MODEL-BASED CLUSTERING OF RIGHT-CENSORED LIFETIME DATA WITH FRAILTIES AND RANDOM COVARIATES

Andrea Cappozzo [1], Chiara Masci[1], Francesca Ieva[12] and Anna Maria Paganoni[1]

[1] MOX, Department of Mathematics, Politecnico di Milano, (e-mail: andrea.cappozzo@polimi.it, chiara.masci@polimi.it, francesca.ieva@polimi.it, anna.paganoni@polimi.it)

[2] Health Data Science Center, Human Technopole, Milano

ABSTRACT: We introduce a new parametric approach for clustering multilevel survival data that accounts for the heterogeneity at baseline and random distributions of the explanatory variables. The proposed method aims to identify clusters of patients with different survival patterns and uncover the impact of the known hierarchy on survival within each cluster. The objective function is maximized using a stochastic EM algorithm tailored to right-censored lifetime data. The proposed methodology can be seen as a generalization of multilevel cluster-weighted modeling for time-to-event outcomes. Promising results are showcased on synthetic data.

KEYWORDS: model-based clustering, survival data, frailty models, EM algorithm, cluster-weighted models

1 Introduction and model formulation

The paper proposes an approach for clustering survival data in which the procedure takes advantage of cluster-wise different random covariates. Additionally, the heterogeneity at the baseline due to a known hierarchy present in the sample (e.g., patients within hospitals) is accounted for in the time-to-event outcome by means of a parametric frailty model. In details, in our proposal a statistical unit is identified by the triplet $(y_{ij}, \delta_{ij}, \mathbf{x}_{ij})$ where:

- y_{ij} is the minimum between the survival time t_{ij} and censoring time c_{ij} for subject i in hospital j,
- $\delta_{ij} = I(t_{ij} \leq c_{ij})$ is the event indicator,
- $\mathbf{x}_{ij} = (\mathbf{u}_{ij}, \mathbf{v}_{ij})$ denotes the vector of covariates with $\mathbf{u}_{ij}$ and $\mathbf{v}_{ij}$ respectively indicating the subset of continuous and categorical predictors for the ij-th unit.

The entire sample is therefore composed by $N = \sum_{j=1}^{J} n_j$ observations among the J hospitals. We further assume that the observed data can be partitioned into G latent clusters independently of the known J groups. The resulting log-likelihood for the considered model reads as follows:

$$
\begin{aligned}
&\ell\left(\{\tau_g, \boldsymbol{\beta}_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \lambda_g, \theta_g\}_{g=1}^{G}\right) = \\
&\sum_{g=1}^{G} \Bigg\{ \sum_{j=1}^{J} \sum_{i \in R_{jg}} \log \tau_g + \sum_{j=1}^{J} \Bigg[\sum_{i \in R_{jg}} \delta_{ij} \left(\log h_0(y_{ij}) + \mathbf{x}_{ij}^{'} \boldsymbol{\beta}_g\right) + \\
&+ \log \Bigg[(-1)^{d_{jg}} \mathcal{L}^{(d_{jg})} \Bigg(\sum_{i \in R_{jg}} H_0(y_{ij}) \exp\left(\mathbf{x}_{ij}^{'} \boldsymbol{\beta}_g\right); \theta_g \Bigg) \Bigg] \Bigg] + \\
&+ \sum_{j=1}^{J} \sum_{i \in R_{jg}} \log \phi(\mathbf{u}_{ij}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) + \sum_{j=1}^{J} \sum_{i \in R_{jg}} \log \psi(\mathbf{v}_{ij}; \lambda_g) \Bigg\}.
\end{aligned}
\tag{1}
$$

