Statistics and Its Interface Volume 13 (2020) 411–422

A semi-parametric joint latent class model with longitudinal and survival data

Yue Liu, Ye Lin, Jianhui Zhou*, and Lei Liu

In many longitudinal studies, we are interested in both repeated measures of a biomarker and time to an event. When there exist heterogeneous patterns of the longitudinal and survival profiles, we propose a latent class joint model to identify subgroups of subjects and study the association between longitudinal and survival outcomes. The model is estimated by maximizing the full likelihood function. We use B-splines to approximate the baseline hazard function which involves a diverging number of parameters. Asymptotic properties of the estimator for the joint latent class model are investigated. We conduct simulation studies to assess the performance of the developed method. A real data example, Mayo Clinic Primary Biliary Cirrhosis Data, is analyzed using the joint modeling approach.

Keywords and phrases: B-splines, Longitudinal measurements, Mixed effects model, Proportional hazards model, Survival outcome.

1. INTRODUCTION

In the Mayo Clinic trial for primary biliary cirrhosis (PBC) of the liver (Lindor et al., 1994), 312 subjects were randomized into the treatment group with the drug D-penicillamine and the placebo group. Their levels of biomarkers were measured during the follow-up, and the first adverse event (transplanted or dead) was also recorded. Our goal is to study the association between a biomarker, serum bilirubin in mg/dl which was found to be a marker of advanced PBC (Shapiro et al., 1979), and the survival outcomes.

Several statistical approaches have been used to study the association between the longitudinal and survival outcomes. One simple solution is to use repeated measures as time-dependent covariates in the Cox proportional hazards model. However, the longitudinal measures are often imbalanced and subject to measurement errors, thus the direct use of repeated measures in the Cox model could lead to a bias toward the null hypothesis (Prentice, 1982). To reduce the bias, Dafni and Tsiatis (1998) used a two-stage approximation approach to obtain parameter estimation. However, due to possible informative censoring caused by the correlated survival outcome, e.g., informative dropout or terminal event, the longitudinal measurements on the observed subjects are different from the ones on the unobserved subjects, resulting in selection bias.

To reduce the bias from the two-stage model, Wulfsohn and Tsiatis (1997) proposed to jointly model data from the longitudinal and the survival processes at the same time. Repeated measures and the survival outcomes are connected by shared random effects. The joint modeling efforts involve the development of likelihood functions for mixed types of measures. Zeng and Cai (2005) developed justification of the asymptotic properties for the maximum likelihood estimators. It was proved that the maximum likelihood estimator is strongly consistent and follows a multivariate normal distribution asymptotically under certain conditions.

Furthermore, when there exists substantial heterogeneity of the longitudinal and survival outcomes in the population, Proust-Lima et al. (2009) proposed a joint model based on latent class approach assuming conditional independence of longitudinal and survival outcome within each latent class. Liu et al. (2015) extended the approach by relaxing the conditional independence and assuming that longitudinal and survival outcomes share random effects within each latent class, but their parametric model is at risk of mis-specifying the survival distributions.

To utilize the full likelihood function to estimate the parameters, we need to specify the baseline hazard functions in the Cox model for survival outcomes. One approach is to impose a parametric distributional assumption, such as Weibull and Gamma, for parsimonious parameterization as in Liu et al. (2015). However, this assumption is difficult to verify in applications, and the parameter estimators are biased if the distribution is mis-specified. We therefore consider nonparametric techniques to approximate baseline hazard functions, which avoids distributional assumption and model mis-specification. In this work, we estimate the parameters in the joint latent class model by approximating the baseline hazards using B-splines, and investigate the asymptotic properties of the estimators in the semi-parametric setting. The development of the asymptotic properties involves a diverging number of parameters.

The rest of the paper is organized as follows. In Section 2, we introduced the model structure and estimation method. In Section 3, asymptotic properties of the developed estimators are derived. In Sections 4 and 5, simulation studies and

*Corresponding author.

an application to the Mayo PBC data are provided. Discussion and summary are presented in Section 6. Technical proofs of the theorems are deferred to Appendix.

2. MODEL AND ESTIMATION

Our proposed joint latent class model (Liu et al., 2015) has the following structure

$$\pi_{ik} = P(R_{ik} = 1) = \frac{\exp(\mathbf{X}_i^T \gamma_k)}{1 + \sum_{k=1}^{K-1} \exp(\boldsymbol{X}_i^T \gamma_k)} \tag{1}$$

$$y_{ij} | (R_{ik} = 1) = \mathbf{Z}_{ij}^T \eta_k + a_{i,k} + \varepsilon_{ij,k} \tag{2}$$

$$h_i(t | R_{ik} = 1) = h_{0,k}(t) \exp(\mathbf{W}_i^T(t)\omega_k + \delta_k a_{i,k}) \tag{3}$$

where π_{ik} in the multinomial logit model (1) denotes the probability that subject i belongs to latent class $k = 1, ..., K$, and $\mathbf{X}_i$ is the covariate vector for subject i to determine the class membership. For identification, we treat the last latent class K as the reference, i.e., $\gamma_K = 0$. Within the latent class k, the longitudinal outcomes y_{ij} are modeled using a linear mixed model with random effect $a_{i,k}$ and covariates $\mathbf{Z}_{ij}$, which could include the visit time t_{ij}, with the corresponding parameters η_k in (2). Meanwhile, the survival outcome is modeled using the Cox proportional hazards model (3) with the same random effect $a_{i,k}$ and the covariates $\mathbf{W}_i(t)$. The coefficients δ_k in (3) connect the two outcomes, and reflect the strength of the connection. Note that all parameters are class-specific for the longitudinal and survival outcomes. Different from Proust-Lima et al. (2009) where only the latent class membership is shared between the longitudinal and survival processes, both the latent class membership and random effect are shared between the two processes in our proposed model. Thus, our model is more general than that of Proust-Lima et al. (2009) since the random effect accounts for both the correlation among longitudinal measurements and the association between the two processes in the proposed model, which is appropriate when the conditional independence between the processes within each class as assumed in Proust-Lima et al. (2009) is violated.

Assuming $\epsilon_{ij,k} \sim N(0, \tau_k^2)$ and $a_{i,k} \sim N(0, \sigma_k^2)$, with known baseline hazard functions $h_{0,k}(t)$, the likelihood function of above joint latent class model is

$$\begin{aligned} &L(\beta \mid \mathbf{X}, \mathbf{Z}, \mathbf{W}(t)) \\ &= \prod_{i=1}^{n} \sum_{k=1}^{K} L_{i,k}(\beta_k \mid \mathbf{X}_i, \mathbf{W}_i(t), \mathbf{Z}_i) \\ &= \prod_{i=1}^{n} \sum_{k=1}^{K} P(R_{ik} = 1) \int \prod_{j=1}^{n_i} f(y_{ij} \mid R_{ik} = 1, a_{i,k}) \\ &\times \left[\exp(\mathbf{W}_i^T(t_i)\omega_k + \delta_k a_{i,k}) h_{0,k}(t_i)\right]^{\Delta_i} \\ &\times \exp\left\{ - \int_0^{t_i} \left[\exp(\mathbf{W}_i^T(t)\omega_k + \delta_k a_{i,k}) h_{0,k}(t)\right] \mathrm{d}t \right\} \\ &\times f(a_{i,k}) \mathrm{d}a_{i,k}, \end{aligned} \tag{4}$$

where the parameter

$$\begin{aligned} \beta = (&\gamma_1, ..., \gamma_K, \eta_1, ..., \eta_K, \omega_1, ..., \omega_K, \delta_1, ..., \delta_K, \\ &\sigma_1^2, ..., \sigma_K^2, \tau_1^2, ..., \tau_K^2) \end{aligned}$$

and β_k is the subset of β associated with the kth class, $f(y_{ij} \mid R_{ik} = 1, a_{i,k}) = \frac{1}{\sqrt{2\pi\tau_k^2}} \exp(-\frac{(y_{ij} - \mathbf{Z}_{ij}^T \eta_k - a_{i,k})^2}{2\tau_k^2})$, $f(a_{i,k}) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp(-\frac{a_{i,k}^2}{2\sigma_k^2})$, and Δ_i is the censoring indicator.

Liu et al. (2015) approximated $h_{0,k}(t)$ by Weibull distribution to simplify the estimation. To avoid potential model mis-specification caused by distributional assumption on $h_{0,k}(t)$, we consider nonparametric modeling of $h_{0,k}(t)$. Specifically, we use B-splines to approximate the baseline hazard functions $h_{0,k}(t)$ in (4), where each $h_{0,k}(t)$ is expressed by

$$h_{0,k}(t) = \sum_{s=1}^{k_n+m} \alpha_{s,k} B_s(t) + e_k(t) \tag{5}$$

where $B_s(t)$ are B-spline basis functions with k_n inner knots on $[0, T]$ and order m, $\alpha_{s,k}$ are the B-spline coefficients for approximation, and $e_k(t)$ is the approximation error that converges uniformly to 0 for any $t \in [0, T]$ under some smoothness condition on $h_{0,k}(t)$. With the approximation to $h_{0,k}(t)$, the log-likelihood function in (4) is approximated as

$$\begin{aligned} &\sum_{i=1}^{n} \log \Bigg\{ \sum_{k=1}^{K} P(R_{ik} = 1) \int \prod_{j=1}^{n_i} f(y_{ij} \mid R_{ik} = 1, a_{i,k}) \\ &\times \left[\exp(\mathbf{W}_i \boldsymbol{\omega}_k + \delta_k a_{i,k}) \sum_{s=1}^{k_n+m} \alpha_{s,k} B_s(t)(t_i)\right]^{\Delta_i} \\ &\exp\left(- \int_0^{t_i} \left[\exp(\mathbf{W}_i \boldsymbol{\omega}_k + \delta_k a_{i,k}) \sum_{s=1}^{k_n+m} \alpha_{s,k} B_s(t)\right] \mathrm{d}t \right) \\ &\times f(a_{i,k}) \mathrm{d}a_{i,k} \Bigg\}. \end{aligned} \tag{6}$$

For a fixed class number K, parameters in (6) are estimated through maximizing the above approximated log full likelihood function. For the integrals of nonlinear functions in (6), we adopt the Gaussian quadrature approach (SAS Proc NLMIXED) to approximate them in computation. The Bayesian Information Criterion (BIC) is used to choose the optimal number of latent classes, which minimizes the BIC values (Lin et al., 2002; Nagin, 1999; Muthen and Shedden, 1999; Nagin and Tremblay, 2001) to balance model complexity and accuracy.

