Kullback-Leibler-Based Discrete Relative Risk Models for Integration of Published Prediction Models with New Dataset

Di Wang DWWANG@UMICH.EDU
Wen Ye WYE@UMICH.EDU
Kevin He* KEVINHE@UMICH.EDU
Department of Biostatistics, University of Michigan, Ann Arbor, MI

Abstract

Existing literature for prediction of time-to-event data has primarily focused on risk factors from a single individual-level dataset. However, these analyses may suffer from small sample sizes, high dimensionality and low signal-to-noise ratios. To improve prediction stability and better understand risk factors associated with outcomes of interest, we propose a Kullback-Leibler-based discrete relative risk modeling procedure to borrow information from existing models. Simulations and real data analysis were conducted to show the advantage of the proposed method compared with those solely based on data from current study or prior information.

Keywords: Survival prediction; Kullback-Leibler distance; Discrete relative risk model; Data integration

1. Introduction

Prior research for predicting survival outcomes has primarily focused on data elements from a single individual-level dataset. These analyses may suffer from small sample sizes, high dimensionality and low signal-to-noise ratios. Moreover, data elements are restricted to those that already exist in the individual-level data. To incorporate prior knowledge and improve prediction performance, Schapire et al. (2005) proposed a boosting algorithm based on the Kullback Leibler (KL) measure of divergence (Kullback and Leibler, 1951) for classification of binary outcomes. More recently, Jiang et al. (2016) proposed a KL-based Lasso approach to improve variable selection for generalized linear models by integrating prior information. While successful, these methods, however, are not applicable for survival analysis with censored data. To fill in the gap and improve the prediction stability of time-to-event data, we utilize the fact that survival time can be viewed as a time-varying binary outcome. We therefore propose a time-dependent Kullback-Leibler discrimination information and develop a discrete relative risk modeling procedure to aggregate the current individual-level data with prior knowledge gathered from previously published prediction models.

[*] Corresponding author

2. Notation

Let T_i denote the failure time of interest and C_i be the censoring time for patient i, $i = 1, \ldots, n$, where n is the total sample size. The observed survival time is $X_i = \min\{T_i, C_i\}$. Let $t_1, \ldots, t_K$ be the distinct failure time and $k = 1, \ldots, K$ be the index of the distinct failure times. Let $\mathcal{D}_k$ denote the set of labels associated with individuals failing at time t_k. The set of labels associated with individuals censored at time t_k is denoted as $\mathcal{C}_k$. Let $\mathcal{R}_k$ denote the at risk set at time t_k. Let $\mathbf{Z}_i$ be an external and possibly time-dependent covariate vector for the i-th subject. Let $\lambda(t_k; \mathbf{Z}_i) = P(T_i = t_k | T_i \geq t_k, \mathbf{Z}_i)$ be the hazard at time t_k for the i-th patient with covariate $\mathbf{Z}_i$. The likelihood function is given by

$$L = \prod_{k=1}^{K} \left\{ \prod_{i \in \mathcal{D}_k} f(t_k; \mathbf{Z}_i) \prod_{i \in \mathcal{C}_k} S(t_k; \mathbf{Z}_i) \right\} = \prod_{k=1}^{K} \left[\prod_{i \in \mathcal{R}_k - \mathcal{D}_k} \{1 - \lambda(t_k; \mathbf{Z}_i)\} \prod_{i \in \mathcal{D}_k} \lambda(t_k; \mathbf{Z}_i) \right], \quad (1)$$

where $S(t_k; \mathbf{Z}_i) = P(T_i > t_k | \mathbf{Z}_i) = \prod_{\ell: t_\ell \leq t_k} \{1 - \lambda(t_\ell; \mathbf{Z}_i)\}$ is the survival function and $f(t_k; \mathbf{Z}_i) = P(T_i = t_k | \mathbf{Z}_i) = S(t_{k-1}; \mathbf{Z}_i) - S(t_k; \mathbf{Z}_i)$ is the density function of subject i at time t_k. Consider a general formulation of the hazard $\lambda(t_k; \mathbf{Z}_i) = g(\eta_k + \mathbf{Z}_i^\top \boldsymbol{\beta})$, where g denotes a monotonically increasing and twice differentiable link function, η_k is the baseline hazard of mortality at time t_k, and $\boldsymbol{\beta}$ denotes a coefficient vector associated with $\mathbf{Z}_i$. The log-likelihood is given by

