

1. Introduction
2. Models and Inference
3. Simulation Studies
4. Real Data Examples
5. Discussion

Semiparametric Hazards Rate Model for Modelling Short-term and Long-term Effects

Guoqing Diao

Department of Statistics
George Mason University

Joint work with Donglin Zeng

Outline

1. Introduction
2. Models and Inference
3. Simulation Studies
4. Real Data Examples
5. Discussion

1. Introduction
2. Models and Inference
3. Simulation Studies
4. Real Data Examples
5. Discussion

1. Introduction

Cox proportional hazards model

- ▶ The Cox model specifies that the hazard function of event time T given covariates \mathbf{X} takes the form

$$\lambda(t|\mathbf{X}) = \lambda(t)e^{\boldsymbol{\beta}^T \mathbf{x}},$$

where $\lambda(t)$ is an unspecified baseline hazard function.

- ▶ The Cox model is the most widely used model in survival analysis. The seminal work of Cox (JRSSB, 1972) was highly cited with 19,641 citations (as of August 1, 2011).
- ▶ However, the assumption of constant relative risk over time is violated in many biomedical and genetic studies.

Example 1: Gastrointestinal Tumor Study

- ▶ The aim was to compare chemotherapy with combined chemotherapy and radiotherapy on the treatment of locally unresectable gastric cancer (Gastrointestinal Tumor Study Group, 1982).
- ▶ There were 45 patients randomly assigned to each treatment arm.
- ▶ Two observations were censored in the chemotherapy group and six were censored in the combined therapy group.

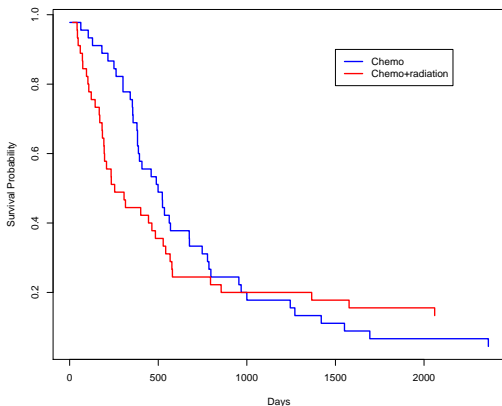


Figure: Kaplan-Meier survival curves from the Gastrointestinal tumor study.

Example 1: Gastrointestinal Tumor Study (Cont.)

What happens if we apply the Cox model to Gastrointestinal tumor study?

- ▶ The log-hazard ratio of chemotherapy versus the combined therapy is estimated at 0.106 with a standard error estimate of 0.223, yielding a p -value of 0.635.
- ▶ The use of proportional hazards model failed to capture the phenomenon of crossing survival curves.
- ▶ Application of the Cox model yielded very misleading results in this situation.

Example 2: COGA Study

- ▶ The Collaborative Study on the Genetics of Alcoholism (COGA) is a genetic family study with the aim of identifying and characterizing genetic factors that affect the susceptibility to alcohol dependence and related phenotypes (Begleiter et al., 1995).
- ▶ Our interest is to investigate genetic effects on the age at onset of alcoholism.
- ▶ After excluding individuals with missing data, the final data set for our analysis consisted of 1,371 individuals, including 626 affected individuals and 745 unaffected individuals.

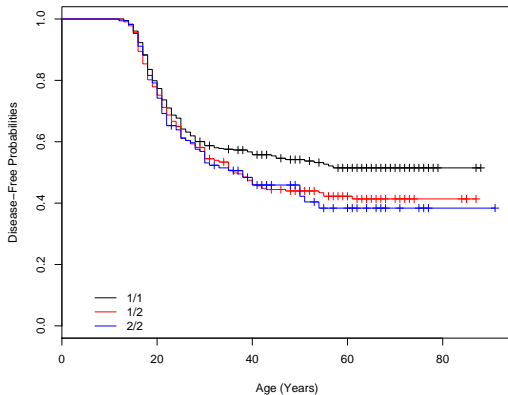


Figure: Genotype-specific Kaplan-Meier survival curves at SNP rs1972373 on chromosome 14 from the COGA study.

Example 2: COGA (Cont.)

What happens if we apply the Cox model to COGA?

- ▶ The log-hazard ratio estimated from the Cox model is 0.083 with a standard error estimate of 0.058, corresponding to a p-value of 0.153.
- ▶ The Cox model failed to detect the long-term effect of SNP rs1972373.