With the selected optimal class number K and the corresponding estimated parameters in the joint latent class model, we use model based classification rule (Magidson and Vermunt, 2004) to classify each subject into

one of the latent classes with specific characteristics. With the following computed class membership probabilities

$$\hat{\pi}_{ik|\hat{\beta},\mathbf{O}_i} = \frac{\hat{\pi}_{ik} L_{i,k}(\hat{\beta}_k \mid \mathbf{O}_i)}{\sum_{k=1}^{K} \hat{\pi}_{ik} L_{i,k}(\hat{\beta}_k \mid \mathbf{O}_i)}, \tag{7}$$

where $\mathbf{O}_i = (\mathbf{X}_i, \mathbf{W}_i(t), \mathbf{Z}_i)$, and $\hat{\pi}_{ik}$ is the prior probability with parameter estimates plugged in (1), we classify subject i to class k' if $\hat{\pi}_{ik'} L_{i,k'}(\hat{\Theta}_{k'})$ is the largest one among $k = 1, ..., K$.

3. ASYMPTOTIC PROPERTIES

In this section, we present the asymptotic properties of the developed estimators by maximizing the approximated log of the full likelihood function in (6). We denote the log of full likelihood function by $\sum_{i=1}^{n} L_{n_i}(Y_{n_i}, \beta)$, and the corresponding approximated log-likelihood function with B-splines by $\sum_{i=1}^{n} \tilde{L}_{n_i}(Y_{n_i}, \theta_n)$, where θ_n denotes all parameters to be estimated in (6) with the dimension p_n. We divide θ_n into $\theta_n = (\beta, \alpha_n)$ with $\alpha_n = (\alpha_{n,1}, \ldots, \alpha_{n,p_{n_2}})$ including the B-spline approximation coefficients for the baseline hazard functions, and β including all the other parameters. We use p_1 to denote the dimension of β, and p_{n_2} to denote the dimension of α_n. Here, we use vector Y_{n_i} to represent all the observed outcomes from the ith subject. Estimators of α_n and β are obtained by maximizing the approximated log-likelihood function in (6).

We present the asymptotic properties of the developed estimators under the following regularity conditions.

A1 The baseline hazard functions $h_{0,k}(t)$ are continuous functions on $[0, T]$, with bounded r-th derivatives on $[0, T]$ for some $r \geq 4$.

A2 The observations Y_{n_i} are distributed independently with density function $f(Y_{n_i}, \beta, h_{0,k})$, which is approximated by $\tilde{f}_n(Y_{n_i}, \theta_n)$, and the first-order derivatives satisfy

$$E_{\theta_n}\left\{\frac{\partial \log \tilde{f}_n(Y_{n_i}, \theta_n)}{\partial \beta_i}\right\} = O_p(\frac{1}{\sqrt{n}}),$$
$$\sup_j E_{\theta_n}\left\{\frac{\partial \log \tilde{f}_n(Y_{n_i}, \theta_n)}{\partial \alpha_{nj}}\right\} = O_p(\frac{1}{\sqrt{nk_n}})$$

for $i = 1, \ldots, p_1$. We assume that there is a large enough open set $w_n \in R^{p_n}$ such that for all $\theta_n \in \omega_n$, the following condition is satisfied

$$E_{\theta_n}\left\{\frac{\partial \log \tilde{f}_n(Y_{n_i}, \theta_n)}{\partial \theta_{nj}} \frac{\partial \log \tilde{f}_n(Y_{n_i}, \theta_n)}{\partial \theta_{nk}}\right\},$$
$$= -E_{\theta_n}\left\{\frac{\partial^2 \log \tilde{f}_n(Y_{n_i}, \theta_n)}{\partial \theta_{nj} \partial \theta_{nk}}\right\},$$

where $j, k = 1, 2, \ldots, p_n$.

A3 We denote the Fisher information matrix as

(8)

$$I_n(\theta_n) = I_n(\beta, \alpha_n)$$
$$= E_{\theta_n}\left[\left\{\frac{\partial \log \tilde{f}_n(Y_{n_i}, \theta_n)}{\partial \theta_n}\right\}\left\{\frac{\partial \log \tilde{f}_n(Y_{n_i}, \theta_n)}{\partial \theta_n}\right\}^T\right],$$

and partition it as

$$I_n(\theta_n) = \begin{pmatrix} I_1(\theta_n) & I_{12}(\theta_n) \\ I_{21}(\theta_n) & I_2(\theta_n) \end{pmatrix}, \tag{9}$$

where $I_1(\theta_n)$ is the partial information matrix for the parameters in β, and $I_2(\theta_n)$ is for the parameters in α_n. We assume that there is a large enough open set $d_n \in R^{p_n}$ which contains the true values of β and α_n, such that as $n \to \infty$, with probability tending to 1, we have

$$0 < C_1 < \lambda_{min}\{I_1(\theta_n)\} \leq \lambda_{max}\{I_1(\theta_n)\} < C_2 < \infty,$$
$$0 < \frac{C_3}{k_n} < \lambda_{min}\{I_2(\theta_n)\} \leq \lambda_{max}\{I_2(\theta_n)\} < \frac{C_4}{k_n} < \infty,$$

for any $\theta_n \in d_n$, and for $j, k = 1, 2, \ldots, p_n$,

$$E_{\theta_n}\left\{\frac{\partial \log \tilde{f}_n(Y_{n_i}, \theta_n)}{\partial \theta_{nj}} \frac{\partial \log \tilde{f}_n(Y_{n_i}, \theta_n)}{\partial \theta_{nk}}\right\}^2 < C_5 < \infty,$$
$$\sup_{j,k} E_{\theta_n}\left\{\frac{\partial^2 \log \tilde{f}_n(Y_{n_i}, \theta_n)}{\partial \beta_j \partial \beta_k}\right\}^2 < C_5 < \infty,$$
$$\sup_{j,k} E_{\theta_n}\left\{\frac{\partial^2 \log \tilde{f}_n(Y_{n_i}, \theta_n)}{\partial \alpha_{nj} \partial \alpha_{nk}}\right\}^2 = O_p(k_n^{-1}),$$

and

$$\sup_{j,k} E_{\theta_n}\left\{\frac{\partial^2 \log \tilde{f}_n(Y_{n_i}, \theta_n)}{\partial \beta_j \partial \alpha_{nk}}\right\}^2 = o_p(k_n^{-3}).$$

A4 Assume that there is a large enough open set $z_n \in R^{p_n}$ which contains the true values of β and α_n, such that

$$\sup_{j,j',j''} E_{\theta_n}\left\{\frac{\partial^3 log \tilde{f}_n(Y_{n_i}, \theta_n)}{\partial \theta_{nj} \partial \theta_{nj'} \theta_{nj''}}\right\}^2 = O_p(k_n^{-1}),$$

for any $\theta_n \in z_n$, and $j, j', j'' = 1, \ldots, p_n$.

Condition A1 assumes the smoothness of the target baseline hazard functions to ensure the B-spline approximation errors converging to 0 uniformly on $[0, T]$. Since we use the approximated log full likelihood function, we assume that the expectation of the first-order derivatives of the approximated log full likelihood functions is small at the parameter θ_n in Condition A2. The second-order derivatives follow

exchangeability in a large enough open set. This condition is the same as in Fan and Peng (2004) when the baseline hazard functions can be exactly represented by B-splines. Otherwise, this exchangeability condition is assumed for the approximated parametric density function, which is easier to verify. In Condition A3, we consider to partition the information matrix into four parts, based on the diverging dimensionality of the parameter α_n in the approximated likelihood function and the fixed dimensionality of the parameter β in the true likelihood function. We also assume the rates of eigenvalues of the partial information matrix $I_1(\theta_n)$ and $I_2(\theta_n)$. Note that the assumed rate of eigenvalues for $I_2(\theta_n)$ is different from the one in Fan and Peng (2004) due to the employment of B-splines, and the other rates in Condition A3 are specified accordingly. Condition A4 regulates the third-order derivatives to be small enough. Similar conditions on the derivatives as in A3 and A4 have been assumed in Fan and Peng (2004). All assumptions in Condition A3 and Condition A4 can be verified in a special case, in which only event times are observed from one class without covariates.

Under the above conditions, we achieve the following asymptotic properties of the developed estimators for our semi-parametric joint latent class model with survival and longitudinal outcomes.

