mathematics

MDPI

Article

Nonparametric Sieve Maximum Likelihood Estimation of Semi-Competing Risks Data

Xifen Huang and Jinfeng Xu *

School of Mathematics, Yunnan Normal University, Kunming 650092, China; 190004@ynnu.edu.cn
* Correspondence: 204203@ynnu.edu.cn

Abstract: In biomedical studies involving time-to-event data, a subject may experience distinct types of events. We consider the problem of estimating the transition functions for a semi-competing risks model under illness-death model framework. We propose to estimate the intensity functions by maximizing a B-spline based sieve likelihood. The method yields smooth estimates without parametric assumptions. Our proposed approach facilitates easy computation of the covariance of the model parameters and yields direct interpretation. Compared with existing approaches, our proposed method requires neither the subjective specification of the frailty distribution nor the Markov or semi-Markov assumption which may be unmet in real applications. We establish the consistency, the convergence rate, and the asymptotic normality of the proposed estimators under some regularity conditions. We also provide simulation studies to assess the finite-sample performance of the proposed modeling and estimation strategy. A real data application is further used to illustrate the proposed methodology.

Keywords: asymptotics; B-spline; illness-death model; Markov model; proportional hazards; semi-competing risks data

MSC: 46N30; 65C60

Citation: Huang, X.; Xu, J. Nonparametric Sieve Maximum Likelihood Estimation of Semi-Competing Risks Data. *Mathematics* **2022**, *10*, 2248. https://doi.org/10.3390/math10132248

Academic Editors: Min Wang, Haijun Gong, Liucang Wu and Songfeng Zheng

Received: 23 May 2022
Accepted: 22 June 2022
Published: 27 June 2022

1. Introduction

In survival analysis, a subject may experience several distinct types of failures. If apart from censoring, the follow up period ends upon the occurrence of the first event, such data are often referred to as competing risks data. This framework consists of survival data where failure may be due to one of a number of competing causes. In some application, with additional information, this notion can be extended to accommodate that of semi-competing risks ([1,2]), where one type of event (terminal event, e.g., death) may censor the other events (non-terminal event, e.g., relapse of the disease), but not vice versa. The framework of semi-competing risk data have been previously discussed in [1,3]. Furthermore, competing risks data can also be regarded as a special type of multitask prediction problem, which simultaneously predicts multiple outcomes from the same set of predictors. A stacking algorithm borrowing information among multiple prediction tasks to improve multivariate prediction performance (MTPS) is recently proposed by [4]. The MTPS is shown to outperform existing multivariate prediction methods.

Recently [5] suggests that semicompeting risks data can also be analyzed using the conventional illness–death compartment model by a subjective specification of the frailty distribution and postulating the Markov or semi-Markov assumption for the conditional transition functions given the covariates and the frailty ([6,7]). However, the subjective specification of the frailty distribution or the Markov or semi-Markov assumption may be unmet in some practical applications, leading to inconsistent estimators. In such cases, alternative (non-Markov) estimators are needed. Furhthemore, their nonparametric maximum likelihood estimation approach may be computational demanding when the sample size is large.

To address the theoretical and numerical challenges in the semiparametric estimation of semi-competing risks model, we employ the B-spline based sieve maximum likelihood approach to simultaneously estimate the regression parameters and transition functions. Covariates are incorporated naturally via proportional hazards assumptions. This approach facilitates easy calculation of the covariance of the model parameters. The proposed spline estimation algorithm requires much less computation than the isotonic type algorithm used in [5] since the size of the step function is much larger than the number of parameters in our proposed B-spline based approach. Under certain regularity conditions, we are able to prove that the estimators of regression parameters is root-n consistent, asymptotically normal and semiparametric efficient.

The rest of the paper is organized as follows. In Section 2, we will introduce our proposed model and estimating approach. In Section 3, we study the asymptotic properties of the proposed estimators. In Section 4, we provide simulation results. An application to colon cancer data is given in Section 5. We then conclude with some discussion in Section 6. All proofs are relegated to the Appendix A.

2. Methodology

2.1. Model and Likelihood Function

For the ith subject, let C_i, X_i, T_{i1}, and T_{i2} denote the censoring, covariate vector, non-terminal event time, and terminal event time, respectively. Define $Y_{i2} = T_{i2} \wedge C_i$, $\delta_{i2} = I(T_{i2} \leqslant C_i)$, $Y_{i1} = T_{i1} \wedge Y_{i2}$, and $\delta_{i1} = I(T_{i1} \leqslant Y_{i2})$. We observe $(Y_{i2}, \delta_{i2}, \delta_{i1}, X_i, i = 1, \ldots, n)$. The hazard functions are defined as below.

