Biometrics 71, 178–187
March 2015

DOI: 10.1111/biom.12221

Simultaneous Variable Selection for Joint Models of Longitudinal and Survival Outcomes

Zangdong He,[1,*] Wanzhu Tu,[1,2] Sijian Wang,[3] Haoda Fu,[4] and Zhangsheng Yu[1]

[1]Department of Biostatistics, Indiana University School of Medicine and Fairbanks School of Public Health, Indianapolis, Indiana, U.S.A.

[2]Regenstrief Institute, Inc., Indianapolis, Indiana, U.S.A.

[3]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, Wisconsin, U.S.A.

[4]Eli Lilly & Company, Indianapolis, Indiana, U.S.A.

email: zanghe@iu.edu

Summary. Joint models of longitudinal and survival outcomes have been used with increasing frequency in clinical investigations. Correct specification of fixed and random effects is essential for practical data analysis. Simultaneous selection of variables in both longitudinal and survival components functions as a necessary safeguard against model misspecification. However, variable selection in such models has not been studied. No existing computational tools, to the best of our knowledge, have been made available to practitioners. In this article, we describe a penalized likelihood method with adaptive least absolute shrinkage and selection operator (ALASSO) penalty functions for simultaneous selection of fixed and random effects in joint models. To perform selection in variance components of random effects, we reparameterize the variance components using a Cholesky decomposition; in doing so, a penalty function of group shrinkage is introduced. To reduce the estimation bias resulted from penalization, we propose a two-stage selection procedure in which the magnitude of the bias is ameliorated in the second stage. The penalized likelihood is approximated by Gaussian quadrature and optimized by an EM algorithm. Simulation study showed excellent selection results in the first stage and small estimation biases in the second stage. To illustrate, we analyzed a longitudinally observed clinical marker and patient survival in a cohort of patients with heart failure.

Key words: ALASSO; Cholesky decomposition; EM algorithm; Gaussian quadrature; Joint models; Mixed effect selection.

1. Introduction

Longitudinal and survival data often arise together in clinical investigations. In a given subject, longitudinally measured clinical markers and patient survival are usually governed by the same latent disease process, and thus are correlated. Separate modeling for the longitudinal and survival outcomes could result in biases in parameter estimation (Faucett and Thomas, 1996). Joint models are therefore recommended to alleviate biases and to ensure valid inference concerning the correlation structure between the two outcomes. In the past two decades, joint models have been studied extensively: Wulfsohn and Tsiatis (1997) proposed a general framework in which the survival component was depicted by a proportional hazard model, and the longitudinal component was accommodated by a linear-growth-curve model. This basic structure was later extended by Xu and Zeger (2001) to a variety of data situations. Other noteworthy method developments and significant data applications were presented by De Gruttola and Tu (1994), Nathoo and Dean (2008), and Albert and Shih (2010). Notably missing in this literature is variable selection. As in any modeling exercise, correct specification of the model and inclusion of the right independent variables are of essential importance, for the preservation of scientific validity. For joint models in particular, random variable selection serves the purpose of justifying the use of shared random effects connecting the longitudinal and survival components.

Traditionally, variable selection has been performed through model comparisons using information-based criteria, such as the Akaike and Bayesian information criteria (AIC and BIC). But such criteria are not always feasible in complex model settings where the number of candidate models is large. As an alternative, penalized likelihood approach has gained popularity since the mid-1990's. Tibshirani (1996) proposed a least absolute shrinkage and selection operator (LASSO) for fixed-effect selection. Asymptotic "oracle" properties of the smoothly clipped absolute deviation (SCAD; Fan and Li, 2001) and the adaptive least absolute shrinkage and selection operator (ALASSO; Zou, 2006) have provided a theoretical assurance for mixed effect selection. Along this line, Fan and Li (2004), Johnson, Lin, and Zeng (2008) and Garcia, Ibrahim, and Zhu (2010) discussed the application of penalized likelihood method to select fixed effect variables in longitudinal model settings. Fan and Li (2002), Zhang and Lu (2007), and Garcia et al. (2010) discussed the selection of fixed effects in survival models. Extending these previous work, Bondell, Krishna, and Ghosh (2010) proposed a method for selecting fixed and random effects in a linear mixed-effects model setting. Most recently, Ibrahim et al. (2011) studied the mixed-effects selection in generalized linear mixed models through an EM algorithm. To the best of our knowledge, no work has been done for simultaneous selection of fixed and random effects in a joint model setting with longitudinal and survival

outcomes. To fill in this methodological gap, we propose a penalized likelihood method with ALASSO penalty for fixed and random effect selection in joint models. We optimize the penalized likelihood using an EM algorithm.

We illustrate the method by analyzing data from an observational study of heart failure patients. The study cohort included 1702 patients with diagnosed congestive heart failure (CHF) between January 1, 2004 and December 31, 2009, identified from a large electronic medical record system. The analytical objective is to assess the effects of medication adherence on disease exacerbation and on patient survival; we also like to assess the correlation between CHF exacerbation and patient mortality. Specifically, we considered two outcomes: the survival outcome is defined as the time from the first recorded CHF diagnosis to mortality, or to December 31, 2009, which ever comes first; the longitudinal outcomes are the repeatedly measured B-type natriuretic peptide (BNP) levels. BNP is a commonly used bedside marker of CHF exacerbation; a higher BNP value indicates fluid volume overload in the left ventricle and increased mortality risk. (Morrison et al., 2002). Although the two outcomes can be modeled individually, separate modeling does not accommodate correlations between BNP and survival. In this article, we consider a joint modeling approach. We consider eight known risk factors and four interaction terms as candidate variables and develop an ALASSO procedure to select the independent variables. In particular, we consider random-effect selection as medical literature rarely avails information on the possible random slopes (e.g., the effect of an independent variable varies across subjects). Misspecification of fixed and random effects for the two outcome variables could result in erroneous inferences.

The remainder of the article is organized as follows. In Section 2, we present the model and the proposed selection method. In Section 3, we report the operating characteristics of the proposed method as observed in a simulation study. In Section 4, we revisit the CHF data analysis. We end the article in Section 5 with a few concluding remarks.

2. Method

2.1. *Model Formulation*

Suppose in a longitudinal study, we observe a survival outcome (t_i, δ_i), and repeated measurements of a continuous outcome $\boldsymbol{y}_i$, for subject $i = 1, \ldots, n$. Here, t_i is the observed event time subject to right censoring, and δ_i is a failure indicator with $\delta_i = 1$ indicating the occurrence of an event of interest, and $\delta_i = 0$ indicating censoring, whereas $\boldsymbol{y}_i$ is an $n_i \times 1$ vector of the n_i repeated measurements. Let $\boldsymbol{X}_{1i} \in \mathbb{R}^{n_i \times p}$ and $\boldsymbol{Z}_{1i} \in \mathbb{R}^{n_i \times q}$ be the fixed and random covariate matrices for the longitudinal outcome, respectively. Similarly, we let $\boldsymbol{x}_{2i} \in \mathbb{R}^{1 \times p}$ and $\boldsymbol{z}_{2i} \in \mathbb{R}^{1 \times q}$ be the fixed and random covariate vectors for the survival outcome. Combining these observations we write $\boldsymbol{O_i} = (\boldsymbol{y}_i, \boldsymbol{X}_{1i}, \boldsymbol{Z}_{1i}, t_i, \delta_i, \boldsymbol{x}_{2i}, \boldsymbol{z}_{2i})$. We assume that the observations $\boldsymbol{O_i}$ are independent across subjects.

Without loss of generality, we herein consider a case where the longitudinal and survival components share the same set of fixed- and random-effect covariates. This model formulation could easily be generalized to situations where the two components have different sets of covariates.

