ARTICLE IN PRESS

Computational Statistics and Data Analysis ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Semiparametric Bayesian joint models of multivariate longitudinal and survival data

Nian-Sheng Tang *, An-Min Tang, Dong-Dong Pan

Department of Statistics, Yunnan University, Kunming 650091, PR China

ARTICLE INFO

Article history:
Received 26 December 2012
Received in revised form 25 November 2013
Accepted 22 February 2014
Available online xxxx

Keywords:
Bayesian case-deletion diagnostic
Centered Dirichlet process prior
Joint models
Longitudinal data
Semiparametric Bayesian analysis
Survival data

ABSTRACT

Joint models for longitudinal and survival data are often used to investigate the association between longitudinal data and survival data in many studies. A common assumption for joint models is that random effects are distributed as a fully parametric distribution such as multivariate normal distribution. The fully parametric distribution assumption of random effects is relaxed by specifying a centered Dirichlet Process Mixture Model (CDPMM) for a general distribution of random effects because of some good properties of CDPMM such as inducing zero mean and continuous probability distribution of random effects. A computationally feasible Bayesian case-deletion diagnostic based on the ϕ-divergence is proposed to identify the potential influential cases in the joint models. Several simulation studies and a real example are used to illustrate our proposed methodologies.

1. Introduction

Joint models for longitudinal and survival data (JMLS) are often employed to investigate the association between longitudinal and survival data in many studies such as HIV/AIDS clinical trials, and have received a lot of attention from both the frequentist and Bayesian perspectives. For example, see De Gruttola and Tu (1994), Tsiatis et al. (1995), Faucett and Thomas (1996), Wulfsohn and Tsiatis (1997), Henderson et al. (2000), Wang and Taylor (2001), Xu and Zeger (2001), Law et al. (2002), Song et al. (2002), Chen et al. (2002), Brown and Ibrahim (2003), Tsiatis and Davidian (2004), Brown et al. (2005), Chi and Ibrahim (2006, 2007), Ding and Wang (2008), Song and Wang (2008), Ye et al. (2008), Hu et al. (2009), Rizopoulos et al. (2009), Albert and Shih (2010), Zhu et al. (2012), among others.

All the above mentioned studies have assumed that the random effects in JMLS are distributed as a fully parametric distribution such as a multivariate normal distribution. However, in some applications, the parametric distribution assumption of random effects in JMLS may be questioned (Brown and Ibrahim, 2003; Rizopoulos and Ghosh, 2011) even though parameter estimators in JMLSs are rather robust to random effect misspecification (Rizopoulos et al., 2008). Therefore, relaxing the parametric distribution assumption of random effects in JMLS have been received a lot of attention in past years (Rizopoulos et al., 2008). For example, Song et al. (2002) developed a semiparametric likelihood approach to JMLS via the EM algorithm based on the assumption that the random effects have a smooth density; Brown and Ibrahim (2003) presented a semiparametric Bayesian approach to JMLS by relaxing the distributional assumptions for the longitudinal model via Dirichlet process (DP) (Ferguson, 1973) priors on the parameters defining the longitudinal model. Recently, Rizopoulos and Ghosh (2011) proposed a new semiparametric multivariate JMLS with the single survival outcome by using a natural cubic spline-based method to flexibly capture the possibly nonlinear shapes of the subject-specific evolutions and

* Corresponding author.
E-mail address: nstang@ynu.edu.cn (N.-S. Tang).

http://dx.doi.org/10.1016/j.csda.2014.02.015
0167-9473/

Please cite this article in press as: Tang, N.-S., et al., Semiparametric Bayesian joint models of multivariate longitudinal and survival data. Computational Statistics and Data Analysis (2014), http://dx.doi.org/10.1016/j.csda.2014.02.015

relax the distribution of the random effects via a DP prior. However, they did not consider (1) multivariate survival outcomes in JMLS; (2) detection of the potential influential observations via Bayesian case-deletion approach which is an important step in the analysis of a data set; (3) flexible continuous distribution for the random effects; and (4) zero means of the random effects using the DP prior. Hence, the main purpose of this paper is to develop a novel Bayesian approach to JMLS for fully addressing the above mentioned problems based on the centered Dirichlet Process Mixture Model (CDPMM) specification of random effect distribution because CDPMM allows for more flexibility in modeling the random effect distribution, which is useful for the efficient estimation and for proper model-based variability estimates of the regression coefficients.

