


Regularized Latent Class Model for Joint Analysis of High-Dimensional Longitudinal Biomarkers and a Time-to-Event Outcome

Jiehuan Sun¹ ,¹ Jose D. Herazo-Maya,² Philip L. Molyneaux,^{3,4} Toby M. Maher,^{3,4} Naftali Kaminski,² and Hongyu Zhao^{1,*}

¹Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, U.S.A.

²Internal Medicine: Pulmonary, Critical Care & Sleep Medicine, Yale School of Medicine, New Haven, Connecticut, U.S.A.

³Fibrosis Research Group, National Heart and Lung Institute, Imperial College, London

⁴Royal Brompton Hospital, Interstitial Lung Disease Unit, London

**email*: hongyu.zhao@yale.edu

SUMMARY. Although many modeling approaches have been developed to jointly analyze longitudinal biomarkers and a time-to-event outcome, most of these methods can only handle one or a few biomarkers. In this article, we propose a novel joint latent class model to deal with high dimensional longitudinal biomarkers. Our model has three components: a class membership model, a survival submodel, and a longitudinal submodel. In our model, we assume that covariates can potentially affect biomarkers and class membership. We adopt a penalized likelihood approach to infer which covariates have random effects and/or fixed effects on biomarkers, and which covariates are informative for the latent classes. Through extensive simulation studies, we show that our proposed method has improved performance in prediction and assigning subjects to the correct classes over other joint modeling methods and that bootstrap can be used to do inference for our model. We then apply our method to a dataset of patients with idiopathic pulmonary fibrosis, for whom gene expression profiles were measured longitudinally. We are able to identify four interesting latent classes with one class being at much higher risk of death compared to the other classes. We also find that each of the latent classes has unique trajectories in some genes, yielding novel biological insights.

KEY WORDS: Fused lasso; Group lasso; High-dimensional longitudinal biomarkers; Joint latent class model; Regularization; Survival outcome.

1. Introduction

In longitudinal studies, it is often of great interest to study the association between longitudinal biomarkers and a time-to-event outcome and/or to utilize the longitudinal biomarker to make prediction of the events. Many methods have been proposed to address these questions, among which the joint modeling methods have been extensively studied in the last two decades. Comprehensive overview of some early work and recent advances on joint models were provided in Tsiatis and Davidian (2004) and Proust-Lima et al. (2014), respectively. In this article, our primary goal is to build a dynamic prediction tool via the joint modeling of a time-to-event outcome and high-dimensional longitudinal biomarkers.

In our motivating study, a cohort of patients with idiopathic pulmonary fibrosis (IPF) were followed longitudinally (Molyneaux et al., 2017). IPF is an incurable lung disease and the median survival time is about three to five years after diagnosis. At each visit, gene expression profiles in the peripheral whole blood were measured. In addition, the time-to-event outcome and important clinical variables were recorded for these patients. Here, we are interested in predicting the risk of death for IPF patients using both the repeatedly measured gene expression profiles and relevant clinical variables.

Standard approaches, other than joint models, are available for risk prediction, such as the landmarking method (e.g., Van Houwelingen, 2007). In landmarking method, a Cox model is usually utilized to model the event time and the longitudinal biomarkers are not modeled. Instead, a last value carried forward approach is adopted to deal with the intermittently measured biomarker. However, the last value carried forward approach could introduce bias in prediction and cannot appropriately account for the measurement errors on the biomarker. In contrast, Joint modeling of longitudinal biomarker and time-to-event outcome could avoid such biases (Proust-Lima et al., 2014). In addition, the dynamic prediction rule based on the joint models can take advantage of all past values of the biomarker while the landmarking method only employs the most recent value. All historical values of biomarker could be more informative than a single most recent value and hence can improve prediction (Rizopoulos, 2011).

There are two broad types of joint modeling methods: the shared random-effects model and the joint latent class model. One major difference between these two types of models is that the longitudinal process and the event process are linked through the latent class in the joint latent class model

rather than the random effects as in the shared random-effects model. Some representative publications on the joint latent class model include Lin et al. (2002), Proust-Lima and Taylor (2009), and Jacqmin-Gadda et al. (2010). Some representative publications on the shared random-effects model include Henderson et al. (2000), Wang and Taylor (2001), Xu and Zeger (2001), and Rizopoulos et al. (2014). Recently, some interesting work has been done to incorporate latent class structure into the shared random-effects model (e.g., Liu et al., 2015).

Generally, a linear mixed-effects model is adopted to model the longitudinal biomarker for both types of models, where the covariates with the random and fixed effects are pre-determined. Although this works well for only one or a few biomarkers, it is subjective and difficult to apply this approach when the number of biomarkers is large, where a data-driven procedure of selecting fixed and random effects is preferable. In this context, a linear mixed-effects model with variable selection feature via the penalized likelihood approach could yield a parsimonious model including most important predictors, with better bias-variance tradeoff and easier interpretation in practice (Chen and Dunson, 2003; Bondell et al., 2010). While the advantages of variable selection via penalization are well recognized in the linear mixed-effects model and many other statistical frameworks (e.g., Tibshirani et al., 2005; Yuan and Lin, 2006; Zou, 2006; Guo et al., 2010), the literature on this topic is rather limited in the area of joint models, possibly due to the lack of clinical studies with both high dimensional longitudinal biomarkers and time-to-event outcomes.

Recently, He et al. (2015) proposed a novel joint modeling approach under the shared random-effects model framework, which allows simultaneous variable selection of fixed and random effects in both the longitudinal submodel and survival submodel. However, this approach is computationally intensive and can only handle up to ten random effects, because the likelihood of the shared random effects model includes integrals over random effects without closed form.

