Competing risks models with two time scales

Angela Carollo[1,2], Hein Putter[2], Paul Eilers[3], Jutta Gampe[1]

[1] Max Planck Institute for Demographic Research, Rostock, Germany
[2] University of Leiden Medical Center, Leiden, The Netherlands
[3] Erasmus University Medical Center, Rotterdam, The Netherlands

E-mail for correspondence: `carollo@demogr.mpg.de`

Abstract: Competing risks models can involve more than one time scale. We propose a model for competing events in which the cause-specific hazards vary smoothly over two times scales. We estimate these two-dimensional hazard functions by P-splines, exploiting the equivalence between hazard smoothing and Poisson regression. As the data are arranged on a grid we can make use of generalized linear array models (GLAM) for efficient computations. We present an application to the study of transitions out of non-marital cohabitation in Germany.

Keywords: Cause-specific hazards; Two-dimensional smoothing; P-splines; GLAM

1 Introduction

Competing risks describe the situation where individuals are at risk of experiencing one of several types of events (Putter et al., 2007). The prototype of a competing risks model is the study of cause-specific mortality. For example, in clinical studies of cancer it is common to analyse mortality from cancer and mortality due to other causes. But competing events are also present in demographic studies: A marriage can end either in divorce or in widowhood, a non-marital cohabitation ends by marriage or separation.
Time is a key quantity in any event history analysis, and it can be recorded over several time scales. For example, after a cancer diagnosis, the risk of death might be studied over time since diagnosis, over age, which is time since birth, or over time since treatment. In demography, age-specific rates are the most common choice, but other time scales, like time since marriage, are often relevant too. All time scales progress at the same speed and differ only in their origin.

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

The key quantities in a competing risks analysis are the cause-specific hazards. They are defined as the instantaneous risk of experiencing an event of a specific type at time t, given no event (of any type) has happened yet. From them the overall survival function, i.e., the probability of no event up to time t, and the cumulative incidence functions, the probability of an event of given type before t, can be derived.

Usually, cause-specific hazards are defined for the same single time scale, and little research has been done on how to handle multiple time scales in competing risks models. Lee et al. (2017) propose a model where two competing causes are modelled on two different time scales and then combined under one of the two scales to estimate the cumulative incidence functions. Carollo et al. (2022) propose a model for a single event where the hazard varies smoothly over two time scales and is estimated by tensor products P-splines. Here we develop this model further for a competing risks setting. Each cause-specific hazard varies over two time scales, and estimation again is achieved by bivariate P-splines smoothing. Therefrom, we calculate the cumulative incidence functions for each cause.

As an application we study transitions out of cohabitation, either into marriage or into separation, for women living in West Germany.

2 A competing risks model with two time scales

Consider individuals in non-marital cohabiting unions. At each point in time a cohabiting individual is at risk of marrying (event 1) or of separating from the current partner (event 2). These transitions can be studied along two time scales, namely $t =$ age and $s =$ duration of cohabitation. A graphical depiction of this process in presented in Figure 1.

The cause-specific hazards for event type $\ell \in \{1, 2\}$ over the two time scales t and s are defined as

$$\lambda_\ell(t,s) = \lim_{\Delta \downarrow 0} \frac{P\,(\ \text{event}_\ell \in \{t + k\Delta, s + k\Delta : 0 \leq k \leq 1\}\,|\,\text{no event before } (t,s)\ )}{\Delta}.$$

Here $t > s$, so the two-dimensional hazards $\lambda_\ell(t,s)$ are only defined in the lower half-open triangle of $\mathbb{R}^2_+$.

The two time scales differ in their origin, which in this example is the age t_0 when the individual enters the cohabitation. This age differs between individuals and in a Lexis diagram individuals move along 45°-lines from $(t_0, 0)$ to $(t_0 + v, v)$ until they leave the risk set (due to event or censoring). This allows to view cause-specific hazards equivalently as two-dimensional functions $\tilde{\lambda}_\ell(u,s)$, where $\tilde{\lambda}_\ell(u = t-s, s) = \lambda_\ell(t,s)$. The $\tilde{\lambda}_\ell(u,s)$ are defined over the full positive quadrant $\mathbb{R}^2_+$.