Alternative methods when PH assumption is not valid

- ▶ Proportional odds model (Bennett, 1983). Still cannot deal with crossing hazards.
- ▶ Cox model with manufactured time-dependent covariates involving interactions between covariates and time in the standard Cox model (Hess, 1994; Therneau and Grambsch, 2000).
- ▶ Time-varying regression coefficients model (Martinussen and Scheike, 2002; Martinussen et al., 2002; among others)

$$\lambda(t|\mathbf{X}) = \lambda(t)e^{\boldsymbol{\beta}^T(t)\mathbf{X}}.$$

In general, nonparametric smoothing is needed to estimate the time varying coefficients.

Alternative methods (Cont.)

- ▶ Yang and Prentice (2005) proposed the following novel semiparametric two-sample model

$$\lambda_T(t) = \frac{\theta_1 \theta_2}{\theta_1 + (\theta_2 - \theta_1) S_C(t)} \lambda_C(t).$$

- ▶ Parameters θ_1 and θ_2 can be interpreted as the short-term and long-term hazard ratios, respectively.
- ▶ Yang and Prentice (2005) developed a pseudo maximum likelihood approach.
- ▶ Extension to general covariates was discussed in Yang and Prentice (2005), but no theoretical or numerical results were available.

What's new?

- ▶ Extend the two-sample model of Yang and Prentice (2005) to accommodate potentially time-dependent continuous or discrete covariates.
- ▶ Develop nonparametric likelihood-based estimation and inference procedures.
- ▶ Establish the large sample properties of the nonparametric maximum likelihood estimators (NPMLEs). In particular, the asymptotic covariance matrix of the NPMLEs of the regression parameters were shown to attain the semiparametric efficiency bound.

1. Introduction
- 2. Models and Inference**
3. Simulation Studies
4. Real Data Examples
5. Discussion

2. Models and Inference

- ▶ Notations: for $i = 1, \dots, n$
 - ▶ T_i : failure time
 - ▶ C_i : censoring time
 - ▶ $Y_i = \min(T_i, C_i)$
 - ▶ $\Delta_i = I(T_i \leq C_i)$
 - ▶ $\mathbf{X}_i(\cdot)$: potentially time-dependent covariates
 - ▶ $\overline{\mathbf{X}}_i(t)$: the history of $\mathbf{X}_i(\cdot)$ over $[0, t]$

Extension to allow for general covariates

- ▶ Model (1):

$$\lambda(t|\mathbf{X}_i) = \frac{e^{(\boldsymbol{\beta}+\boldsymbol{\gamma})^T \mathbf{x}_i}}{e^{\boldsymbol{\beta}^T \mathbf{x}_i} F(t) + e^{\boldsymbol{\gamma}^T \mathbf{x}_i} S(t)} \lambda(t),$$

where S and F are baseline survival and cumulative distribution functions, respectively.

- ▶ Parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be interpreted as short-term and long-term log-hazard ratios, respectively.
- ▶ Reduces to Cox model when $\boldsymbol{\beta} = \boldsymbol{\gamma}$; and reduces to proportional odds model when $\boldsymbol{\gamma} = \mathbf{0}$.

Extension to allow for time-dependent covariates

- ▶ Model (2):

$$\Lambda(t|\bar{\mathbf{X}}_i(t)) = \int_0^t \frac{e^{(\boldsymbol{\beta}+\boldsymbol{\gamma})^T \mathbf{X}_i(s)}}{e^{\boldsymbol{\beta}^T \mathbf{X}_i(s)} F(s) + e^{\boldsymbol{\gamma}^T \mathbf{X}_i(s)} S(s)} d\Lambda(s),$$

where $\Lambda(t|\bar{\mathbf{X}}(t))$ is the cumulative hazard function given $\bar{\mathbf{X}}(t)$.

- ▶ Reduces to Model (1) if $\mathbf{X}(\cdot)$ is time-independent.

Likelihood function for (β, γ, Λ)

$$\prod_{i=1}^n \left[\frac{e^{(\beta+\gamma)^T \mathbf{x}_i(Y_i)} \Lambda'(Y_i)}{e^{\beta^T \mathbf{x}_i(Y_i)} F(Y_i) + e^{\gamma^T \mathbf{x}_i(Y_i)} S(Y_i)} \right]^{\Delta_i} e^{-\Lambda(Y_i | \bar{\mathbf{x}}_i(Y_i))},$$

where $\Lambda'(t)$ is the first derivative of $\Lambda(t)$.