$$\lambda_1(t_1) = \lim_{\Delta \to 0} P[T_1 \in [t_1, t_1 + \Delta) | T_1 \geq t_1, T_2 \geq t_1] / \Delta, \tag{1}$$

$$\lambda_2(t_2) = \lim_{\Delta \to 0} P[T_2 \in [t_2, t_2 + \Delta) | T_1 \geq t_2, T_2 \geq t_2] / \Delta, \tag{2}$$

$$\lambda_{12}(t_2|t_1) = \lim_{\Delta \to 0} P[T_2 \in [t_2, t_2 + \Delta) | T_1 = t_1, T_2 \geq t_2] / \Delta, \tag{3}$$

where $0 < t_1 < t_2$. In general, $\lambda_{12}(t_2|t_1)$ can depend on both t_1 and t_2 (see Remark 1 for more detailed discussions). Let $\Lambda_1(t) = \int_0^t \lambda_1(x)dx$ and $\Lambda_2(s) = \int_0^s \lambda_2(x)dx$. Specifically, the probability measure P refers to the joint distribution of (T_1, T_2, C) in the unconditional case. In the conditional case, the probability measure P refers to the joint distribution of (T_1, T_2, C) given X. For the unconditional case, the likelihood function $L(\theta)$ then takes the form

$$\prod_{i=1}^{n} \lambda_1(Y_{i1})^{\delta_{i1}} \lambda_2(Y_{i2})^{(1-\delta_{i1})\delta_{i2}} \lambda_{12}(Y_{i2}|Y_{i1})^{\delta_{i1}\delta_{i2}} \exp\Big(- \Lambda_1(Y_{i1}) - \Lambda_2(Y_{i1}) - \int_{Y_{i1}}^{Y_{i2}} \lambda_{12}(s|Y_{i1})ds) \Big), \tag{4}$$

where $\theta = (\beta_1, \beta_2, \beta_3, \lambda_{10}, \lambda_{20}, \lambda_{30})$ will be specified as follows.

For the case with q dimension covariates X, the conditional transition rate functions are defined as follows:

$$\lambda_1(t_1|X = x) = \lambda_{10}(t_1) \exp(\beta_1^T x), \tag{5}$$

$$\lambda_2(t_2|X = x) = \lambda_{20}(t_2) \exp(\beta_2^T x), \tag{6}$$

$$\lambda_{12}(t_2|t_1, X = x) = \lambda_{12,0}(t_2|t_1) \exp(\beta_3^T x). \tag{7}$$

Note that both x and X refer to the covariates where X denote the random variable and x refers to its observed values. The Equations (5)–(7) are the conditional transition functions of T_1 and T_2 (given $X = x$) while the Equations (1)–(3) are the unconditional transition functions of T_1 and T_2.

To simplify the notation, denote $\lambda_3(t, s) = \lambda_{12}(s|t)$, $\lambda_{30}(t, s) = \lambda_{12,0}(s|t)$, $\beta = (\beta_1^T, \beta_2^T, \beta_3^T)^T$, $\beta_0 = (\beta_{10}^T, \beta_{20}^T, \beta_{30}^T)^T$. Note that in our modeling approach, λ_{30} depends on two parameters t and s.

2.2. Sieve Space Θ_n for the Parameters $(\beta_1, \beta_2, \beta_3, \lambda_{10}, \lambda_{20}, \lambda_{30})$

We propose a sieve space consisting of B-splines for $\lambda_{j0}(j = 1, 2, 3)$ in maximizing (4). We suppose that Y_1 and Y_2 have compact supports (say $[0, 1]$) and that $\|\beta\| \leq M$ for a known constant M. Rewrite $\lambda_{10}(t) = \exp(g_{10}(t)), \lambda_{20}(s) = \exp(g_{20}(s)), \lambda_{30}(t, s) = \exp(g_{30}(t, s))$. Let $\psi = (g_1, g_2, g_3)$ and $\psi_0 = (g_{10}, g_{20}, g_{30})$. A sieve space consisting of B-splines is defined for these new parameters as follows: First, we obtain an extended partition with equal length $1/K_n$ for the interval $[0, 1]$:

$$\Delta = \{s_{-m} = \cdots = s_{-1} = 0 = s_0 < s_1 < \cdots < s_{K_n} = 1 = \cdots = s_{K_n+m}\},$$

where m (independent of the sample size n) and $K_n = O(n^\nu)(0 < \nu < 1/2)$ are two integers to be chosen later. Note that m and K_n are two parameters often used in B-spline modeling where m indicates the smoothness of the basis function. Let $N_n = K_n + m$ and $\{N_j^m(s)\}_{j=1}^{N_n}$ be a normalized B-spline basis associated with Δ (see [8]). Then the sieve space for the parameters $\theta = (\beta, \psi(t, s))$ is defined as

$$\begin{aligned}\Theta_n = \Big\{&\theta_n = (\beta, \psi_n(s, t)) : \psi_n(s, t) = (g_{1n}(t), g_{2n}(s), g_{3n}(s, t), \|\beta\| \leq M,\\ &g_{1n}(t) = \sum_{i=1}^{m+K_n} \alpha_i N_i^m(t), g_{2n}(s) = \sum_{i=1}^{m+K_n} \eta_i N_i^m(s),\\ &g_{3n}(s, t) = \sum_{i_1, i_2=1}^{m+K_n} \gamma_{i_1, i_2} N_{i_1}^m(s) N_{i_2}^m(t), \max_{1\leq i\leq m+K_n} |\alpha_i| \leq M_n,\\ &\max_{1\leq i\leq m+K_n} |\eta_i| \leq M_n, \max_{1\leq i_1, i_2\leq m+K_n} |\gamma_{i_1, i_2}| \leq M_n\Big\},\end{aligned} \tag{8}$$

where $M_n \leq (2m - 1)/(2m'(2m + 1))$ with a constant m' arbitrarily close to m.

