Bayesian Variable Selection under the Proportional Hazards Mixed-effects Model

Kyeong Eun Lee*and Yongku Kim†and Ronghui Xu‡

*Department of Statistics, Kyungpook National University, Daegu, 702-701, Korea

†Department of Statistics Yeungnam University, Kyungsan, Korea

‡Division of Biostatistics and Bioinformatics, Department of Family and Preventive Medicine and Department of Mathematics, University of California, San Diego, USA

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. In this paper we develop a stochastic search variable selection (SSVS) approach under the proportional hazards mixed-effects model (PHMM). 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.

Key Words: correlated survival data, MCMC, model selection, multi-center clinical trial, proportional hazards mixed-effects model, stochastic search variable selection

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 hazard 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 a $q \times 1$ subvector of $\mathbf{x}_{ij}$ except possibly a '1' for random 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 NPMLE. A comprehensive comparison of these methods can be found in Gamst *et al.* (2009). Although it is reasonably clear 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 $i = 1, \cdots, n$, and observations $j = 1, \cdots, n_i$, denote t_{ij} the observed, possibly right-censored failure time, δ_{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. 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: $g(T_{ij}) = -\mathbf{x}'_{ij}\boldsymbol{\beta} - \mathbf{z}'_{ij}\mathbf{b}_i + e_{ij}$, where $e_{ij} = e_{0ij} - \epsilon_{ij}$, and e_{0ij} has a fixed, known extreme value distribution with $\text{Var}(e_{0ij}) = 1.645$.

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 $\mathbf{X} = (\mathbf{X}'_1, \mathbf{X}'_2, \ldots, \mathbf{X}'_n)'$, $\mathbf{Z} = \text{diag}\{\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_\text{n}\}$, $\mathbf{b} = (\mathbf{b}'_1, \mathbf{b}'_2, \ldots, \mathbf{b}'_n)'$, 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:

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

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 a $n \times n$ identity matrix.

The SSVS uses latent binary variables $\boldsymbol{\gamma} = (\gamma_1, ...\gamma_p)$ to indicate the inclusion or exclusion of a fixed effect: $\gamma_k = 1$ if $\beta_k \neq 0$ and 0 otherwise, $k = 1, ..., p$. 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 to sample from the posterior distribution of $\boldsymbol{\gamma} = (\gamma_1, ...\gamma_p)$. This gives the marginal probability of inclusion for each fixed effect, and the details are given next. For selection of the random effects, which is equivalent to selection of its covariance matrix structure, a re-parameterization is applied which then makes it similar to the selection of the fixed effects, and is also described in full details in the following.

2.1 Prior specification

The priors for $\boldsymbol{\gamma}$, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and σ^2 are:

$$\gamma_k \sim \text{Bernoulli}(\pi_k), \;\; k = 1, ..., p; \tag{4}$$

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}} | \boldsymbol{\gamma} \sim N\left(\mathbf{0}, \sigma^2(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})^{-1}/g\right); \tag{5}$$

$$g \sim G(1/2, N/2); \tag{6}$$

$$\sigma^2 \sim \frac{1}{\sigma^2}. \tag{7}$$

In the above, the γ_k's are assumed to be a priori independent with $P(\gamma_k = 1) = \pi_k$, $0 \leq \pi_k \leq 1$, for $k = 1, ..., p$. In practice we may take $\pi_k = 0.5$ if there is no prior knowledge about whether a fixed effect should be included, or we may take $\pi_k = 1$ if we want to force a fixed effect into the model. When all $\pi_k = 0.5$ ($k = 1, ..., p$), it is clear that each model $\boldsymbol{\gamma}$ for the fixed effects has a prior probability equal to 2^{-p}. The prior variance of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is taken to be proportional to $\sigma^2(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})^{-1}$, as it results in a fast computing algorithm for the Gibbs sampler; this is also called Zellner's g-prior (Zellner, 1986; Smith and Kohn, 1996). Finally, the improper prior for σ^2 is commonly used such that $\log(\sigma^2)$ is uniform.

