ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Semiparametric spatial model for interval-censored data with time-varying covariate effects

Yue Zhang [a], Bin Zhang [b,*]

[a] Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, PR China

[b] Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, United States

ARTICLE INFO

Article history:
Received 3 May 2017
Received in revised form 23 January 2018
Accepted 30 January 2018
Available online xxxx

Keywords:
Cox model
Interval censoring
Reversible jump Markov chain Monte Carlo
Smoking cessation data
Spatial correlation
Time-varying coefficient

ABSTRACT

Cox regression is one of the most commonly used methods in the analysis of interval-censored failure time data. In many practical studies, the covariate effects on the failure time may not be constant over time. Time-varying coefficients are therefore of great interest due to their flexibility in capturing the temporal covariate effects. To analyze spatially correlated interval-censored time-to-event data with time-varying covariate effects, a Bayesian approach with dynamic Cox regression model is proposed. The coefficient is estimated as a piecewise constant function and the number of jump points estimated from the data. A conditional autoregressive distribution is employed to model the spatial dependency. The posterior summaries are obtained via an efficient reversible jump Markov chain Monte Carlo algorithm. The properties of our method are illustrated by simulation studies as well as an application to smoking cessation data in southeast Minnesota.

1. Introduction

In medical studies, it is common that patients are only examined periodically at specific points. In such cases, the event time of interest cannot be observed exactly, but is known to fall between certain visits, i.e., the event time is known to fall within a certain time interval. This type of data is known as interval-censored data. For example, if we are interested in the time to smoking relapse of smokers who have ever quit smoking, but the smokers are monitored on an annual basis, then the exact time to smoking relapse is hard to record. Thus, the data are interval-censored, i.e. the time is only known to fall between two visits. Examples of interval-censored data can be found in De Gruttola and Lagakos (1989), Jewell et al. (1994), and Sun (1996).

Different models have been introduced to analyze interval-censored data, among which the proportional hazards (PH) model is the most popular. The PH model, also referred to as the Cox model (Cox, 1972), specifies the multiplicative effect of the covariates on the hazard function of the failure time. Many approaches have been developed under the Cox model over the past several decades (Finkelstein, 1986; Pan, 1999, 2000; Cai and Betensky, 2003). However, they assume that the effect of a risk factor does not change over time. This assumption does not always hold and may lead to biased estimates for covariate effects that very over time. In order to capture the temporal covariate effects, Sinha et al. (1999) treated the unobserved exact times as a latent variable and sampled from their full conditional posterior distribution via Gibbs sampling. Wang et al. (2013) applied reversible jump Markov chain Monte Marlo (MCMC) to automatically determine the

* Corresponding author.
E-mail address: Bin.Zhang@cchmc.org (B. Zhang).

https://doi.org/10.1016/j.csda.2018.01.017
0167-9473/

Please cite this article in press as: Zhang Y., Zhang B., Semiparametric spatial model for interval-censored data with time-varying covariate effects. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.01.017.

dimension of coefficients as well as the baseline hazard function. However, these methods only considered the case where the subjects are independent, which may be violated in many applications.

Zhang et al. (2017) introduced a shared gamma frailty to model the dependence among individuals within the same cluster. However, a limitation is that the model assumes that frailties across clusters are independent, which may not hold in many applications. There exists a large body of studies on the spatially correlated time-to-event data, such as Banerjee et al. (2003), Banerjee and Carlin (2004), and Pan et al. (2014). However, most studies assume a constant coefficient.

This study is motivated by a geographically referenced smoking cessation data set consisting of 223 subjects from 51 zip code areas in southeast Minnesota. The objective is to estimate the effect of prognostic factors on the time to relapse to smoking, after adjusting for the spatial dependency among clusters (zip code areas). Each zip code area forms a spatial cluster. The event of interest is relapse to smoking, which was interval-censored because subjects were only monitored at annual visits for 5 years.

We propose a Bayesian dynamic Cox model to estimate time-varying coefficients for spatially correlated interval-censored data. The proposed model contributes to the current literature on interval-censored data in the following ways: (a) it allows for a flexible estimation of the baseline hazard as well as the temporal effects of risk factors via a Bayesian discretized semi-parametric model; (b) it includes gamma frailties across spatial clusters (based on zip codes) to model heterogeneities across clusters; and (c) it includes a conditional autoregressive (CAR) distribution as the prior for the frailties to adjust for the spatial correlation across clusters.

We compare the proposed model with two other commonly employed models: one is with constant coefficients and spatially correlated frailties, and the other is with time-varying coefficients and independent frailties. In both simulation studies and real data applications, we demonstrate that the proposed model improves the model fitting and leads to better estimation for both the dynamic effects of risk factors and the dependence across spatial clusters.