For the longitudinal outcome, we consider the following linear mixed-effects model:

$$\boldsymbol{y}_i = \boldsymbol{X}_{1i}\boldsymbol{\beta}_1 + \boldsymbol{Z}_{1i}\boldsymbol{\Gamma}_1\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \tag{1}$$

where $\boldsymbol{\beta}_1 = (\beta_{10}, \beta_{11}, \ldots, \beta_{1p})^T$ is the coefficient vector, and β_{10} is the intercept. $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})^T \sim N_{n_i}(0, \sigma^2 \boldsymbol{I}_{n_i})$ is the measurement error vector, and $\boldsymbol{b_i} \in \mathbb{R}_1^q$ is a q-dimensional random effect vector following a multivariate normal distribution $N_q(0, \boldsymbol{I}_q)$, with $\boldsymbol{I_q}$ as a $q \times q$ identity matrix. $\boldsymbol{\Gamma}_1$ is a $q \times q$ lower triangular matrix and $\boldsymbol{\Gamma}_1\boldsymbol{b}_i$ follows $N_q(0, \boldsymbol{D}_1)$. Thus $\boldsymbol{\Gamma}_1$ represents a Cholesky decomposition of the covariance matrix $\boldsymbol{D}_1$.

For the survival outcome, we consider a frailty model, defined as follows:

$$h(t_i) = h_0(t_i)\exp(\boldsymbol{x}_{2i}\boldsymbol{\beta}_2 + \boldsymbol{z}_{2i}\boldsymbol{\Gamma}_2\boldsymbol{b}_i), \tag{2}$$

where $h_0(t_i)$ is the baseline hazard function, and $\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{2p})^T$ is the coefficient vector. $\boldsymbol{\Gamma}_2\boldsymbol{b}_i$ follows $N_q(0, \boldsymbol{D}_2)$, and $\boldsymbol{\Gamma}_2$ is a Cholesky decomposition of the $q \times q$ matrix $\boldsymbol{D}_2$.

2.2. *Variable Selection Using Penalized Likelihood*

To select fixed and random effects, we propose a penalized likelihood to simultaneously identify the non-zero elements in $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\Gamma}_1\boldsymbol{b}_i, \boldsymbol{\Gamma}_2\boldsymbol{b}_i)$. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \boldsymbol{\phi})$ be the collection of all the unknown parameters, where $\boldsymbol{\phi}$ denotes parameters other than $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2)$. Writing the density function of $(\boldsymbol{y_i}, t_i, \boldsymbol{b_i})$ as $f(\boldsymbol{y_i}, t_i, \boldsymbol{b_i}|\boldsymbol{X}_{1i}, \boldsymbol{Z}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{z}_{2i}, h_0(t_i), \delta_i, \boldsymbol{\theta})$, we have the following log-marginal likelihood for $\boldsymbol{\theta}$:

$$l_o(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int f_y(\boldsymbol{y}_i|\boldsymbol{X}_{1i}, \boldsymbol{Z}_{1i}, \boldsymbol{b}_i, \boldsymbol{\theta}) \times f_s(t_i, \delta_i|\boldsymbol{x}_{2i}, \boldsymbol{z}_{2i}, h_0(t_i), \boldsymbol{b}_i, \boldsymbol{\theta}) f_b(\boldsymbol{b}_i)\, \mathrm{d}\boldsymbol{b}_i, \tag{3}$$

where $f_b(\boldsymbol{b_i})$ is a q-variate normal density function for $\boldsymbol{b_i}$. Functions $f_s(\cdot)$ and $f_y(\cdot)$ are the conditional density functions of the survival time and repeated measurements when $\boldsymbol{b_i}$ is given, respectively. We note that in the absence of restrictions on the baseline hazard $h_0(t_i)$, the maximum of the marginal likelihood is infinity. To remedy the deficiency, one could parameterize $h_0(t_i)$ with a parametric distribution. For example, a natural choice is to use a Weibull distribution with a baseline hazard $h_0(t_i) = \alpha\lambda t_i^{\alpha-1}$, where α is the shape parameter and λ is the scale parameter. Alternatively, one could use a piece-wise constant baseline hazard by dividing the study period into m intervals and assuming $h_0(t)$ to be a constant within each interval as $h_0(t) = h_k, t_{k-1} < t \leq t_k, k = 1, \ldots, m$, where t_ks are knots defining the intervals. This piece-wise constant baseline hazard have been shown to perform well by Feng, Wolfe, and Port (2005).

To select fixed and random effects simultaneously, we consider a penalized likelihood $PL(\boldsymbol{\theta}) = \frac{1}{n} l_o(\boldsymbol{\theta}) - \kappa_{\lambda_1}(\boldsymbol{\beta_1}) - \kappa_{\lambda_2}(\boldsymbol{\beta_2}) - \kappa_{\lambda_3}(\boldsymbol{D_1}) - \kappa_{\lambda_4}(\boldsymbol{D_2})$. The penalty terms $\kappa_{\lambda_1}(\boldsymbol{\beta_1})$ and $\kappa_{\lambda_2}(\boldsymbol{\beta_2})$ control for the sparsity of estimates of $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ so that the fixed effects are selected. The penalty terms $\kappa_{\lambda_3}(\boldsymbol{D_1})$ and $\kappa_{\lambda_4}(\boldsymbol{D_2})$ control for the sparsity of estimates of $\boldsymbol{D_1}$ and $\boldsymbol{D_2}$ to select the random effects. The penalty functions $\kappa_{\lambda_j}(\cdot)$, for $j = 1, 2, 3, 4$, could be the adaptive LASSO, or the

smoothly clipped absolute deviation (SCAD). For the fixed-effect selection, we define the adaptive LASSO penalties as $\kappa_{\lambda_1}(\boldsymbol{\beta_1}) = \lambda_1 \sum_{j=1}^{p} \omega_{\beta_{1j}} |\beta_{1j}|$ and $\kappa_{\lambda_2}(\boldsymbol{\beta_2}) = \lambda_2 \sum_{k=1}^{p} \omega_{\beta_{2k}} |\beta_{2k}|$, where λ_1 and λ_2 are tuning parameters that control the degree of penalties; $\omega_{\beta_{1j}}, \omega_{\beta_{2k}}$ are the corresponding positive weights for penalties $|\beta_{1j}|$ and $|\beta_{2k}|$. The summation in $\kappa_{\lambda_1}(\boldsymbol{\beta_1}) = \lambda_1 \sum_{j=1}^{p} \omega_{\beta_{1j}} |\beta_{1j}|$ starts from 1 as we are not interested in selecting intercept β_{10}. Some of the estimates of $\hat{\beta}_{1j}$ and $\hat{\beta}_{2k}$ will be zero since $|\beta_{1j}|$ and $|\beta_{2k}|$ are singular when $|\beta_{1j}| = 0$ and $|\beta_{2k}| = 0$.

For the random-effect selection, we note that $\boldsymbol{D_1 = \Gamma_1 \Gamma_1^T}$ and $\boldsymbol{D_2 = \Gamma_2 \Gamma_2^T}$. Let $\boldsymbol{\gamma}_{1m}$ and $\boldsymbol{\gamma}_{2l}$ be the mth and lth row vectors of $\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2$, respectively. In fact, $\boldsymbol{\gamma}_{1m}\boldsymbol{\gamma}_{1m}^T = \boldsymbol{D}_{1mm}$ and $\boldsymbol{\gamma}_{2l}\boldsymbol{\gamma}_{2l}^T = \boldsymbol{D}_{2ll}$ are the variance components of the mth and lth elements of the random effects $\boldsymbol{\Gamma_1 b_i}$ and $\boldsymbol{\Gamma_2 b_i}$. We form the penalty terms for the random effects in a group manner so that the estimates of elements of the entire vectors $\boldsymbol{\gamma}_{1m}$ and $\boldsymbol{\gamma}_{2l}$ are either all zero or at least one of the estimates is nonzero. The group penalties on $\boldsymbol{\gamma_{1m}}$ and $\boldsymbol{\gamma_{2l}}$ will ensure selection for the covariance structure due to the following connection of covariance matrices $\boldsymbol{D_1, D_2}$ and the Cholesky decomposition matrices $\boldsymbol{\Gamma_1, \Gamma_2}$ (Wang, Song, and Zhu, 2010):

$$\begin{aligned} \boldsymbol{\gamma_{1m}} = 0 \Leftrightarrow D_{1mm} = 0, D_{1mh} = D_{1hm} = 0, \quad \forall h \\ \boldsymbol{\gamma_{2l}} = 0 \Leftrightarrow D_{2ll} = 0, D_{2lh} = D_{2hl} = 0, \quad \forall h. \end{aligned} \tag{4}$$

From (4), it follows that if $\boldsymbol{\gamma_{1m}} = 0$, then the diagonal element D_{1mm}, the variance of the random effect $(\boldsymbol{\Gamma_1 b_i})_m$, is zero. Furthermore, for any $h \neq m$, the off-diagonal element $D_{1mh} = D_{1hm} = 0$ implies that the covariance between $(\boldsymbol{\Gamma_1 b_i})_m$ and all other random effects are zero. Thus, the random effect $(\boldsymbol{\Gamma_1 b_i})_m$ in longitudinal component is to be excluded from the model and the positive-definiteness of $\boldsymbol{D_1}$ will be preserved. This applies to the random-effect selection in the survival component as well, which is to shrink the whole vector $\boldsymbol{\gamma_{2l}}$ to zero.

