Non- and Semi- Parametric Modeling in Survival analysis *

Jianqing Fan
Department of ORFE
Princeton University
Princeton, NJ 08544, USA
E-mail: jqfan@princeton.edu

Jiancheng Jiang
Department of Mathematics and Statistics
University of North Carolina
Charlotte, NC 28223, USA
E-mail: jjiang1@uncc.edu

Abstract

In this chapter, we give a selective review of the nonparametric modeling methods using Cox's type of models in survival analysis. We first introduce Cox's model (Cox 1972) and then study its variants in the direction of smoothing. The model fitting, variable selection, and hypothesis testing problems are addressed. A number of topics worthy of further study are given throughout this chapter.

Keywords and Phrases. Censoring, Cox's model, failure time, likelihood, modeling, nonparametric smoothing.

1 Introduction

Survival analysis is concerned with studying the time between entry to a study and a subsequent event and becomes one of the most important fields in statistics. The techniques developed in survival analysis are now applied in many fields, such as biology (survival time), engineering (failure time), medicine (treatment effects or the efficacy of drugs), quality control (lifetime of component), credit risk modeling in finance (default time of a firm).

An important problem in survival analysis is how to model well the conditional hazard rate of failure times given certain covariates, because it involves frequently asked questions about whether or not certain independent variables are correlated with the survival or failure times. These problems have presented a significant challenge to statisticians in the last 5 decades, and their importance

*The authors are partly supported by NSF grants DMS-0532370, DMS-0704337 and NIH R01-GM072611.

has motivated many statisticians to work in this area. Among them is one of the most important contributions, the proportional hazards model or Cox's model and its associated partial likelihood estimation method (Cox, 1972), which stimulated a lot of works in this field. In this chapter we review related work along this direction using the Cox type of models and open an academic research avenue for interested readers. Various estimation methods are considered, a variable selection approach is studied, and a useful inference method, the generalized likelihood ratio (GLR) test, is employed to address hypothesis testing problems for the models. Several topics worthy of further study are laid down in the discussion section.

The remainder of this chapter is organized as follows. We consider univariate Cox's type of models in Section 2 and study multivariate Cox's type of models using the marginal modeling strategy in Section 3. Section 4 focuses on model selection rules, Section 5 is devoted to validating Cox's type of models, and Section 6 discusses transformation models (extensions to Cox's models). Finally, we conclude this chapter in the discussion section.

2 Cox's Type of Models

Model Specification. The celebrated Cox model has provided a tremendously successful tool for exploring the association of covariates with failure time and survival distributions and for studying the effect of a primary covariate while adjusting for other variables. This model assumes that, given a q-dimensional vector of covariates $\mathbf{Z}$, the underlying conditional hazard rate (rather than expected survival time T),

$$\lambda(t|\mathbf{z}) = \lim_{\Delta t \to 0+} \frac{1}{\Delta t} P\{t \leq T < t + \Delta t | T \geq t, \mathbf{Z} = \mathbf{z}\},$$

is a function of the independent variables (covariates):

$$\lambda(t|\mathbf{z}) = \lambda_0(t)\Psi(\mathbf{z}), \tag{1}$$

where $\Psi(\mathbf{z}) = \exp(\psi(\mathbf{z}))$ with the form of the function $\psi(\mathbf{z})$ known such as $\psi(\mathbf{z}) = \boldsymbol{\beta}^T\mathbf{z}$, and $\lambda_0(t)$ is an unknown baseline hazard function. Once the conditional hazard rate is given, the condition survivor function $S(t|\mathbf{z})$ and conditional density $f(t|\mathbf{z})$ are also determined. In general, they have the following relationship:

$$S(t|\mathbf{z}) = \exp(-\Lambda(t|\mathbf{z})), \quad f(t|\mathbf{z}) = \lambda(t|\mathbf{z})S(t|\mathbf{z}), \tag{2}$$

where $\Lambda(t|\mathbf{z}) = \int_0^t \lambda(t|\mathbf{z})dt$ is the cumulative hazard function. Since no assumptions are made about the nature or shape of the baseline hazard function, the Cox regression model may be considered to be a semiparametric model.

The Cox model is very useful for tackling with censored data which often happen in practice. For example, due to termination of the study or early withdrawal from a study, not all of the survival times $T_1, \cdots, T_n$ may be fully observable. Instead one observes for the i^{th} subject an event time $X_i = \min(T_i, C_i)$, a censoring indicator $\delta_i = I(T_i \leq C_i)$, as well as an associated vector of covariates $\mathbf{Z}_i$. Denote the observed data by $\{(\mathbf{Z}_i, X_i, \delta_i) : i = 1, \cdots, n\}$ which is an i.i.d. sample from the population $(\mathbf{Z}, X, \delta)$ with $X = \min(T, C)$ and $\delta = I(T \leq C)$. Suppose that the random variables T and C are positive and continuous. Then by Fan, Gijbels, and King (1997), under the Cox model (1),

$$\Psi(x) = \frac{E\{\delta|\mathbf{Z} = \mathbf{z}\}}{E\{\Lambda_0(X)|\mathbf{Z} = \mathbf{z}\}}, \tag{3}$$

where $\Lambda_0(t) = \int_0^t \lambda_0(u)\, du$ is the cumulative baseline hazard function. Equation (3) allows one to estimate the function Ψ using regression techniques if $\lambda_0(t)$ is known.

The likelihood function can also be derived. When $\delta_i = 0$, all we know is that the survival time $T_i \geq C_i$ and the probability for getting this is

$$P(T_i \geq C_i|\mathbf{Z}_i) = P(T_i \geq X_i|\mathbf{Z}_i) = S(X_i|\mathbf{Z}_i),$$

whereas when $\delta_i = 1$, the likelihood of getting T_i is $f(T_i|\mathbf{Z}_i) = f(X_i|\mathbf{Z}_i)$. Therefore the conditional (given covariates) likelihood for getting the data is

$$L = \prod_{\delta_i=1} f(X_i|\mathbf{Z}_i) \prod_{\delta_i=0} S(X_i|\mathbf{Z}_i) = \prod_{\delta_i=1} \lambda(X_i|\mathbf{Z}_i) \prod_i S(X_i|\mathbf{Z}_i), \tag{4}$$

and using (2), we have

$$\begin{aligned} L &= \sum_{\delta_i=1} \log(\lambda(X_i|\mathbf{Z}_i)) - \sum_i \Lambda(X_i|Z_i). \\ &= \sum_i \delta_i \log(\lambda(X_i|\mathbf{Z}_i)) - \sum_i \Lambda(X_i|\mathbf{Z}_i). \end{aligned} \tag{5}$$

For proportional hazards model (1), we have specifically

$$L = \sum_i \delta_i \log(\lambda_0(X_i)\Psi(Z_i)) - \sum_i \Lambda_0(X_i)\Psi(Z_i). \tag{6}$$

Therefore, when both $\psi(\cdot)$ and $\lambda_0(\cdot)$ are parameterized, the parameters can be estimated by maximizing the likelihood (6).

Estimation. The likelihood inference can be made about the parameters in model (1) if the baseline $\lambda_0(\cdot)$ and the risk function $\psi(\cdot)$ are known up to a vector of unknown parameters $\boldsymbol{\beta}$ (Aitkin and Clayton, 1980), i.e.

$$\lambda_0(\cdot) = \lambda_0(\cdot; \boldsymbol{\beta}); \quad \text{and} \quad \lambda(\cdot) = \lambda_0(\cdot; \boldsymbol{\beta}).$$

When the baseline is completely unknown and the form of the function $\psi(\cdot)$ is given, inference can be based on the partial likelihood (Cox, 1975). Since the full likelihood involves both $\boldsymbol{\beta}$ and $\lambda_0(t)$, Cox decomposed the full likelihood into a product of the term corresponding to identities of successive failures and the term corresponding to the gap times between any two successive failures. The first term inherits the usual large-sample properties of the full likelihood and is called the partial likelihood.

The partial likelihood can also be derived from counting process theory (see for example Andersen, Borgan, Gill, and Keiding 1993) or from a profile likelihood in Johansen (1983). In the following we introduce the latter.

Example 1 [The partial likelihood as profile likelihood; Fan, Gijbel, and King (1997)] Consider the case that $\psi(\mathbf{z}) = \psi(\mathbf{z};\boldsymbol{\beta})$. Let $t_1 < \cdots < t_N$ denote the ordered failure times and let (i) denote the label of the item failing at t_i. Denote by R_i the risk set at time t_i-, that is $R_i = \{j : X_j \geq t_i\}$. Consider the least informative nonparametric modeling for $\Lambda_0(\cdot)$, that is, $\Lambda_0(t)$ puts point mass θ_j at time t_j in the same way as constructing the empirical distribution:

$$\Lambda_0(t;\theta) = \sum_{j=1}^{N} \theta_j I(t_j \leq t). \tag{7}$$

Then

$$\Lambda_0(X_i;\theta) = \sum_{j=1}^{N} \theta_j I(i \in R_j). \tag{8}$$

Under the proportional hazards model (1), using (6), the log likelihood is

$$\begin{aligned}\log L &= \sum_{i=1}^{n}[\delta_i\{\log\lambda_0(X_i;\theta) + \psi(Z_i;\boldsymbol{\beta})\} \\ &\quad -\Lambda_0(X_i;\theta)\exp\{\psi(Z_i;\boldsymbol{\beta})\}].\end{aligned} \tag{9}$$

Substituting (7) and (8) into (9), one establishes that

$$\begin{aligned}\log L &= \sum_{j=1}^{n}[\log\theta_j + \psi(Z_{(j)};\boldsymbol{\beta})] \\ &\quad -\sum_{i=1}^{n}\sum_{j=1}^{N}\theta_j I(i \in R_j)\exp\{\psi(Z_i;\boldsymbol{\beta})\}.\end{aligned} \tag{10}$$

Maximizing $\log L$ with respect to θ_j leads to the following Breslow estimator of the baseline hazard [Brewlow (1972, 1974)]

$$\hat{\theta}_j = \Big[\sum_{i \in R_j} \exp\{\psi(Z_i;\boldsymbol{\beta})\}\Big]^{-1}. \tag{11}$$

Substituting (11) into (10), we obtain

$$\max_{\lambda_0} \log L = \sum_{i=1}^{n}\Big(\psi(\mathbf{Z}_{(i)};\boldsymbol{\beta}) - \log\Big[\sum_{j \in R_i}\exp\{\psi(\mathbf{Z}_j;\boldsymbol{\beta})\}\Big]\Big) - N$$

This leads to the log partial likelihood function (Cox 1975)

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Big(\psi(\mathbf{Z}_{(i)}; \boldsymbol{\beta}) - \log\Big[\sum_{j \in R_i} \exp\{\psi(\mathbf{Z}_j; \boldsymbol{\beta})\} \Big] \Big). \tag{12}$$

An alternative expression is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Big(\psi(\mathbf{Z}_{(i)}; \boldsymbol{\beta}) - \log\Big[\sum_{j=1}^{n} Y_j(X_i) \exp\{\psi(\mathbf{Z}_j; \boldsymbol{\beta})\} \Big] \Big),$$

where $Y_j(t) = I(X_j \geq t)$ is the survival indicator on whether the j-th subject survives at the time t.

The above partial likelihood function is a profile likelihood and is derived from the full likelihood using the least informative nonparametric modeling for $\Lambda_0(\cdot)$, that is, $\Lambda_0(t)$ has a jump θ_i at t_i. $\diamond$

Let $\hat{\boldsymbol{\beta}}$ be the partial likelihood estimator of $\boldsymbol{\beta}$ maximizing (12) with respect to $\boldsymbol{\beta}$. By standard likelihood theory, it can be shown that (see for example Tsiatis 1981) the asymptotic distribution $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is multivariate normal with mean zero and a covariance matrix which may be estimated consistently by $(n^{-1} I(\hat{\boldsymbol{\beta}}))^{-1}$, where

$$I(\boldsymbol{\beta}) = \int_0^{\tau} \Big[\frac{S_2(\boldsymbol{\beta}, t)}{S_0(\boldsymbol{\beta}, t)} - \Big(\frac{S_1(\boldsymbol{\beta}, t)}{S_0(\boldsymbol{\beta}, t)} \Big)^{\otimes 2} \Big] \, dN(t)$$

and for $k = 0, 1, 2$,

$$S_k(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} Y_i(t) \psi'(\mathbf{Z}_i; \boldsymbol{\beta})^{\otimes k} \exp\{\psi(\mathbf{Z}_i; \boldsymbol{\beta})\},$$

where $N(t) = 1(X \leq t, \delta = 1)$, and $\mathbf{x}^{\otimes k} = 1, \mathbf{x}, \mathbf{x}\mathbf{x}^T$, respectively for $k = 0, 1$ and 2.

Since the baseline hazard Λ_0 does not appear in the partial likelihood, it is not estimable from the likelihood. There are several methods for estimating parameters related to Λ_0. One appealing estimate among them is the Breslow estimator (Breslow 1972, 1974)

$$\hat{\Lambda}_0(t) = \int_0^{\tau} \Big[\sum_{i=1}^{n} Y_i(s) \exp\{\mathbf{Z}_i^{\tau} \hat{\boldsymbol{\beta}}\} \Big]^{-1} \Big\{ \sum_{i=1}^{n} dN_i(s) \Big\}, \tag{13}$$

where $N_i(s) = 1(X_i \leq t, \delta_i = 1)$.

