Bayesian Variable Selection in the Proportional Hazard Model with Application to DNA Microarray Data

Kyeon Eun Lee[1] **Bani K. Mallick**[2]

[1]*Department of Statistics, Kyungpook National University, Daegu, Korea*

[2]*Department of Statistics, Texas A&M University, Texas, U.S.A.*

Email: artlee@knu.ac.kr, bmallick@stat.tamu.edu

ABSTRACT: In this paper we consider the well-known semiparametric proportional hazards (PH) models for survival analysis. These models are usually used with few covariates and many observations (subjects). But, for a typical setting of gene expression data from DNA microarray, we need to consider the case where the number of covariates p exceeds the number of samples n. For a given vector of response values which are times to event (death or censored times) and p gene expressions (covariates), we address the issue of how to reduce the dimension by selecting the significant genes. This approach enable us to estimate the survival curve when $n << p$. In our approach, rather than fixing the number of selected genes, we will assign a prior distribution to this number. The approach creates additional flexibility by allowing the imposition of constraints, such as bounding the dimension via a prior, which in effect works as a penalty. To implement our methodology, we use a Markov Chain Monte Carlo (MCMC) method. We demonstrate the use of the methodology to diffuse large B-cell lymphoma (DLBCL) complementary DNA (cDNA) data.

1 INTRODUCTION

The advent of DNA microarrays makes possible to provide thousands of gene expression at once (Duggan *et al.*, 1999; Schena *et al.*, 1995) but the main difficulty with microarray data analysis is that the sample size is so small compared to the dimension of the problem (the number of genes). The number of genes for a single individual is usually in the thousands and there are few individuals in the data set. Models for such data be complicated, and computational methods are generally intensive. In this paper we are considering a situation when survival times of (for example) cancer patients are of interest. In this setting, it is of interest to identify the significant genes which are controlling the survival time of the patients. Also we want to estimate the patient survival probabilities after controlling for other covariates such as levels of clinical risk. In this paper, we suggest a gene selection technique using a Bayesian model based variable selection approach for survival data. Typical Bayesian variable selection methods are based on the assumptions of Gaussian distributions for the likelihood and use of mixture priors to obtain marginal distributions (George and McCulloch, 1993). We extend these models to the data context where the responses are time to event. We address the issue of how to select the significant genes as well as assess the survival curves using the Cox proportional hazards model where the sample size n is much more smaller than the number of variables (genes) p. We generalize the Gaussian mixture prior approach in this non-Gaussian framework. For non-Gaussian data it is well known that conjugate priors do not exist for the regression coefficients. The computations are then potentially much harder particularly when sampling the dimension of the model. This is due to possibly strong posterior correlation between the elements of regression parameters such that adding or removing a variable can result in a large drop in the model likelihood unless careful update proposals are made to the coefficients to accommodate the change. Hence, without tailored updates to regression parameter mixing in the MCMC sampler can be poor as moves are rarely accepted. The construction of good proposals is not trivial and depends on both the form of the model as well as on the data. In this paper we exploit the use of a random residual component within the model. The use of a residual component is consistent with the belief that there may be unexplained sources of variation in the data perhaps due to explanatory variables that were not recorded in the original study. By adopting a Gaussian residual effect many of the conditional distributions for the model parameters will be of standard form which greatly aids in the computations. We consider one cDNA data set, B-cell lymphoma data set (Alizadeh *et al.*, 2000) and identify a set of responsible genes which explain the survival time.