To perform group penalties on vectors $\boldsymbol{\gamma_{1m}}$ and $\boldsymbol{\gamma_{2m}}$, we first summarize the penalties using L_2-norm: $||\boldsymbol{\gamma}_{1m}|| = (\boldsymbol{\gamma}_{1m}\boldsymbol{\gamma}_{1m}^T)^{1/2}$ and $||\boldsymbol{\gamma}_{2l}|| = (\boldsymbol{\gamma}_{2l}\boldsymbol{\gamma}_{2l}^T)^{1/2}$ for $m, l = 2, \ldots, q$. Following Yuan and Lin (2006), the adaptive LASSO penalties are defined as: $\kappa_{\lambda_3}(\boldsymbol{D_1}) = \lambda_3 \sum_{m=2}^{q} \omega_{\boldsymbol{\gamma}_{1m}} ||\boldsymbol{\gamma}_{1m}||$ and $\kappa_{\lambda_4}(\boldsymbol{D_2}) = \lambda_4 \sum_{l=2}^{q} \omega_{\boldsymbol{\gamma}_{2l}} ||\boldsymbol{\gamma}_{2l}||$. We use adaptive LASSO penalties in the simulation study. Note that the summation starts from $m = 2, l = 2$, as we keep the random intercepts in both the longitudinal and survival components without eliminating the possible minimal within-cluster correlation. λ_3 and λ_4 are the positive tuning parameters, and $\omega_{\gamma_{1m}}, \omega_{\gamma_{2l}}$, are the positive weights associated with penalties on $||\boldsymbol{\gamma}_{1m}||$ and $||\boldsymbol{\gamma}_{2l}||$. Let $p(\boldsymbol{\theta}) = \lambda_1 \sum_{j=1}^{p} \omega_{\beta_{1j}} |\beta_{1j}| + \lambda_2 \sum_{k=1}^{p} \omega_{\beta_{2k}} |\beta_{2k}| + \lambda_3 \sum_{m=2}^{q} \omega_{\boldsymbol{\gamma_{1m}}} ||\boldsymbol{\gamma_{1m}}|| + \lambda_4 \sum_{m=2}^{q} \omega_{\boldsymbol{\gamma_{2l}}} ||\boldsymbol{\gamma_{2l}}||$, and the penalized likelihood with the adaptive LASSO penalties can be written as

$$pl(\boldsymbol{\theta}) = \frac{1}{n} l_o(\boldsymbol{\theta}) - p(\boldsymbol{\theta}). \tag{5}$$

Penalized likelihood with SCAD penalties could be constructed by substituting the penalty terms in (5) using SCAD. The estimator of $\boldsymbol{\theta}$ can be obtained by maximizing (5).

2.3. *EM Algorithm for Optimization of the Penalized Likelihood*

To maximize the penalized likelihood (5), we use an EM algorithm. We start with the penalized log-complete likelihood for $(\boldsymbol{O_i}, \boldsymbol{b_i})$ for $i = 1, \ldots, n$, which is

$$\begin{aligned} pl_c(\boldsymbol{\theta}) &= \frac{1}{n} \sum_{i=1}^{n} \log f(\boldsymbol{y}_i, t_i, \delta_i, \boldsymbol{b}_i | \boldsymbol{\theta}) - p(\boldsymbol{\theta}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \{\log[f_y(\boldsymbol{y}_i | \boldsymbol{b}_i, \boldsymbol{\theta})] + \delta_i \log[h(t_i | \boldsymbol{b}_i, \boldsymbol{\theta})] \\ &\quad + \log[S(t_i | \boldsymbol{b}_i, \boldsymbol{\theta})] + \log[f_b(\boldsymbol{b}_i | \boldsymbol{\theta})]\} - p(\boldsymbol{\theta}). \end{aligned} \tag{6}$$

In Equation (6), $S(\cdot)$ is the survival function of t_i conditional on $\boldsymbol{b_i}$. Let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)^T$ and $\boldsymbol{\omega} = (\omega_{\beta_{1j}}, \omega_{\beta_{2k}}, \omega_{\gamma_{1m}}, \omega_{\gamma_{2l}})^T$. We denote $\boldsymbol{g}_{c,i1} = (\boldsymbol{y}_i, \boldsymbol{X}_{1i}, \boldsymbol{Z}_{1i}, \boldsymbol{b}_i)$, $\boldsymbol{g}_{c,i2} = (t_i, \delta_i, \boldsymbol{x}_{2i}, \boldsymbol{z}_{2i}, \boldsymbol{b}_i)$, and $\boldsymbol{g}_{c,i} = (\boldsymbol{y}_i, \boldsymbol{X}_{1i}, \boldsymbol{Z}_{1i}, t_i, \delta_i, \boldsymbol{x}_{2i}, \boldsymbol{z}_{2i}, \boldsymbol{b}_i)$ as the complete data for longitudinal, survival and both components, respectively, and $\boldsymbol{g}_{o,i1} = (\boldsymbol{y}_i, \boldsymbol{X}_{1i}, \boldsymbol{Z}_{1i})$, $\boldsymbol{g}_{o,i2} = (t_i, \delta_i, \boldsymbol{x}_{2i}, \boldsymbol{z}_{2i})$, and $\boldsymbol{g}_{o,i} = (\boldsymbol{y}_i, \boldsymbol{X}_{1i}, \boldsymbol{Z}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{z}_{2i})$ as the corresponding observed data.

2.3.1. E-step. We first derive the E-step of the EM algorithm for the given $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$. Assuming that we have estimates $\boldsymbol{\theta}^{(s)}$ from the (s)th iteration of the maximization step, we take the expectation of the penalized log-complete likelihood conditional on $\boldsymbol{\theta}^{(s)}$ and $\boldsymbol{g}_{oi}$, for $i = 1, \ldots, n$ and obtain the following penalized Q-function:

$$\begin{aligned} Q_{\lambda,\omega}(\boldsymbol{\theta} | \boldsymbol{\theta}^{(s)}) &= \frac{1}{n} \sum_{i=1}^{n} \{E[\log f_y(\boldsymbol{g}_{c,i1}, \boldsymbol{\theta}) | (\boldsymbol{g}_{o,i}, \boldsymbol{\theta}^{(s)})] \\ &\quad + E[\delta_i \log h(\boldsymbol{g}_{c,i2}, \boldsymbol{\theta}) | (\boldsymbol{g}_{o,i}, \boldsymbol{\theta}^{(s)})] \\ &\quad + E[\log S(\boldsymbol{g}_{c,i2}, \boldsymbol{\theta}) | (\boldsymbol{g}_{o,i}, \boldsymbol{\theta}^{(s)})]\} \\ &\quad - p(\boldsymbol{\theta}) + \frac{1}{n} \sum_{i=1}^{n} E[\log f_b(\boldsymbol{b}_i) | (\boldsymbol{g}_{o,i}, \boldsymbol{\theta}^{(s)})]. \end{aligned} \tag{7}$$