Theorem 1 Assuming A2-A3 and $k_n^2/n \to 0$ as $n \to \infty$, there exists a local maximizer $\hat{\beta}$ and $\hat{\alpha}_n$ of $\tilde{L}_n(\beta, \alpha_n)$ such that

$$\|\hat{\alpha}_n - \alpha_n\|_2 = O_p(k_n/\sqrt{n}) \text{ and } \|\hat{\beta} - \beta\|_2 = O_p(1/\sqrt{n}) .$$

For estimator of the coefficients in the parametric part of the joint latent class model, we have the following asymptotic normality.

Theorem 2 Assuming A2-A3 and $k_n^2/n \to 0$ as $n \to \infty$, the $\sqrt{n}$-consistent estimator $\hat{\beta}$ in Theorem 1 has the following asymptotic distribution

$$\sqrt{n} I_1^{1/2}(\beta, \alpha_n)(\hat{\beta} - \beta) \to_d N(0, I_{p_1 \times p_1}) .$$

For the nonparametric part in the joint latent class model, we have the following asymptotic normality for the estimator of the B-spline approximation coefficients.

Theorem 3 Assuming A2-A4 and $k_n^7/n \to 0$ as $n \to \infty$, the $\sqrt{n}/k_n$-consistent estimator $\hat{\alpha}_n$ in Theorem 1 has the following asymptotic distribution

$$\sqrt{n} A_n I_2^{1/2}(\beta, \alpha_n)(\hat{\alpha}_n - \alpha_n) \to_d N(0, G) ,$$

where A_n is a $q \times p_{n_2}$ matrix such that $A_n A_n^T \to G$, and G is a $q \times q$ nonegative symmetric matrix.

By Theorem 3, we have the asymptotic normal distribution for $\hat{\alpha}_n$. Accordingly, we obtain the following property for the baseline hazard function estimator in the survival part.

Theorem 4 Assuming A1-A4, $k_n^7/n \to 0$, and $n/k_n^{1+2r} \to 0$ as $n \to \infty$, we have, for any $t \in [0, T]$,

$$(n/k_n)^{1/2}(\hat{h}_{0,k}(t) - h_{0,k}(t)) \to_d N(0, \sigma_k^2(t)) ,$$

for each k, where $\hat{h}_{0,k}(t) = B_n^T(t)\hat{\alpha}_{n,k}$, $\sigma_k^2(t) = \lim_{n\to\infty} \frac{1}{k_n} B_n^T(t) I_{2,k}^{-1}(\theta_n) B_n(t)$, and $I_{2,k}(\theta_n)$ is the partial information matrix for $\alpha_{n,k}$.

To approximate the unknown baseline hazard functions by B-splines, the number of inner knots k_n needs to go to infinity as $n \to \infty$. Therefore, the asymptotic theories involve a diverging number of parameters. Theorem 1 states the convergence rates of the estimators for β in the original model and for α_n induced by B-splines in the approximation model. By maximizing the approximated full log-likelihood function, we achieve the $\sqrt{n}$-consistent estimator for β, and further show that the estimator has asymptotically normal distribution in Theorem 2. Since the dimensionality of α_n goes to infinity as n increases, we use the matrix A_n to project the parameter vector onto a space with fixed dimensionality in Theorem 3 as in Fan and Peng (2004). Using the asymptotic distribution of $\hat{\alpha}_n$ obtained in Theorem 3, we obtain the point-wise asymptotic normality of the B-spline estimator for the baseline hazard functions in Theorem 4.

4. SIMULATION STUDIES

In this section, simulation studies are conducted to assess the performance of the developed method, and the performance is compared with the existing parametric method in Liu et al. (2015). For each subject, a covariate x_i is generated from the uniform distribution $U(0, 1)$. Longitudinal responses are generated at time points $0, 1, ..., 14$. A noninformative censoring scheme is applied to the generated times to event using the generated censoring times from the uniform distribution $12 + U(1, 3)$, which yields a censoring rate around 20%. Longitudinal measurements before censoring time are used for estimation. For generating times to event, the baseline hazard functions are specified as the Log-logistic distributions. Each dataset with 2 latent classes is generated as follows.

Latent Class Part: $\text{logit}(p_i) = 0 - 0.5x_i$.

Class 1:
Longitudinal Part: $y_{ij,1} = -1 - x_i - t_{ij} + a_{i,1} + \varepsilon_{ij,1}$.
Survival Part: $h_{i,1}(t_i) = h_{0,1}(t_i) \exp(x_i - a_{i,1})$, where the baseline hazard follows a Log-logistic distribution with $\alpha = 4$ and $\lambda = \frac{1}{2000}$, i.e., $h_{0,1}(t_i) = \frac{4x^{4-1}/2000}{1+x^4/2000}$.

Class 2:
Longitudinal Part: $y_{ij,2} = 1 + x_i + t_{ij} + a_{i,2} + \varepsilon_{ij,2}$.

Table 1. Summary of simulation results. SE is the empirical standard error of the parameter estimates; SEM is the mean of the standard error estimates; CP is the coverage probability of the 95% confidence interval

Method		Weibull					B-spline			
Description	Parameter	True Value	Bias	SE	SEM	CP	Bias	SE	SEM	CP
Latent Class Part	Intercept	0	−0.0107	0.2055	0.2019	93.5%	−0.0100	0.2056	0.2019	93.5%
	X_i	−0.5	0.0094	0.3623	0.3536	94.3%	0.0086	0.3624	0.3536	94.3%
Class 1	Longitudinal Part									
	Intercept	−1.0	0.0169	0.1879	0.1169	76.3%	−0.0035	0.1635	0.1438	89.8%
	X_i	−1.0	0.0103	0.3248	0.2072	76.3%	0.0149	0.2797	0.2577	92.0%
	Month	−1.0	−0.0019	0.0075	0.0076	95.0%	0.0001	0.0075	0.0076	94.8%
	$\text{Var}(a_{i,1})$	1.0	−0.0014	0.1376	0.0964	82.0%	−0.0147	0.1243	0.1112	89.8%
	$\text{Var}(\varepsilon_{ij,1})$	0.5	< 0.0001	0.0224	0.0220	95.3%	−0.0008	0.0224	0.0219	95.3%
	Survival Part									
	X_i	1.0	0.1094	0.4661	0.3649	87.5%	−0.0079	0.3817	0.3821	93.0%
	δ_1	−1.0	−0.1023	0.1304	0.1203	85.8%	0.0234	0.1147	0.1163	94.8%
Class 2	Longitudinal Part									
	Intercept	1.0	−0.0005	0.1593	0.1461	91.3%	0.0034	0.1553	0.1485	93.5%
	X_i	1.0	0.0111	0.2658	0.2419	92.0%	0.0068	0.2542	0.2474	94.3%
	Month	1.0	0.0011	0.0060	0.0065	96.0%	$< -3e\text{-}5$	0.0060	0.0065	96.5%
	$\text{Var}(a_{i,2})$	1.0	−0.0176	0.1207	0.1076	90.8%	−0.0152	0.1168	0.1090	92.8%
	$\text{Var}(\varepsilon_{ij,2})$	1.0	0.0019	0.0328	0.0333	96.0%	0.0009	0.0326	0.0332	96.0%
	Survival Part									
	X_i	−1.0	−0.0644	0.4146	0.3887	93.5%	−0.0211	0.3908	0.3830	94.8%
	δ_2	1.0	0.1160	0.1317	0.1274	91.0%	0.0493	0.1250	0.1253	95.5%

Survival Part: $h_{i,2}(t_i) = h_{0,2}(t_i)\exp(-x_i + a_{i,2})$, where the baseline hazard follows a Log-logistic distribution with $\alpha = 3$ and $\lambda = \frac{1}{250}$, i.e., $h_{0,2}(t_i) = \frac{3x^{3-1}/250}{1+x^3/250}$.

We specify $a_{i,1} \sim^{iid} N(0,1)$, $a_{i,2} \sim^{iid} N(0,1)$, $\varepsilon_{ij,1} \sim^{iid} N(0,0.5)$, $\varepsilon_{ij,2} \sim^{iid} N(0,1)$, and they are independent of each other.

For the integrals in (6), we use SAS PROC NLMIXED with 50 quadrature points to approximate them. We use B-splines with 4 inner quantile knots and order 3 to approximate the baseline hazard functions. For comparison, we also estimate the parameters using parametric method in Liu et al. (2015) with Weibull distribution for the hazard functions.

We generate 400 data sets with sample size 400 in this simulation study, and estimate the parameters using both the developed semi-parametric method and the parametric method in Liu et al. (2015). The simulation results assuming $K = 2$ are shown in Table 1. Compared with the parametric method assuming mis-specified Weibull distribution for the baseline hazard functions, our method with B-spline approximation to those functions has both smaller biases and smaller standard errors for parameter estimation, especially for the survival part. All the coverage probabilities of the constructed 95% confidence intervals from our method are closer to the nominal value than those from the parametric method for the survival part. For the survival part, the parametric method assuming Weibull distribution yields biased parameter estimates and results in poor coverage probabilities, showing that the developed semi-parametric method outperforms the parametric method when baseline hazard functions are mis-specified.

We also calculate mis-classification rate for each of the 400 generated data sets. The mis-classification rate for each data set is calculated as the percentage of the subjects being classified into the wrong class based on (7). The average mis-classification rate is 0.03%, showing that the classification based on the developed method is very reliable.

