\{\beta, \gamma(\delta, 0)\} &= \frac{\partial l_c\{\beta, \gamma(\delta, 0)\}}{\partial \beta},\ S_\gamma\{\beta, \gamma(\delta, 0)\} = \frac{\partial l_c\{\beta, \gamma(\delta, 0)\}}{\partial \gamma},\\
I_{\beta\beta}\{\beta, \gamma(\delta, 0)\} &= -\frac{\partial^2 l_c\{\beta, \gamma(\delta, 0)\}}{\partial \beta \partial \beta^T},\ I_{\beta\gamma}\{\beta, \gamma(\delta, 0)\} = -\frac{\partial^2 l_c\{\beta, \gamma(\delta, 0)\}}{\partial \beta \partial \gamma^T},\\
I_{\gamma\beta}\{\beta, \gamma(\delta, 0)\} &= -\frac{\partial^2 l_c\{\beta, \gamma(\delta, 0)\}}{\partial \gamma \partial \beta^T},\ I_{\gamma\gamma}\{\beta, \gamma(\delta, 0)\} = -\frac{\partial^2 l_c\{\beta, \gamma(\delta, 0)\}}{\partial \gamma \partial \gamma^T}.
\end{aligned}
$$

By the mean value expansion, we have

$$
\begin{aligned}
\begin{pmatrix} S_\beta\{\widehat{\beta}, \gamma(\widehat{\delta}, 0)\} \\ S_\gamma\{\widehat{\beta}, \gamma(\widehat{\delta}, 0)\} \end{pmatrix} &= \begin{pmatrix} S_\beta\{\beta_0, \gamma(\delta_0, 0)\} \\ S_\gamma\{\beta_0, \gamma(\delta_0, 0)\} \end{pmatrix} \\
&\quad - \begin{bmatrix} I^*_{\beta\beta} & I^*_{\beta\gamma} H \\ I^*_{\gamma\beta} & I^*_{\gamma\gamma} H \end{bmatrix} \begin{pmatrix} \widehat{\beta} - \beta_0 \\ \widehat{\delta} - \delta_0 \end{pmatrix}, \qquad (3.10)
\end{aligned}
$$

where $I^*_{\beta\beta} = I_{\beta\beta}\{\beta^*, \gamma(\delta^*, 0)\}$ and similarly for $I^*_{\beta\gamma}$, $I^*_{\gamma\beta}$ and $I^*_{\gamma\gamma}$; β^* is some value between β_0 and $\widehat{\beta}$; and δ^* is some value between δ_0 and $\widehat{\delta}$. Because $(\widehat{\beta}, \widehat{\delta})$ are the maximum partial likelihood estimators under H_0, we have

$$
\begin{aligned}
0 &= \begin{pmatrix} \partial l_c(\widehat{\beta}, \widehat{\delta})/\partial \beta \\ \partial l_c(\widehat{\beta}, \widehat{\delta})/\partial \delta \end{pmatrix} = \begin{pmatrix} S_\beta\{\widehat{\beta}, \gamma(\widehat{\delta}, 0)\} \\ H^T S_\gamma\{\widehat{\beta}, \gamma(\widehat{\delta}, 0)\} \end{pmatrix} \\
&\overset{(3.10)}{=} \begin{pmatrix} S_\beta\{\beta_0, \gamma(\delta_0, 0)\} \\ H^T S_\gamma\{\beta_0, \gamma(\delta_0, 0)\} \end{pmatrix} - \begin{bmatrix} I^*_{\beta\beta} & I^*_{\beta\gamma} H \\ H^T I^*_{\gamma\beta} & H^T I^*_{\gamma\gamma} H \end{bmatrix} \begin{pmatrix} \widehat{\beta} - \beta_0 \\ \widehat{\delta} - \delta_0 \end{pmatrix}.
\end{aligned}
$$

Therefore,

$$\begin{pmatrix} \widehat{\beta}-\beta_0 \\ \widehat{\delta}-\delta_0 \end{pmatrix} = \begin{bmatrix} I^*_{\beta\beta} & I^*_{\beta\gamma}H \\ H^T I^*_{\gamma\beta} & H^T I^*_{\gamma\gamma}H \end{bmatrix}^{-1} \begin{pmatrix} S_\beta\{\beta_0,\gamma(\delta_0,0)\} \\ H^T S_\gamma\{\beta_0,\gamma(\delta_0,0)\} \end{pmatrix}. \quad (3.11)$$

Note that β is $p \times 1$ and γ is $r \times 1$. From (3.11) we obtain

$$\begin{pmatrix} \widehat{\beta}-\beta_0 \\ \widehat{\delta}-\delta_0 \end{pmatrix} = \begin{bmatrix} I^*_{\beta\beta} & I^*_{\beta\gamma}H \\ H^T I^*_{\gamma\beta} & H^T I^*_{\gamma\gamma}H \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{I}_p & 0_{p\times r} \\ 0_{1\times p} & H^T \end{bmatrix} \times \begin{pmatrix} S_\beta\{\beta_0,\gamma(\delta_0,0)\} \\ S_\gamma\{\beta_0,\gamma(\delta_0,0)\} \end{pmatrix}, \quad (3.12)$$

where $\mathcal{I}_k$ denotes the $k \times k$ identity matrix.