In this article, we propose a regularized joint latent class model, which allows variable selection of fixed and random effects in the longitudinal submodel. The major difference between our proposed model and a standard joint latent class model lies in the longitudinal submodel. Specifically, the selection of the random effects is performed by imposing an adaptive group lasso penalty (Yuan and Lin, 2006; Zou, 2006) on the reparameterized covariance matrix of the random effects using Cholesky decomposition. An adaptive pairwise fused lasso penalty (Tibshirani et al., 2005) is imposed on each of the length K vectors of fixed effects to induce variable selection on the fixed effects. The estimation of the parameters in our proposed model is performed using an Expectation-Maximization (EM) algorithm.

The remainder of the article is organized as follows. Section 2 details our proposed model. Section 3 describes our model selection criterion and survival prediction approach. Section 4 presents the performance of our proposed method in simulation studies. Section 5 illustrates the application of our proposed model on the IPF dataset. Section 6 concludes the article with some remarks.

2. Statistical Model

Similar to a standard joint latent class model, our proposed model is comprised of three ingredients, that is the class membership probability, survival submodel, and longitudinal submodel. We describe each ingredient following the model formulation in Lin et al. (2002).

2.1. Class Membership Probability

Let $\mathbf{c}_i = (c_{i1}, \dots, c_{iK})'$ be the class membership of subject i with $c_{ik} = 1$ if subject i belongs to latent class k , for $i = 1, \dots, N$, where K is the total number of latent classes and N is the total number of subjects. Let $\mathbf{v}_i = (v_{i1}, \dots, v_{im})'$ be the vector of m covariates related to the class membership for subject i . The class membership \mathbf{c}_i is assumed to follow a multinomial distribution with probabilities relating to the covariates \mathbf{v}_i via a logit model as

$$P(c_{ik} = 1 | \mathbf{v}_i) = \frac{\exp(\mathbf{v}_i' \boldsymbol{\eta}_k)}{\sum_{j=1}^K \exp(\mathbf{v}_i' \boldsymbol{\eta}_j)}, \quad (1)$$

where $\boldsymbol{\eta}_k$ is the vector of class-specific coefficients and $\boldsymbol{\eta}_1 = \mathbf{0}$ for identifiability.

2.2. Survival Submodel

Let (T_i, Δ_i) be the survival outcome of subject i with T_i being the observed survival time and Δ_i being the censoring indicator, where $\Delta_i = 1$ indicates that subject i is censored. Let $\mathbf{u}_i = (u_{i1}, \dots, u_{is})'$ be the vector of s covariates related to the risk for subject i . Then, the risk of the event for subject i conditional on the class membership is specified by a proportional hazards model as

$$h_i(t | c_{ik} = 1, \mathbf{u}_i) = \lambda_k(t) \exp(\mathbf{u}_i' \boldsymbol{\gamma}), \quad (2)$$

where $\lambda_k(t)$ is the class-specific hazard function with corresponding cumulative hazard function $\Lambda_k(t)$ and $\boldsymbol{\gamma}$ is the vector of coefficients shared by all latent classes. It is easy to accommodate class-specific covariates effects (i.e., $\boldsymbol{\gamma}_k$) and more details on this are provided in the Discussion Section. Due to the small sample size in the IPF dataset, the class-specific hazard function $\lambda_k(t)$ is taken to be Weibull hazard function with shape and scale parameters (α_k, θ_k) in all of our analyses such that $\lambda_k(t) = \alpha_k t^{\alpha_k - 1} / \theta_k^{\alpha_k}$.

2.3. Longitudinal Submodel

Let $\mathbf{y}_{ig} = (y_{ig1}, \dots, y_{ign_i})'$ be the vector of gene expression values for gene g of subject i , for $g = 1, \dots, G$, where n_i is the number of visits for subject i and G is the total number of genes. Let $\mathbf{X}_{ig} = (\mathbf{x}_{ig1}, \dots, \mathbf{x}_{ign_i})'$, where \mathbf{x}_{igj} is the vector of p_g covariates of fixed effects for gene g of the j th measurement of subject i . Let $\mathbf{Z}_{ig} = (\mathbf{z}_{ig1}, \dots, \mathbf{z}_{ign_i})'$, where \mathbf{z}_{igj} is the vector of q_g covariates of random effects for gene g of the j th measurement of subject i .

By concatenating these variables from all G genes, we let $\mathbf{Y}_i = (\mathbf{y}_{i1}', \dots, \mathbf{y}_{iG}')'$, \mathbf{X}_i be the block diagonal matrix with the g th block being \mathbf{X}_{ig} , and \mathbf{Z}_i be the block diagonal matrix with the g th block being \mathbf{Z}_{ig} . Then, the longitudinal gene expression profiles of subject i are modeled using a linear mixed-effects

model as

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\Psi} \mathbf{c}_i + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad (3)$$

where $\mathbf{b}_i \sim \text{MVN}(\mathbf{0}, \mathbf{D})$ denotes the vector of $q = \sum_g q_g$ random effects, the covariance matrix \mathbf{D} is unstructured capturing the correlations among covariates and genes, $\boldsymbol{\epsilon}_i \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{I}_{n_i})$ denotes the vector of $G n_i$ random noises, the covariance matrix $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \dots, \sigma_G^2)$ with σ_g^2 being the variance of the noises in gene g , the symbol \otimes denotes Kronecker product, \mathbf{I}_{n_i} denotes the $n_i \times n_i$ identity matrix, and $\boldsymbol{\Psi} = (\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_K)$ with $\boldsymbol{\psi}_k$ being the vector of $p = \sum_g p_g$ class-specific fixed effects.

To facilitate the selection of the random effects, we reparametrize the covariance matrix \mathbf{D} as $\boldsymbol{\Gamma} \boldsymbol{\Gamma}'$, where $\boldsymbol{\Gamma}$ is a lower triangular matrix with positive diagonal elements. Rationales on this reparametrization will be provided in Section 2.5. Then, the linear mixed-effects model specified in equation (3) can be rewritten as

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\Psi} \mathbf{c}_i + \mathbf{Z}_i \boldsymbol{\Gamma} \mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad (4)$$

where the vector of random effects \mathbf{b}_i now follows $\text{MVN}(\mathbf{0}, \mathbf{I}_q)$. We stick to this formulation in the following discussions, if no explicit reference is provided.