The study of this paper is motivated by a data set from a clinical trial conducted by the International Breast Cancer Study Group (IBCSG) (Chi and Ibrahim, 2006). In the IBCSG trial, each premenopausal woman with node-positive breast cancer was randomly assigned in a 2×2 factorial design to receive either the adjuvant chemotherapy or the reintroduction of three single courses of delayed chemotherapy. Different therapeutic procedure may have a direct affect on disease-free-survival (DFS) and overall survival (OS), as well as age, estrogen receptor (ER) status (negative/positive) and the number of positive nodes of the tumor, and the toxicity of a therapeutic procedure may adversely affect a patient's qualify of life (QOL), which is specifically related to DFS and OS. Four indicators of health-related QOL, including physical well-being (lousy-good), mood (miserable–happy), appetite (none-good) and perceived coping ("How much effort does it cost you to cope with your illness?" (a great deal-none)), were assessed at baseline and at months 3 and 18 after randomization. Under the normality assumption of random effects, Chi and Ibrahim (2006) proposed a joint likelihood approach to jointly model multidimensional QOL and the bivariate failure time random variables DFS and OS based on the subset of the IBCSG data set, which were collected from $n = 832$ patients from Switzerland, Sweden and New Zealand/Australia; Zhu et al. (2012) also considered the Bayesian local influence assessment of JMLS for the data set. However, to the best of our knowledge, there is little work done on developing some novel Bayesian case-deletion diagnostic measures to detect the potential influential observations for the data set that was fitted by JMLS using the CDPMM to allow flexible continuous distribution for the random effects.

Bayesian case-deletion diagnostics for detecting the potential influential observations (or sets of observations) have been proposed for many models such as normal linear regression models, mixed-effects models, generalized linear mixed models and survival models based on the Kullback–Leibler (K–L) divergence and the conditional predictive ordinate (CPO). For example, see Carlin and Polson (1991), Zeger and Karim (1991), Bradlow and Zaslavsky (1997), Weiss and Cho (1998), Spiegelhalter et al. (2002), Cho et al. (2009), Fong et al. (2010), Jackson et al. (2012). But, extending these existing Bayesian case-deletion diagnostics to our considered JMLS has computational challenge because of the complexity of the considered models and the unknown distribution of the random effects. To overcome the above mentioned difficulties, Markov chain Monte Carlo (MCMC) algorithm is employed to develop a Bayesian case-deletion influence diagnostic to assess the effect of cases (or sets of observations) on estimations of parameters based on the ϕ-divergence in the paper. A computationally feasible formula for the proposed Bayesian case-deletion diagnostic is also presented because the closed-form for the ϕ-divergence is not available.

The rest of this article is organized as follows. In Section 2, we describe a general semiparametric JMLS by using a centered Dirichlet Process Mixture Model (CDPMM) to specify the distribution of the random effects. Section 3 develops a Bayesian MCMC algorithm to make Bayesian inference on the JMLS by using the stick-breaking prior presentation for the centered DP prior and the blocked Gibbs sampler, together with the Metropolis–Hastings algorithm. Also, a Bayesian case-deletion diagnostic measure is proposed to detect the potential influential observations based on the ϕ-divergence in Section 3. Four simulation studies and an example are used to illustrate our proposed methodologies in Section 4. Some concluding remarks are given in Section 5. Technical details are presented in the Appendix A.

2. Model and notation

2.1. Notation

Consider the data from n independent individuals. For each individual, we consider K longitudinal responses and M time-to-event outcomes. Suppose that y_{ijk} is the observation of the kth longitudinal response measured at time t_{ijk} for the ith individual and $\mathbf{Y}_{ik} = (y_{i1k}, \ldots, y_{in_{ik}k})^T$ is the observed longitudinal process for the kth response for $i = 1, \ldots, n$, $k = 1, \ldots, K$ and $j = 1, \ldots, n_{ik}$. For the ith individual, we observe the event time $T_{im} = \min(T^*_{im}, C_{im})$ and the event indicator $\delta_{im} = \mathbf{1}(T^*_{im} \leq C_{im})$ for the mth time-to-event outcome for $m = 1, \ldots, M$, where $\mathbf{1}(A)$ is an indicator function of an event A, and T^*_{im} and C_{im} are the true survival time and the censoring time, respectively.

2.2. Generalized linear mixed longitudinal models

Let $\boldsymbol{b}_i = (\boldsymbol{b}^T_{i1}, \ldots, \boldsymbol{b}^T_{iK})^T$ be time-independent random effects underlying both the longitudinal and survival processes for the ith individual. It is usually assumed that $\boldsymbol{b}_i$'s have zero mean (Zhu et al., 2012), which facilitates the interpretation of other fixed effects in the joint model, and all components of the longitudinal outcomes and the time-to-event outcomes given $\boldsymbol{b}_i$ are conditionally independent. Specifically, we suppose that $y_{i1k}, \ldots, y_{in_{ik}k}$ given $\boldsymbol{b}_{ik}$ are conditionally independent and each y_{ijk} given $\boldsymbol{b}_{ik}$ follows the following exponential family distribution

$$p(y_{ijk}|\boldsymbol{b}_{ik}, \phi_k) = \exp[\phi_k^{-1}\{y_{ijk}\vartheta_{ijk} - s(\vartheta_{ijk})\} + c(y_{ijk}, \phi_k)], \quad (1)$$

Please cite this article in press as: Tang, N.-S., et al., Semiparametric Bayesian joint models of multivariate longitudinal and survival data. Computational Statistics and Data Analysis (2014), http://dx.doi.org/10.1016/j.csda.2014.02.015