# Frailty multi-state models using maximum penalised partial likelihood estimation

Federico Rotolo and Catherine Legrand

**Abstract** We propose a way to incorporate frailties into multi-state models to account for clustering in complex survival data. Estimation is possible via maximum penalized partial likelihood. This model can both account for risk differences between groups and model the risk of different events simultaneously.

**Key words:** survival analysis, frailty models, multi-state models, partial likelihood

## 1 Introduction

The focus of many clinical trials is on the time to some event of interest and on how given factors can accelerate or delay the process under study. In such contexts, the Cox proportional hazards model [2] is one of the most popular regression models, and many extensions have been developed for more complex problems. Multi-state models allow a deep understanding of duration data in the presence of several endpoints, linked to each other [7]. Frailty models [3, 10] allow to analyse duration data accounting for dependence among clustered data, e. g. in multicenter clinical trials with patients from several hospitals.

The integration of frailties into multi-state models can allow to account for clustering in the framework of multi-state structures. Some applied works have appeared in recent years [12, 1, 5], while theoretical research is still taking its first steps [8].

We propose a way to incorporate frailties into Markov multi-state models. Transforming data from wide to long format [6] permits to express the likelihood as that

Federico Rotolo, e-mail: federico.rotolo@stat.unipd.it
Department of Statistical Sciences, University of Padova
Via Cesare Battisti 241, 35121 Padova, Italy.

Catherine Legrand, e-mail: catherine.legrand@uclouvain.be
Institut de Statistique, Biostatistique et Sciences Actuarielles, Université catholique de Louvain
Voie du Roman Pays, 20, 1348 Louvain-la-Neuve, Belgium.

of a frailty model with stratification and left truncation. A semiparametric estimation approach is then possible, based on the EMPL procedure [4] by Horny, extending maximum penalized partial likelihood estimation [9].

## 2 Background

### Frailty models

Consider $n$ subjects, grouped in $H$ clusters of sizes $n_h$, with $\sum_{h=1}^{H} n_h = n$. For each subject $i$ in cluster $h$, the time variable $T_{hi}$ contains the event or censoring time; $\mathsf{N}_{hi}(t)$ is the counting process, being 0 until the time of the event and 1 afterwards; $\mathsf{Y}_{hi}(t)$ is the at-risk stochastic process, that is 1 if the subject is at risk at time $t^-$ and 0 otherwise.

The conditional hazard of subject $i$ in cluster $h$, with covariates vector $\mathbf{x}_{hi}$, is then

$$\lambda_{hi}(t) = u_h \lambda_0(t) \exp\{\mathbf{x}_{hi}^\top \boldsymbol{\beta}\}, \qquad h = 1,\ldots,H, \quad i = 1,\ldots,n_h, \tag{1}$$

with $\lambda_0(\cdot)$ the baseline hazard, $\boldsymbol{\beta}$ the regression coefficients, $u_h$ the frailty term, unobservable realisation of $U_h \overset{iid}{\sim} f_U(\cdot;\boldsymbol{\theta})$.

In a semiparametric estimation approach, no parametric form is assumed for the baseline risk and the full loglikelihood is profiled w.r.t. it. The so-obtained penalized partial loglikelihood, expressed in terms of counting processes, is

$$\ell_{\mathsf{PP}}(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{u}) = \sum_{h=1}^{H} \Bigg\{ \log f_U(u_h;\boldsymbol{\theta}) + \sum_{i=1}^{n_h} \int_0^\infty \Big[\log u_h + \mathbf{x}_{hi}^\top \boldsymbol{\beta} - \log \mathsf{Y}^{(0)}(\boldsymbol{\beta},\mathbf{u},t)\Big]\, \mathrm{d}\mathsf{N}_{hi}(t) \Bigg\}, \tag{2}$$

with $\mathsf{Y}^{(0)}(\boldsymbol{\beta},\mathbf{u},t) = \sum_{k=1}^{H} \sum_{j=1}^{n_h} u_k \exp\{\mathbf{x}_{kj}^\top \boldsymbol{\beta}\} \mathsf{Y}_{kj}(t)$ the weighted risk set at time $t$ and where $\mathbf{u} = (u_1,\ldots,u_H)$ is the vector of the frailties. Then, maximum penalized partial likelihood estimation [9] is obtained by alternating the maximisation of (2) on $\boldsymbol{\beta}$ and $\mathbf{u}$ for fixed $\boldsymbol{\theta}$ and the maximisation on $\boldsymbol{\theta}$ for fixed $\boldsymbol{\beta}$ and $\mathbf{u}$.

