

Abstract

Many studies employ the analysis of time-to-event data that incorporates competing risks and right censoring. Most methods and software packages are geared towards analyzing data that comes from a continuous failure time distribution. However, failure-time data may sometimes be discrete either because time is inherently discrete or due to imprecise measurement. In this work, we introduce a **novel estimation procedure for discrete-time survival analysis with competing events**. The proposed approach offers two key advantages over existing procedures: first, it accelerates the estimation process; second, it allows for straightforward integration and application of widely used regularized regression and screening methods.

Methods

Notation and Definitions

 $i=1, \dots, N$ iid observations

 $Z^{p \times 1}$ - vector of p covariates

 $T \in \{1, \dots, d\}$ - event time

 $C \in \{1, \dots, d\}$ - censoring time

 $X = \min(T, C)$ - observed time

 $J \in \{1, \dots, M\}$ - event type

Logit-link event-specific hazard functions:

$$\lambda_j(t|Z) = \frac{\exp(\alpha_{jt} + Z^T \beta_j)}{1 + \exp(\alpha_{jt} + Z^T \beta_j)}$$

 Our goal is to estimate α_{jt} and β_j .

Lee et al. (2018) approach

Lee et al. (2018) expanded the data and use a generalized linear model to repeated binary outcomes.

 $\delta_{jit} = 1$ if $J=j$ occurred for i at $X=t$

Original Data				Expanded Data					
i	X_i	δ_i	Z_i	i	X_i	δ_{1it}	δ_{2it}	$1 - \delta_{1it} - \delta_{2it}$	Z_i
1	2	1	Z_1	1	1	0	0	1	Z_1
				1	2	1	0	0	Z_1
2	3	2	Z_2	2	1	0	0	1	Z_2
				2	2	0	0	1	Z_2
				2	3	0	1	0	Z_2
3	3	0	Z_3	3	1	0	0	1	Z_3
				3	2	0	0	1	Z_3
				3	3	0	0	1	Z_3

 The event-specific conditional likelihoods of the expanded data are denoted by $L_j^C(\beta_j)$

The Proposed Two-Step Approach

Step 1. Each β_j is estimated separately by maximizing L_j^C - a "likelihood" based on a conditional logistic regression, while stratifying the expanded data according to X and given the number of observed events within each stratum.

Step 2. Given the $\hat{\beta}_j$ s, solve $M \cdot d$ estimation equations:

$$\hat{\alpha}_{jt} = \operatorname{argmin}_a \left\{ \frac{1}{Y(t)} \sum_{i=1}^n I(X_i \geq t) \frac{\exp(a + Z_i^T \hat{\beta}_j)}{1 + \exp(a + Z_i^T \hat{\beta}_j)} - \frac{N_j(t)}{Y(t)} \right\}^2$$

 $Y(t)$ - risk set size at time t
 $N_j(t)$ - number of observed events of type j at time t

 Separated estimation of β_j and α_{jt} allows us to use, for example:

 1. **Penalized regression** in Lagrangian form by minimizing

$$-\log L_j^C(\beta_j) + \eta_j P(\beta_j)$$

 $P(\cdot)$ - a penalty function, $\eta_j > 0$ - a shrinkage tuning parameter.

 2. **Screening method** for Cox regression (Zhao and Li, 2012).

 Maximize L_j^C for each covariate, one at a time. The final model is the set of covariates whose absolute standardized estimated coefficients exceed a predetermined threshold.

Evaluation Measures

Let,

$$\pi_{ij}(t) = \widehat{\Pr}(T_i = t, J_i = j | Z_i)$$

$$D_{ij}(t) = I(T_i = t, J_i = j)$$

 $AUC_j(t)$ is defined as

 $\Pr(\pi_{ij}(t) > \pi_{mj}(t) | D_{ij}(t) = 1, D_{mj}(t) = 0, T_m \geq t)$ an integrated AUC_j