The lower part of (3.10) yields

$$S_\gamma\{\widehat{\beta},\gamma(\widehat{\delta},0)\} = S_\gamma\{\beta_0,\gamma(\delta_0,0)\} - (I^*_{\gamma\beta} \;\; I^*_{\gamma\gamma}H) \begin{pmatrix} \widehat{\beta}-\beta_0 \\ \widehat{\delta}-\delta_0 \end{pmatrix}. \quad (3.13)$$

Substituting (3.12) into (3.13), we get

$$S_\gamma\{\widehat{\beta},\gamma(\widehat{\delta},0)\} = \left\{ (0_{r\times p} \; \mathcal{I}_r) - (I^*_{\gamma\beta} \;\; I^*_{\gamma\gamma}H) \begin{bmatrix} I^*_{\beta\beta} & I^*_{\beta\gamma}H \\ H^T I^*_{\gamma\beta} & H^T I^*_{\gamma\gamma}H \end{bmatrix}^{-1} \times \begin{bmatrix} \mathcal{I}_p & 0_{p\times r} \\ 0_{1\times p} & H^T \end{bmatrix} \right\} \begin{pmatrix} S_\beta\{\beta_0,\gamma(\delta_0,0)\} \\ S_\gamma\{\beta_0,\gamma(\delta_0,0)\} \end{pmatrix}. \quad (3.14)$$

Denote by γ_0 the true value of γ. Note that under H$_0$, $\gamma(\delta_0,0)=\gamma_0$; $\widehat{\beta}$ and $\widehat{\delta}$ are consistent estimators for β_0 and δ_0. As $\gamma(\cdot,\cdot)$ is continuous, both $\gamma(\widehat{\delta},0)$ and $\gamma(\delta^*,0)$ are consistent estimators for γ_0. By well-known results, e.g., Andersen and Gill (1982, Theorem 3.2), there exists a nonnegative definite matrix V such that

$$n^{-1/2}\begin{pmatrix} S_\beta\{\beta_0,\gamma(\delta_0,0)\} \\ S_\gamma\{\beta_0,\gamma(\delta_0,0)\} \end{pmatrix} \quad \xrightarrow{d} \quad N(0,V), \text{ and}$$

$$n^{-1}\widehat{I} \text{ and } n^{-1}I^* \quad \xrightarrow{p} \quad V = \begin{bmatrix} V_{\beta\beta} & V_{\beta\gamma} \\ V_{\gamma\beta} & V_{\gamma\gamma} \end{bmatrix}, \qquad (3.15)$$

where $\widehat{I}$ and I^* are matrices with elements $I_{\beta\beta}$, $I_{\beta\gamma}$, $I_{\gamma\beta}$ and $I_{\gamma\gamma}$ evaluated at $(\widehat{\beta},\widehat{\delta})$ and (β^*,δ^*), respectively.

From the second result in (3.15), it is easy to see the second term on the right hand side of (3.7) converges in probability to $-\frac{1}{2}\text{tr}(nV_{\gamma\gamma}\Sigma)$. Now from the first result and (3.14), we have

$$\begin{aligned} & n^{-1/2}S_\gamma\{\widehat{\beta},\gamma(\widehat{\delta},0)\} \\ = & \left\{(0_{r\times p}\ \mathcal{I}_r) \ - \ (n^{-1}I^*_{\gamma\beta} \ \ n^{-1}I^*_{\gamma\gamma}H)\begin{bmatrix} n^{-1}I^*_{\beta\beta} & n^{-1}I^*_{\beta\gamma}H \\ n^{-1}H^TI^*_{\gamma\beta} & n^{-1}H^TI^*_{\gamma\gamma}H \end{bmatrix}^{-1}\right. \\ & \left.\times\begin{bmatrix} \mathcal{I}_p & 0_{p\times r} \\ 0_{1\times p} & H^T \end{bmatrix}\right\} \times \ n^{-1/2}\begin{pmatrix} S_\beta\{\beta_0,\gamma(\delta_0,0)\} \\ S_\gamma\{\beta_0,\gamma(\delta_0,0)\} \end{pmatrix} \xrightarrow{d} N(0, WVW^T), \end{aligned}$$

where

$$W \ = \ (0_{r\times p}\ \mathcal{I}_r) \ - \ (V_{\gamma\beta} \ \ V_{\gamma\gamma}H)\begin{bmatrix} V_{\beta\beta} & V_{\beta\gamma}H \\ H^TV_{\gamma\beta} & H^TV_{\gamma\gamma}H \end{bmatrix}^{-1}\begin{bmatrix} \mathcal{I}_p & 0_{p\times r} \\ 0_{1\times p} & H^T \end{bmatrix}.$$

After some algebra, one can show that

$$WVW^T = V_{\gamma\gamma} - (V_{\gamma\beta}\ V_{\gamma\gamma}H)\begin{bmatrix} V_{\beta\beta} & V_{\beta\gamma}H \\ H^TV_{\gamma\beta} & H^TV_{\gamma\gamma}H \end{bmatrix}^{-1}\begin{pmatrix} V_{\beta\gamma} \\ H^TV_{\gamma\gamma} \end{pmatrix}.$$

In practice we may consistently estimate the unknown matrix V by the matrix $\widehat{V}$ found by substituting $\widehat{I}_{\beta\beta}/n$, $\widehat{I}_{\beta\gamma}/n$, $\widehat{I}_{\gamma\beta}/n$, $\widehat{I}_{\gamma\gamma}/n$ for $V_{\beta\beta}$, $V_{\beta\gamma}$, $V_{\gamma\beta}$, $V_{\gamma\gamma}$, respectively, where $\widehat{I}_{\beta\beta} = I_{\beta\beta}\{\widehat{\beta}, \gamma(\widehat{\delta}, 0)\}$ and similarly for the others.

As a special case, note that if S_i is the only covariate in the Cox model, then we have the simplified form $W = (0_{r\times p} \;\; \mathcal{I}_r - V_{\gamma\gamma}H(H^T V_{\gamma\gamma}H)^{-1}H^T)$ and $WVW^T = V_{\gamma\gamma} - V_{\gamma\gamma}H(H^T V_{\gamma\gamma}H)^{-1}H^T V_{\gamma\gamma}$.