Hypothesis testing. After fitting the Cox model, one might be interested in checking if covariates really contribute to the risk function, for example, checking if the coefficient vector $\boldsymbol{\beta}$ is zero. More generally, one considers the hypothesis testing problem

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0.$$

From the asymptotic normality of the estimator $\hat{\boldsymbol{\beta}}$, it follows that the asymptotic null distribution of the Wald test statistic

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T I(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

is the chi-squared distribution with q degrees of freedom. Standard likelihood theory also suggests that the partial likelihood ratio test statistic

$$\lambda_{n1} = 2[\ell(\hat{\boldsymbol{\beta}}) - \ell(\boldsymbol{\beta}_0)] \tag{14}$$

and the score test statistic

$$T_n = U(\boldsymbol{\beta}_0)^T I^{-1}(\boldsymbol{\beta}_0) U(\boldsymbol{\beta}_0)$$

have the same asymptotic null distribution as the Wald statistic, where $U(\boldsymbol{\beta}_0) = \ell'(\boldsymbol{\beta}_0)$ is the score function (see for example, Andersen et al., 1993).

Cox's models with time-varying covariates. The Cox model (1) assumes that the hazard function for a subject depends on the values of the covariates and the baseline. Since the covariates are independent of time, the ratio of the hazard rate functions of two subjects is constant over time. Is this assumption reasonable?

Consider, for example, the case with age included in the study. Suppose we study survival time after heart transplantation. Then it is possible that age is a more critical factor of risk right after transplantation than a later time. Another example is given in Lawless (1982, page 393) with the amount of voltage as covariate which slowly increases over time until the electrical insulation fails. In this case, the impact of the covariate clearly depends on time. Therefore, the above assumption does not hold, and we have to analyze survival data with time-varying covariates.

Although the partial likelihood in (12) was derived for the setting of the Cox model with non-time-varying covariates, it can also be derived for the Cox model with time-varying covariates if one uses the counting process notation. For details, see marginal modeling of multivariate data using the Cox type of models in Section 3.1.

More about Cox's models. For the computational simplicity of the partial likelihood estimator, Cox's model has already been a useful case study for formal semiparametric estimation theory (Begun, Hall, Huang, and Wellner 1982; Bickel, Klaassen, Ritov, and Wellner 1993; Oakes 2002). Moreover, due to the derivation of the partial likelihood from profile likelihood (see Example 1), Cox's model has been considered as an approach to statistical science in the sense that "*it formulates scientific questions or quantities in terms of parameters γ in a model $f(y;\gamma)$ representing the underlying scientific mechanisms (Cox, 1997); partition the parameters $\gamma = (\theta, \eta)$ into a subset of interest θ and other nuisance parameters η necessary to complete the probability distribution (Cox and Hinkley, 1974); develops methods of inference about the scientific quantities that* *depend as little as possible upon the nuisance parameters (Barndorff-Nielsen and Cox, 1989); and thinks critically about the appropriate conditional distribution on which to base inferece*" (Zeger, Diggle, and Liang 2004).

Although Cox's models have driven a lot of statistical innovations in the past four decades, scientific fruit will continue to be born in the future. This motivates us to explore some recent development for Cox's models using the nonparametric idea and hope to open an avenue of academic research for interested readers.

2.1 Cox's models with unknown nonlinear risk functions

Misspecification of the risk function ψ may happen in the previous parametric form $\psi(\cdot, \boldsymbol{\beta})$, which could create a large modeling bias. To reduce the modeling bias, one considers nonparametric forms of ψ. Here we introduce such an attempt from Fan, Gijbels, and King (1997).

For easy exposition, we consider only the case with $q = 1$:

$$\lambda(t|z) = \lambda_0(t) \exp\{\psi(z)\}, \tag{15}$$

where z is one dimensional. Suppose the form of $\psi(z)$ in model (15) is not specified and the p^{th} order derivative of $\psi(z)$ at the point z exists. Then by the Taylor expansion,

$$\psi(Z) \approx \psi(z) + \psi'(z)(Z - z) + \cdots + \frac{\psi^{(p)}(z)}{p!}(Z - z)^p,$$

for Z in a neighborhood of z. Put

$$\tilde{\mathbf{Z}} = \{1, Z - z, \cdots, (Z - z)^p\}^\tau \text{ and } \tilde{\mathbf{Z}}_i = \{1, Z_i - z, \cdots, (Z_i - z)^p\}^\tau,$$

where τ denotes the transpose of a vector throughout this chapter. Let h be the bandwidth controlling the size of the neighborhood of x and K be a kernel function with compact support $[-1, 1]$ for weighting down the contribution of remote data points. Then for $|Z - z| \leq h$, as $h \to 0$,

$$\psi(Z) \approx \tilde{\mathbf{Z}}^\tau \boldsymbol{\alpha},$$

where

$$\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \cdots, \alpha_p)^\tau = \{\psi(z), \psi'(z), \cdots, \psi^{(p)}(z)/p!\}^\tau.$$

By using the above approximation and incorporating the localizing weights, the local (log) likelihood is obtained from (9) as

$$\begin{aligned} \ell_n(\boldsymbol{\beta}, \theta) &= n^{-1} \sum_{i=1}^{n} \Big[\delta_i \{\log \lambda_0(X_i; \theta) + \tilde{\mathbf{Z}}_i^\tau \boldsymbol{\alpha}\} \\ &\qquad - \lambda_0(X_i; \theta) \exp(\tilde{\mathbf{Z}}_i^\tau \boldsymbol{\alpha}) \Big] K_h(Z_i - x), \end{aligned} \tag{16}$$

where $K_h(t) = h^{-1}K(t/h)$. Then using the least-informative nonparametric model (7) for the baseline hazard and the same argument as for (12), we obtain the local log partial likelihood

$$\sum_{i=1}^{N} K_h(Z_{(i)} - z)\Big(\tilde{\mathbf{Z}}_{(i)}^{\tau}\boldsymbol{\alpha} - \log\Big[\sum_{j\in R_i} \exp\{\tilde{\mathbf{Z}}_{(j)}^{\tau}\boldsymbol{\alpha}\}K_h(Z_j - z)\Big]\Big). \tag{17}$$

Maximizing the above function with respect to $\boldsymbol{\alpha}$ leads to an estimate $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$. Note that the function value $\psi(z)$ is not directly estimable; (17) does not involve the intercept α_0 since it cancels out. The first component $\hat{\alpha}_1 = \hat{\psi}'(z)$ estimates $\psi'(z)$. It is evident from model (15) that $\psi(z)$ is only identifiable up to a constant. By imposing the condition $\psi(0) = 0$, the function $\psi(z)$ can be estimated by

$$\hat{\psi}(z) = \int_0^z \hat{\psi}'(t)\, dt.$$

According to Fan, Gijbels, and King (1997), under certain conditions, the following asymptotic normality holds for $\hat{\psi}'(z)$:

$$\sqrt{nh^3}\{\hat{\psi}'(z) - \psi'(z) - b_n(z)\} \overset{D}{\rightarrow} N(0, v_n^2(z)),$$

where

$$b_n(z) = \frac{1}{6}h^2 \int t^3 K_1^*(t)\, dt\psi^{(3)}(z)$$

and

$$v_n^2(z) = \sigma^2(z)f^{-1}(z)\int K_1^*(t)^2\, dt$$

with $K_1^*(t) = tK(t)/\int t^2K(t)\, dt$ and $\sigma^2(z) = E\{\delta|Z = z\}^{-1}$.

With the estimator of $\psi(\cdot)$, using the same argument as for (13), one can estimate the baseline hazard by

$$\hat{\Lambda}_0(t) = \int_0^{\tau} \Big[\sum_{i=1}^{n} Y_i(s)\exp\{\hat{\psi}(Z_i)\}\Big]^{-1}\Big\{\sum_{i=1}^{n} dN_i(s)\Big\}. \tag{18}$$

Inference problems associated with the resulting estimator include constructing confidence intervals and hypothesis tests, which can be solved via standard nonparametric techniques but to our knowledge no rigor mathematical theory exists in the literature. A possible test method can be developed along the line of the generalized likelihood ratio (*GLR*) tests in Section 5, and theoretical properties of the resulting tests are to be developed.

For multiple covariates cases, the above modeling method is applicable without any difficulty if one employs a multivariate kernel as in common nonparametric regression. See §2.2 for further details. However, a fully nonparametric specification of $\psi(\cdot)$ with large dimensionality q may cause the "curse of dimensionality" problem. This naturally leads us to consider some dimension reduction techniques.

2.2 Partly linear Cox's models

The partly linear Cox's model is proposed to alleviate the difficulty with a saturated specification of the risk function and takes the form

$$\lambda(t|\mathbf{z}) = \lambda_0(t)\Psi(\mathbf{z}_1, \mathbf{z}_2), \tag{19}$$

where $\lambda_0(\cdot)$ is an unspecified baseline hazard function and

$$\Psi(\mathbf{z}_1, \mathbf{z}_2) = \exp\{\psi_1(\mathbf{z}_1; \boldsymbol{\beta}) + \psi_2(\mathbf{z}_2)\},$$

where the form of the function $\psi_1(\mathbf{z}_1; \boldsymbol{\beta})$ is known up to an unknown vector of finite parameters $\boldsymbol{\beta}$, and $\psi_2(\cdot)$ is an unknown function. This model inherents nice interpretation of the finite parameter $\boldsymbol{\beta}$ in model (1) while modeling possible nonlinear effects of the $d \times 1$ vector of covariates $\mathbf{z}_2$. In particular, when there is no parametric component, the model reduces to the aforementioned full nonparametric model in §2.1. Hence, in practice, the number of components in $\mathbf{z}_2$ is small.

The parameters $\boldsymbol{\beta}$ and function $\psi_2(\mathbf{z}_2)$ can be estimated using the profile partial likelihood method. Specifically, as argued in the previous section, the function ψ_2 admits the linear approximation

$$\psi_2(\mathbf{Z}_2) \approx \psi_2(\mathbf{z}_2) + \psi_2'(\mathbf{z}_2)^\tau(\mathbf{Z}_2 - \mathbf{z}_2) \equiv \boldsymbol{\alpha}^\tau \tilde{\mathbf{Z}}_2$$

when $\mathbf{Z}_2$ is close to $\mathbf{z}_2$, where $\boldsymbol{\alpha} = \{\psi_2(\mathbf{z}_2), \psi_2'(\mathbf{z}_2)^\tau\}^\tau$ and $\tilde{\mathbf{Z}}_2 = \{1, (\mathbf{Z}_2 - \mathbf{z}_2)^\tau\}^\tau$. Given $\boldsymbol{\beta}$, we can estimate the function $\psi_2(\cdot)$ by maximizing the local partial likelihood

$$\begin{aligned} \ell_n(\alpha) \quad = \quad & \sum_{i=1}^{N} K_H(\mathbf{Z}_{2(i)} - \mathbf{z}_2)\Big(\psi_1'(\mathbf{Z}_{1(i)}; \boldsymbol{\beta}) + \tilde{\mathbf{Z}}_{2(i)}^\tau \boldsymbol{\alpha} \\ & - \log\Big[\sum_{j \in R_i} \exp\{\psi_1(\mathbf{Z}_{1(j)}; \boldsymbol{\beta}) + \tilde{\mathbf{Z}}_{2(j)}^\tau \boldsymbol{\alpha}\} K_H(\mathbf{Z}_{2j} - \mathbf{z}_2)\Big]\Big), \end{aligned} \tag{20}$$

where $K_H(\mathbf{z}_2) = |H|^{-1}K(H^{-1}\mathbf{z}_2)$ with $K(\cdot)$ being a d-variate probability density (the kernel) with unique mode 0 and $\int uK(u)du = 0$, and H is a nonsingular $d \times d$ matrix called the bandwidth matrix (see for example Jiang and Doksum 2003). For expressing the dependence of the resulting solution on $\boldsymbol{\beta}$, we denote it by $\hat{\boldsymbol{\alpha}}(\mathbf{z}_2; \boldsymbol{\beta}) = \{\hat{\psi}_2(\mathbf{z}_2; \boldsymbol{\beta}), \hat{\psi}_2'(\mathbf{z}_2; \boldsymbol{\beta})\}$. Substituting $\hat{\psi}_2(\cdot; \boldsymbol{\beta})$ into the partial likelihood, we obtain the profile partial likelihood of $\boldsymbol{\beta}$

$$\begin{aligned} \ell_n(\boldsymbol{\beta}) \quad = \quad & \sum_{i=1}^{n}\Big(\psi_1(\mathbf{Z}_{1(i)}; \boldsymbol{\beta}) + \hat{\psi}_2(\mathbf{Z}_{2(i)}; \boldsymbol{\beta}) \\ & - \log\Big[\sum_{j \in R_i} \exp\{\psi_1(\mathbf{Z}_{1j}; \boldsymbol{\beta}) + \hat{\psi}_2(\mathbf{Z}_{2j}; \boldsymbol{\beta})\}\Big]\Big). \end{aligned} \tag{21}$$

Maximizing (21) with respect to $\boldsymbol{\beta}$ will lead to an estimate of $\boldsymbol{\beta}$. We denote by $\hat{\boldsymbol{\beta}}$ the resulting estimate. The estimate of function $\psi_2(\cdot)$ is simply $\hat{\psi}_2(\cdot; \hat{\boldsymbol{\beta}})$.