$$AUC_j = \sum_t AUC_j(t) w_j(t), \quad w_j(t) = \frac{N_j(t)}{\sum_t N_j(t)}$$

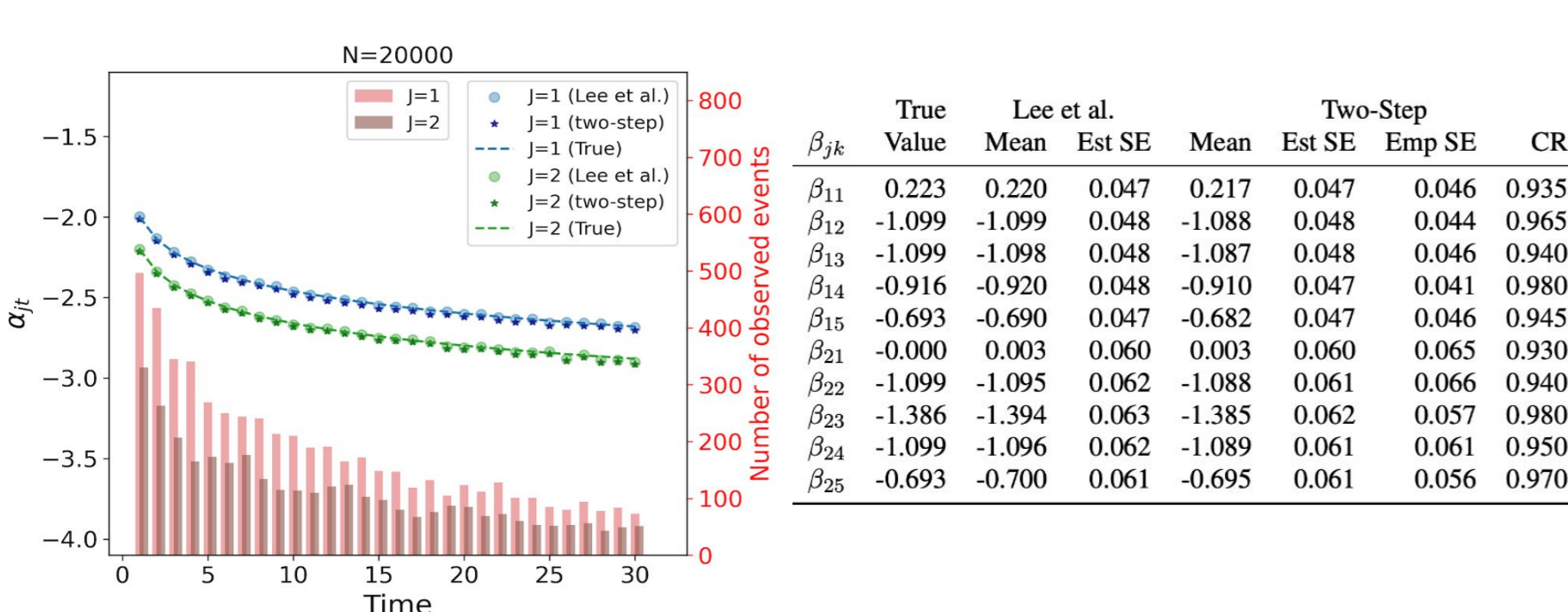
and a global AUC

$$AUC = \sum_j AUC_j v_j, \quad v_j = \frac{\sum_t N_j(t)}{\sum_{j=1}^M \sum_t N_j(t)}$$

Simulations

Estimation Performance Comparison

A comprehensive simulation study indicates that Lee et al. method and the proposed method perform similarly in terms of bias and standard errors. The empirical coverage rates of 95% Wald-type confidence intervals for each regression coefficient, are reasonably close to 95%. For example:



The Proposed Approach with LASSO

 $N = 20000$ observations, $p = 100$ covariates, $J = 2$ competing events.