3.8 Tables and Figures

Table 3.1: Empirical sizes of nominal 0.05-level spline tests for proportional hazards of S_i in the model $\lambda(t|S_i) = \lambda_0(t)\exp\{S_i\delta_0\}, i = 1, 2, \cdots, n$, expressed as percent. $\lambda_0(t) = 1$; values of S_i are equally spaced on the interval $(0, 1)$ with an equal number of subjects having each distinct S_i values. Results are based on 2000 simulations for each scenario.

	Number of	True value of δ_0					
Censoring	distinct	$n = 100$			$n = 200$		
distribution	S_i values	0	1	2	0	1	2
Unit exponential	2	5.10	5.70	6.10	6.20	5.40	4.95
	4	5.70	6.05	5.10	5.60	4.65	4.85
	10	5.70	6.30	5.95	6.40	5.00	5.30
	20	5.60	6.35	5.85	6.40	4.75	4.85
	50	5.90	6.20	6.00	6.45	4.65	4.60
	100	5.70	6.60	5.95	6.35	4.65	4.60
	200				6.40	4.90	4.70
Uniform (0,2)	2	5.20	4.45	5.20	5.60	4.60	4.35
	4	5.55	4.55	4.55	4.85	4.75	4.25
	10	5.35	4.10	5.20	5.00	4.45	4.75
	20	5.30	4.30	4.50	4.95	4.95	4.75
	50	5.35	4.15	4.90	4.85	4.70	4.60
	100	5.40	4.30	4.90	4.80	4.65	4.45
	200				4.80	4.85	4.55

Table 3.2: Estimated powers of nominal 0.05-level tests for proportional hazards of S_i in the model $\lambda(t|S_i) = \lambda_0(t)\exp\{S_i\gamma(t)\}, i = 1, 2, \cdots, n$, expressed as percent. $\lambda_0(t) = 1$; S_i is a single binary covariate defining two groups of equal size; $\gamma(t)$ is the true alternative; $n = 200$. Censoring distribution is uniform on $(0, 2)$. Tests and alternatives are as described in the text. Results are based on 1000 simulations for each scenario.

	Alternative				
Test	Curve 1	Curve 2	Curve 3	Curve 4	Curve 5
Spline	90.8	78.4	47.6	37.3	28.6
Linear	90.5	78.8	51.4	10.1	30.4
Quadratic	79.7	65.3	50.0	13.8	36.6
Log	93.3	75.8	37.4	32.1	15.5
Optimal	93.3	81.7	51.4	91.5	46.6

Table 3.3: Empirical sizes of nominal 0.05-level spline tests for covariate effects of S_i in the model $\lambda(t|S_i) = \lambda_0(t)$ (no effect) and $\lambda(t|S_i) = \lambda_0(t)\exp\{S_i\}$ (linear effect), $i = 1, 2, \cdots, n$, expressed as percent. $\lambda_0(t) = 1$; values of S_i are as in Table 3.1. Results are based on 2000 simulations for each scenario.

	Number of	Null hypothesis			
Censoring	distinct	$n = 100$		$n = 200$	
distribution	S_i values	No effect	Linear effect	No effect	Linear effect
Unit exponential	4	5.25	4.65	5.10	4.90
	10	5.20	4.35	5.00	4.60
	20	5.15	4.60	5.05	4.50
	50	5.05	4.45	4.95	4.60
	100	5.15	4.25	5.00	4.80
	200			4.95	4.70
Uniform (0,1.5)	4	4.90	4.80	4.50	4.65
	10	5.30	5.15	5.10	5.05
	20	5.00	5.50	4.60	4.90
	50	5.05	5.60	4.50	4.95
	100	5.00	5.70	4.70	4.95
	200			4.70	4.85

Table 3.4: Estimated powers of nominal 0.05-level tests for covariate effects of S_i in the model $\lambda(t|S_i) = \lambda_0(t)\exp\{\gamma(S_i)\}, i = 1, 2, \cdots, n$, expressed as percent. $\lambda_0(t) = 1$; values of S_i are equally spaced on the interval $[-1.719, 1.719]$ with step 0.0173; $\gamma(S_i)$ is the true alternative; $n = 200$. Censoring distribution is uniform on $(0, 1.5)$. Tests and alternatives are as described in the text. Results are based on 1000 simulations for each scenario.

Null		Alternative					
hypothesis	Test	E	L	S1	Q	C	S2
No effect	Spline	74.0	72.4	73.8	23.1	5.8	16.2
	Linear	74.4	71.5	68.9	4.5	4.3	4.2
	Quadratic	71.5	60.4	84.1	73.6	5.9	44.7
	Cubic	67.2	55.5	84.1	67.7	6.2	38.5
	Optimal	81.6	74.2	96.3	81.7	92.0	93.7
Linear effect	Spline	12.8	4.9	56.0	80.7	7.7	65.4
	Quadratic	13.7	4.9	54.0	81.7	6.9	58.5
	Cubic	12.0	7.5	54.0	73.7	6.7	46.4
	Optimal	14.2	10.5	78.3	81.7	91.9	93.8