The remainder of this paper is organized as follows. Section 2 introduces the proposed model and the two comparison models. It also discusses the associated likelihood functions for each of the three models. Section 3 specifies the priors on the regression coefficients, the frailties, and the other parameters in the models. Section 4 includes posterior inference details. Section 5 presents results from simulation studies. Smoking cessation data are analyzed in Section 6. Conclusions and discussions are provided in Section 7.

2. Model specification

Let $T_{i,j}$ denote the survival time for the jth subject in the ith cluster, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m_i$. Consider a Cox model with time-varying regression coefficients conditional on a Q-dimensional vector of covariates, $\mathbf{x}_{i,j}$, and the unobserved frailty random variable ω_i for the ith cluster. The hazard function can be written as

$$\lambda(t|\omega_i, \mathbf{x}_{i,j}) = \lambda_0(t)\exp(\mathbf{x}_{i,j}^T\boldsymbol{\beta}(t) + \omega_i),$$

where $\lambda_0(.)$ is an unknown baseline hazard function common to all subjects, $\mathbf{x}_{i,j}$ is the $Q \times 1$ covariate vector for the jth subject in the ith cluster and $\boldsymbol{\beta}(t)$ is the Q-dimensional regression coefficient function of main interest. The frailty ω_i can be either independent or correlated. In this study, we consider three models specified as follows

Model 1: $\boldsymbol{\beta}(t)$ is constant, ω_i's are spatially correlated.
Model 2: $\boldsymbol{\beta}(t)$ is time-varying, ω_i's are independent.
Model 3: $\boldsymbol{\beta}(t)$ is time-varying, ω_i's are spatially correlated.

For interval-censored data, the unobserved event time $T_{i,j}$ is located in an observed time interval $(L_{i,j}, R_{i,j}]$. The contribution of subject j in ith cluster to the observed data likelihood is then

$$\Pr(T_{i,j} \in (L_{i,j}, R_{i,j}]|\omega_i, \mathbf{x}_{i,j}) = \Pr(T_{i,j} > L_{i,j}|\omega_i, \mathbf{x}_{i,j}) - \Pr(T_{i,j} > R_{i,j}|\omega_i, \mathbf{x}_{i,j}).$$

Due to the unavailability of the partial likelihood, coefficient estimation is challenging since we have to estimate the coefficients and the baseline simultaneously (Sun, 2007). In this study, we borrowed the idea of the Bayesian discretized semiparametric model, which was proposed by Sinha et al. (1999) and also studied by Kim et al. (2007) and Wang et al. (2013). The basic idea is to generate the augmented event time in its observed interval and compute the estimates of parameters piecewisely. Once the unobserved event time $T_{i,j}$'s are given, the interval-censored data reduce to right-censored data. We assume that $\lambda_0(t)$ and $\boldsymbol{\beta}(t)$ are left continuous step functions, where both the number and locations of the jumps are random and to be estimated. Let $k = 1, 2, \ldots, K$ denote all the ordered grids and $0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_K < \infty$ be the corresponding time points. Here we assume the time points $(\tau_k, k = 1, 2, \ldots, K)$ contain all potential jump points. The length of each time interval may be taken to be sufficiently small so that the hazard and coefficient functions can be appropriately estimated. Let $dN_{i,j,k}$ indicate whether or not the event time $T_{i,j}$ falls within the kth interval, i.e., $dN_{i,j,k} = \mathbb{1}(T_{i,j} \in (\tau_{k-1}, \tau_k])$. Let $Y_{i,j,k}$ be the at-risk variable defined as $Y_{i,j,l} = 1$ for $l < k$, $Y_{i,j,l} = 0$ for $l > k$, and $Y_{i,j,k} = (T_{i,j} - \tau_{k-1})/\Delta_k$ for $l = k$, where $\Delta_k = \tau_k - \tau_{k-1}$ is the width of the kth interval. Denote $\lambda_k = \lambda_0(\tau_k)$ and $\boldsymbol{\beta}_k = \boldsymbol{\beta}(\tau_k)$ as the baseline hazard function and the coefficient function evaluated at each time point. Thus, the augmented likelihood function for jth subject of ith cluster is

$$\ell_{i,j}\left(\Theta|\{dN_{i,j,k}, Y_{i,j,k}\}_{k=1}^K, \omega_i, \mathbf{x}_{i,j}\right) = \prod_{k=1}^{K}\left\{\lambda_k\exp(\mathbf{x}_{i,j}^T\boldsymbol{\beta}_k + \omega_i)\right\}^{dN_{i,j,k}}\exp\left\{-\Delta_k\lambda_k\exp(\mathbf{x}_{i,j}^T\boldsymbol{\beta}_k + \omega_i)Y_{i,j,k}\right\}, \quad (1)$$

Please cite this article in press as: Zhang Y., Zhang B., Semiparametric spatial model for interval-censored data with time-varying covariate effects. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.01.017.