We write

$$E[H(\boldsymbol{b_i}) | (\boldsymbol{g}_{o,i}, \boldsymbol{\theta}^{(s)})] = \int H(\boldsymbol{b_i}) f_b(\boldsymbol{b_i} | \boldsymbol{g}_{o,i}, \boldsymbol{\theta}^{(s)}) \, \mathrm{d}\boldsymbol{b_i} \tag{8}$$

for each of $H(\boldsymbol{b_i}) = \log f_y(\boldsymbol{g}_{c,i1}, \boldsymbol{\theta})$, $H(\boldsymbol{b_i}) = \delta_i \log h(\boldsymbol{g}_{c,i2}, \boldsymbol{\theta})$, and $H(\boldsymbol{b_i}) = \log S(\boldsymbol{g}_{c,i2}, \boldsymbol{\theta})$. Because integral (8) is intractable, we approximate it by using a multivariate Gaussian quadrature method (Pinheiro and Bates, 1995). Since $\boldsymbol{b_i} \sim N(0, \boldsymbol{I}_q)$, if we choose k quadrature points in each dimension, there will be k^q vector nodes of $q \times 1$ dimension. Let $\boldsymbol{b}_l^{'} = (b_{l,1}^{'}, b_{l,2}^{'}, \ldots, b_{l,q}^{'})$ denote the lth node, and w_l the corresponding quadrature weight, for $l = 1, \ldots, k^q$, integral in (8) can be approximated by

$$\tilde{E}\{H(\boldsymbol{b_i}) | (\boldsymbol{g}_{o,i}, \boldsymbol{\theta}^{(s)})\} \approx \sum_{l=1}^{k^q} w_l H(\boldsymbol{b}_l^{\prime}) f_b(\boldsymbol{b}_l^{\prime} | \boldsymbol{g}_{o,i}, \boldsymbol{\theta}^{(s)}). \tag{9}$$

We therefore obtain the approximated penalized Q-function in the $(s+1)$th iteration

$$\tilde{Q}_{\lambda,\omega}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \frac{1}{n}\sum_{i=1}^{n}\{\tilde{E}[\log f_y(\boldsymbol{g}_{c,i1},\boldsymbol{\theta})|(\boldsymbol{g}_{o,i},\boldsymbol{\theta}^{(s)})] + \tilde{E}[\delta_i \log h(\boldsymbol{g}_{c,i2},\boldsymbol{\theta})|(\boldsymbol{g}_{o,i},\boldsymbol{\theta}^{(s)})] + \tilde{E}[\log S(\boldsymbol{g}_{c,i2},\boldsymbol{\theta})|(\boldsymbol{g}_{o,i},\boldsymbol{\theta}^{(s)})]\} - p(\boldsymbol{\theta}). \quad (10)$$

The last term $\frac{1}{n}\sum_{i=1}^{n} E[\log f_b(\boldsymbol{b}_i)|(\boldsymbol{g}_{o,i},\boldsymbol{\theta}^{(s)})$ in (7) does not involve any unknown parameters, thus could be omitted from the optimization.

2.3.2. M-step. We maximize (10) with respect to the fixed- and random-effect parameters alternatively. When $(\boldsymbol{\Gamma_1}, \boldsymbol{\Gamma_2}, \boldsymbol{\phi})$ are fixed, we maximize (10) with respect to $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, and the penalty function involving L_1 penalty terms can be solved by applying the LARS/LASSO algorithm (Efron et al., 2004) and the SCAD penalties could be solved according to Fan and Li (2001). When $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\phi})$ are fixed, we maximize (10) with respect to $\boldsymbol{(\Gamma_1, \Gamma_2)}$. Following Lin and Zhang (2006) and Wang et al. (2010), we transform the optimization problem to a two-step equivalent objective function involving quadratic penalty term that is easier to solve. Specifically, let

$$\tilde{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \frac{1}{n}\sum_{i=1}^{n}\{\tilde{E}[\log f_y(\boldsymbol{g}_{c,i1},\boldsymbol{\theta})|(\boldsymbol{g}_{o,i},\boldsymbol{\theta}^{(s)})] + \tilde{E}[\delta_i \log h(\boldsymbol{g}_{c,i2},\boldsymbol{\theta})|(\boldsymbol{g}_{o,i},\boldsymbol{\theta}^{(s)})] + \tilde{E}[\log S(\boldsymbol{g}_{c,i2},\boldsymbol{\theta})|(\boldsymbol{g}_{o,i},\boldsymbol{\theta}^{(s)})]\},$$

then for any given $\hat{\boldsymbol{\beta}}$ and $(\boldsymbol{\lambda}, \boldsymbol{\omega})$, the following two optimization problems with respect to $\boldsymbol{\gamma}$s achieve the same solution:

$$\tilde{Q}(\hat{\boldsymbol{\beta}}, \Gamma_1, \Gamma_2|\boldsymbol{\theta}^{(s)}) - \lambda_3\sum_{m=2}^{q}\omega_{\gamma_{1m}}||\boldsymbol{\gamma_{1m}}|| - \lambda_4\sum_{l=2}^{q}\omega_{\gamma_{2l}}||\boldsymbol{\gamma_{2l}}|| \quad (11)$$

$$\tilde{Q}(\hat{\boldsymbol{\beta}}, \Gamma_1, \Gamma_2|\boldsymbol{\theta}^{(s)}) - \sum_{m=2}^{q}\zeta_{1m}^2 - \frac{1}{4}\sum_{m=2}^{q}\frac{(\lambda_3\omega_{\gamma_{1m}})^2}{\zeta_{1m}^2}||\boldsymbol{\gamma_{1m}}||^2 - \sum_{l=2}^{q}\eta_{2l}^2 - \frac{1}{4}\sum_{l=2}^{q}\frac{(\lambda_4\omega_{\gamma_{2l}})^2}{\eta_{2l}^2}||\boldsymbol{\gamma_{2l}}||^2. \quad (12)$$

Let $(\hat{\boldsymbol{\gamma}}_{1m}, \hat{\boldsymbol{\gamma}}_{2l})$ be the maximizer of (11), and $(\tilde{\zeta}_{1m}, \tilde{\boldsymbol{\gamma}}_{1m}, \tilde{\eta}_{2l}, \tilde{\boldsymbol{\gamma}}_{2l})$ be the maximizer of (12), then we have

$$\hat{\boldsymbol{\gamma}}_{\boldsymbol{1m}} = \tilde{\boldsymbol{\gamma}}_{\boldsymbol{1m}}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{2l}} = \tilde{\boldsymbol{\gamma}}_{\boldsymbol{2l}} \quad (13)$$

$$\tilde{\zeta}_{1m} = \sqrt{\frac{\lambda_3\omega_{\gamma_{1m}}}{2}||\hat{\boldsymbol{\gamma}}_{\boldsymbol{1m}}||}, \tilde{\eta}_{2l} = \sqrt{\frac{\lambda_4\omega_{\gamma_{2l}}}{2}||\hat{\boldsymbol{\gamma}}_{\boldsymbol{2l}}||}. \quad (14)$$

(13) and (14) imply that one can optimize (12) iteratively with respect to $(\boldsymbol{\gamma_{1m}}, \boldsymbol{\gamma_{2l}})$ and (ζ_{1m}, η_{2l}), instead of directly maximizing (12). Maximizing (12) with respect to $(\boldsymbol{\gamma_{1m}}, \boldsymbol{\gamma_{2l}})$ when (ζ_{1m}, η_{2l}) is given is similar to a generalized ridge regression. When $(\boldsymbol{\gamma_{1m}}, \boldsymbol{\gamma_{2l}})$ is given, (ζ_{1m}, η_{2l}) could be easily computed from (14).

Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}, \zeta_{1m}, \eta_{2l})$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2, \boldsymbol{\phi})$ are defined in Section 2.2. We propose the expectation conditional maximization procedures to optimize the penalized likelihood and the details are described in the Supplementary Materials (Web Appendix 1). The typical values for the weights are selected as: $\omega_{\beta_{1j}} = |\hat{\beta}_{1j}^*|^{-1}, \omega_{\beta_{2k}} = |\hat{\beta}_{2k}^*|^{-1}, \omega_{\gamma_{1m}} = \sqrt{m}||\hat{\boldsymbol{\gamma}}_{\boldsymbol{1m}}^*||^{-1}, \omega_{\gamma_{2l}} = \sqrt{l}||\hat{\boldsymbol{\gamma}}_{\boldsymbol{2l}}^*||^{-1}$, where $\hat{\beta}_{1j}^*, \hat{\beta}_{2k}^*, \hat{\boldsymbol{\gamma}}_{\boldsymbol{1m}}^*, \hat{\boldsymbol{\gamma}}_{\boldsymbol{2l}}^*$ are the unpenalized MLEs (Zou, 2006; Ibrahim et al., 2011) and $\sqrt{m}, \sqrt{l}$ are the normalizing constants for penalty parameters γ_{1m}, γ_{2l} to accommodate the varying sizes of $\boldsymbol{\gamma_{1m}}, \boldsymbol{\gamma_{2l}}$.