The quantities $h_0(\cdot)$ and $H_0(\cdot)$ denote the baseline hazard and cumulative hazard functions, and $\mathcal{L}^{(q)}$ is the q-th derivative of the Laplace transform of the frailty distribution. Depending on the chosen baseline and/or frailty term, the formulation in (1) encompasses a general family of parametric mixture frailty models. With $\phi(\cdot)$ and $\psi(\cdot)$ we respectively identify the densities of a multivariate Gaussian and independent multinomial distributions (one for each categorical variable), needed to incorporate the cluster-wise different contribution of the covariates. Further, d_{jg} is the total number of observed events assigned to cluster g belonging to hospital j, and R_{jg} contains the indexes of the observations in cluster g and hospital j. The remaining terms are model parameters that need to be estimated from the sample. In details, τ_g represents the mixing proportion for cluster g, with $\tau_g \geq 0$ for all g and $\sum_{g=1}^{G} \tau_g = 1$. The vector of regression coefficients is denoted with $\boldsymbol{\beta}_g$, while θ_g is the heterogeneity parameter for $g = 1, \ldots, G$. Lastly, parameters for the conditionally independent multinomial distributions within each cluster are compactly identified with λ_g, and $\boldsymbol{\mu}_g$, $\boldsymbol{\Sigma}_g$ denote the mean vector and the covariance matrix of the continuous covariates.

Maximization of (1) is carried out by means of a stochastic EM algorithm tailored to right-censored lifetime data (Bordes & Chauveau, 2016). The proposed methodology extends the work in Berta & Vinciotti, 2019 by considering a time-to-event outcome, leveraging on recent advances in the efficient estimation of parametric frailty models (Munda *et al.*, 2012). To this extent, the goal of the proposed procedure is twofold. On the one hand, we aim to identify G clusters of patients with different survival patterns. On the other hand, within each cluster we wish to uncover the different impact the known hierarchy has on the survival. Promising results are reported for synthetic data, as described in the next section.

Table 1. *BIC and ARI for several choices of baseline, number of clusters and frailty densities in the Multilevel time-to-event cluster-weighted model.*

G	Baseline	Frailty	BIC	ARI
2	Exponential	None	-3795.60	0.84
2	Exponential	Gamma	-3756.93	0.84
2	Weibull	None	-3279.04	0.93
2	Weibull	Gamma	**-2586.46**	**0.95**
3	Exponential	None	-3767.83	0.73
3	Exponential	Gamma	-3597.16	0.67
3	Weibull	None	-3193.45	0.82
3	Weibull	Gamma	-2810.98	0.80

2 Results on simulated data

We assess the performance of the proposed procedure on a two components ($G = 2$) synthetic population simulated with the `genfrail` function of the `frailtySurv R` package (Monaco *et al.*, 2018). The data generating process includes $n_j = 40$ for all $j = 1,\ldots,J$ and $J = 10$, resulting in a sample whose size is equal to $N = 400$. The baseline hazard has a parametric Weibull distribution, while a Gamma density is used to simulate the frailty term in the equally sized clusters. The survival time depends on two continuous covariates, whose distribution is multivariate Gaussian with cluster-wise different mean vectors and equal covariance matrix. Model results are reported in Table 1 in which several specifications for the baseline and frailty densities are considered. The comparison includes also an option with fixed effects only, denoted with Frailty equals to None in the table. We observe that the model with Weibull baseline, Gamma frailty and true number of clusters outperforms the competing methods in both goodness of fit and clustering performance, showcasing higher values in both Bayesian Information Criterion and Adjusted Rand Index metrics.

3 Conclusion

The proposed approach provides a flexible method for analyzing right-censored lifetime data with random covariates and frailties, making it a valuable tool for applications in personalized medicine and hospitals evaluation. Some analyses are currently being accomplished on this regard and they will be the object of future work.

References

Berta, Paolo, & Vinciotti, Veronica. 2019. Multilevel logistic cluster-weighted model for outcome evaluation in health care. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **12**(5), 434–443.

Bordes, Laurent, & Chauveau, Didier. 2016. Stochastic EM algorithms for parametric and semiparametric mixture models for right-censored lifetime data. *Computational Statistics*, **31**(4), 1513–1538.

Monaco, John V., Gorfine, Malka, & Hsu, Li. 2018. General Semi-parametric Shared Frailty Model: Estimation and Simulation with frailty-Surv. *Journal of Statistical Software*, **86**(4).

Munda, Marco, Rotolo, Federico, & Legrand, Catherine. 2012. parfm : Parametric Frailty Models in R. *Journal of Statistical Software*, **51**(11).