$$\ell(\theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} Y_i(t_k) \left[\delta_i(t_k) \log \left\{ \frac{g(\eta_k + \mathbf{Z}_i^\top \boldsymbol{\beta})}{1 - g(\eta_k + \mathbf{Z}_i^\top \boldsymbol{\beta})} \right\} + \log\{1 - g(\eta_k + \mathbf{Z}_i^\top \boldsymbol{\beta})\} \right], \quad (2)$$

where $\theta = (\eta^\top, \beta^\top)^\top$, $Y_i(t_k) = I(X_i \geq t_k)$ is the at-risk indicator and $\delta_i(t_k) = I(T_i = t_k)$ is the death indicator at time t_k. Common choices for the link function g include complementary log-log (grouped relative risk model), log (discrete relative risk model), and logit (discrete logistic model).

3. KL discriminatory information for discrete time-to-event model

To extend the classical KL discrimination information to time-to-event data, we utilize the fact that the discrete failure time model can be viewed as a sequence of Bernoulli trials. Conditional on the event $T_i \geq t_k$, we define the KL discrimination information at time t_k as

$$\begin{aligned} D_{KL}(P_1, P_2; t_k, \mathbf{Z}_i) =& P_1(T_i = t_k | T_i \geq t_k, \mathbf{Z}_i) \log \left\{ \frac{P_1(T_i = t_k | T_i \geq t_k, \mathbf{Z}_i)}{P_2(T_i = t_k | T_i \geq t_k, \mathbf{Z}_i)} \right\} \\ &+ P_1(T_i > t_k | T_i \geq t_k, \mathbf{Z}_i) \log \left\{ \frac{P_1(T_i > t_k | T_i \geq t_k, \mathbf{Z}_i)}{P_2(T_i > t_k | T_i \geq t_k, \mathbf{Z}_i)} \right\}, \end{aligned}$$

which is a measure of disparity between two failure time distributions P_1 and P_2. Note that $D_{KL}(P_1, P_2; t_k, \mathbf{Z}_i)$ can also be considered as a measure of divergence between two probability distributions of a time-varying binary outcome $P_1(\Delta_{ik} = 1)$ and $P_2(\Delta_{ik} = 1)$, where $\Delta_{ik} \overset{d}{=} I(T_i = t_k | T_i \geq t_k, \mathbf{Z}_i)$ denotes the time-varying binary outcome for subject i

at time t_k, and d stands for distribution. That is,

$$D_{KL}(P_1, P_2; t_k, \mathbf{Z}_i) = P_1(\Delta_{ik} = 1) \log \left\{ \frac{P_1(\Delta_{ik} = 1)}{P_2(\Delta_{ik} = 1)} \right\} + P_1(\Delta_{ik} = 0) \log \left\{ \frac{P_1(\Delta_{ik} = 0)}{P_2(\Delta_{ik} = 0)} \right\}.$$

4. KL-based integration

We now apply the proposed KL discriminatory information to integrate discrete time-to-event models. Suppose P_1 is corresponding to a historical discrete time-to-event model, with parameters $\widetilde{\theta} = (\widetilde{\eta}^\top, \widetilde{\beta}^\top)^\top$ obtained from a previous study. Suppose P_2 is corresponding to the discrete time-to-event model, with parameters θ, over subjects in the current dataset.