To specify the priors for the variance components, Chen and Dunson (2003) considered a modified Cholesky Decomposition of $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \boldsymbol{\Psi}\boldsymbol{\Omega}\boldsymbol{\Omega}'\boldsymbol{\Psi}, \tag{8}$$

where $\boldsymbol{\Psi} = \text{diag}(\psi_1, \ldots, \psi_q)$, and $\boldsymbol{\Omega}$ is a lower triangular matrix with diagonal elements equal to 1. When $\psi_l = 0$ in $\boldsymbol{\Psi}$, the l-th diagonal element of $\boldsymbol{\Sigma}$ is also equal to 0, implying that the l-th random effect is excluded from model (1). The off-diagonal elements of $\boldsymbol{\Omega}$, denoted by $\boldsymbol{\omega}$, represent the dependency among the random effects. Using decomposition (8) we have $\mathbf{W} = \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} + \mathbf{Z}\left(\mathbf{I}_n \bigotimes \boldsymbol{\Psi}\boldsymbol{\Omega}\right)\mathbf{a} + \boldsymbol{\epsilon}$, where $\mathbf{a} = (\mathbf{a}'_1, \mathbf{a}'_2, \ldots, \mathbf{a}'_n)'$, $\mathbf{a}_i \sim N(\mathbf{0}, \mathbf{I}_q)$. Kinney and Dunson (2007) further considered the parameter-expansion (PX) approach of Gelman (2006) for variance components. The over-parameterization in PX reduces dependence among the parameters in a hierarchical model and improves the Gibbs convergence (Liu *et al.*, 1998). Using the PX approach (3) becomes

$$\mathbf{W} = \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}} + \mathbf{Z}\big(\mathbf{I}_n \bigotimes \mathbf{A}\boldsymbol{\Omega}\big)\boldsymbol{\xi} + \boldsymbol{\epsilon}, \;\; \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_N), \; \boldsymbol{\xi}_i \sim N(\mathbf{0}, \mathbf{D}), \tag{9}$$

where $\mathbf{A} = \text{diag}(\alpha_1, \ldots, \alpha_q)$, $\mathbf{D} = \text{diag}(d_1, \ldots, d_q)$, and $\boldsymbol{\xi} = (\boldsymbol{\xi}_1', \boldsymbol{\xi}_2', \ldots, \boldsymbol{\xi}_n')'$. Following Kinney and Dunson (2007) the priors are:

$$\alpha_l \sim \text{ZI-N}^+(0, 1, p_{l0}), \;\; l = 1, ..., q; \tag{10}$$

$$\boldsymbol{\omega}|\boldsymbol{\alpha} \sim N(\boldsymbol{\omega}_0, \mathbf{V}_{\boldsymbol{\omega}}); \tag{11}$$

$$d_l \sim IG\left(\frac{1}{2}, \frac{N}{2}\right), \tag{12}$$

where ZI-N$^+(0, 1, p_{l0})$ represents the mixture distribution putting point mass p_{l0} on $\alpha_l = 0$, and probability $1 - p_{l0}$ on $N^+(0, 1)$ which is the positive part of N(0,1). Just like for the fixed effects $\boldsymbol{\beta}$, we can set the hyperparameters $p_{l0} = 0.5$ for equal prior probabilities to include or exclude a random effect, or we can set $p_{l0} = 1$ to force a random effect in the model. For the other hyperparameters we set $\boldsymbol{\omega}_0 = \mathbf{0}$ and $\mathbf{V}_{\boldsymbol{\omega}} = 0.5\mathbf{I}$.

Finally for the baseline cumulative hazard function $\Lambda_0(t)$ it is common to use a Gamma process (GP) prior (Kalbfleisch, 1978; Clayton, 1991; Ibrahim *et al.*, 2001):

$$\Lambda_0(t) \sim GP(a\Lambda^*(t), a), \tag{13}$$

where Λ^* is the mean process, and a is a weight parameter about the mean. Typically Λ^* is assumed to be a known parametric cumulative hazard function with hyperparameters, and $\lambda^* = d\Lambda^*/dt$ denotes its corresponding hazard function. When there are no random effects in the proportional hazards model and a is close to zero, the resulting marginal posterior of $\boldsymbol{\beta}$ is approximately proportional to the partial likelihood of Cox (1975), while as $a \to \infty$ the Gamma process is effectively replaced by Λ^*, and it becomes the likelihood function of $(\boldsymbol{\beta}, \Lambda^*)$ (Ibrahim *et al.*, 2001). Here we take $\Lambda^*(t) = \eta t^\kappa$ from the Weibull distribution. Following Lee and Mallick (2004) we fix $a = 10$. Since Λ^* is the mean process of the baseline cumulative hazard function, we estimate the hyperparameters η and κ from the data by fitting a Weibull regression model including all covariates.

