Latent Factor Modeling of Lymphoma Survival Times

By Steve Kang, Sam Yin, Lucy Lu

Abstract

In this case study, we examine length of survival and gene expression levels for 240 subjects. In order to reduce number of covariates, we first conduct independent screening with Cox proportional model. Afterwards, we impute censored data with a Gibbs Sampler. As for modeling and prediction, we use both Unweighted and Weighted Latent Factor Models with 9 latent factors. The result indicates that Unweighted Latent Factor Model performs the best by root mean squared error (RMSE).

Independent Screening

The large number of predictors (7,399 gene expression levels) in this data set necessitates variable selection. Independent screening with Cox proportional hazard model is applied for each covariate:

$$\lambda_i(t) = \lambda_0(t)\exp(\beta_j x_{ij}),\ i = 1, ..., n,\ j = 1, ..., m$$

where $\lambda_i(t)$ is the hazard rate for subject i at time t $\lambda_0(t)$ is the baseline hazard when no effect of explanatory variable (gene expression levels) is present, and β_j is the coefficient for x_{ij}, the j^{th} gene of subject i. The threshold for variable selection is determined by False Discovery Rate (FDR) correction which controls for the overall Type I error rate in multiple comparisons. The procedure (also referred to as Benjamini–Hochberg procedure) is as follows:

1. Sort the p-values $p_1, p_2, \ldots, p_m$ for all m models in ascending order;
2. For $\alpha = 0.05$, find the largest k such that $p_k \leq \frac{k}{m}\alpha$;
3. Reject the null hypotheses corresponding to $p_1, p_2, \ldots, p_k$, i.e., select the first k corresponding predictors.

The above procedure identifies 77 genes which will be used in imputation and modeling.

Data Imputation

The follow-up period in the original study was not able to cover the life span of many patients. As a result, the survival time of 102 out of 240 subjects are right-censored. The following Gibbs sampler is used for data imputation: for $t = 1, 2, \ldots, T$

1. Sample $\tilde{Y}_i^{(t)} \sim TruncNorm(\boldsymbol{\beta}^T\mathbf{X}_i, \tau^{-1}; L = \tilde{Y}_i^{(0)})$ for all censored Y_i's (where $\tilde{Y}_i^{(0)}$ is based on the observed survival time),
2. Update $\boldsymbol{\beta} \sim MVN(\Sigma(\Sigma_0^{-1}\boldsymbol{\beta} + \tau\mathbf{X}^T\mathbf{Y}), \Sigma)$ where $\Sigma = (\Sigma_0^{-1} + \tau\mathbf{X}^T\mathbf{X})^{-1}$ and $\tau \sim Gamma(a + \frac{n}{2}, b + \frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}))$.

The parameters are initialized as $\boldsymbol{\beta} \sim MVN(\boldsymbol{\beta}_0, \Sigma_0)$ and $\tau \sim Gamma(a, b)$ where $\boldsymbol{\beta}_0 = \mathbf{0}, \Sigma_0 = \mathcal{I}, a = b = 1$. The median of the sampled $\mathbf{Y}_i^{(t)}$'s are imputed.

Latent Factor Model

Assume that several groups of correlated genes form pathways "meta genes"and collectively affect the survival time of subjects. These pathways can be modeled as k latent factors (k << p, where p is the number of explanatory variables in the original data set), and the covariance matrix of $\mathbf{X}$ can be decomposed as $\Sigma = \Lambda\Lambda^T + \Psi$ where Λ has dimension $p \times k$ and Ψ is a $p \times p$ diagonal matrix. Given the latent factors, the gene expression levels and the survival time are conditionally independent.
This latent factor model is formulated as follows:

$$x_{ij} = \lambda_j\eta_i + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \phi_j),$$
$$y_i = \alpha + \eta_i^T\beta + \nu_i, \nu_i \sim N(0, \sigma^2),$$

where y_i is the log survival time of the i^{th} subject, λ_j is the j^{th} row of the factor loading matrix Λ, η_i is the level of latent factors for the i^{th} subject. One modification to the likelihood function is upweighting y_i: $\mathcal{L}(\mathbf{X}, \mathbf{Y}|-) = \prod_{i=1}^n\{[\prod_{j=1}^p(P(X_{ij}|-))]P(\log(Y_i)|-)^w\}(w \geq 1)$.
The model was implemented with JAGS package in RStudio. The training set consists of the first 160 subjects, and the test set the other 80. The prediction of log survival time is based on the posterior means of parameters. $\hat{Y} \sim N(\hat{\alpha} + \mathbf{X}\hat{\Lambda}(\hat{\Lambda}^T\hat{\Lambda})^{-1}\hat{\beta}, \hat{\sigma}^2)$.

Result

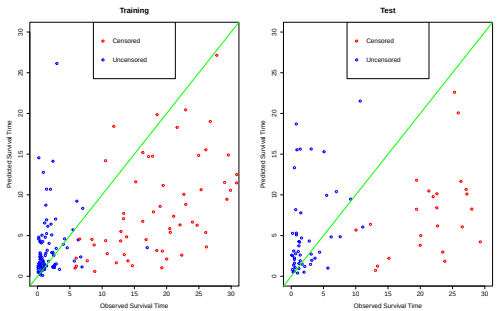

Figure 1: Observed vs Predicted Survival Times, Unweighted Latent Factor Model

Model	Training	Test
Unweighted	10.395	11.376
Weighted (w = 1)	22.314	20.874
Weighted (w = 50)	49.651	19.838
Weighted (w = 100)	115.056	26.850

Table 1: Comparison of Root Mean Squared Errors

Discussion

The unweighted latent factor model is the best performing of all; however, it is likely to underestimate the imputed data (Figure 1). This can be due to the limitation of the model or the inadequacy of single imputation.

The two features we tried to implement in the weighted models are upweighting the likelihood contribution of survival times, and enforcing sparsity by applying restricting priors to factor loadings. The reason they did not give satisfactory results can be due to imperfect tuning process (tuning parameters include number of latent factors, weighting, and sparsity control) and the size of data set.

Note that for the weighted models, training RMSEs are higher than test RMSEs. One possible reason is that some extreme predictions blow up errors, and the training set has more extreme predictions because they have more data points than the test set.