### Multi-state models

Markov multi-state models with proportional hazards can be expressed in terms of the transition-specific hazard for transitions of type $q$, subject $i$

$$\lambda_{qi}(t) = \lambda_{q0}(t) \exp\{\boldsymbol{\beta}^\top \mathbf{x}_{qi}(t)\}, \qquad q = 1,\ldots,Q, \quad i = 1,\ldots,n \tag{3}$$

with $\lambda_{q0}(t)$ the baseline hazard for transitions of type $q$, $\boldsymbol{\beta}$ the vector of the stacked transition-specific coefficient vectors $\boldsymbol{\beta}_q$, and $\mathbf{x}_{qi}(t)$ a vector of transition-specific covariates derived from the covariates vector $\mathbf{x}_i$ by transforming data from wide to long format [6].

For each transition type $q \in \{1,\dots,Q\}$ and each subject $i \in \{1,\dots,n\}$, $\mathsf{N}_{qi}(t)$ is the counting process being 0 at the beginning and increasing by 1 whenever the subject experiences an event of type $q$; $\mathsf{Y}_{qi}(t)$ is the at-risk stochastic process, being 1 if the subject is at risk of event $q$ at time $t^-$, 0 otherwise.

Once multi-state data are transformed into long format, the model can be treated as a Cox model stratified on transition types and with left truncation at the times of entering each state. Then, semiparametric inference for regression parameters is based on the partial likelihood, i.e. the likelihood profiled w.r.t. the baseline hazards $\lambda_{10}(\cdot),\dots,\lambda_{Q0}(\cdot)$ [11]. For the multi-state model (3), the partial loglikelihood is

$$\ell_{\mathsf{P}}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{q=1}^{Q}\int_0^\infty \left[\boldsymbol{\beta}^\top \mathbf{x}_{qi}(t) - \log \mathsf{Y}_q^{(0)}(\boldsymbol{\beta},t)\right] \mathrm{d}\mathsf{N}_{qi}(t), \tag{4}$$

with $\mathsf{Y}_q^{(0)}(\boldsymbol{\beta},t) = \sum_{i=1}^{n} \exp\{\boldsymbol{\beta}^\top \mathbf{x}_{qi}(t)\}\mathsf{Y}_{qi}(t)$ the weighted risk set at time $t$.

## 3 Multi-state models with nested frailties

In order to account for clustering in a multi-state framework, we consider now a multi-state model with the addition of two frailties, one ($v_h$) for the cluster and one ($w_{qh}$) for the transition type within the cluster. We define the conditional transition-specific hazard as

$$\lambda_{qhi}(t) = v_h w_{qh} \lambda_{q0}(t) \exp\{\boldsymbol{\beta}^\top \mathbf{x}_{qhi}(t)\}, \qquad q=1,\dots,Q, \quad h=1,\dots,H, \quad i=1,\dots,n_h, \tag{5}$$

with $u_{qh} = v_h w_{qh}$ and assuming that

$$V_h \overset{iid}{\sim} f_V(v;\theta_V), \qquad W_{qh} \overset{iid}{\sim} f_W(w;\theta_q) \qquad \text{and} \qquad V_h \perp\!\!\!\perp W_{qj},\ \forall q,h,j.$$

Then, with $U_{qh} = V_h W_{qh}$, we have that

$$\begin{aligned} &U_{qh} \perp\!\!\!\perp U_{kj},\ h \neq j,\\ &U_{qh} \not\perp\!\!\!\perp U_{kh},\ \text{with } \mathsf{Cor}(U_{qh},U_{kh}) > 0. \end{aligned}$$