2.4. Conditional Independence and Likelihood

One computationally attractive feature of the joint latent class model is the conditional independence assumption, that is the event process and longitudinal process are independent conditional on the latent class membership. Because of the conditional independence structure, the random effects \mathbf{b}_i only appear in the longitudinal submodel, which can be integrated out analytically and hence makes it relatively easier to compute the likelihood than the shared random-effects model.

Let $\boldsymbol{\Xi} = \{\{\eta_k, \theta_k, \alpha_k, \boldsymbol{\psi}_k\}_{k=1}^K, \boldsymbol{\gamma}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}\}$ denote all the unknown parameters and $\mathbf{W}_i = \{\mathbf{X}_i, \mathbf{Z}_i, \mathbf{u}_i, \mathbf{v}_i\}$ denote all covariates of subject i . Then, the probability of the observed data of subject i can be written as

$$\begin{aligned} P(\mathbf{Y}_i, T_i, \Delta_i | \mathbf{W}_i; \boldsymbol{\Xi}) &= \sum_{k=1}^K P(\mathbf{Y}_i, T_i, \Delta_i, c_{ik}=1 | \mathbf{W}_i; \boldsymbol{\Xi}), \\ &= \sum_{k=1}^K P(c_{ik}=1 | \mathbf{W}_i; \boldsymbol{\Xi}) \times P(\mathbf{Y}_i | c_{ik}=1, \mathbf{W}_i; \boldsymbol{\Xi}) \\ &\quad \times P(T_i, \Delta_i | c_{ik}=1, \mathbf{W}_i; \boldsymbol{\Xi}), \\ &= \sum_{k=1}^K \left[\frac{\exp(\mathbf{v}_i' \boldsymbol{\eta}_k)}{\sum_{j=1}^K \exp(\mathbf{v}_i' \boldsymbol{\eta}_j)} \right. \\ &\quad \times \Phi(\mathbf{Y}_i; \mathbf{X}_i \boldsymbol{\psi}_k, \boldsymbol{\Sigma} \otimes \mathbf{I}_{n_i} + \mathbf{Z}_i \boldsymbol{\Gamma} \boldsymbol{\Gamma}' \mathbf{Z}_i') \\ &\quad \times \left\{ \lambda_k(T_i) \exp(\mathbf{u}_i' \boldsymbol{\gamma}) \right\}^{\Delta_i} \\ &\quad \left. \times \exp \left\{ -\exp(\mathbf{u}_i' \boldsymbol{\gamma}) \Lambda_k(T_i) \right\} \right], \quad (5) \end{aligned}$$

where the second equality holds because of the conditional independence assumption and $\Phi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the density function of a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The log-likelihood of the observed data can be written as

$$\mathcal{L}_o(\boldsymbol{\Xi}; \{\mathbf{Y}_i, T_i, \Delta_i, \mathbf{W}_i\}_{i=1}^N) = \sum_{i=1}^N \log \left\{ P(\mathbf{Y}_i, T_i, \Delta_i | \mathbf{W}_i; \boldsymbol{\Xi}) \right\}. \quad (6)$$

2.5. Variable Selection via Regularization

To select covariates with fixed and/or random effects, we consider a regularized likelihood approach. For the selection of covariates with fixed effects, we propose to add a penalty function of $\boldsymbol{\Psi}$ to the log-likelihood. Specifically, we choose the adaptive pairwise fused lasso penalty, which has been adopted in a clustering algorithm by Guo et al. (2010), as follows,

$$\mathcal{R}(\boldsymbol{\Psi}) = \sum_{j=1}^p \sum_{1 \leq k < k' \leq K} \tau_1 \zeta_{1jkk'} |\boldsymbol{\Psi}_{jk} - \boldsymbol{\Psi}_{jk'}| + \sum_{j=1}^p \sum_{k=1}^K \tau_2 \zeta_{2jk} |\boldsymbol{\Psi}_{jk}|, \quad (7)$$

where τ_1 and τ_2 are tuning parameters and $\zeta_{1jkk'}$ and ζ_{2jk} are the corresponding positive weights for the penalty terms. These weights can be chosen based on some preliminary analysis so that larger effects are penalized less and smaller effects are penalized more, which can avoid selection inconsistency and alleviate estimation bias. For example, a rough estimate $\hat{\boldsymbol{\Psi}}$ of $\boldsymbol{\Psi}$ can be obtained by optimizing the unpenalized likelihood function in equation (6) and $\zeta_{1jkk'}$ and ζ_{2jk} can be chosen as $1/|\hat{\boldsymbol{\Psi}}_{jk} - \hat{\boldsymbol{\Psi}}_{jk'}|$ and $1/|\hat{\boldsymbol{\Psi}}_{jk}|$, respectively. Further details on the adaptive penalty can be found in Zou (2006).

For the selection of covariates with random effects, we add an adaptive group lasso penalty function of $\boldsymbol{\Gamma}$ to the log-likelihood (Yuan and Lin, 2006), which can be written as

$$\mathcal{R}(\boldsymbol{\Gamma}) = \sum_{j \in \mathcal{S}} \tau_3 \zeta_{3j} \|\boldsymbol{\Gamma}_j\|, \quad (8)$$

where $\|\cdot\|$ denotes the ℓ_2 norm, τ_3 is a tuning parameter, ζ_{3j} is the corresponding positive weight for the j_{th} penalty term, $\boldsymbol{\Gamma}_j$ denotes the j_{th} row of $\boldsymbol{\Gamma}$, and \mathcal{S} denotes the set of indices for rows of $\boldsymbol{\Gamma}$ for penalty (the rows corresponding to random intercepts are excluded from \mathcal{S} so that we can account for correlations among the repeated measures). Because of the property of the group lasso penalty, some rows of $\boldsymbol{\Gamma}$ can be shrunk to zero. This allows the selection of covariates with random effects, because it was shown in Wang et al. (2010) that $\boldsymbol{\Gamma}_j = \mathbf{0} \iff \mathbf{D}_j = \mathbf{D}_j = \mathbf{0}$, which indicates that the j_{th} covariate of random effect can be excluded from the model if $\boldsymbol{\Gamma}_j = \mathbf{0}$ (\mathbf{D}_j and \mathbf{D}_j denote the j_{th} row and column of \mathbf{D} , respectively). Please see Web Appendix A.1 for our structural assumption on matrix \mathbf{D} .