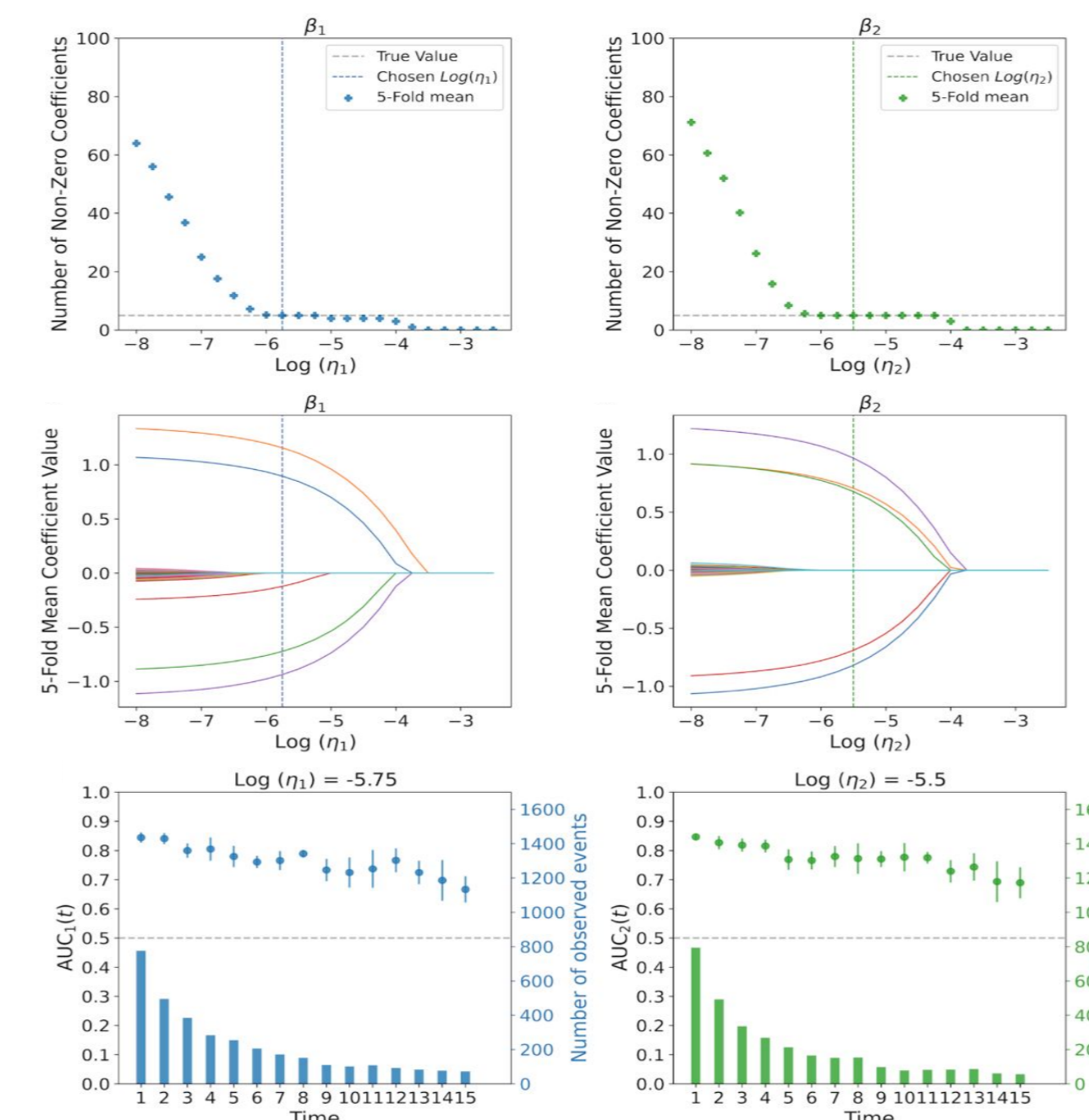
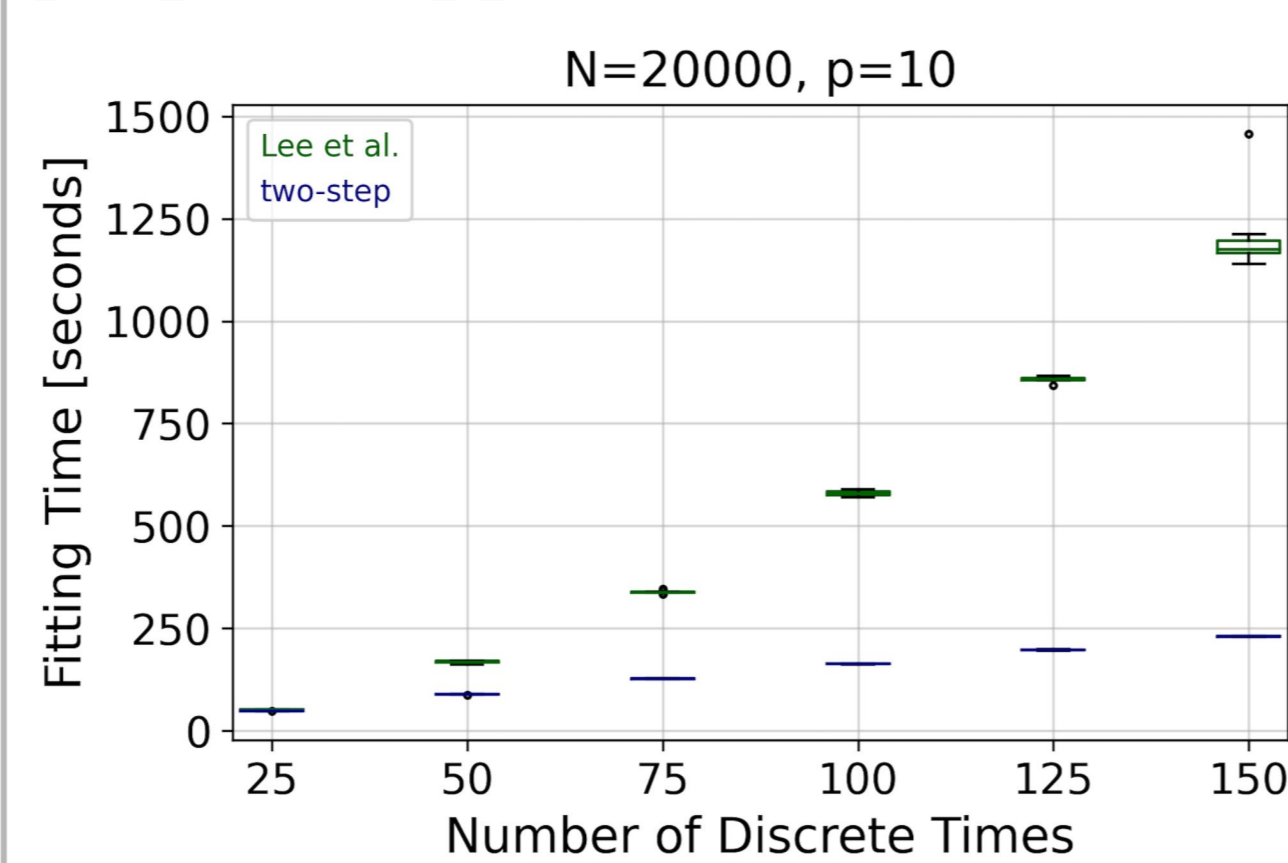
 The true model parameters β_{1k} and β_{2k} are non-zero only for $k = \{1, \dots, 5\}$.