2.2 The likelihood

Conditional on the random effects, we can integrate out $\Lambda_0(t) \sim \text{Gamma}(a\Lambda^*(t), a)$ at each t, and obtain the likelihood of the survival data $\mathbf{Y}$ marginalized over the prior distribution of the baseline hazard function (Lee and Mallick, 2004). Let $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\omega}, \mathbf{d}, \sigma^2, g)$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)'$ and $\mathbf{d} = (d_1, \ldots, d_q)'$. The resulting likelihood is

$$L(\mathbf{Y}|\mathbf{W}) = \exp\left\{-\sum_{i=1}^{n}\sum_{j=1}^{n_i} aB_{ij}\Lambda^*(t_{ij})\right\}\prod_{i=1}^{n}\prod_{j=1}^{n_i}\{a\lambda^*(t_{ij})B_{ij}\}^{\delta_{ij}}, \tag{14}$$

where $B_{ij} = -\log\{1 - \exp(W_{ij})/(a + A_{ij})\}$, $A_{ij} = \sum_{kl \in R(t_{ij})} \exp(W_{kl})$, and $R(t_{ij})$ is the set of individuals at risk at time $t_{ij}-$ $(j = 1, \ldots, n_i; i = 1, \ldots, n)$. Note that the above likelihood involves $\boldsymbol{\theta}$ only through $\mathbf{W}$. This is the likelihood that will be used to derive the posterior distributions below.

2.3 Posterior computation

Based on the previous description, we can obtain the posterior distribution of interest by

$$p(\boldsymbol{\theta}, \mathbf{W}\big|\mathbf{Y}) \propto L(\mathbf{Y}\big|\mathbf{W})p(\mathbf{W}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{15}$$

As mentioned before $\mathbf{W}$ is an intermediate quantity that associates the predictors with the survival outcome, and here it is viewed more like a parameter in the posterior computation.

To draw inferences about all the parameters of interest as well as model selection, Gibbs samplers or Metropolis-within-Gibbs algorithms are typically implemented. To compute the model posterior distribution, we consider the composite parameter space method of Green and O'Hagan (1998), and tailor it to the context of candidate models with fixed and random effect structures. During an iteration of the procedure, parameters belonging to one part of the model are updated using a standard method, such as a Gibbs or Metropolis-Hastings step, while the other parameters are left unchanged. Our scheme moves around among the indicators for the fixed and the random effects, and the parameters for the fixed and the random effects, as detailed below.

1. Move from a selection of the fixed effects to the next selection of fixed effects by a standard MCMC step. The selection of fixed effects is indexed by the latent binary variables $\boldsymbol{\gamma} = (\gamma_1, ...\gamma_p)$ to indicate the inclusion or exclusion of a fixed effect.

2. Update all fixed effect parameters by a standard MCMC procedure, holding all other parameters unchanged. That is, generate $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ from the full conditional distribution.

3. Move from a selection of ranodm effects to another selection of random effects by a standard MCMC step. Just like for the fixed effect, the index for the random effects is determined by $\alpha_l = 0$ or not, $l = 1, ..., q$.

4. Update all random effect parameters, holding all other parameters unchanged.

The proof of convergence properties as shown by Green and O'Hagan carries over to the algorithm above. All the relevant posterior computations are given in the Appendix. Note that we update each γ_k individually, $k = 1, ..., p$. Here we actually integrate out $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ in (9). A similar approach integrating out both $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and σ^2 was used in Smith and Kohn (1996) to accelerate the convergence of the MCMC chain. We investigated both approaches, and the results were similar. Posteriors for both approaches are given in the Appendix. Each α_l is also updated individually, $l = 1, ..., q$. The zero-inflated truncated normal prior for α_l yields a conjugate posterior.