For any $\theta_i = (\beta_i, \psi_i) \in \Theta$ $(i = 1, 2)$, we define a distance $d(\theta_1, \theta_2) = \|\beta_1 - \beta_2\| + \|\psi_1 - \psi_2\|_2$.

Remark 1. *Here we assume that the transition intensity $\lambda_{30}(\cdot)$ depends on both t_1 and t_2. A semi-Markov process specifies that $\lambda_{30}(t_1, t_2) = h_2(t_2 - t_1)$. However, it is important to note that in either Markov or semi-Markov approaches, λ_{30} depends on only one parameter, corresponding to the special cases of our modeling approach where λ_{30} can flexibly depend on two parameters.*

2.3. Maximization

Let P_n, P denote the empirical measure and the true probability measure of $(\delta_1, \delta_2, Y_1, Y_2, X)$, respectively. We maximize the function

$$\begin{aligned}l_n(\beta, \psi) = P_n l(\theta; W_i) = P_n l(\beta, \psi; W_i) = P_n\Big\{&\delta_{1i}\Big[X_i^T\beta_1 + g_1(Y_{1i})\Big] + (1 - \delta_{1i})\delta_{2i}\Big[X_i^T\beta_2\\ &+g_2(Y_{2i})\Big] + \delta_{1i}\delta_{2i}\Big[X_i^T\beta_3 + g_3(Y_{1i}, Y_{2i})\Big] - \Lambda_1(Y_{1i}) - \Lambda_2(Y_{2i})\\ &- \int_{Y_{1i}}^{Y_{2i}} \exp(g_3(Y_{1i}, s))ds\Big\}\end{aligned} \tag{9}$$

over the sieve space Θ_n.

For the knot selection, we let $m = 3$ and use the Bayesian information criterion

$$\mathrm{BIC}(N_n) = l_n(\hat{\beta}, \hat{\psi}) + \frac{\log n}{n}(3N_n + 3q)$$

to choose K_n which minimizes the criterion function.

3. Theoretical Properties

In this section, we establish the theoretical properties of our spline-based modeling strategy under the following regularity conditions.

Assumptions

(A1) Y_1 and Y_2 have compact supports (say $[0,1]$) and X has bounded support in $\mathbb{R}^q$ where q is the dimension of X. Moreover, if there exists a constant c_0 and a constant vector $\tilde{\gamma}$ such that $\gamma^\top X = c_0$ almost surely, then $c_0 = 0$ and $\tilde{\gamma} = 0$.
(A2) $\beta_0 \in \mathcal{B}$, where $\mathcal{B}$ is a compact set of $\mathbb{R}^{3q}$ with nonempty interior. λ_{10} and $\lambda_{20} \in \mathcal{H}_r$, and $\lambda_{30} \in \mathcal{C}_r$.
(A3) $K_n = O(n^\nu)$ where ν satisfies the restrictions $0.25/r < \nu < 0.5$.
(A4) $r \geq 2$ where r is the measure of smoothness of λ_j in definitions of $\mathcal{H}_r$ and $\mathcal{C}_r$.

We first establish the strong consistency for the estimated model parameters.

Theorem 1. *Under Assumptions A1–A3, $\hat{\beta}$ are strong consistent estimators of the true coefficients β_0, and $\|\hat{\lambda}_1 - \lambda_{10}\|_2 \longrightarrow 0$, $\|\hat{\lambda}_2 - \lambda_{20}\|_2 \longrightarrow 0$, $\|\hat{\lambda}_3 - \lambda_{30}\|_2 \longrightarrow 0$ almost surely.*

Next, we obtain the convergence rates for the proposed estimators.

Theorem 2. *Under Assumptions A1–A3, it holds that*

$$\|\hat{\lambda}_1 - \lambda_{10}\|_2 + \|\hat{\lambda}_2 - \lambda_{20}\|_2 + \|\hat{\lambda}_3 - \lambda_{30}\|_2 = O_p(n^{-r\nu} + n^{-(1/2-\nu)}).$$

This theorem implies that if $v = 1/(2+2r)$, $\|\hat{\lambda}_3 - \lambda_{30}\|_2 = O_p(n^{-r/(2r+2)})$, which is the optimal convergence rate in the non-parametric regression setting for bivariate function estimation by [9].

To derive the limiting distribution of the proposed estimators, establish the asymptotic normality, we calculate the directional derivative of the log-likelihood in the associate functional spaces as follows.

