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

Variable selection for high-dimensional genomic data with censored outcomes using group lasso prior

Kyu Ha Lee [a,b,*], Sounak Chakraborty [c], Jianguo Sun [c]

[a] Epidemiology and Biostatistics Core, The Forsyth Institute, Cambridge, MA, USA
[b] Department of Oral Health Policy and Epidemiology, Harvard School of Dental Medicine, Boston, MA, USA
[c] Department of Statistics, University of Missouri-Columbia, Columbia, MO, USA

ARTICLE INFO

Article history:
Received 8 March 2016
Received in revised form 19 February 2017
Accepted 20 February 2017
Available online xxxx

Keywords:
Accelerated failure time model
Bayesian lasso
Gibbs sampler
Group lasso
Penalized regression

ABSTRACT

The variable selection problem is discussed in the context of high-dimensional failure time data arising from the accelerated failure time model. A data augmentation approach is employed in order to deal with censored survival times and to facilitate prior-posterior conjugacy. To identify a set of grouped relevant covariates, a shrinkage prior distribution is specified for the regression coefficients mimicking the effect of group lasso penalty. It is noted that unlike the corresponding frequentist method, a Bayesian penalized regression approach cannot shrink the estimates of coefficients to exact zeros in general. Toward resolving the issue, a two-stage thresholding method that exploits the scaled neighborhood criterion and the Bayesian information criterion is devised. Simulation studies are performed to assess the robustness and performance of the proposed method in terms of variable selection accuracy and predictive power. The method is successfully applied to a set of microarray data on the individuals diagnosed with diffuse large B-cell lymphoma. In addition, an R package called `psbcGroup`, which can be downloaded freely from CRAN, is developed for the implementation of the methods.

1. Introduction

Survival analysis with high-dimensional covariates has received substantial attention in the last few years. With the increasing ability for obtaining gene expression data for patients, genetic signatures have started to play a more important role than pathological outcomes in the study of the molecular portrait of a disease and the prediction of potential survival time (Sotiriou and Piccart, 2007). A comprehensive molecular and genetic profiling of the disease has a huge potential toward developing personalized medicine and targeted therapies.

In survival analysis, the most commonly used regression model is perhaps the proportional hazards (PH) model (Cox, 1972). On the other hand, it is well-known that the PH model has some restrictions and may not fit the data well sometimes. Also in some contexts, the proportionality structure assumed by the model makes it hard to interpret regression coefficients (Wei, 1992; Hernán, 2010; Uno et al., 2015). To address these, many other models have been proposed and among them, one is the accelerated failure time (AFT) model (Prentice, 1978; Buckley and James, 1979; Koul et al., 1981). One important feature of the AFT model is that it provides a more direct physical interpretation on regression parameters (Collett, 2003) since it has a structure similar to the ordinary linear regression model and directly links the failure time of interest to covariates.

* Correspondence to: 245 First St, Cambridge, MA 02142, USA. Fax: +1 617 262 4021.
E-mail address: klee@forsyth.org (K.H. Lee).

http://dx.doi.org/10.1016/j.csda.2017.02.014
0167-9473/

Please cite this article in press as: Lee, K.H., et al., Variable selection for high-dimensional genomic data with censored outcomes using group lasso prior. Computational Statistics and Data Analysis (2017), http://dx.doi.org/10.1016/j.csda.2017.02.014

Variable selection in the high-dimensional setting has recently received a great deal of attention and many methods have been developed for it. Among them, one of the early and commonly used approach is the lasso penalized approach (Tibshirani, 1996) and by following it, other authors proposed a few other penalty functions including elastic-net (Zou and Hastie, 2005), grouped lasso (Yuan and Lin, 2006), and fused lasso (Tibshirani et al., 2005). Note that most of these and other existing methods were developed for either continuous response or binary response data, and also all of the penalization methods mentioned above are not based on any probabilistic framework. In addition, there are several penalization methods developed in the Bayesian paradigm (Park and Casella, 2008; Bornn et al., 2010; Kyung et al., 2010), which has the advantage that allows one to incorporate substantive prior information in the analysis. Also a Bayesian framework enables one to obtain the posterior predictive distribution and thus to readily quantify the uncertainty related to the prediction. Nevertheless, there only exists limited literature on Bayesian variable selection for high-dimensional data with the response variable of interest being a failure time or suffering right-censoring.