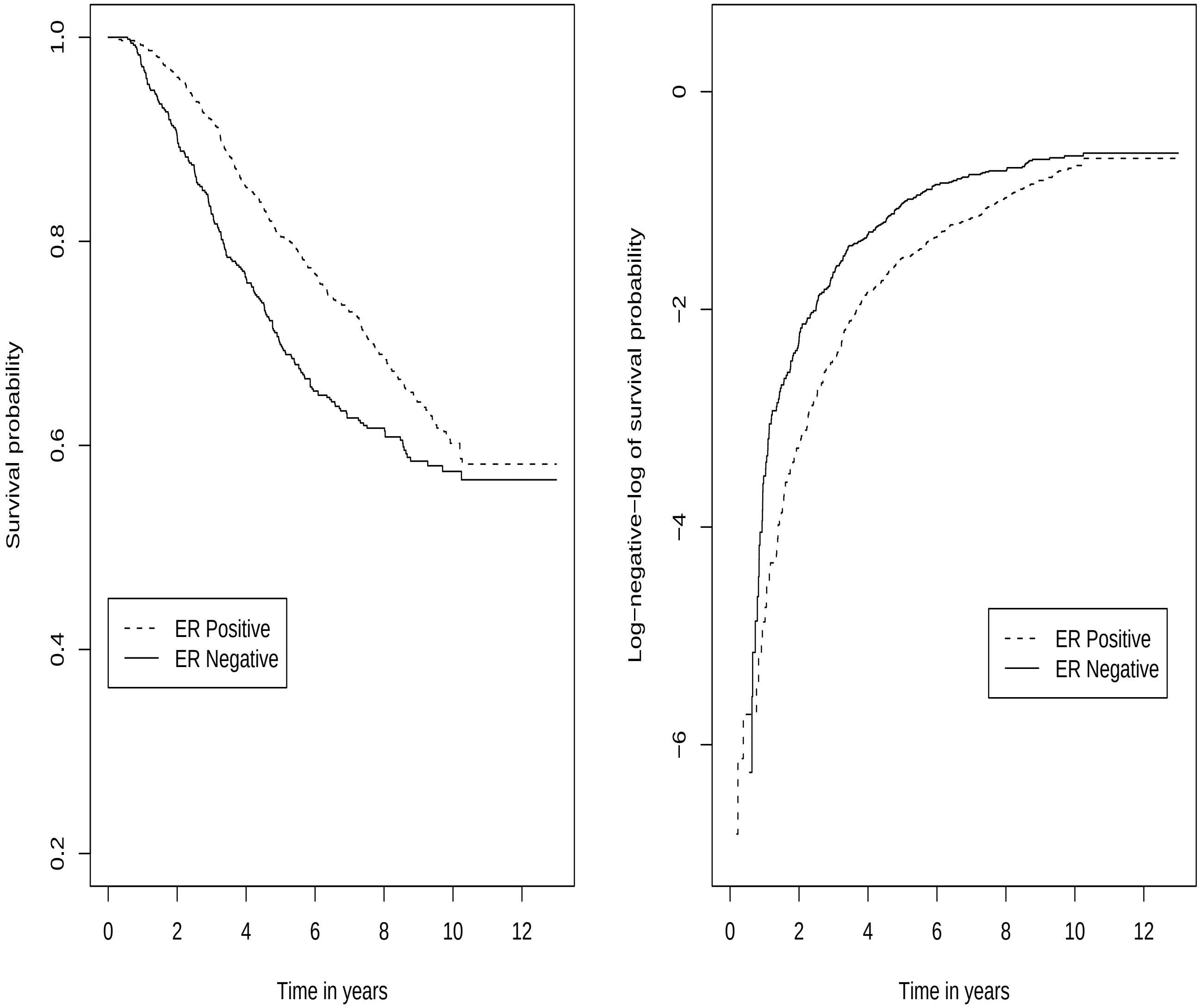

Figure 3.1: CALGB 8541: Survival and log-negative-log of survival distribution by ER status. Survival distribution was estimated by using the Kaplan-Meier method.

