Computational Statistics and Data Analysis 56 (2012) 209–216

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

A composite likelihood approach for spatially correlated survival data

Jane Paik [a,*], Zhiliang Ying [b]

[a] Department of Medicine, Stanford University, Stanford, CA 94305, United States
[b] Department of Statistics, Columbia University, New York, NY 10025, United States

ARTICLE INFO

Article history:
Received 29 September 2010
Received in revised form 7 July 2011
Accepted 7 July 2011
Available online 23 July 2011

Keywords:
Spatial dependence
Pairwise joint likelihood
Marginal likelihood
Event times
Consistency
Asymptotic normality
Censoring

ABSTRACT

The aim of this paper is to provide a composite likelihood approach to handle spatially correlated survival data using pairwise joint distributions. With e-commerce data, a recent question of interest in marketing research has been to describe spatially clustered purchasing behavior and to assess whether geographic distance is the appropriate metric to describe purchasing dependence. We present a model for the dependence structure of time-to-event data subject to spatial dependence to characterize purchasing behavior from the motivating example from e-commerce data. We assume the Farlie–Gumbel–Morgenstern (FGM) distribution and then model the dependence parameter as a function of geographic and demographic pairwise distances. For estimation of the dependence parameters, we present pairwise composite likelihood equations. We prove that the resulting estimators exhibit key properties of consistency and asymptotic normality under certain regularity conditions in the increasing-domain framework of spatial asymptotic theory.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Multivariate time-to-event data are subject to spatial correlation in familial and multicenter clinical trials in biomedical sciences, region-wide disease studies in epidemiology and e-commerce studies in marketing (Li and Lin, 2006; Henderson et al., 2002; Li and Ryan, 2002; Banerjee et al., 2003). The practical interest lies in the dependence between the survival outcomes in the geographic domain of interest. In standard geo-statistical practice, the variance and correlation structure of uncensored data are modeled through a parametric covariance function and the parameters are estimated by maximum likelihood (Cressie, 1993). The estimation of parameters poses a challenge, since using a full likelihood approach is computationally burdensome due to high-dimensional integrals. An added challenge in the analysis of time-to-event data prone to spatial correlation is the presence of censoring.

Two widely used methods in modeling associations between failure times are frailty and copula models. Nielsen et al. (1992), Klein (1992), Murphy (1995, 1996), and Parner (1998) investigated the estimation and inference for the frailty model. A comprehensive review can be found in Andersen et al. (1993) and Hougaard (2000). The frailty model-based approach is appealing for family studies since it accommodates the dependencies among relatives by assuming a shared frailty (Parner, 1998; Bandeen-Roche and Liang, 1996; Hsu and Gorfine, 2006). Models developed for clustered data may not fully allow for spatially correlated data. Motivated by a study of asthma onset in Boston, Li and Ryan (2002) used an extended frailty model to take into account the spatial correlation structure. Henderson et al. (2002) investigated survival of leukemia in northwest England by using a multivariate gamma frailty model with a covariance structure allowing for spatial effects.

Genest and MacKay (1986), Oakes (1989) and Shih and Louis (1995) modeled the association of bivariate failure times using copula functions. An attractive feature of the copula model is that the margins do not depend on the choice of the

* Corresponding author.
E-mail address: janepaik@stanford.edu (J. Paik).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.07.004

dependency structure. As a result, one can model and estimate the dependency and margins separately. In a multivariate setting, Li and Lin (2006) developed a method for analyzing survival data correlated in a region by specifying a multivariate normal copula and allowing for a spatial correlation structure in the parameters. They provided an estimating equation approach, avoiding the full likelihood that can be intractable when spatially correlated survival outcomes are involved.

Composite likelihood as proposed in Lindsay (1998) is convenient in the setting where the full likelihood is difficult to construct. Earlier, Besag (1974) considered a similar approach for spatial data. Cox and Reid (2004) provided a general framework for the composite likelihood approach to inference. Composite likelihood methods have been used in multivariate analysis of various types such as non-normal spatial data (Heagerty and Lele, 1998; Varin et al., 2005) and binary correlated data (LeCessie and Van Houwelingen, 1994; Kuk and Nott, 2000). Kuk (2007) considered a weighted composite likelihood for clustered data. Varin (2008) provided a survey of composite likelihood applications.

