A high dimensional mixture model for time-to-event data

Simon Bussy[1], Stephane Gaiffas[2], Agathe Guilloux[1,2] and Anne-Sophie Jannot[3]

[1]Theoretical and applied statistics laboratory, Pierre and Marie Curie University, Paris, France
[2]Ecole Polytechnique's Applied Mathematics Center, Paris, France
[3]INSERM 22^{nd} team, Georges-Pompidou European Hospital, Paris, France

Objectives

The main focuses will be to:

1. Introduce the censored mixture model for duration
2. Present the maximum likelihood techniques used for inference
3. Introduce the QNEM algorithm developed
4. Illustrate the method with a simulation study and on real datasets

Introduction

Based on right censored survival event time $T^c \in \mathbb{N}^*$ (for instance rehospitalization, relapse or death), and features $X \in \mathbb{R}^d$ corresponding to clinical data recorded during hospitalization, we want to construct a score for a patient by assessing his early event occurrence risk. The goal is first to construct this score for physicians that would help them to decide if a patient can be released or not from hospital, and second to study the effect of any covariates.

We consider a model with a binary latent variable $Z = 0$ or 1 for patients with low or high risk of early event occurrence respectively, that depends on clinical variables X.

For physicians, the variable Z can be viewed as the indicator that a patient should stay longer at the hospital or not. Conditionally on this latent state, we suppose that the time distribution before the next event is different, leading to a mixture of responses in the distribution of the duration before the next event

$$f_T(t) = \pi_\beta(x) f_0(t;\alpha_0) + (1-\pi_\beta(x)) f_1(t;\alpha_1)$$

with

$$\pi_\beta(x) = \mathbb{P}[Z=0|X=x] = \frac{1}{1+e^{-x^\top\beta}}$$

and $\beta \in \mathbb{R}^d$ being a vector of coefficients to estimate, that quantifies the impact of each covariates on the probability that patient belongs to the low-risk or the high-risk population.

A censored mixture model

In practice, we are dealing with censored data. To take into account this phenomenon, let's introduce the variable $C \in \mathbb{N}^*$ being the time when the individual leaves the target cohort. The survival variable T^c and the censoring indicator δ are then defined by

$$T^c = T \wedge C,$$
$$\delta = \mathbb{1}_{\{T \leq C\}}.$$

Then, under the hypothesis that T and C are conditionally independent given Z and X, and that C is independent of Z and X, one can derive the likelihood of the model $\ell_n(\theta)$ that we want to maximize, where $\theta = (\alpha_0, \alpha_1, \beta)$ are the parameters to infer.

Inference

In order to avoid overfitting and to improve the prediction power of our model, we use Elastic-Net regularization (Zou 2005), by minimizing the objective

$$-\ell_n(\theta) + \gamma\big((1-\eta)\|\beta\|_1 + \frac{\eta}{2}\|\beta\|_2^2\big). \qquad (1)$$

To handle this optimization problem, we will derive a novel generalized EM algorithm.

Then, depending on the chosen laws f_0 and f_1, the M-step could either be explicit for the updates of α_0 and α_1, or obtained using a minimization algorithm if not. The update for β requires the minimization of a convex problem, where we used the L-BGFS-B algorithm.

$$\ell_n(\theta) = n^{-1}\sum_{i=1}^n \log\Big[\big\{\pi_\beta(x_i) f_0(t_i^c;\alpha_0) + \big(1-\pi_\beta(x_i)\big) f_1(t_i^c;\alpha_1)\big\}\overline{G}(t_i^{c-})\Big]^{\delta_i}$$
$$\times \Big[\big\{\pi_\beta(x_i)\overline{F}_0(t_i^{c-};\alpha_0) + \big(1-\pi_\beta(x_i)\big)\overline{F}_1(t_i^{c-};\alpha_1)\big\} g(t_i^c)\Big]^{1-\delta_i}$$

Convergence of the QNEM algorithm

Under reasonable constraints on f_0 and f_1, every cluster point $\overline{\theta}$ of the sequence $\{\theta^{(l)}; l = 0,1,2,\dots\}$ generated by the QNEM algorithm is a stationary point of the criterion function in (1).

Simulation Study and results on real datasets

The following figures compare the performances of the 3 considered models in terms of $AUC(t)$ mean curves.