The regularized log-likelihood function can then be written as

$$\mathcal{L}_o(\Xi) \equiv \mathcal{L}_o(\Xi; \{Y_i, T_i, \Delta_i, \mathbf{W}_i\}_{i=1}^N) - \mathcal{R}(\Psi) - \mathcal{R}(\Gamma), \quad (9)$$

and the estimation of parameters in our proposed model is done by maximizing this regularized log-likelihood function via an EM algorithm. Please see Web Appendix A.2, A.3, and A.4 for detailed implementation of our method.

3. Statistical Inference

3.1. Model Selection

In our proposed model, a data-driven method for determining tuning parameters $\{\tau_1, \tau_2, \tau_3\}$ and the number of latent classes K is essential. To do this, we consider a BIC-type criterion following Guo et al. (2010) as

$$\text{BIC}(K, \{\tau_1, \tau_2, \tau_3\}) = -2\mathcal{L}_o(\hat{\Xi}; \{Y_i, T_i, \Delta_i, \mathbf{W}_i\}_{i=1}^N) + d_1 \log(N_1) + d_2 \log(N_2), \quad (10)$$

where $\hat{\Xi}$ are the parameter estimates using the number of latent classes K and tuning parameters $\{\tau_1, \tau_2, \tau_3\}$, d_1 and d_2 are the degrees of freedom as defined below, $N_1 = N$, and $N_2 = G(\sum_{i=1}^N n_i)$. The degrees of freedom in the survival submodel and class membership probability are defined as $d_1 = (K-1)m + 2K + s$ and the degrees of freedom in the longitudinal submodel are defined as $d_2 = G + \sum_{j=1}^p \text{DNZ}(\hat{\Psi}_j) + \text{NZL}(\hat{\Gamma})$, where $\text{DNZ}(\cdot)$ and $\text{NZL}(\cdot)$ count the number of distinct non-zero parameters in a vector and the number of non-zero parameters in the lower triangular part of a matrix (including the diagonal elements), respectively.

3.2. Dynamic Prediction

Consider a new subject i free of event at time point t_l , let $\mathbf{Y}_i^{(t_l)}$ and T_i^* denote the observed gene expression profiles of subject i up to time t_l and the true survival time of subject i , respectively. Then, for any time point $t_u \geq t_l$, the probability of the event occurring between t_l and t_u for subject i can be calculated as

$$\begin{aligned} \rho_i(t_u, t_l) &\equiv P(T_i^* \leq t_u | T_i^* \geq t_l, \mathbf{Y}_i^{(t_l)}, \mathbf{W}_i; \Xi), \\ &= \sum_{k=1}^K P(T_i^* \leq t_u | c_{ik} = 1, T_i^* \geq t_l, \mathbf{Y}_i^{(t_l)}, \mathbf{W}_i; \Xi) \\ &\quad P(c_{ik} = 1 | T_i^* \geq t_l, \mathbf{Y}_i^{(t_l)}, \mathbf{W}_i; \Xi), \\ &= \sum_{k=1}^K \frac{P(t_l \leq T_i^* \leq t_u | c_{ik} = 1, \mathbf{W}_i; \Xi)}{P(T_i^* \geq t_l | c_{ik} = 1, \mathbf{W}_i; \Xi)} \\ &\quad \frac{P(T_i^* \geq t_l, \mathbf{Y}_i^{(t_l)} | c_{ik} = 1, \mathbf{W}_i; \Xi) P(c_{ik} = 1 | \mathbf{W}_i; \Xi)}{\sum_{j=1}^K P(T_i^* \geq t_l, \mathbf{Y}_i^{(t_l)} | c_{ij} = 1, \mathbf{W}_i; \Xi) P(c_{ij} = 1 | \mathbf{W}_i; \Xi)}, \end{aligned} \quad (11)$$

where each term can be easily calculated using the corresponding formula in equation (5). To evaluate the predictive

performance, we adopt a similar measure as in Rizopoulos et al. (2014), that is the area under the receiver operating characteristic curve (AUC) defined as

$$\text{AUC}(t_u, t_l) = P\{\rho_i(t_u, t_l) < \rho_j(t_u, t_l) | T_i^* > T_j^* > t_l\}. \quad (12)$$

Please see Web Appendix A.5 for details on calculation of the AUC.

3.3. Inference

Given the complexity of our proposed model as well as the regularization, we adopt the bootstrap approach to do inference as previously done for regularized survival model in Sinnott and Cai (2016). Specifically, for the a_{th} bootstrap, we optimize the weighted regularized log-likelihood written as

$$\mathcal{L}_{or}^{(a)}(\Xi) \equiv \sum_{i=1}^N \kappa_i^{(a)} \log\{P(\mathbf{Y}_i, T_i, \Delta_i | \mathbf{W}_i; \Xi)\} - \mathcal{R}(\Psi) - \mathcal{R}(\Gamma), \quad (13)$$

where the positive weights $\kappa_i^{(a)}$ are independently generated from a distribution with both mean and variance being one and these weights play the role of resampling. In this article, we generate the weights using a Beta distribution, that is $\kappa_i^{(a)} \sim 4\text{Beta}(0.5, 1.5)$, as done in Sinnott and Cai (2016). Please see Web Appendix A.6 for details on construction of 95% CI based on bootstrap.