For survival data, the composite likelihood approach has been used in the analysis of clustered data in familial studies. Parner (2001) modeled the marginal distribution of pairs of failure times using shared frailty models and constructed a pseudo log-likelihood function by adding the pairwise likelihood contributions. Andersen (2004) specified joint survivor functions with copula models and estimated the marginal hazard and association parameters via composite likelihood. Tibaldi et al. (2004a,b) considered composite marginal likelihood inference for multivariate survival data using a Plackett–Dale (Plackett, 1965) model. For estimation, Zhao and Joe (2005) considered a two-stage approach and proposed the use in a multivariate setting in frailty and copula models. These methods are not suited for spatial correlation among the observations since they assume clusters to be independent.

The methodology in this paper is motivated by a problem in e-commerce data, where marketing research has generated much interest in ascertaining whether there is any spatial clustering in purchasing behavior among new customers (Bell and Song, 2007; Bradlow et al., 2005). We present a model for the dependence structure of failure times subject to spatial correlation. We use the Farlie–Gumbel–Morgenstern bivariate family as the survivor function. In order to address the question whether the spatial clustering of purchasing behavior depends on only geographical distances, we model the dependence parameter as a function of Euclidean distances and other pairwise distances. To estimate parameters, we follow a composite likelihood approach for the analysis of spatially correlated survival data. Following Cox and Reid (2004), we use a univariate marginal likelihood to estimate parameters in the hazard function and pairwise composite likelihood for the dependence parameters. The resulting pairwise composite likelihood has a convenient form. We consider the case in which the marginal distribution of failure times follows the parametric Weibull family or the Cox proportional hazards model. We use a two-stage estimation procedure to estimate parameters from the marginal likelihood and then use composite likelihood to estimate dependence parameters. We prove that the resulting estimators exhibit key properties of consistency and asymptotic normality under certain regularity conditions in the increasing-domain spatial asymptotic framework.

In Section 2 we present the model and we present the estimation procedure of composite likelihood along with the asymptotic properties of the estimators. In Section 3 we discuss an application to a marketing study of e-commerce data and conclude with final remarks in Section 4.

2. Method

2.1. Notation and model

In a spatial region of interest, consider a total of n subjects who are followed up to failure or censoring. Let T be the failure time and C the censoring time. Let Z be a p-vector of covariates. Conditional on Z, T and C are assumed to be independent. Let $(T_i, C_i, Z_i, i = 1, \ldots, n)$ be n copies of (T, C, Z). For the ith subject, one can only observe $(\widetilde{T}_i, Z_i, \delta_i, i = 1, \ldots, n)$ where $\widetilde{T}_i = min(T_i, C_i)$ and $\delta_i = 1(T_i \leq C_i)$. Denote by $\lambda_0(t)$ the baseline hazard function and $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ the baseline cumulative hazard function. Let $\Lambda_i(t)$ denote the cumulative hazard function, and let β be a p-dimensional regression coefficient vector. In addition to the covariates, each subject's geographic coordinates (latitude and longitude) are observed. We are interested in the marginal effect of the covariate vector Z on the hazard function as well as the dependence structure among failure times.

A first step towards modeling the dependence structure is to model pairwise joint distributions. We propose using the Farlie–Gumbel–Morgenstern (FGM) family of distributions (Morgenstern, 1956; Gumbel, 1960; Farlie, 1960), given by

$$F_{ij}(t_i, t_j) = F_i(t_i)F_j(t_j)[1 + \xi_{ij}\{1 - F_i(t_i)\}\{1 - F_j(t_j)\}],$$

where $F_i(t_i)$ is the marginal survival function of T_i, and $F_{ij}(t_i, t_j) = P(T_i \geqslant t_i, T_j \geqslant t_j)$. There is a restriction on ξ_{ij} such that $0 \leq |\xi_{ij}| \leq 1$. The parameter ξ_{ij} can be viewed as a measure of dependence. We consider the case in which the marginals follow the Weibull family $F_i(t_i) = \exp[-t_i^{\gamma} \exp(\alpha + \beta' Z_i)]$.

To parameterize the bivariate joint distribution, it suffices to parameterize ξ_{ij} and the univariate marginal survival functions $F_i(t_i)$. To this end, we propose that the dependence parameter ξ_{ij} is itself a function of geographic distance as well as other arbitrary pairwise distances. Specifically, let $\xi_{ij} = \xi_{ij}(\psi; w_{ij})$, where w_{ij} could include the Euclidean distance and certain pairwise characteristics and ψ is a vector of parameters to be estimated.

Specifically, let d_{ij} be the Euclidean distance between the spatial locations of subjects i and j, where $d_{ii} = 0$ by definition, and z_{ij} is a function of demographic variables for units i and j, e.g. an indicator of whether subjects i and j reside in

Download English Version:

https://daneshyari.com/en/article/416512

Download Persian Version:

https://daneshyari.com/article/416512

Daneshyari.com