The size of latent classes is chosen based on the BIC values. Empirically, we calculate BIC for each proposed size K, and find optimal value to minimize BIC. For 100 generated data sets, we select the optimal size among $K = 1, 2, 3$. The BIC value for $K = 2$ are the smallest for each of the 100 data sets. The developed method correctly identifies the number of latent classes using the BIC criterion.

To evaluate the numerical performance of the proposed method with higher censoring rate close to the real data in Section 5, we ran a simulation study with a higher censoring rate about 45%. There are 100 data sets generated using the same method, except for generating censoring times from the uniform distribution $U(3, 15)$. As in Table 1, the developed method yields smaller biases for parameter estimation and improved coverage probabilities of 95% confidence intervals.

5. APPLICATION

In this section, we apply the developed method to the Mayo Clinic Primary Biliary Cirrhosis Data, collected between 1974 and 1984, in primary biliary cirrhosis (PBC) of

Table 2. Results for 2 latent classes

Description	Parameter	Estimates	SE	P-value
Latent Class Part	Intercept	0.5310	0.1438	0.0003
	Age	−0.1984	0.1454	0.1736
	Gender	0.6301	0.4419	0.1549
Class 1	Longitudinal Part			
	Intercept	0.3787	0.0696	< .0001
	Time	0.1841	0.0067	< .0001
	Trt	-0.2564	0.0738	0.0006
	Age	0.1424	0.0547	0.0097
	Gender	0.0619	0.0852	0.4678
	$\text{Var}(a_{i,1})$	0.9336	0.0647	< .0001
	$\text{Var}(\varepsilon_{ij,1})$	0.2033	0.0097	< .0001
	Survival Part			
	Trt	-0.3739	0.3111	0.2304
	Age	0.5908	0.1155	< .0001
	Gender	−0.2362	0.2635	0.3708
	δ_1	1.3660	0.1867	< .0001
Class 2	Longitudinal Part			
	Intercept	−0.7068	0.0355	< .0001
	Time	0.0108	0.0044	0.0134
	Trt	0.0110	0.0388	0.7778
	Age	−0.0258	0.0249	0.3013
	Gender	0.1650	0.0489	0.0008
	$\text{Var}(a_{i,2})$	0.1688	0.0143	< .0001
	$\text{Var}(\varepsilon_{ij,2})$	0.0511	0.0042	< .0001
	Survival Part			
	Trt	-3.4331	2.1611	0.1132
	Age	3.8183	1.4530	0.0090
	Gender	3.9116	1.3694	0.0046
	δ_2	2.0962	2.4136	0.3858

the liver, to study the association between level of serum bilirubin and hazard of death. There are 312 subjects in our study. We use log(serBlilir) as the longitudinal responses due to data skewness, and consider time as a predictor along with other covariates including treatment (1 for treatment; 0 for placebo), age, and gender (1 for male; 0 for female) in the longitudinal part. We use time to first adverse event (transplanted or dead) as the survival outcome, and consider treatment, age and gender as covariates in the survival part. The censoring rate is 54%, and 51% of the subjects are assigned to the treatment group. Around 12% of the subjects are male. Participants' ages range from 26 to 78.

5.1 Model fitting

We fit our joint latent class model of longitudinal measures, log(serBilir), and survival outcomes, death or transplant, with the size of latent classes varying from 1 to 3. Gaussian quadrature with 50 quadrature points is used for parameter estimation. As in the simulation study, we approximate the baseline hazard functions using B-splines with order 3 and 4 inner quantile knots. The BIC value calculated from the 2 latent classes model is 3601.9, which is smaller than that from 1 class model (4383.4). We cannot obtain valid results from the 3 classes model, indicating the current sample size cannot afford the complexity of the 3 classes model. Thus, we choose $K = 2$ as the optimal size of latent classes for this data example. Due to the same limited sample size issue, we didn't include the interaction between time and treatment in our model. The model setup is as follows.

Latent class part: $logit(p_i) = \alpha_0 + \alpha_1 age + \alpha_2 gender$.

Class $k = 1, 2$:

$y_{ij}|(R_{ik} = 1) = \beta_{0,k} + \beta_{1,k} time + \beta_{2,k} trt + \beta_{3,k} age + \beta_{4,k} gender + a_{i,k} + \varepsilon_{ij,k}$;

$h_i(t|R_{ik} = 1) = h_{0,k}(t) \exp(\gamma_{1,k} trt + \gamma_{2,k} age + \gamma_{3,k} gender + \delta_k a_{i,k})$.

Table 2 summarizes parameter estimates with $K = 2$. Based on the results, patients are divided into two latent classes: a high risk group (Class 1) and a low risk group (Class 2). High risk group consists of younger, more male patients. Their serum bilirubin levels are generally higher and increases relatively faster over time. Therefore this group generally has higher hazard in terms of survival outcome. This is also the group that responds well to the treatment for lowering serum bilirubin, possibly due to the already high and increasing levels. The other group includes a larger

Table 3. Results for one latent classes

Parameter	Estimates	SE	P-value
Longitudinal Part			
Intercept	−0.1195	0.0443	0.0073
Time	0.0872	0.0038	< .0001
Trt	−0.1019	0.0511	0.0469
Age	0.0504	0.0252	0.0463
Gender	0.4513	0.0785	< .0001
Var(a)	0.8972	0.0406	< .0001
Var(ε_{ij})	0.1952	0.0068	< .0001
Survival Part			
Trt	-0.2736	0.1781	0.1254
Age	0.5491	0.0958	< .0001
Gender	0.5653	0.2521	0.0256
δ	1.5653	0.1189	< .0001

number of elderly female patients. Their serum bilirubin levels are relatively low and stable over time. Treatment effect is not significant for reducing the serum bilirubin levels in this group. It is also not significant in increasing the survival probabilities in either group, which is consistent with medical findings (Gong et al., 2004). The effect of age and gender within each group also exhibit different patterns. In class 1, the age at enrollment is associated not only with the serum bilirubin level but the overall survival probabilities as well. Whereas in Class 2, age at enrollment only affect survival probabilities but not the biomarker level. Gender is a significant factor only in Class 2 where male patients have higher biomarker levels and lower survival probabilities.

We also summarize the results from fitting the one class joint model in Table 3. Same as our findings from fitting the model with two latent classes, the treatment effect is not significant for the survival model in either class. However, in this model, log(serBilir) increases over time and could be lowered significantly by the drug. As shown above, this is not true in the low risk latent class. Similarly for age and gender, without taking into consideration of the heterogeneity, we will not be able to identify their different patterns in different latent classes.

5.2 Classification

According to the posterior probabilities calculated by (7), 63% of patients are classified to the first class. To better understand how patients are divided into two classes, we plot the longitudinal trajectories and the survival probabilities for each class marginally. Figure 1 plots the smoothed mean curves of log(serBilir) for patients in the classified two classes respectively, and of all patients for the one class joint model. The red (black) curve represents the first (second) class. The green curve represents the log(serBilir) level as one class, which is in between of the two latent classes. We can see that the serum bilirubin level is higher in the first class, corresponding to a lower survival rate in Figure 2 in which the Kaplan-Meier estimates of the survival

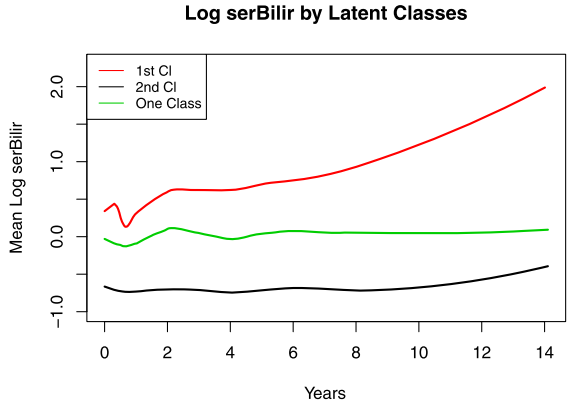

Figure 1. Mean log(serBilir) trajectories by posterior classification.

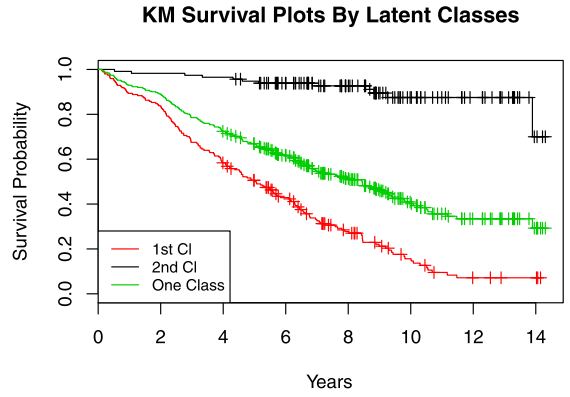

Figure 2. Predicted survival functions by posterior classification.