- ▶ Replace $\Lambda'(Y_i)$ with $\Lambda\{Y_i\}$, the jump size of $\Lambda(\cdot)$ at Y_i , we obtain the nonparametric likelihood.
- ▶ Use optimization algorithm to estimate $\theta \equiv (\beta, \gamma)$ and jump sizes of $\Lambda(\cdot)$ simultaneously.
- ▶ Denote the NPMLEs by $(\hat{\theta}_n, \hat{\Lambda}_n)$.

Large sample properties of NPMLEs

Theorem 1. Under certain regularity conditions, $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \rightarrow 0$ and $\sup_{t \in [0, \tau]} |\widehat{\Lambda}_n(t) - \Lambda_0(t)| \rightarrow 0$ almost surely, where $\|\cdot\|$ is the Euclidean norm.

Theorem 2. Under conditions (C1)-(C5), the random element $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0, \widehat{\Lambda}_n - \Lambda_0)$ converges weakly to a zero mean Gaussian process in the metric space $l^\infty(\mathcal{H})$, where

$$\mathcal{H} = \{(\mathbf{h}_1, \mathbf{h}_2, h_3) : \mathbf{h}_1 \in R^p, \mathbf{h}_2 \in R^p, h_3 \text{ is a function on } [0, \tau]; \\ \|\mathbf{h}_1\| \leq 1, \|\mathbf{h}_2\| \leq 1, |h_3|_V \leq 1\}$$

and $|h_3|_V$ denotes the total variation of h_3 in $[0, \tau]$. Furthermore, $\widehat{\boldsymbol{\theta}}_n$ is asymptotically efficient.

1. Introduction
2. Models and Inference
- 3. Simulation Studies**
4. Real Data Examples
5. Discussion

3. Simulation Studies

We generate data from Model (2):

- ▶ X_i is a uniform(-1,1) random variable.
- ▶ $\Lambda(t) = t$.
- ▶ We consider two scenarios for the value of regression parameters: (a) $(\beta, \gamma) = (-0.5, 0.5)$; and (b) $(\beta, \gamma) = (0, 0.5)$.
- ▶ Censoring time is set to be the minimum of 2 and a uniform(0,4) variable.
- ▶ Censoring rate is approximately 29% for each of the above four scenarios.
- ▶ We generate 1,000 replicates for each setting.

Table 1. Summary statistics for the NPMLEs under scenario (a)

n	Parameter	Mean	SE	SEE	CP
200	β	-0.512	0.291	0.288	0.954
	γ	0.496	0.400	0.389	0.940
	$\Lambda(0.5)$	0.504	0.059	0.059	0.954
	$\Lambda(1.0)$	1.012	0.104	0.101	0.947
	β_{PH}	-0.107	0.153	0.147	-
	β_{PO}	-0.287	0.231	0.225	-
400	β	-0.504	0.202	0.202	0.953
	γ	0.494	0.277	0.274	0.946
	$\Lambda(0.5)$	0.502	0.043	0.042	0.936
	$\Lambda(1.0)$	1.006	0.073	0.071	0.946
	β_{PH}	-0.105	0.105	0.103	-
	β_{PO}	-0.284	0.159	0.159	-

Table 2. Summary statistics for the NPMLEs under scenario (b)

n	Parameter	Mean	SE	SEE	CP
200	β	-0.009	0.284	0.282	0.956
	γ	0.501	0.398	0.395	0.947
	$\Lambda(0.5)$	0.506	0.059	0.060	0.952
	$\Lambda(1.0)$	1.014	0.104	0.102	0.944
	β_{PH}	0.193	0.149	0.147	-
	β_{PO}	0.207	0.227	0.225	-
400	β	-0.004	0.193	0.198	0.960
	γ	0.498	0.272	0.278	0.953
	$\Lambda(0.5)$	0.503	0.042	0.042	0.948
	$\Lambda(1.0)$	1.007	0.071	0.072	0.939
	β_{PH}	0.194	0.104	0.103	-
	β_{PO}	0.209	0.157	0.158	-

Table 3. Mean squared errors of the proposed NPMLEs and the pseudo-maximum likelihood estimators (PMLEs) of Yang and Prentice (2005) for (β, γ)

n	(β, γ)	PMLE		NPMLE		PMLE/NPMLE	
		$\hat{\beta}$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{\gamma}$
200	(-0.5, 0.5)	0.0483	0.0595	0.0355	0.0537	1.360	1.107
	(-0.5, 0.0)	0.0410	0.0607	0.0313	0.0543	1.310	1.119
	(0.0, 0.5)	0.0303	0.0503	0.0296	0.0516	1.025	0.974
	(0.5, 0.5)	0.0345	0.0638	0.0299	0.0598	1.152	1.068
400	(-0.5, 0.5)	0.0226	0.0319	0.0167	0.0279	1.357	1.140
	(-0.5, 0.0)	0.0160	0.0296	0.0145	0.0284	1.101	1.041
	(0.0, 0.5)	0.0133	0.0250	0.0131	0.0253	1.013	0.989
	(0.5, 0.5)	0.0145	0.0297	0.0137	0.0285	1.061	1.039

