Computational Statistics and Data Analysis 75 (2014) 53–65

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Bayesian variable selection under the proportional hazards mixed-effects model

Kyeong Eun Lee [a], Yongku Kim [a], Ronghui Xu [b,c,*]

[a] Department of Statistics, Kyungpook National University, Daegu, 702-701, Republic of Korea

[b] Division of Biostatistics and Bioinformatics, Department of Family and Preventive Medicine, University of California, San Diego, USA

[c] Department of Mathematics, University of California, San Diego, USA

ARTICLE INFO

Article history:
Received 28 February 2013
Received in revised form 4 February 2014
Accepted 9 February 2014
Available online 15 February 2014

Keywords:
Correlated survival data
MCMC
Model selection
Multi-center clinical trial
Proportional hazards mixed-effects model
Stochastic search variable selection

ABSTRACT

Over the past decade much statistical research has been carried out to develop models for correlated survival data; however, methods for model selection are still very limited. A stochastic search variable selection (SSVS) approach under the proportional hazards mixed-effects model (PHMM) is developed. The SSVS method has previously been applied to linear and generalized linear mixed models, and to the proportional hazards model with high dimensional data. Because the method has mainly been developed for hierarchical normal mixture distributions, it operates on the linear predictor under the Cox type models. The PHMM naturally incorporates the normal distribution via the random effects, which enables SSVS to efficiently search through the candidate variable space. The approach was evaluated through simulation, and applied to a multi-center lung cancer clinical trial data set, for which the variable selection problem was previously debated upon in the literature.

1. Introduction

Correlated survival data arise in various practical applications including multi-center clinical trials, genetic studies, and recurrent events. In many such applications the data consist of clusters and observations within the clusters. A number of statistical methods have been developed over the last decade to analyze such data. The proportional hazards mixed-effects model (PHMM) was proposed by Ripatti and Palmgren (2000) and Vaida and Xu (2000) to model clustered survival data, which allows cluster specific random effects of arbitrary covariates. Suppose that T_{ij} is the random variable representing the failure time of individual j in cluster i. The PHMM assumes that the hazard function of T_{ij} follows

$$\lambda_{ij}(t) = \lambda_0(t) \exp\left(\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{b}_i\right), \tag{1}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\mathbf{b}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ is a $q \times 1$ vector of cluster specific random effects, $\mathbf{x}_{ij}$ is a $p \times 1$ vector of covariates, and $\mathbf{z}_{ij}$ is typically a $q \times 1$ subvector of $\mathbf{x}_{ij}$, except that $\mathbf{z}_{ij}$ is allowed to contain an element of '1' for a random cluster effect on the baseline hazard.

Under model (1) various inference procedures have been proposed in the literature. Ripatti and Palmgren (2000) considered a penalized partial likelihood approach, which is similar to the penalized quasi-likelihood (PQL) under the generalized linear mixed models. Vaida and Xu (2000) proposed a nonparametric maximum likelihood estimator (NPMLE), obtained using a Monte Carlo EM algorithm. Cortiñas-Abrahantes et al. (2007) considered a Laplace EM algorithm for the

* Correspondence to: 9500 Gilman Drive, La Jolla, CA 92093-0112, USA. Tel.: +1 858 534 6380; fax: +1 858 534 5273.
E-mail address: rxu@ucsd.edu (R. Xu).

http://dx.doi.org/10.1016/j.csda.2014.02.009
0167-9473/

NPMLE. A comprehensive comparison of these methods can be found in Gamst et al. (2009). Although it is reasonably clear to see the advantages and limitations of the different inference procedures, only very recently attention has started to focus on model selection. Under model (1) this concerns the selection of fixed as well as random effects.

Xu et al. (2009) considered the likelihood ratio test under model (1), as well as a profile Akaike information criterion for model selection. Donohue et al. (2011) developed a conditional Akaike information criterion, where the focus is on the estimation of the fixed as well as the random effects. Under the special case of frailty models where $\mathbf{z}_{ij}$ is restricted to either 0 or 1, Fan and Li (2002) considered selection of the fixed effects. Gray (1995) and Commenges and Andersen (1995) developed score tests for no random effects in the frailty model, although it is also possible to generalize the score tests to test for no random effects of additional covariates under model (1) via stratification (Gray, 2006). Dunson and Chen (2004) also considered selection of random effects under the gamma frailty model, using a Bayesian approach. Interestingly Dunson and Chen (2004) arrived at a different conclusion from the score tests of Gray (1995), on the data from a multi-center clinical trial in lung cancer, which will be further discussed in this paper.

Stochastic search variable selection (George and McCulloch, 1993, SSVS) is an approach based on the Bayesian hierarchical normal mixture setup under a regression model, where latent variables are used to indicate the inclusion or exclusion of a potential predictor. It uses Gibbs sampler to sample from a multinomial distribution on the set of possible subset choices, and the promising subsets of predictors are identified as those with high posterior probabilities. As will be described below, SSVS avoids the overwhelming problem of calculating the posterior probabilities of all 2^p subsets, and is computationally fast and efficient. The SSVS method has been extended to linear and generalized linear mixed models (Chen and Dunson, 2003; Kinney and Dunson, 2007), and to survival models (Lee and Mallick, 2004). Because of its ability to select among a larger number of potential predictors, it has been applied to high dimensional data including genomics and other complex disease risk factor studies (Beattie et al., 2002; Lee et al., 2003; Swartz et al., 2008; Lin and Huang, 2008).