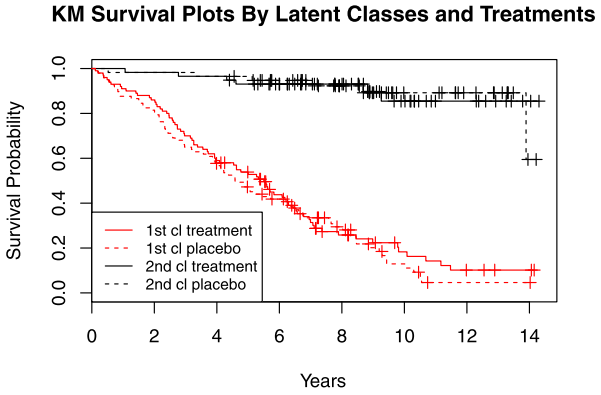

Figure 3. Predicted survival functions by treatment groups.

functions are plotted for the classes. Figure 3 plots the estimated survival functions for treatment and placebo groups in each class, showing that treatment is not significant for survival rate in the two latent classes, which is consistent with our parameter estimates. Since we are able to approx-

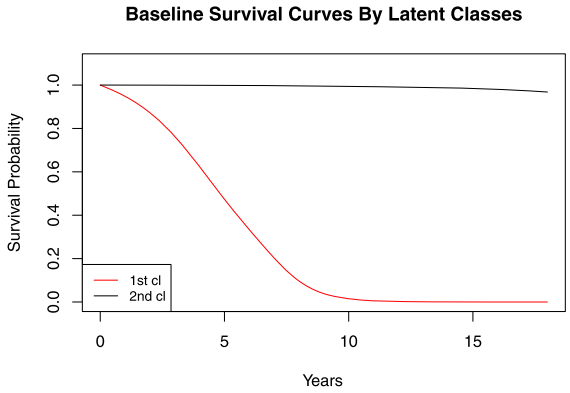

Figure 4. Fitted baseline survival functions by latent classes.

imate baseline hazard function using B-splines, we also plot the smoothed baseline survival functions without any covariates in Figure 4, which shows that they have similar pattern as the empirical Kaplan-Meier curve in Figure 2.

6. DISCUSSION

We develop a semi-parametric latent class model to analyze the longitudinal and survival outcomes jointly. The developed method utilizes data efficiently, and discovers the underlying or hidden patterns that one class model fails to identify. Our model does not make assumption on the form of the baseline hazard functions, avoiding potential model mis-specification of parametric models. Furthermore, the covariate effects are allowed to vary across classes. Posterior classification can be used to study different longitudinal and survival patterns, providing a good prognostic tool. The simulation results suggest that the estimation method performs well with finite sample size, and outperforms the parametric method when the baseline hazard function is mis-specified. PBC data is analyzed as a real data example to illustrate the strength of our method in modeling and identifying heterogeneity in joint longitudinal and survival data.

One challenge in the implementation of the proposed method is the computation feasibility, in particular with more random effects terms. We adopt Gaussian quadrature for a balance of accuracy of approximating the integral with respect to random effects and the computational burden. Other computationally simpler approach, e.g., Laplace approximation (Breslow and Clayton 1993), may provide a possible approach for more complicated models though at the cost of lower accuracy.

APPENDIX A

A.1 Proof of theorems

Proof of Theorem 1: Let $a_n = k_n/\sqrt{n}$, $b_n = 1/\sqrt{n}$. It suffices to show that, for any $\epsilon > 0$, there exist constants C_1 and C_2, such that, for large n, we have

$$P\left\{\sup_{\|\mathbf{u}\|=C_1,\|\mathbf{v}\|=C_2} \tilde{L}_n(\beta+b_n\mathbf{u},\alpha_n+a_n\mathbf{v}) < \tilde{L}_n(\beta,\alpha_n)\right\} \geq 1-\epsilon\,. \tag{10}$$

Denoting the derivatives of the approximated likelihood function as

$$\nabla^T\tilde{L}_n(\beta,\alpha_n) = \left(\nabla_1^T\tilde{L}_n(\beta,\alpha_n), \nabla_2^T\tilde{L}_n(\beta,\alpha_n)\right)^T, \tag{11}$$

$$\nabla^2\tilde{L}_n(\beta,\alpha_n) = \begin{pmatrix} \nabla_1^2\tilde{L}_n(\beta,\alpha_n) & \nabla_{12}^2\tilde{L}_n(\beta,\alpha_n) \\ \nabla_{21}^2\tilde{L}_n(\beta,\alpha_n) & \nabla_2^2\tilde{L}_n(\beta,\alpha_n)\end{pmatrix}, \tag{12}$$

where $\nabla_1\tilde{L}_n(\beta,\alpha_n)$ and $\nabla_1^2\tilde{L}_n(\beta,\alpha_n)$ are the partial derivatives for β, $\nabla_{12}^2\tilde{L}_n(\beta,\alpha_n)$ and $\nabla_{21}^2\tilde{L}_n(\beta,\alpha_n)$ are for β and α_n, and $\nabla_2\tilde{L}_n(\beta,\alpha_n)$ and $\nabla_2^2\tilde{L}_n(\beta,\alpha_n)$ are for α_n, we have

$$\begin{aligned}
&\tilde{L}_n(\beta+b_n\mathbf{u},\alpha_n+a_n\mathbf{v}) - \tilde{L}_n(\beta,\alpha_n)\\
=\;& b_n\nabla_1^T\tilde{L}_n(\beta,\alpha_n)\mathbf{u} + a_n\nabla_2^T\tilde{L}_n(\beta,\alpha_n)\mathbf{v}\\
&+\frac{1}{2}b_n^2\mathbf{u}^T\nabla_1^2\tilde{L}_n(\beta^*,\alpha_n^*)\mathbf{u} + \frac{1}{2}a_nb_n\mathbf{v}^T\nabla_{21}^2\tilde{L}_n(\beta^*,\alpha_n^*)\mathbf{u}\\
&+\frac{1}{2}a_nb_n\mathbf{u}^T\nabla_{12}^2\tilde{L}_n(\beta^*,\alpha_n^*)\mathbf{v} + \frac{1}{2}a_n^2\mathbf{v}^T\nabla_2^2\tilde{L}_n(\beta^*,\alpha_n^*)\mathbf{v}\\
=\;& I_{u1}+I_{v1}+I_{u2}+I_{vu}+I_{uv}+I_{v2}\,,
\end{aligned}$$

where the point (β^*,α_n^*) lies between $(\beta+b_n\mathbf{u},\alpha_n+a_n\mathbf{v})$ and (β,α_n).

Given A2 and $b_n = 1/\sqrt{n}$, we have

$$\begin{aligned}
|I_{u1}| &\leq b_n\|\nabla_1\tilde{L}_n(\beta,\alpha_n)\|_2\|\mathbf{u}\|_2 = b_nO_p(\sqrt{n})\|\mathbf{u}\|_2\\
&= nb_n^2O_p(1)\|\mathbf{u}\|_2\,.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
|I_{v1}| &= |a_n\nabla_2^T\tilde{L}_n(\beta,\alpha_n)\mathbf{v}| \leq a_n\|\nabla_2\tilde{L}_n(\beta,\alpha_n)\|_2\|\mathbf{v}\|_2\\
&= a_nO_p(\sqrt{n})\|\mathbf{v}\|_2\,.
\end{aligned}$$

For I_{u2}, we have that, with probability tending to 1,

$$\begin{aligned}
I_{u2} &= \frac{1}{2}nb_n^2\mathbf{u}^T\{\frac{1}{n}\nabla_1^2\tilde{L}_n(\beta^*,\alpha_n^*) + I_1(\beta^*,\alpha_n^*)\}\mathbf{u}\\
&\quad -\frac{1}{2}nb_n^2\mathbf{u}^TI_1(\beta^*,\alpha_n^*)\mathbf{u}
\end{aligned}$$

is dominated by $-\frac{1}{2}nb_n^2\mathbf{u}^TI_1(\beta^*,\alpha_n^*)\mathbf{u}$ since $\|\frac{1}{n}\nabla_1^2\tilde{L}_n(\beta^*,\alpha_n^*) + I_1(\beta^*,\alpha_n^*)\|_2 = o_p(1)$. Given A_3 and a large C_1, I_{u1} is dominated by I_{u2}.

For I_{v2}, we have

$$\begin{aligned}
I_{v2} &= \frac{1}{2}na_n^2\mathbf{v}^T\{\frac{1}{n}\nabla_2^2\tilde{L}_n(\beta^*,\alpha_n^*) + I_2(\beta^*,\alpha_n^*)\}\mathbf{v}\\
&\quad -\frac{1}{2}na_n^2\mathbf{v}^TI_2(\beta^*,\alpha_n^*)\mathbf{v}
\end{aligned}$$

Given A_3, we have $\|I_2(\beta^*,\alpha_n^*)\|_2 = O_p(k_n^{-1/2})$. For $\frac{1}{n}\nabla_2^2\tilde{L}_n(\beta^*,\alpha_n^*) + I_2(\beta^*,\alpha_n^*)$, we have that, for any $\epsilon>0$, by Chebyshev's inequality and Condition A_3,

$$\begin{aligned}
& P\left(\left\|\frac{1}{n}\nabla_2^2\tilde{L}_n(\beta^*,\alpha_n^*) + I_2(\beta^*,\alpha_n^*)\right\|_2 \geq \frac{\varepsilon}{k_n^{1/2}}\right) \\
\leq\ & \frac{k_n}{n^2\varepsilon^2}E\sum_{j,k=1}^{p_{n_2}}\left\{\frac{\partial^2\tilde{L}_n(\beta^*,\alpha_n^*)}{\partial\alpha_{nj}\partial\alpha_{nk}} - E\frac{\partial^2\tilde{L}_n(\beta^*,\alpha_n^*)}{\partial\alpha_{nj}\partial\alpha_{nk}}\right\}^2 \\
\leq\ & \frac{k_n n}{n^2\varepsilon^2}\sum_{j,k=1}^{p_{n_2}} var\left(\frac{\partial^2\tilde{f}_n(Y_{n_i},\theta_n)}{\partial\alpha_{nj}\partial\alpha_{nk}}\right) \\
=\ & O_p(\frac{k_n^2}{n\varepsilon^2}) \to 0.
\end{aligned}$$