2.4. *Tuning Parameter Selection and Two-Stage Estimation*

A data-driven method for determining tuning parameters is essential for variable selection. Criteria such as generalized cross-validation, k-fold cross validation, AIC, BIC, or GIC have been used as the objective scores to minimize over a preselected grid of tuning parameters. BIC is known to be consistent in the model selection (Shao, 1997; Pu and Niu, 2006). Wang, Li, and Leng (2009) showed that selecting tuning parameters via BIC consistently yielded the true model in the linear model setting. Ibrahim et al. (2011) showed that selecting tuning parameters for mixed-effects selection via BIC-type IC_Q criterion also consistently yielded true models in generalized linear mixed models; their simulation study further showed that the approach worked well in finite sample situations. Thus, we propose to use the BIC-type criterion to determine the values of tuning parameters, where

$$\mathrm{BIC}_{\boldsymbol{\lambda}} = -2l_o(\hat{\boldsymbol{\theta}}) + \log(n) \times df_{\boldsymbol{\lambda}}, \quad (15)$$

In (15), $\hat{\boldsymbol{\theta}}$ are the estimators obtained from penalized likelihood under the given $\boldsymbol{\lambda}$, and $l_o(\hat{\boldsymbol{\theta}})$ is the value of the observed likelihood $l_o(\boldsymbol{\theta})$ at the estimates $\hat{\boldsymbol{\theta}}$. The solution is chosen to minimize the $\mathrm{BIC}_{\boldsymbol{\lambda}}$ criterion. In this BIC-type criterion, the total sample size n is used. We take d, the total number of non-zero estimates of $\hat{\theta}$ as the degree of freedom $df_{\boldsymbol{\lambda}}$. In the linear model, d is an unbiased estimator of $df_{\boldsymbol{\lambda}}$. Our simulation shows this criterion works well, as suggested by Pu and Niu (2006).

To reduce the estimation bias, we propose a two-stage process. In the first stage, we focus on variable selection and use the penalized likelihood method to select the model that minimizes the BIC value. In the second stage, we re-estimate parameters using selected variables without penalty for selection, to reduce the estimation bias.

3. Simulation Study

3.1. *Data Generation*

We conduct a simulation study to examine the performance of the proposed method. We generate data under six different scenarios.

Table 1
Selection frequency of mixed effects in longitudinal and survival components for Scenarios 1–4

Fixed effect selection								
	Sel. freq. (%) for longitudinal component				Sel. freq. (%) for survival component			
Scenarios	$X_{1,1}$ Non-zero	$X_{1,2}$ Zero	$X_{1,3}$ Non-zero	$X_{1,4}$ Zero	$X_{2,1}$ Non-zero	$X_{2,2}$ Zero	$X_{2,3}$ Zero	$X_{2,4}$ Non-zero
1	100	0	100	0	100	1	3	100
2	100	0	100	0	100	3	4	100
3	100	0	100	0	100	0	1	100
4	100	0	100	0	100	1	1	100
Random effect selection								
	Sel. freq. (%) for longitudinal component				Sel. freq. (%) for survival component			
Scenarios	$Z_{1,1}$ Non-zero	$Z_{1,2}$ Zero	$Z_{1,3}$ Non-zero	$Z_{1,4}$ Zero	$Z_{2,1}$ Non-zero	$Z_{2,2}$ Zero	$Z_{2,3}$ Zero	$Z_{2,4}$ Non-zero
1	100	8	100	11	99	7	19	84
2	100	2	100	6	100	10	13	92
3	100	2	100	10	100	5	12	97
4	100	1	100	10	100	6	8	99

For Scenarios 1–4, we generate the longitudinal outcome Y_{ij} from the following model:

$$Y_{ij} = 1 + 1X_{1ij,1} + 0X_{1ij,2} + 3X_{1ij,3} + 0X_{1ij,4} + b_{li,0} + b_{li,1}Z_{1ij,1} + b_{li,2}Z_{1ij,2} + b_{li,3}Z_{1ij,3} + b_{li,4}Z_{1ij,4} + \epsilon_{ij} \tag{16}$$

and the failure time from a Weibull distribution with the hazard function:

$$\lambda_i(t) = \lambda_0(t)\exp(1x_{2i,1} + 0x_{2i,2} + 0x_{2i,3} + 1x_{2i,4} + b_{si,0} + b_{si,1}z_{2i,1} + b_{si,2}z_{2i,2} + b_{si,3}z_{2i,3} + b_{si,4}z_{2i,4}), \tag{17}$$

for $i = 1, \ldots, 250$, $j = 1, \ldots, 5$, where $\lambda_0(t) = \alpha\lambda t^{\alpha-1}$ with $\alpha = 2$, and $\lambda = \exp(1) = 2.718$.

Random effect vector $\boldsymbol{b_i}$ is independently generated from $N(0, \boldsymbol{I_5})$. $\boldsymbol{b_{li}} = (b_{li,0}, b_{li,1}, b_{li,2}, b_{li,3}, b_{li,4})$ is obtained from $\boldsymbol{b_{li}} = \boldsymbol{\Gamma_1 b_i}$ and $\boldsymbol{b_{si}} = (b_{si,0}, b_{si,1}, b_{si,2}, b_{si,3}, b_{si,4})$ is obtained from $\boldsymbol{b_{si}} = \boldsymbol{\Gamma_2 b_i}$, where $\boldsymbol{\Gamma_1} = \sigma_D \boldsymbol{R_1}$ and $\boldsymbol{\Gamma_2} = \sigma_D \boldsymbol{R_2}$, with lower triangular matrix $\boldsymbol{R_1} = \{1; \frac{1}{2}, \frac{1}{2}; 0, 0, 0; \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}; 0, 0, 0, 0, 0\}$ and $\boldsymbol{R_2} = \{1; \frac{1}{2}, \frac{1}{2}; 0, 0, 0; 0, 0, 0, 0; \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\}$. Covariates $X_{1ij,1} = Z_{1ij,1}$, $X_{1ij,2} = Z_{1ij,2}$, $X_{1ij,4} = Z_{1ij,4}$, and $x_{2i,1} = z_{2i,1}$, $x_{2i,2} = z_{2i,2}$, $x_{2i,4} = z_{2i,4}$ are generated as independent $N(0, 1)$ variables; $X_{1ij,3} = Z_{1ij,3}$ and $x_{2i,3} = z_{2i,3}$ are binary variables with equal probability taking value 0 or 1. The measurement error $\epsilon_{ij} \sim i.i.d.N(0, 1)$. Censoring time is independently generated from an exponential distribution to achieve the desired censoring percentage.

In Scenario 1, we set σ_D to $\sqrt{0.5}$ and censoring percentage to 30%; in Scenario 2, we set σ_D to $\sqrt{1}$ and censoring percentage to 30%; in Scenario 3, we set σ_D to $\sqrt{0.5}$ and censoring percentage to 10%; in Scenario 4, we set σ_D to $\sqrt{1}$ and censoring percentage to 10%.

We additionally simulated data settings where there are higher proportions of censoring (Scenario 5) and larger numbers of random effects (Scenario 6). We describe the data generation schemes and simulation results from those settings in the Supplemental Materials (Web Appendices 2–3, Web Table 1–6).

For each scenario, we generate 100 data sets and apply the proposed method to select the non-zero fixed or random effects in the first-stage model. After obtaining the selected variables, we fit the second-stage model including only the selected effects. The tuning parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are determined by minimizing the BIC criterion, as defined in (15). The model without variable selection is also fitted for comparison.