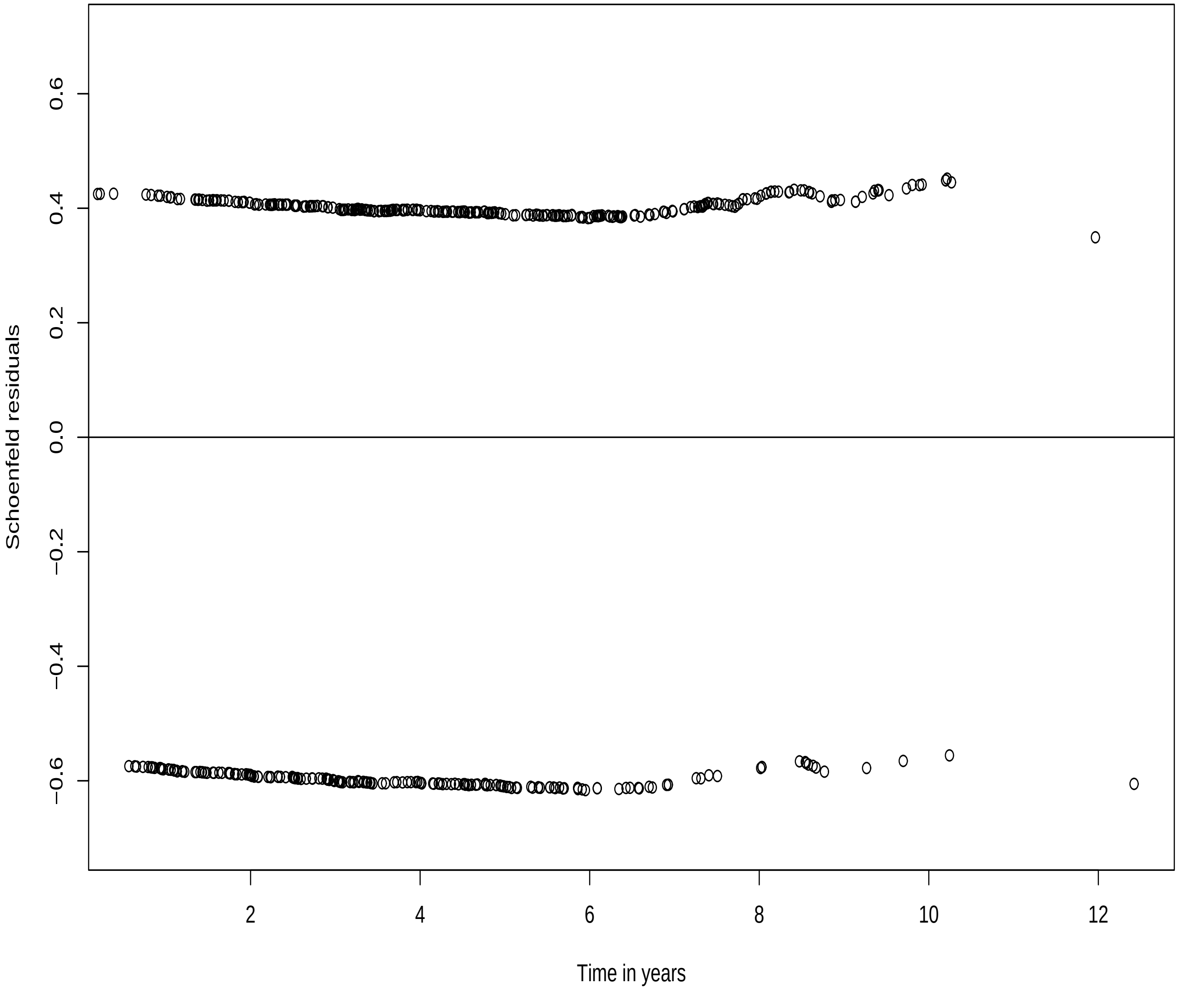

Figure 3.2: CALGB 8541: Schoenfeld (1982) residuals of ER status, obtained by using SAS PROC PHREG. Residuals above and below the horizontal line are for patients with ER positive and those with ER negative, respectively.

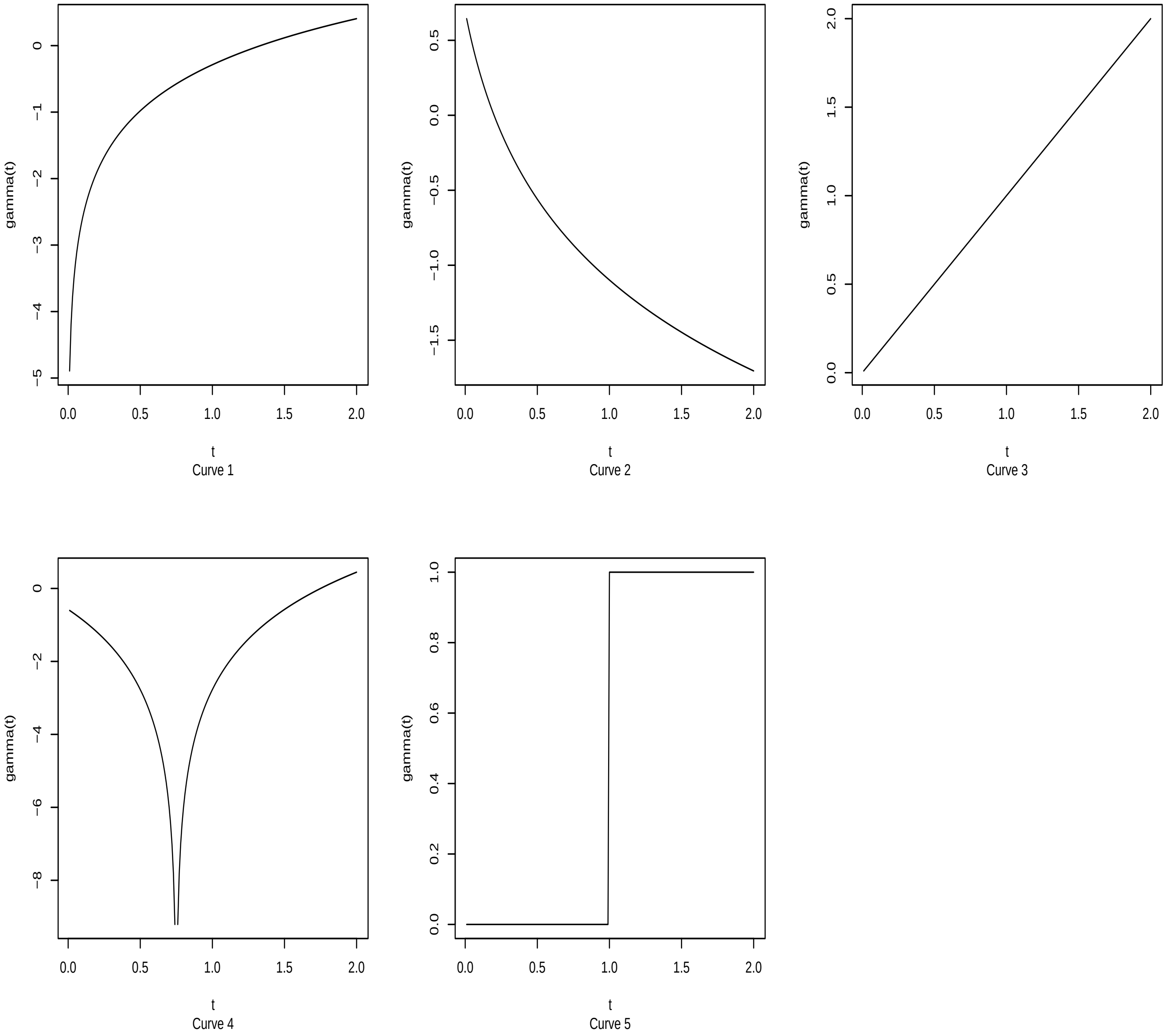

Figure 3.3: Plots of curves used in the simulation study evaluating powers of the tests for proportional hazards. Curve 1: $\gamma(t) = \log\{.75t\}$; curve 2: $\gamma(t) = \log\{2/(1+5t)\}$; curve 3: $\gamma(t) = \log\{e^t\} = t$; curve 4: $\gamma(t) = \log\{(t-.75)^2\}$; curve 5: $\gamma(t) = \log\{e^{I(t\geq 1)}\} = I(t \geq 1)$.