By an argument similar to that in Cai, Fan, Jiang, and Zhou (2007), it can be shown that the profile partial likelihood estimation provides a root-n consistent estimator of $\boldsymbol{\beta}$ (see also Section 3). This allows us to estimate the nonparametric component ψ_2 as well as if the parameter $\boldsymbol{\beta}$ were known.

2.3 Partly linear additive Cox's models

The partly linear model (19) is useful for modeling failure time data with multiple covariates, but for high-dimensional covariate $\mathbf{z}_2$, it still suffers from the so-called "curse-of-dimensionality" problem in high-dimensional function estimation. One of the methods for attenuating this difficulty is to use the additive structure for the function $\psi_2(\cdot)$ as in Huang (1999), which leads to the partly linear additive Cox model. It specifies the conditional hazard of the failure time T given the covariate value $(\mathbf{z}, \mathbf{w})$ as

$$\lambda\{t|\mathbf{z}, \mathbf{w}\} \quad = \quad \lambda_0(t) \exp\{\psi(\mathbf{z}; \boldsymbol{\beta}) + \phi(\mathbf{w})\}, \tag{22}$$

where $\phi(\mathbf{w}) = \phi_1(w_1) + \cdots + \phi_J(w_J)$. The parameters of interest are the finite parameter vector $\boldsymbol{\beta}$ and the unknown functions ϕ_j's. The former measures the effect of the treatment variable vector $\mathbf{z}$, and the latter may be used to suggest a parametric structure of the risk. This model allows one to explore nonlinearity of certain covariates, avoids the "curse-of-dimensionality" problem inherent in the saturated multivariate semiparametric hazard regression model (19), and retains the nice interpretability of the traditional linear structure in Cox's model (Cox 1972) . See the discussions in Hastie and Tibshirani (1990).

Suppose that observed data for the ith subject is $\{X_i, \delta_i, \mathbf{W}_i, \mathbf{Z}_i\}$, where X_i is the observed event time for the ith subject, which is the minimum of the potential failure time T_i and the censoring time C_i, δ_i is the indicator of failure, and $\{\mathbf{Z}_i, \mathbf{W}_i\}$ is the vector of covariates. Then the log partial likelihood function for model (22) is

$$\ell(\boldsymbol{\beta}, \phi) = \sum_{i=1}^{n} \delta_i \Big\{ \psi(\mathbf{Z}_i; \boldsymbol{\beta}) + \phi(\mathbf{W}_i) - \log \sum_{j \in \mathcal{R}_i} r_j(\boldsymbol{\beta}, \phi) \Big\}, \tag{23}$$

where

$$r_j(\boldsymbol{\beta}, \phi) = \exp\{\psi(\mathbf{Z}_j; \boldsymbol{\beta}) + \phi(\mathbf{W}_j)\}.$$

Since the partial likelihood has no finite maximum over all parameters $(\boldsymbol{\beta}, \phi)$, it is impossible to use the maximum partial likelihood estimation for $(\boldsymbol{\beta}, \phi)$ without any restrictions on the function ϕ.

Now let us introduce the polynomial-spline based estimation method in Huang (1999). Assume that $\mathbf{W}$ takes values in $\mathcal{W} = [0, 1]^J$. Let

$$\underline{\xi} = \{0 = \xi_0 < \xi_1 < \cdots < \xi_K < \xi_{K+1} = 1\}$$

be a partition of $[0, 1]$ into K subintervals $I_{Ki} = [\xi_i, \xi_{i+1}),\ i = 0, \ldots, K-1$, and $I_{KK} = [\xi_K, \xi_{K+1}]$, where $K \equiv K_n = O(n^v)$ with $0 < v < 0.5$ being a positive integer such that

$$h \equiv \max_{1 \le k \le K+1} |\xi_k - \xi_{k-1}| = O(n^{-v}).$$

Let $\mathcal{S}(\ell, \underline{\xi})$ be the space of polynomial splines of degree $\ell \geq 1$ consisting of functions $s(\cdot)$ satisfying:

(i) the restriction of $s(\cdot)$ to I_{Ki} is a polynomial of order $\ell - 1$ for $1 \leq i \leq K$;

(ii) for $\ell \geq 2$, s is $\ell - 2$ times continuously differentiable on $[0, 1]$.

According to Schumaker (1981, page 124), there exists a local basis $B_i(\cdot)$, $1 \leq i \leq q_n$ for $\mathcal{S}(\ell, \underline{\xi})$ with $q_n = K_n + \ell$, such that for any $\phi_{nj}(\cdot) \in S(\ell, \underline{\xi})$,

$$\phi_{nj}(w_j) = \sum_{i=1}^{q_n} b_{ji} B_i(w_j), \quad 1 \leq j \leq J.$$

Put

$$B(w) = (B_1(w), \ldots, B_{q_n}(w))^\tau, \ \ \mathbf{B}(\mathbf{w}) = (B^\tau(w_1), \ldots, B^\tau(w_J))^\tau,$$

$$\mathbf{b}_j = (b_{j1}, \ldots, b_{jq_n})^\tau, \ \ \mathbf{b} = (\mathbf{b}_1^\tau, \ldots, \mathbf{b}_J^\tau)^\tau.$$

Then $\phi_{nj}(w_j) = \mathbf{b}_j^\tau B(w_j)$ and $\phi_n(\mathbf{w}) \equiv \sum_{j=1}^J \phi_{nj}(w_j) = \mathbf{b}^\tau \mathbf{B}(\mathbf{w})$. Under regular smoothness assumptions, ϕ_j's can be well approximated by functions in $\mathcal{S}(\ell, \underline{\xi})$. Therefore, by (23), we have the logarithm of an approximated partial likelihood

$$\ell(\beta, \mathbf{b}) = \sum_{i=1}^n \delta_i \Big\{ \psi(\mathbf{Z}_i; \boldsymbol{\beta}) + \phi_n(\mathbf{W}_i) - \log \sum_{j \in \mathcal{R}_i} \exp[\psi(\mathbf{Z}_j; \boldsymbol{\beta}) + \phi_n(\mathbf{W}_j)] \Big\}, \quad (24)$$

where

$$\phi_n(\mathbf{W}_i) = \sum_{j=1}^J \phi_{nj}(W_{ji})$$

with W_{ji} being the jth component of $\mathbf{W}_i$, for $i = 1 \ldots, n$. Let $(\hat{\beta}, \hat{\mathbf{b}})$ maximize the above partial likelihood (24). Then an estimator of $\phi(\cdot)$ at point $\mathbf{w}$ is simply the $\hat{\phi}(\mathbf{w}) = \sum_{j=1}^J \hat{\phi}_j(w_j)$ with $\hat{\phi}_j(w_j) = \hat{\mathbf{b}}_j^\tau B(w_j)$.

As shown in Huang (1999), when $\psi(\mathbf{z}; \boldsymbol{\beta}) = \mathbf{z}^\tau \boldsymbol{\beta}$, the estimator $\hat{\boldsymbol{\beta}}$ achieves $\sqrt{n}$-consistency. That is, under certain conditions,

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= n^{-1/2} I^{-1}(\boldsymbol{\beta}) \sum_{i=1}^n l^*_{\boldsymbol{\beta}}(X_i, \delta_i, \mathbf{Z}_i, \mathbf{W}_i) + o_p(1) \\ &\xrightarrow{d} N(0, I^{-1}(\boldsymbol{\beta})), \end{aligned}$$

where $I(\boldsymbol{\beta}) = E[l^*_{\boldsymbol{\beta}}(X, \Delta, \mathbf{Z}, \mathbf{W})]^{\otimes 2}$ is the information bound and

$$l^*_{\boldsymbol{\beta}}(X, \delta, \mathbf{Z}, \mathbf{W}) = \int_0^\tau (\mathbf{Z} - a^*(t) - h^*(\mathbf{W}))\, dM(t)$$

is the efficient score for estimation of $\boldsymbol{\beta}$ in model (22), where $h^*(\mathbf{w}) = h_1^*(w_1) + \cdots + h_J^*(w_J)$ and $(a^*, h_1^*, \ldots, h_J^*)$ is the unique L_2 functions that minimize

$$E\{\delta\|\mathbf{Z} - a(X) - h_1(W_1) - \cdots - h_J(W_J)\|^2\},$$

where

$$M(t) = \delta 1\{X \leq t\} - \int_0^t 1\{X \geq u\} \exp[\mathbf{Z}'\boldsymbol{\beta} + \phi(\mathbf{W})]\, d\Lambda_0(u)$$

is the usual counting process martingale.

Since the estimator, $\hat{\boldsymbol{\beta}}$, achieves the semiparametric information lower bound and is asymptotically linear, it is asymptotically efficient among all the regular estimators (see Bickel, Klaassen, Ritov, and Wellner 1993). However, the information lower bound cannot be consistently estimated, which makes inference for $\boldsymbol{\beta}$ difficult in practice. Further, the asymptotic distribution of the resulting estimator $\hat{\phi}$ is hard to derive. This makes it difficult to test if ϕ admits a certain parametric form.

The resulting estimates are easy to implement. Computationally, the maximization problem in (24) can be solved via the existing Cox regression program, for example **coxph** and **bs** in Splus software [for details, see Huang (1999)]. However, the number of parameters is large and numerical stability in implementation arises in computing the partial likelihood function. An alternative approach is to use the profile partial likelihood method as in Cai *et al.* (2007) (see also §3.2). The latter solves many much smaller local maximum likelihood estimation problems.

With the estimators of $\boldsymbol{\beta}$ and $\phi(\cdot)$, one can estimate the cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ by a Breslow's type of estimators:

$$\hat{\Lambda}_0(t) = \int_0^t \big[\sum_{i=1}^n Y_i(u) \exp\{\psi(\mathbf{Z}_i; \hat{\boldsymbol{\beta}}) + \hat{\phi}(\mathbf{W}_i)\}\big]^{-1} \sum_{i=1}^n dN_i(u),$$

where $Y_i(u) = 1(X_i \geq u)$ is the at-risk indicator and $N_i(u) = 1(X_i < u, \Delta_i = 1)$ is the associated counting process.

3 Multivariate Cox's Type of Models

The above Cox type of models are useful for modeling univariate survival data. However, multivariate survival data often arise from case-control family studies and other investigations where either two or more events occur for the same subject, or from identical events occurring to related subjects such as family members or classmates. Since failure times are correlated within cluster (subject or group), the independence of failure times assumption in univariate survival analysis is violated. Developing Cox's type of models to tackle with such kind of data is in need.

Three types of models are commonly used in the multivariate failure time literature: overall intensity process models, frailty models, and marginal hazard models. In general, the overall hazard models deal with the overall intensity, which is defined as the hazard rate given the history of the entire cluster (Andersen and Gill 1982). Interpretation of the parameters in an overall hazard model is conditioned on the failure and censoring information of every individual in the cluster. Consequently, most attention over the past two decades has been confined to marginal hazard models and frailty models. The frailty model considers the conditional hazard given the unobservable frailty random variables, which is particularly useful when the association of failure types within a subject is of interest (see Hougaard 2000). However, such models tend to be restrictive with respect to the types of dependence that can be modeled and model fitting is usually cumbersome. When the correlation among the observations is unknown or not of interest, the marginal hazard model approach which models the "population-averaged" covariate effects has been widely used (see Wei, Lin and Weissfeld 1989, Lee, Wei and Amato 1992, Liang, Self and Chang 1993, Lin 1994, Cai and Prentice 1995, Prentice and Hsu 1997, Spiekerman and Lin 1998, and Cai, Fan, Jiang, and Zhou 2007 among others).

Suppose that there are n subjects and for each subject there are J failure types. Let T_{ij} denote the potential failure time, C_{ij} the potential censoring time, $X_{ij} = min(T_{ij}, C_{ij})$ the observed time, and $\mathbf{Z}_{ij}$ the covariate vector for the j^{th} failure type of the i^{th} subject ($i = 1, \cdots, n; j = 1, \cdots, J$). Let Δ_{ij} be the indicator which equals 1 if X_{ij} is a failure time and 0 otherwise. Let $\mathcal{F}_{t,ij}$ represent the failure, censoring and covariate information for the jth failure type as well as the covariate information for the other failure types of the ith subject up to time t. The marginal hazard function is defined as

$$\lambda_{ij}(t) = h^{-1} \lim_{h \downarrow 0} P[t < T_{ij} \leq t + h | T_{ij} > t, \mathcal{F}_{t,ij}].$$

The censoring time is assumed to be independent of the failure time conditioning on the covariates.

There are various methods to model the marginal hazard rates of multivariate failure times. In general, different methods employ different marginal models. We here introduce the methods leading to nonparametric smoothing in our research papers.