3.2. *Simulation Results*

For Scenarios 1–4, we present the fixed- and random-effect selection results in Table 1, fixed-effect estimation results in Table 2, and random-effect estimation results in Table 3. For fixed effects, the average correct selection rates are 100% for both non-zero and zero effects in longitudinal component, and 100% for non-zero and 98% for zero effects in survival component. The longitudinal fixed-effect estimates do not show severe biases in the first-stage estimation, and the biases are further reduced to less than 1% in the second-stage estimation. The survival fixed-effect estimates show 15–25% biases in the first-stage estimation, and the biases are reduced to below 4% in the second-stage estimation.

For random effects, the average rates of correct selection are 100% for non-zero and 94% for zero effects in longitudinal component, and approximately 96% for non-zero and 90% for zero effects in survival component. For non-zero random effects, the estimates in longitudinal component have biases ranging from 8% to 17% in the first-stage estimation, and the biases are reduced to below 6% in the second-stage estimation. The survival non-zero random effect estimates show

Table 2
Estimation of fixed effects $\beta_{1,j}$ and $\beta_{2,j}$ in longitudinal and survival components for Scenarios 1–4

		$\hat{\beta}_{1,j} \pm$ SE (Coverage probability) for longitudinal component[a]				
		Intercept	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$
Scenarios	True value β	1	1	0	3	0
1	W/O selection $\hat{\beta}$	$1.002 \pm 0.076(87\%)$	$1.007 \pm 0.075(81\%)$	-0.003	$3.009 \pm 0.077(96\%)$	0.000
	1st stage $\hat{\beta}$	$1.006 \pm 0.071(94\%)$	$0.989 \pm 0.072(88\%)$	0.000	$2.988 \pm 0.074(97\%)$	0.000
	2nd stage $\hat{\beta}$	$0.999 \pm 0.068(95\%)$	$1.003 \pm 0.067(85\%)$	0.000	$3.003 \pm 0.070(98\%)$	0.000
2	W/O selection $\hat{\beta}$	$0.995 \pm 0.108(76\%)$	$0.998 \pm 0.111(66\%)$	-0.005	$2.999 \pm 0.111(89\%)$	0.001
	1st stage $\hat{\beta}$	$1.002 \pm 0.099(79\%)$	$0.979 \pm 0.103(74\%)$	0.000	$2.979 \pm 0.096(89\%)$	0.000
	2nd stage $\hat{\beta}$	$0.994 \pm 0.108(82\%)$	$0.996 \pm 0.108(75\%)$	0.000	$2.997 \pm 0.094(94\%)$	0.000
3	W/O selection $\hat{\beta}$	$1.000 \pm 0.076(84\%)$	$1.005 \pm 0.075(80\%)$	-0.003	$3.007 \pm 0.078(96\%)$	-0.001
	1st stage $\hat{\beta}$	$1.007 \pm 0.070(96\%)$	$0.989 \pm 0.072(91\%)$	0.000	$2.988 \pm 0.074(98\%)$	0.000
	2nd stage $\hat{\beta}$	$0.998 \pm 0.070(93\%)$	$1.002 \pm 0.066(88\%)$	0.000	$3.003 \pm 0.07(98\%)$	0.000
4	W/O selection $\hat{\beta}$	$0.994 \pm 0.113(72\%)$	$1.002 \pm 0.111(63\%)$	-0.005	$3.003 \pm 0.109(89\%)$	0.001
	1st stage $\hat{\beta}$	$1.004 \pm 0.102(85\%)$	$0.983 \pm 0.103(78\%)$	0.000	$2.982 \pm 0.098(91\%)$	0.000
	2nd stage $\hat{\beta}$	$0.995 \pm 0.108(81\%)$	$1.000 \pm 0.104(73\%)$	0.000	$2.999 \pm 0.091(95\%)$	0.000
		$\hat{\beta}_{2,j} \pm$ SE (Coverage probability) for survival component[a]				
		Intercept	$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	$X_{2,4}$
Scenarios	True value β	—	1	0	0	1
1	W/O selection $\hat{\beta}$		$1.140 \pm 0.275(86\%)$	-0.009	0.010	$1.141 \pm 0.308(81\%)$
	1st stage $\hat{\beta}$		$0.854 \pm 0.224(59\%)$	-0.001	-0.005	$0.853 \pm 0.255(62\%)$
	2nd stage $\hat{\beta}$		$1.000 \pm 0.263(89\%)$	0.001	0.014	$0.996 \pm 0.285(80\%)$
2	W/O selection $\hat{\beta}$		$1.110 \pm 0.370(79\%)$	-0.023	-0.013	$1.137 \pm 0.424(79\%)$
	1st stage $\hat{\beta}$		$0.751 \pm 0.312(35\%)$	0.001	-0.007	$0.778 \pm 0.348(40\%)$
	2nd stage $\hat{\beta}$		$0.990 \pm 0.417(83\%)$	0.017	0.015	$1.030 \pm 0.636(81\%)$
3	W/O selection $\hat{\beta}$		$1.157 \pm 0.256(84\%)$	0.005	0.040	$1.145 \pm 0.250(88\%)$
	1st stage $\hat{\beta}$		$0.878 \pm 0.187(63\%)$	-0.001	-0.003	$0.876 \pm 0.187(66\%)$
	2nd stage $\hat{\beta}$		$1.023 \pm 0.158(98\%)$	0.000	0.003	$1.021 \pm 0.161(92\%)$
4	W/O selection $\hat{\beta}$		$1.113 \pm 0.259(84\%)$	0.000	0.054	$1.132 \pm 0.299(85\%)$
	1st stage $\hat{\beta}$		$0.792 \pm 0.207(53\%)$	0.000	0.001	$0.810 \pm 0.245(51\%)$
	2nd stage $\hat{\beta}$		$1.008 \pm 0.187(91\%)$	0.003	0.002	$1.013 \pm 0.239(86\%)$

[a] $\hat{\beta}s$ are the averages of estimates over the 100 data sets; SE is the empirical standard error of the 100 $\hat{\beta}s$; For each data set, the 95% confidence interval based on the parameter and standard error estimates is calculated and the corresponding coverage probabilities for the true value over the 100 data sets are included in the parentheses. SE and coverage probability are only reported for non-zero variables.

up to 42% biases in the first-stage estimation; in the second stage, the biases are reduced to less than 8%. For zero random effects, both the first- and second-stage estimates in longitudinal component have biases below 2%. The survival zero random effect estimates generally have less than 10% biases in both stage estimations.

Simulation results for settings with higher proportions of censoring and larger numbers of random effects are reported in the online Supplemental Materials (Web Table 1–6). Briefly, we find that the probabilities of correct selection remain excellent for those data settings.

One consequence of including more random effects is the increased computing time. The complexity of Gaussian quadrature increases exponentially with the dimension of the random effect vector. In this research, we used three quadrature points. With three quadrature points, each data set in Scenarios 1–4 took approximately 20 minutes to complete the first stage variable selection under one tuning parameter; and it took another 10 minutes in the second stage estimation. When we increased the number of random effects to 8 (as in Scenario 6), the computation time increased to 10 and 5 hours, respectively. The computing time is estimated on a single CPU (Intel(R) Xeon(R) CPU E7- 4830 @ 2.13GHz) and 4 GB memory in the Unix system. The total computing time depended on the number of tuning parameters. Other factors, such as the shape of the likelihood function could also influence the approximation accuracy of Gaussian quadrature and the computing time.