Denote V as the linear span of $\Theta_0 - \theta_0$, where θ_0 denote the true value of $\theta = (\beta, \psi)$ and Θ_0 denote the true parameter space. Let $l(\theta; W)$ be the log-likelihood for a sample of size one and $\delta_n = n^{-r\nu} + n^{-(1/2-\nu)}$. For any $\theta \in \{\theta \in \Theta_0 : \|\theta - \theta_0\| = O(\delta_n)\}$, define the first order directional derivative of $l(\theta; W)$ at the direction $v \in V$ as

$$\dot{l}(\theta; W) = \frac{dl(\theta + sv; W)}{ds}\Big|_{s=0},$$

and the second order directional derivative as

$$\ddot{l}(\theta; W) = \frac{d^2 l(\theta + sv + \tilde{s}\tilde{v}; W)}{d\tilde{s}ds}\Big|_{s=0}\Big|_{\tilde{s}=0} = \frac{d\dot{l}(\theta + \tilde{s}\tilde{v}; W)}{d\tilde{s}}\Big|_{\tilde{s}=0}.$$

Define the Fisher inner product on the space V as

$$< v, \tilde{v} >= P\Big\{\dot{l}(\theta; W)[v]\dot{l}(\theta; W)[\tilde{v}]\Big\}$$

and the Fisher norm for $v \in V$ as $\|v\|^{1/2} =< v, v >$. Let $\bar{V}$ be the closed linear span of V under the Fisher norm. Then $(\bar{V}, \|\cdot\|)$ is a Hilbert space.

Define the smooth functional of θ as

$$\gamma(\theta) = b'\beta + \int_0^1 \phi_1(t)\lambda_1(t)dt + \int_0^1 \phi_2(s)\lambda_2(s)ds + \int_0^1\int_0^1 \phi_3(t,s)\lambda_3(t,s)dtds,$$

where b is any vector of $3q$ dimension with $\|b\| \leq 1$, $\phi_i \in \mathcal{H}_r[0,1], i = 1,2$ $\lambda_3 \in \mathcal{C}_r[0,1]^2$. For any $v \in V$, we denote

$$\dot{\gamma}(\theta_0)[v] = \frac{d\gamma(\theta_0 + sv)}{ds}\Big|_{s=0}$$

whenever the right hand-side limit is well defined and assume:
(A5) for any $v \in \bar{V}$, $\gamma(\theta_0 + sv)$ is continuously differentiable in $s \in [0,1]$ near $s = 0$, and

$$\|\dot{\gamma}(\theta_0)\| = \sup_{v \in \bar{V}: \|v\| > 0} \frac{|\dot{\gamma}(\theta_0)[v]|}{\|v\|} < \infty.$$

Note that $\gamma(\theta) - \gamma(\theta_0) = \dot{\gamma}(\theta_0)[\theta - \theta_0]$. Under Assumption A5, by the Riesz representation theorem, there exists $v^* \in \bar{V}$ such that $\dot{\gamma}(\theta_0)[v] =< v^*, v >$ for all $v \in \bar{V}$ and $\|v^*\|^2 = \|\dot{\gamma}(\theta_0)\|$.

Theorem 3. *Suppose suppose $r > 2$ and assumptions A1–A3, A5 hold, then $n^{1/2}(\gamma(\hat{\theta}) - \gamma(\theta)) \to N(0, \|\dot{\gamma}(\theta_0)\|^2)$ in distribution and and $\gamma(\hat{\theta})$ is semiparametrically efficient.*

Remark 2. Inference about $\hat{\beta}$. *Theorem 3 offers ease of inference procedure, especially for the regression parameter β. Set $\phi_j(\cdot) = 0 (j = 1,2,3)$, then Theorem 3 yields that $n^{1/2} b'(\hat{\beta} - \beta_0) \to N(0, b' \Sigma_{\beta\beta} b)$, and thus*

$$n^{1/2}(\hat{\beta} - \beta_0) \to N(0, \Sigma_{\beta\beta}),$$

by Gramer-Wold device, one can establish semiparametricefficiency of $\hat{\beta}$. where $\Sigma_{\beta\beta}$ can be consistently estimated using the inverse of the Hessian matrix.

Remark 3. Inference about $\lambda_j(\cdot)(j = 1,2,3)$. *For $\lambda_j(\cdot)(j = 1,2)$, let $b = 0$ and $\phi_k(k \neq j) = 0$, then Theorem 3 yields that*

$$n^{1/2} \int_0^1 \phi_j(w)(\hat{\lambda}_j(w) - \lambda_{j0}(w))dw \to N(0, \sigma^2_{\lambda_j}),$$

where $\sigma^2_{\lambda_j} (j = 1,2)$ can be consistently estimated by using the delta method or some resampling methods. Similarly inference can be done for $\lambda_3(t,s)$: Let $b = 0$, $\phi_1(\cdot) = 0$, $\phi_2(\cdot) = 0$, then Theorem 3 yields that

$$n^{1/2} \int_0^1 \int_0^1 \phi_j(t,s)(\hat{\lambda}_3(t,s) - \lambda_{30}(t,s))dtds \to N(0, \sigma^2_{\lambda_3}),$$

where $\sigma^2_{\lambda_3}$ can be consistently estimated by using the delta method or some resampling methods. The above results can be used to check the linear (quadratic) effect of $t_j (j = 1,2)$, or to check whether $\lambda_3(t_1, t_2)$ is an additive form of t_1 and t_2.