3. Simulation Experiments

We simulated data under model (1) for various numbers of clusters and cluster sizes (n, n_i). Here we show the results with relatively small n and n_i, to illustrate the type of sample sizes required for the SSVS to properly select the fixed and the random effects. We set $\lambda_0(t) = 1$. Censoring was generated as Uniform$(0, \tau)$, where τ was chosen so that about 20% of the observations were censored in each case. We had $p = 4$ potential covariates, and $\mathbf{x}_{ij} = \left(x_{i1}, x_{i2}, x_{i3}, x_{i4}\right)'$ where each component of $\mathbf{x}$ was generated independently from Uniform (-2, 2). For the random effects, we had $q = 3$, and $\mathbf{z}_{ij} = \left(1, x_{i1}, x_{i2}\right)'$. The true value of the parameters were $\boldsymbol{\beta} = \left(0.8, 0.4, 0.4, 0\right)'$, and $\boldsymbol{\Sigma} = \text{diag}(0.4, 0.2, 0)$. In the tables we used subscript 0, 1 and 2 to indicate the random effects for the baseline hazard, x_1 and x_2, respectively. We also gave the empirical variances of the simulated random effects in parenthesis in addition to the true values of $\boldsymbol{\Sigma}$; the accuracy of the estimated variances can be better reflected when compared to these empirical variances than to the true values. We used non-informative prior for selecting any of the fixed or random effects, that is, $\pi_k = p_{l0} = 0.5$, $k = 1, ..., p$, $l = 1, ..., q$. The MCMC consisted of 10,000 iterations, with the first 10% for burn-in.

The SSVS gives the marginal posterior probability for selecting each of the fixed and random effects. It also gives the posterior probability of each potential model. In Tables 1 and 2 we present the results for the top three selected models B1, B2 and B3, as well as the

averaged estimates from all models and the corresponding 95% credibility intervals. The sample sizes were 20, 30 and 50 clusters, with cluster sizes 10 and 20, respectively.

For the smallest sample size of 20 clusters with 10 observations each, the top one-third of Table 1 shows that all of the top three models missed the random effect on x_1 which had a variance of 0.2. The first two fixed effects were selected with marginal probability one, while the third fixed effect was selected with probability 0.764. The random effect on the baseline hazard which had a variance of 0.4 was selected with probability 0.635. The 95% credibility intervals contained the true values of all parameters except β_3 and Σ_{11}.

When the number of clusters increased to 30 in the middle of Table 1, the true model was chose with probability 0.678. All the 'true' fixed effects were chosen with probability one, and two random effects were chosen with probability 0.77 and 0.934, respectively. There was a slight over selection of the 4th fixed effects with probability 0.212. The 95% credibility intervals contained the true values of all parameters except Σ_{00}, which was under-estimated.

The results for 50×10 are in the bottom one-third of Table 1, where the 4th fixed effect was over selected with probability 0.418. But the true random effects were selected with much higher probabilities (1 and 0.91, respectively) than the previous two scenarios, and the 95% credibility intervals contained the true values of all parameters except Σ_{00}, which was over-estimated in this case.

When the cluster size increased to 20 observations per cluster, even with only 20 clusters in Table 2, the selection results were quite good: the true model was selected with probability 0.891, all the true fixed and random effects were selected with probability one or very close to one (0.968 for Σ_{11}), the null fixed (β_4) and random (Σ_{22}) effects were selected with very low probabilities, and the 95% credibility intervals contained the true values of all parameters.

Finally with larger numbers of clusters as in the middle and bottom of Table 2, the results were even better, with generally tighter credibility intervals, and the true model being selected with probability 0.94 when there were 50 clusters of 20 observations each.

4. An Example

We apply our proposed model to a multi-center advanced stage non-small cell lung cancer clinical trial data which was analyzed in Gray (1994) and Vaida and Xu (2000). The study was conducted by the Eastern Coorperative Oncology Group. There were two randomized treatment arms: a standard chemotherapy (CAV) and an alternating regimens (CAV-HEM) where cycles of CAV were alternated with HEM. The outcome of interest was overall survival. Five binary covariates were found to be significantly associated with survival in the previous published analyses: treatment assignment, presence or absence of bone metastases, presence or absence of liver metastases, performance status at study entry (ambulatory or not), and whether there was weight loss prior to study entry. Gray (1995) found significant institution-to-institution variation in the treatment effects using a score test under the frailty model. Vaida and Xu (2000) fitted model (1) to the data with potential random effects for all five covariates, and found that those for bone metastases were even stronger than the random effects for treatment, while the variances of the random effects of the rest three covariates converged towards zero. Dunson and Chen (2004) considered selection of frailty terms using a Bayesian approach by putting a mixture prior on the frailty variances with point mass at zero and inverse Gamma, and concluded that after accounting for the random bone metastases effects, there was no direct evidence of institutional variation in treatment effects. This then led to a correspondence by Gray (2006) pointing out the statistical significance of the random treatment effects by a score test even after accounting for

the random bone metastases effects, together with a reply by Dunson and Chen who did a separate analysis to support their original conclusion published in 2004.