Generally, mis-selection rate increases as the censoring rate increases or the variance magnitude σ_D decreases, since

Table 3
Estimation of random effects $\sqrt{D_{1kk}}$ and $\sqrt{D_{2kk}}$ in longitudinal and survival components for Scenarios 1–4

		$\sqrt{\hat{D}_{1kk}}$ for Longitudinal component[a]					$\sqrt{\hat{D}_{2kk}}$ for Survival component[a]				
		Intercept$_1$	$Z_{1,1}$	$Z_{1,2}$	$Z_{1,3}$	$Z_{1,4}$	*Intercept*$_2$	$Z_{2,1}$	$Z_{2,2}$	$Z_{2,3}$	$Z_{2,4}$
Scenarios	True value $\sqrt{D_{kk}}$	0.707	0.707	0	0.707	0	0.707	0.707	0	0	0.707
1	W/O selection $\sqrt{\hat{D}_{kk}}$	0.658	0.668	0.112	0.788	0.133	0.770	0.799	0.441	0.775	0.944
	1st stage $\sqrt{\hat{D}_{kk}}$	0.824	0.763	0.003	0.771	0.001	0.710	0.556	0.031	0.073	0.491
	2nd stage $\sqrt{\hat{D}_{kk}}$	0.695	0.690	0.017	0.735	0.020	0.698	0.718	0.054	0.170	0.708
3	W/O selection $\sqrt{\hat{D}_{kk}}$	0.658	0.665	0.102	0.790	0.130	0.757	0.793	0.342	0.644	0.886
	1st stage $\sqrt{\hat{D}_{kk}}$	0.828	0.750	0.002	0.741	0.001	0.727	0.585	0.006	0.028	0.440
	2nd stage $\sqrt{\hat{D}_{kk}}$	0.695	0.690	0.005	0.734	0.016	0.690	0.731	0.021	0.071	0.725
		Intercept$_1$	$Z_{1,1}$	$Z_{1,2}$	$Z_{1,3}$	$Z_{1,4}$	*Intercept*$_2$	$Z_{2,1}$	$Z_{2,2}$	$Z_{2,3}$	$Z_{2,4}$
	True value $\sqrt{D_{kk}}$	1	1	0	1	0	1	1	0	0	1
2	W/O selection $\sqrt{\hat{D}_{kk}}$	0.877	0.913	0.105	1.103	0.136	1.034	1.094	0.552	0.870	1.262
	1st stage $\sqrt{\hat{D}_{kk}}$	1.143	1.085	0.000	1.140	0.000	0.923	0.715	0.043	0.017	0.580
	2nd stage $\sqrt{\hat{D}_{kk}}$	0.941	0.955	0.004	1.035	0.010	0.970	0.995	0.083	0.146	1.059
4	W/O selection $\sqrt{\hat{D}_{kk}}$	0.873	0.909	0.097	1.093	0.135	1.000	1.046	0.391	0.714	1.163
	1st stage $\sqrt{\hat{D}_{kk}}$	1.140	1.080	0.000	1.132	0.000	0.956	0.783	0.014	0.010	0.591
	2nd stage $\sqrt{\hat{D}_{kk}}$	0.938	0.953	0.003	1.032	0.015	0.923	0.975	0.038	0.053	1.006

[a] $\sqrt{\hat{D}_{1kk}}$ and $\sqrt{\hat{D}_{2kk}}$ are the averages of estimates over the 100 data sets.

smaller variance σ_D means less resolution between non-zero and zero random effects. The mis-selection subsequently leads to larger estimation bias. The influence of censoring rate on selection accuracy is greater than that of variance. Increased number of random effects does not necessarily lead to worse selection accuracy, but it tends to slightly increase estimation bias, which may be due to the reduced approximation accuracy of Gaussian quadrature method. The estimates from the model without variable selection generally have more biases than the second-stage estimates, especially for the zero effects. In summary, we contend that the proposed variable selection and estimation method works well even under high proportions of censoring and large number of random effects. The two stage procedure ensures good selection performance in the first stage and reduced biased parameter estimation in the second stage.

4. Data Application

To illustrate the method, we analyzed observational data from the CHF study. As previously stated, the main purpose of the investigation is to assess the effects of medication adherence on disease exacerbation and patient survival. For the survival outcome, we modeled the time from the first recorded CHF diagnosis to patient mortality, which could be censored on December 31, 2009. For the longitudinal outcome, we modeled the repeatedly measured BNP levels as markers of disease exacerbation. Because the distribution of BNP skewed strongly to the right, we used the logarithmic-transformed BNP (log(BNP)) in the model. Medication adherence, the independent variable of primary interest, was the average proportion of days covered (PDC) by all prescribed medications within each patient (Choudhry et al., 2009). Besides PDC, seven other risk factors were considered, including systolic blood pressure (SBP), diastolic blood pressure (DBP), BMI, gender, age at CHF diagnosis date (IndexAge), number of comorbidities (NumComorbid), and number of medications taken (NumMed). We also considered interactions among SBP, DBP, BMI, PDC, and gender.

In the study sample, 58.3% of the subjects were females and the average BMI was 32.7 (kg/m^2). The average age for the study cohort at the CHF diagnosis date was 62.7 years. On average, the study subjects had 5.1 comorbidities and took 8.4 medications with a mean PDC of 0.327. Among the covariates, concurrently measured SBP (mean: 134.8 mmHg; SD: 24.2 mmHg) and DBP (mean: 77.0 mmHg; SD: 16.0 mmHg) were recorded at the time of BNP assessment; the remaining variables were collected as baseline covariates. The censoring percentage was 64.1%, and median time to death was 4115 days (11.3 years).

For longitudinally measured BNP levels, we use linear mixed-effects model $\log(\text{BNP})_{ij} = \boldsymbol{x}_{1,ij}\boldsymbol{\beta}_1 + \boldsymbol{z}_{1,ij}\boldsymbol{\Gamma}_1\boldsymbol{b}_i + \varepsilon_{ij}$ for $i = 1, \ldots, 1702$, and $j = 1, \ldots, n_i$. We let $\boldsymbol{x}_{1,ij} = (1, \text{DBP}_{ij}, \text{SBP}_{ij}, \text{BMI}_i, \text{PDC}_i, \text{Gender}_i, \text{DBP}_{ij}\times\text{Gender}_i, \text{SBP}_{ij}\times\text{Gender}_i, \text{BMI}_i \times \text{Gender}_i, \text{PDC}_i \times \text{Gender}_i, \text{NumComorbid}_i, \text{NumMed}_i, \text{IndexAge}_i)$ be the design matrix of the fixed effects and $\boldsymbol{z}_{1,ij} = (1, \text{DBP}_{ij}, \text{SBP}_{ij}, \text{BMI}_i, \text{PDC}_i)$ be the design matrix of the random effects. We assume that $\boldsymbol{b}_i$ follows $N(0, \boldsymbol{I}_5)$ and we let $\varepsilon_{ij} \sim i.i.d. N(0, \sigma^2)$ be the measurement error.

For mortality, we assume that the survival time t_i follows a Weibull distribution. We use a proportional hazard model $h(t_i) = h_0(t_i)\exp(\boldsymbol{x}_{2,i}\boldsymbol{\beta}_2 + \boldsymbol{z}_{2,i}\boldsymbol{\Gamma}_2\boldsymbol{b}_i)$, with baseline hazard $h_0(t_i) = \alpha\lambda t_i^{\alpha-1}$ for $i = 1, \ldots, 1702$, where α is the shape

Table 4
Results for the heart failure patient data analysis

	Longitudinal component		Survival component	
	Fixed effect[a]	Variance component[b]	Fixed effect[a]	Variance component[b]
Intercept	5.0042 ± 0.2321	2.6735	—	0.9657
DBP	0.0145 ± 0.0016	0	0	0
SBP	0	0.0133	0	0
BMI	-0.0299 ± 0.0030	0	0	0
PDC	0	2.8857	0	1.7911
Gender	0	—	0	—
DBP × Gender	0	—	0	—
SBP × Gender	0	—	0	—
BMI × Gender	0	—	0	—
PDC × Gender	0	—	0	—
Num. of comorbidities	0.1197 ± 0.0196	—	0	—
Num. of drugs	0	—	-0.1163 ± 0.0121	—
Index age	-0.0033 ± 0.0022	—	0.0044 ± 0.0024	—

[a] Estimate of $\boldsymbol{\beta}_1 \pm \text{SE}$ and $\boldsymbol{\beta}_2 \pm \text{SE}$.
[b] Estimate of $\text{diag}(\sqrt{\boldsymbol{D}_1})$ and $\text{diag}\ (\sqrt{\boldsymbol{D}_2})$.

parameter and λ is the scale parameter. We let $\boldsymbol{x}_{2,i} = (1, \text{DBP}_{i1}, \text{SBP}_{i1}, \text{BMI}_i, \text{PDC}_i, \text{Gender}_i, \text{DBP}_{i1} \times \text{Gender}_i, \text{SBP}_{i1} \times \text{Gender}_i, \text{BMI}_i \times \text{Gender}_i, \text{PDC}_i \times \text{Gender}_i, \text{NumComorbid}_i, \text{NumMed}_i, \text{IndexAge}_i)$ be the design matrix for the fixed effects and $\boldsymbol{z}_{2,i} = (1, \text{DBP}_{i1}, \text{SBP}_{i1}, \text{BMI}_i, \text{PDC}_i)$ be the design matrix for the random effects. Given the random effect $\boldsymbol{b}_i$, we assume that $\log(\text{BNP})_{ij}$, $\log(\text{BNP})_{ij'}$ and t_i are conditionally independent.