4. Simulation Study

We conducted simulations to investigate finite sample performance of the proposed estimator. In the simulation, we let

$$\lambda_{10}(t_1) = \frac{1}{1 + 2t_1},$$

$$\lambda_{20}(t_2) = \frac{1}{1 + 2t_2},$$

$$\lambda_{30}(t_2|t_1) = \frac{2}{1 + t_1 + t_2}.$$

By calculation, it is clear that the stipulated transition functions do not follow the transition functions from the models involving the frailty distribution and Markov or semi-Markov modells ([1,5]). It is therefore of interest to examine whether the proposed spline-based estimation procedure still yields reliable and accurate estimates for this scenario

which cannot be tackled by existing approaches. We report results with one covariate, X, having a uniform. distribution between 0 and 0.5. We consider $\beta_j = 1, -1, 0.5, j = 1, 2, 3$, and $n = 200$ and 400. The censoring time was simulated from from a uniform distribution on $(0, \tau)$ with $\tau = 50$. We compute the spline based semiparametric maximum likelihood estimate using the cubic B-spline and estimate the standard error of the estimated regression parameter using the inverse of the Hessian matrix. For the B-spline, the number of knots K_n or equivalently $N_n = (K_n + m)$ is chosen using BIC defined in Section 2.3. Tables 1–3 presents the estimation bias (BIAS), standard deviations (STD), the mean of the estimated standard error of the estimated regression parameter(ESE) and the coverage proportion of the 95 percent confidence intervals (CP) based on 500 replicates.

Table 1. Simulation results for $(\beta_{10}, \beta_{20}, \beta_{20}) = (1, 1, 1)$.

		BIAS	STD	ESE	CP
$n = 200$	$\beta_1 = 1$	0.021	0.233	0.219	0.953
	$\beta_2 = 1$	−0.016	0.230	0.263	0.954
	$\beta_3 = 1$	0.026	0.281	0.219	0.986
$n = 400$	$\beta_1 = 1$	0.017	0.166	0.159	0.963
	$\beta_2 = 1$	−0.013	0.167	0.164	0.960
	$\beta_3 = 1$	0.018	0.122	0.141	0.965

Table 2. Simulation results for $(\beta_{10}, \beta_{20}, \beta_{20}) = (-1, -1, -1)$.

		BIAS	STD	ESE	CP
$n = 200$	$\beta_1 = -1$	−0.015	0.244	0.225	0.956
	$\beta_2 = -1$	0.019	0.232	0.239	0.962
	$\beta_3 = -1$	−0.014	0.269	0.284	0.961
$n = 400$	$\beta_1 = -1$	−0.013	0.144	0.165	0.961
	$\beta_2 = -1$	0.014	0.158	0.164	0.945
	$\beta_3 = -1$	−0.013	0.197	0.185	0.980

Table 3. Simulation results for $(\beta_{10}, \beta_{20}, \beta_{20}) = (0.5, 0.5, 0.5)$.

		BIAS	STD	ESE	CP
$n = 200$	$\beta_1 = 0.5$	0.017	0.230	0.205	0.966
	$\beta_2 = 0.5$	−0.013	0.221	0.219	0.965
	$\beta_3 = 0.5$	0.016	0.182	0.218	0.945
$n = 400$	$\beta_1 = 0.5$	0.008	0.172	0.155	0.941
	$\beta_2 = 0.5$	−0.011	0.132	0.152	0.954
	$\beta_3 = 0.5$	0.012	0.125	0.157	0.938

From Tables 1–3, we can see (a) the proposed estimates have very small biases; (b) standard deviations of the estimates shrink at approximately the $\sqrt{n}$ rate; (c) the estimated standard deviations are very close to those of the original estimates; the 95 percent confidence intervals provide adequate coverage probabilities. It can be seen that the proposed modeling strategy and estimation procedure can yield reliable and accurate estimates and exhibit direct and good interpretation in practice.

5. A Real Data Example

As our proposed B-spline based modeling strategy does not involve the subjective specification of the frailty distribution and do not require the Markov or semi-Markov assumption which may be unmet in real applications, it is hence more flexible than existing approaches in practice. To illustrate this point, we now apply the illness-death model presented in Section 2 to the colon cancer data. It is of interest to examine whether the time spent in state 1 (past) is related to the transition function from state 2 into state 3. For answering this question, we consider a working model $\lambda_3(t, s) = \exp(\xi t)\lambda(s)$. It

translates to test $H_0 : \xi = 0$. This can be done using the usual likelihood ratio statistic. The results obtained for the colon cancer study show that the effect of time spent in state 1 is significant (p-value < 0.05). This allows us to conclude that the Markov assumption may be unsatisfactory for the colon cancer data set. This further demonstrate the stringent assumptions required by existing approaches may be unmet in practice which calls for the need of our proposed methodology.

For illustrative purposes, we only consider one covariates: Lev+5-FU treatment. Our interest centers on understanding the effect of Lev+5-FU treatment and nonparametricall modelling transition functions in different states. Table 4 reports the estimates of the regression coefficients along with standard errors and p-values. From Figures 1 and 2, we can see our proposed model and estimation procedure yield the estimated transition functions with direct and good interpretation. It stipulates quantitatively how the hazard functions of the time to terminal event and the time to non-terminal event evolves over time and shed lights on the disease progression and death risks for colon cancer patients with and without relapse of the cancer. We plot the estimated the transition functions in Figure 2.