Here we take another look at the data using the SSVS approach. We consider the 22 institutions with more than 7 enrolled subjects each; this gives a total of 546 patients, and the actual numbers of patients per institution are between 11 and 56. Since there has been consensus in the literature about the significance of the fixed effects for all five covariates, we set the prior probabilities $\pi_k = 1$, $k = 1, ..., 5$. We then consider six potential random effects, on the baseline hazard as well as the five covariates, and set the prior probabilities for the random effects to be $p_{l0} = 0.5$, $l = 0, 1, ..., 5$. Like in the simulations we run 10,000 MCMC iterations, with the first 20% for burn-in. The results of SSVS are shown in Table 3. The top model with only random bone metastases effects is chosen 99.4% of the time, while the random treatment effect has basically an inclusion probability of zero.

To better understand the behavior of SSVS in this case, we carry out further simulations in Tables 4 and 5 mimicking the lung cancer data. The covariates as well as the sample sizes including the number of clusters and the numbers of observations in each cluster for both tables are the same as in the lung cancer data. Recall that in the simulations of Section 3 all the covariates were continuously distributed as Uniform(-2, 2), with a variance of 4/3. For binary (0, 1) covariates, however, the variance is only 1/4. We can only compare the strength of any effect when the corresponding covariates are on the same scale, since we can otherwise always multiple the effect by a non-zero constant and divide the covariate by the same constant and the model is unchanged. In Tables 4 the strength of the random effects as reflected in their variances Σ_{11} and Σ_{22} are comparable to those estimated from the lung cancer data, while in Table 5 they are increased to be equivalent to those for the Uniform(-2,2) covariates as in Section 3 ($0.2 \times 16/3 = 16/15$, $0.4 \times 16/3 = 32/15$). It is clear from the two tables that when both random effects are strong as in Table 5, the SSVS will select both random effects with probability one; but with the level of strength as in the lung cancer data, the SSVS selects only the stronger of the two random effects.

The above investigation might provide some explanation for the discrepancy between the frequentist score test as mentioned before and the Bayesian variable selection for the lung cancer data. While the score test detects significant institutional variation in treatment effects after having accounted for the random bone metastases effects, the random treatment effects are weak such that in simulation studies the Bayesian variable selection chooses not to model it. From the point of view of model selection, it then depends on the criterion that is important to the question of concern according to which one chooses to model the random treatment effect or not.

5. Discussion

In this paper we have developed the Bayesian SSVS approach for selection of fixed as well as random effects under the PHMM. To apply the SSVS, we have added the ϵ_{ij}'s to the linear predictor in the PHMM which expands the model to allow for individual heterogeneity. Our simulation results show that this approach works well even when the data have no such heterogeneity. For the prior distribution of $\sigma^2 = \text{Var}(\epsilon_{ij})$, we have also considered truncated inverse-Gamma, the simulation results (data not shown) depended on the range of truncation and were generally not better than the uniform prior described in details in this paper.

For estimation under the PHMM using maximum likelihood, the EM-type algorithm, although have been shown to be numerically stable and accurate (Gamst *et al.*, 2009), is also known to be slow to converge when the variance of a random effect is very close to zero. The variable selection method developed in this paper may help to 'declare' zero for

those random effects.

There are many approaches to model selection in general, although few have been adapted under the PHMM. The SSVS has the advantage of subsequent Bayesian model averaging that takes into account model uncertainty and selection bias. George and Foster (2000) discussed the connection between Bayes variable selection and other commonly used types of information criteria such as AIC, BIC, etc. The explicit connection under the PHMM is an open problem to be explored.

Our simulation experiments were carried out with moderate sample sizes. A notable phenomenon has been that the cluster sizes appear to have more impact on the performance of SSVS than the number of clusters: larger cluster sizes have substantially improved the variable selection. We note that the approach has not worked well for clusters as small as 5 observations each, and this precludes its application to certain data types such as from twins. Model selection under the PHMM for those cases still remains an area to be further studied.

Acknowledgement

The research of Kyeong Eun Lee was supported by Basic Science Research Program through the National Research Foundation (NRF) of Korea funded by the Ministry of Education, Science and Technology (2010-0023305). The research of Ronghui Xu was partially supported by the United States National Institutes of Health NCRR Clinical and Translational Science Award 1UL1RR031980.