Data analytical results are presented in Table 4. For longitudinally measured BNP, our procedure selects DBP, BMI, NumComorbid, and IndexAge as non-zero fixed effects; SBP and PDC as non-zero random effects. For the survival outcome, NumMed and IndexAge are selected as the non-zero fixed effects; PDC as non-zero random effect. The residual plots (Web Figure 1) show no violation of basic model assumptions for the two outcomes. The selected model has a smaller BIC value than the full model and a reduced model including all fixed effects and random intercept.

The effects of the selected variables on the outcomes are in expected directions. In the longitudinal model, DBP is positively associated with BNP ($\beta = 0.0145$) (greater diastolic dysfunction is associated with increased BNP level). BMI exhibits a significant negative association with BNP. For each unit of increase in BMI, log-BNP level decreases by 0.0299 ($\beta = -0.0299$). This result is not surprising as patients at advanced stage of CHF (indicated by greater BNP values) tend to have deteriorated health and much reduced body weight. Interestingly, blood pressure is not found to be associated with the survival outcome, which is influenced more strongly by the number of medications. Patients taking more medications have reduced mortality risk ($\beta = -0.1163$). Patients who are older at CHF diagnosis tend to have significantly increased mortality risk ($\beta = 0.0044$). PDC, our primary variable of interest, has non-zero random effects in both longitudinal (SD = 2.8857) and survival (SD = 1.7911) components, which implies that medication adherence is the underlying latent process influencing both the BNP level and patient survival, and further suggests that the effects of medication adherence on the outcomes may vary across subjects. The shared random intercepts in the longitudinal component (SD = 2.6735), and in the survival component (SD = 0.9657) are also non-zero, which implies a strong within-patient correlation between the two outcomes as well. The heart failure patient data and code for its variable selection are provided online in the Supplemental Materials.

5. Discussion

Despite the increasing popularity of joint models in practical data analysis, few variable selection tools are available for identifying appropriate models. In this article, we propose a method that simultaneously selects random and fixed effects in a joint model setting. For random effect selection, we apply a Cholesky parametrization to the covariance matrix of random effects and use a group penalty, as previous studies have done (Bondell et al., 2010; Ibrahim et al., 2011). This parametrization has made the mixed-effects selection easily adaptable in the complicated joint model settings. Our simulation study shows that the proposed method could correctly identify important fixed and random effects simultaneously, even in the presence of a high proportion of censoring and a large number of random effects. The two-stage model fitting process has helped to control the estimation biases caused by the inclusion of penalty.

A major challenge of using penalized likelihood for variable selection is the computational complexity. The observed likelihood or the E-step in the EM algorithm involves analytically intractable integration. The MCMC method for integral approximation is computationally intensive. Laplace approximation could be a useful alternative, as it has been shown to offer improved computation efficiency at the expense of extra estimation bias (Ye, Lin, and Taylor, 2008). The Gaussian quadrature method used in the current study exhibits excellent stability (at the threshold of 10^{-7}, our simulation shows a 100% convergence rate in Scenarios 2–6, and 92% convergence rate in Scenario 1; generally, the simulation converges within 60 iterations). As we have demonstrated in the simulation,

the proposed method can easily handle up to eight random effects. A possible alternative of Gaussian quadrature is the pseudo-adaptive Gauss–Hermite quadrature rule, which has been shown to be faster in computation with comparable accuracy in the joint model setting (Rizopoulos, 2012). In practice, considering the fact that most biomedical applications use random effects to accommodate structured data dependency, thus will have a relatively small numbers of random effects, we contend that the proposed method is likely adequate for most applications. Additionally, as we have demonstrated through simulation, the number of quadrature points has limited impact on the accuracy of model selection. As a result, for complicated models one could use a smaller number of quadrature points to enhance computational efficiency in the first stage, and then increase the number of quadrature points in the second stage to achieve desired estimation accuracy. Comparing our simulation results with the reported performance in linear mixed-effects models (Bondell et al., 2010), generalized linear mixed models (Ibrahim et al., 2011), and survival models (Zhang and Lu, 2007), we note that our method has achieved comparable selection and estimation accuracy.

In summary, we show that penalized likelihood method can be used for variable selection in joint model settings. The procedure can be modified for the simultaneous mixed-effects selection in other bi-component models. Our research has demonstrated, through a real data example, that the proposed method provides a useful tool for practical data analysis. The method is easy to implement and it is efficient in computation.

6. Supplementary Materials

Web Appendices 1–3 and Web Tables 1–6 referenced in Section 3; Example data set, Web Figure 1 and computational code referenced in Section 4 are available with this paper at the *Biometrics* website on Wiley Online Library.

Acknowledgements

This research is supported by grant RO1-HL095086 from the National Heart, Lung and Blood Institute, and a grant from Merck & Company.

References

Albert, P. S. and Shih, J. H. (2010). On estimating the relationship between longitudinal measurements and time-to-event data using a simple two-stage procedure. *Biometrics* **66**, 983–987.

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* **66**, 1069–1077.

Choudhry, N. K., Shrank, W. H., Levin, R. L., Lee, J. L., Jan, S. A., Brookhart, M. A., and Solomon, D. H. (2009). Measuring concurrent adherence to multiple related medications. *The American Journal of Managed Care* **15**, 457.

De Gruttola, V. and Tu, X. M. (1994). Modelling progression of cd4-lymphocyte count and its relationship to survival time. *Biometrics*, **50**, 1003–1014.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics* **32**, 407–499.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

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

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association* **99**, 710–723.

Faucett, C. L. and Thomas, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Statistics in Medicine* **15**, 1663–1685.

Feng, S., Wolfe, R. A., and Port, F. K. (2005). Frailty survival model analysis of the national deceased donor kidney transplant dataset using Poisson variance structures. *Journal of the American Statistical Association* **100**, 728–728.

Garcia, R. I., Ibrahim, J. G., and Zhu, H. (2010). Variable selection for regression models with missing data. *Statistica Sinica* **20**, 149.

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics* **67**, 495–503.

Johnson, B. A., Lin, D. Y., and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association* **103**, 672–680.

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics* **34**, 2272–2297.

Morrison, L. K., Harrison, A., Krishnaswamy, P., Kazanegra, R., Clopton, P., and Maisel, A. (2002). Utility of a rapid b-natriuretic peptide assay in differentiating congestive heart failure from lung disease in patients presenting with dyspnea. *Journal of the American College of Cardiology* **39**, 202–209.

Nathoo, F. and Dean, C. (2008). Spatial multistate transitional models for longitudinal event data. *Biometrics* **64**, 271–279.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* **4**, 12–35.

Pu, W. and Niu, X.-F. (2006). Selecting mixed-effects models based on a generalized information criterion. *Journal of Multivariate Analysis* **97**, 733–758.

Rizopoulos, D. (2012). Fast fitting of joint models for longitudinal and event time data using a pseudo-adaptive Gaussian quadrature rule. *Computational Statistics & Data Analysis* **56**, 491–501.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica* **7**, 221–242.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, **58**, 267–288.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **71**, 671–683.

Wang, S., Song, P. X., and Zhu, J. (2010). Doubly regularized reml for estimation and selection of fixed and random effects in linear mixed-effects models. *The University of Michigan Department of Biostatistics Working Paper Series*. http://biostats.bepress.com/umichbiostat/paper89.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330–339.

Xu, J. and Zeger, S. L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **50**, 375–387.

Ye, W., Lin, X., and Taylor, J. M. (2008). A penalized likelihood approach to joint modeling of longitudinal measurements and time-to-event data. *Statistics and Interface* **1**, 33–45.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **68**, 49–67.

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for Cox's proportional hazards model. *Biometrika* **94**, 691–703.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association* **101**, 1418–1429.

Received July 2013. *Revised June* 2014. *Accepted June* 2014.