2 Variable Selection Model

Let T_i be the survival time (observed or censored) for the ith patient and X_{ij}s are the $p+1$ covariates associated with it. Usually X_{i0} indicates the binary or multi-category phenotype covariate and other X_{ij}'s are p gene expressions from DNA microarray data, which is continuous in nature.

$$\begin{bmatrix} \text{Survivaltime} \\ t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix} \begin{bmatrix} \text{Category} & \text{Gene 1} & \text{Gene 2} & \cdots & \text{Gene p} \\ X_{10} & X_{11} & X_{12} & \cdots & X_{1p} \\ X_{20} & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{n0} & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

The Cox's proportional hazards model (Cox, 1972) assumes that the hazard function consists of two parts: baseline hazard function and nonnegative function of covariates. It is given by

$$h(t|x) = h_0(t)\exp(W),$$

where $h_0(t)$ is the baseline hazard function and $W = x'\beta$ where β is a vector of regression coefficients. The Weibull model in the previous section is a special case of Cox's proportional hazard model with $h_0(t) = \alpha t^{\alpha-1}$. Due to indetermination of baseline hazard function, the proportional hazards (PH) model has adequately adaptable for many applications(Kalbfleisch, 2002). Kalbfleisch (1978) suggested the nonparametric Bayesian method for the PH model. We apply Bayesian variable selection approach to this model. We overcame the computation difficulties by including a random residual component as

$$W_i = x_i'\beta + \varepsilon_i, \quad \varepsilon_i \sim N(0,\sigma^2)$$

where $\mathbf{X}$ is the design matrix with ith column $\mathbf{X}_i$. This introduction of ε enables to generate smaples from full conditionals of all other parameters which is consistent with the belief that there may be unexplained source of variation in the data perhaps departure from the assumption of linearity. Assume that T_i is an independent random variable with conditional survival function

$$P(T_i \geq t_i|W_i,\Lambda) = \exp\{-\Lambda(t_i)\exp(W_i)\}\,(i = 1,\cdots,n).$$

Kalbfleisch(1978) suggested Gamma process (GP) prior for the baseline cumulative hazard function $\Lambda(t)$. The assumption is $\Lambda(t) \sim GP(a\Lambda^*, a)$ where Λ^* is the mean process and a is a weight parameter about the mean (Ibrahim *et al.*, 2001). Kalbfleisch(1978) showed that if $a \approx 0$, the likelihood is approximately proportional to the partial likelihood and if $a \to \infty$, the limit of likelihood is same to the likelihood when the gamma process is replaced by Λ^*. Since $\Lambda(t) \sim \text{Gam}(a\Lambda^*(t), a)$ for given t, the unconditional marginal survival function is obtained by direct integration:

$$\begin{aligned} P(T_i \geq t|W) &= e^{(-re^W)}\frac{r^{a\Lambda^*(t)-1}}{\Gamma(a\Lambda^*(t))a^{-a\Lambda^*(t)}}e^{(-ar)}dr \\ &= \left(\frac{a}{a+\exp(W)}\right)^{a\Lambda^*}. \end{aligned}$$

The joint survival function conditional on Λ is

$$P(T_1 \geq t_1,\cdots,T_n \geq t_n|\mathbf{W},\Lambda) = \exp\{-\sum\Lambda(t_i)\exp(W_i)\}.$$

Using a property of Gamma process, Kalbfleisch (1978) showed that the likelihood with some right censoring is

$$L(\mathbf{W}|\mathbf{D}) = \exp\{-\sum aB_i\Lambda^*(t_i)\}\prod_1^n\{a\lambda^*(t_i)B_i\}^{\nu_i}$$

where

$$\nu_i = \begin{cases} 0 & \text{if } t_i \text{ is right censored} \\ 1 & \text{if } t_i \text{ is a death time} \end{cases},$$

$A_i = \sum_{l\in R(t_i)}\exp(\mathbf{W}_l)$ $(j = 1,\cdots,n)$, $R(t_i)$ is the set of individuals at risk at time $t_i - 0$, $B_i = -\log\{1-\exp(W_i)/(a+A_i)\}$ and $\mathbf{D} = (n,t,\nu)$ denotes the observed data. Now we construct the Gaussian mixture for *beta* to perform the variable selection procedure. Define γ to be an arbitrary $p \times 1$ vector of indicator variables with ith elemenent γ_i such that $\gamma_i = 0$ if $\beta_i = 0$ (the gene is not selected) and $\gamma_i = 1$ if $\beta_i \neq 0$ (the gene is selected). Given γ, let β_γ consists of all nonzero elements of β and let $\mathbf{X}_\gamma$ be the columns of $\mathbf{X}$ corresponding to those elements of γ that are equal to one. To complete the hierarchical model we need to make the prior assumptions:

1. Given γ, the prior for β_γ will be $N(0, c(\mathbf{X}_\gamma'\mathbf{X}_\gamma)^{-1})$, where c is a positive scale factor specified by the user. Smith and Kohn (1996) suggested to choose c between 10 and 100 for linear model problems. We will fix $c = 100$, so that the prior of β_γ, given γ, contains little information about β_γ.

2. The γ_i will be assumed to be a priori independent with $P(\gamma_i = 1) = \pi_i$. The value of π_i will be chosen to be small which will restrict the number of genes in the model. For example, if we have 3000 total number of genes and want to allow only 15 genes due to small sample size then will fix $\pi \equiv 0.005$ to achieve the purpose. In addition, if we have prior knowledge that some genes are more important than others, we can incoporate this easily by assigning larger values of π.

So the prior distributions for variable selection is as follows:

$$\begin{aligned} [\mathbf{W}|\beta_\gamma] &\sim \text{MN}(\mathbf{X}_\gamma\beta_\gamma, \sigma^2\mathbf{I}) \\ [\beta_\gamma] &\sim \text{MN}(0, c\sigma^2(\mathbf{X}_\gamma'\mathbf{X}_\gamma)^{-1}) \\ [\gamma_i] &\sim \text{Bernoulli}(\pi_i) \\ [\sigma^2] &\sim \text{InverseGamma}\left(a_0, \frac{b_0}{2}\right). \end{aligned}$$

where MN is multivariate normal distribution and IG is inverse Gamma distribution. The full condition distributions are presented in the appendix.

3 Example

We applied these methods for finding a set of responsible genes which explain the survival function to a Diffuse Large B-cell lymphoma (DLBCL) data set (Alizadeh *et al.*, 2000). Diffuse large B-cell lymphoma (DLBCL) is one of subtypes of non-Hodgkin's lymphoma. But still patients with this disease had diverse responses to current therapy. So Alizadeh *et al.* (2000) proposed that there should be some different forms of DLBCL and discovered two distinct forms of DLBCL, activated B-like DLBCL and GC-B like, using DNA microarray experiment and hierarchical clustering. They showed that these two subgroups of DLBCL were differentiated from each other by distinct gene expressions of hundreds of different genes and had different survival time patterns. There are 40 patients and expression level measurement for 4513 genes for each patient. We consider the fixed binary covariate $\mathbf{X}_0$ as $X_{i0} = 1$ if ith sample is Activated B-like and $X_{i0} = 0$ if other case for $i = 1,\cdots,40$. Also we have the expression level measurement for a set of genes, so X_{ij} is the normalized log scale measurement of the expression level of jth gene for the ith sample, where $i = 1,\cdots,40$ and $j = 1,\cdots,4513$. To develop the semiparametric model, we choose the baseline function Λ^* as Weibull distribution for the Gamma process, that is, $\Lambda^*(t) = \eta_0 t^{\kappa_0}$. We choose moderate value of the hyperparameter as $a = 10$. The estimates of hyperparameters, η_0 and κ_0, are obtained using estimates of intercept and scale in Survreg function (survReg (formula=Surv (y, censor) $\sim$ 1, dist="weibull") in S+. For our computational convenience, 1000 genes are preselected by a two-sample t-test. We consider several frequent subsets from the MCMC chain and the

top two-genes model comes out to be the best subset with respect to Bayes factor. The survival function in the Cox proportional hazards model is