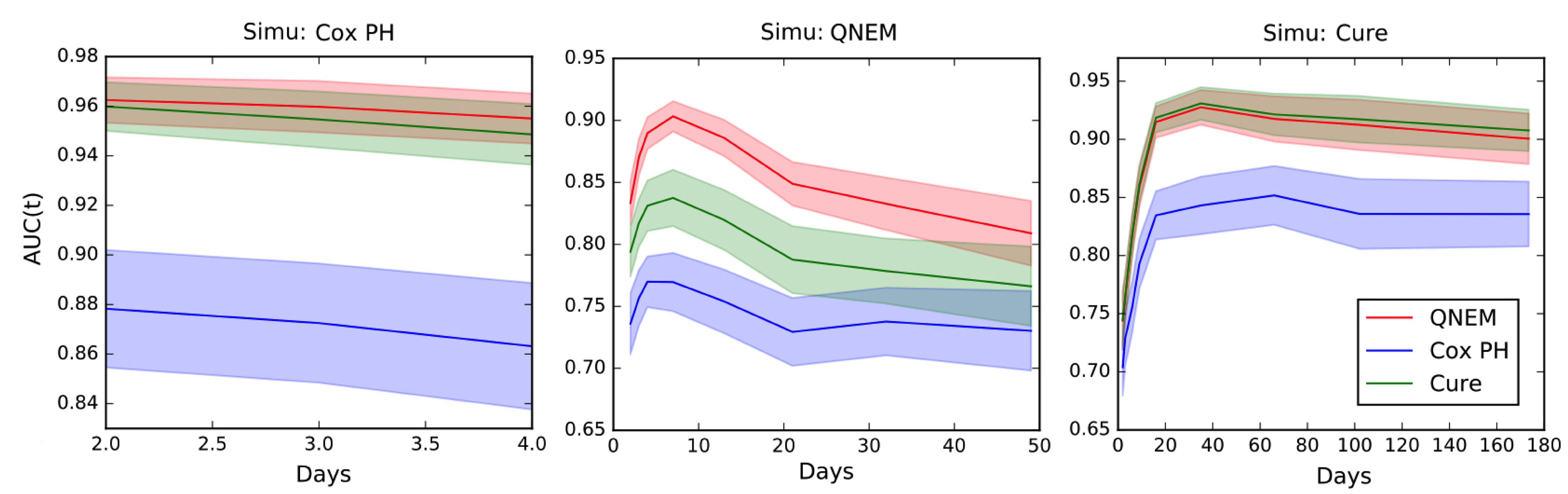

Figure 1: $AUC(t)$ mean curves comparisons after 100 consecutive simulations

C-index comparisons on two real datasets:

- Primary Biliary Cirrhosis (PBC) dataset: $(n = 312, d = 17)$
- Echocardiogram dataset: $(n = 130, d = 8)$

Models	PBC	Echocardiogram
QNEM	**0.874**	**0.774**
Cure	0.863	0.750
Cox PH	0.780	0.712

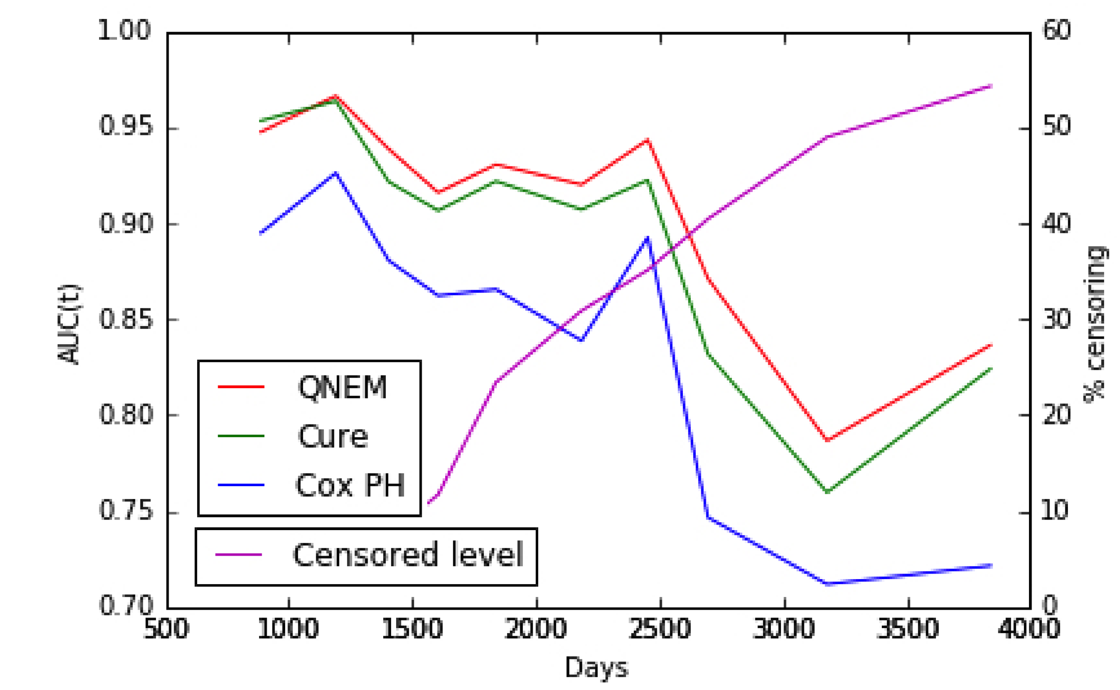

Figure 2: $AUC(t)$ curves comparisons on the PBC dataset

Conclusion

The proposed methodology gives better results than the state-of-the-art survival algorithms, namely the cure model (Farewell 1982) and the Cox PH model (Cox 1972), for multiple considered datasets. We also provide a robust implementation of the QNEM algorithm in high dimension.

References

Cox, D.R. (1972). Regression Models and Life-Tables (with Discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220.

Farewell, V.T . (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* **38**, 1041–1046.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* **67**, 301–320.

Contact Information

Simon Bussy, PhD Student
Email: simon.bussy@upmc.fr