3.1 Marginal modeling using Cox's models with linear risks

Failure rates differ in both baseline and coefficients. Wei, Lin and Weissfeld (1989) proposed a marginal modeling approach for multivariate data. Specifically, for the jth type of failure of the ith subject, they assume that the hazard function $\lambda_{ij}(t)$ takes the form

$$\lambda_{ij}(t) = \lambda_{0j}(t) \exp\{\boldsymbol{\beta}_j^{\tau} \mathbf{Z}_{ij}(t)\}, \tag{25}$$

where $\lambda_{0j}(t)$ an unspecified baseline hazard function and $\boldsymbol{\beta}_j = (\beta_{1j}, \cdots, \beta_{pj})'$ is the failure-specific regression parameter. Now, let $R_j(t) = \{l : X_{lj} \geq t\}$, that is, the set of subjects at risk just prior to time t with respect to the jth type of failure. Then the jth failure-specific partial likelihood (Cox 1972; Cox 1975) is

$$L_j(\boldsymbol{\beta}) = \prod_{i=1}^{n} \Big[\frac{\exp\{\boldsymbol{\beta}^\tau \mathbf{Z}_{ij}(X_{ij})\}}{\sum_{l \in R_j(X_{ij})} \exp\{\boldsymbol{\beta}^\tau \mathbf{Z}_{lj}(X_{ij})\}} \Big]^{\Delta_{ij}}; \tag{26}$$

see also (12). Note that only the terms $\Delta_{ij} = 1$ contribute to the product of (26). The maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}_j$ for $\boldsymbol{\beta}_j$ is defined as the solution to the score equation

$$\partial \log L_j(\boldsymbol{\beta}) / \partial \boldsymbol{\beta} = 0. \tag{27}$$

Using the counting process notation and the martingale theory, Wei, Lin and Weissfeld (1989) established the asymptotic properties of the estimates $\hat{\boldsymbol{\beta}}_j$'s, which show that the estimator $\hat{\boldsymbol{\beta}}_j$ is consistent for $\boldsymbol{\beta}_j$ and the estimators $\hat{\boldsymbol{\beta}}_j$'s are generally correlated. For readers' convenience, we summarize their argument in the following two examples. The employed approach to proving normality of the estimates is typical and can be used in other situations. Throughout the remainder of this chapter, for a column vector $\mathbf{a}$, we use $\mathbf{a}^{\otimes k}$ to denote 1, $\mathbf{a}$, and the matrix $\mathbf{a}\mathbf{a}'$, respectively for $k = 0, 1$, and 2.

Example 2 (*Score Equation in Counting Process Notation*). Let $N_{ij}(t) = 1\{X_{ij} \le t, \Delta_{ij} = 1\}$, $Y_{ij}(t) = 1\{X_{ij} \ge t\}$, and $M_{ij}(t) = N_{ij}(t) - \int_0^t Y_{ij}(u) \lambda_{ij}(u)\, du$. Then the log partial likelihood for the j^{th} type of failure evaluated at time t is

$$\ell_j(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} \int_0^t \boldsymbol{\beta}^\tau \mathbf{Z}_{ij}(u)\, dN_{ij}(u) - \int_0^t \log \Big[\sum_{i=1}^{n} Y_{ij}(u) \exp(\boldsymbol{\beta}^\tau \mathbf{Z}_{ij}(u)) \Big]\, d\bar{N}_j(u),$$

where $\bar{N}_j(u) = \sum_{i=1}^n N_{ij}(u)$. It is easy to see that the score equation (27) is

$$U_j(\boldsymbol{\beta}, t) = \sum_{i=1}^{n} \int_0^t \mathbf{Z}_{ij}(u)\, dN_{ij}(u) - \int_0^t S_j^{(1)}(\boldsymbol{\beta}, u) / S_j^{(0)}(\boldsymbol{\beta}, u)\, d\bar{N}_j(u) = 0, \tag{28}$$

where and thereafter for $k = 0, 1, 2$

$$S_j^{(k)}(\boldsymbol{\beta}, u) = n^{-1} \sum_{i=1}^{n} Y_{ij}(u) \mathbf{Z}_{ij}(u)^{\otimes k} \exp\{\boldsymbol{\beta}' \mathbf{Z}_{ij}(u)\}.$$

Example 3. (*Asymptotic Normality of the Estimators*). By (28),

$$U_j(\boldsymbol{\beta}_j, t) = \sum_{i=1}^{n} \int_0^t \mathbf{Z}_{ij}(u)\, dM_{ij}(u) - \int_0^t S_j^{(1)}(\boldsymbol{\beta}_j, u) / S_j^{(0)}(\boldsymbol{\beta}_j, u)\, d\bar{M}_j(u), \tag{29}$$

where $\bar{M}_j(u) = \sum_{i=1}^n M_{ij}(u)$. For $k = 0, 1$, let

$$s_j^{(k)}(\boldsymbol{\beta}, t) = E\Big[Y_{1j}(t) \mathbf{Z}_{1j}(t)^{\otimes k} \exp\{\boldsymbol{\beta}' \mathbf{Z}_{1j}(t)\} \Big].$$

Using the Taylor expansion of $U_j(\hat{\boldsymbol{\beta}}_j, \infty)$ around $\boldsymbol{\beta}$, one obtains that

$$n^{-1/2}U_j(\boldsymbol{\beta}_j, \infty) = \hat{A}_j(\boldsymbol{\beta}^*)\sqrt{n}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j),$$

where $\boldsymbol{\beta}^*$ is on the line segment between $\hat{\boldsymbol{\beta}}_j$ and $\boldsymbol{\beta}_j$, and

$$\hat{A}_j(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n} \Delta_{ij}\Big[\frac{S_j^{(2)}(\boldsymbol{\beta}, X_{ij})}{S_j^{(0)}(\boldsymbol{\beta}, X_{ij})} - \Big(\frac{S_j^{(1)}(\boldsymbol{\beta}, X_{ij})}{S_j^{(0)}(\boldsymbol{\beta}, X_{ij})}\Big)^{\otimes 2}\Big].$$

Note that for any $\boldsymbol{\beta}$,

$$n^{-1/2}\int_0^\infty \Big\{S_j^{(1)}(\boldsymbol{\beta}, u)/S_j^{(0)}(\boldsymbol{\beta}, u) - s_j^{(1)}(\boldsymbol{\beta}, u)/s_j^{(0)}(\boldsymbol{\beta}, u)\Big\}\, d\bar{M}_j(u) \to 0$$

in probability. It follows from (29) that

$$\begin{aligned} n^{-1/2}U_j(\boldsymbol{\beta}_j, \infty) \quad = \quad & n^{-1/2}\sum_{i=1}^{n}\int_0^\infty \Big\{\mathbf{Z}_{ij}(u)\, dM_{ij}(u) \\ & - \int_0^\infty \frac{s_j^{(1)}(\boldsymbol{\beta}_j, u)}{s_j^{(0)}(\boldsymbol{\beta}_j, u)}\, dM_{ij}(u)\Big\} + o_p(1), \qquad (30) \end{aligned}$$

which is asymptotically normal with mean zero. By the consistency of $\hat{A}_j(\boldsymbol{\beta})$ to a matrix $A_j(\boldsymbol{\beta})$ and by the asymptotic normality of $n^{-1/2}U_j(\boldsymbol{\beta}_j, \infty)$, one obtains that

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) \quad = \quad & A_j(\boldsymbol{\beta})^{-1}n^{-1/2}\sum_{i=1}^{n}\int_0^\infty \Big\{\mathbf{Z}_{ij}(u)\, dM_{ij}(u) \\ & - \int_0^t \frac{s_j^{(1)}(\boldsymbol{\beta}_j, u)}{s_j^{(0)}(\boldsymbol{\beta}_j, u)}\, dM_{ij}(u)\Big\} + o_p(1). \qquad (31) \end{aligned}$$

Then by the multivariate martingale central limit theorem, for large n, $(\hat{\boldsymbol{\beta}}_1^\tau, \cdots, \hat{\boldsymbol{\beta}}_J^\tau)^\tau$ is approximately normal with mean $(\boldsymbol{\beta}_1^\tau, \cdots, \boldsymbol{\beta}_J^\tau)^\tau$ and covariance matrix $D = (D_{jl})$, $j, l = 1, \ldots, J$, say. The asymptotic covariance matrix between $\sqrt{n}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l)$ is given by

$$D_{jl}(\boldsymbol{\beta}_j, \boldsymbol{\beta}_l) = A_j^{-1}(\boldsymbol{\beta}_j)E\{w_{j1}(\boldsymbol{\beta}_j)w_{l1}(\boldsymbol{\beta}_l)^\tau\}A_l^{-1}(\boldsymbol{\beta}_l),$$

where

$$w_{j1}(\boldsymbol{\beta}_j) = \int_0^\infty \{\mathbf{Z}_{1j}(t) - s_j^{(1)}(\boldsymbol{\beta}_j, t)/s_j^{(0)}(\boldsymbol{\beta}_j, t)\}\, dM_{1j}(t).$$

Wei, Lin and Weissfeld (1989) also gave a consistent empirical estimate of the covariance matrix D. This allows for simultaneous inference about the $\boldsymbol{\beta}_j$'s.

Failure rates differ only in the baseline. Lin (1994) proposed to model the j^{th} failure time using marginal Cox's model:

$$\lambda_{ij}(t) = \lambda_{0j}(t)\exp\{\boldsymbol{\beta}^\tau \mathbf{Z}_{ij}(t)\}. \tag{32}$$

For model (25), if the coefficients $\boldsymbol{\beta}_j$ are all equal to $\boldsymbol{\beta}$, then it reduces to model (32), and each $\hat{\boldsymbol{\beta}}_j$ is a consistent estimate of $\boldsymbol{\beta}$. Naturally, one can use a linear combination of the estimates,

$$\hat{\boldsymbol{\beta}}(\omega) = \sum_{j=1}^{J} \omega_j \hat{\boldsymbol{\beta}}_j \tag{33}$$

to estimate $\boldsymbol{\beta}$, where $\sum_{j=1}^{J} \omega_j = 1$. Using the above joint asymptotic normality of $\hat{\boldsymbol{\beta}}_j$'s, Wei, Lin and Weissfeld (1989) computed the variance of $\hat{\boldsymbol{\beta}}(\omega)$ and employed the weight $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_J)^\tau$ minimizing the variance. Specifically, let Σ be the covariance matrix of $(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_J)^\tau$. Then

$$\mathrm{Var}(\hat{\beta}(\boldsymbol{\omega})) = \boldsymbol{\omega}'\Sigma\boldsymbol{\omega}.$$

Using Langrange's multiplication method, one can find the optimal weight:

$$\hat{\boldsymbol{\omega}} = (\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1}\Sigma^{-1}\mathbf{1}.$$

If all of the observations for each failure type are independent, the partial likelihood for model (32) is (see Cox 1975)

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{j=1}^{J} L_j(\boldsymbol{\beta}) \\ &= \prod_{j=1}^{J}\prod_{i=1}^{n}\Big\{\frac{\exp\{\boldsymbol{\beta}^\tau \mathbf{Z}_{ij}\}}{\sum_{l\in R_j(X_{ij})}\exp\{\boldsymbol{\beta}^\tau \mathbf{Z}_{lj}\}}\Big\}^{\Delta_{ij}} \\ &= \prod_{j=1}^{J}\prod_{i=1}^{n}\Big\{\frac{\exp\{\boldsymbol{\beta}^\tau \mathbf{Z}_{ij}\}}{\sum_{l=1}^{n} Y_{lj}(X_{ij})\exp\{\boldsymbol{\beta}^\tau \mathbf{Z}_{lj}\}}\Big\}^{\Delta_{ij}}, \end{aligned} \tag{34}$$

where $L_j(\boldsymbol{\beta})$ is given by (26) and $Y_{lj}(t) = I(X_{lj} \geq t)$. Since the observations within a cluster are not necessarily independent, we refer to (34) as pseudo-partial likelihood. Note that

$$\log L(\boldsymbol{\beta}) = \sum_{j=1}^{J} \log L_j(\boldsymbol{\beta}), \quad \text{and} \quad \frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{j=1}^{J} \frac{\partial \log L_j(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}.$$

Therefore, the pseudo-partial likelihood merely aggregates J consistent estimation equations to yield a more powerful estimation equation without using any dependent structure.

Maximizing (34) leads to an estimator $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. We call this estimation method "pseudo-partial likelihood estimation". Following the argument in Example 3, it is easy to derive the asymptotic normality of $\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$. For large n and small J, Lin (1994) gave the covariance matrix estimation formula for $\tilde{\boldsymbol{\beta}}$. It is interesting to compare the efficiency of $\tilde{\boldsymbol{\beta}}$ with respect to $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\omega}})$, which is left as an exercise for interested readers.

3.2 Marginal modeling using Cox's models with nonlinear risks

The marginal Cox's models with linear risks provide a convenient tool for modeling the effects of covariates on the failure rate, but as we stressed in Section 2.1 they may yield large modeling bias if the underlying risk function is not linear. This motivated Cai, Fan, Zhou, and Zhou (2007) to study the following Cox model with a nonlinear risk:

$$\lambda_{ij}(t) = \lambda_{0j}(t)\exp\{\boldsymbol{\beta}(V_{ij}(t))^{\tau}\mathbf{Z}_{ij}(t) + g(V_{ij}(t))\}, \qquad (35)$$

where $\boldsymbol{\beta}(\cdot)$ is the regression coefficient vector that may be a function of the covariate V_{ij} , $g(\cdot)$ is an unknown nonlinear effect of V_{ij}. Model (35) is useful for modeling the nonlinear effect of V_{ij} and possible interaction between covariates V_{ij} and $\mathbf{Z}_{ij}$. A related work has been done in Cai and Sun (2003) using the time-varying coefficient Cox model for univariate data with $J = 1$.