Proposition 1 *Ignoring terms not involving θ, the KL measure $D_{KL}(P_1, P_2; t_k, \boldsymbol{Z}_i)$ between the historical model and the current model for subject i at time t_k is proportional to $-\widetilde{\ell}_{ik}(\theta)$, where*

$$\widetilde{\ell}_{ik}(\theta) = \tilde{\delta}_i(t_k) \log \left\{ \frac{g(\eta_k + \boldsymbol{Z}_i^\top \boldsymbol{\beta})}{1 - g(\eta_k + \boldsymbol{Z}_i^\top \boldsymbol{\beta})} \right\} + \log\{1 - g(\eta_k + \boldsymbol{Z}_i^\top \boldsymbol{\beta})\},$$

with $\tilde{\delta}_i(t_k) = \tilde{\lambda}(t_k; \boldsymbol{Z}_i) = g(\tilde{\eta}_k + \boldsymbol{Z}_i^\top \widetilde{\boldsymbol{\beta}})$ being the predicted outcome for subject i at time t_k based on the risk factors in the current data and the parameters from the historical model.

Combining the original log-likelihood (2) and the time-dependent KL measure, we define the resulting weighted log-likelihood function to link the historical model and the local dataset

$$\ell_\lambda(\theta) = \ell(\theta) - \lambda \tilde{\ell}(\theta), \tag{3}$$

where $\tilde{\ell}(\theta) = \sum_{i=1}^n \sum_{k=1}^K Y_i(t_k) \tilde{\ell}_{ik}(\theta)$. Here λ is a tuning parameter weighing the relative importance of the prior model to the local data and is determined using cross-validation. In the extreme case of $\lambda = 0$, the penalized log-likelihood is reduced to the log-likelihood based on the new data. In contrast, when $\lambda = \infty$, the model is equivalent to the historical model. Note that $\ell_\lambda(\theta)$ is proportional to the following objective function

$$\sum_{i=1}^n \sum_{k=1}^K Y_i(t_k) \left[\frac{\delta_i(t_k) + \lambda \tilde{\delta}_{ik}}{1 + \lambda} \log \left\{ \frac{g(\eta_k + \mathbf{Z}_i^\top \boldsymbol{\beta})}{1 - g(\eta_k + \mathbf{Z}_i^\top \boldsymbol{\beta})} \right\} + \log\{1 - g(\eta_k + \mathbf{Z}_i^\top \boldsymbol{\beta})\} \right], \tag{4}$$

for which the parameter estimation can be obtained similarly to the standard discrete relative risk models.

5. Simulation

To assess the performance of the proposed KL-based discrete modeling procedure, we conducted simulation studies to compare the following four models: the prior model using prior information only, the local model fitted by data of current study only, the stacked model fitted by stacked regression, and the KL model fitted by proposed KL-based discrete modeling procedure. Generally, the proposed modeling procedure works for log-log, log and logit link functions. Here, we used logit link as an example function to conduct the simulation studies.

Suppose $\boldsymbol{\beta}_0 = (\beta_0, \beta_1, \ldots, \beta_{p_0})^\top$ is the vector of coefficients from a previously published model, and $\boldsymbol{\beta}_l = (\beta_0, \beta_1, \ldots, \beta_{p_n})^\top$ is the vector of coefficients from which the current data is generated. We assume that the prior model and the current data share the same set of baseline hazard of mortality η_k at each discrete time point t_k. Then the current data was generated by $\boldsymbol{Z} \sim MVN(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ was a first-order autoregressive (AR1) correlation matrix with the auto-correlation parameter 0.5. For each time point t_k, the event indicator $Y_i(t_k)$ for each subject i in the at risk set $\mathcal{R}_k$ was generated by Bernoulli$(logit^{-1}(\eta_k + \boldsymbol{Z}_i^\top \boldsymbol{\beta}_l))$. We deleted subject i from the at risk set $\mathcal{R}_{T>t_k}$ for future time points if $Y_i(t_k) = 1$ at t_k; otherwise, we kept it. Latent censoring times were generated from a discrete uniform(1, 30) and truncated by an administrative censoring at time point 10.