From the cause-specific hazards $\lambda_\ell(t,s)$ or $\tilde{\lambda}_\ell(u,s)$, respectively, we obtain the cumulated cause-specific hazards

$$\Lambda_\ell(t,s) = \int_0^s \lambda_\ell(t_0 + v, v)\,dv \qquad \text{or} \qquad \tilde{\Lambda}_\ell(u,s) = \int_0^s \tilde{\lambda}_\ell(u,v)\,dv,$$

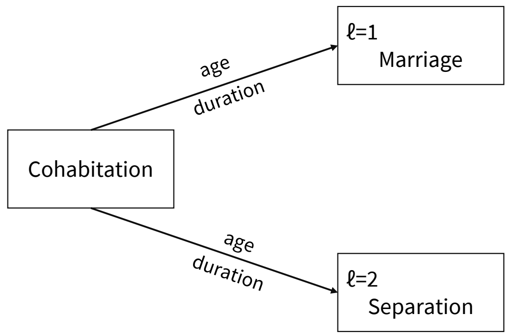

FIGURE 1. Competing risks process for the transitions out of cohabitation by age and duration of the cohabitation.

and the overall survival function

$$S(t,s) = \exp\left\{-\sum_{\ell=1}^{2} \Lambda_\ell(t,s)\right\} \qquad \text{or} \qquad \tilde{S}(u,s) = \exp\left\{-\sum_{\ell=1}^{2} \tilde{\Lambda}_\ell(u,s)\right\},$$

where $u = t - s$. The cumulative incidence functions (CIF) are

$$I_\ell(t,s) = \int_0^s \lambda_\ell(t_0 + v, v)\, S(t_0 + v, v)\, dv \ ; \ \tilde{I}_\ell(u,s) = \int_0^s \tilde{\lambda}_\ell(u,v)\, \tilde{S}(u,v)\, dv.$$

In the context of our application, $\tilde{I}_\ell(u,s)$ is the cumulative probability of marriage ($\ell = 1$) or separation ($\ell = 2$) within s years after start of cohabitation for a subject who entered cohabitation at age u.
To estimate the cause-specific hazard surfaces $\tilde{\lambda}_\ell(u,s)$ we divide the (u,s)-plane into $J \times K$ small bins (squares) and within each bin we count the number of events of type ℓ, denoted by $y_{jk}^{(\ell)}$, and the exposure times r_{jk}. The $y_{jk}^{(\ell)}$ are assumed to be realizations of Poisson variates with means

$$\tilde{\mu}_{jk}^{(\ell)} = r_{jk} \cdot \tilde{\lambda}_{jk}^{(\ell)} = r_{jk} \cdot \exp\{\tilde{\eta}_{jk}^{(\ell)}\}. \tag{1}$$

The $\tilde{\lambda}_{jk}^{(\ell)}$ represent the cause-specific hazard $\tilde{\lambda}_\ell(u,s)$ evaluated at the center of bin (j,k).
The log-hazards $\tilde{\eta}_\ell(u,s)$ are assumed to be smooth functions and they are modelled as sums of tensor products of B-splines. Difference penalties on the coefficients, one in the row- and one in the column-direction, ensure smoothness of the estimates (Currie et al., 2004).

Due to the binning the data, event counts and at-risk times, are on a regular grid and are naturally arranged as $J \times K$ matrices

$$Y_\ell = [y_{jk}^{(\ell)}] \qquad \text{and} \qquad R = [r_{jk}].$$

Correspondingly, we denote $M_\ell = [\tilde{\mu}_{jk}^{(\ell)}]$ and $E_\ell = [\tilde{\eta}_{jk}^{(\ell)}]$. The tensor products are formed from two marginal B-splines bases of size m and $\breve{m}$, respectively, along the u- and s-axis. The two basis matrices are denoted by B, which is $J \times m$, and by $\breve{B}$, which is $K \times \breve{m}$. The coefficients are arranged in the matrix $A_\ell = [\alpha_{fg}^{(\ell)}]$, where $f = 1, \dots, m$ and $g = 1, \dots, \breve{m}$.
The log-hazard can then be expressed in a compact way as $E_\ell = B \cdot A_\ell \cdot \breve{B}^{\mathrm{T}}$. The penalty on the coefficients in A_ℓ is constructed from two matrices D and $\breve{D}$ that form differences (of order d) of the columns of a matrix and it is controlled by two smoothing parameters ρ and $\breve{\rho}$ to allow anisotropic smoothing:

$$\mathrm{pen}(\rho, \breve{\rho}) = \rho \, ||DA_\ell||_{\mathrm{F}}^2 + \breve{\rho} \, ||A_\ell \breve{D}^{\mathrm{T}}||_{\mathrm{F}}^2$$

($||.||_{\mathrm{F}}^2$ represents the sum of all squared elements of a matrix.)
The objective function to minimize resulting from (1) is

$$\begin{aligned} \mathrm{Q}_\ell &= dev(M_\ell; Y_\ell) + \mathrm{pen}(\rho, \breve{\rho}) \\ &= 2 \sum_{j=1}^{J} \sum_{k=1}^{K} \left(y_{jk}^{(\ell)} \ln(y_{jk}^{(\ell)} / \tilde{\mu}_{jk}^{(\ell)}) - (y_{jk}^{(\ell)} - \tilde{\mu}_{jk}^{(\ell)}) \right) + \mathrm{pen}(\rho, \breve{\rho}) \end{aligned}$$

which leads to normal equations that can be solved, for given ρ and $\breve{\rho}$, in a penalized Poisson IWLS scheme (in compact notation):

$$\left[(\breve{B} \otimes B)^{\mathrm{T}} W_\ell' (\breve{B} \otimes B) + P \right] \alpha_\ell' = (\breve{B} \otimes B)^{\mathrm{T}} W_\ell' z_\ell'.$$

Here, $P = \rho(\breve{I} \otimes D^{\mathrm{T}} D) + \breve{\rho}(\breve{D}^{\mathrm{T}} \breve{D} \otimes I)$, with I and $\breve{I}$ identity matrices of appropriate dimension, W_ℓ is a diagonal matrix of weights, α_ℓ is the vector of coefficients, z_ℓ is the working variable and the prime symbol indicates the current value in the iteration.
The special format of Y_ℓ, R and E_ℓ allows to employ generalized linear array methods (Currie et al., 2006) for efficient computations. We determine the optimal values for the smoothing parameters by numerically optimizing the AIC of the model as a function of $(\rho, \breve{\rho})$.
Once the coefficients $\hat{A}_\ell = [\hat{\alpha}_{fg}^{(\ell)}]$ are obtained, we can evaluate the estimated $\tilde{\eta}_\ell(u, s)$ and hence also the cause-specific hazards $\tilde{\lambda}_\ell(u, s)$ on a detailed grid. Consequently, the cumulative hazards $\tilde{\Lambda}_\ell(u, s)$, the overall survival function $\tilde{S}(u, s)$ and the CIF $\tilde{I}_\ell(u, s)$ can be obtained by simple numerical integration (rectangle or trapezoid rule) with sufficient accuracy.

3 Application: transitions out of cohabitation

We applied this approach to study transitions out of cohabitation, either to marriage or to separation, for West German women over age and length of the cohabitation. The data come from the 11$^{\text{th}}$ wave of the German Family Panel (pairfam). The pairfam is a longitudinal panel study for researching issues of family and relationships dynamics in Germany (Huinink et al., 2011).

For this application we selected all women living in West Germany who at the time of their first interview were already in a non-marital cohabiting union, or have entered one at some time during the study period (2008-2019). We follow the trajectories of these cohabiting unions from either the start of the cohabitation, or the time of the first interview, until marriage, separation or end of the follow-up period. For each individual in the sample we know the age at beginning of the cohabitation, the age at the event or censoring time, and the duration of the cohabitation (which is left truncated if the cohabiting union was already formed by the time the individual entered the study).