Similar to (34), the pseudo partial likelihood for model (35) is

$$L(\boldsymbol{\beta}(\cdot), g(\cdot)) = \prod_{j=1}^{J}\prod_{i=1}^{n}\Big\{\frac{\exp\{\boldsymbol{\beta}(V_{ij})^{\tau}\mathbf{Z}_{ij} + g(V_{ij})\}}{\sum_{l\in R_j(X_{ij})}\exp\{\boldsymbol{\beta}(V_{lj})^{\tau}\mathbf{Z}_{lj} + g(V_{lj})\}}\Big\}^{\Delta_{ij}}. \qquad (36)$$

The pseudo-partial likelihood (34) can be regarded as parametric counterpart of (36). The log-pseudo partial likelihood is given by

$$\begin{aligned}\log L(\boldsymbol{\beta}(\cdot), g(\cdot)) \quad = \quad & \sum_{j=1}^{J}\sum_{i=1}^{n}\Delta_{ij}\Big\{\boldsymbol{\beta}(V_{ij})^{\tau}\mathbf{Z}_{ij} + g(V_{ij}) \\ & - \log\sum_{l\in R_j(X_{ij})}\exp\{\boldsymbol{\beta}(V_{lj})^{\tau}\mathbf{Z}_{lj} + g(V_{lj})\}\Big\}. \qquad (37)\end{aligned}$$

Assume that all functions in the components of $\boldsymbol{\beta}(\cdot)$ and $g(\cdot)$ are smooth so that they admit Taylor's expansions: for each given v and u, where u is close to v,

$$\begin{aligned}\boldsymbol{\beta}(u) \quad &\approx \quad \boldsymbol{\beta}(v) + \boldsymbol{\beta}'(v)(u - v) \equiv \boldsymbol{\delta} + \boldsymbol{\eta}(u - v), \\ g(u) \quad &\approx \quad g(v) + g'(v)(u - v) \equiv \alpha + \gamma(u - v). \qquad (38)\end{aligned}$$

Substituting these local models into (36), we obtain a similar local pseudo-partial likelihood to (17):

$$\ell(\boldsymbol{\xi}) \quad = \quad \sum_{j=1}^{J}\sum_{i=1}^{n}K_h(V_{ij} - v)\Delta_{ij}$$

$$\times\Big\{\boldsymbol{\xi}^\tau \mathbf{X}_{ij}^* - \log\Big(\sum_{l \in R_j(X_{ij})} \exp(\boldsymbol{\xi}^\tau \mathbf{X}_{lj}^*) K_h(V_{lj} - v)\Big)\Big\}, \qquad (39)$$

where $\boldsymbol{\xi} = (\boldsymbol{\delta}^\tau, \boldsymbol{\eta}^\tau, \gamma)^\tau$ and $\mathbf{X}_{ij}^* = (\mathbf{Z}_{ij}^\tau, \mathbf{Z}_{ij}^\tau(V_{ij} - v), (V_{ij} - v)))^\tau$. The kernel function is introduced to confine the fact that the local model (38) is only applied to the data around v. It gives a larger weight to the data closer to the point v.

Let $\hat{\boldsymbol{\xi}}(v) = (\boldsymbol{\delta}(v)^\tau, \boldsymbol{\eta}(v)^\tau, \hat{\gamma}(v))^\tau$ be the maximizer of (39). Then $\hat{\boldsymbol{\beta}}(v) = \hat{\boldsymbol{\delta}}(v)$ is a local linear estimator for the coefficient function $\boldsymbol{\beta}(\cdot)$ at the point v. Similarly, an estimator of $g'(\cdot)$ at the point v is simply the local slope $\hat{\gamma}(v)$, that is, the curve $g(\cdot)$ can be estimated by integration of the function $g'(v)$. Using the counting process theory incorporated with nonparametric regression techniques and the argument in Examples 2 and 3, Cai, Fan, Zhou, and Zhou (2007) derived asymptotic normality of the resulting pseudo-likelihood estimates

An alternative estimation approach is to fit a varying coefficient model for each failure type, that is, for event type j , to fit the model

$$\lambda_{ij}(t) = \lambda_{0j}(t) \exp\{\boldsymbol{\beta}_j(V_{ij}(t))^\tau \mathbf{Z}_{ij}(t) + g_j(V_{ij}(t))\}, \qquad (40)$$

resulting in $\hat{\boldsymbol{\xi}}_j(v)$ for estimating $\boldsymbol{\xi}_j(v) = (\boldsymbol{\beta}_j^\tau(v), \boldsymbol{\beta}_j'(v)^\tau, g_j'(v))^\tau$. Under model (35), we have $\boldsymbol{\xi}_1 = \boldsymbol{\xi}_2 = \cdots = \boldsymbol{\xi}_J$. Thus, as in (33), we can estimate $\boldsymbol{\xi}(v)$ by a linear combination

$$\hat{\boldsymbol{\xi}}(v; \boldsymbol{\omega}) = \sum_{j=1}^J \omega_j \boldsymbol{\xi}_j(v)$$

with $\sum_{j=1}^J \omega_j = 1$. The weights can be chosen in a similar way to (34). For details, see the reference above.

3.3 Marginal modeling using partly linear Cox's models

The fully nonparametric modeling of the risk function in the previous section is useful for building nonlinear effects of covariates on the failure rate, but it could lose efficiency if some covariates' effects are linear. To gain efficiency and to retain nice interpretation of the linear Cox models, Cai, Fan, Jiang, and Zhou (2007) studied the following marginal partly linear Cox model:

$$\lambda_{ij}(t) = \lambda_{0j}(t) \text{exp}[\boldsymbol{\beta}^\tau \mathbf{W}_{ij}(t) + g(Z_{ij}(t))], \qquad (41)$$

where $Z_{ij}(\cdot)$ is a main exposure variable of interest whose effect on the logarithm of the hazard might be non-linear; $\mathbf{W}_{ij}(\cdot) = (W_{ij1}(\cdot), \cdots, W_{ijq}(\cdot))^\tau$ is a vector of covariates that have linear effects; $\lambda_{0j}(\cdot)$ is an unspecified baseline hazard function; and $g(\cdot)$ is an unspecified smooth function. For d-dimensional variable $\mathbf{Z}_{ij}$, one can use an additive version $g(\mathbf{Z}) = g_1(Z_1) + \cdots + g(Z_d)$ to replace the above function $g(\cdot)$ for alleviating the difficulty with curse of dimensionality.

Like model (32), model (41) allows a different set of covariates for different failure types of the subject. It also allows for a different baseline hazard function for different failure types of the subject. It is useful when the failure types in a subject have different susceptibilities to failures. Compared with model (32), model (41) has an additional nonlinear term in the risk function. A related class of marginal models is given by restricting the baseline hazard functions in (41) to be common for all the failure types within a subject, i.e.,

$$\lambda_{ij}(t) = \lambda_0(t)\exp[\boldsymbol{\beta}^\tau \mathbf{W}_{ij}(t) + g(Z_{ij}(t))]. \tag{42}$$

While this model is more restrictive, the common baseline hazard model (42) leads to more efficient estimation when the baseline hazards are indeed the same for all the failure types within a subject. Model (42) is very useful for modeling clustered failure time data where subjects within clusters are exchangeable.

Denote by $\mathcal{R}_j(t) = \{i : X_{ij} \geq t\}$ the set of subjects at risk just prior to time t for failure type j. If failure times from the same subject were independent, then the logarithm of the pseudo partial likelihood for (41) is (see Cox 1975)

$$\ell(\boldsymbol{\beta}, g(\cdot)) = \sum_{j=1}^{J}\sum_{i=1}^{n} \Delta_{ij}\big\{\boldsymbol{\beta}^\tau \mathbf{W}_{ij}(X_{ij}) + g(Z_{ij}(X_{ij})) - R_{ij}(\boldsymbol{\beta}, g)\big\}, \tag{43}$$

where $R_{ij}(\boldsymbol{\beta}, g) = \log\Big(\sum_{l\in\mathcal{R}_j(X_{ij})} \exp[\boldsymbol{\beta}^\tau \mathbf{W}_{lj}(X_{ij}) + g(Z_{lj}(X_{ij}))]\Big)$. The pseudo partial likelihood estimation is robust against the mis-specification of correlations among failure times, since we neither require that the failure times are independent nor specify a dependence structure among failure times.

Assume that $g(\cdot)$ is smooth so that it can be approximated locally by a polynomial of order p. For any given point z_0, by Taylor's expansion,

$$g(z) \approx g(z_0) + \sum_{k=1}^{p} \frac{g^{(k)}(z_0)}{k!}(z - z_0)^k \equiv \alpha + \boldsymbol{\gamma}^\tau \tilde{\mathbf{Z}}, \tag{44}$$

where $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_p)^\tau$ and $\tilde{\mathbf{Z}} = \{z - z_0, \cdots, (z - z_0)^p\}^\tau$. Using the local model (44) for the data around z_0 and noting that the local intercept α cancels in (43), we obtain a similar version of the logarithm of the local pseudo-partial likelihood in (17):

$$\begin{aligned}\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}) \;&=\; \sum_{j=1}^{J}\sum_{i=1}^{n} K_h(Z_{ij}(X_{ij}) - z_0)\Delta_{ij} \\ &\quad \times\big[\boldsymbol{\beta}^\tau \mathbf{W}_{ij}(X_{ij}) + \boldsymbol{\gamma}^\tau \tilde{\mathbf{Z}}_{ij}(X_{ij}) - R^*_{ij}(\boldsymbol{\beta}, \boldsymbol{\gamma})\big],\end{aligned} \tag{45}$$

where

$$R^*_{ij}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \log\Big(\sum_{l\in\mathcal{R}_j(X_{ij})} \exp[\boldsymbol{\beta}^\tau \mathbf{W}_{lj}(X_{ij}) + \boldsymbol{\gamma}^\tau \tilde{\mathbf{Z}}_{lj}(X_{ij})]K_h(Z_{lj}(X_{ij}) - z_0)\Big),$$

and $\tilde{\mathbf{Z}}_{ij}(u) = \{Z_{ij}(u) - z_0, \cdots, (Z_{ij}(u) - z_0)^p\}^\tau$.

Let $(\hat{\boldsymbol{\beta}}(z_0), \hat{\boldsymbol{\gamma}}(z_0))$ maximize the local pseudo-partial likelihood (45). Then, an estimator of $g'(\cdot)$ at the point z_0 is simply the first component of $\hat{\boldsymbol{\gamma}}(z_0)$, namely $\hat{g'}(z_0) = \hat{\gamma}_1(z_0)$. The curve $\hat{g}$ can be estimated by integration on the function $\hat{g'}(z_0)$ using the trapezoidal rule by Hastie and Tibshirani (1990). To assure the identifiability of $g(\cdot)$, one can set $g(0) = 0$ without loss of generality.

Since only the local data are used in the estimation of $\boldsymbol{\beta}$, the resulting estimator for $\boldsymbol{\beta}$ cannot be root-n consistent. Cai, Fan, Jiang, and Zhou (2007) referred to $(\hat{\boldsymbol{\beta}}(z_0), \hat{\boldsymbol{\gamma}}(z_0))$ as the naive estimator and proposed a profile likelihood based estimation method to fix the drawbacks of the naive estimator. Now let us introduce this method.

For a given $\boldsymbol{\beta}$, we obtain an estimator $\hat{g}^{(k)}(\cdot, \boldsymbol{\beta})$ of $g^{(k)}(\cdot)$, and hence $\hat{g}(\cdot, \boldsymbol{\beta})$, by maximizing (45) with respect to $\boldsymbol{\gamma}$. Denote by $\hat{\boldsymbol{\gamma}}(z_0, \boldsymbol{\beta})$ the maximizer. Substituting the estimator $\hat{g}(\cdot, \boldsymbol{\beta})$ into (43), one can obtain the logarithm of the profile pseudo-partial likelihood:

$$\begin{aligned} \ell_p(\boldsymbol{\beta}) \quad = \quad & \sum_{j=1}^{J} \sum_{i=1}^{n} \Delta_{ij} \Big\{ \beta^\tau \mathbf{W}_{ij} + \hat{g}(Z_{ij}, \boldsymbol{\beta}) \\ & - \log\Big(\sum_{l \in \mathcal{R}_j(X_{ij})} \exp[\boldsymbol{\beta}^\tau \mathbf{W}_{lj} + \hat{g}(Z_{lj}, \boldsymbol{\beta})] \Big) \Big\}. \end{aligned} \qquad (46)$$

Let $\hat{\boldsymbol{\beta}}$ maximize (46) and $\hat{\boldsymbol{\gamma}} = \hat{\boldsymbol{\gamma}}(z_0, \hat{\boldsymbol{\beta}})$. Then the proposed estimator for the parametric component is simply $\hat{\boldsymbol{\beta}}$ and for the nonparametric component is $\hat{g}(\cdot) = \hat{g}(\cdot, \hat{\boldsymbol{\beta}})$.

Maximizing (46) is challenging since the function form $\hat{g}(\cdot, \boldsymbol{\beta})$ is implicit. The objective function $\ell_p(\cdot)$ is non-concave. One possible way is to use the backfitting algorithm, which iteratively optimizes (45) and (46). More precisely, given $\boldsymbol{\beta}_0$, optimize (45) to obtain $\hat{g}(\cdot, \boldsymbol{\beta}_0)$. Now, given $\hat{g}(\cdot, \boldsymbol{\beta}_0)$, optimize (46) with respect to $\boldsymbol{\beta}$ by fixing the value of $\boldsymbol{\beta}$ in $\hat{g}(\cdot, \boldsymbol{\beta})$ as $\boldsymbol{\beta}_0$, and iterate this until convergence. An alternative approach is to optimize (46) by using the Newton-Raphson method, but ignore the computation of $\frac{\partial^2}{\partial \boldsymbol{\beta}^2} \hat{g}(\cdot, \boldsymbol{\beta})$, i.e. setting it to zero in computing the Newton-Raphson updating step.