Appendix

The Appendix is available from the authors and not included here due to page limitation.

References

Beattie, S. D., Fong, D. K. H., and Lin, D. K. J. (2002). A two-stage bayesian model selection strategy for supersaturated designs. *Technometrics*, **44**, 55–63.

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics*, **59**, 762–769.

Clayton, D. G. (1991). A Monte Carlo method for Bayesian inference in frailty models. *Biometrics*, **47**, 467–485.

Commenges, D. and Andersen, P. (1995). Score test of homogeneity for survival data. *Lifetime Data Analysis*, **1**, 145–156.

Cortiñas-Abrahantes, J., Legrand, C., Burzykowski, T., Janssen, P., Ducrocq, V., and Duchateau, L. (2007). Comparison of different estimation procedures for proportional hazards model with random effects. *Computational Statistics and Data Analysis*, **51**, 3913–3930.

Cox, D. (1975). Partial likelihood. *Biometrika*, **62**, 269–276.

Donohue, M. C., Overholser, R., Xu, R., and Vaida, F. (2011). Conditional Akaike information under generalized linear and proportional hazards mixed models. *Biometrika*, **98**, 685–700.

Dunson, D. B. and Chen, Z. (2004). Selecting factors predictive of heterogeneity in multivariate event time data. *Biometrics*, **60**, 352–358.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics*, **30**, 74–99.

Gamst, A., Donohue, M., and Xu, R. (2009). Asymptotic properties and empirical evaluation of the *npmle* in the proportional hazards mixed-effects model. *Statistica Sinica*, **19**, 997–1011.

Gelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, **1**, 515–533.

George, E. I. and Foster, D. P. (2000). Calibration and empirical bayes variable selection. *Biometrika*, **87**, 731–747.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.

Gray, R. (1994). A Bayesian analysis of institutional effects in a multicenter cancer clinical trial. *Biometrics*, **50**, 244–253.

Gray, R. (1995). Tests for variation over groups in survival data. *Journal of the American Statistical Association*, **90**, 198–203.

Gray, R. (2006). Correspondence (Re: Dunson and Chen, 2004). *Biometrics*, **62**, 623–624.

Green, P. and O'Hagan, A. (1998). Model choice with mcmc on product spaces without using pseudo-priors. *Nottingham University Statistics Research Report 98-01*.

Ibrahim, J. G., Chen, M., and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer-Verlag, New York.

Kalbfleisch, J. D. (1978). Nonparametric bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series B*, **40**, 214–221.

Kinney, S. K. and Dunson, D. B. (2007). Fixed and random effects selection in linear and logistic models. *Biometrics*, **63**, 690–698.

Kosorok, M. R., Lee, B. L., and Fine, J. P. (2001). Semiparametric inference for proportional hazards frailty regresssion models. Technical report, Department of Biostatistics, University of Wisconsin.

Lee, K. E. and Mallick, B. K. (2004). Bayesian methods for variable selection in survival models with application to dna microarray data. *Sankhya: The Indian Journal of Statistics*, **66**, 756–778.

Lee, K. E., Sha, N., Dougherty, E. R., Vannucci, M., and Mallick, B. K. (2003). Gene selection: a bayesian variable selection approach. *Bioinformatics*, **19**, 90–97.

Lin, E. and Huang, L. C. (2008). Identification of significant genes in genomics using bayesian variable selection methods. *Advances and Applications in Bioinformatics and Chemistry*, **1**, 13–18.

Liu, C., B, R. D., and Wu, Y. N. (1998). Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika*, **85**, 755–770.

O'Quigley, J. and Stare, J. (2002). Proportional hazards models with frailties and random effects. *Statistics in Medicine*, **21**, 3219–33.

Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, **56**, 1016–1022.

Smith, M. and Kohn, R. (1996). Nonparametric regression using bayesian variable selection. *Journal of Econometrics*, **75**, 317–343.

Swartz, M. D., Yu, R. K., and Shete, S. (2008). Finding factors influencing risk: comparing variable selection methods applied to logistic regression models of cases and controls. *Statistics in Medicine*, **27**, 6158–6174.

Vaida, F. and Xu, R. (2000). Proportional hazards model with random effects. *Statistics in Medicine*, **19**, 3309–3324.