Furthermore, to illustrate the computational advantage of our proposed approach, for the real data application, the existing frailty-model approach will require the number of parameters $(3 + 413 + 1 = 417)$. However, our proposed B-spline approach only require $(m + K_n) * 3 + 3 = (4 + 8) * 3 + 3 = 39$ parameters. Hence, the computational cost is substantially reduced while our approach is more flexible than existing approaches because it does not require the subjective specification of the frailty distribution and the Markov or semi-Markov assumption.

Table 4. Estimated regression coefficients and their standard errors for the colon data.

Transition	Parameters	Estimate	Standard Error	p-Value
12	β_1	-0.513	0.119	1.6×10^{-5}
13	β_1	-0.028	0.379	0.469
23	β_1	0.738	0.130	7.0×10^{-9}

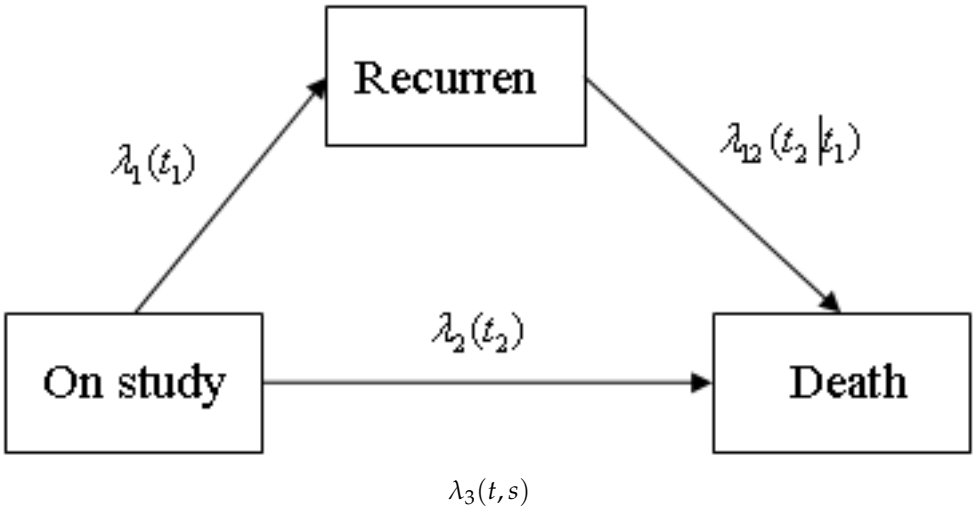
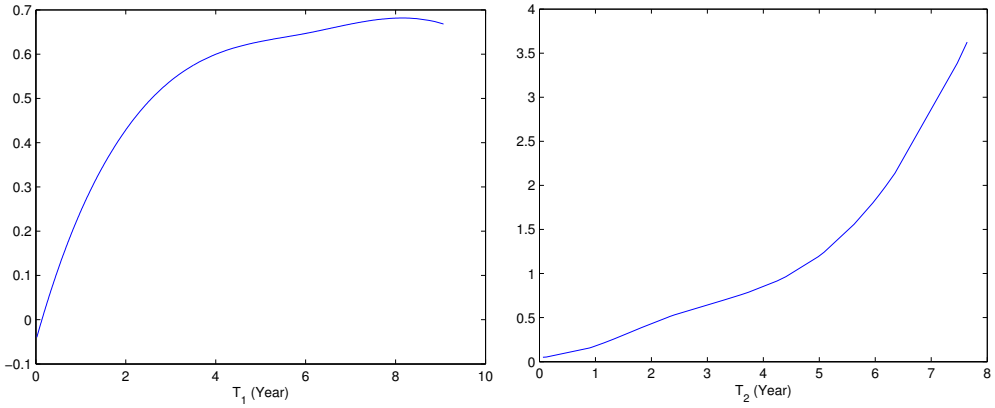
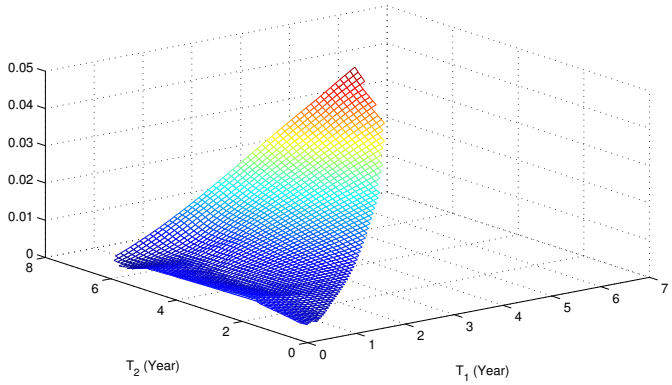


Figure 1. Compartment model for semicompeting risks data.



(**a**) $\lambda_1(t)$ (**b**) $\lambda_2(t)$



(**c**) $\lambda_3(t,s)$

Figure 2. Estimated transition functions for the colon cancer data.