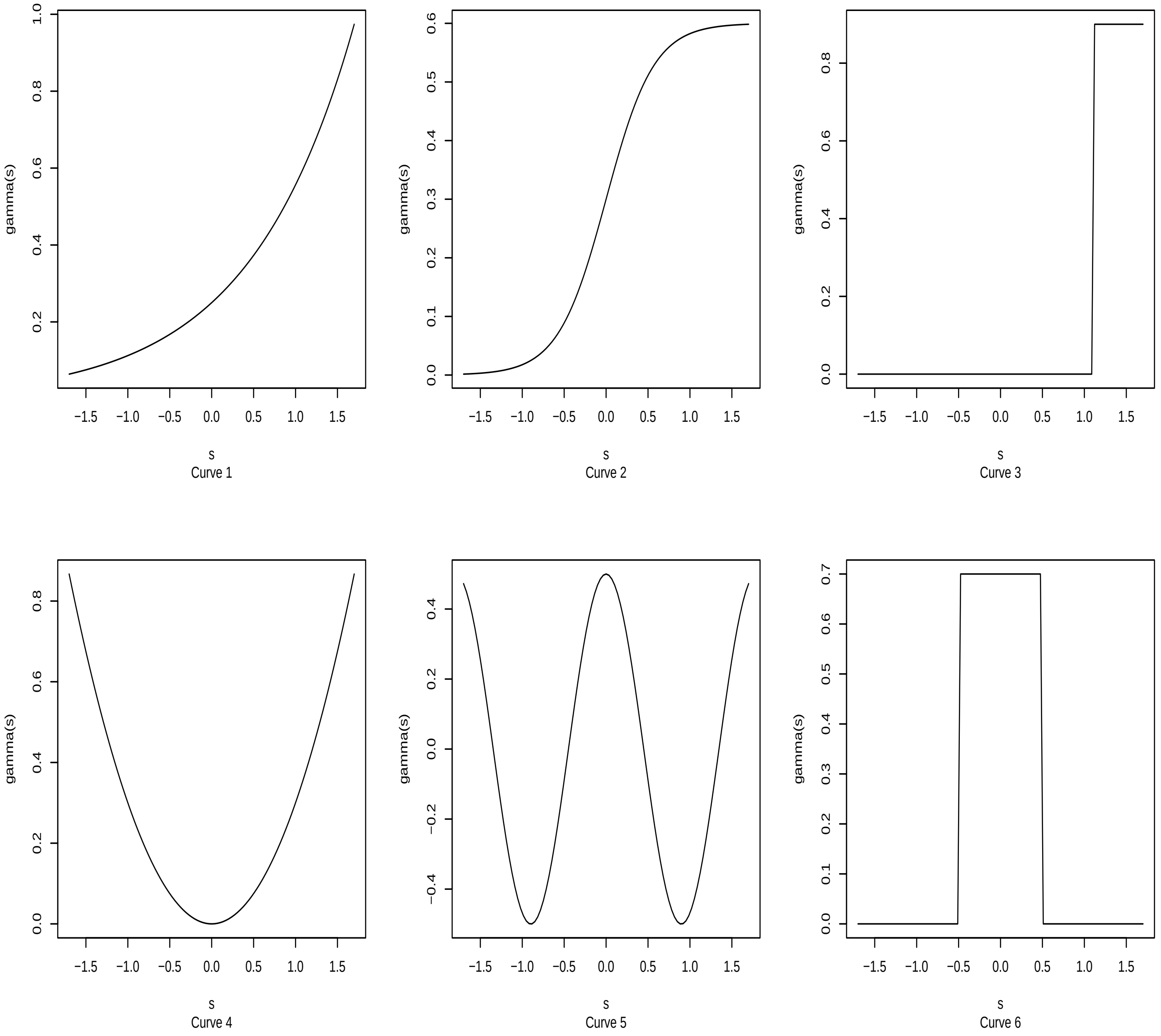

Figure 3.4: Plots of curves used in the simulation study evaluating powers of the tests for covariate effects. Curve 1 (E): $\gamma(s) = .25\exp\{.8s\}$; curve 2 (L): $\gamma(s) = .6\exp\{3.5s\}/(1+\exp\{3.5s\})$; curve 3 (S1): $\gamma(s) = .9I(s > 1.1)$; curve 4 (Q): $\gamma(s) = .3s^2$; curve 5 (C): $\gamma(s) = .5\cos(3.5s)$; curve 6 (S2): $\gamma(s) = .7I(|s| < .5)$.