In the following we develop the SSVS under the general PHMM (1), for selection of both fixed and random effects of arbitrary covariates. There has been no Bayesian approach to this problem in the literature, which has the advantage of subsequent model averaging that can take into account model uncertainty and selection bias. In Section 3 we examine the performance of SSVS using simulations. We apply the approach to the multi-center lung cancer clinical trial data set that was previously analyzed in Gray (1995) and Dunson and Chen (2004) in Section 4. The last section contains further discussion, and all the posterior computation details are given in the Appendix.

2. Variable selection under the PHMM

For clusters $i = 1, \ldots, n$, and observations $j = 1, \ldots, n_i$, denote t_{ij} the observed, possibly right-censored failure time, $\delta_{ij} = 1$ if t_{ij} is an observed failure time, and 0 otherwise. Let N be the total number of observations, that is, $N = \sum_{i=1}^{n} n_i$.

Under model (1) $\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{b}_i$ is the linear predictor, or the prognostic index, which determines the relative risk of an individual. It is an intermediate quantity analogous to the response in a linear model, which in this case associates the predictors with the ultimate survival outcome. Since the SSVS was initially developed for the hierarchical normal mixture distributions, Lee and Mallick (2004) considered adding a small random quantity $\epsilon_{ij} \sim N(0, \sigma^2)$ to the linear predictor. The resulting model is then

$$\lambda_{ij}(t) = \lambda_0(t) \exp(\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{b}_i + \epsilon_{ij}). \tag{2}$$

The ϵ_{ij}'s may be viewed as an individual heterogeneity term which can improve the fit of the model to the data (O'Quigley and Stare, 2002). But the consideration here is mainly computational, because it simplifies the posterior computation as described below and allows the Gibbs sampler to efficiently search through the model space. We should still consider data as generated under model (1), while model (2) is a working model; this is also reflected in our simulations later: while data were generated under model (1), we follow the approach described below to do variable selection and estimation. We should mention that the identifiability of model (2) is similar to the individual frailty models considered in Kosorok et al. (2001), and can also be more intuitively seen from the equivalent transformation model formulation: $h(T_{ij}) = -\mathbf{x}_{ij}'\boldsymbol{\beta} - \mathbf{z}_{ij}'\mathbf{b}_i + e_{ij}$, where $e_{ij} = e_{0ij} - \epsilon_{ij}$, e_{0ij} has a fixed, known extreme value distribution with $\mathrm{Var}(e_{0ij}) = 1.645$, and $h(t) = \log \Lambda_0(t)$ where $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ is the cumulative baseline hazard function.

For notational purposes, let $\mathbf{X}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{in_i})'$, $\mathbf{Z}_i = (\mathbf{z}_{i1}, \mathbf{z}_{i2}, \ldots, \mathbf{z}_{in_i})'$, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{in_i})'$ for $i = 1, 2, \ldots, n$. Also let $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_n')'$, $\mathbf{Z} = \mathrm{diag}\{\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n\}$, $\mathbf{b} = (\mathbf{b}_1', \mathbf{b}_2', \ldots, \mathbf{b}_n')'$, and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1', \boldsymbol{\epsilon}_2', \ldots, \boldsymbol{\epsilon}_n')'$. Finally let $W_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{b}_i + \epsilon_{ij}$, $\mathbf{W} = (W_{11}, W_{12}, \ldots, W_{nn_n})'$, $\mathbf{t} = (t_{11}, \ldots, t_{nn_n})'$, $\boldsymbol{\delta} = (\delta_{11}, \ldots, \delta_{nn_n})'$, and $\mathbf{Y} = (\mathbf{t}, \boldsymbol{\delta})$ which denotes the observed survival data. Then we have:

$$\mathbf{W} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_N), \ \mathbf{b} \sim N(\mathbf{0}, \mathbf{I}_n \otimes \boldsymbol{\Sigma}), \tag{3}$$

where $\boldsymbol{\Sigma}$ is positive semi-definite as it may include variance components that should be excluded from the final selected models, $\bigotimes$ denotes the Kronecker product, and $\mathbf{I}_n$ denotes an $n \times n$ identity matrix.

The SSVS uses latent binary variables $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$ to indicate the inclusion or exclusion of a fixed effect: $\gamma_k = 1$ if $\beta_k \neq 0$ and 0 otherwise, $k = 1, \ldots, p$. Given $\boldsymbol{\gamma}$, let $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ consist of all nonzero elements of $\boldsymbol{\beta}$, and let $\mathbf{X}_{\boldsymbol{\gamma}}$ be the columns of $\mathbf{X}$ corresponding to the elements of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. After specifying the prior distribution for $\boldsymbol{\gamma}$, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and other parameters, one uses the observed data likelihood and Markov chain Monte Carlo (MCMC) to sample from the posterior distribution of