As shown in Cai, Fan, Jiang, and Zhou (2007), the resulting estimator $\hat{\boldsymbol{\beta}}$ is root-n consistent and its asymptotic variance admits a sandwich formula, which leads to a consistent variance estimation for $\hat{\boldsymbol{\beta}}$. This furnishes a practical inference tool for the parameter $\boldsymbol{\beta}$. Since $\hat{\boldsymbol{\beta}}$ is root-n consistent, it does not affect the estimator of the nonparametric component g. If the covariates $(\mathbf{W}_{1j}^\tau, Z_{1j})^\tau$ for different j are identically distributed, then the resulting estimate $\hat{g}$ has the same distribution as the estimate in Section 2.1. That is, even though the failure types within subjects are correlated, the profile likelihood estimator of $g(\cdot)$ performs as well as if they were independent. Similar phenomena were also discovered in nonparametric regression models (see Masry and Fan 1997; Jiang and Mack 2001).

With the estimators of $\boldsymbol{\beta}$ and $g(\cdot)$, one can estimate the cumulative baseline hazard function $\Lambda_{0j}(t) = \int_0^t \lambda_{0j}(u)du$ under mild conditions by a consistent estimator:

$$\hat{\Lambda}_{0j}(t) = \int_0^t \big[\sum_{i=1}^n Y_{ij}(u)\exp\{\hat{\boldsymbol{\beta}}^\tau \mathbf{W}_{ij}(u) + \hat{g}(Z_{ij}(u))\}\big]^{-1} \sum_{i=1}^n dN_{ij}(u), \qquad (47)$$

where $Y_{ij}(u) = 1(X_{ij} \geq u)$ is the at-risk indicator and $N_{ij}(u) = 1(X_{ij} \leq u, \Delta_{ij} = 1)$ is the associated counting process.

3.4 Marginal modeling using partly linear Cox's models with varying coefficients

The model (41) is useful for modeling nonlinear covariate effects, but it cannot deal with possible interaction between covariates. This motivated Cai, Fan, Jiang, and Zhou (2008) to consider the following partly linear Cox model with varying coefficients:

$$\lambda_{ij}(t) = \lambda_{0j}(t)\text{exp}\{\boldsymbol{\beta}^\tau \mathbf{W}_{ij}(t) + \boldsymbol{\alpha}(V_{ij}(t))^\tau \mathbf{Z}_{ij}(t)\}, \qquad (48)$$

where $\mathbf{W}_{ij}(\cdot) = (W_{ij1}(\cdot), \cdots, W_{ijq}(\cdot))^\tau$ is a vector of covariates that has linear effects on the logarithm of the hazard, $\mathbf{Z}_{ij}(\cdot) = (Z_{ij1}(\cdot), \cdots, Z_{ijp}(\cdot))^\tau$ is a vector of covariates that may interact with some exposure covariate $V_{ij}(\cdot)$; $\lambda_{0j}(\cdot)$ is an unspecified baseline hazard function; and $\alpha(\cdot)$ is a vector of unspecified coefficient functions. Model (48) is useful for capturing nonlinear interaction between covariates V and $\mathbf{Z}$. This kind of phenomenon often happens in practice. For example, in the aforementioned FHS study, V would represent the calendar year of birthdate, $\mathbf{W}$ would consist of confounding variables such as gender, blood pressure, cholesterol level and smoking status, etc, and $\mathbf{Z}$ would contain covariates possibly interacting with V such as the body mass index (BMI). In this example, one needs to model possible complex interaction between the BMI and the birth cohort.

As before we use $\mathcal{R}_j(t) = \{i : X_{ij} \geq t\}$ to denote the set of the individuals at risk just prior to time t for failure type j. If failure times from the same subject were independent, then the partial likelihood for (48) is

$$L(\boldsymbol{\beta}, \alpha) = \prod_{j=1}^J \prod_{i=1}^n \left\{ \frac{\exp\{\boldsymbol{\beta}^\tau \mathbf{W}_{ij}(X_{ij}) + \boldsymbol{\alpha}(V_{ij}(X_{ij}))^\tau \mathbf{Z}_{ij}(X_{ij})\}}{\sum_{l \in \mathcal{R}_j(X_{ij})} \exp\{\boldsymbol{\beta}^\tau \mathbf{W}_{lj}(X_{ij}) + \boldsymbol{\alpha}(V_{lj}(X_{ij}))^\tau \mathbf{Z}_{lj}(X_{ij})\}} \right\}^{\Delta_{ij}}. \quad (49)$$

For the case with $J = 1$, if the coefficient functions are constant, the partial likelihood above is just the one in Cox's model (Cox 1972). Since failure times from the same subject are dependent, the above partial likelihood is actually again a pseudo-partial likelihood.

Assume that $\boldsymbol{\alpha}(\cdot)$ is smooth so that it can be approximated locally by a linear function. Denote by $f_j(\cdot)$ the density of V_{1j}. For any given point $v_0 \in \cup_{j=1}^J \text{supp}(f_j)$, where $\text{supp}(f_j)$ denotes the support of $f_j(\cdot)$, by Taylor's expansion,

$$\boldsymbol{\alpha}(v) \approx \boldsymbol{\alpha}(v_0) + \boldsymbol{\alpha}'(v_0)(v - v_0) \equiv \boldsymbol{\delta} + \boldsymbol{\eta}(v - v_0). \qquad (50)$$

Using the local model (50) for the data around v_0, we obtain the logarithm of the local pseudo-partial likelihood [see also (17)]:

$$\begin{aligned}\ell(\boldsymbol{\beta},\boldsymbol{\gamma}) &= \sum_{j=1}^{J}\sum_{i=1}^{n} K_h(V_{ij}(X_{ij})-v_0)\Delta_{ij} \\ &\quad \times\big\{\boldsymbol{\beta}^T\mathbf{W}_{ij}(X_{ij})+\boldsymbol{\gamma}^T\mathbf{U}_{ij}(X_{ij},v_0)-R^*_{ij}(\boldsymbol{\beta},\boldsymbol{\gamma})\big\}, \qquad (51)\end{aligned}$$

where $\mathbf{U}_{ij}(u,v_0)=\{\mathbf{Z}_{ij}(u)^T,\mathbf{Z}_{ij}(u)^T(V_{ij}(u)-v_0)\}^T$, $\boldsymbol{\gamma}=(\boldsymbol{\delta}^T,\boldsymbol{\eta}^T)^T$ and

$$R^*_{ij}(\boldsymbol{\beta},\boldsymbol{\gamma})=\log\big(\sum_{l\in\mathcal{R}_j(X_{ij})}\exp[\boldsymbol{\beta}^T\mathbf{W}_{lj}(X_{ij})+\boldsymbol{\gamma}^T\mathbf{U}_{lj}(X_{ij},v_0)]K_h(V_{lj}(X_{ij})-v_0)\big).$$

Let $(\hat{\boldsymbol{\beta}}(v_0),\hat{\boldsymbol{\gamma}}(v_0))$ maximize the local pseudo-partial likelihood (51). Then, an estimator of $\boldsymbol{\alpha}(\cdot)$ at the point v_0 is simply the local intercept $\hat{\boldsymbol{\delta}}(v_0)$, namely $\hat{\boldsymbol{\alpha}}(v_0)=\hat{\boldsymbol{\delta}}(v_0)$. When v_0 varies over a grid of prescribed points, the estimates of the functions are obtained. Since only the local data are used in the estimation of $\boldsymbol{\beta}$, the resulting estimator for $\boldsymbol{\beta}$ cannot be $\sqrt{n}$-consistent. Let us refer to $(\hat{\boldsymbol{\beta}}(v_0),\hat{\alpha}(v_0))$ as a naive estimator.

To enhance efficiency of estimation, Cai, Fan, Jiang and Zhou (2008) studied a profile likelihood similar to (46). Specifically, for a given $\boldsymbol{\beta}$, they obtained an estimator of $\hat{\boldsymbol{\alpha}}(\cdot,\boldsymbol{\beta})$ by maximizing (51) with respect to $\boldsymbol{\gamma}$. Substituting the estimator $\hat{\boldsymbol{\alpha}}(\cdot,\boldsymbol{\beta})$ into (49), they obtained the logarithm of the profile pseudo-partial likelihood:

$$\begin{aligned}\ell_p(\boldsymbol{\beta}) &= \sum_{j=1}^{J}\sum_{i=1}^{n}\Delta_{ij}\Big\{\boldsymbol{\beta}^T\mathbf{W}_{ij}+\hat{\boldsymbol{\alpha}}(V_{ij},\boldsymbol{\beta})^T\mathbf{Z}_{ij} \\ &\quad -\log\Big(\sum_{l\in\mathcal{R}_j(X_{ij})}\exp[\boldsymbol{\beta}^T\mathbf{W}_{lj}+\hat{\boldsymbol{\alpha}}(V_{lj},\boldsymbol{\beta})^T\mathbf{Z}_{lj}]\Big)\Big\}. \qquad (52)\end{aligned}$$

Let $\hat{\boldsymbol{\beta}}$ maximize (52). The final estimator for the parametric component is simply $\hat{\boldsymbol{\beta}}$ and for the coefficient function is $\hat{\boldsymbol{\alpha}}(\cdot)=\hat{\boldsymbol{\alpha}}(\cdot,\hat{\boldsymbol{\beta}})$. The idea in §3.3 can be used to compute the profile pseudo-partial likelihood estimator.

The resulting estimator $\hat{\boldsymbol{\beta}}$ is root-n consistent and its asymptotic variance admits a sandwich formula, which leads to a consistent variance estimation for $\hat{\boldsymbol{\beta}}$. Since $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent, it does not affect the estimator of the nonparametric component $\boldsymbol{\alpha}$. If the covariates $(\mathbf{W}^\tau_{1j},Z_{1j})^\tau$ for different j are identically distributed, then even though the failure types within subjects are correlated, the profile likelihood estimator of $\boldsymbol{\alpha}(\cdot)$ performs as well as if they were independent [see Cai, Fan, Jiang, and Zhou (2008)].

With the estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}(\cdot)$, one can estimate the cumulative baseline hazard function $\Lambda_{0j}(t)=\int_0^t\lambda_{0j}(u)du$ by a consistent estimator:

$$\hat{\Lambda}_{0j}(t)=\int_0^t\Big[\sum_{i=1}^{n}Y_{ij}(u)\exp\{\hat{\boldsymbol{\beta}}^T\mathbf{W}_{ij}(u)+\hat{\boldsymbol{\alpha}}(V_{ij}(u))^T\mathbf{Z}_{ij}(u)\}\Big]^{-1}\sum_{i=1}^{n}dN_{ij}(u),$$

where $Y_{ij}(\cdot)$ and $N_{ij}(u)$ are the same in §3.3.

4 Model Selection on Cox's Models

For Cox's type of models, different estimation methods have introduced for estimating the unknown parameters/functions. However, when there are many covariates, one has to face up to the variable selection problems.

Different variable selection techniques in linear regression models have been extended to the Cox model. Examples include the LASSO variable selector in Tibshirani (1997), the Bayesian variable selection method in Faraggi and Simon (1998), the nonconcave penalised likelihood approach in Fan and Li (2002), the penalised partial likelihood with a quadratic penalty in Huang and Harrington (2002), and the extended BIC-type variable selection criteria in Bunea and McKeague (2005).

In the following we introduce a model selection approach from Cai, Fan, Li, and Zhou (2005). It is a penalised pseudo-partial likelihood method for variable selection with multivariate failure time data with a growing number of regression coefficients. Any model selection method should ideally achieve two targets: to efficiently estimate the parameters and to correctly select the variables. The penalised pseudo-partial likelihood method integrates them together. This kind of idea appears in Fan & Li (2001, 2002).

Suppose that there are n independent clusters and that each cluster has K_i subjects. For each subject, J types of failure may occur. Let T_{ijk} denote the potential failure time, C_{ijk} the potential censoring time, $X_{ijk} = min(T_{ijk}, C_{ijk})$ the observed time, and $\mathbf{Z}_{ijk}$ the covariate vector for the j^{th} failure type of the k^{th} subject in the ith cluster. Let Δ_{ijk} be the indicator which equals 1 if X_{ijk} is a failure time and 0 otherwise. For the failure time in the case of the j^{th} type of failure on subject k in cluster i, the marginal hazards model is taken as

$$\lambda_{ijk}\{t|\mathbf{Z}_{ijk}(t)\} = \lambda_{0j}(t)\exp\{\boldsymbol{\beta}^T\mathbf{Z}_{ijk}(t)\}, \tag{53}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{d_n})^T$ is a vector of unknown regression coefficients, d_n is the dimension of $\boldsymbol{\beta}$, $\mathbf{Z}_{ijk}(t)$ is a possibly external time-dependent covariate vector, and $\lambda_{0j}(t)$ are unspecified baseline hazard functions.