Xu, R., Vaida, F., and Harrington, D. (2009). Using profile likelihood for semiparametric model selection with application to proportional hazards mixed models. *Statistica Sinica*, **19**, 819–842.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distribution. *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti*, pages 233–243.

Table 1: Simulation results with $n_i = 10$; B1, B2 and B3 are top three selected models.

	Parameter	True Value	B1	B2	B3	Estimate	95% CI	Pr(Inclusion)
$n = 20$	β_1	0.8	0.882	0.964	0.828	0.896	(0.738, 1.103)	1.000
	β_2	0.4	0.459	0.489	0.482	0.473	(0.313, 0.628)	1.000
	β_3	0.4	0.226	0.282	-	0.191	(0.000, 0.382)	0.764
	β_4	0	-	-	-	0.001	(-0.007, 0.028)	0.121
	Σ_{00}	0.4 (0.19)	0.313	-	0.327	0.203	(0.000, 0.605)	0.635
	Σ_{11}	0.2 (0.21)	-	-	-	0.001	(0.000, 0.000)	0.013
	Σ_{22}	0	-	-	-	0.000	(0.000, 0.000)	0.006
	Pr(Selection)		0.385	0.276	0.157			
$n = 30$	β_1	0.8	0.768	0.764	0.719	0.758	(0.603, 0.871)	1.000
	β_2	0.4	0.409	0.409	0.362	0.401	(0.305, 0.475)	1.000
	β_3	0.4	0.384	0.371	0.367	0.382	(0.295, 0.492)	1.000
	β_4	0	-	-0.075	-0.118	-0.021	(-0.142, 0.000)	0.212
	Σ_{00}	0.4 (0.17)	0.237	0.198	-	0.179	(0.000, 0.387)	0.770
	Σ_{11}	0.2 (0.17)	0.222	0.254	0.228	0.210	(0.000, 0.370)	0.934
	Σ_{22}	0	-	-	-	0.000	(0.000, 0.000)	0.001
	Pr(Selection)		0.678	0.091	0.091			
$n = 50$	β_1	0.8	0.850	0.867	0.796	0.851	(0.715, 0.999)	1.000
	β_2	0.4	0.464	0.471	0.447	0.465	(0.381, 0.541)	1.000
	β_3	0.4	0.500	0.496	0.438	0.494	(0.394, 0.597)	1.000
	β_4	0	-	0.095	-	0.039	(0.000, 0.144)	0.418
	Σ_{00}	0.4 (0.43)	0.843	0.798	0.666	0.808	(0.502, 1.233)	1.000
	Σ_{11}	0.2 (0.16)	0.306	0.306	-	0.279	(0.000, 0.465)	0.911
	Σ_{22}	0	-	-	-	0.000	(0.000, 0.000)	0.001
	Pr(Selection)		0.514	0.396	0.067			

Table 2: Simulation results with $n_i = 20$; B1, B2 and B3 are top three selected models.

	Parameter	True Value	B1	B2	B3	Estimate	95% CI	Pr(Inclusion)
$n = 20$	β_1	0.8	0.726	0.715	0.748	0.726	(0.564, 0.922)	1.000
	β_2	0.4	0.426	0.431	0.460	0.428	(0.351, 0.518)	1.000
	β_3	0.4	0.360	0.381	0.336	0.361	(0.283, 0.473)	1.000
	β_4	0	-	-0.012	-	-0.001	(0.000, 0.000)	0.078
	Σ_{00}	0.4 (0.41)	0.527	0.551	0.539	0.530	(0.259, 0.982)	1.000
	Σ_{11}	0.2 (0.22)	0.212	0.226	-	0.206	(0.071, 0.423)	0.968
	Σ_{22}	0	-	-	-	0.000	(0.000, 0.000)	0.001
	Pr(Selection)		0.891	0.077	0.030			
$n = 30$	β_1	0.8	0.775	0.771	0.776	0.775	(0.627, 0.940)	1.000
	β_2	0.4	0.403	0.397	0.405	0.402	(0.315, 0.483)	1.000
	β_3	0.4	0.469	0.476	0.501	0.470	(0.388, 0.559)	1.000
	β_4	0	-	-0.011	-	-0.001	(0.000, 0.000)	0.081
	Σ_{00}	0.4 (0.52)	0.718	0.721	0.707	0.718	(0.396, 1.174)	1.000
	Σ_{11}	0.2 (0.17)	0.242	0.256	0.274	0.244	(0.121, 0.425)	1.000
	Σ_{22}	0	-	-	0.002	0.000	(0.000, 0.000)	0.002
	Pr(Selection)		0.917	0.081	0.002			
$n = 50$	β_1	0.8	0.819	0.820	-	0.819	(0.685, 0.974)	1.000
	β_2	0.4	0.398	0.400	-	0.398	(0.330, 0.472)	1.000
	β_3	0.4	0.422	0.428	-	0.423	(0.354, 0.493)	1.000
	β_4	0	-	-0.004	-	-0.000	(0.000, 0.000)	0.060
	Σ_{00}	0.4 (0.29)	0.295	0.300	-	0.295	(0.176, 0.458)	1.000
	Σ_{11}	0.2 (0.28)	0.291	0.301	-	0.292	(0.189, 0.434)	1.000
	Σ_{22}	0	-	-		-	(0.000, 0.000)	0.000
	Pr(Selection)		0.940	0.060	0.000			