$$S(t|W) = P(T \geq t|W) = \left(\frac{a}{a+\exp(W)}\right)^{a\Lambda^*}$$

and we exploited the posterior samples for this model to get the Monte Carlo estimate of the function. The posterior estimates of survival curves (solid line) with 5th and 95th survival estimates (dotted line) based on the top two genes are superimposed on the Kaplan-Meier (1958) estimates (dash-dotted line) of survival functions (Figure 1). These plots show that this model is a good fit to both of the subgroup of patients. Rather than a single, parsimonious model, the biologists may interested to bigger families of genes to study relationships and functions. We presented some selected genes based on the marginal frequencies in Table 1. Some of the identified genes are already known to be biologically significant. Since MAPK10 (mitogen-activated protein kinase 10) is connected to TNF (tumor necrosis factor)-a signaling pathway (Decraene *et al.*, 2002), its expression is directly related to tumor. Rimokh *et al.* (1993) showed that FVT1 (follicular variant-translocation gene) is highly expressed in some T-cell malignancies. It would take part in the tumoral process. WASIP (Wiskott-Aldrich syndrome protein-interacting protein) is known to play a role in cortical actin assembly related to lymphocyte function by Ramesh *et al.* (1997).

Heat maps have become popular in the microarray literature, Eisen *et al.* (1998), as graphical representations of the primary data where each point is associated with a color that reflects its value. Increasingly, positive values are represented with reds of increasing intensity and increasing negative values with greens of increasing intensity. A heatmap based on the top two genes in Figure 2 shows that these gene expression pattern is related to survival times and it is distinct between two groups.

Freq	Clone ID	Gene Name
1014	1355868	
910	290230	Interferon consensus sequence binding protein 1
405	814260	Follicular lymphoma variant translocation 1
289	1353111	Phorbolin-like protein MDS019
180	683069	EST
156	1340233	FGD1 family, member 3
156	23173	Mitogen-activated protein kinase 10
154	814601	
132	1335070	ESTs
132	1303587	
124	1337701	WASPI protein
121	824198	Homo sapiens FUSE binding protein 1

Table 1: Responsible Genes Found for Estimating the Survival Function DLBCL Data

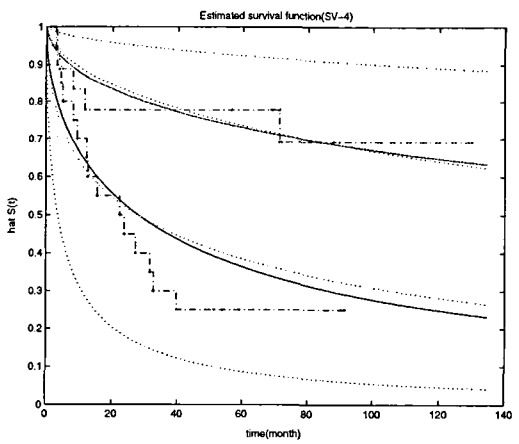

Figure 1: Survival function for DLBCL data using semiparametric hazards model

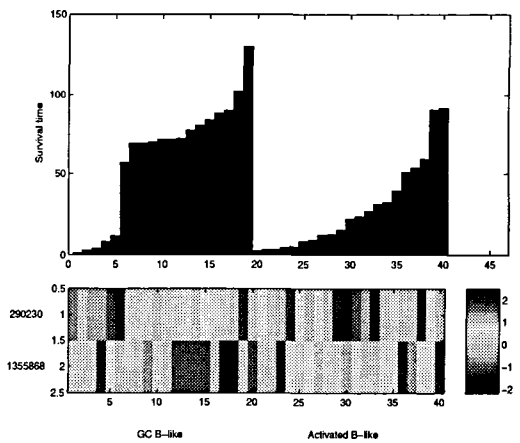

Figure 2: Heat Map with Survival Time for DLBCL data

4 Discussion

We have proposed a Bayesian model for variable selection in the proportional hazard model with specific application to analyze Microarray data. We obtain a nice estimate of the survival curves with an extremely small number of genes. On the other hand, bigger families of genes can be useful to biologists to study the relationship and functions. Information on the size of models for prediction can be easily included in our Bayesian search of good models. The method has flexibility of allowing the location of larger sets of genes, via the inspection of the best visited models or the marginal probabilities of single genes, as we have illustrated.