4. Simulation Studies

4.1. Simulation Settings

In our simulation study, the data are generated according to equations (1), (2), and (4) (Please see Web Appendix A.7 for details on the data generation mechanism.). Let α be the parameter controlling the censoring rate and r be the proportion of non-informative genes. We design four simulation scenarios by varying the configurations for (α, r) and the forms of the class-specific baseline hazard functions so that we can see how the performance of our proposed method is influenced by these factors.

- Scenario 1: We select $(\alpha, r) = (6, 0.2)$ so that the censoring rate is about 15% and most of the covariates of fixed effects are informative about the latent classes. The class-specific baseline hazard functions are specified to be the Weibull hazard functions with shape and scale parameters being (4,2), (5,3), and (6,4), respectively.
- Scenario 2: Compared to Scenario 1, we change α to 3.8 so that the censoring rate is increased to about 40%.
- Scenario 3: Compared to Scenario 1, the class-specific baseline hazard functions are changed to be the Gamma hazard functions with shape and scale parameters being (10,10), (20,10), and (30,10), respectively.
- Scenario 4: Compared to Scenario 1, r is increased to 0.4 so that the number of informative genes is reduced.

In all simulation settings, we consider the number of genes to be 20 ($G = 20$), the total number of subjects to be 150 ($N =$

150), and the number of latent classes to be three ($K = 3$). For each scenario, we simulate 100 datasets. For each dataset, 50 subjects are generated additionally for each latent class as an independent testing dataset.

4.2. Simulation Results

We call our proposed method rJLCM (short name for regularized Joint Latent Class Model) from here on in our manuscript. We compare the performance of rJLCM to other joint modeling approaches. Specifically, we compare rJLCM to two variants of rJLCM, denoted by rJLCM-trueK and JLCM, where JLCM is rJLCM run with no penalties and rJLCM-trueK is rJLCM run by fixing the number of latent classes at the true value (i.e., $K = 3$). We include a shared random-effects model for comparison by adopting the R package *JM* to perform the inference (Please see Web Appendix A.8 for details on implementation). The longitudinal biomarkers in all methods are modeled using the same covariates of fixed and random effects, that is $\mathbf{z}_{igj} = \mathbf{x}_{igj} = (1, t_{ij}, t_{ij}^2)$.

We compare the predictive performance of all methods using $\text{AUC}(t_l, t_u)$, where the time point t_l is chosen to be the largest value so that all subjects in the testing dataset are at risk and $t_u = t_l + 0.1$ (other values for t_u such as $t_u = t_l + 0.5$ give similar results). In addition, we are also interested in how well rJLCM-trueK, rJLCM, and JLCM can assign subjects to the correct latent classes, for which we adopt the adjusted Rand index to evaluate the performance (Hubert and Arabie, 1985). We run rJLCM and JLCM with K ranging from 2 to 5 and the K values selected by BIC are recorded.

From Table 1, we can see that rJLCM-trueK has better clustering performance (i.e., larger Rand indices) than rJLCM in all scenarios, since rJLCM-trueK starts with the true number of latent classes. The clustering performance of rJLCM is better than JLCM in the first three scenarios, where the ratios of informative genes are large and the latent class structures are easy to be captured in these scenarios. The rJLCM selects the correct number of latent classes ($K = 3$) for most cases while JLCM usually underestimates the number

of latent classes, suggesting that removing uninformative variables about the latent classes using regularization can improve the clustering performance. However, in Scenario 4, where the ratio of informative genes is small and it is difficult to identify the true underlying latent classes, the clustering performance of rJLCM is similar to that of rJLCM, both of which are worse than rJLCM-trueK since it uses the true number of latent classes.

In terms of predictive performance (AUC), rJLCM-trueK is slightly better than rJLCM, which outperforms JLCM in the first three scenarios. This is expected as the clustering performance influences the predictive performance. One exception is that the AUCs are similar for rJLCM-trueK, rJLCM, and JLCM in Scenario 4, where the ratio of noise genes is large and hence it is difficult to identify the true underlying latent classes. This suggests that inclusion of too many noise genes in the model can diminish the predictive performance, although the penalization can remove some of the noise genes. In addition, the performance of JM is worse than rJLCM, suggesting that it is important to use all rather than one informative gene to make prediction. It seems that both the censoring rate and misspecification of baseline hazard function do not have a large impact on the clustering and predictive performance comparing Scenarios 2 and 3 to Scenario 1, respectively, suggesting that our proposed method is robust. We also observe that rJLCM outperforms another standard approach based on the last value carried forward approach and that rJLCM performs reasonably well in variable selection (Please see Web Appendix A.8 for these results).

We next study the performance of bootstrap for inference in rJLCM. From Table 2, we can see that the 95% CIO constructed using bootstrap have coverage probabilities close to 0.95, suggesting that bootstrap is a promising tool to do inference in rJLCM. Also, we can see that the biases are relatively small, even although we have included regularization in our model. Possible explanations for the small bias might be that the regularization part is in the longitudinal submodel which has little effect on the estimation of parameters in the survival

Table 1

Comparisons in performance of rJLCM-trueK, rJLCM, JLCM, and JM using 100 simulated datasets for each of the four scenarios. The number in each of the cells for Rand and AUC indicates the average value of adjusted Rand indices and AUCs across 100 simulated datasets, respectively, with standard deviation in the parenthesis. The numbers in each of the cells for K indicate the times of each K (ranging from 2 to 5) that is selected by BIC.