A few methods have been proposed for variable selection based on right-censored failure time data based on the Cox PH model (Tibshirani, 1997; Gui and Li, 2005; Bøvelstad et al., 2007; Tibshirani, 2009) and a nice comparative study for the methods can be found in Bøvelstad et al. (2007). Huang et al. (2006) and Engler and Li (2009) discussed some penalized approaches for the failure time data arising from the AFT model with a large number of predictors. Also in the Bayesian paradigm, Lee et al. (2011, 2015) developed variable selection methods for the PH model by using various shrinkage priors. Note that another traditional way to impose sparsity in the Bayesian paradigm is to specify spike and slab (a mixture of Normal and point mass distribution at zero) priors on regression coefficients. This is otherwise known as the stochastic search variable selection technique, studied extensively in George and McCulloch (1993, 1997) and applied in many applications (Brown et al., 1998; Sha et al., 2006; Hernández-Lobato et al., 2013; Narisetty and He, 2014; Lee et al., in press). Newcombe et al. (2014) proposed a Bayesian variable selection method based on a reversible jump implementation of the Weibull model. Most recently, Zhang et al. (in press) developed a variable selection approach using lasso under AFT model, which studied the grouping structure among errors.

In the following, we will consider the analysis of right-censored failure time data arising from the AFT model and present a Bayesian method with the use of a shrinkage prior and the focus on the identification of a subset of important covariates related to the failure time of interest in the high dimensional situation ($n \ll p$). In the method, to induce the sparsity of the model and group variable selection, we will utilize the scale mixture of normal and gamma distributions for regression coefficients that mimics the role of group lasso penalty and a data augmentation technique for the imputation of the censored failure times (Tanner and Wong, 1987; Komárek and Lesaffre, 2007). The prior specification along with the data augmentation approach can facilitate the prior-posterior conjugacy in the proposed Bayesian framework, and thus we can estimate the posterior distributions of parameters via typical Gibbs sampling without employing any complex Monte Carlo methods. Also the tuning parameter that controls the sparsity will be automatically adjusted and updated by specifying a hyperprior distribution. Note that although useful for shrinking regression parameters corresponding to irrelevant covariates toward zero, the Bayesian lasso does not have the ability to produce exact zero estimates like frequentist counterpart. In order to perform the variable selection, we will develop a two-stage thresholding method that accommodates both the posterior distribution of regression parameters and a goodness-of-fit.

The remainder of the paper is organized as follows. We will begin in Section 2 with introducing a motiving example, the diffuse large B-cell lymphoma (DLBCL) microarray study. In particular, some preliminary analysis results are provided and suggest that the PH model may not fit the data appropriately. Section 3 will present the proposed Bayesian inference procedure and Section 4 will discuss the variable selection problem and describe a thresholding method. In Section 5, simulation studies are conducted to assess the performance of the proposed methodology, especially the robustness, and the results suggest that the approach works well in practice. Section 6 applies the methodology to the DLBCL microarray study and some discussion and future research directions are given in Section 7.

2. Diffuse large B-cell lymphoma microarray study

In this paper, we consider data collected from a DLBCL microarray study. Since molecular features of the tumors influences the mortality of patients with DLBCL after chemotherapy, Rosenwald et al. (2002) developed a molecular predictor of survival by using the gene-expression profiles of the lymphomas. In the study, a Lymphochip cDNA microarray with $p = 7399$ probes were used to monitor $n = 240$ patients. The survival times for 102 subjects were right-censored. We applied the 10 nearest-neighbors method for missing gene expressions in DLBCL data set.

Rosenwald et al. (2002) identified individual genes whose expression are correlated with patients' survival by using a Cox PH model. The PH model assumes constant hazard ratios over time. However, when the proportionality assumption is violated, the application of the Cox PH model has the potential to bias the results that may lead to the loss of power in estimation and inference of the prognostic factors on mortality (Therneau et al., 1990). Therefore, we check the PH assumption by adopting the Weibull heteroscedastic hazards regression model (Hsieh, 2001; Nikulin et al., 2006). This approach permits the shape parameter of the Weibull baseline hazard function to depend upon covariate values. Let T denote survival time and $\boldsymbol{x}$ a p-dimensional covariate vector. Specifically, we consider the following heteroscedastic Weibull model:

$$\lambda(t_i; \boldsymbol{x}_i, \boldsymbol{z}_i) = \eta\kappa_0 e^{\boldsymbol{z}_i^\top \boldsymbol{\beta}^*} t_i^{\kappa_0 e^{\boldsymbol{z}_i^\top \boldsymbol{\beta}^*} - 1} e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}^0}, \quad (1)$$

Please cite this article in press as: Lee, K.H., et al., Variable selection for high-dimensional genomic data with censored outcomes using group lasso prior. Computational Statistics and Data Analysis (2017), http://dx.doi.org/10.1016/j.csda.2017.02.014