By analogy to multi-state models, the partial loglikelihood $\ell_{\mathsf{P}}(\boldsymbol{\beta};\mathbf{v},\mathbf{w})$ can be obtained, conditionally on the frailties $\mathbf{v} = (v_1,\dots,v_H)^\top$ and $\mathbf{w} = (v_{11},\dots,w_{QH})^\top$, once the data are transformed into long format. Then, as in the case of semi-parametric frailty models, the penalised partial loglikelihood is $\ell_{\mathsf{PP}}(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{v},\mathbf{w}) = \ell_{\mathsf{P}}(\boldsymbol{\beta};\mathbf{v},\mathbf{w}) + \log f_{\mathbf{V}}(\mathbf{v};\theta_V) + \log f_{\mathbf{W}}(\mathbf{w};\theta_1,\dots,\theta_Q)$. For the model (5) it is given by

$$\ell_{\mathsf{PP}}(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{v},\mathbf{w}) = \sum_{h=1}^{H}\Bigg\{\log f_V(v_h;\theta_V) + \sum_{q=1}^{Q}\Bigg\{\log f_W(w_{qh};\theta_q) + \\ \sum_{i=1}^{n_h}\int_0^\infty \Big[\log\big(v_h w_{qh}\big) + \boldsymbol{\beta}^\top \mathbf{x}_{qhi}(t) - \log \mathsf{Y}_q^{(0)}(\boldsymbol{\beta},\mathbf{v},\mathbf{w},t)\Big]\,\mathrm{d}\mathsf{N}_{qhi}(t)\Bigg\}\Bigg\}, \quad (6)$$

where $\mathsf{Y}_q^{(0)}(\boldsymbol{\beta},\mathbf{v},\mathbf{w},t) = \sum_{k=1}^{H} v_k w_{qk} \sum_{j=1}^{n_k} \exp\{\boldsymbol{\beta}^\top \mathbf{x}_{qkj}(t)\}\mathsf{Y}_{qkj}(t)$ is the weighted risk set at time $t$, conditional on the frailties. This likelihood corresponds to the penalised partial loglikelihood of a nested frailty model with stratification and left truncation; strata are the transition types and truncation times are the times of entering each state. Then, semiparametric estimation can then be performed via EMPL algorithm proposed by [4] for multilevel frailty models.

**Acknowledgements** The research work was done during a visit of FR at the *ISBA*, *Université Catholique de Louvain*. FR acknowledges financial support from grant 60A15-3814/11, University of Padova, CL from IAP research network P6/03 of the Belgian Government (Belgian Science Policy), and from the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique', granted by the 'Académie universitaire Louvain'.

## References

1. Bhattacharyya, M., Klein, J.P.: A random effects model for multistate survival analysis with application to bone marrow transplants. Mathematical Biosciences **194**(1), 37–48 (2005)
2. Cox, D.R.: Regression models and life-tables. Journal of the Royal Statistical Society. Series B **34**(2), 187–220 (1972)
3. Duchateau, L., Janssen, P.: The frailty model. Springer (2008)
4. Horny, G.: Inference in mixed proportional hazard models with K random effects. Statistical Papers **50**(3), 481–499 (2009)
5. Ma, T., Joly, I., Raux, C.: A shared frailty semi-parametric markov renewal model for travel and activity time-use pattern analysis. In: Proceedings of the $12^{th}$ WCTR – Lisbon (2010)
6. Putter, H.: Tutorial in biostatistics: Competing risks and multi-state models, analyses using the `mstate` package. Companion file for the `mstate` package (2011)
7. Putter, H., Fiocco, M., Geskus, R.B.: Tutorial in biostatistics: competing risks and multi-state models. Statistics in Medicine **26**(11), 2389–430 (2007)
8. Putter, H., van Houwelingen, H.C.: Frailties in multi-state models: Are they identifiable? do we need them? Statistical Methods in Medical Research (2011)
9. Therneau, T.M., Grambsch, P.M., Pankratz, V.S.: Penalized survival models and frailty. Journal of Computational and Graphical Statistics **12**(1), 156–175 (2003)
10. Wienke, A.: Frailty Models in Survival Analysis. Taylor and Francis (2010)
11. de Wreede, L., Fiocco, M., Putter, H.: The `mstate` package for estimation and prediction in non- and semi-parametric multi-state and competing risks models. Computer Methods and Programs in Biomedicine **99**(3), 261–74 (2010)
12. Yen, A., Chen, T., Duffy, S., Chen, C.: Incorporating frailty in a multi-state model: application to disease natural history modelling of adenoma-carcinoma in the large bowel. Statistical Methods In Medical Research **19**(5), 529–546 (2010)