Similar to (34), the logarithm of a pseudo-partial likelihood function for model (53) is

$$\ell(\boldsymbol{\beta}) = \sum_{j=1}^{J}\sum_{i=1}^{n}\sum_{k=1}^{K_i}\Delta_{ijk}\Big(\boldsymbol{\beta}^T\mathbf{Z}_{ijk}(X_{ijk}) - R(\boldsymbol{\beta})\Big), \tag{54}$$

where $R(\boldsymbol{\beta}) = \log\Big[\sum_{l=1}^{n}\sum_{g=1}^{K_i} Y_{ljg}(X_{ijk})\exp\{\boldsymbol{\beta}^T\mathbf{Z}_{ljg}(X_{ijk})\}\Big]$ and $Y_{ljg}(t) = I(X_{ljg} \geq t)$ is the survival indicator on whether the g^{th} subject in the l^{th} cluster surviving at time t. To balance modeling bias and estimation variance, many traditional variable selection criteria have resorted to the use of penalised likelihood, including the AIC (Akaike, 1973) and BIC (Schwarz, 1978). The penalised pseudo-partial likelihood for model (53) is defined as

$$L(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - n\sum_{j=1}^{d_n} p_{\lambda_j}(|\beta_j|), \tag{55}$$

where $p_{\lambda_j}(|\beta_j|)$ is a given nonnegative function called a penalty function with λ_j as a regularisation or tuning parameter. The tuning parameters can be chosen subjectively by data analysts or objectively by data themselves. In general, large values of λ_j result in simpler models with fewer selected variables. When $K_i = 1$, $J = 1$, $d_n = d$, and $\lambda_j = \lambda$, it reduces to the penalized partial likelihood in Fan and Li (2002).

Many classical variable selection criteria are special cases of (55). An example is the L_0 penalty (or entropy penalty)

$$p_\lambda(|\theta|) = 0.5\lambda^2 1(|\theta| \neq 0).$$

In this case, the penalty term in (55) is merely $0.5n\lambda^2 k$, with k being the number of variables that are selected. Given k, the best fit to (55) is the subset of k variables having the largest likelihood $\ell(\boldsymbol{\beta})$ among all subsets of k variables. In other words, the method corresponds to the best subset selection. The number of variables depends on the choice of λ. The AIC (Akaike, 1973), BIC (Schwarz, 1978), ϕ-criterion (Shibata, 1984), and RIC (Foster & George, 1994) correspond to

$$\lambda = (2/n)^{1/2},\ \{\log(n)/n\}^{1/2},\ [\log\{\log(n)\}]^{1/2},\ \text{and}\ \{\log(d_n)/n\}^{1/2},$$

respectively. Since the entropy penalty function is discontinuous, one requires to search over all possible subsets to maximise (55). Hence it is very expensive computationally. Furthermore, as analysed by Breiman (1996), best-subset variable selection suffers from several drawbacks, including its lack of stability.

There are several choices for continuous penalty functions. The L_1 penalty, defined by $p_\lambda(|\theta|) = \lambda|\theta|$, results in the LASSO variable selector (Tibshirani, 1996). The smoothly clipped absolute deviation (SCAD) penalty, defined by

$$p'_\lambda(\theta) = \lambda I(|\theta| \leq \lambda) + \frac{(a\lambda - \theta)_+}{a-1} I(\theta > \lambda), \tag{56}$$

for some $a > 2$ and $\lambda > 0$, with $p_\lambda(0) = 0$. Fan and Li (2001) recommended $a = 3.7$ based on a risk optimization consideration. This penalty improves the entropy penalty function by saving computational cost and resulting in a continuous solution to avoid unnecessary modelling variation. Furthermore, it improves the L_1 penalty by avoiding excessive estimation bias.

The penalised pseudo-partial likelihood estimator, denoted by $\hat{\boldsymbol{\beta}}$, maximises (55). For certain penalty functions, such as the L_1 penalty and the SCAD penalty, maximising $L(\boldsymbol{\beta})$ will result in some vanishing estimates of coefficients and make their associated variables be deleted. Hence, by maximising $L(\boldsymbol{\beta})$, one selects a model and estimates its parameters simultaneously.

Denote by β_0 the true value of β with the nonzero and zero components β_{10} and β_{20}. To emphasize the dependence of λ_j on the sample size n, λ_j is written as λ_{jn}. Let s_n be the dimension of β_{10},

$$a_n = \max_{1\leq j\leq s_n}\{|p'_{\lambda_{jn}}| : \beta_{j0} \neq 0\},\ \text{and}\ b_n = \max_{1\leq j\leq s_n}\{|p''_{\lambda_{jn}}| : \beta_{j0} \neq 0\}.$$

As shown in Cai, Fan, Li, and Zhou (2005), under certain conditions, if $a_n \to 0$, $b_n \to 0$ and $d_n^4/n \to 0$, as $n \to \infty$, then with probability tending to one, there exists a local maximizer $\hat{\beta}$ of $L(\beta)$, such that

$$\|\hat{\beta} - \beta_0\| = O_p(\sqrt{d_n}(n^{-1/2} + a_n)).$$

Furthermore, if $\lambda_{jn} \to 0$, $\sqrt{n/d_n}\lambda_{jn} \to \infty$, and $a_n = O(n^{-1/2})$, then with probability tending to 1, the above consistent local maximizer $\hat{\beta} = (\hat{\beta}_1^\tau, \hat{\beta}_2^\tau)^\tau$ must be such that

(i) $\hat{\beta}_2 = 0$ and

(ii) for any nonzero constant $s_n \times 1$ vector c_n with $c_n^\tau c_n = 1$,

$$\sqrt{n} c_n^\tau \Gamma_{11}^{-1/2}(A_{11} + \Sigma)\{\hat{\beta}_1 - \beta_{10} + (A_{11} + \Sigma)^{-1} b\} \xrightarrow{\mathcal{D}} N(0, 1),$$

where A_{11} and Γ_{11} consist of the first s_n columns and rows of $A(\beta_{10}, 0)$ and $\Gamma(\beta_{10}, 0)$, respectively (see the aforementioned paper for details of notation here).

The above result demonstrates that the resulting estimators have the oracle property. For example, with the SCAD penalty, we have $a_n = 0$, $b = 0$ and $\Sigma = 0$ for sufficiently large n. Hence, by the above result,

$$\sqrt{n} c_n^\tau \Gamma_{11}^{-1/2} A_{11}(\hat{\beta}_1 - \beta_{10}) \xrightarrow{\mathcal{D}} N(0, 1).$$

The estimator $\hat{\beta}_1$ shares the same sampling property as the oracle estimator. Furthermore, $\hat{\beta}_2 = 0$ is the same as the oracle estimator that knows in advance that $\beta_2 = 0$. In other words, the resulting estimator can correctly identify the true model, as if it were known in advance.

Further study in this area includes extending the above model selection method to other Cox's type of models, such as the partly linear models in Sections 2.3, 3.3 and 3.4.

5 Validating Cox's Type of Models

Even though different Cox's type of models are useful for exploring the complicate association of covariates with failure rates, there is a risk that misspecification of a working Cox model can create large modeling bias and lead to wrong conclusions and erroneous forecasting. It is important to check whether certain Cox's models fit well a given data set.

In parametric hypothesis testing, the most frequently used method is the likelihood ratio inference. It compares the likelihoods under the null and alternative models. See for example the likelihood ratio statistic in (14). The likelihood ratio tests are widely used in the theory and practice of statistics. An important fundamental property of the likelihood ratio tests is that their asymptotic null distributions are independent of nuisance parameters in the null model. It is natural to extend the likelihood ratio tests to see if some nonparametric components in Cox's type of models are of certain parametric forms. This allows us to validate some nested Cox's models.

In nonparametric regression, a number of authors constructed the generalized likelihood ratio (*GLR*) tests to test if certain parametric/nonparametric null models hold and showed that the resulting tests share a common phenomenon, the Wilks phenomenon called in Fan, Zhang, and Zhang (2001). For details, see the reviewing paper of Fan and Jiang (2007). In the following, we introduce an idea of the *GLR* tests for Cox's type of models.

Consider, for example, the partly linear additive Cox model in (22):

$$\lambda\{t|\mathbf{z},\mathbf{w}\} \quad = \quad \lambda_0(t)\exp\{\mathbf{z}^\tau\boldsymbol{\beta}+\phi_1(w_1)+\cdots+\phi_J(w_J)\}, \tag{57}$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters and ϕ_i's are unknown functions. If one is interested in checking the significance of covariates, the following two null models may be considered:

$$H_a: \ \mathbf{A}\boldsymbol{\beta}=0 \ \text{ versus } \ H_1: \ \mathbf{A}\boldsymbol{\beta}\neq 0 \tag{58}$$

and

$$\begin{aligned} &H_b: \ \phi_1(w_1)=\cdots=\phi_d(w_d)=0 \ \text{ versus} \\ &H_1: \ \phi_1(w_1)\neq 0,\ldots, \text{ or } \phi_d(w_d)\neq 0, \end{aligned} \tag{59}$$

for $d=1,\ldots,J$. The former (58) tests the linear hypothesis on the parametric components, including the significance of a subset of variables, and the latter (59) tests the significance of the nonparametric components.

Under model (57), the maximum partial likelihood is

$$\ell(H_1)=\sum_{i=1}^n\delta_i\Big\{\mathbf{Z}_i^\tau\hat{\boldsymbol{\beta}}+\hat{\phi}(\mathbf{W}_i)-\log\sum_{j\in\mathcal{R}_i}\exp[\mathbf{Z}_j^\tau\hat{\boldsymbol{\beta}}+\hat{\phi}(\mathbf{W}_j)]\Big\},$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\phi}(\mathbf{W}_i)=\sum_{j=1}^J\hat{\phi}_j(W_{ji})$ are estimators in Section 2.3. For the null model (58), the maximum partial likelihood is

$$\ell(H_a)=\sum_{i=1}^n\delta_i\Big\{\mathbf{Z}_i^\tau\hat{\boldsymbol{\beta}}_a+\hat{\phi}_a(\mathbf{W}_j)-\log\sum_{j\in\mathcal{R}_i}\exp[\mathbf{Z}_j^\tau\hat{\boldsymbol{\beta}}_a+\hat{\phi}_a(\mathbf{W}_j)]\Big\},$$

where $\hat{\phi}_a(\mathbf{W}_j)=\sum_{k=1}^J\hat{\phi}_k(W_{kj})$ is the estimate based on polynomial splines under the null model. For the null model (59), the maximum partial likelihood is

$$\ell(H_b)=\sum_{i=1}^n\delta_i\Big\{\mathbf{Z}_i^\tau\hat{\boldsymbol{\beta}}_b+\hat{\phi}_b(\mathbf{W}_j)-\log\sum_{j\in\mathcal{R}_i}\exp[\mathbf{Z}_j^\tau\hat{\boldsymbol{\beta}}_b+\hat{\phi}_b(\mathbf{W}_j)]\Big\},$$

where $\hat{\phi}_b(\mathbf{W}_j) = \sum_{k=d+1}^{J} \hat{\phi}_k(W_{kj})$ is again the polynomial-spline based estimate under H_b. The *GLR* statistics can be defined as

$$\lambda_{n,a} = \ell(H_1) - \ell(H_a)$$

and

$$\lambda_{n,b} = \ell(H_1) - \ell(H_b),$$

respectively for the testing problems (58) and (59).

Since the estimation method for ϕ is efficient, we conjecture that the Wilks phenomenon holds (see Bickel 2007). That is, asymptotic null distribution of $\lambda_{n,a}$ is expected to be the Chi-square distribution with l degrees of freedom (see Fan and Huang 2005; Fan and Jiang 2007). Hence, the critical value can be computed by either the asymptotic distribution or simulations with nuisance parameters' values taken to be reasonable estimates under H_0. It is also demonstrated that one can proceed to the likelihood ratio test as if the model were parametric. For the test statistic $\lambda_{n,b}$, we conjecture that Wilks' phenomenon still exists. However, it is challenging to derive the asymptotic null distribution of the test statistic.

Similar test problems also exist in other Cox's type of models. More investigations along this direction are needed.

6 Transformation Models

Although Cox's type of models are very useful for analyzing survival data, the proportionality hazard assumption may not hold in applications. As an alternative to Cox's model (1), the following model

$$\lambda(t|Z(t)) = \lambda_0(t) + Z(t)'\boldsymbol{\beta}$$

postulates an additive structure on the baseline and the covariates' effects. This model is called an additive hazards model and has received much attention in statistics. See, for example, Lin and Ying (1994), Kulich and Lin (2000), and Jiang and Zhou (2007), among others. A combination of the multiplicative and additive hazards structures was proposed by Lin and Ying (1995), which takes the form

$$\lambda(t|Z_1(t), Z_2(t)) = \lambda_0(t)\exp(\boldsymbol{\beta}_1' Z_1(t)) + \boldsymbol{\beta}_2' Z_2(t),$$

where $Z_1(t)$ and $Z_2(t)$ are different covariates of $Z(t)$. It may happen in practice that the true hazard risks are neither multiplicative nor additive. This motivated Zeng, Yin, and Ibrahim (2005) to study a class of transformed hazards models by imposing both an additive structure and a known transformation $G(\cdot)$ on the hazard function, that is,

$$G(\lambda(t|Z(t)) = \lambda_0(t) + \boldsymbol{\beta}' Z(t), \qquad (60)$$

where $G(\cdot)$ is a known and increasing transformation function. Essentially, model (60) is a partial linear regression model for the transformed hazard function. In particular, within the family of the Box-Cox transformations

$$G(x) = \begin{cases} (x^s - 1)/s, & \text{if } s > 0, \\ \log(s), & \text{if } s = 0, \end{cases}$$

model (60) is the additive hazards model when $s = 1$ and the Cox model when $s = 0$. Since the model (60) allows a much broader class of hazard patterns than those of the Cox proportional hazards model and the additive hazards model, it provides us more flexibility in modeling survival data. The sieve maximum likelihood method can be used to estimate the model parameters, and the resulting estimators of parameters are efficient in the sense that their variances achieve the semiparametric efficiency bounds. For details, see Zeng, Yin, and Ibrahim (2005). Further work along this topic includes variable selection using the SCAD introduced before, hypothesis testing for the model parameters, and extensions to multivariate data analysis, among others, to which interested readers are encouraged to contribute.