Scenarios		rJLCM-trueK	rJLCM	JLCM	JM
1	Rand	0.99 (0.01)	0.92 (0.18)	0.53 (0.12)	NA
	K	(0, 100, 0, 0)	(15, 85, 1, 0)	(98, 2, 0, 0)	NA
	AUC	0.82 (0.02)	0.81 (0.02)	0.80 (0.02)	0.76 (0.03)
2	Rand	0.99 (0.01)	0.88 (0.22)	0.53 (0.11)	NA
	K	(0, 100, 0, 0)	(23, 77, 0, 0)	(97, 3, 0, 0)	NA
	AUC	0.84 (0.02)	0.83 (0.03)	0.81 (0.02)	0.76 (0.03)
3	Rand	0.99 (0.02)	0.94 (0.15)	0.57 (0.15)	NA
	K	(0, 100, 0, 0)	(11, 89, 1, 0)	(92, 8, 0, 0)	NA
	AUC	0.85 (0.02)	0.84 (0.03)	0.81 (0.02)	0.74 (0.02)
4	Rand	0.75 (0.17)	0.45 (0.08)	0.45 (0.08)	NA
	K	(0, 100, 0, 0)	(100, 0, 0, 0)	(100, 0, 0, 0)	NA
	AUC	0.79 (0.03)	0.78 (0.02)	0.78 (0.02)	0.74 (0.03)

Table 2

Coverage probability of the 95% CI constructed using 100 bootstraps and the biases in the estimation using simulated data from Scenario 1. The numbers in the parenthesis indicate standard deviation. These results are averaged over the all parameters of the same type.

Targets	Values
Coverage probability of 95% CI for γ	96.5%
Biases in estimating γ	0.034 (0.12)
Coverage probability of 95% CI for $\rho_i(t_l + 0.1, t_l)$	96.3% (0.02)
Biases in estimating $\rho_i(t_l + 0.1, t_l)$	0.000084 (0.0006)

submodel, since the longitudinal model affects the survival model only through the class membership and some bias in the estimation of the parameters in the longitudinal submodel might not have large effects on the class membership probabilities.

5. IPF Data Analysis

In this section, we apply rJLCM to the IPF dataset as briefly described in Section 1 and compare its performance to JLCM and JM. IPF is a highly lethal lung disease and lung transplantation is a plausible option for late-stage patients who

two batches among these 57 patients. For the first batch of patients recruited in our study, their gene expression data were measured up to one year while we only measured the baseline gene expression data for the second batch of patients. All patients were followed up to record their survival outcome until death or censored. It is widely known that age, gender, and percent predicted forced vital capacity (FVC) are three important clinical variables related to the survival outcome of IPF patients (Ley et al., 2012). In our dataset, the average age of the subjects is 67.42 with standard deviation 7.99, the proportion of males is 0.67, and the average FVC is 74.06 with standard deviation 21.87.

For the gene expression dataset, it is widely recognized that most genes are redundant and selecting informative genes before clustering analysis can improve the results (Hastie et al., 2000). Following the same idea, we calculate variance for each gene using all data from all subjects and then select 20 genes with the largest variances out of about 20,000 genes for our analysis, as they explain a large proportion of the variance in the data, which is likely related to patient heterogeneities. Then, all covariates, including age, gender, FVC, and gene expression data, are standardized to have mean 0 and variance 1 in all analyses. Our goal is to build a risk prediction model using these three clinical variables and the longitudinal gene expression profiles.

We run rJLCM and JLCM with the number of latent classes K varying from 2 to 5. Each of the submodels is specified as follows.

$$P(c_{ik} = 1 | \text{age}_i, \text{gender}_i, \text{FVC}_i) = \frac{\exp(\text{age}_i \eta_{k1} + \text{gender}_i \eta_{k2} + \text{FVC}_i \eta_{k3})}{\sum_{j=1}^K \exp(\text{age}_i \eta_{j1} + \text{gender}_i \eta_{j2} + \text{FVC}_i \eta_{j3})},$$

$$h_i(t | c_{ik} = 1, \text{age}_i, \text{gender}_i, \text{FVC}_i) = \lambda_k(t) \exp(\text{age}_i \gamma_1 + \text{gender}_i \gamma_2 + \text{FVC}_i \gamma_3),$$

$$Y_i = X_i \Psi c_i + Z_i \Gamma b_i + \epsilon_i,$$

are at high risk of death. Therefore, it is important to calibrate the risk of death for IPF patients to guide and deliver precision care. In the IPF dataset, there are totally 57 patients with the median number of visits being 3, among which 23 patients died and 34 patients were censored. There are

where Y_i indicates the longitudinal gene expression profiles of subject i , X_{ig} is a matrix with the j th row being $x_{igj} = (1, t_{ij}, t_{ij}^2)$ (t_{ij} is the j th visit time point of subject i), and $Z_i = X_i = \text{bdiag}(X_{i1}, \dots, X_{iG})$ is a block-diagonal matrix. In all analyses,

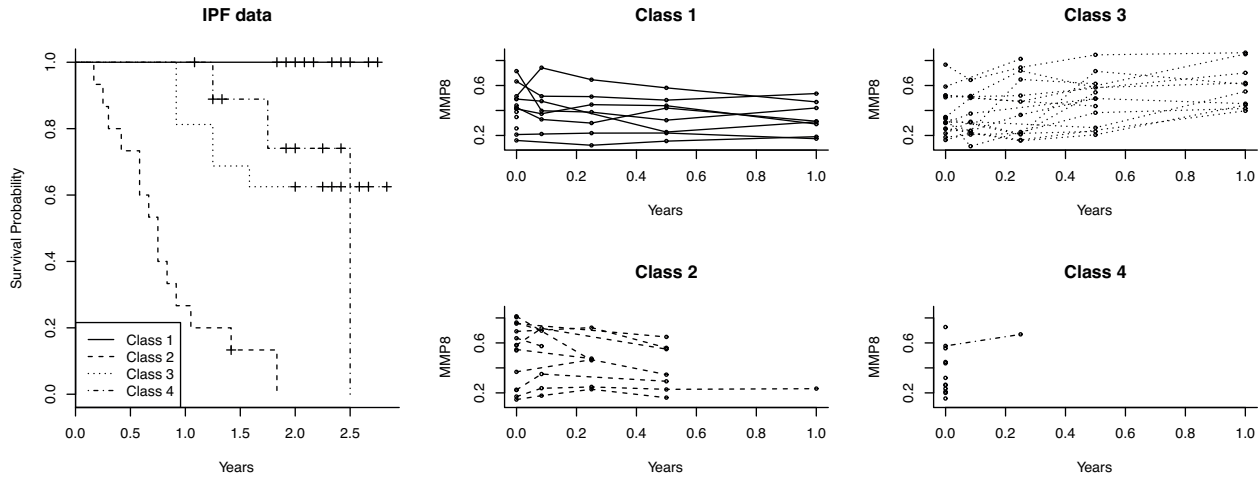


Figure 1. The survival curves for the four latent classes identified in the IPF dataset with corresponding trajectories of gene MMP8.