6. Concluding Remarks

In this paper, we proposed an spline-based sieve semiparametric maximum likelihood method for semi-competing risks data. This method reduces the dimensionality of the estimation problem using the splines and therefore releases the numerical burden of the computation. This approach allow essily infer for both regression parameters and transition functions. It should be a straightforward task to apply the method presented here to allow for non-linear relationships between continuous predictors and survival in the multi-state framework ([6,10] and others). Simulations showed that the new estimator may behave very good. For illustration purposes we used a real dataset from a clinical trail for colon cancer. Competing risks data can also be regarded as a special type of multitask prediction problem. In such a field, the most state-of-the-art method is MTPS [4], which currently does not support predicting survival outcomes. Following their approaches, it would be worthwhile studying the stacked algorithm for prediction with multivariate survival outcomes including competing risks and semi-competing risks data.

Author Contributions: Conceptualization, J.X.; methodology, X.H. and J.X.; software, X.H.; formal analysis, X.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proofs of Theorem 1, Theorem 2, and Theorem 3

This section contains the proofs for Theorems 1–3. Some empirical process theorems developed in [11] will be repeatedly used. Throughout the following proofs, we denote

$Pf = \int f(x)dP(x)$ and $P_n f = n^{-1}\sum_{i=1}^{n} f(X_i)$, the empirical process indexed by function $f(X)$.

Appendix A.1. Proof of Theorem 1

By applying the inequality (31) in [12] (p. 31), we have

$$\sup_{\theta\in\Theta_n} \left|P_n l(\theta;W) - Pl(\theta;W)\right| \to 0, a.s. \tag{A1}$$

Let

$$\zeta_{1n} = \sup_{\theta\in\Theta_n} |P_n l(\theta;W) - Pl(\theta;W)|, \tag{A2}$$

$$\zeta_{2n} = P_n l(\theta_0;W) - Pl(\theta_0;W). \tag{A3}$$

Denote $K_\epsilon = \{\theta : d(\theta,\theta_0) \geq \epsilon, \theta \in \Theta_n\}$.

$$\begin{aligned} \inf_{K_\epsilon} Pl(\theta;W) &= \inf_{K_\epsilon}\Big\{Pl(\theta;W) - P_n l(\theta;W) + P_n l(\theta;W)\Big\} \\ &\leq \zeta_{1n} + \inf_{K_\epsilon} P_n l(\theta;W). \end{aligned} \tag{A4}$$

If $\hat{\theta}_n \in K_\epsilon$, we have

$$\begin{aligned} &\inf_{K_\epsilon} P_n l(\theta;W) = P_n l(\hat{\theta};W) \\ \leq\ & P_n l(\theta_0;W) = P_n l(\theta_0;W) - Pl(\theta_0;W) + Pl(\theta_0;W) \\ =\ & \zeta_{2n} + Pl(\theta_0;W). \end{aligned} \tag{A5}$$

By condition A3, we obtain that $\inf_{K_\epsilon} Pl(\theta;W) - Pl(\theta_0;W) = \delta_\epsilon > 0$. It completes the proof.

Appendix A.2. Proof of Theorem 2

Noticing

$$E_P\|n^{1/2}(P_n - P)\|_{\mathcal{F}_\eta} \leq C J_\eta(\varepsilon,\mathcal{F}_\eta,\|\cdot\|_2)\left\{1 + \frac{J_\eta(\varepsilon,\mathcal{F}_\eta,\|\cdot\|_2)}{\eta^2 n^{1/2}}\right\}, \tag{A6}$$

where $J_\eta(\varepsilon,\mathcal{F}_\eta,\|\cdot\|_2) = \int_0^\eta \{1 + \log N_{[]}(\varepsilon,\mathcal{F}_\eta,\|\cdot\|_2)\}^{1/2} d\varepsilon \leq CN^{1/2}\eta$. The right-hand side of (A6) yields $\phi_n(\eta) = C(N^{1/2}\eta + N/n^{1/2})$. It is easy to see that $\phi_n(\eta)/\eta$ decreasing in η, and $r_n^2\phi_n(1/r_n) = r_n N^{1/2} + r_n^2 N/n^{1/2} < 2n^{1/2}$, where $r_n = N^{-1/2}n^{1/2} = n^{-\nu+1/2}$, $0 < \nu < 1/2$. Hence $n^{-\nu+1/2}d(\hat{\theta},\theta_{n0}) = O_P(1)$ by Theorem 3.2.5 of [11]. This, together with $d(\theta_{n0},\theta_0) = O_p(n^{-r\nu})$ (see Theorem 12.7 in [8], yields that $d(\hat{\theta},\theta_0) = O_p(n^{-(1/2-\nu)} + n^{-r\nu})$. This completes the proofs.

Appendix A.3. Proof of Theorem 3

Let ε_n be any positive sequence satisfying $\varepsilon_n = o(n^{-1/2})$. For any $v^* \in \Theta_0$, by [8], Theorem 12.7, there exists $\Pi_n v^* \in \Theta_n$ such that $\|\Pi_n v^* - v^*\| = o(1)$ and $\delta_n\|\Pi_n v^* - v^*\| = o(n^{-1/2})$. Also define $r[\theta - \theta_0; W] = l(\theta;W) - l(\theta_0;W) - \dot{l}(\theta;W)[\theta - \theta_0]$. Then by definition of $\hat{\theta}$, we have

By (A1) and Chebyshev inequality, independent and identical distribution data, and $\|\Pi_n v^* - v^*\| = o(1)$, we have $I_1 = o_p(n^{-1/2})$.