We considered six different models which were clustered into two scenarios in the simulation studies:

Scenario 1: Current data and historical model share the same set of predictors:

Model (a): $\boldsymbol{\beta}_l = \boldsymbol{\beta}_0$;
Model (b): $\boldsymbol{\beta}_l =$reverse$(\boldsymbol{\beta}_0) = (\beta_{p_0}, \beta_{p_0-1}, \ldots, \beta_0)^\top$;
Model (c): $\boldsymbol{\beta}_l = \boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$, where $\epsilon_j \sim N(0, 0.25)$, $j = 1, \ldots, p_0$;

Scenario 2: Current data contains additional new predictors than historical model:

Model (d): $\boldsymbol{\beta}_l = (\boldsymbol{\beta}_0, 0.2\boldsymbol{\beta}_0)$;
Model (e): $\boldsymbol{\beta}_l = (\boldsymbol{\beta}_0, 0.5\boldsymbol{\beta}_0)$;
Model (f): $\boldsymbol{\beta}_l = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_0)$.

For scenario 1, model (a) and (b) simulated the situations when current data came from exactly the same or completely different distribution from the prior model, respectively; and the current data in model (c) was generated from a model which was similar to the prior model. For scenario 2, model (d), (e) and (f) simulated the situations where the additional new predictors in the current data were of different importance relative to the prior model, which were managed by adjusting magnitudes of the new predictors. Moreover, we set the local sample size $n_l = 300$, number of prior predictors $p_0 = 10$, number of additional new predictors $p_n = 10$ and the range of tuning parameters to be $\lambda \in [0, 10]$.

The tuning parameter λ was selected by 5-fold cross validation on the current data with average empirical log likelihood as the metric of model performance. After determining λ, we compared the proposed KL-based discrete modeling procedure with the prior model, local model and stacked model. In order to make a fair comparison, we evaluated the models on the hold-out external validation dataset, which was simulated from the same distribution as the current data. The best model achieved the maximal log likelihood on the external validation dataset. The simulation studies were replicated 100 times.

Figure 1 shows that the KL-based discrete modeling procedure achieved the best performance among four models under all scenarios. Specifically, the KL-based discrete modeling procedure favors the cases where the current data is similar to the prior model. Even under extreme situations when the current data was generated from completely different model from the prior model (Fig. 1 (b)) or missing important predictors (Fig. 1 (f)), the proposed

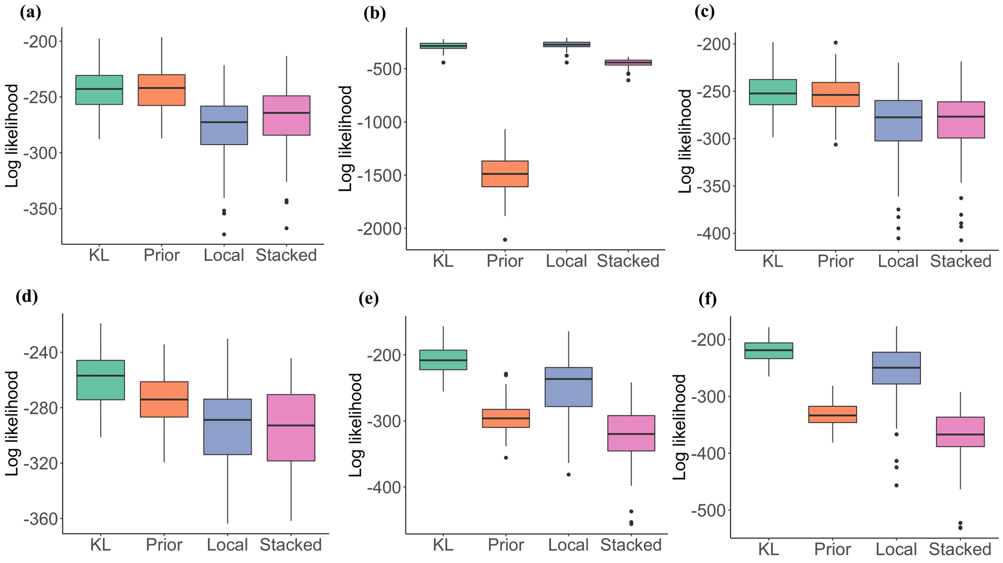

Figure 1: Simulation results of KL-based discrete modeling procedure (green) compared with prior (red), local (purple) and stacked regression (pink) models. (a)-(f) presents results for model setting (a)-(f) respectively.

modeling procedure did not result in worse predictions. We also presented the best tuning parameter λ determined by the proposed modeling procedure in the simulation studies. The KL-based discrete modeling procedure tend to select larger λ when the current data was more similar to the prior model. In addition, it did not incorporate misleading prior information which was not relevant to the local data by selecting an extremely small λ or setting it to be 0 (Figure 2).