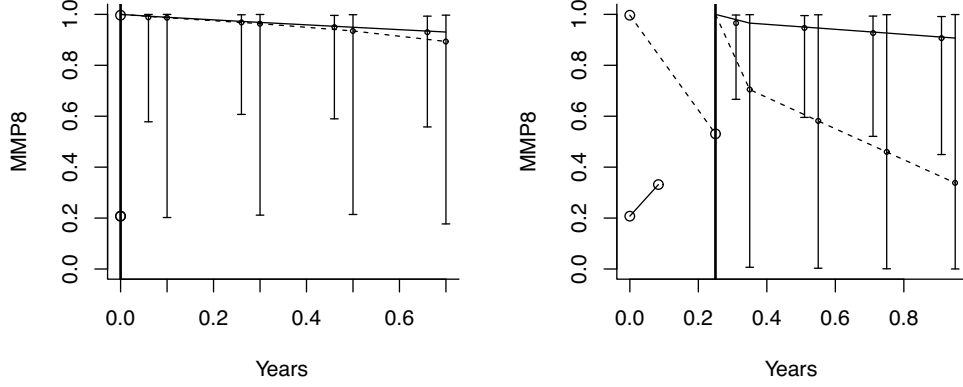


Figure 2. Dynamic prediction plot with both the trajectories of the gene MMP8 (on the left) and the predicted survival probabilities (on the right). Dashed line and solid line indicate different patients and the error bars represent 95% CI.

we use quadratic term in time to model the trajectories of genes, which should be adequate, as shown in the right panel of Figure 1.

Based on the results of rJLCM, the number of latent classes is selected to be four. The survival curves for the four latent classes are shown in the left panel of Figure 1, where we can see patients in Class 2 have the worst survival outcomes while the other three classes have similar survival outcomes. After looking into the estimates of Ψ , we find that the coefficients for the linear term of time for gene MMP8 differ a lot among Classes 1, 2, and 3. From the trajectories of gene MMP8 for Classes 1, 2, and 3, as shown in the right panel of Figure 1, we can see that the trend in trajectory of MMP8 is stable for Class 1, decreasing for Class 2, and increasing for Class 3. These suggest that decreased expression value of MMP8 in the blood might indicate increased risk of death. The trajectories of MMP8 for Class 4 appear similar to that of Class 3. In fact, trajectories of some other genes for Class 4 differ from that of the other classes, as the latent classes are determined by the informative variables from all genes.

To compare the predictive performance of rJLCM, JLCM, and JM, we calculate the $AUC(t_l, t_u)$ for the three methods using cross-validation. Specifically, we hold out each pair of patients whose survival outcomes are comparable as the testing dataset and the remaining patients are treated as the training dataset. For each comparable pair (patients i and j), the t_l is chosen as the minimum of the last time points when the gene expression data were measured of the two patients (i.e., $t_l = \min(t_{in_i}, t_{in_j})$) and t_u is fixed at $t_l + 0.1$ years. Based on the results of 100 comparable pairs, the AUCs are 0.79, 0.72, and 0.45 for rJLCM, JLCM, and JM, respectively, suggesting that the prediction accuracy of rJLCM is improved over that of the other two methods. The AUC is usually greater than 0.5 while the AUC of JM is less than 0.5, which is possibly due to randomness.

To show the dynamic prediction visually, we plot survival probabilities for one comparable pair (i.e., two patients in the testing dataset) based on their baseline observations and observations up to 0.25 years. Specifically, for a given t_l , we estimate $\rho(t_u, t_l)$ for $t_u = (t_l, t_l + 0.1, t_l + 0.3, \dots, t_l + 0.7)$ and plot values of $1 - \rho(t_u, t_l)$, as shown in Figure 2. In fact, for

MMP8, only the slopes of the trend differ across different classes, based on the estimate of Ψ . Therefore, as shown in the left panel of Figure 2, which is based on the baseline observations only, the phenomenon that the estimated survival probabilities for the two patients start to diverge around 0.7 years could be mainly due to the differences in the baseline clinical variables. When more measurements in the longitudinal gene expression profiles are taken, the estimated survival probabilities for the two patients start to diverge around 0.4 years and the estimated survival probabilities for the patient in dashed lines drop drastically as time approaches 1.0 years, which could be mainly due to the fact that the decreasing trend of MMP8 for this patient resembles that of Class 2 (as shown in Figure 1) and around 80% of the patients in Class 2 died before 1.0 years. In contrast, the increasing trend of MMP8 for the patient in solid lines resembles that of Class 3 and only about 20% of the patients in Class 3 died before 1.0 years, which makes the estimated probabilities stable for this patient. In fact, the patient in dashed line died at 0.92 years while the patient in solid lines was censored at 2.67 years. These suggest that it is critical to update the measurements of the gene expression profiles over time, which could provide more accurate predictions. However, based on the wide 95% CI for the survival probabilities, which is very likely due to the small sample size in our dataset, we cannot conclude that the patient in solid lines is significantly better than the patient in dashed lines. We also observe that the estimates for the parameters in the survival submodel are similar to that from a standard Cox model, but with larger standard errors due to the complexity of our method (Please see Web Appendix A.10 for these results).

6. Discussion

In this article, we have proposed a novel regularized joint modeling approach under the framework of joint latent class model to deal with the high dimensional longitudinal biomarkers. Specifically, we utilize the penalized likelihood approach to allow selection of covariates with fixed effects, covariates with random effects, and informative covariates for the latent

classes. We showed that our proposed method has improved clustering and predictive performance over other existing methods through both simulation studies and real data application.

