Cox Mixture Model with High Dimensional Data to Predict Patients Lifespan

Eliz Peyraud, Julien Jacques, Guillaume Metzler

COX MIXTURE MODEL WITH HIGH DIMENSIONAL DATA TO PREDICT PATIENTS LIFESPAN

Eliz Peyraud, Julien Jacques, Guillaume Metzler

Univ Lyon, Univ Lyon 2, UR ERIC, France

MOTIVATIONS

Organ transplantation is the only therapeutic treatment available today to effectively treat terminal illnesses (failure, cancer, etc.).
The use of organ transplants is constantly on the increase, but results are still mixed when it comes to patient survival after transplantation.
The evaluation of transplant results can be analyzed to understand the factors of failure and success, and thus improve matching between donor and recipient.

BASIS FOR SURVIVAL ANALYSIS

Survival analysis is a commonly used method in the medical field, in particular when estimating a patient's lifespan following surgery.
Two dates are used to define survival time :

- Original date
- Last follow-up date.

Let t_i be the survival time of individual i. The probability for an patient of surviving to at least time T is then

$$S(t_i|x_i) = P(T > t_i).$$

CENSORING MECHANISM

For each individual i we define the binary variable δ_i such as:

$$\delta_i = \begin{cases} 0 \text{ if the individual is censored} \\ 1 \text{ otherwise.} \end{cases}$$

Correspond to the patient's exit from the follow-up program (before having been able to observe their potential death).

PROPORTIONAL HAZARD REGRESSION IN HIGH DIMENSION

The Cox Proportional hazards model is the most commonly used semi-parametric model to describe the phenomenon of events occurring over time, as a function of covariates:

$$\lambda(t, X) = \lambda_0(t) \exp(\beta^T X)$$

Estimation of β by maximum likelihood yields the survival function:

$$S(t \mid X) = \exp\left(-\int_0^t \lambda(u, X)\, du \exp(\beta^T X)\right).$$

In the case of high dimensionality, we seek to reduce the number of variables while maintaining the interpretability of the results. We therefore choose to apply an L1 penalty to the likelihood function of the Cox model.

$$l_{\text{pen}}(\beta) = l(\beta) - \eta \sum_{i=1}^{p} |\beta_i|$$

The strength of the penalty is determined by the η parameter, which can be set or optimized (usually by cross validation).

DATA STUDY

We focus on a database of 178 127 patients who have undergone a liver transplant in their lifetime. For each transplant, a wide range of information is available :

- health of the donor
- health of the recipient
- organ preservation

After pre-processing, this information represents 4139 mixed variables.

To perform the penalized Cox model, we divided the population into several subgroups of individuals with similar characteristics. The results, available in Table 1, show that depending on the population studied, the factors influencing post-operation survival differ. In a context of high heterogeneity, a unique model cannot be applied to the whole population.

MIXTURE OF COX MODEL

For a mixture with $g \in N$ components, the survival function is given by:

$$S(t \mid X) = \sum_{h=1}^{g} \pi_h S_h(t \mid X).$$

The likelihood of the mixture model is defined as the product of the density functions for uncensored individuals and the survival functions for censored individuals:

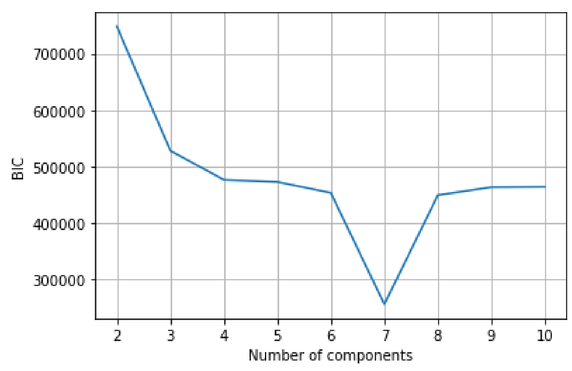
$$L(\boldsymbol{\beta}, \mathbf{x}, \mathbf{t}, \boldsymbol{\delta}) = \prod_{i=1}^{n} f(x_i, t_i, \beta)^{\delta_i} S(x_i, t_i, \beta)^{1-\delta_i},$$

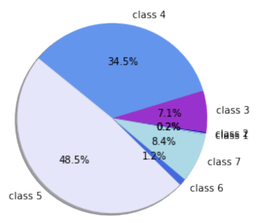
$$= \prod_{i=1}^{n} \left(\sum_{h=1}^{g} \pi_h f_h(x_i, t_i, \beta_h)\right)^{\delta_i} \times \left(\sum_{h=1}^{g} \pi_h S_h(x_i, t_i, \beta_h)\right)^{1-\delta_i}$$

where $(\mathbf{x}, \mathbf{t}, \boldsymbol{\delta}) = (x_1, ..., x_n, t_1, ..., t_n, \delta_1, ..., \delta_n)$ and $(\boldsymbol{\beta}) = (\beta_1, ..., \beta_g)$
An **EM algorithm** is then used to estimate all the model parameters (proportional hazard regression parameters and mixture proportions).

DATA STUDY 2

We are now looking to automatically identify patient clusters from a small number of variables with a mixture of Cox models.
The number of sub-clusters identified is chosen by BIC optimization criterion:





This choice allows us to classify individuals into 7 groups of respective proportions. Working with specialists, we can then identify similarities between patients in the same group and define new homogeneous populations.

PERSPECTIVES

We are currently working on a model that can handle both dimension reduction and data clustering. However, such a model represents a significant computational cost.
Moreover, we would ultimately like to have a model capable of prediction. For the moment, we can only predict the survival of a new patient if we have a priori knowledge of the cluster to which it belongs (which is not always the case). One solution would be to switch to an expert mixture model, which would enable us to estimate, at the same time as a patient's lifespan, his cluster belonging (in practice, the proportions of the mixture would become dependent on patient variables).

	Children	**Senior**	**Women**	**Men**	**Patients w/ cancer**
HLA status - immunocompatibility	×	×	×	×	×
Candidate's max weight		×	×	×	×
Infectious status - Antibody presence		×			×
Donor's age	×		×	×	
Graft preservation (organ rinsing)		×	×	×	×
Donor's West Nile test		×		×	

References

[1] Haiqun Lin et al. *Modeling survival data: extending the Cox model*. 2002.
[2] Shu Kay Ng et al. *Mixture modelling for medical and health sciences*. Chapman and Hall/CRC, 2019.
[3] Rosen O. et al. "Mixtures of proportional hazards regression models". In: *Statistics in Medicine* 18.9 (1999), pp. 1119–1131.
[4] Cox D. R. "Regression models and life-tables". In: *Journal of the Royal Statistical Society* 34.2 (1972), pp. 187–202.
[5] R. Tibshirani. "The lasso method for variable selection in the Cox model". In: *Statistics in medicine* 16.4 (1997), pp. 385–395.

Contact: eliz.peyraud@univ-lyon2.fr