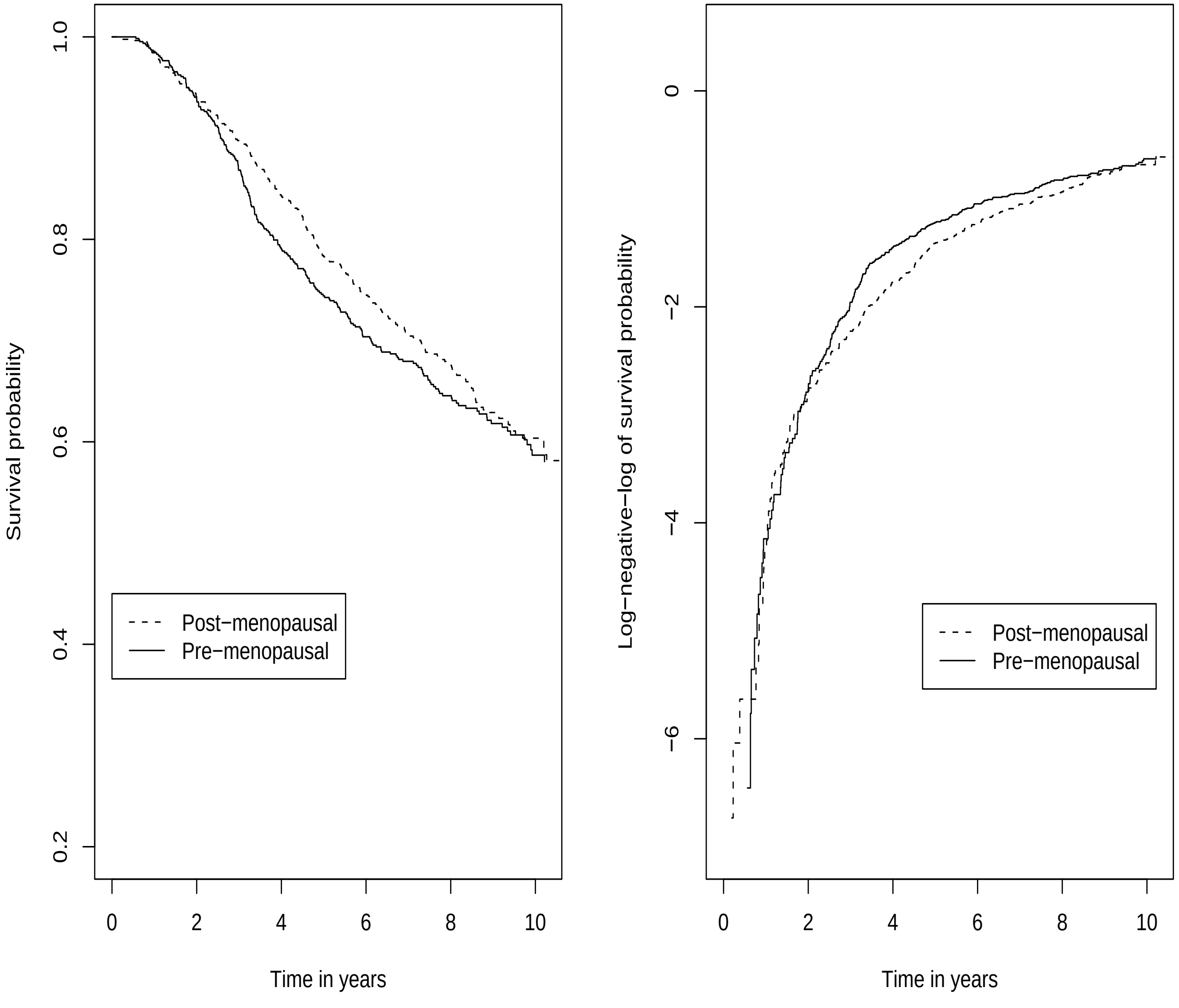

Figure 3.5: CALGB 8541: Survival and log-negative-log of survival distribution by menopausal status. Survival distribution was estimated by using the Kaplan-Meier method.