Therefore, $\|\frac{1}{n}\nabla_2^2\tilde{L}_n(\beta^*,\alpha_n^*) + I_2(\beta^*,\alpha_n^*)\|_2 = o_p(k_n^{-1/2})$, I_{v2} is dominated by $-\frac{1}{2}na_n^2\mathbf{v}^T I_2(\beta^*,\alpha_n^*)\mathbf{v}$, and and I_{v1} is dominated by I_{v2} for a large C_2. Similarly, we have $\|\frac{1}{n}\nabla_{12}^2\tilde{L}_n(\beta^*,\alpha_n^*) + I_{12}(\beta^*,\alpha_n^*)\|_2 = o_p(k_n^{-1})$ and $\|\frac{1}{n}\nabla_{21}^2\tilde{L}_n(\beta^*,\alpha_n^*) + I_{21}(\beta^*,\alpha_n^*)\|_2 = o_p(k_n^{-1})$, and $a_nb_n\mathbf{u}^T[\frac{1}{n}\nabla_{12}^2\tilde{L}_n(\beta^*,\alpha_n^*) + I_{12}(\beta^*,\alpha_n^*)]\mathbf{v}$ is and $a_nb_n\mathbf{v}^T[\frac{1}{n}\nabla_{21}^2\tilde{L}_n(\beta^*,\alpha_n^*) + I_{21}(\beta^*,\alpha_n^*)]\mathbf{u}$ are dominated by $b_n^2\mathbf{u}^T I_1(\beta^*,\alpha_n^*)\mathbf{u} + a_n^2\mathbf{v}^T I_2(\beta^*,\alpha_n^*)\mathbf{v}$ for large C_1 and C_2.

Therefore, $L_n(\beta+b_n\mathbf{v},\alpha_n+a_n\mathbf{u}) - \tilde{L}_n(\beta,\alpha_n)$ is dominated by

$$\begin{aligned}
& -\frac{1}{2}n\{b_n^2\mathbf{u}^T I_1(\beta^*,\alpha_n^*)\mathbf{u} + a_n^2\mathbf{v}^T I_2(\beta^*,\alpha_n^*)\mathbf{v} \\
& + a_nb_n\mathbf{v}^T I_{21}(\beta^*,\alpha_n^*)\mathbf{u} + a_nb_n\mathbf{u}^T I_{12}(\beta^*,\alpha_n^*)\mathbf{v}\} \\
& = -\frac{1}{2}n\begin{pmatrix} b_n\mathbf{u} \\ a_n\mathbf{v}\end{pmatrix}^T I_n(\beta^*,\alpha_n^*)\begin{pmatrix} b_n\mathbf{u} \\ a_n\mathbf{v}\end{pmatrix} < 0
\end{aligned}$$

which implies that, as $n\to\infty$, we have, with probability tending to 1,

$$\tilde{L}_n(\beta+b_n\mathbf{u},\alpha_n+a_n\mathbf{v}) - \tilde{L}_n(\alpha_n,\beta) < 0\,.$$

Theorem 1 is proved.

Proof of Theorem 2: According to Lemma 1, we have

$$\begin{aligned}
& \sqrt{n}I_1^{-1/2}(\beta,\alpha_n)I_1(\beta,\alpha_n)(\hat{\beta}-\beta) \\
=\ & \frac{1}{\sqrt{n}}I_1^{-1/2}(\beta,\alpha_n)\nabla_1\tilde{L}_n(\beta,\alpha_n) + o_p\left(I_1^{-1/2}(\beta,\alpha_n)\right).
\end{aligned}$$

Assuming $\phi_{n_i} = \frac{1}{\sqrt{n}}I_1^{-1/2}(\beta,\alpha_n)\nabla_1\tilde{L}_{n_i}(\beta,\alpha_n)$, for $\varepsilon>0$, we have

$$\begin{aligned}
& \sum_{i=1}^n E\left(\|\phi_{n_i}\|_2^2 1\{\|\phi_{n_i}\|_2>\varepsilon\}\right) \\
& \leq n\left\{E\|\phi_{n_1}\|_2^4\right\}^{1/2}\left\{P(\|\phi_{n_1}\|_2>\varepsilon)\right\}^{1/2}.
\end{aligned}$$

Given $A3$, we have

$$\begin{aligned}
P(\|\phi_{n_i}\|_2\geq\varepsilon) &\leq \frac{E\|I_1^{-1/2}(\beta,\alpha_n)\nabla_1\tilde{L}_{n_i}(\beta,\alpha_n)\|_2^2}{n\varepsilon^2} \\
&= O_p(1/n)\,,
\end{aligned}$$

and

$$\begin{aligned}
E\|\phi_{n_i}\|_2^4 &= \frac{1}{n^2}E\left\|I_1^{-1/2}(\beta,\alpha_n)\nabla_1\tilde{L}_{n_i}(\beta,\alpha_n)\right\|_2^4 \\
&= O_p(1/n^2)\,.
\end{aligned}$$

Therefore, we have $\sum_{i=1}^n E\left(\|\phi_{n_i}\|_2^2 1\{\|\phi_{n_i}\|_2\geq\varepsilon\}\right)\to 0$. Letting $s_n^2 = var(\sum_{i=1}^n\phi_{n_i}) = nvar(\phi_{n_1})$, we have

$$\begin{aligned}
s_n^2 &= var\left(I_1^{-1/2}(\beta,\alpha_n)\nabla_1\tilde{L}_{n_1}(\beta,\alpha_n)\right) \\
&= I_1^{-1/2}(\beta,\alpha_n)\,var\left(\nabla_1\tilde{L}_{n_1}(\beta,\alpha_n)\right)I_1^{-1/2}(\beta,\alpha_n) \\
&\to I\,.
\end{aligned}$$

According to the Lindeberg-Feller central limit theorem, $\frac{1}{\sqrt{n}}I_1^{-1/2}(\beta,\alpha_n)\nabla_1\tilde{L}_n(\hat{\beta},\alpha_n)$ has an asymptotic normal distribution $N(0,I)$, implying $\sqrt{n}I_1^{1/2}(\beta,\alpha_n)(\hat{\beta}-\beta)\to_d N(0,I)$. Theorem 2 is proved.

Proof of Theorem 3: According to Lemma 2, we have

$$\begin{aligned}
& \sqrt{n}A_nI_2^{-1/2}(\beta,\alpha_n)I_2(\beta,\alpha_n)(\hat{\alpha}_n-\alpha_n) \\
=\ & \frac{1}{\sqrt{n}}A_nI_2^{-1/2}(\beta,\alpha_n)\nabla_2\tilde{L}_n(\beta,\alpha_n) \\
+\ & o_p\left(A_nI_2^{-1/2}(\beta,\alpha_n)\frac{1}{\sqrt{k_n}}\right).
\end{aligned}$$

Letting $\eta_{n_i} = \frac{1}{\sqrt{n}}A_nI_2^{-1/2}(\beta,\alpha_n)\nabla_2\tilde{L}_{n_i}(\beta,\alpha_n)$, we have

$$\begin{aligned}
& \sum_{i=1}^n E\left(\|\eta_{n_i}\|_2^2 1\{\|\eta_{n_i}\|_2>\varepsilon\}\right) \\
& \leq n\left\{E\|\eta_{n_1}\|_2^4\right\}^{1/2}\left\{P(\|\eta_{n_1}\|_2>\varepsilon)\right\}^{1/2}.
\end{aligned}$$

Given $A3$, we have

$$\begin{aligned}
P(\|\eta_{n_1}\|_2\geq\varepsilon) &\leq \frac{E\left\|A_nI_2^{-1/2}(\beta,\alpha_n)\nabla_2\tilde{L}_{n_i}(\beta,\alpha_n)\right\|_2^2}{n\varepsilon^2} \\
&= O_p(1/n)
\end{aligned}$$

and

$$\begin{aligned}
& E\|\eta_{n_1}\|_2^4 = \frac{1}{n^2}E\|A_nI_2^{-1/2}(\beta,\alpha_n)\nabla_2\tilde{L}_{n_i}(\beta,\alpha_n)\|_2^4 \\
=\ & \frac{1}{n^2}E\Big[\nabla_2^T\tilde{L}_{n_i}(\beta,\alpha_n)I_2^{-1/2}(\beta,\alpha_n)A_n^TA_nI_2^{-1/2}(\beta,\alpha_n) \\
& \qquad \times\nabla_2\tilde{L}_{n_i}(\beta,\alpha_n)\Big]^2 \\
=\ & O_p(k_n^2/n^2).
\end{aligned}$$

Therefore, we have $\sum_{i=1}^{n} E\left(\|\eta_{n_i}\|_2^2 1\{\|\eta_{n_i}\|_2 \geq \varepsilon\}\right) \to 0$. Letting $\tilde{s}_n^2 = var(\sum_{i=1}^{n}\eta_{n_i}) = nvar(\eta_{n_1})$, we have

$$\tilde{s}_n^2 = A_n I_2^{-1/2}(\beta,\alpha_n) var\left(\nabla_2 \tilde{L}_{n_1}(\beta,\alpha_n)\right) I_2^{-1/2}(\beta,\alpha_n) A_n^T \to G\,.$$

According to the Lindeberg-Feller central limit theorem, $\frac{1}{\sqrt{n}} A_n I_2^{-1/2}(\beta,\alpha_n)\nabla_2 \tilde{L}_n(\beta,\alpha_n)$ has an asymptotic normal distribution $N(0,G)$, implying $\sqrt{n}A_n I_2^{1/2}(\beta,\alpha_n)(\hat{\alpha}_n - \alpha_n) \to_d N(0,G)$. Theorem 3 is proved.