5 Appendix

5.1 Conditional Distributions in the Cox's Proportional Model

The full conditional distribution of **W** is:

$$p(\mathbf{W}|\mathbf{D},\beta_\gamma,\sigma^2) \propto \exp\{-\sum aB_i\Lambda^*(t_i)\}\prod_1^n\{a\lambda^*(t_i)B_i\}^{\nu_i}$$

$$\times \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{W}-\mathbf{X}_\gamma\beta_\gamma)'(\mathbf{W}-\mathbf{X}_\gamma\beta_\gamma)\right\}$$

In order to get the $p(\gamma|\mathbf{W},\sigma^2)$, we need to integrate out β_γ and the approach is similar to the Weibull regression situation. The marginal distribution of γ given $\mathbf{W}$ and σ^2 is

$$\begin{aligned} p(\gamma|\mathbf{W},\sigma^2) &\propto p(\mathbf{W}|\gamma,\sigma^2)p(\gamma) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}S(\gamma)\right\}\prod_{i=1}^{n}\pi_i^{\gamma_i}(1-\pi_i)^{1-\gamma_i} \end{aligned}$$

where $S(\gamma) = \mathbf{W}'\mathbf{W} - \frac{c}{1+c}\mathbf{W}'\mathbf{X}_\gamma(\mathbf{X}_\gamma'\mathbf{X}_\gamma)^{-1}\mathbf{X}_\gamma'\mathbf{W}$. Now rather than drawing γ as a vector better to draw component wise from $p(\gamma_i|\lambda,\gamma_{ray*}$ and the conditional distribution of σ^2 is Inverse-gamma.

REFERENCES

[1] A. Alizadeh, *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[2] D. R. Cox. Regression models and life tables. *J. R. Statist. Soc. B* 34:187–220, 1972.

[3] C. Decraene, B. Brugg, M. Ruberg, Eveno, *et al.* Identification of genes involved in ceramide-dependent neuronal apoptosis using cDNA arrays, *Genome Biol.* 3(8): research0042.1-research0042.22, 2002.

[4] D. J. Duggan. Expression profiling using cdna micrarrays. *Nature Genetics*,21:10–14, 1999.

[5] E. George, and R. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88:881–889, 1993.

[6] J. G. Ibrahim,M. H. Chen, and D. Sinha. Bayesian Survival Analysis. *Springer*, 2001.

[7] J. D. Kalbfleisch. Non-parametric Bayesian Analysis of Survival Time Data. *J. R. Statist. Soc. B*, 40(2): 214–221,1978.

[8] J. D. Kalbfleisch, and R. L. Prentice. The Statistical Analysis of Failure Time Data. *Wiley-interscience*, 2nd edition, 2002.

[9] E. L. Kaplan, and P. Meier. Nonparametric estimation fro incomplete observations. *J. Am. Stat. Assoc.*, 53:457–481, 1958.

[10] N. Ramesh, I. M. Anton, J. H. Hartwig, R. S. Geha. WIP, a protein associated with wiskott-aldrich syndrome protein, induces actin polymerization and redistribution in lymphoid cells, *Proc Natl Acad Sci* 1997 Dec 23; 94(26):14671–14676, 1997.

[11] R. Rimokh, M. Gadoux, M. F. Bertheas, F. Berger, M. Garoscio, G. Deleage, D. Germain, J. P. Magaud, FVT-1, a novel human transcription unit affected by variant translocation t(2;18)(p11;q21) of follicular lymphoma. *Blood*. 81(1):136–142, 1993.

[12] M. Schena, *et al.* Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, 1995.

[13] M. Smith, and R. Kohn. Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75:317–344, 1997.