Table 3: Selection of random effects for the lung cancer data

Effect	Variable	B1	B2	B3	Estimate	95% CI	Pr(Inclusion)
Fixed	Treatment	-0.098	-0.144	-0.060	-0.098	(-0.180, -0.013)	1.000
	Bone	0.243	0.335	0.180	0.243	(0.090, 0.394)	1.000
	Liver	0.352	0.299	0.317	0.352	(0.244, 0.530)	1.000
	P.S.	-0.343	-0.379	-0.367	-0.344	(-0.461, -0.158)	1.000
	W.L.	0.349	0.365	0.334	0.350	(0.192, 0.501)	1.000
Random	Baseline				0.000	(0.000, 0.000)	0.000
	Treatment				0.000	(0.000, 0.000)	0.000
	Bone	0.114	0.131	0.141	0.115	(0.029, 0.263)	1.000
	Liver			0.001	0.000	(0.000, 0.000)	0.002
	P.S.				0.000	(0.000, 0.000)	0.001
	W.L.		0.001		0.000	(0.000, 0.000)	0.003
	Selection	0.994	0.003	0.002			

Table 4: Simulated lung cancer data with weak random effects

Parameter	True Value	B1	B2	B3	Estimate	95% CI	Pr(Inclusion)
β_1	-0.5	-0.578	-0.549	-0.659	-0.567	(-0.776, -0.394)	1.000
β_2	0.3	0.297		0.373	0.155	(0.000, 0.518)	0.517
β_3	0.5	0.480	0.470	0.583	0.479	(0.332, 0.674)	1.000
β_4	-0.8	-0.733	-0.649	-0.761	-0.695	(-0.918, -0.511)	1.000
β_5	0.3	0.512	0.490	0.491	0.503	(0.334, 0.725)	1.000
Σ_{00}	0				0.000	(0.000, 0.000)	0.001
Σ_{11}	0.1			0.089	0.002	(0.000, 0.000)	0.023
Σ_{22}	0.2	0.306	0.432	0.277	0.365	(0.092, 0.762)	0.996
Σ_{33}	0				0.000	(0.000, 0.000)	0.008
Σ_{44}	0				0.000	(0.000, 0.000)	0.002
Σ_{55}	0				0.000	(0.000, 0.000)	0.004
Selection		0.495	0.466	0.010			

Table 5: Simulated lung cancer data with strong random effects

Parameter	True Value	B1	B2	B3	Estimate	95% CI	Pr(Inclusion)
β_1	1	0.911	1.013	0.939	0.921	(0.543, 1.352)	1.000
β_2	-2	-1.708	-1.433	-1.641	-1.692	(-2.219, -0.947)	1.000
β_3	0		-0.117		-0.010	(-0.084, 0.000)	0.084
β_4	-1	-1.110	-1.105	-1.121	-1.110	(-1.269, -0.924)	1.000
β_5	0			0.051	0.004	(0.000, 0.009)	0.069
Σ_{00}	0				0.000	(0.000, 0.000)	0.001
Σ_{11}	16/15	0.961	1.142	0.862	0.970	(0.320, 2.009)	1.000
Σ_{22}	32/15	3.414	3.812	3.438	3.430	(1.814, 5.916)	1.000
Σ_{33}	0				0.002	(0.000, 0.000)	0.023
Σ_{44}	0				0.001	(0.000, 0.000)	0.023
Σ_{55}	0				0.002	(0.000, 0.000)	0.030
Selection		0.793	0.073	0.052			