We first estimate the cause-specific hazards along age and duration of the cohabitation for West German women by dividing the transformed positive quadrant into a grid of 62 by 54 bins. Then, we compute the cubic marginal B-splines by placing a knot every 3 bins circa, resulting in 20 and 17 B-splines on the u and s axis respectively.

The cause-specific hazards estimated from the data, given the optimal smoothing parameters, are represented in the transformed positive quadrant in Figure 2.

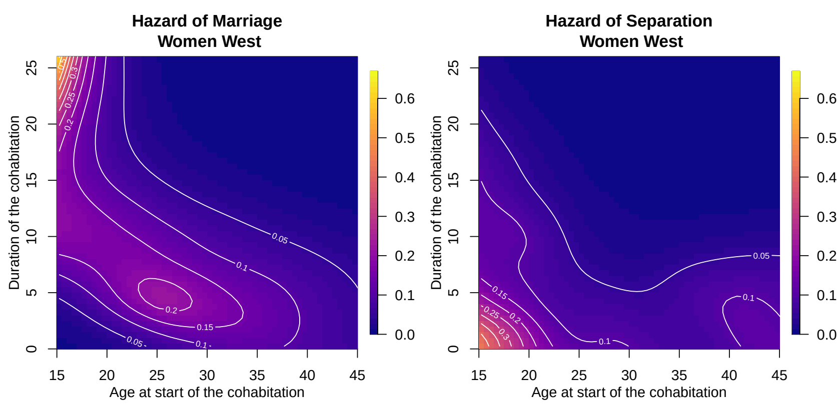

FIGURE 2. Cause-specific hazards of marriage (left panel) and separation (right panel) by age at entry into cohabitation and duration of cohabitation.

From these, we calculated the estimated cumulative incidence functions, shown in Figure 3.

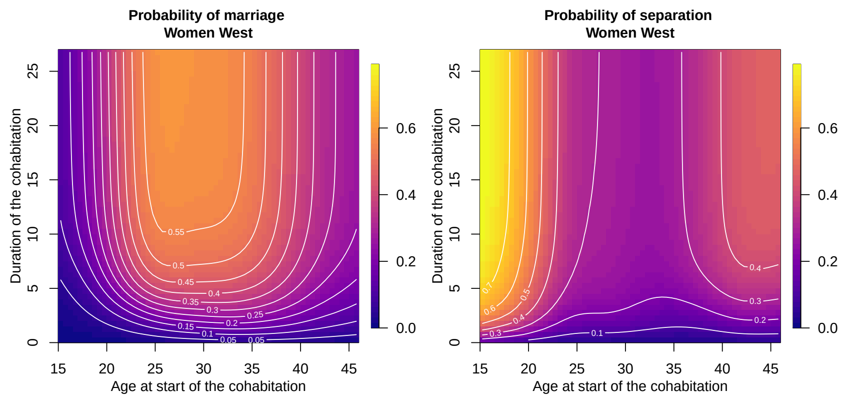

FIGURE 3. Cumulative incidence functions of marriage (left panel) and separation (right panel) by age at entry into cohabitation and duration of cohabitation.

References

Carollo, A., Putter, H., Eilers, P.H.C. and Gampe, J. (2022). Smooth hazards with multiple time scales. (preprint)

Currie, I.D., Durban, M. and Eilers, P.H.C. (2004). Smoothing and forecasting mortality rates. *Stat. Modelling*, **4**, 279-298.

Currie, I.D., Durban, M. and Eilers, P.H.C. (2006). Generalized linear array models with applications to multidimensional smoothing. *JRSS B*, **68**, 259-280.

Huinink, J. et al. (2011). Panel Analysis of Intimate Relationships and Family Dynamics (pairfam): Conceptual framework and design. *Zeitschrift f"ur Familienforschung*, **1/2011**, 77-101.

Lee, M., Gouskova, N.A., Feuer, E.J and Fine, J.P. (2017). On the choice of time scales in competing risks predictions. *Biostatistics*, **18**, 15 – 31.

Putter, H., Fiocco, M. and Geskus, R.B. (2007). Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine*, **26**, 2389 – 2430.