Proof of Theorem 4: Given a fixed k, for $h_{0,k}(t)$, there exist $\alpha_{n,k}$, such that

$$\sup_{t\in[0,T]} |B_n^T(t)\alpha_{n,k} - h_{0,k}(t)| \leq O_p(k_n^{-r})\,,$$

for $r \geq 3$ in Condition A1. Given $n/k_n^{1+2r} \to 0$, we have

$$\begin{aligned} & \sqrt{n/k_n}(B_n^T(t)\hat{\alpha}_{n,k} - h_{0,k}(t)) \\ = \;& \sqrt{n/k_n}[B_n^T(t)\hat{\alpha}_{n,k} - B_n^T(t)\alpha_{n,k} + B_n^T(t)\alpha_{n,k} - h_{0,k}(t)] \\ = \;& \sqrt{n/k_n}[B_n^T(t)\hat{\alpha}_{n,k} - B_n^T(t)\alpha_{n,k}] + o_p(1)\,. \end{aligned}$$

From Theorem 3, we have $\sqrt{n}A_n I_2^{1/2}(\beta,\alpha_n)(\hat{\alpha}_n - \alpha_n) \to_d N(0,G)$. Consequently, we have $\sqrt{n}A_{n,k} I_{2,k}^{1/2}(\beta,\alpha_n)(\hat{\alpha}_{n,k} - \alpha_{n,k}) \to_d N(0,G_k)$, where $A_{n,k}$, $I_{2,k}$, and G_k are the corresponding sub-matrices related to the sub-vector $\alpha_{n,k}$. Theorem 4 is proved by letting $A_{n,k} = \frac{1}{\sqrt{k_n}} B_n^T(t) I_{2,k}^{-1/2}(\beta,\alpha_n)$ and $\sigma_k^2(t) = \lim_{n\to\infty} A_{n,k}A_{n,k}^T = \lim_{n\to\infty} \frac{1}{k_n} B_n^T(t) I_{2,k}^{-1}(\beta,\alpha_n) B_n(t)$.

A.2 Lemmas

Lemma 1: Under the conditions in Theorem 2, we have that

$$I_1(\beta,\alpha_n)(\hat{\beta}-\beta) = \frac{1}{n}\nabla_1 \tilde{L}_n(\beta,\alpha_n) + o_p(1/\sqrt{n})\,.$$

Proof of Lemma 1: Based on the Taylor expansion of $\nabla_1 \tilde{L}_n(\beta,\hat{\alpha}_n)$ at $\hat{\beta}$, we have

$$-\frac{1}{n}\nabla_1 \tilde{L}_n(\beta,\hat{\alpha}_n) = \frac{1}{n}\nabla_1^2 \tilde{L}_n(\beta^*,\hat{\alpha}_n)(\hat{\beta}-\beta)\,,$$

where β^* lies between $\hat{\beta}$ and β. For any $\varepsilon > 0$, by Chebyshev's inequality and Condition A4, we have

$$\begin{aligned} & P\left(\left\|\frac{1}{n}\nabla_1^2 \tilde{L}_n(\beta^*,\hat{\alpha}_n) + I_1(\beta,\alpha_n)\right\|_2 \geq \varepsilon\right) \\ \leq\; & \frac{1}{n^2\varepsilon^2} E \sum_{i,j=1}^{p_1} \left\{ \frac{\partial^2 \tilde{L}_n(\beta^*,\hat{\alpha}_n)}{\partial\beta_i\partial\beta_j} - E\frac{\partial^2 \tilde{L}_n(\beta^*,\hat{\alpha}_n)}{\partial\beta_i\partial\beta_j} \right. \\ & \left. + E\frac{\partial^2 \tilde{L}_n(\beta^*,\hat{\alpha}_n)}{\partial\beta_i\partial\beta_j} - E\frac{\partial^2 \tilde{L}_n(\beta,\alpha_n)}{\partial\beta_i\partial\beta_j} \right\}^2 \\ \leq\; & 2\frac{1}{n^2\varepsilon^2} E \sum_{i,j=1}^{p_1} \left\{ \left[\frac{\partial^2 \tilde{L}_n(\beta^*,\hat{\alpha}_n)}{\partial\beta_i\partial\beta_j} - E\frac{\partial^2 \tilde{L}_n(\beta^*,\hat{\alpha}_n)}{\partial\beta_i\partial\beta_j} \right]^2 \right. \\ & \left. + \left[E\frac{\partial^2 \tilde{L}_n(\beta^*,\hat{\alpha}_n)}{\partial\beta_i\partial\beta_j} - E\frac{\partial^2 \tilde{L}_n(\beta,\alpha_n)}{\partial\beta_i\partial\beta_j} \right]^2 \right\} \\ \leq\; & 2\frac{1}{n^2\varepsilon^2} \left\{ n \sum_{i,j=1}^{p_1} var\left(\frac{\partial^2 \tilde{L}_{n_1}(\beta^*,\hat{\alpha}_n)}{\partial\beta_i\partial\beta_j} \right) \right. \\ & \left. + \sum_{i,j=1}^{p_1} n^2 E \left[\frac{\partial^2 \tilde{L}_{n_1}(\beta^*,\hat{\alpha}_n)}{\partial\beta_i\partial\beta_j} - \frac{\partial^2 \tilde{L}_{n_1}(\beta,\alpha_n)}{\partial\beta_i\partial\beta_j} \right]^2 \right\} \\ =\; & O_p(\frac{1}{n\varepsilon^2}) + \frac{2}{n^2\varepsilon^2} \sum_{i,j=1}^{p_1} n^2 E \left[\nabla_1^T \frac{\partial^2 \tilde{L}_{n_1}(\beta^{**},\alpha_n^{**})}{\partial\beta_i\partial\beta_j}(\beta^{**}-\beta) \right. \\ & \left. + \nabla_2^T \frac{\partial^2 \tilde{L}_{n_1}(\beta^{**},\alpha_n^{**})}{\partial\beta_i\partial\beta_j}(\alpha_n^{**}-\alpha_n) \right]^2 \\ \leq\; & O_p(\frac{1}{n\varepsilon^2}) + \frac{2}{n^2\varepsilon^2} \sum_{i,j=1}^{p_1} n^2 E \left\| \nabla \frac{\partial^2 \tilde{L}_{n_1}(\beta^{**},\alpha_n^{**})}{\partial\beta_i\partial\beta_j} \right\|_2^2 \\ & \times \|\theta_n^{**} - \theta_n\|_2^2 \\ =\; & O_p(\frac{k_n^2}{n\varepsilon^2}) \to 0\,, \end{aligned}$$

where $(\beta^{**},\alpha_n^{**})$ lies between $(\beta^*,\hat{\alpha}_n)$ and (β,α_n), indicating $\left\|\frac{1}{n}\nabla_1^2 \tilde{L}_n(\beta^*,\hat{\alpha}_n) + I_1(\beta,\alpha_n)\right\|_2 = o_p(1)$. Given $\hat{\beta}-\beta = O_p(1/\sqrt{n})$ from Theorem 1, we have $-\frac{1}{n}\nabla_1 \tilde{L}_n(\beta,\hat{\alpha}_n) = \frac{1}{n}\nabla_1^2 \tilde{L}_n(\beta^*,\hat{\alpha}_n)(\hat{\beta}-\beta) = -I_1(\beta,\alpha_n)(\hat{\beta}-\beta) + o_p(1/\sqrt{n})$. Since $\left\|\frac{1}{n}\nabla_1 \tilde{L}_n(\beta,\hat{\alpha}_n) - \frac{1}{n}\nabla_1 \tilde{L}_n(\beta,\alpha_n)\right\|_2 = o_p(1/\sqrt{n})$ due to Condition A3, we have $-\frac{1}{n}\nabla_1 \tilde{L}_n(\beta,\alpha_n) = -I_1(\beta,\alpha_n)(\hat{\beta}-\beta) + o_p(1/\sqrt{n})$. Lemma 1 is proved.