Our proposed method is built upon the latent class model framework, which has some disadvantages. Firstly, it is not straightforward to assess the association between the biomarker and the survival outcome in the latent class model. So, if the primary goal is to do hypothesis testing rather than risk prediction as in our article, the shared random-effects model could be more suitable. Having said that, our proposed model can, to some extent, allow us to select which genes are indirectly related to the survival outcome, since the genes affect the latent classes and the latent classes affect the survival outcome. In addition, the latent classes may not be easy to interpret in some situations. Moreover, the conditional independence assumption is a strong assumption and may not always hold. There are ways to test the conditional independence assumption (Jacqmin-Gadda et al., 2010), however, it relies on the model setup and works for certain types of alternative hypotheses.

As mentioned in Section 2.5 and Web Appendix A.1, we impose some structural assumption on the covariance matrix \mathbf{D} to improve model efficiency. In our current specification of the covariance matrix, the $(G \times G)$ covariance matrix among random intercepts is unstructured. This works well when the number of genes (G) is about 50 based on our experience. Although the number of informative genes related to the outcome of interest is not large for most cases, it is possible that the number of informative genes increases to the scale of hundreds or even thousands. In these cases, the specification of the covariance matrix \mathbf{D} to be a block diagonal matrix with each block corresponding to the random effects for each gene could be more practical. Based on an additional simulation, we find that this extended version works well when the number of genes is 500 and has improved performance over other existing methods (Please see Web Appendix A.9 for details).

In our current framework, we adopt a proportional hazards model with Weibull baseline hazard function for the survival submodel, because of the small sample size in the IPF dataset. The survival submodel can be easily extended to accommodate more flexible scenarios when the sample size becomes moderate to large. For example, the class-specific baseline hazard functions can be specified as piecewise-constant hazard functions to allow more flexibility; the vector of coefficients ($\boldsymbol{\gamma}$) in the survival submodel can also be class-specific; the random effects can also be class-specific in the longitudinal submodel. Based on an additional simulation, we find that the extended version with class-specific parameters in the survival submodel can further improve the prediction if those parameters are truly class-specific (Please see Web Appendix A.9 for details).

7. Supplementary Materials

Web Appendices referenced in Sections 2.5, 3.2, 3.3, 4.1, 4.2, 5, and 6 are available with this article at the *Biometrics* website on Wiley Online Library. The data and R codes used in this article are available at the *Biometrics* website on Wiley Online Library.

ACKNOWLEDGEMENTS

Jiehuan Sun and Hongyu Zhao were supported in part by the National Institutes of Health grants R01 GM59507 and P01 CA154295. Jose D. Herazo-Maya was supported by the Pulmonary Fibrosis Foundation and the Robert Wood Johnson Foundation under the Harold Amos Medical Faculty Development Program. Toby M. Maher was supported by an NIHR Clinician Scientist Fellowship (NIHR Ref: CS-2013-13-017). Naftali Kaminski was supported in part by the National Institutes of Health grants U01 HL122626, UH3 HL123886, R01 HL127349, and U01 HL112707. The contents are solely the responsibility of the authors and do not necessarily represent the official view of NIH.

REFERENCES

- 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.
- Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics* **59**, 762–769.
- Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2010). Pairwise variable selection for high-dimensional model-based clustering. *Biometrics* **66**, 793–804.
- Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., et al. (2000). ‘Gene shaving’ as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology* **1**, 1–3.
- He, Z., Tu, W., Wang, S., Fu, H., and Yu, Z. (2015). Simultaneous variable selection for joint models of longitudinal and survival outcomes. *Biometrics* **71**, 178–187.
- Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1**, 465–480.
- Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification* **2**, 193–218.
- Jacqmin-Gadda, H., Proust-Lima, C., Taylor, J. M., and Commenges, D. (2010). Score test for conditional independence between longitudinal outcome and time to event given the classes in the joint latent class model. *Biometrics* **66**, 11–19.
- Ley, B., Ryerson, C. J., Vittinghoff, E., Ryu, J. H., Tomassetti, S., Lee, J. S., et al. (2012). A multidimensional index and staging system for idiopathic pulmonary fibrosis. *Annals of Internal Medicine* **156**, 684–691.
- 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.
- Liu, Y., Liu, L., and Zhou, J. (2015). Joint latent class model of survival and longitudinal data: An application to CPCRA study. *Computational Statistics & Data Analysis* **91**, 40–50.
- Molyneaux, P. L., Willis-Owen, S. A., Cox, M. J., James, P., Cowman, S., Loebinge, M., et al. (2017). Host-microbial interactions in idiopathic pulmonary fibrosis. *American Journal of Respiratory and Critical Care Medicine* **195**, 1640–1650.
- Proust-Lima, C., Séne, M., Taylor, J. M., and Jacqmin-Gadda, H. (2014). Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research* **23**, 74–90.
- Proust-Lima, C. and Taylor, J. M. (2009). Development and validation of a dynamic prognostic tool for prostate cancer

- recurrence using repeated measures of posttreatment PSA: A joint modeling approach. *Biostatistics* **10**, 535–549.
- Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics* **67**, 819–829.
- Rizopoulos, D., Hatfield, L. A., Carlin, B. P., and Takkenberg, J. J. (2014). Combining dynamic predictions from joint models for longitudinal and time-to-event data using Bayesian model averaging. *Journal of the American Statistical Association* **109**, 1385–1397.
- Sinnott, J. A. and Cai, T. (2016). Inference for survival prediction under the regularized cox model. *Biostatistics* **17**, 692–707.
- Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **67**, 91–108.
- Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statistica Sinica* **14**, 809–834.
- Van Houwelingen, H. C. (2007). Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics* **34**, 70–85.
- Wang, S., Song, P., 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.
- Wang, Y. and Taylor, J. M. G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. *Journal of the American Statistical Association* **96**, 895–905.
- 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.
- 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.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

Received June 2017. Revised July 2018. Accepted August 2018.