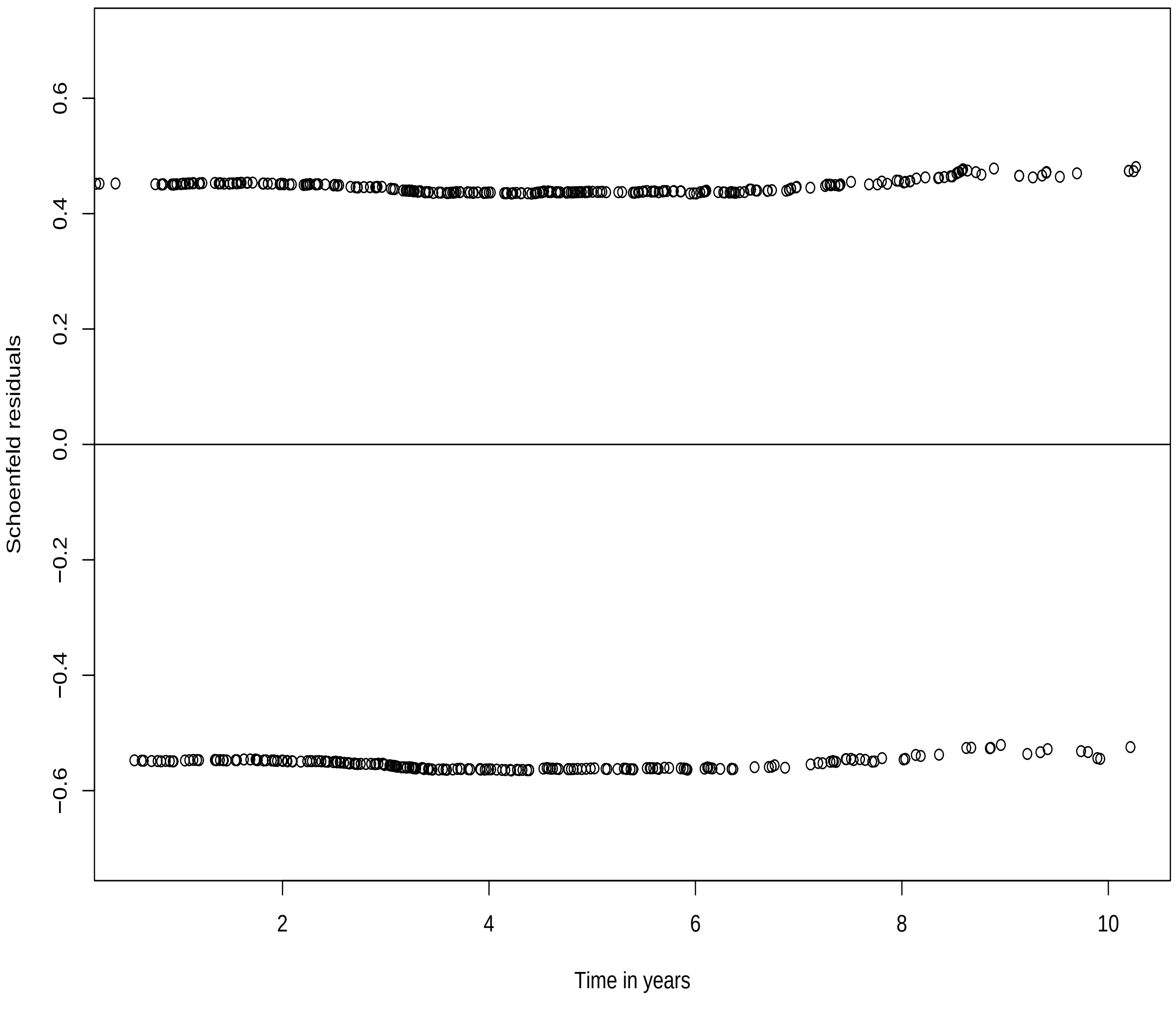

Figure 3.6: CALGB 8541: Schoenfeld (1982) residuals of menopausal status, obtained by using SAS PROC PHREG. Residuals above and below the horizontal line are for patients post-menopausal and those pre-menopausal, respectively.

Bibliography

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Annals of Statistics* **10**, 1100-1120.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9-25.

Cox, D. D. and Chang, Y. (1990). Iterated state space algorithms and cross validation for generalized smoothing splines. *Technical report 49, Department of Statistics, University of Illinois, Champion, IL.*

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

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik* **31**, 377-403.

Davies, R. B. (1980). The distribution of a linear combination of χ^2 random variables. *Applied Statistics* **29**, 323-333.

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

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

Gray, R. J. (1994). Spline-based tests in survival analysis. *Biometrics* **50**, 640-652.

Gray, R. J. (2000). Estimation of regression parameters and the hazard function in transformed linear survival models. *Biometrics* **56**, 571-576.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society, Series B* **46**, 149-192.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review* **55**, 245-259.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized*

Linear Models. London: Chapman and Hall.

Gu, C. (1990). Adaptive spline smoothing in non-Gaussian regression models. *Journal of the American Statistical Association* **85**, 801-807.

Gu, C. (1992). Cross validating non-Gaussian data. *Journal of the Computational and Graphical Statistics* **1**, 169-179.

Gu, C. and Xiang, D. (2001). Cross validating non-Gaussian data: generalized approximate cross-validation revisited. *Journal of the Computational and Graphical Statistics* **10**, 581-591.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**, 320-340.

Hastie, T. and Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics* **46**, 1005-1016.

Kimeldorf, G. and Wahba, G. (1971). Some results on Tachebycheffian spline functions. *Journal of Mathematical Analysis and Applications* **33**, 82-95.

Klein, J. P. and Moeschberger, M. L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer.

Kohn, R., Ansley, C. F., and Tharm, D. (1991). The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *Journal of the American Statistical Association* **86**, 1042-1050.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society, Series B* **61**, 381-400.

McCullagh, P. and Nelder, P. A. (1989). *Generalized Linear Models (2nd ed.).* London: Chapman and Hall.

O'Sullivan, F. (1988). Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal on Scientific and Statistical Computing* **9**, 531-542.

O'Sullivan, F., Yandell, B., and Raynor, W. (1986). Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association* **81**, 96-103.

Pettitt, A. N. and Bin Daud, I. (1990). Investigating time dependencies in Cox's proportional hazards model. *Applied Statistics* **39**, 313-329.

Raudenbush, S. W., Yang, M.-L., and Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate

Laplace approximation. *Journal of the Computational and Graphical Statistics* **9**, 141-157.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika* **69**, 239-241.

Speed, T. (1991). Discussion on "BLUP is a good thing: the estimation of random effects" by Robinson, G. K. *Statistical Science* **6**, 15-51.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model.* New York: Springer.

Thompson, R. (1985). Discussion on "some aspects of the spline smoothing approach to non-parametric regression curve fitting" by B. W. Silverman. *Journal of the Royal Statistical Society, Series B* **47**, 43-44.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics* **13**, 1378-1402.

Wahba, G. (1990). *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics, Philadelphia, PA.

Wahba, G. and Wold, S. (1975). Some new mathematical methods for variational objective analysis using splines and cross validation. *Communications in Statistics* **4**, 1-17.

Xiang, D. and Wahba, G. (1996). A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statistica Sinica* **6**, 675-692.

Zhang, D., Lin, X., Raz, J. and Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association* **93**, 710-719.

Zhang, D. and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics* **4**, 57-74.

Zucker, D. M. and Karr, A. F. (1990). Nonparametric survival analysis with time-dependent covariate effects: A penalized partial likelihood approach. *Annals of Statistics* **18**, 329-353.