For I_2, we have

$$\begin{aligned} I_2 &= (P_n - P)\Big\{l(\hat{\theta}; W) - l(\hat{\theta} \pm \varepsilon_n \Pi_n v^*; W) \pm \varepsilon_n \dot{l}(\theta_0; W)[\Pi_n v^*]\Big\} \\ &= \mp\varepsilon_n (P_n - P)\Big\{\dot{l}(\tilde{\theta}; W) - \dot{l}(\theta_0; W)[\Pi_n v^*]\Big\}, \end{aligned}$$

where $\tilde{\theta}$ lies between $\hat{\theta}$ and $\hat{\theta} \pm \varepsilon_n \Pi_n v^*$. It follows that $\{\dot{l}(\theta; W)[\Pi_n v^*] : \|\theta - \theta_0\| = O(\delta_n)\}$ is Donsker class. Therefore, by Theorem 2.11.23 of [11], we have $I_2 = \varepsilon_n \times o_p(n^{-1/2})$.

It follows that $\delta_n \|\Pi_n v^* - v^*\| = o(n^{-1/2})$, and $\|\Pi_n v^*\|^2 \to \|v^*\|^2$. Combing the above facts, together with $P\dot{l}(\theta_0; W[v^*]) = 0$, we can establish that

$0 \leq P_n\{l(\hat{\theta}; W) - l(\hat{\theta} \pm \varepsilon_n \Pi_n v^*; W)\} = \mp\varepsilon_n P_n \dot{l}(\theta_0; W)[v^*] \pm \varepsilon_n < \hat{\theta} - \theta_0, v^* > + \varepsilon_n \times o_p(n^{-1/2}) = \mp\varepsilon_n (P_n - P)\{\dot{l}(\theta_0; W)[v^*]\} \pm \varepsilon_n < \hat{\theta} - \theta_0, v^* > + \varepsilon_n \times o_p(n^{-1/2})$.

Therefore, we obtain $\sqrt{n} < \hat{\theta} - \theta_0, v^* >= \sqrt{n}(P_n - P)\{\dot{l}(\theta_0; W)[v^*]\} + o_p(1) \to N(0, \|v^*\|^2)$, where the asymptotic normality is guaranteed by Central limits Theorem and the the asymptotic variance being equal to $\|v^*\|^2 = \|\dot{l}(\theta_0; W)\|^2$. This, together with A5 imply $n^{1/2}(\gamma(\hat{\theta}) - \gamma(\theta_0)) = n^{1/2} < \hat{\theta} - \theta_0, v^* > + o_p(1) \to N(0, \|v^*\|^2)$ in distribution. The semiparametric efficiency can be established by applying the result of [13].

References

1. Fine, J.P.; Jiang, H.; Chappell, R. On semi-competing risks data. *Biometrika* **2001**, *88*, 907–919. [CrossRef]
2. Wang, W. Estimating the association parameter for copula models under dependent censoring. *J. R. Stat. Soc. Ser. Stat. Methodol.* **2003**, *65*, 257–273. [CrossRef]
3. Day, R.; Bryant, J.; Lefkopoulou, M. Adaptation of bivariate frailty models for prediction, with application to biological markers as prognostic indicators. *Biometrika* **1997**, *84*, 45–56. [CrossRef]
4. Xing, L.; Lesperance, M.L.; Zhang, X. Simultaneous prediction of multiple outcomes using revised stacking algorithms. *Bioinformatics* **2020**, *36*, 65–72. [CrossRef]
5. Xu, J.; Kalbfleisch, J.D.; Tai, B. Statistical analysis of illness–death processes and semicompeting risks data. *Biometrics* **2010**, *66*, 716–725. [CrossRef] [PubMed]
6. Andersen, P.K.; Borgan, O.; Gill, R.D.; Keiding, N. *Statistical Models Based on Counting Processes*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012.
7. Kalbfleisch, J.D.; Prentice, R.L. *The Statistical Analysis of Failure Time Data*; John Wiley & Sons: Hoboken, NJ, USA, 2011.
8. Schumaker, L. *Spline Functions: Basic Theory*; Cambridge University Press: Cambridge, UK, 2007.
9. Stone, C.J. Optimal global rates of convergence for nonparametric regression. *Ann. Stat.* **1982**, *10*, 1040–1053. [CrossRef]
10. Meira-Machado, L.; de Uña-Álvarez, J.; Cadarso-Suárez, C.; Andersen, P.K. Multi-state models for the analysis of time-to-event data. *Stat. Methods Med. Res.* **2009**, *18*, 195–222. [CrossRef]
11. Wellner, J. *Weak Convergence and Empirical Processes: With Applications to Statistics*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
12. Pollard, D. *Convergence of Stochastic Processes*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012.
13. Bickel, P.J.; Kwon, J. Inference for semiparametric models: Some questions and an answer. *Stat. Sin.* **2001**, *11*, 863–886.