 Regularization parameters η_1, η_2 were chosen using 5-fold CV and maximize global-AUC.

 At the optimal η_1, η_2 the correct non-zero parameters were identified.

Fitting Time Comparison

The mean fitting time for different numbers of discrete times d . As d increases, the advantage of the proposed approach increases as well.



Length of Hospital Stay in ICU

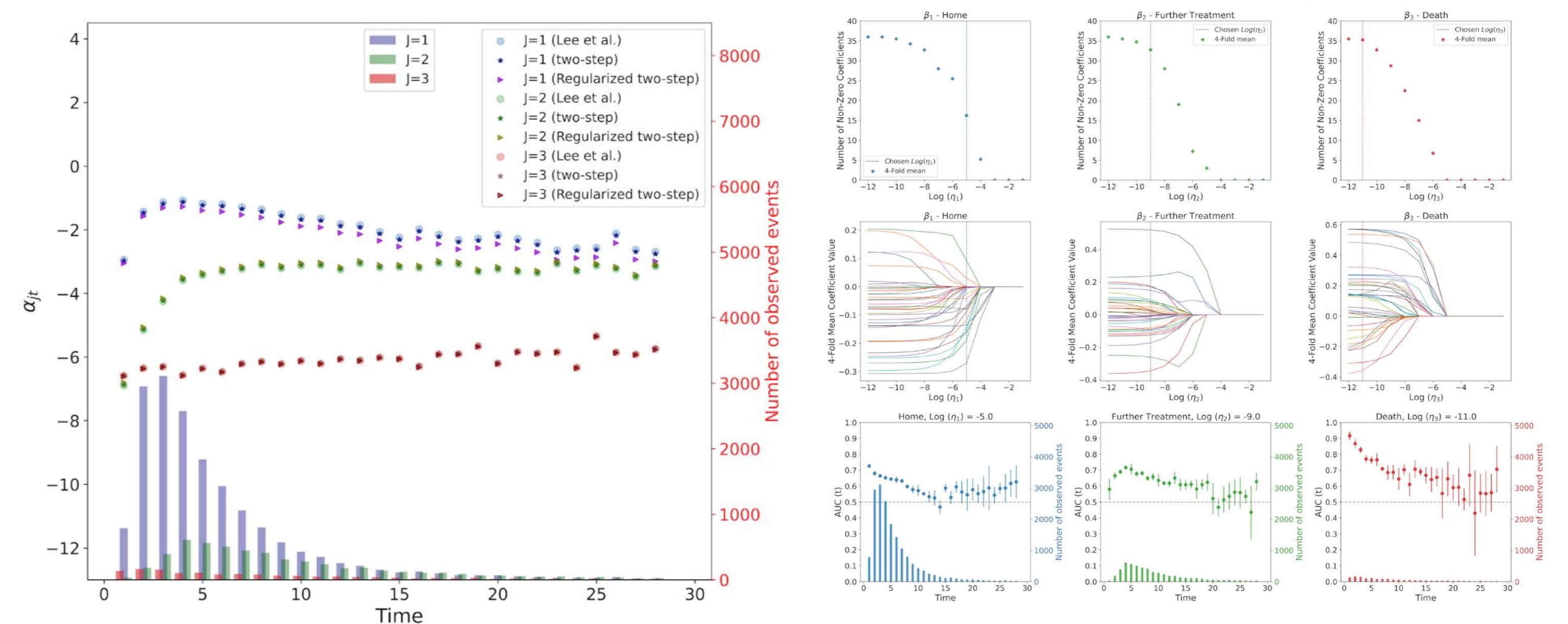
Goal: Estimating the length of stay (LOS) in days of patients hospitalized in the intensive care unit (ICU) upon arrival.

Dataset: $N = 25,170$ emergency admissions (MIMIC-IV) with $p = 36$ covariates (characteristics, lab tests from the first 24 hours)

Competing events: Home discharge ($J = 1$, 69.0%), further treatment elsewhere ($J = 2$, 21.4%), and in-hospital death ($J = 3$, 6.1%)

Censoring ($J = 0$): Left against medical advice (1%), LOS > 28 (2.5%).

Approaches: Lee et al. (2018), two-step, and two-step with LASSO.



Similar α_{jt}, β_j using Lee et al. and two-step without regularization.

With LASSO regularization the number of predictors β_j was reduced, but the corresponding α_{jt} remained highly similar.

Conclusion

This work provides a new estimation procedure for a semi-parametric logit-link survival model of discrete-time with competing events.

We show that the proposed approach:

- Performs well in terms of empirical bias and coverage rates.
- Significantly faster than existing methods, by separating the estimation of α_{jt} and β_j .
- Allows including modern machine-learning model-selection procedures, such as regularization and screening.

We provide PyDTS, an open-source Python software which implements our approach among other tools for discrete-time survival analysis.

References

- [1] Meir, Tomer and Gorfine, Malka, *Discrete-time Competing-Risks Regression with or without Penalization*, arXiv preprint, (2023).
- [2] Meir, Tomer* and Gutman, Rom* and Gorfine, Malka, *PyDTS: A Python Package for Discrete-Time Survival (Regularized) Regression with Competing Risks*, arXiv preprint, (2022).
- [3] Lee, Minjung and Feuer, Eric J. and Fine, Jason P., *On the analysis of discrete time competing risks data*, Biometrics, (2018).