6. Real Data Analysis

We used Estimated Post-Transplant Survival (EPTS) model and a local kidney transplant data as an example to illustrate how to use KL-based discrete modeling procedure to integrate previously published prediction model and new dataset. The raw EPTS score was derived from a Cox proportional hazards model using Scientific Registry of Transplant Recipients (SRTR) kidney transplant data. For simplicity, only 4 predictors were included in the raw EPTS score model: candidate's age in years, duration on dialysis in years, current diagnosis of diabetes, and whether the candidate has a prior organ transplant. Since the baseline survival information wasn't reported by the EPTS model, we applied estimated baseline survival information using kidney transplant data obtained from the U.S. Organ Procurement and Transplantation Network (OPTN) (https://optn.transplant.hrsa.gov/data/). A total of 80,019 patients which included all adult patients who received kidney transplant between January 2005 and January 2013 with deceased donor type were used in the estima-

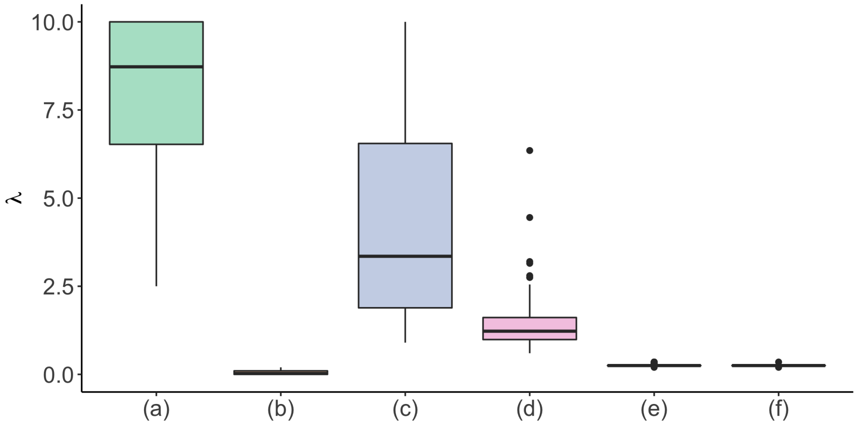

Figure 2: Selected tuning parameter λ for the best fitting of KL-based discrete modeling procedure. (a)-(f) represents model setting (a)-(f) respectively.

tion. Specifically, we fit a discrete relative risk model including the same set of predictors as the EPTS model and obtained the parameter estimates for each week within the first year after receiving transplants. Thus, our prior model was the combination of EPTS model and estimated baseline survival information by week.

Table 1: The log likelihood of models fitted for the real data example.

		Scenario 1		
Model	KL	Prior	Local	Stacked
Log likelihood	-358.47	-398.66	-395.27	-398.88
		Scenario 2		
Model	KL	Prior	Local	Stacked
Log likelihood	-358.45	-398.66	-409.47	-412.98

The current kidney transplant data we used was the University of Michigan Medical Center (MIUM) kidney transplant dataset. We considered two different scenarios regarding to predictors of current data: Scenario 1, which included the same set of predictors as EPTS model; and Scenario 2, which included two additional predictors of comorbidities (whether candidate has previous malignancy, and presence of pre-transplant peripheral vascular disease) than EPTS model. In this real data analysis, we applied log-log link function in the proposed modeling procedure. As shown in Table 1, KL-based discrete modeling procedure had the best performance under both scenarios. Specifically, using the same set of predictors, the model fitted by current data only achieved a slightly better performance than prior model, which indicates that the prior model lacks accuracy when applied to this specific current dataset. However, the log likelihood of the local model decreased substantially when including additional predictors, which shows that the model fitted by current dataset only is unstable. In summary, KL-based discrete modeling procedure provides a more stable and accurate prediction than other models.