Let $S(\cdot|\mathbf{Z})$ be the survival function of T conditioning on a vector of covariates $\mathbf{Z}$. Cox's model can be rewritten as

$$\log[-\log\{S(t|\mathbf{Z})\}] = H(t) + \mathbf{Z}'\boldsymbol{\beta}, \tag{61}$$

where H is an unspecified strictly increasing function. An alternative is the proportional odds model (Pettitt 1982; Bennett 1983):

$$-\text{logit}\{S(t|\mathbf{Z})\} = H(t) + \mathbf{Z}'\boldsymbol{\beta}. \tag{62}$$

Thus, a natural generalisation of (61) and (62) is

$$G\{S(t|\mathbf{Z})\} = H(t) + \mathbf{Z}'\boldsymbol{\beta}, \tag{63}$$

where $G(\cdot)$ is a known decreasing function. It is easy to see that model (63) is equivalent to

$$G(T) = -\boldsymbol{\beta}'Z + e, \tag{64}$$

where e is a random error with distribution function $F = 1 - G^{-1}$. For the noncensored case, the above model was studied by Cuzick (1988) and Bickel and Ritov (1997). For model (64) with possibly right censored observations, Cheng, Wei and Ying (1995) studied a class of estimating functions for the regression parameter $\boldsymbol{\beta}$.

A recent extension to model (64) is considered by Ma and Kosorok (2005), which takes the form

$$H(T) = \boldsymbol{\beta}'Z + f(W) + e,$$

where f is an unknown smooth function. This model obviously extends the partly linear Cox model (22) and model (64). Penalized maximum likelihood estimation has been investigated by Ma and Kosorok (2005) for the current status data, which shows that the resulting estimator of $\boldsymbol{\beta}$ is semiparametrically efficient while the estimators of H and f are $n^{1/3}$-consistent. Since the estimation method is likelihood based, the variable selection method and the GLR test introduced before are applicable to this model. Rigor theoretical results in this direction are to be developed.

7 Concluding remarks

Survival analysis is an important field in the theory and practice of statistics. The techniques developed in survival analysis have penetrated many disciplines such as the credit risk modeling in finance. Various methods are available in the literature for studying the survival data. Due to the limitation of space and time, we touch only the partial likelihood ratio inference for Cox's type of models. It is demonstrated that the non- and semi- parametric models provide various flexibility in modeling survival data. For analysis of asymptotic properties of the nonparametric components in Cox's type of models, counting processes and their associated martingales play an important role. For details, interested readers can consult with Fan, Gijbels, and King (2007) and Cai, Fan, Jiang, and Zhou (2007).

There are many other approaches to modeling survival data. Parametric methods for censored data are covered in detail by Kalbfleisch and Prentice (1980, Chapters 2 and 3) and by Lawless (1982, Chapter 6). Semiparametric models with unspecified baseline hazard function are studied in Cox and Oakes (1984). Martingale methods are also used to study the parametric models (Borgan 1984) and the semiparametric models (Fleming and Harrington 2005; Andersen et al, 1993).

References

[1] H. Akaike. Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika 60 (1973), 255–65.

[2] P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding, Statistical Models Based on Counting Processes, Springer-Verlag, New York 1993.

[3] M. Aitkin and D. G. Clayton. The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. Appl. Statist. 29 (1980), 156–163.

[4] O. E. Barndorff-Nielsen and D. R. Cox. Asymptotic Techniques for Use in Statistics. Chapman & Hall, 1989, page 252.

[5] J. M. Begun, W. J. Hall, W.-M. Huang, and J. A. Wellner. Information and asymptotic efficiency in parametric-nonparametric models. Ann. Statist. 11 (1982), 432–452.

[6] S. Bennetts. Analysis of survival data by the proportional odds model. Statist. Med. 2 (1983), 273-7.

[7] P. J. Bickel. Contribution to the discussion on the paper by Fan and Jiang, “Nonparametric inference with generalized likelihood ratio tests”. Test 16 (2007), 445–447.

[8] P. J. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner, Efficient and Adaptive Estimation in Semiparametric Models, Johns Hopkins University Press, Baltimore 1993.

[9] P. J. Bickel and Y. Ritov. Local asymptotic normality of ranks and covariates in transformation models. In ”Festschrift for Lucien Le Cam” (eds. D. Pollard, E. Torgersen and G. L. Yang) 43–54. Spring, New york 1997.

[10] L. Breiman. Heuristics of instability and stabilization in model selection. Ann. Statist. 24 (1996), 2350–83.

[11] N. E. Breslow. Contribution to the discussion on the paper by D.R. Cox, “Regression and life tables”. J. Royal Statist. Soc. B 34 (1972), 216–217.

[12] N. E. Breslow. Covariance analysis of censored survival data. Biometrics 30 (1974) 89–99.

[13] Bunea, F. and I. W. McKeague. Covariate selection for semiparametric hazard function regression models. J. Mult. Anal. 92 (2005), 186–204.

[14] J. Cai, J. Fan, J. Jiang, and H. Zhou. Partially Linear Hazard Regression for Multivariate Survival Data. Jour. Amer. Statist. Assoc. 102 (2007), 538–551.

[15] J. Cai, J. Fan, J. Jiang, and H. Zhou. Partially Linear Hazard Regression with Varying-coefficients for Multivariate Survival Data. J. Roy. Statist. Soc. B 70 (2008), 141-158.

[16] J. Cai, J. Fan, R. Li, and H. Zhou. Variable selection for multivariate failure time data. Biometrika 92 (2005), 303–316.

[17] J. Cai, J. Fan, H. Zhou, and Y. Zhou. Marginal hazard models with varying-coefficients for multivariate failure time data. The Annals of Statistics 35 (2007), 324–354

[18] J. Cai and R. L. Prentice. Estimating equations for hazard ratio parameters based on correlated failure time data, *Biometrika* 82 (1995), 151–164.

[19] Z. Cai and Y. Sun. Local linear estimation for time-dependent coefficients in Cox’s regression models. *Scandinavian Journal of Statistics* 30 (2003), 93–111.

[20] S. C. Cheng, L. J. Wei and Z. Ying. Analysis of transformation models with censored cata. Biometrika 82 (1995), 835–845.

[21] D. R. Cox. Regression models and life-tables (with discussion). J. Roy. Statist. Soc. B 34 (1972), 187–220.

[22] D. R. Cox. Partial likelihood. Biometrika 62 (1975), 269–276.

[23] D. R. Cox. The current position of statistics: a personal view, (with discussion). International Statistical Review 65 (1997), 261–276.

[24] D. R. Cox and D.V. Hinkley. Theoretical Statistics, London: Chapman and Hall, 1974.

[25] J. Cuzick. Rank regression. Aniz. Statist. 16 (1988), 1369–89.

[26] J. Fan, I. Gijbels, and M. King. Local likelihood and local partial likelihood in hazard regression. The Annals of Statistics 25 (1997), 1661–1690.

[27] J. Fan and J. Jiang. Nonparametric inference with generalized likelihood ratio tests (with discussions). Test 16 (2007), 409–478.

[28] J. Fan and T. Huang. Profile Likelihood Inferences on semiparametric varying-coefficient partially linear models. Bernoulli 11 (2005), 1031–1057.

[29] J. Fan and R. Li. Variable selection via penalized likelihood. Journal of American Statistical Association 96 (2001) 1348–1360.

[30] J. Fan and R. Li. Variable selection for Cox's proportional hazards model and frailty model. Ann. Statist. 30 (2002), 74–99.

[31] D. Faraggi and R. Simon. Bayesian variable selection method for censored survival data. Biometrics 54 (1998), 1475–85.

[32] T. R. Fleming and D. P. Harrington, Counting Processes and Survival Analysis, John Wiley & Sons, New Jersey 2005.

[33] D. P. Foster and E. I. George. The risk inflation criterion for multiple regression. Ann. Statist. 22 (1994), 1947–75.

[34] T. Hastie and R. Tibshirani. Generalized Additive Models. Chapman and Hall, London 1990.

[35] P. Hougaard. Analysis of Multivariate Survival Data. Springer, New York 2000.

[36] J. Huang. Efficient estimation of the partly linear additive Cox model. The Annals of Statistics 27 (1999), 1536–1563.

[37] J. Huang and D. Harrington. Penalised partial likelihood regression for right-censored data with bootstrap selection of the penalty parameter. Biometrics 58 (2002), 781–91.

[38] J. Jiang and K. A. Doksum. Empirical Plug-in Curve and Surface Estimates, In "Mathematical and Statistical Methods in Reliability", eds. B. H. Lindqvist and K. A. Doksum. Ser. Qual. Reliab. Eng. Stat. 7, 433–453. World Scientific Publishing Co., River Edge, New Jersey 2003.

[39] J. Jiang and Y.P. Mack. Robust local polynomial regression for dependent data. Statistica Sinica 11 (2001), 705–722.

[40] J. Jiang and H. Zhou. Additive Hazards Regression with Auxiliary Covariates. Biometrika 94 (2007), 359–369.

[41] S. Johansen. An extension of Cox's regression model. International Statistical Review 51 (1983), 258–262.

[42] J.D. Kalbfleisch and R.L. Prentice. The Statistical Analysis of Failure Time Data. Wiley, New York 2002.

[43] M. Kulich and D. Y. Lin. Additive hazards regression with co- variate measurement error. J. Am. Statist. Assoc. 95 (2000), 238–248.

[44] J. F. Lawless. Statistical Models and Methods for Lifetime Data. Wiley, New York 1982.

[45] E.W. Lee, L.J. Wei, and D.A. Amato. Cox-type regression analysis for large numbers of small groups of correlated failure time observations, *Survival Analysis: State of the Art. J. P. Klein and P. K. Goel (eds.)*, Kluwer Academic Publishers 1992, 237–247.

[46] K.Y. Liang, S.G. Self, and Y. Chang. Modeling marginal hazards in multivariate failure time data, J. Roy. Statist. Soc. B 55 (1993), 441–453.

[47] D. Y. Lin. Cox regression analysis of multivariate failure time data: The marginal approach. Statistics in Medicine 13 (1994), 2233–2247.

[48] D. Y. Lin and Z. Ying. Semiparametric analysis of the additive risk model. Biometrika 81 (1994), 61–71.

[49] D. Y. Lin and Z. Ying. Semiparametric analysis of general additive-multiplicative hazard models for counting processes," Annals of Statistics 23 (1995), 1712–1734.

[50] S. Ma and M. R. Kosorok. Penalized log-likelihood estimation for partly linear transformation models with current status data. Ann. Statist. 33 (2005), 2256–2290.

[51] E. Masry and J. Fan. Local polynomial estimation of regression functions for mixing processes. Scandinavian Journal of Statistics 24 (1997), 165–179.

[52] D. Oakes. Survival analysis, In "*Statistics in the 21st Century*", eds. A. E. Raftery, M. A. Tanner, and M. T. Wells. Monographs on Statistics and Applied Probability 93, 4–11. Chapman & Hall, London 2002.

[53] A. N. Pettitt. Inference for the linear model using a likelihood based on ranks. J. R. Statist. Soc. B 44 (1982), 234–243.

[54] R. L. Prentice and L. Hsu. Regression on hazard ratios and cross ratios in multivariate failure time analysis. Biometrka 84 (1997), 349–363.

[55] G. Schwarz. Estimating the dimension of a model. Ann. Statist. 6 (1978), 461–464.

[56] L. Schumaker. Spline Functions: Basic Theory. Wiley, New York 1981.

[57] R. Shibata. Approximation efficiency of a selection procedure for the number of regression variables. Biometrika 71(1984), 43–49.

[58] C.F. Spiekerman and D.Y. Lin. Marginal regression models for multivariate failure time data, Jour. Amer. Statist. Assoc. 93 (1998), 1164–1175.

[59] R. Tibshirani. The lasso method for variable selection in the Cox model. Statist. Med. 16 (1997), 385–395.

[60] A.A. Tsiatis. A large sample study of Cox's regression model. The Annals of Statistics 9 (1981), 93–108.

[61] L. J. Wei, D. Y. Lin, and L. Weissfeld. Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. J. Am. Statist. Assoc. 84 (1989), 1065–1073.

[62] Zeger, Diggle, and Liang (2004). A Cox model for biostatistics of the future. Johns Hopkins University, Dept. of Biostatistics Working Papers.

[63] D. Zeng, G. Yin, and J. G. Ibrahim. Inference for a Class of Transformed Hazards Models. J. Am. Statist. Assoc. 100 (2005), 1000-1008.