Lemma 2: Under the conditions in Theorem 3, we have

$$I_2(\beta,\alpha_n)(\hat{\alpha}_n-\alpha_n) = \frac{1}{n}\nabla_2 \tilde{L}_n(\beta,\alpha_n) + o_p(\frac{1}{\sqrt{nk_n}})\,.$$

Proof of Lemma 2: Based on the Taylor expansion of $\nabla_2 \tilde{L}_n(\hat{\beta},\alpha_n)$ at $\hat{\alpha}_n$, we have

$$-\frac{1}{n}\nabla_2 \tilde{L}_n(\hat{\beta},\alpha_n) = \frac{1}{n}\nabla_2^2 \tilde{L}_n(\hat{\beta},\alpha_n^*)(\hat{\alpha}_n-\alpha_n)\,, \tag{13}$$

where α_n^* lies between α_n and $\hat{\alpha}_n$. For any $\varepsilon < 0$, by Chebyshev's inequality, we have

$$\begin{aligned} & P\left(\left\|\frac{1}{n}\nabla_2^2 \tilde{L}_n(\hat{\beta},\alpha_n^*) + I_2(\beta,\alpha_n)\right\|_2 \geq \frac{\varepsilon}{\sqrt{k_n^3}}\right) \\ \leq\; & \frac{k_n^3}{n^2\varepsilon^2} E \sum_{i,j=1}^{p_{n_2}} \left\{ \frac{\partial^2 \tilde{L}_n(\hat{\beta},\alpha_n^*)}{\partial\alpha_{n_i}\partial\alpha_{n_j}} - E\frac{\partial^2 \tilde{L}_n(\hat{\beta},\alpha_n^*)}{\partial\alpha_{n_i}\partial\alpha_{n_j}} \right. \end{aligned}$$

$$+ E\frac{\partial^2 \tilde{L}_n(\hat{\beta},\alpha_n^*)}{\partial\alpha_{n_i}\partial\alpha_{n_j}} - E\frac{\partial^2 \tilde{L}_n(\beta,\alpha_n)}{\partial\alpha_{n_i}\partial\alpha_{n_j}}\Bigg\}^2$$

$$\le 2\frac{k_n^3}{n^2\varepsilon^2} E\sum_{i,j=1}^{p_{n_2}}\left\{\left[\frac{\partial^2 \tilde{L}_n(\hat{\beta},\alpha_n^*)}{\partial\alpha_{n_i}\partial\alpha_{n_j}} - E\frac{\partial^2 \tilde{L}_n(\hat{\beta},\alpha_n^*)}{\partial\alpha_{n_i}\partial\alpha_{n_j}}\right]^2\right.$$

$$\left. + \left[E\frac{\partial^2 \tilde{L}_n(\hat{\beta},\alpha_n^*)}{\partial\alpha_{n_i}\partial\alpha_{n_j}} - E\frac{\partial^2 \tilde{L}_n(\beta,\alpha_n)}{\partial\alpha_{n_i}\partial\alpha_{n_j}}\right]^2\right\}$$

$$\le 2\frac{k_n^3}{n^2\varepsilon^2}\left\{n\sum_{i,j=1}^{p_{n_2}} var\left(\frac{\partial^2 \tilde{L}_{n_1}(\hat{\beta},\alpha_n^*)}{\partial\alpha_{n_i}\partial\alpha_{n_j}}\right)\right.$$

$$\left. + \sum_{i,j=1}^{p_{n_2}} n^2 E\left[\frac{\partial^2 \tilde{L}_{n_1}(\hat{\beta},\alpha_n^*)}{\partial\alpha_{n_i}\partial\alpha_{n_j}} - \frac{\partial^2 \tilde{L}_{n_1}(\beta,\alpha_n)}{\partial\alpha_{n_i}\partial\alpha_{n_j}}\right]^2\right\}$$

$$= O_p(\frac{k_n^4}{n\varepsilon^2}) + \frac{2k_n^3}{n^2\varepsilon^2}\sum_{i,j=1}^{p_{n_2}} n^2 E\left[\nabla_1^T\frac{\partial^2 \tilde{L}_{n_1}(\beta^{**},\alpha_n^{**})}{\partial\alpha_{n_i}\partial\alpha_{n_j}}(\beta^{**}-\beta)\right.$$

$$\left. + \nabla_2^T\frac{\partial^2 \tilde{L}_{n_1}(\beta^{**},\alpha_n^{**})}{\partial\alpha_{n_i}\partial\alpha_{n_j}}(\alpha_n^{**}-\alpha_n)\right]^2$$

$$\le O_p(\frac{k_n^4}{n\varepsilon^2}) + \frac{2k_n^3}{n^2\varepsilon^2}\sum_{i,j=1}^{p_{n_2}} n^2 E\left\|\nabla\frac{\partial^2 \tilde{L}_{n_1}(\beta^{**},\alpha_n^{**})}{\partial\alpha_{n_i}\partial\alpha_{n_j}}\right\|_2^2$$

$$\times \|\theta_n^{**}-\theta_n\|_2^2$$

$$= O_p(\frac{k_n^4}{n\varepsilon^2}) + \frac{2k_n^3}{n^2\varepsilon^2}k_n^2 n^2 O_p(k_n^{-1}k_n\frac{k_n^2}{n})$$

$$= O_p(\frac{k_n^7}{n\varepsilon^2}) \to 0\,,$$

where $(\beta^{**},\alpha_n^{**})$ lies between $(\hat{\beta},\alpha_n^*)$ and (β,α_n), indicating $\|\frac{1}{n}\nabla_2^2\tilde{L}_n(\hat{\beta},\alpha_n^*) + I_{n2}(\beta,\alpha_n)\|_2 = o_p(\frac{1}{\sqrt{k_n^3}})$. Given $\|\hat{\alpha}_n - \alpha_n\|_2 = O_p(k_n/\sqrt{n})$ from Theorem 1, we have $-\frac{1}{n}\nabla_2\tilde{L}_n(\hat{\beta},\alpha_n) = \frac{1}{n}\nabla_2^2\tilde{L}_n(\hat{\beta},\alpha_n^*)(\hat{\alpha}_n - \alpha_n) = -I_{n2}(\beta,\alpha_n)(\hat{\alpha}_n - \alpha_n) + o_p(1/\sqrt{nk_n})$. Since $\left\|\frac{1}{n}\nabla_2\tilde{L}_n(\hat{\beta},\alpha_n) - \frac{1}{n}\nabla_2\tilde{L}_n(\beta,\alpha_n)\right\|_2 = o_p(1/\sqrt{nk_n})$ due to Condition A3, we have $-\frac{1}{n}\nabla_2\tilde{L}_n(\beta,\alpha_n) = -I_{n2}(\beta,\alpha_n)(\hat{\alpha}_n - \alpha_n) + o_p(1/\sqrt{nk_n})$. Lemma 2 is proved.

ACKNOWLEDGEMENTS

This research is partly supported by the NIAAA grant RC1 AA019274 and AHRQ grant R01 HS020263.

Received 5 July 2019

REFERENCES

Breslow, N. E. and Clayton, D. G. (1993), 'Approximate inference in generalized linear mixed models', *Journal of the American Statistical Association*, **88**, 9–25. MR1394064

Dafni, U. G. and Tsiatis, A. A. (1998), 'Evaluating surrogate markers of clinical outcome when measured with error', *Biometrics*, **54**, 1445–1462.

Fan, J. and Peng, H. (2004), 'Nonconcave penalized likelihood with a diverging number of parameters', *Annals of Statistics*, **32**, 928–961. MR2065194

Gong, Y., Frederiksen, S. L., and Gluud, C. (2004), 'D-penicillamine for primary biliary cirrhosis', *The Cochrane Database of Systematic Reviews*, CD004789.

Lin, H., Turnbull, B. W., McCulloch, C. E., and Slate, E. H. (2002), 'Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer', *Journal of the American Statistical Association*, **97**, 53–65. MR1947272

Lindor, K. D., Dickson, E. R., Baldus, W. P., Jorgensen, R. A., Ludwig, J., Murtaugh, P. A., Harrison, J. M., Wiesner, R. H., Anderson, M. L., Lange, S. M., LeSage, G., Rossi, S. S., and Hofman, A. F. (1994), 'Ursodeoxycholic acid in the treatment of primary biliary cirrhosis', *Gastroenterology*, **106**, 1284–1290.

Liu, Y., Liu, L., and Zhou, J. (2015), 'Joint latent class model of survival and longitudinal data: An application to CPCRA study', *Computational Statistics and Data Analysis*, **91**, 40–50. MR3368004

Magidson, J. and Vermunt, J. K. (2004), 'Latent class models. In: Kaplan, D., editor. The Sage handbook of quantitative methodology for the social sciences', *Sage Publications, Inc.*, 175–98.

Muthen, B. and Shedden, K. (1999), 'Finite mixture modeling with mixture outcomes using the EM algorithm', *Biometrics*, **55**, 463–469.

Nagin, D. S. (1999), 'Analyzing developmental trajectories: A semi-parametric, group-based approach', *Psychological Methods*, **4**, 139–157.

Nagin, D. S. and Tremblay, R. E. (2001), 'Analyzing developmental trajectories of distinct but related behaviors: A group-based method', *Psychological Methods*, **6**, 18–34.

Prentice, R. L. (1982), 'Covariate measurement errors and parameter estimation in a failure time regression model', *Biometrika*, **69**, 331–342. MR0671971

Proust-Lima, C., Joly, P., Dartigues, J.-F., and Jacqmin-Gadda, H. (2009), 'Joint modeling of multivariate longitudinal outcomes and a time-to-event: A nonlinear latent class approach', *Computational Statistics and Data Analysis*, **53**, 1142–1154. MR2657078

Shapiro, J. M., Smith, H., and Schaffner, F. (1979), 'Serum bilirubin: a prognostic factor in primary biliary cirrhosis', *Gut*, **20**, 137–140.

Wulfsohn, M. S. and Tsiatis, A. A. (1997), 'A joint model for survival and longitudinal data measured with error', *Biometrics*, **53**, 330–339. MR1450186

Zeng, D. and Cai, J. (2005), 'Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time', *The Annals of Statistics*, **33**, 2132–2163. MR2211082

Yue Liu
Takeda Pharmaceutical Company Limited
Cambridge, MA
USA
E-mail address: yl7z@virginia.edu

Ye Lin
Department of Statistics
University of Virginia
Charlottesville, VA
USA
E-mail address: yl5kn@virginia.edu

Jianhui Zhou
Department of Statistics
University of Virginia
Charlottesville, VA
USA
E-mail address: jz9p@virginia.edu

Lei Liu
Division of Biostatistics
School of Medicine
Washington University in St. Louis
St. Louis, MO
USA
E-mail address: lei.liu@wustl.edu