7. Summary

In this paper, we proposed a Kullback-Leibler-based discrete modeling procedure and illustrated the performance of the proposed method by simulation studies and a real data example. We developed a time-dependent KL discrimination information which extends the classical KL discrimination information to time-to-event data, and defined a weighted likelihood function to link the prior information and the data from current study. Both simulation studies and the real data results showed advantages of the proposed modeling procedure. In addition, the real data example indicates that the proposed method also works well for situations when prior information and current data come from different underlying models. The proposed Kullback-Leibler-based discrete relative risk modeling procedure is sufficiently flexible to incorporate prior information from multiple data sources and improve prediction stability.

References

Y. Jiang, H. He, and H. Zhang. Variable selection with prior information for generalized linear models via the prior lasso method. *Journal of the American Statistical Association*, 111:355–376, February 2016.

S. Kullback and R.A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, March 1951.

R.E. Schapire, M. Rochery, M. Rahim, and N. Gupta. Boosting with prior knowledge for call classification. *IEEE transactions on speech and audio processing*, 13:174–181, March 2005.

Appendix A. Proof of Proposition 1

With P_1 corresponding to a historical discrete time-to-event model and P_2 corresponding to the discrete time-to-event model in the current dataset,

$$\begin{aligned}
D_{KL}(P_1, P_2; t_k, \mathbf{Z}_i) =& P_1(T_i = t_k | T_i \geq t_k, \mathbf{Z}_i) \log \left\{ \frac{P_1(T_i = t_k | T_i \geq t_k, \mathbf{Z}_i)}{P_2(T_i = t_k | T_i \geq t_k, \mathbf{Z}_i)} \right\} \\
&+ P_1(T_i > t_k | T_i \geq t_k, \mathbf{Z}_i) \log \left\{ \frac{P_1(T_i > t_k | T_i \geq t_k, \mathbf{Z}_i)}{P_2(T_i > t_k | T_i \geq t_k, \mathbf{Z}_i)} \right\} \\
=& \tilde{\lambda}(t_k; \mathbf{Z}_i) \log \left\{ \frac{\tilde{\lambda}(t_k; \mathbf{Z}_i)}{\lambda(t_k; \mathbf{Z}_i)} \right\} + \{1 - \tilde{\lambda}(t_k; \mathbf{Z}_i)\} \log \left\{ \frac{1 - \tilde{\lambda}(t_k; \mathbf{Z}_i)}{1 - \lambda(t_k; \mathbf{Z}_i)} \right\} \\
\propto& - \tilde{\lambda}(t_k; \mathbf{Z}_i) \log\{\lambda(t_k; \mathbf{Z}_i)\} - \{1 - \tilde{\lambda}(t_k; \mathbf{Z}_i)\} \log\{1 - \lambda(t_k; \mathbf{Z}_i)\} \\
=& - g(\tilde{\eta}_k + \mathbf{Z}_i^\top \widetilde{\boldsymbol{\beta}}) \log\{g(\eta_k + \mathbf{Z}_i^\top \boldsymbol{\beta})\} - \{1 - g(\tilde{\eta}_k + \mathbf{Z}_i^\top \widetilde{\boldsymbol{\beta}})\} \log\{1 - g(\eta_k + \mathbf{Z}_i^\top \boldsymbol{\beta})\} \\
=& - \left[\tilde{\delta}_i(t_k) \log \left\{ \frac{g(\eta_k + \mathbf{Z}_i^\top \boldsymbol{\beta})}{1 - g(\eta_k + \mathbf{Z}_i^\top \boldsymbol{\beta})} \right\} + \log\{1 - g(\eta_k + \mathbf{Z}_i^\top \boldsymbol{\beta})\} \right]
\end{aligned}$$