1. Introduction
2. Models and Inference
3. Simulation Studies
- 4. Real Data Examples**
5. Discussion

4. Real Data Examples

Example 1: Gastrointestinal Tumor Study

- ▶ Let $X_i = 0.5$ for the combined therapy group and $X_i = -0.5$ for the chemotherapy group.
- ▶ $\hat{\beta}_n = 1.76, \widehat{S.E.}(\hat{\beta}_n) = 0.582, p\text{-value} = 0.0025, 95\%CI = (0.62, 2.90)$
- ▶ $\hat{\gamma}_n = -1.59, \widehat{S.E.}(\hat{\gamma}_n) = 0.509, p\text{-value} = 0.0018, 95\%CI = (-2.59, -0.59)$
- ▶ Testing of proportional hazards assumption:
p - value = 0.0006

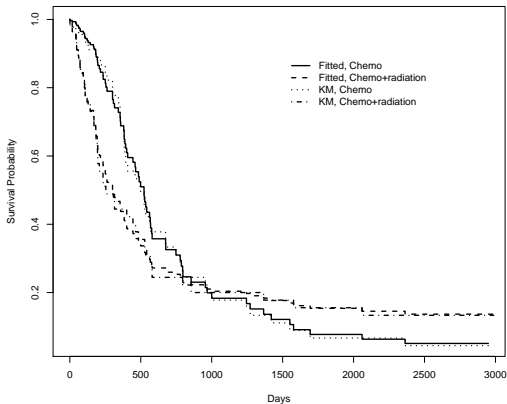


Figure: Kaplan-Meier (KM) and model-fitted (Fitted) survival curves from the Gastrointestinal tumor study.

Example 2: COGA Study

- ▶ $X_{i1} = 1$ for male and 0 for female.
- ▶ X_{i2} = the number of allele type '2' at SNP rs1972373.
- ▶ Both covariates were then centered at their means.
- ▶ $\hat{\beta}_{n,1} = 0.866, \widehat{S.E.}(\hat{\beta}_{n,1}) = 0.147, p - \text{value} < 0.0001$
- ▶ $\hat{\gamma}_{n,1} = 1.993, \widehat{S.E.}(\hat{\gamma}_{n,2}) = 0.367, p - \text{value} < 0.0001$
- ▶ $\hat{\beta}_{n,2} = -0.06, p - \text{value} = 0.479$
- ▶ $\hat{\gamma}_{n,2} = 0.683, p - \text{value} = 0.015$
- ▶ Testing of proportional hazards assumption:
p - value = 0.016 and p - value = 0.027 for gender and genotype score, respectively.

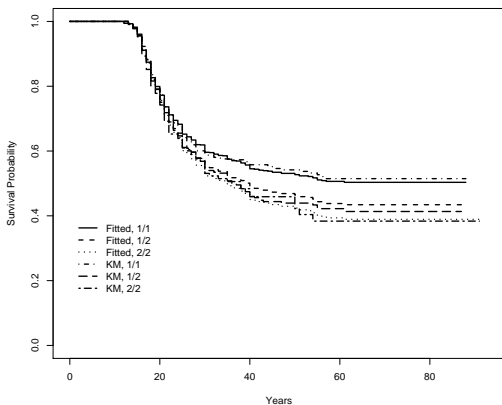


Figure: Kaplan-Meier (KM) and model-fitted (Fitted) survival curves from the COGA study.

1. Introduction
2. Models and Inference
3. Simulation Studies
4. Real Data Examples
- 5. Discussion**

5. Discussion

- ▶ We propose a semiparametric general hazards model which allows for non-constant hazard ratios over time and accommodates time-dependent covariates.
- ▶ We establish the large sample properties of the NPMLEs.
- ▶ We may use random effects to account for correlations among subjects in a family or cluster.
- ▶ A user-friendly computer program is available at <http://mason.gmu.edu/~gdiao/software>.

1. Introduction
2. Models and Inference
3. Simulation Studies
4. Real Data Examples
5. Discussion

Acknowledgment

This work was supported by National Cancer Institute Grant CA150698-01.